

Revised Edition

A Textbook of Engineering Physics

For the Students of B.E.,
B.Tech., B.Arch. and B.Sc. (Engg.)



Dr. M.N. AVADHANULU
Dr. P.G. KSHIRSAGAR

S. CHAND

A TEXTBOOK OF **ENGINEERING PHYSICS**

[For the Students of B.E., B.Tech., B.Arch., B.Sc., (Engg.)]

Dr. M.N. Avadhanulu

M.Sc., Ph.D.

*Ex-Principal, Om College of Engineering of Wardha
Former Professor and Head, Department of Physics*

Kavikulguru Institute of Technology & Science

Ramtek-441 106, Dist. Nagpur (M.S.)

and

Dr. P.G. Kshirsagar

M.Sc., Ph.D.

*Formerly Head of the Department of Applied Physics,
Visvesvaraya National Institute of Technology
NAGPUR*

[Revised Edition]



S. CHAND & COMPANY PVT. LTD.

(AN ISO 9001 : 2008 COMPANY)

RAM NAGAR, NEW DELHI-110 055



S. CHAND & COMPANY PVT. LTD.

(An ISO 9001 : 2008 Company)

Head Office: 7361, RAM NAGAR, NEW DELHI - 110 055

Phone: 23672080-81-82, 9899107446, 9911310888 Fax: 91-11-23677446

Shop at: schandgroup.com; e-mail: info@schandgroup.com

Branches :

AHMEDABAD	: 1st Floor, Heritage, Near Gujarat Vidhyapeeth, Ashram Road, Ahmedabad - 380 014, Ph: 27541965, 27542369, ahmedabad@schandgroup.com
BENGALURU	: No. 6, Ahuja Chambers, 1st Cross, Kumara Krupa Road, Bengaluru - 560 001, Ph: 22268048, 22354008, bangalore@schandgroup.com
BHOPAL	: Bajaj Tower, Plot No. 2&3, Lala Lajpat Rai Colony, Raisen Road, Bhopal - 462 011, Ph: 4274723, 4209587, bhopal@schandgroup.com
CHANDIGARH	: S.C.O. 2419-20, First Floor, Sector - 22-C (Near Aroma Hotel), Chandigarh -160 022, Ph: 2725443, 2725446, chandigarh@schandgroup.com
CHENNAI	: No.1, Whites Road, Opposite Express Avenue, Royapettah, Chennai - 600014 Ph. 28410027, 28410058, chennai@schandgroup.com
COIMBATORE	: 1790, Trichy Road, LGB Colony, Ramanathapuram, Coimbatore -6410045, Ph: 2323620, 4217136 coimbatore@schandgroup.com (Marketing Office)
CUTTACK	: 1st Floor, Bhartia Tower, Badambadi, Cuttack - 753 009, Ph: 2332580; 2332581, cuttack@schandgroup.com
DEHRADUN	: 1st Floor, 20, New Road, Near Dwarka Store, Dehradun - 248 001, Ph: 2711101, 2710861, dehradun@schandgroup.com
GUWAHATI	: Dilip Commercial (1st floor), M.N. Road, Pan Bazar, Guwahati - 781 001, Ph: 27388111, 2735640 guwahati@schandgroup.com
HYDERABAD	: Padma Plaza, H.No. 3-4-630, Opp. Ratna College, Narayanaguda, Hyderabad - 500 029, Ph: 27550194, 27550195, hyderabad@schandgroup.com
JAIPUR	: 1st Floor, Nand Plaza, Hawa Sadak, Ajmer Road, Jaipur - 302 006, Ph: 2219175, 2219176, jaipur@schandgroup.com
JALANDHAR	: Mai Hiran Gate, Jalandhar - 144 008, Ph: 2401630, 5000630, jalandhar@schandgroup.com
KOCHI	: Kachapilly Square, Mullassery Canal Road, Ernakulam, Kochi - 682 011, Ph: 2378740, 2378207-08, cochin@schandgroup.com
KOLKATA	: 285/J, Bipin Bihari Ganguli Street, Kolkata - 700 012, Ph: 22367459, 22373914, kolkata@schandgroup.com
LUCKNOW	: Mahabeer Market, 25 Gwynne Road, Aminabad, Lucknow - 226 018, Ph: 4076971, 4026791, 4065646, 4027188, lucknow@schandgroup.com
MUMBAI	: Blackie House, IInd Floor, 103/5, Walchand Hirachand Marg, Opp. G.P.O., Mumbai - 400 001, Ph: 22690881, 22610885, mumbai@schandgroup.com
NAGPUR	: Karnal Bagh, Near Model Mill Chowk, Nagpur - 440 032, Ph: 2720523, 2777666 nagpur@schandgroup.com
PATNA	: 104, Citicentre Ashok, Mahima Palace , Govind Mitra Road, Patna - 800 004, Ph: 2300489, 2302100, patna@schandgroup.com
PUNE	: 291, Flat No.-16, Ganesh Gayatri Complex, IInd Floor, Somwarpeth, Near Jain Mandir, Pune - 411 011, Ph: 64017298, pune@schandgroup.com (Marketing Office)
RAIPUR	: Kailash Residency, Plot No. 4B, Bottle House Road, Shankar Nagar, Raipur - 492 007, Ph: 2443142, Mb. : 09981200834, raipur@schandgroup.com (Marketing Office)
RANCHI	: Flat No. 104, Sri Draupadi Smriti Apartments, (Near of Jaipal Singh Stadium) Neel Ratan Street, Upper Bazar, Ranchi - 834 001, Ph: 2208761, ranchi@schandgroup.com (Marketing Office)
SILIGURI	: 122, Raja Ram Mohan Roy Road, East Vivekanandapally, P.O., Siliguri, Siliguri -734001, Dist., Jalpaiguri, (W.B.) Ph. 0353-2520750 (Marketing Office) siliguri@schandgroup.com
VISAKHAPATNAM	: No. 49-54-15/53/8, Plot No. 7, 1st Floor, Opp. Radhakrishna Towers, Seethammadhara North Extn., Visakhapatnam - 530 013, Ph-2782609 (M) 09440100555, visakhapatnam@schandgroup.com (Marketing Office)

© 1992, Dr. M.N. Avadhanulu and Dr. P.G. Kshirsagar

All rights reserved. No part of this publication may be reproduced or copied in any material form (including photo copying or storing it in any medium in form of graphics, electronic or mechanical means and whether or not transient or incidental to some other use of this publication) without written permission of the copyright owner. Any breach of this will entail legal action and prosecution without further notice.

Jurisdiction : All disputes with respect to this publication shall be subject to the jurisdiction of the Courts, tribunals and forums of New Delhi, India only.

First Edition 1992

Subsequent Editions and Reprints 1993, 94, 95, 98, 2000, 2001 (Twice), 2003, 2004, 2005, 2006 (Twice), 2007, 2008, 2009 (Twice), 2010, 2011 (Twice); 2012

Thoroughly Revised Edition 2014

ISBN : 81-219-0817-5

Code : 10B 131

PRINTED IN INDIA

**By Rajendra Ravindra Printers Pvt. Ltd., 7361, Ram Nagar, New Delhi -110 055
and published by S. Chand & Company Pvt. Ltd., 7361, Ram Nagar, New Delhi -110 055.**

Preface to the Revised Edition

A Textbook of Engineering Physics is originally designed to serve as a textbook as well as reference book for two semester course in Engineering Physics. The book is written with two distinct objectives: First to provide a single source of information and the second to present the principles of Physics as relevant to the B.E./B.Tech. students in an easy-to-understand style. In this edition, a new chapter number 40 namely “Geometrical Optics” has been added to make the book still more useful to the students. The requirements of the students are given priority and the material is moulded in a more student-friendly style. However, the spirit of Physics is not sacrificed at any stage and the expectations of teachers are held high at every step. It is generally felt that Physics is one more body of facts thrust on engineering students who are already burdened with a heavy syllabus and evolved through the efforts of rational thinkers who have been interested to know; why, what and how of natural phenomena.

Engineering has emerged as the application of their understanding for the benefit of human society at large. Thus Physics is the foundation on which stands the elaborate structure of technology. The main purpose of teaching Physics to Engineering undergraduates is to acquaint the budding engineers with the thread of development and the urge that underlies the presentation of the material in this book, so that they can apply this knowledge beneficially in their later pursuits.

The authors sincerely hope that this book will assist the students in learning the principles of Physics more effectively.

Enough care is taken to eliminate printing mistakes. However, some mistakes might have crept in inadvertently. The authors appeal to the readers to point out such left-out mistakes. The authors are also highly indebted to the teachers in various engineering institutions who have been extending unstinted support to this book.

M.N.AVADHANULU

mna2005@rediffmail.com

Disclaimer : While the authors of this book have made every effort to avoid any mistake or omission and have used their skill, expertise and knowledge to the best of their capacity to provide accurate and updated information. The authors and S. Chand does not give any representation or warranty with respect to the accuracy or completeness of the contents of this publication and are selling this publication on the condition and understanding that they shall not be made liable in any manner whatsoever. S.Chand and the author expressly disclaim all and any liability/responsibility to any person, whether a purchaser or reader of this publication or not, in respect of anything and everything forming part of the contents of this publication. S. Chand shall not be responsible for any errors, omissions or damages arising out of the use of the information contained in this publication.

Further, the appearance of the personal name, location, place and incidence, if any; in the illustrations used herein is purely coincidental and work of imagination. Thus the same should in no manner be termed as defamatory to any individual.

Preface to the Ninth Revised Edition

“A Textbook of Engineering Physics” is written with two distinct objectives : to provide a single source of information for engineering undergraduates of different specializations and provide them a solid base in physics. Successive editions of the book incorporated topics as required by students pursuing their studies in various universities. In this new edition the contents are fine-tuned, modernized and updated at various stages.

Physics is not an isolated body of theories which merely serve vocational usefulness. What has been achieved in physics has sooner or later made tremendous impact on the technological growth of our society. To become active participants in the technological revolution, one has to necessarily acquaint himself with the methods of science. Mechanical memorizing of certain definitions and derivations does not belong to the method of science and as such is of little value to the student. The main purpose of teaching physics to engineering undergraduates is to equip them with an understanding of the “scientific method”, so that they may use the training beneficially in their higher pursuits. An earnest attempt is made in this direction right from the first edition of this book by blending careful presentation of fundamental concepts and methods of physics.

This edition retains the original theme of emphasis on concepts with less mathematical formalism. The practical applications are discussed at each stage. The question bank given at the end of each chapter is updated. At a number of places, points for refinement are noticed and those have been incorporated. We have gladly received and carefully considered suggestions from professors and students who have used earlier editions. Further suggestions for improvement of the quality and quantity of the content are most welcome.

M.N.AVADHANULU
mna2005@rediffmail.com

Acknowledgement

The authors offer their special thanks to Smt. Nirmala Gupta, Chairperson & Managing Director, Shri Amit Gupta, C.E.O., Shri Naveen Joshi, Executive vice-president (Publishing), Shri Bhagirath Kaushik, Vice president (Sales and Marketing), S.Chand & Company Ltd. and Shri Vijay, Branch Manager, Nagpur and their dedicated team for all their efforts in bringing out this book nicely and in time.

M.N.AVADHANULU
mna2005@rediffmail.com

Dedicated to
My Mother,
Maternal Uncles
Shri Mullapudi Suryanarayana,
Dr. Mullapudi Subba Rao and
Shri Mullapudi Satyanarayana,
and to
My Wife Suvarchala

Books are not paper and words but interaction with
thinkers on a one-to-one basis, not of one generation but
separated by hundreds and thousands of years

—Thomas Carlyle

Contents

<i>Chapters</i>	<i>Pages</i>
1. OSCILLATIONS AND WAVES	1 – 37
1.1 Introduction 1; 1.2 Oscillations 1; 1.3 Simple Harmonic Motion 2; 1.4 Free Oscillations 9; 1.5 Damped Oscillations 10; 1.6 Forced Oscillations 13; 1.7 Resonance 15; 1.8 Coupled Oscillations 16; 1.9 Waves 17; 1.10 Types of Waves 21; 1.11 Reflection and Transmission of Waves at a Boundary 23; 1.12 Principle of Superposition 26; 1.13 Stationary Waves 28; 1.14 Superposition of two Perpendicular Shms 31; 1.15 Dispersion 34;	
2. ELECTROSTATICS	38 - 64
2.1 Introduction 38; 2.2 Electric Charges 38; 2.3 Coulomb's Law 38; 2.4 Principle of Superposition 40; 2.5 Electric Field 40; 2.6 Computation of Electric Field in Some Specific Cases 41; 2.7 Electrostatic Potential 46; 2.8 Equipotential Surfaces 49; 2.9 Electric Field is a Conservative Field 50; 2.10 Potential at a Point Due to a Group Of Point Charges 51; 2.11 Computation of Electric Potential in Some Specific Cases 51; 2.12 Flux 53; 2.13 Solid Angle 54; 2.14 Gauss' Law of Electrostatics in Free Space 54; 2.15 Divergence of Electric Field 55; 2.16 Differential Form of Gauss's Law 56; 2.17 Derivation of Coulomb's Law From Gauss Law 56; 2.18 Applications of Gauss's Law 57; 2.19 Gauss' Law of Electrostatics in a (Dielectric) Medium 61; 2.20 Electric Displacement Vector 62	
3 MAGNETOSTATICS AND ELECTRODYNAMICS	65 – 78
3.1 Magnetic Field 65; 3.2 Magnetic Flux Density 66; 3.3 Biot-Savart Law 66; 3.4 Ampere's Law 67; 3.5 Gauss's Law for Magnetism 68; 3.6 Magnetic Scalar Potential 69; 3.7 Magnetic Vector Potential 69; 3.8 Faraday's Laws of Induction 70; 3.9 Lenz's Law 71; 3.10 Integral Form of Faraday's Law 72; 3.11 Equation of Continuity 73; 3.12 Displacement Current 74; 3.13 Maxwell's Equations 76; 3.14 Maxwell's Equations in Integral Form 77	
4. ELECTROMAGNETIC WAVES	79 – 93
4.1 Introduction 79; 4.2 Electromagnetic Waves 79; 4.3 Electromagnetic Wave Equations 80; 4.4 Maxwell's Wave Equations for Free Space 81; 4.5 Uniform Plane Waves 82; 4.6 Electromagnetic Energy Density 84; 4.7 The Poynting Theorem 85; 4.8 The Poynting Vector 86; 4.9 Wave Propagation in a Lossy Medium 87; 4.10 Conductors and Dielectrics 88	
5. LIGHT	94 – 130
5.1 Introduction 94; 5.2 Nature of Light 94; 5.3 The Velocity of Light 95; 5.4 Optical Medium 95; 5.5 Homogeneous Isotropic Medium 97; 5.6 Reflection and Refraction	

98; 5.7 Total Internal Reflection 99; 5.8 Reflectivity and Transmissivity 100; 5.9 Absorption 101; 5.10 Wave Front and the Ray 102; 5.11 Mathematical Representation of a Plane Wave 103; 5.12 Light is an Electromagnetic Wave 106; 5.13 Visible Range 111; 5.14 Optical Path Length 112; 5.15 Phase Change and path Difference 113; 5.16 The Principle of Superposition 114; 5.17 Interference of Light Waves 115; 5.18 Young's Double Slit Experiment 120; 5.19 Wave Trains—Light From Common Sources 121; 5.20 Coherence 122; 5.21 Double Slit Experiment Again 126; 5.22 Dispersion 127; 5.23 Scattering 128	131 – 172
6. INTERFERENCE	131 – 172
6.1 Introduction 131; 6.2 Interference 131; 6.3 Conditions for Observing Sustained Interference 133; 6.4 Techniques Of Obtaining Interference 133; 6.5 Review of Important Concepts 134; 6.6 Fresnel Biprism 135; 6.7 Thin Film Interference 141; 6.8 Plane Parallel Film 142; 6.9 Variable Thickness (Wedge-Shaped) Film 146; 6.10 Colours in Thin Films 151; 6.11 Newton's Rings 151; 6.12 Applications of Interference 158; 6.13 Michelson's Interferometer 163; 6.14 Applications of Michelson Interferometer 166; 6.15 Moire Fringes 168	
7. DIFFRACTION	173 – 197
7.1 Introduction 173; 7.2 Diffraction 173; 7.3 Distinction Between Interference and Diffraction 175; 7.4 The two Types of Diffraction 175; 7.5 Fraunhofer Diffraction at a Single Slit 176; 7.6 Fraunhofer Diffraction at Double Slit 182; 7.7 Diffraction Due to N-Slits—Diffraction Grating (Normal Incidence) 186; 7.8 Plane Diffraction Grating - Theory 186; 7.9 Resolving Power 193; 7.10 Resolving Power of a Plane Transmission Grating 194;	
8. POLARIZATION	198 – 236
8.1 Introduction 198; 8.2 Polarization 198; 8.3 Unpolarized and Polarized Light 199; 8.4 Natural Light is Unpolarized Light 200; 8.5 Types of Polarization 201; 8.6 Production of Plane Polarized Light 204; 8.7 Polaroid Sheets 209; 8.8 Polarizer and Analyzer 209; 8.9 Malus' Law 211; 8.10 Anisotropic Crystals 212; 8.11 Double Refraction in Calcite Crystal 214; 8.12 Nicol Prism 217; 8.13 Effect of Polarizer on Light of Different Polarizations 219; 8.14 Phase Difference Between E-Ray and O-Ray 219; 8.15 Superposition of Waves Linearly Polarised at Right Angles 221; 8.16 Retarders 224; 8.17 Production of Elliptically Polarized Light 227; 8.18 Production of Circularly Polarized Light 228; 8.19 Analysis of Polarized Light 229; 8.20 Applications of Polarized Light 230	
9. OPTICAL ACTIVITY	237 – 252
9.1 Introduction 237; 9.2 Optical Rotation 237; 9.3 Specific Rotation 238; 9.4 Fresnel's Explanation 238; 9.5 Polarimeter 239; 9.6 Electro-Optic and Magneto-Optic Effects 242; 9.7 Electro-Optic Effects 242; 9.8 Magneto-Optic Effects 244; 9.9 Anisotropy Induced by Mechanical Strain 245; 9.10 Photoelasticity 245	
10 OPTICAL FIBRES	253 – 295
10.1 Introduction 253; 10.2 Optical Fibre 53; 10.3 Total Internal Reflection 257; 10.4 Propagation of Light Through an Optical Fibre 257; 10.5 Fractional Refractive Index Change 261; 10.6 Numerical Aperture 262; 10.7 Skip Distance and Number of Total Internal Reflections 263; 10.8 Modes of Propagation 264; 10.9 Types of Rays 265; 10.10 Classification of Optical Fibres 266; 10.11 The Three Types of Fibres 267;	

10.12 Materials 269; 10.13 V-Number 270; 10.14 Fabrication 273; 10.15 Splicing 273; 10.16 Losses in Optical Fibre 275; 10.17 Bandwidth 283; 10.18 Characteristics of the Fibres 283; 10.19 Applications 285; 10.20 Fibre Optic Communication System 287; 10.21 Merits of Optical Fibres 289; 10.22 Fibre Optic Sensors 289	
11. ARCHITECTURAL ACOUSTICS	296 – 321
11.1 Introduction 296; 11.2 Sound 296; 11.3 Classification of Sound 298; 11.4 Characteristics of Musical Sound 298; 11.5 Weber-Fechner Law 299; 11.6 Sound Intensity Level - Decibel 300; 11.7 Human Audiogram 302; 11.8 Phon 302; 11.9 Sound Reflection 303; 11.10 Reverberation Time 304; 11.11 Sound Absorption 305; 11.12 Sabine's Formula for Reverberation Time 306; 11.13 Reverberation Theory 307; 11.14 Determination of Absorption Coefficient 311; 11.15 Factors Affecting Acoustics of Buildings and their Remedies 312; 11.16 Acoustic Design of a Hall 315	
12. ULTRASONICS	322 – 344
12.1 Introduction 322; 12.2 Production of Ultrasonic Waves 322; 12.3 Piezoelectric Effect 324; 12.4 Detection of Ultrasonic Waves 326; 12.5 Properties of Ultrasonic Waves 327; 12.6 Cavitations 327; 12.7 Types of Ultrasonic Waves 327; 12.8 Determination of Velocity of Ultrasonic Waves 328; 12.9 Measurement of Elastic Constants in Liquids 330; 12.10 Determination of Velocity of Ultrasonic Waves in Solids 331; 12.11 Measurement of Elastic Constants in Solids 332; 12.12 Industrial Applications 332; 12.13 Ultrasonic Testing 335; 12.14 Modes of Display 337; 12.15 Medical Applications—Sonography 338; 12.16 Ultrasound Scanner 339; 12.17 Ultrasonic Blood Flow Meter 341; 12.18 Other Medical Applications 342	
13. ELECTRON EMISSION	345 – 351
13.1 Introduction 345; 13.2 Work Function 345; 13.3 Electron Emission 346; 13.4 Thermionic Emission 347; 13.5 Photoelectric Emission 349; 13.6 Field Emission 350; 13.7 Secondary Emission 350	
14. ELECTRON BALLISTICS	352 – 381
14.1 Introduction 352; 14.2 Electric Field 352; 14.3 Motion of an Electron in a Uniform Electric Field 353; 14.4 Uniform Magnetic Field 360; 14.5 Motion of an Electron in a Uniform Magnetic Field 361; 14.6 Magnetostatic Deflection 366; 14.7 Lorentz Equation 368; 14.8 Crossed Electric and Magnetic Field Configuration 368; 14.9 Velocity Selector 369; 14.10 Parallel Electric and Magnetic Field Configuration 370; 14.12 Charge of the Electron 372; 14.13 Mass of the Electron 375; 14.14 Radius of the Electron 375; 14.15 Positive Rays 376; 14.16 Thomson's Parabola Method 377	
15. ELECTRON OPTICS	382 – 403
15.1 Introduction 382; 15.2 Bethe's Law 382; 15.3 Electron Lens 384; 15.4 Focussing by Uniform Magnetic Fields 386; 15.5 Focusing by Axially Symmetric Magnetic Field 387; 15.6 Cathode Ray Tube 388; 15.7 Electromagnetic Deflection Type Crt 392; 15.8 Cathode Ray Oscilloscope 393; 15.9 Applications 398; 15.10 Other Applications of an Electron Beam 401; 15.11 Motion of Charged Particles in a Nonuniform Magnetic Field 401; 15.12 The Magnetic Bottle 402	
16. ELEMENTS OF THERMODYNAMICS	404 – 429
16.1 Introduction 404; 16.2 Concept Of Temperature 404; 16.3 Heat 405; 16.4 Thermodynamics 406; 16.5 Terminology 407; 16.6 Work 411; 16.7 Heat in Thermodynamics 414; 16.8 Comparison of Heat and Work 414; 16.9 Internal Energy	

415; 16.10 Law of Conservation of Energy 415; 16.11 First Law of Thermodynamics 416; 16.12 Applications of the First Law 416; 16.13 Heat Engine 418; 16.14 The Carnot Cycle 420; 16.15 Heat Pump 423; 16.16 Second Law of Thermodynamics 425; 16.17 Entropy 426; 16.18 Third Law of Thermodynamics 428	
17. THERMOELECTRICITY	430 – 449
17.1 Introduction 430; 17.2 Seebeck Effect 430; 17.3 Thermocouple 431; 17.4 Thermoelectric Series 432; 17.5 Variation of Thermoelectric E.M.F. With Temperature 432; 17.6 The Peltier Effect 434; 17.7 The Thomson Effect 435; 17.8 E.M.F. in a Thermocouple 438; 17.9 The Thermoelectric Power 438; 17.11 Relation Between Thomson Coefficient and Thermoelectric Power 440; 17.12 The Thermoelectric Laws 443; 17.13 Applications of Thermocouple 443; 17.14 Figure-Of-Merit, Z 444; 17.15 Thermoelectric Power Generation 445; 17.16 Thermoelectric Cooling 446; 17.17 The Thermoelectric Coolers 446	
18. SPECIAL THEORY OF RELATIVITY	450 – 482
18.1 Introduction 450; 18.2 Space, Time And Motion 450; 18.3 Frame of Reference 451; 18.4 Inertial Frames of Reference 451; 18.5 Non-Inertial Reference Frame 452; 18.6 Galileo's Principle Of Relativity 452; 18.7 Galilean Transformations 452; 18.8 The Ether 455; 18.9 Michelson-Morley Experiment 455; 18.10 Failure of Galilean Transformations 458; 18.11 Einstein's Principle of Relativity 459; 18.12 The Lorentz Transformations 460; 18.13 Consequences of Special Relativity 466; 18.14 Simultaneity of Events 466; 18.15 Length Contraction 468; 18.16 the Time Dilation 470; 18.17 The Twin Paradox 473; 18.18 The Relativistic Mass 474; 18.19 The Relativistic Momentum 475; 18.20 Kinetic Energy 475; 18.21 Mass-Energy Equivalence 476; 18.22 Relation Between Momentum and Energy 477	
19. ATOMIC PHYSICS	483 – 552
19.1 Introduction 483; 19.2 Wave-Picture of Radiation—Energy Flow is Continuous 484; 19.3 Blackbody Radiation 484; 19.4 Planck's Quantum Hypothesis – Energy is Quantized 489; 19.5 Particle Picture of Radiation – Radiation Is A Stream of Photons 490; 19.6 Photoelectric Effect 492; 19.7 X-Rays 495; 19.8 Generation of X-Rays 495; 19.9 X-Ray Spectrum 496; 19.10 Origin of Continuous X-Ray Spectrum 497; 19.12 Compton Scattering 499; 19.13 Pair Production 506; 19.14 Wave-Particle Duality 507; 19.15 Spectral Lines 508; 19.16 Atomic Structure 510; 19.17 Bohr's Model Of Atom 510; 19.18 Frank-Hertz Experiment 513; 19.19 Energy Level Diagram 515; 19.20 Electron Shells 517; 19.21 Characteristic X-Ray Spectrum 518; 19.22 Moseley's Law 519; 19.23 The Sommerfeld Relativistic Atom Model 522; 19.24 The Vector Atom Model 526; 19.25 Applications of the Vector Atom Model 535; 19.27 Zeeman Effect 538; 19.28 The Stern-Gerlach Experiment 543; 19.29 Anomalous Zeeman Effect 545; 19.30 Paschen-Back Effect 547; 19.31 Stark Effect 548	
20. QUANTUM MECHANICS	553 – 616
20.1 Introduction 553; 20.2 De Broglie Hypothesis 554; 20.3 De Broglie's Justification of Bohr's Postulate 555; 20.4 De Broglie Waves are Insignificant in Case of Macro-Bodies 557; 20.5 Properties of Matter Waves 558; 20.6 Davisson–Germer Experiment 558; 20.8 Velocity of De Broglie Waves 560; 20.9 Wave Packet – Represents a Microparticle 561; 20.10 Applications of De Broglie Waves 564; 20.11 Heisenberg Uncertainty Principle 568; 20.12 Elementary Proof of Uncertainty Principle Using De Broglie Wave Concept 571; 20.13 Implication of Uncertainty	

Principle 571; **20.14** Uncertainty Principle is Not Significant in Case of Macro-Bodies 572; **20.15** Thought Experiments 573; **20.16** Applications of Uncertainty Principle 574; **20.17** Wave Function and Probability Interpretation 577; **20.18** Schrödinger Wave Equation 579; **20.19** The Free Particle 582; **20.20** Potential Energy Step 584; **20.21** Rectangular Potential Barrier 586; **20.22** Infinite Potential Well 591; **20.23** Extension to Three-Dimensional Case 596; **20.24** Harmonic Oscillator 601; **20.25** The Wave Mechanical Model of Atom 601; **20.26** The Transition From Deterministic to Probabilistic Nature 605; **20.27** Superposition Principle 606; **20.28** Observables and Operators 606; **20.29** Important Operators of Quantum Mechanics 608; **20.30** Expectation Values 609

21. ATOMIC NUCLEUS AND NUCLEAR ENERGY **617 – 658**

21.1 Introduction 617; **21.2** The Atomic Nucleus 617; **21.3** Isotopes 618; **21.4** The Nuclear Force 618; **21.5** Static Properties of Nucleus 620; **21.6** Mass Defect 621; **21.7** Binding Energy 622; **21.8** Nuclear Models 624; **21.9** Natural Radioactivity 626; **21.10** Radioactive Decay 626; **21.11** Radioactive Series 627; **21.12** Law of Radioactive Decay 627; **21.13** Activity 629; **21.14** Half-Life 629; **21.15** Average Life Time 630; **21.16** Units of Activity 631; **21.17** Induced Radioactivity 632; **21.19** Nuclear Reactions 635; **21.20** Q-Value 636; **21.21** Nuclear Reaction Cross-Section 638; **21.22** Neutrons and Neutron Induced Reactions 640; **21.23** Nuclear Fission 641; **21.24** Nuclear Chain Reaction 644; **21.25** Nuclear Energy 646; **21.26** Nuclear Reactors 648; **21.27** Nuclear Power Plant 651; **21.28** Nuclear Fusion 651; **21.30** Controlled Thermonuclear Reactions 653; **21.31** Fusion Reactor 655

22. COSMIC RAYS AND ELEMENTARY PARTICLES **659 – 669**

22.1 Introduction 659; **22.2** Primary Cosmic Rays 659; **22.3** Secondary Osmic Rays 660; **22.4** Origin of Cosmic Rays 660; **22.5** Altitude Effect 660; **22.6** Latitude Effect 661; **22.7** Longitude Effect 662; **22.8** East-West Effect 662; **22.9** The Positron 662; **22.10** Pair Production 662; **22.11** Cosmic Ray Showers 663; **22.12** The Mesons 663; **22.13** Elementary Particles 664; **22.14** Classification of Elementary Particles 664; **22.15** Basic Forces in Nature 664; **22.16** Classification of Elementary Particles Basing on The Basic Forces 665; **22.17** Antiparticles 666; **22.18** Leptons 666; **22.19** Hadrons 667; **22.20** Resonances 668; **22.21** The Quark Model 668; **22.22** Other Models 669;

23. NUCLEAR INSTRUMENTS **670 – 700**

23.1 Introduction 670; **23.2** Geiger-Muller Counter 670; **23.3** The Wilson Cloud Chamber 672; **23.4** Bubble Chamber 674; **23.5** Spark Chamber 675; **23.6** Scintillation Counter 675; **23.7** Solid State Detectors 676; **23.8** Cerenkov Detector 676; **23.9** Mass Spectrographs 677; **23.10** Aston Mass Spectrograph 677; **23.11** Dempster Mass Spectrograph 680; **23.12** Bainbridge Mass Spectrograph 682; **23.13** Particle Accelerators 684; **23.14** Drift Tube Accelerator 685; **23.15** Cyclotron 687; **23.16** Synchrocyclotron 692; **23.17** Betatron 693; **23.18** Electron Synchrotron 697; **23.19** Proton Synchrotron 698

24. LASERS **701 – 738**

24.1 Introduction 701; **24.2** Interaction of Light With Matter and the Three Quantum Processes 701; **24.3** Einstein Coefficients and their Relations 706; **24.4** Light Amplification 708; **24.5** Meeting the Three Requirements 709; **24.6** Components of Laser 711; **24.7** Lasing Action 712; **24.8** Pumping Methods 713; **24.9** Threshold

Condition for Lasing 715; 24.10 Modes of the Laser Beam 717; 24.11 Types of Lasers 719; 24.12 Laser Beam Characteristics 732; 24.13 Applications 733	
25. HOLOGRAPHY	739 – 753
25.1 Introduction 739; 25.2 Principle of Holography 740; 25.3 Coaxial Holography 741; 25.4 Off-Axis Holography 742; 25.5 Theory 742; 25.6 Holograms 743; 25.7 Important Properties of a Hologram 744; 25.8 Classification of Holograms 745; 25.9 Applications 749; 25.10 Medical Applications of Holography 752;	
26. CRYSTAL STRUCTURES	754 – 797
26.1 Introduction 754; 26.2 Classification of Solids 754; 26.3 Space Lattice 756; 26.4 Crystal Structure 757; 26.5 Unit Cell 757; 26.6 Bravais Lattices 758; 26.7 Symmetries in Crystals 761; 26.8 Calculation of Parameters of a Cubic Lattice 763; 26.9 Body Centred Cubic (Bcc) Cell 766; 26.10 Face Centred Cubic (Fcc) Cell 768; 26.11 Hcp Structure 770; 26.12 Atom Positions in Cubic Unit Cells 772; 26.13 Indices of Crystallographic Direction 773; 26.14 Lattice Planes and Miller Indices 774; 26.15 Interplanar Spacing in a Cubic Lattice 776; 26.16 Atomic Packing 778; 26.17 Voids 781; 26.18 Ionic Solids 782; 26.19 Diamond Cubic Structure 783; 26.20 Zns Structure 784; 26.21 Polymorphism and Allotropy 784; 26.22 Graphite Structure 784; 26.23 Crystal Structure Analysis 785; 26.24 Braggs' Law 786; 26.25 Braggs' Spectrometer 789; 26.26 Powder Crystal Method 792; 26.27 Rotating Crystal Method 793	
27. CRYSTAL DEFECTS	798 – 813
27.1 Introduction 798; 27.2 Crystal Defects 798; 27.3 Point Defects 798; 27.4 Vacancies 799; 27.5 Energy Of Formation of Vacancy in a Metallic Crystal 799; 27.6 Schottky Defect 801; 27.7 Interstitials 802; 27.8 Equilibrium Concentration of Schottky Defects in an Ionic Crystal 802; 27.9 Frenkel Defect 804; 27.10 Equilibrium Concentration of Frenkel Defects in an Ionic Crystal 804; 27.11 Impurities 807; 27.12 Electronic Defects 808; 27.13 Effect of Point Defects 808; 27.14 Line Defects 808; 27.15 Burgers Vector 810; 27.16 Planar Defects or Surface Defects 811; 27.17 Volume Defects 812	
28. CONDUCTORS	814 – 834
28.1 Introduction 814; 28.2 Electrical Conduction 814; 28.3 Classification of Materials 815; 28.4 Free Electron Model of Solids 816; 28.5 Classical Free Electron Theory of Metals 817; 28.6 Drift Velocity 818; 28.7 Electrical Conductivity 818; 28.8 Mobility 820; 28.9 Relaxation Time 820; 28.10 Thermal Conductivity 821; 28.11 Wiedemann-Franz Law 822; 28.12 Lorentz Number 823; 28.13 Resistance 824; 28.14 Drawbacks of Classical Free Electron Theory 825; 28.15 Quantum Free Electron Theory 826; 28.16 Density of Energy States 826; 28.17 Carrier Concentration in Metals 829; 28.18 Fermi Energy, E_f 830; 28.19 Fermi-Dirac Distribution Function 831; 28.20 Quantum Free Electron Theory of Electrical Conduction 831; 28.21 Failure of Quantum Free Electron Theory 833	
29. BAND THEORY OF SOLIDS	835 – 852
29.1 Introduction 835; 29.2 The Band Theory of Solids—A Qualitative Explanation 835; 29.3 The Band Theory of Solids—Quantum Mechanical Explanation 837; 29.4 Energy Band Structure of a Solid 838; 29.5 Electrical Conduction From the View Point of Band Theory 840; 29.6 Energy Band Diagram 840; 29.7 Classification of	

Solids 840; 29.8 Energy Band Diagrams for Some Typical Solids 842; 29.9 Energy Band Structure of a Conductor 843; 29.10 Energy Band Structure of an Insulator 849; 29.11 Energy Band Structure of a Semiconductor 849; 29.12 Effective Mass 850;	
30. SEMICONDUCTORS	853 – 900
30.1 Introduction 853; 30.2 Crystal Structure 853; 30.3 Intrinsic Semiconductor 854; 30.4 Correlation Between Crystal Lattice and Energy Band Descriptions 856; 30.5 Holes 857; 30.6 Generation and Recombination 859; 30.7 Intrinsic Conductivity 860; 30.8 Carrier Concentrations 861; 30.9 Intrinsic Carrier Concentration 864; 30.10 The Fraction of Electrons in the Conduction Band 866; 30.11 Fermi Level in Intrinsic Semiconductor 867; 30.12 Variation of Intrinsic Conductivity with Temperature 870; 30.13 Determination of Band Gap 871; 30.14 Limitations of Intrinsic Semiconductor 872; 30.15 Extrinsic Semiconductors 872; 30.16 N-Type Semiconductor 873; 30.17 P-Type Semiconductor 877; 30.18 Band Diagrams of Extrinsic Semiconductors at 0K and 300K 880; 30.19 Extrinsic Conductivity 880; 30.20 Law of Mass Action 882; 30.21 Charge Neutrality Condition 883; 30.23 Fermi Level in Extrinsic Semiconductors 885; 30.24 Variation of Fermi Level with Impurity Concentration 886; 30.25 Drift and Diffusion Currents 887; 30.26 Minority Carrier Diffusion 889; 30.27 Compound Semiconductors 890; 30.28 Hall Effect 891	
31. SEMICONDUCTOR DIODES	901 – 937
31.1 Introduction 901; 31.2 P-N Junction Diode 901; 31.3 P-N Junction Under Forward Bias 909; 31.4 P-N Junction Under Reverse Bias 913; 31.5 The Diode Equation 914; 31.6 Voltage-Ampere Characteristic 915; 31.7 Applications 917; 31.8 Zener Diode 919; 31.9 Varactor Diode 921; 31.10 Light Emitting Diode (Led) 923; 31.11 Photodetectors 925; 31.12 Solar Cell 930; 31.13 Light Sources for Fiber Optic Systems 933	
32. BIPOLAR JUNCTION TRANSISTOR	938 – 951
32.1 Introduction 938; 32.2 Transistor Structure 938; 32.3 Schematic Representation 939; 32.4 Formation of Depletion Regions 939; 32.5 Energy Band Diagram of Unbiased Transistor 940; 32.6 Biasing the Transistor 941; 32.7 Circuit Configurations 941; 32.8 Action of the Bias 942; 32.9 Transistor Action 943; 32.10 Roles of Emitter, Base and Collector 945; 32.11 Relation Between Currents in CB Configuration 946; 32.12 Energy Band Diagram of a Transistor Biased in Normal Mode 947; 32.13 Common Emitter Configuration 948; 32.14 Current Relations in CE Configuration 949; 32.15 Transistor as an Amplifier 950	
33. DIELECTRICS	952 – 993
33.1 Introduction 952; 33.2 Dielectrics 952; 33.3 Dielectric Constant 953; 33.4 Dielectric Polarization 953; 33.5 Gauss Law 954; 33.6 Dielectric Susceptibility 955; 33.7 The Three Field Vectors 955; 33.8 Relation Between E_R And X 956; 33.9 Relation Between P And E 956; 33.10 Induced Dipoles 957; 33.11 Permanent Dipoles 958; 33.12 Nonpolar And Polar Dielectrics 959; 33.13 Polarization-An Atomic View 960; 33.14 Types Of Polarization 961; 33.15 Temperature Dependence Of Polarization 968; 33.16 Frequency Dependence Of Total Polarization 969; 33.17 The Internal Field In Solids 970; 33.18 Lorentz Field 971; 33.19 Clausius-Mosotti Equation 973; 33.20 Dielectric Loss 974; 33.21 Dielectric Breakdown 978; 33.22 Applications 979; 33.23 Piezoelectricity 981; 33.24 Ferroelectricity 984; 33.25 Pyroelectricity 988; 33.26 Materials 988; 33.27 Applications 989	
34. MAGNETIC MATERIALS	994 – 1030
34.1 Introduction 994; 34.2 Terms and Definitions 994; 34.3 Relation Between M_R and X 996; 34.4 Origin of Magnetization 996; 34.5 Classification of Magnetic	

Materials 997; 34.6 Diamagnetic Materials 997; 34.7 Paramagnetic Materials 1001; 34.8 Ferromagnetic Materials 1005; 34.9 Magnetostriction 1014; 34.10 Antiferromagnetism 1014; 34.11 Ferrimagnetism 1015; 34.12 Ferrites 1016; 34.13 Hysteresis Loss 1016; 34.14 Soft and Hard Magnetic Materials 1017; 34.15 Magnetic Materials and their Applications 1019; 34.16 Magnetic Devices 1021	
35. SUPERCONDUCTIVITY	1031 – 1052
35.1 Introduction 1031 35.2 Superconductivity 1031; 35.3 Materials (Low T_c Materials) 1032; 35.4 Properties of Superconductors 1034; 35.5 Other External Factors that Affect Superconductivity 1040; 35.6 Type-I and Type-II Superconductors 1040; 35.7 BCS Theory 1042; 35.8 Josephson Effect 1043; 35.9 High Superconductors 1045; 35.10 Applications 1046;	
36. MODERN ENGINEERING MATERIALS	1053 – 1080
36.1 Introduction 1053; 36.2 Metallic Glasses 1053; 36.3 Liquid Crystals 1058; 36.4 Shape Memory Alloys 1065; 36.5 Biomaterials 1076;	
37. NON DESTRUCTIVE TESTING	1081 – 1096
37.1 Introduction 1081; 37.2 Types of Defects 1081; 37.3 Methods of NDT 1082; 37.4 Visual Inspection 1082; 37.5 Liquid/Dye Penetrant Testing 1082; 37.6 Magnetic Particle Testing 1084; 37.7 Eddy Current Testing 1085; 37.8 Ultrasonic Inspection Method 1085; 37.9 Advantages 1092; 37.10 X-Ray Radiography 1092; 37.11 X-Ray Fluoroscopy 1095; 37.12 Comparison of Conventional and Real-Time Radiography 1095;	
38. VACUUM TECHNOLOGY	1097 – 1109
38.1 Introduction 1097; 38.2 Vacuum 1097; 38.3 Units of Vacuum 1097; 38.4 Vacuum Ranges 1098; 38.5 Production of Vacuum 1098; 38.6 Classification of Vacuum Pumps 1098; 38.7 Rotary Oil Pumps 1098; 38.8 Diffusion Pump 1100; 38.9 Turbomolecular Pumps 1101; 38.10 Cryopumps 1102; 38.11 Vacuum Gauges 1103; 38.11.1 Thermocouple Gauge 1103; 38.12 Vacuum Technology 1106; 38.13 Applications of Vacuum 1106; 38.14 High Vacuum Systems 1107; 38.15 Thin Film Deposition 1108;	
39. NANOTECHNOLOGY	1110 – 1154
39.1 Introduction 1110; 39.2 Nanoscale 1111; 39.3 The Significance of the Nanoscale 1111; 39.4 Nanotechnology 1112; 39.5 What is Molecular Nanotechnology? 1112; 39.6 Nanotechnologies in the Past 1113; 39.7 Four Generations of Nanotechnology Development 1114; 39.8 Why Nanotechnology? 1114; 39.9 Production Techniques 1115; 39.10 Tools 1118; 39.11 Nanomaterials 1121; 39.12 Nanolayers 1121; 39.13 Nanoparticles 1126; 39.14 Applications of Nanomaterials 1134; 39.15 Carbon Nanomaterials 1137; 39.16 Fullerenes 1137; 39.17 Carbon Nanotubes 1139; 39.18 Nanowires 1144; 39.19 Quantum Dots 1145; 39.20 Dendrimers 1146; 39.21 Nanocomposites 1146 39.22 Scaling Laws 1147 39.23 Nano Devices and Nanomachines 1153	
40. GEOMETRICAL OPTICS	1155 – 1189
40.1 Introduction 1155; 40.2 Thin Lenses 1155; 40.3 Coaxial Lens Systems 1158; 40.4 Cardinal Points 1159; 40.5. Definitions and Properties of Cardinal Points And Planes 1159; 40.6 Construction of Image Using Cardinal Points 1163; 40.7 Nodal Slide 1165; 40.8 Equivalent Focal Length of A Coaxial System of Two Thin Lenses 1168; 40.9 Cardinal Points of A Coaxial System of Two Thin Lenses 1169; 40.10 Eyepieces 1171; 40.11 Huygens Eyepiece 1172; 40.12 Ramsden Eyepiece 1175; 40.13 Comparison Of Ramsden Eyepiece With Huygens Eyepiece 1178	

CHAPTER

1

Oscillations and Waves

1.1 INTRODUCTION

Oscillatory motion is a repeating motion, which occurs extensively in nature. Oscillations (vibrations) and waves pervade all sciences. We are familiar with several general examples of oscillatory motion such as fluttering of tree leaves in a gentle breeze, swinging of a swing, the beating of the heart etc. Vibrations of atoms in a solid, electric and magnetic fields in light waves and radio waves, large structures swaying due to earth quake etc have an oscillatory nature. Electrical and mechanical oscillations as well as vibrations in structures are every day topics in the world of engineering. A repeating and periodic disturbance (oscillation) moving through a medium from one location to another gives rise to a wave. Waves are also encountered extensively in sciences and technology. Water waves, waves on a string, sound waves, radio waves, microwaves, light waves, and earthquake waves are a few of the examples. Therefore, the study of oscillations and waves constitutes the core topic in engineering and technology.

1.2 OSCILLATIONS

We all know that when a constant force acts on a body, the body moves with a *constant acceleration*. Rectilinear motion and uniform circular motion are examples of such a motion. However, when the force acting on the body varies in time, the acceleration of the body will also change with time. *Oscillatory motion is one type of motion that can occur when a body is subjected to a force that varies in time.*

The first systematic observations of oscillations were made by Galileo, who had observed certain rhythm in the swinging of chandeliers in the cathedral at Pisa. He determined the time taken for the oscillations with the help of his pulse. He found to his surprise that the time taken by each oscillation was constant. Later, this property of constant time period for oscillations was exploited in making pendulum clocks.

A pendulum serves as the simplest example of an oscillating body (Fig. 1.1). When a pendulum bob is at the lowest position, the pendulum is in a *state of rest* or it is said to be in its *equilibrium position*. The forces that act on the pendulum bob are the force of gravity and the force of tension in the string. When the bob is displaced from this equilibrium position, the gravitational force gives rise to a vertical component, directed along the string and a tangential component, directed

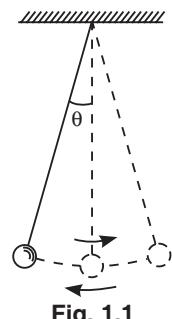


Fig. 1.1

perpendicular to the string (see Fig. 1.6). The tangential component pulls the bob continually back to the equilibrium position and therefore it is known as the **restoring force**. The bob will not come to the state of rest immediately. When the bob reaches the midpoint, the restoring force vanishes. However, the **inertia** property causes the bob to overshoot the equilibrium position and the motion continues. Once again, the restoring force comes into action but with a change in its direction. The restoring force, which is maximal at the extreme positions of the bob, stops the bob and pulls it back toward the equilibrium position. The result is a continuing *oscillatory motion* of the bob back and forth along an arc. Thus, the constant play between the restoring force and inertia property is responsible for the oscillatory motion. The oscillatory motion is **periodic** and repeats itself in equal intervals of time.

In general, an **oscillation** is a periodic fluctuation in the value of a physical quantity above and below some equilibrium value. In mechanical oscillations, the body undergoes linear or angular displacement whereas non-mechanical oscillations involve the variation of quantities such as voltage or current in electrical circuits or the electric and magnetic fields in TV signals, light waves, UV-rays and X-rays.

1.3 SIMPLE HARMONIC MOTION

Any motion, which repeats itself at regular intervals according to a sinusoidal law, is called a **harmonic motion**. The oscillations of a simple pendulum or the motion of a mass m under a restoring force is an *idealized* model of harmonic motion. In these cases, the force is *directly proportional* to displacement. The oscillatory motion in which the force is directly proportional to the displacement is called **simple harmonic motion (S.H.M.)**. Since force is proportional to the displacement, the acceleration is not constant but varies with time. S.H.M. is thus a non-uniformly accelerated motion. Hence the equations of motion with constant acceleration are not applicable to simple harmonic motion.

1.3.1 Equation of Simple Harmonic Motion

We now obtain the expressions for displacement, velocity, and acceleration of a body moving with simple harmonic motion.

For studying simple harmonic motion, we consider a block of mass m attached to a spring (see Fig. 1.2). When the mass is pulled and left to it, it oscillates about its equilibrium position. The directed distance of the mass from its equilibrium position is called its **displacement**. The restoring force F acting on the body is due to the *stiffness* of the spring and is given by Hooke's law.

$$F = -kx \quad (1.1)$$

where x is the displacement from the equilibrium position. k is called the elastic constant which represents the force required to displace the mass one unit of distance.

The negative sign in the expression indicates that the force F is opposite to the displacement. When the mass is pulled to right, the spring gets stretched. Then, x is positive and the force is negative and is directed to the left. When x is negative, the spring is compressed and F is positive and directed to the right.

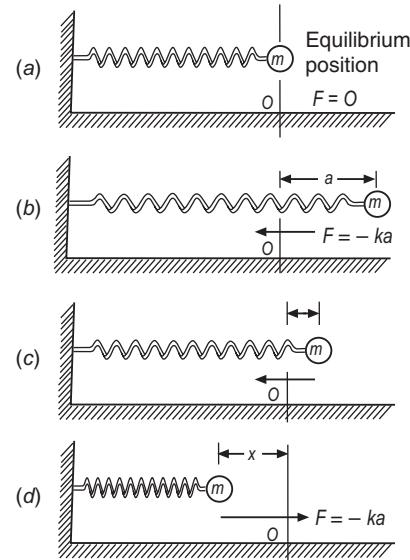


Fig. 1.2

According to Newton's second law, the restoring force produces acceleration. Thus,

$$\begin{aligned} F &= -kx = ma \\ (\text{restoring force}) &\quad (\text{inertial force}) \\ \therefore m \frac{d^2x}{dt^2} &= -kx \\ \text{or } \frac{d^2x}{dt^2} + \frac{k}{m}x &= 0 \end{aligned} \quad (1.2)$$

This equation is merely another way of writing Newton's second law and it is known as the *differential equation of simple harmonic motion*.

Putting $k/m = \omega^2$ into the above equation, we get

$$\frac{d^2x}{dt^2} + \omega^2 x = 0 \quad (1.3)$$

where ω is called the angular frequency and is given by $\omega = \sqrt{\frac{k}{m}}$. The time period of oscillations is given by

$$T = \sqrt{\frac{m}{k}} \quad (1.4)$$

It is seen from equ. (1.4) that the time period of the mass is independent of the amplitude. Secondly, for a given elastic constant, the period increases with the increase in the mass of the block; a heavier mass oscillates more slowly. For a given mass, the period decreases as k increases; a stiffer spring causes quicker oscillations.

All simple harmonic oscillators obey a differential equation of the form (1.3).

1.3.2 Characteristics of SHM

(i) Displacement: The general solution of the differential equation (1.3) is given by

$$x = A e^{i\omega t} + B e^{-i\omega t} \quad (1.5)$$

where A and B are unknown constants to be determined from the initial conditions. This solution of the harmonic force equation is known as the exponential form of the solution. The general solution can be simplified as

$$x = A \sin(\omega t + \phi) \quad (1.6)$$

Thus, the displacement of the body at any instant is given by equn. (1.6). Therefore, we can say that if the displacement x of a particle relative to the origin of the coordinate system is given as a function of time as in equn. (1.6), then the motion is simple harmonic. x varies periodically between the values $-A$ and $+A$. A is the maximum value of the displacement and is known as the **amplitude** of the oscillation. It may be noted that the amplitude A is constant. The quantity $(\omega t + \phi)$ is called the *phase angle* and ϕ is the initial phase, that is the phase at $t = 0$.

(ii) Velocity: By differentiating equ. (1.6) we can find the velocity of a particle moving with SHM. Thus,

$$v = \frac{dx}{dt} = \omega A \cos(\omega t + \phi) \quad (1.7)$$

v varies periodically between the values $+\omega A$ and $-\omega A$. When the magnitude of the displacement is greatest, the velocity is zero; and when the displacement is least that is at the midpoint of the motion, the velocity has its greatest magnitude.

(iii) Acceleration: We can find the acceleration of the oscillating particle by differentiating the expression for v . Thus,

$$a = \frac{dv}{dt} = -\omega^2 A \sin(\omega t + \phi) \quad (1.8)$$

Acceleration varies periodically between the values $+\omega^2 A$ and $-\omega^2 A$. We can combine equ. (1.8) and equ. (1.6) to give

$$a = -\omega^2 x \quad (1.9)$$

In simple harmonic motion the acceleration is proportional and opposite to the displacement. When the displacement has its greatest positive value, the acceleration has its greatest negative value and vice versa. When the displacement is zero, acceleration is also zero. Basing on the equations (1.1) and (1.9), we can make a general statement that *whenever the force acting on a body is linearly proportional to the displacement and in the opposite direction, the body will execute simple harmonic motion.*

Fig. 1.3 illustrates displacement x , velocity v , and acceleration a as functions of time.

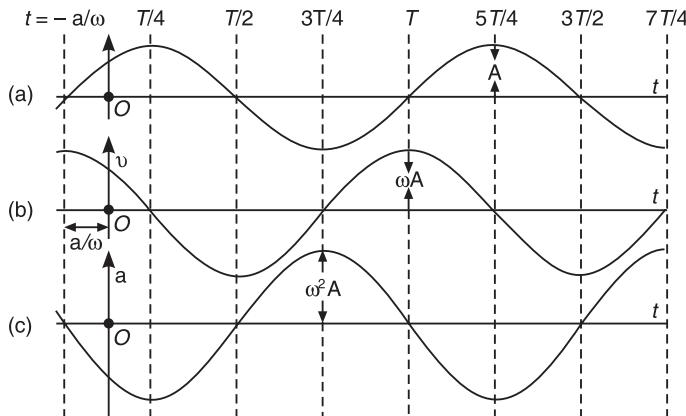


Fig. 1.3

(iv) Time period: The time period T is the time taken for one complete oscillation. The sine function (1.6) repeats itself whenever the quantity in parenthesis increases by 2π . Thus, if we start at $t = 0$, $\phi = 0$ and the time T is given by

$$\omega T = 2\pi$$

$$\therefore T = \frac{2\pi}{\omega} = 2\pi \sqrt{\frac{m}{k}} \quad (1.10)$$

Note that the period of motion is determined by the mass m and elastic constant k . It does not depend on the amplitude of the oscillation. It means for any given values of m and k , whether the amplitude is large or smaller, the time for one complete oscillation is the same.

(v) Frequency: The frequency, v , is the number of complete oscillations per second. It is the reciprocal of the time period.

$$v = \frac{1}{2\pi} \sqrt{\frac{k}{m}} \quad (1.11)$$

Note that the frequency and time period are independent of the amplitude A .

(vi) Phase: The angle $(\omega t + \phi)$ is called the phase of the oscillation. It represents the state of the oscillation of the body by specifying the *position and direction of motion* of the body. The constant angle ϕ is called the phase constant, which is determined

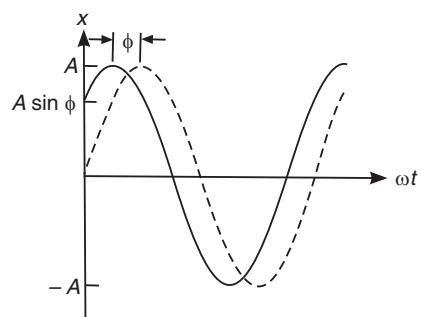


Fig. 1.4

uniquely by the initial displacement and velocity of the particle. The constant ϕ and amplitude A tell us what the displacement was at time $t = 0$. The phase of the oscillation is useful in comparing the motions of two bodies.

1.3.3 Three Conditions for the Occurrence of Simple Harmonic Oscillations

In case of mechanical oscillators, three conditions must be satisfied for the occurrence of simple harmonic oscillations.

- (i) There must be a position of stable equilibrium.
- (ii) There must be no dissipation of energy.
- (iii) The acceleration should be proportional to the displacement and opposite in direction.

1.3.4 Energy

A body executing simple harmonic oscillations is called a **simple harmonic oscillator**. A simple harmonic oscillator possesses potential energy as well as kinetic energy. The elastic property of the oscillating system (spring) stores potential energy. That is, the potential energy is possessed by virtue of its displacement from the equilibrium position and the inertia property (mass) stores kinetic energy; thus, the kinetic energy is due to its velocity. As the system oscillates, there is a continuous conversion of potential energy into kinetic energy and vice versa. If no dissipative forces are present, the total energy is conserved.

The kinetic energy of the particle is

$$\begin{aligned} E_k &= \frac{1}{2}mv^2 = \frac{1}{2}m\omega^2 A^2 \cos^2(\omega t + \phi) \\ &= \frac{1}{2}m\omega^2 A^2 [1 - \sin^2(\omega t + \phi)] \\ &= \frac{1}{2}m\omega^2 (A^2 - x^2) \\ &= \frac{1}{2}k(A^2 - x^2) \end{aligned} \quad (1.12)$$

The potential energy, U , is given by

$$U = - \int_0^x F dx = \int_0^x kx dx = \frac{1}{2}kx^2 \quad (1.13)$$

According to equ. (1.13) the potential energy is proportional to the square of the displacement. It implies that the mass is in a **potential well** created by the spring. All simple harmonic oscillations are characterized by such a *parabolic potential well* (See Fig. 1.5).

The total energy of the simple harmonic oscillator is

$$\begin{aligned} E &= K.E. + P.E. \\ &= \frac{1}{2}k(A^2 - x^2) + \frac{1}{2}kx^2 \\ \text{or} \quad E &= \frac{1}{2}kA^2 = \text{constant.} \end{aligned} \quad (1.14)$$

Thus, *the total energy of a simple harmonic oscillator does not depend on time and is a constant of the motion. Further, it is proportional to the square of the amplitude of the oscillation.*

Fig. 1.5 shows the kinetic energy E_k , potential energy E_p and total energy E of the oscillator plotted against the displacement x . The horizontal line represents the total energy E , which is constant and does not vary with x . The potential energy curve $E_p(x)$ is parabolic with respect to x and is symmetric about the position of equilibrium, $x = 0$. The kinetic energy curve E_k is also parabolic with respect to x and is symmetric about the position of equilibrium, $x = 0$. One curve is inverted with respect to the other, which indicates the 90° phase difference between the displacement and the velocity. The horizontal line intersects the potential energy curve at $x = -A$ and $x = A$, where the energy is entirely potential. At these points $v = 0$ and there is no kinetic energy. At the equilibrium position, $x = 0$ and P.E. = 0, so that the total energy is in the form of kinetic energy. At any value of x between $-A$ and $+A$, the vertical distance from the x -axis to the parabola is U ; since $E = \text{K.E.} + \text{P.E.}$, the remaining vertical distance up to the horizontal line is K.E. It is thus seen that energy is continuously being transferred between potential energy stored in the spring and the kinetic energy of the mass. Note that the maximum values of the potential and kinetic energies are equal.

We now look at two oscillating systems that exhibit simple harmonic motion, namely, a simple pendulum and a torsional pendulum.

1.3.5 The Simple Pendulum

A simple pendulum consists of a point mass, m suspended by a light string of length L . When the point is pulled to one side of its equilibrium position and is released, it oscillates about the equilibrium position. The motion occurs in a vertical plane and is driven by the force of gravity. The path of the point mass is not a straight line but the arc of a circle with radius L equal to the length of the string. In this case the restoring force must be proportional to x or to θ since $x = L\theta$. The forces acting on the mass are the tension, T acting along the string and the weight mg . In Fig. 1.6 we represent the forces on the mass in terms of tangential and radial components. The restoring force F is the tangential component of the net force.

$$F = -mg \sin \theta \quad (1.15)$$

Note that F always acts towards the equilibrium position and opposite to the displacement.

The restoring force is provided by gravity; the tension T merely acts to make the point mass move in an arc. The restoring force is proportional to $\sin \theta$ and hence the motion is not simple harmonic. However, when the angle is small, $\sin \theta \approx \theta$. With this approximation

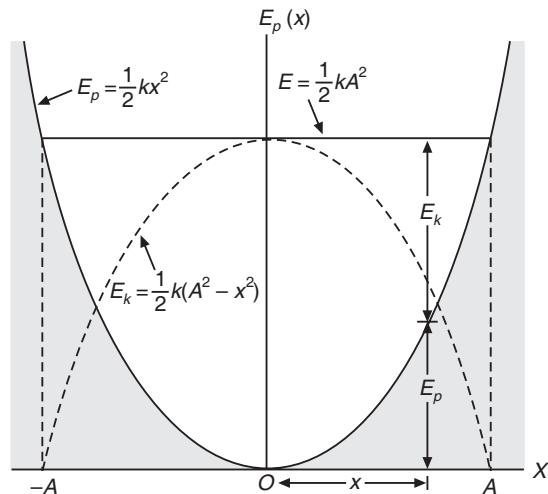


Fig. 1.5

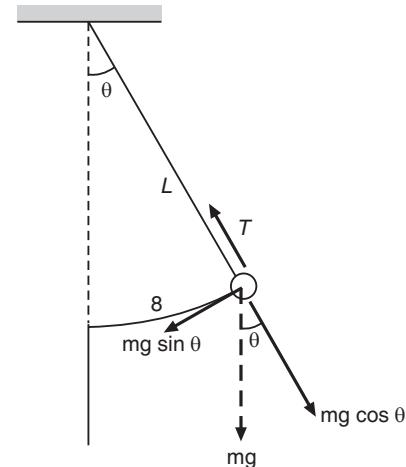


Fig. 1.6

$$F = -mg \theta = -\frac{mg}{L}x. \quad (1.16)$$

The force constant is

$$k = mg/L. \quad (1.17)$$

It follows from equ. (1.17) that the angular frequency is $\omega = \sqrt{k/m} = \sqrt{\frac{mg/L}{m}} = \sqrt{g/L}$

The corresponding frequency and time period relations are

$$\nu = \frac{\omega}{2\pi} = \frac{1}{2\pi} \sqrt{\frac{g}{L}} \quad (1.18)$$

and

$$T = \frac{1}{\nu} = 2\pi \sqrt{\frac{L}{g}} \quad (1.19)$$

1.3.6 Torsional Pendulum

A torsional pendulum consists of a disk or rod suspended at the end of a wire (see Fig. 1.7). When the end of the wire is twisted by an angle θ , the restoring torque τ arises which obeys Hooke's law.

$$\tau = -\kappa\theta \quad (1.20)$$

where κ is called the torsional constant. If the wire is twisted and released, the oscillating system is called a torsional pendulum. The rotational form of Newton's second law is

$$\tau = I\alpha$$

or

$$-\kappa\theta = I \frac{d^2\theta}{dt^2}$$

which may be rewritten as

$$\frac{d^2\theta}{dt^2} + \frac{\kappa}{I}\theta = 0 \quad (1.21)$$

This is the equation of a simple harmonic oscillator whose angular frequency is

$$\omega = \sqrt{\frac{\kappa}{I}}$$

and time period is

$$T = 2\pi \sqrt{\frac{I}{\kappa}} \quad (1.22)$$

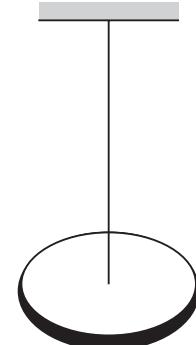


Fig. 1.7

The balance wheel in a clock or wristwatch is an example of a torsional pendulum.

Example 1.1: The displacement equation of a particle describing simple harmonic motion is $x = 0.01 \sin 100\pi(t + 0.005)$ meter, where x is the displacement of the particle at any instant t . Calculate the amplitude, periodic time, maximum velocity and displacement at the time of the motion.

Solution: The displacement of the particle is,

$$x = 0.01 \sin 100\pi(t + 0.005)$$

The general equation of SHM is,

$$x = a \sin(\omega t + \varphi)$$

Comparing the two equations, we get

Amplitude, $a = 0.01$ metre

$$\omega = 100\pi \quad \therefore \text{Periodic time } T = \frac{2\pi}{\omega} = \frac{2\pi}{100\pi} = 0.02 \text{ seconds}$$

The velocity of the particle at the displacement x is given by, $v = \omega\sqrt{a^2 - x^2}$

The maximum velocity is given by v_{\max} when $x = 0$

$$\therefore v_{\max} = \omega a = 100 \times 3.14 \times 0.01 = 3.14 \text{ m/s}$$

The displacement at the time of start ($t = 0$) is

$$x = 0.01 \sin 100\pi (0.005)$$

or

$$x = 0.01 \sin \frac{\pi}{2} = 0.01 \text{ metre}$$

Example 1.2: A body of mass 0.05 kg executes SHM. When the displacement from the centre of motion is 0.04 m, the force acting on the body is 18×10^{-3} N. If the maximum velocity is 2 m/s, find the amplitude and acceleration.

Solution: The acceleration is given by,

$$a = -\omega^2 x$$

The force acting on the body is given by,

$$F = m \times a = m\omega^2 x$$

Accordingly, $0.05 \times \omega^2 \times 0.04 = 18 \times 10^{-3}$

$$\therefore \omega^2 = \frac{18 \times 10^{-3}}{0.05 \times 0.04} = 9 \quad \text{or} \quad \omega = 3 \text{ rad./s}$$

$$\text{Now, } v_{\max} = \omega a \quad \text{or} \quad a = \frac{v_{\max}}{\omega}$$

$$\therefore \text{Amplitude, } a = \frac{2}{3} \text{ m} = 0.667 \text{ m}$$

$$\text{The maximum acceleration} = \omega^2 a = 9 \times 0.667 = 6 \text{ m/s}^2$$

Example 1.3: A particle of mass 5 gm executes SHM and has amplitude of 8 cm. If it makes 16 vibrations per second, find its maximum velocity and energy at mean position.

Solution: In case of SHM, $x = a \sin (\omega t + \phi)$

The maximum velocity, $v_{\max} = \omega a = 2\pi n a$

$$\therefore v_{\max} = 2 \times 3.14 \times 10 \times 8 = 803.8 \text{ cm/s} \\ = 8.038 \text{ m/s}$$

The energy at mean position is entirely kinetic.

$$\therefore E = \frac{1}{2} m v_{\max}^2 = \frac{1}{2} \times 5 \times 10^{-3} \text{ kg} \times (8.038)^2 = 0.16 \text{ J}$$

Example 1.4: The displacement of particle executing SHM is given by, $x = 20 \sin \left(\frac{2\pi t}{T} + \phi \right)$. The period of vibration is 60 seconds. At $t = 0$, the displacement of the particle is 1 cm. Find

(a) The initial phase

(b) The phase angle when the displacement is 3 cm.

(c) The phase difference between any two positions of the particle 10 seconds apart.

Solution: (a) When $t = 0$, $x = 1 \text{ cm}$

$$\therefore 1 = 20 \sin (0 + \phi) \quad \text{or} \quad \sin \phi = \frac{1}{20} = 0.05$$

$$\therefore \phi = \sin^{-1}(0.05) = 2^\circ 52'$$

$$(b) \text{ In this case, } 3 = 20 \sin\left(\frac{2\pi}{t} + \phi\right)$$

$$\therefore \text{Phase angle}\left(\frac{2\pi}{t} + \phi\right) = \sin^{-1}\left(\frac{3}{20}\right) = \sin^{-1}(0.15) = 8^\circ 38'$$

Let α_1 and α_2 be the phase angles at times t_1 and t_2 respectively such that $(t_1 - t_2) = 10 \text{ sec.}$

$$\therefore \alpha_1 = \left(\frac{2\pi t_1}{T} + \phi\right) \text{ and } \alpha_2 = \left(\frac{2\pi t_2}{T} + \phi\right)$$

$$(c) \text{ Phase difference} = \alpha_1 - \alpha_2 = \alpha_1 = \frac{2\pi}{T}(t_1 - t_2) = \frac{2\pi \times 10}{60} = \frac{\pi}{3} \text{ radians}$$

Example 1.5: A spring is stretched by 8 cm by a force of 10 N. Find the force constant. What will be the period of 4.0 kg mass suspended by it?

Solution: $F = kx$, where k is a force constant.

$$\therefore k = \frac{F}{x} = \frac{10 \text{ N}}{0.08 \text{ m}} = 125 \text{ N/m}$$

The time period T , is given by,

$$T = 2\pi \sqrt{m/k} = 2\pi \sqrt{4.0 \text{ kg} / 125 \text{ N/m}} = 1.12 \text{ s}$$

1.4 FREE OSCILLATIONS

Free oscillations are oscillations that appear in a system as a result of a single initial deviation of the system from its state of stable equilibrium. When a pendulum is displaced from its equilibrium position and left to the action of internal forces, it undergoes *free oscillations* with the frequency given by,

$$v = \frac{1}{2\pi} \sqrt{\frac{g}{L}}$$

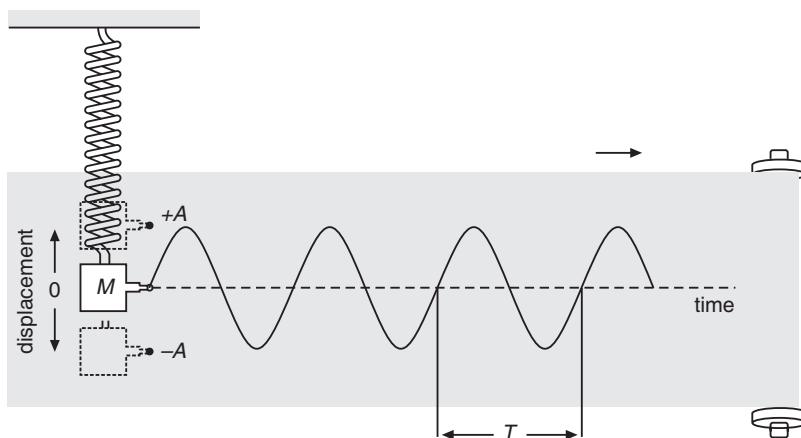


Fig. 1.8

The time period and frequency of the simple pendulum depend *only* on the length of the string and the acceleration due to gravity and is independent of the mass. The frequency will be the same as long as the pendulum does not experience resistance to its motion. The frequency with which the pendulum oscillates freely at its own is called its **natural frequency**. Thus, if no resistance is offered to the motion of any oscillating body by air friction or other forces, the body will keep on oscillating indefinitely at its **natural frequency**, as shown in Fig. 1.8.

Such an oscillator is called an *ideal* oscillator. The period of oscillations of an ideal oscillator is independent of the amplitude and a characteristic property of the oscillation. The ideal systems are frictionless and energy is not dissipated away and hence the total mechanical energy and the amplitude remain constant.

Example 1.6: A body of mass 4.9 kg hangs from a spring and oscillates with a period of 0.6 seconds. How much will the spring shorten when the body is removed?

Solution: $T = 2\pi\sqrt{m/k}$

$$\therefore 0.6 = 2\pi\sqrt{4.9/k} \quad \text{or} \quad k = 536.8 \text{ N/m}$$

Further

$$k = F/x \quad \text{or} \quad x = F/k$$

$$\therefore x = \frac{4.9 \text{ kg} \times 9.8 \text{ m/s}^2}{536.8 \text{ N/m}} = 0.089 \text{ m} = 8.9 \text{ cm}$$

Example 1.7: A mass $M = 2 \text{ kg}$ hangs from a vertical spring. When a mass $m = 0.6 \text{ kg}$ is gently added, the spring is further stretched by 4 cm. Now m is removed and M is set into oscillations. Calculate the period of oscillations.

Solution: Let x be the displacement of the spring with mass M .

Now $F = kx, \quad \therefore Mg = kx$

or $2g = kx$

When an extra mass $m (= 0.6 \text{ kg})$ is added, the spring is further stretched by 0.04 m.

Hence, $(M+m) \times g = k(x+0.4)$

or $(2 + 0.06) \times g = kx + 0.4 \quad k = 2g + 0.04 \quad k$

$$\therefore 2.06 \times g = 2g + 0.04 \quad \text{or} \quad k = \frac{0.6 \times 9.8}{0.04} = 14.7 \text{ N/m}$$

Now $T = 2\pi\sqrt{M/k} = 2\pi\sqrt{2/14.7} = 2.317 \text{ sec.}$

1.5 DAMPED OSCILLATIONS

An *ideal* pendulum, once set into motion, continues to oscillate between two spatial positions forever without a decrease in amplitude. In actual practice, no body can oscillate for an indefinite time. If we watch an oscillating pendulum, we shall find that its amplitude of oscillation goes on decreasing due to resistance offered both at the supports and by the surrounding air; and ultimately it stops. The frictional force always opposes the motion of the body, whether it is going away from the equilibrium position or it is returning towards the equilibrium position. Hence the energy given in the initial displacement is converted slowly but continuously into heat in doing work against friction. This energy is never returned to the body. This phenomenon is called the **energy dissipation**. As a result of energy dissipation, the amplitude of oscillation of the body diminishes with each oscillation. When the whole of the initial energy of the oscillating body is dissipated, the amplitude of oscillation becomes zero. The phenomenon of decay in the amplitude of oscillations is known as **damping**. Damped oscillations are not sinusoidal, but are much more complex. The period is no longer

a characteristic property of the oscillation, but depends on the amplitude. For example, a pendulum immersed in water exhibits damped oscillations. On the other hand, if it is immersed in a viscous medium such as oil, there will be no oscillations at all.

Damping force is resistive: it opposes motion (i.e. is always in opposite direction to motion). To explain the damping dynamically, we may assume that in addition to the restoring force $F = -kx$, there is a *damping force* that is opposed to the velocity. Friction and viscosity are such kind of forces. A damped system is subjected to the following two forces:

- (i) A *restoring force* proportional to displacement but oppositely directed and
- (ii) A *frictional force* proportional to the velocity but oppositely directed.

We write the damping force as $F' = -bv$ (1.23)

where b is a constant that depends on the medium and the shape of the body.

The resultant force on the body is

$$F + F' = -kx - bv \quad (1.24)$$

Therefore the equation of motion of the body is

$$\begin{aligned} ma &= -kx & -bv \\ (\text{inertial force}) & (\text{restoring force}) & (\text{damping force}) \\ m \frac{d^2x}{dt^2} &= -kx - b \frac{dx}{dt} \\ m \frac{d^2x}{dt^2} + b \frac{dx}{dt} + kx &= 0 \\ \frac{d^2x}{dt^2} + \frac{b}{m} \frac{dx}{dt} + \frac{k}{m} x &= 0 \\ \frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} + \omega_0^2 x &= 0 \end{aligned} \quad (1.25)$$

where $\gamma = b/m$ is the *damping coefficient* of the system and $\omega_0 = \sqrt{k/m}$ is the *natural frequency* of the system. Equ. (1.26) is a differential equation. A simple solution of the differential equation is given by

$$x = A_0 e^{-\gamma t/2m} \cos(\omega t + \phi) \quad \text{for } \omega_0^2 > \gamma^2 \quad (1.27)$$

The angular frequency of damped oscillations is given by

$$\omega = \sqrt{\omega_0^2 - \left(\frac{\gamma}{2m}\right)^2} \quad (1.28)$$

It is clear that the damped angular frequency ω is less than the natural angular frequency $\omega_0 = \sqrt{\frac{k}{m}}$.

Damping plays a beneficial role. For example, the shock absorbers in a car provide a velocity dependent damping force so that when the car goes over a bump, it does not continue bouncing forever.

1.5.1 Weak Damping

In equ. (1.28), ω will be a real and positive quantity if the condition $\frac{\gamma}{2m} < \omega_0$ is satisfied.

When ω is real, the damping force is weaker than the restoring force and the oscillations are weakly damped. It is said that the oscillations are **underdamped**. The condition represents a

simple harmonic motion with amplitude $Ae^{-\gamma t/2m}$. The motion differs from the undamped motion in two ways. First, the amplitude of the oscillations is not constant but falls slowly with time over many oscillations, as shown in Fig. 1.9. We may treat the damped oscillations as nearly sinusoidal with progressively diminishing amplitude. Secondly, weakly damped systems have angular frequencies close to the system's natural frequency. The time period of damped oscillation is the time interval between two consecutive maximum displacements.

It has been found that although friction affects the manner in which the amplitude decreases it has practically no effect on the period of damped oscillations.

The motion of a pendulum in air and the electric oscillations of an LCR circuit belong to this category. In electric circuits containing inductance, capacitance and resistance, there is a natural frequency of oscillation and the resistance plays the role of the damping constant γ . It is usually desirable to minimize damping, but damping can never be prevented completely.

1.5.2 Heavy Damping

When the damping is very large such that $\gamma > 2m\omega_0$, the solution for the eqn. (1.26) assumes the form

$$x(t) = C_1 e^{-\gamma t} + C_2 e^{-\gamma t} \quad (1.29)$$

In this case, ω is imaginary and there are no oscillations. The system once displaced, returns to its equilibrium position quite slowly without any oscillations (see Fig. 1.10). The motion is termed as **heavy** or **over-damped motion**. As the damping increases, the time taken by the body to reach equilibrium also increases.

For example, if a door closing mechanism is heavily damped, when released from the open position the door slowly closes, moving to the equilibrium position without oscillating.

1.5.3 Critical Damping

When $\gamma = 2m\omega_0$, we get $\omega = 0$ and again, there is no oscillation. The motion is known as **critical damped motion**. A critically damped system approaches equilibrium as fast as possible without any overshoot or oscillation (Fig. 1.11). This is a very desirable attribute in many mechanical and electrical systems.

Critical damping is used in the construction of many pointer type instruments where the pointer moves and comes to stationary position in a very short time. Such instruments are called *dead-beat*. Such behaviour is desirable in electrical meters and the like where we would like to note a steady reading as soon as the meter is connected in a circuit. In the moving coil galvanometer, the ammeter and the voltmeter, the current carrying coil is wound on a metallic frame so that the induced eddy currents in the frame make the motion dead-beat. On the other

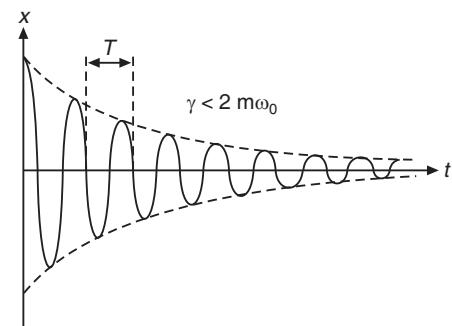


Fig. 1.9

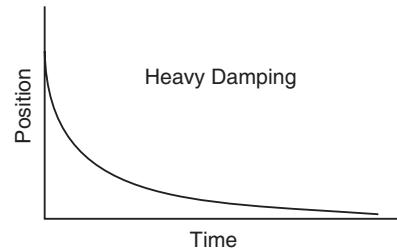


Fig. 1.10

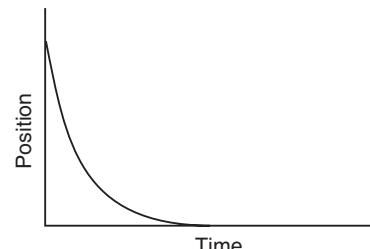


Fig. 1.10

hand, in the ballistic galvanometer where weak damping condition is to be fulfilled, the coil is wound on a non-metallic frame.

Another familiar example is that of **shock absorbers** provided in a motorbike. A shock absorber often combines a spring with a sealed container of fluid. Shock absorbers lessen the jolts of a bumpy trail. To understand this in more detail, let us consider what happens when a bike equipped with such a shock absorber hits a bump. The force from the bump compresses the spring, with the result that less of the force from the bump passes to the rest of the bike and the rider. The spring then supplies a restoring force. In the absence of any other force, the rider and bike would in principle then move forever in simple harmonic motion. However, inside a shock absorber, the spring moves a piston in a sealed cylinder of fluid. The fluid supplies the *damping force*, which is greater than the minimum needed to prevent oscillations. The vehicle returns to equilibrium without oscillating.

1.6 FORCED OSCILLATIONS

The energy of a damped oscillator decreases in time as a result of the dissipative force. It is possible to compensate for the energy loss by applying an external force, which supplies energy to the oscillator. The oscillations produced when an *external oscillatory force* is applied to a body subject to an elastic force are known as **forced oscillations**. For instance, you could pull a child on a swing up to a certain height, then let it go and wait for the motion to die away. But this is not the only possibility; we could also repeatedly push the swing at any frequency we like and watch what happens. In this case, we have produced **forced oscillations**. There are now two frequencies in the problem: the **natural frequency** ω_0 of the free oscillations, and the **driving frequency** ω_f of the forced oscillations. *Forced oscillations may be defined as the oscillations in which the body oscillates with a frequency other than its natural frequency under the action of an external periodic force.*

The forces that act on the body undergoing forced oscillations are as follows:

- (i) A restoring force F_1 , that is proportional to the displacement and oppositely directed;
- (ii) A damping force F_2 , that is proportional to the velocity but oppositely directed; and
- (iii) A **driving force** (external periodic force), $F_3 = F_o \sin \omega_f t$.

The total force acting on the body is given by

$$F = F_1 + F_2 + F_3 = -kx - \gamma v + F_o \sin \omega_f t$$

According to Newton's second law this force must be equal to the product of the mass and acceleration. Therefore the equation of the motion of the body is

$$\begin{aligned} ma &= -kx & -\gamma v &+ F_o \sin \omega_f t \\ (\text{inertial force}) & (\text{restoring force}) & (\text{damping force}) & (\text{external force}) \\ \therefore m \frac{d^2x}{dt^2} &= -kx - \gamma \frac{dx}{dt} + F_o \sin \omega_f t \end{aligned} \quad (1.30)$$

Rearranging the terms in the above equation, we get

$$\frac{d^2x}{dt^2} + \frac{\gamma}{m} \frac{dx}{dt} + \frac{k}{m} x = \frac{F_o}{m} \sin \omega_f t \quad (1.31)$$

Equ. (1.31) is a differential equation which can be solved by standard techniques. Instead of going into the mathematical details, we confine here to the results. When the force is first applied, the motion is complex. The system wants to vibrate with a natural frequency ω_0 but is being forced by a driving frequency ω_f . Therefore, the motion is initially a superposition of free damped oscillations and forced oscillations. This initial motion is referred to as **transitory behaviour**. However, the transitory behaviour decays exponentially. After the

free oscillations die out, only the forced oscillations remain. We then say that the motion has reached a **steady-state condition** and the oscillator oscillates with constant amplitude (see Fig. 1.12b). In the steady state, the energy input per cycle equals the energy lost per cycle.

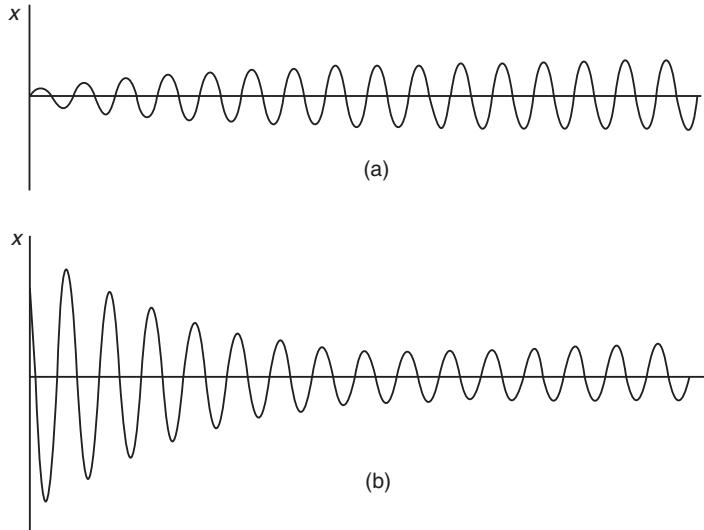


Fig. 1.12

In the steady-state the body is forced to oscillate with the angular frequency ω_f of the applied force. Therefore, the steady state solution to equ. (1.31) is

$$x = A \sin(\omega_f t - \alpha) \quad (1.32)$$

where A is the amplitude of the forced oscillations and α is the phase angle between the displacement x and the external force F_3 . Both the amplitude A and the initial phase α are not arbitrary constants but are fixed quantities that depend on the frequency ω_f of the applied force. The larger the difference between ω_f and the natural frequency ω_0 , the smaller the amplitude of the forced oscillations because it is more difficult for the oscillator to respond to the applied force when the forcing frequency is not near the natural frequency. It can be shown that the amplitude of the forced oscillations is given by

$$A = \frac{F_0 / m}{\sqrt{\left(\omega_0^2 - \omega_f^2\right)^2 + (\gamma\omega_f / m)^2}} \quad (1.33)$$

At constant values of F_0 , m and γ , the amplitude of free oscillations depends on the ratio of the frequencies of the driving force ω_f and of the free undamped oscillations, ω_0 . The variation of the amplitude with ω_f at different damping factors γ is shown in Fig. 1.13.

Equ. (1.33) indicates that the forced oscillations are not damped but are of constant amplitude. It means that the external agent overcomes the damping forces and provides the energy necessary to maintain the oscillations.

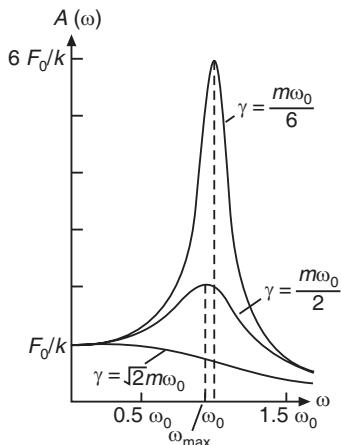


Fig. 1.13

1.6.1 Distinction between Free and Forced Oscillations

- Free oscillations are oscillations executed by a body without being acted upon by an external force. They occur due to the elastic forces and inertia of the system. In contrast, the forced oscillations occur due to the action of a periodic force applied externally.
- Free oscillations diminish gradually due to the damping forces. Forced oscillations persist as long as the applied periodic force acts on the body.
- The frequency of free oscillations depends on the mass, and elasticity of the body. The frequency of forced oscillations does not depend on any of such factors and is equal to the frequency of the applied periodic force.
- Free oscillations may occur with any amplitude. Due to the effects of damping the amplitude goes on decreasing. In case of forced oscillations, the amplitude is small except in the vicinity of resonance frequency.

1.7 RESONANCE

Referring to Fig. 1.13, we observe that each driving frequency is characterized by its own amplitude. As the driving frequency ω_f is increased, the amplitude rises until it reaches a maximum at ω_{\max} and at further high frequencies the amplitude again decreases. When the frequency of the driving force is near the natural frequency ω_0 of the oscillating system, the oscillation amplitude becomes very large (see Fig. 1.13).

The dramatic increase in amplitude near the natural frequency is called **resonance** and the frequency ω_0 is called the **resonance frequency** of the system. The reason for large amplitude oscillation is that the rate of energy transfer from the applied force to the forced oscillator is a maximum. Resonance occurs whenever a system is subject to an external action that varies periodically with time and with the proper frequency. At the resonance frequency, the external force and the velocity of the particle are in phase. As a result, the power transfer to the oscillator has its maximum value. At frequencies above or below the resonance value, the force and velocity are not in phase and hence the power transfer is lower.

Note the following:

- When damping is small, the amplitude of forced oscillations grows with increasing ω_f and at $\omega_f = \omega_0$, the amplitude of the oscillations becomes equal to infinity. Further, the resonance curve is sharp, that is the amplitude falls off rapidly on either side of the resonant frequency.
- When damping is large, the amplitude falls off very slowly on either side of the resonant frequency.
- When the oscillator responds to a number of close by frequencies near the resonant value, the resonance is flat.
- At resonance, the velocity is in phase with the applied force. Since the rate of work done on oscillator by the applied force is F_v , this quantity is positive when F and v are in phase and represents a favourable condition for transfer of energy to the oscillator.
- At resonance, the oscillating system continuously absorbs energy from the agent applying external periodic force.

The phenomenon of resonance appears in many areas of physics.

- Tuning of a radio receiver involves matching of the frequency of the tuned circuit with that of the radio waves. Only when the resonance condition is reached, we listen clearly the sounds transmitted by a particular radio station.
- Resonance absorption of radiation by atoms takes place when the frequency of the incident light waves equals the natural frequency of the atom.

- In a cyclotron, particles are accelerated to high energies only when the frequency of electric field accelerating the particles is equal to the frequency of revolution of the particle in magnetic field acting perpendicular to the particle path.

A dramatic example of resonance occurred when a company of soldiers was marching in step across a bridge in St.Petersburg. The bridge collapsed. The period of free oscillation of the bridge coincided with the period of an ordinary marching step and resonance took place; it caused swinging of the bridge with very high amplitude leading to the ultimate collapse. Another instance was that of the Tacoma Narrows Bridge in Washington State in 1940. The wind blowing through the Tacoma Narrows broke up into vortices, which provided puffs of wind that shook the bridge at a frequency that matched one of its natural vibrational frequencies. In a couple of hours the amplitude became so large that the centre span collapsed due to resonance.

1.7.1 Sharpness of Resonance

When the driving force is increased or decreased from resonant frequency, the amplitude falls off from the maximum value. The term sharpness of resonance refers to the rate of fall of amplitude with the change in frequency of the driving force, on either side of the resonant frequency. The power absorbed is maximum at resonance for small values of damping. If P_r is the power absorbed at resonance, P is the power absorbed at any frequency v . A plot between P/P_r versus v is shown in Fig. 1.14. The frequency values, on either side of v_0 , at which the power absorbed is half of the maximum are called **half-power points**. The frequency difference between these two half-power points is called the **bandwidth** of the oscillator. Thus,

$$\text{Bandwidth, } \Delta v = v_2 - v_1 \quad (1.34)$$

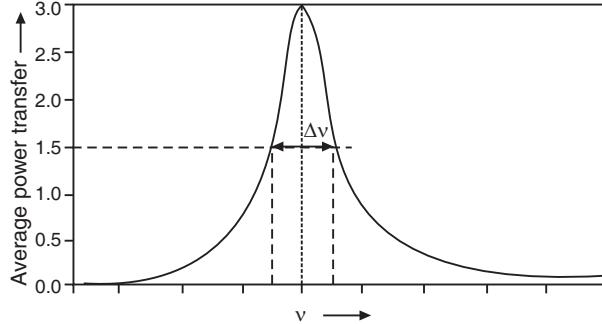


Fig. 1.14

1.8 COUPLED OSCILLATIONS

In physics we come across a variety of coupled systems that can oscillate. For example, a solid body is composed of many atoms or molecules. Every atom behaves like an oscillator, which vibrates about an equilibrium position. Each atom affects the motion of its neighbouring atom and thus the atoms of the solid are coupled together. A **coupled oscillator** is a system of two or more connected bodies that oscillates within a set of stable, predictable patterns. The number of “ways” in which a coupled system can oscillate is determined by the number of coupled bodies. Once released, a pendulum can only have one frequency that describes its motion. Two pendulums coupled together can have two oscillating frequencies. A particular motion of coupled oscillators is called a **normal mode** of oscillation. Normal modes correspond to the case where the two bodies move with the same frequency and maintain a constant phase difference. In one normal mode, the two oscillators move in phase and in the second normal mode, they move out of phase. Usually, any coupled system will exhibit a mixing of all possible modes.

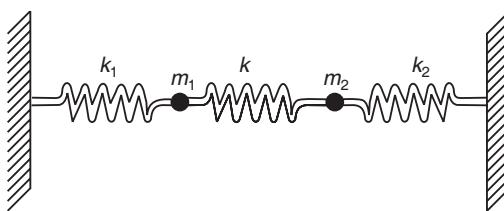


Fig. 1.15

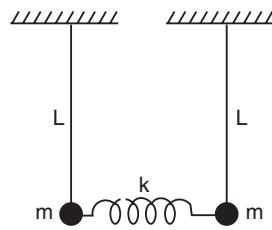


Fig. 1.16

An example of coupled oscillators is two identical pendulums coupled by a spring k , as shown in Fig. 1.16. When the motion of the coupled oscillators occurs in one of the normal modes, their energies and the oscillation amplitudes remain constant. In the general case, the amplitude of each oscillator does not remain constant. The displacement of two coupled oscillators with time is shown in Fig. 1.16. It is clear from Fig. 1.16 that when one pendulum is displaced and is released, its amplitude begins to decrease while the second pendulum starts oscillating with increasing amplitude. Then the trend reverses and the amplitude of the second pendulum decreases and that of the first pendulum increases and this kind of exchange goes on.

A good example of coupled oscillators is the vibration of atoms in a molecule. The various modes of vibration of the linear tri-atomic molecule CO_2 are depicted in Fig. 1.17 and those of the non-linear molecule H_2O are shown in Fig. 1.18. Fig. 1.17 (a) depicts the oxygen atoms oscillating in phase, while the carbon atom moves in the opposite direction. Fig. 1.17 (b) depicts the oxygen atoms oscillating out of phase while the carbon atom remaining fixed. Fig. 1.17 (c) shows the oxygen atoms oscillating in a direction perpendicular to the line joining them, which leads to the bending of the molecule. Fig. 1.18 shows the normal modes of vibration of the H_2O molecule.

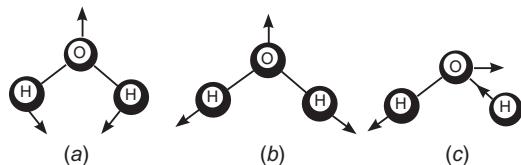


Fig. 1.17



Fig. 1.18

1.9 WAVES

The equations (1.3) and (1.5) describe the simple harmonic motion of a single point. Whenever a disturbance is created in a material medium, such as due to a stone thrown into still water, a local displacement of particles from equilibrium is caused. When a particle in a medium is disturbed, intermolecular forces provide the restoring force. The interactions of the particle with the next adjacent particle tend to return the former to its original position, and the latter begins to oscillate. In so doing, it affects the adjacent molecules, which are in turn set into oscillation. In such a situation, we find that something that happens at A at t_1 causes a similar happening at B at a later time t_2 . This is referred to as propagation of the disturbance. A medium is a material, which supports the propagation of the disturbance. When a continuous and repetitive disturbance passes through a medium, it gives rise to a continuous propagation of energy. In this situation, a series of particles are set into identical oscillations in succession.

Such a repeating and periodic disturbance, which propagates through a medium from one location to another, is called a **wave**. When a wave is present in a medium, the individual particles of the medium are *only temporarily* displaced from their equilibrium position. There is always a force acting upon the particles, which restores them to their original position. Note that the disturbance may take any of a number of shapes, from a finite width pulse to an infinitely long sine wave. A pulse is a single disturbance moving through a medium from one location to another location.

When we observe ripples in a pond, what we see actually is a rearrangement of the surface of water. Without water, there could be no wave. The interesting point here is that a wave transports its energy without transporting matter. The energy supplied by a stone (the agent of disturbance) is transferred from one point to another. Energy is transported through the medium, yet the water molecules are not transported. Therefore, waves are said to be an **energy transport phenomenon**. In conclusion, a wave can be described as *any disturbance, which travels through the medium due to the repeated periodic motion of the particles (of the medium) about their mean position and transporting energy from one location (its source) to another location without transporting matter*. Note that any wave moving through a medium has a source. Somewhere along the medium, there was an initial displacement of one of the particles.

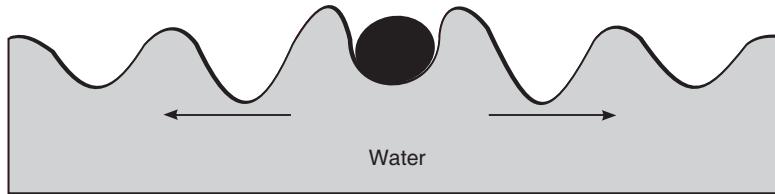


Fig. 1.19

We are familiar with waves on water surface. When a pebble is thrown into still water of a pond, ripples are produced which expand in the form of circles. The water wave has a *crest* and a *trough* and travels from one location to another. One crest is followed by a second crest which is followed by a third crest. Every crest is separated by a trough to create an alternating pattern of crests and troughs (Fig. 1.19). This mental picture of water waves is highly useful for understanding the nature of a wave. In the case of water waves, we identify the wave motion with the help of crests and troughs travelling away from the centre of disturbance. Waves such as those we see on the surface of water, which move away from the centre of disturbance are called **travelling waves** or **progressive waves**.

1.9.1 Travelling Waves

If a snapshot of a progressive wave is taken at any instant, we observe a wave profile as in Fig. 1.20. It consists of a sequence of waveforms.

Any wave is characterized by the following parameters.

(a) **Time Period, T:** If a point is chosen and the wave profile is observed as it passes this point, then the profile is seen to repeat at equal intervals of time. This repeat time is known as the *time period* of the wave.

(b) **Wavelength, λ :** The distance between the corresponding points, such as two successive crests, in successive waveforms is called the *wavelength*.

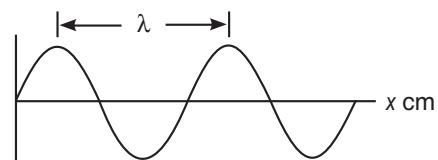


Fig. 1.20

(c) Amplitude, A: The maximum displacement in a waveform is known as the *amplitude*.

(d) Velocity, v: Each time the source (of disturbance) vibrates once, the wave moves forward a distance λ . If there are v vibrations in one second, the wave moves forward a distance of ' $v\lambda$ '. The distance that the wave moves in one second is the *velocity* of the wave, v . Thus,

$$v = v\lambda \quad (1.35)$$

where

$$v = \frac{1}{T} \quad (1.36)$$

(e) Phase angle, ϕ : The displacement of particles in the medium and the direction of their displacement change from point to point along the wave. The quantity, which represents the displacement, is called the *phase of the vibration*, ϕ . The phase may be expressed in terms of degrees or radians; or as the ratio of time t to the time period T ; or as the ratio of the distance x to the wavelength λ . The ratios t/T and x/λ are fractional numbers and have a maximum value of 1. When expressed in terms of radians (or degrees), the maximum value that the phase can take is 2π radians (or 360°).

(f) Intensity, I: The energy transferred on an average by a wave in unit time, through a unit area perpendicular to its propagation direction, is known as the *intensity* of the wave. It is established that the intensity of a wave is directly proportional to the square of the amplitude of the wave. Thus,

$$I \propto |A|^2 \quad (1.37)$$

1.9.2 Wave Equation

Equation of motion of an object is the equation that gives the position of the object as a function of time. We obtain the entire picture of wave motion only when we consider the harmonic motion of a series of points in the medium. As the oscillations are communicated from point to point, the points in the medium will be in different states of oscillation at different times. The displacement of a particle in the medium is therefore a function of space coordinates as well as a function of time. We denote the displacement by y . Thus,

$$y = f(x, t) \quad (1.38)$$

The displacement y is sometimes called the **wave function**.

Let us consider the case of a one-dimensional wave moving along + x -axis, as in Fig. 1.21.

We first consider the displacement as a *function of time*, at the position $x = 0$. Then,

$$y = f(t)$$

Since the oscillations are sinusoidal, we can describe the displacement y in terms of *time* as

$$y = A \sin \omega t$$

or

$$y = A \sin 2\pi v t \quad (1.39)$$

The wave is travelling forward to the right with a velocity, say v . Then after time t , the wave has moved through the distance $x = v t$. The displacement at x can be represented by

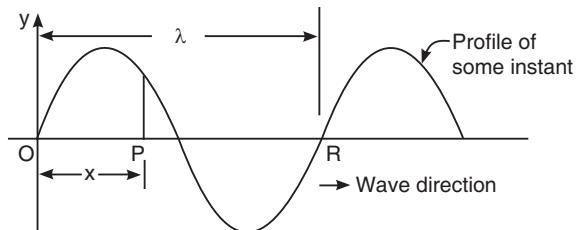


Fig. 1.21

$$\begin{aligned}y &= f(x - vt) \\v &= v\lambda\end{aligned}\tag{1.40}$$

Also $v = \frac{x}{t}$. Therefore, $v\lambda = x/t$.

or $v = \frac{x}{\lambda t}$ (1.41)

We can rewrite the relation (1.39) using (1.41) as

$$y = A \sin 2\pi \left(\frac{x}{\lambda} \right) \tag{1.42}$$

This describes the displacement in terms of *space*.

Using the equations (1.40) and (1.42), we can describe the displacement of any point on a harmonic wave in terms of both space and time as

$$y = A \sin \left[\frac{2\pi}{\lambda} (x - vt) \right] \tag{1.43}$$

This equation gives the relationship between the space and time dependence of disturbances in a medium. It is seen from the above that the wave is periodic in both space and time.

The equation (1.43) may be rewritten as

$$y = A \sin k(x - vt) \tag{1.44}$$

where $k = \frac{2\pi}{\lambda}$.

k is known as **propagation constant or wave number**.

The equation (1.43) may further be rewritten as

$$y = A \sin (kx - \omega t) \tag{1.45}$$

The above equation can be made independent of the system of coordinates by converting it into vector form. Let vector \mathbf{k} have a magnitude equal to the wave number k and a direction parallel to the positive direction of the x -axis. Such a vector is called a **wave vector**. Using \mathbf{k} into equ. (1.45), we get

$$y(x, t) = A \sin (\mathbf{k} \cdot \mathbf{x} - \omega t)$$

In the most general case, where r is any arbitrary direction, we replace x by r and write

$$y(r, t) = A \sin (\mathbf{k} \cdot \mathbf{r} - \omega t) \tag{1.45a}$$

1.9.3 General Wave Equation

To know how the displacement y varies as a function of space x and time t we have to do partial differentiation of y with respect to x and y in equ. (1.44).

$$\frac{\partial y}{\partial x} = \frac{2\pi}{\lambda} A \cos \left[\frac{2\pi}{\lambda} (x - vt) \right] \tag{1.46a}$$

$$\frac{\partial y}{\partial t} = -\frac{2\pi v}{\lambda} A \cos \left[\frac{2\pi}{\lambda} (x - vt) \right] \tag{1.46b}$$

Combining both these equations and eliminating equal factors, we get

$$\frac{\partial y}{\partial x} = -\frac{1}{v} \frac{\partial y}{\partial t} \tag{1.47}$$

If we take the second derivatives, it will hold for any sinusoidal wave, independent of the direction of travel, either $-x$ or $+x$.

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2} \quad \text{or} \quad \frac{\partial^2 y}{\partial t^2} = v^2 \frac{\partial^2 y}{\partial x^2}$$

We replace y by the more general term ξ , which stands for any disturbance.

$$\frac{\partial^2 \xi}{\partial t^2} = v^2 \frac{\partial^2 \xi}{\partial x^2} \quad (1.48)$$

This is the *one-dimensional wave equation*. It connects the variations in space and time to the velocity of propagation of the wave.

If we are to include waves propagating in any direction, we need to extend the right hand term to the y and z -axes, and replace it by

$$\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2} + \frac{\partial^2 \xi}{\partial z^2}$$

Using the Laplacian operator $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$, we can write the equation as

$$\frac{\partial^2 \xi}{\partial t^2} = v^2 \Delta^2 \xi \quad (1.49)$$

This is the *general three-dimensional wave equation*.

Example 1.8: The wave function for a light wave is given by,

$$E(z, t) = 10^3 \sin \pi(3 \times 10^6 x - 9 \times 10^{14} t).$$

Determine the speed, wavelength and frequency of the wave.

Solution: The given equation resembles the general equation,

$$E(z, t) = E_0 \sin k(x - vt) \quad (a)$$

The given equation may be written as

$$E(z, t) = 10^3 \sin 3 \times 10^6 \pi(x - 3 \times 10^8 t) \quad (b)$$

Comparing (b) with (a), we find that,

$$v = 3 \times 10^8 \text{ m/s} \quad \text{and} \quad k = 3 \times 10^6 \pi/\text{m}.$$

$$\text{As } k = \frac{2\pi}{\lambda}, \quad \lambda = \frac{2\pi}{k} = \frac{2\pi}{3 \times 10^6 \pi} = 6666 \text{ \AA}.$$

$$\text{Frequency, } v = \frac{\nu}{\lambda} = \frac{3 \times 10^8}{6666 \times 10^{-10}} = 4.5 \times 10^{14} \text{ Hz.}$$

Example 1.9: A progressive sinusoidal wave is represented by $y(x, t) = A \sin [(0.2 \text{ m}^{-1})x - (0.4 \text{ s}^{-1})t + \pi/6]$ where x and t are in meter and second respectively. Determine the speed of propagation of the wave. (B.P.U.T. 2004)

Solution: Here, we know that,

$$\begin{aligned} y &= A \sin (\omega t - kx + \Phi) \\ \therefore \omega &= 0.4 \text{ s}^{-1}, \quad k = 0.2 \text{ m}^{-1} \quad \text{and} \quad \phi = \pi/6 \end{aligned}$$

$$\text{Now the speed of propagation of the wave is given by, } v = \frac{\omega}{k} \quad \therefore \quad v = \frac{0.4}{0.2} = 2 \text{ m/s}$$

1.10 TYPES OF WAVES

Waves occur in many shapes and forms. One way to classify waves is on the basis of the direction of movement of the individual particles of the medium with respect to the direction

in which the waves travel. On this basis we can distinguish basically two types of waves: transverse waves, and longitudinal waves.

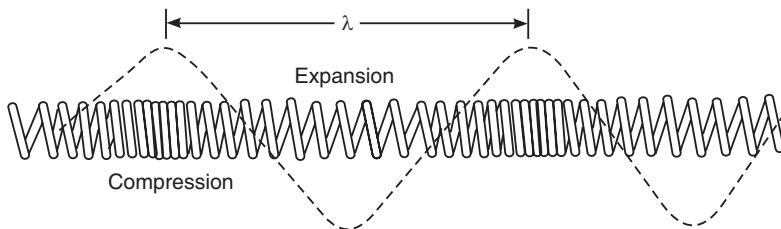


Fig. 1.22

A **longitudinal wave** is a wave in which particles of the medium move in a direction *parallel* to the direction, along which the wave moves. When a disturbance occurs in an elastic medium it generates an *elastic wave*. The elastic wave is a train of periodically alternating regions of compression and rarefaction; the particles of the medium being displaced in the direction of propagation of the wave (see Fig. 1.22). A wave, in which the displacement of each point of the medium is *back and forth* along the direction of propagation of the wave, is called a *longitudinal wave* or *compressional wave*. Note that longitudinal waves are always characterized by particle motion being *parallel* to wave motion. For example, if one end of a spring is pushed and then pulled, a longitudinal disturbance will propagate along the spring. Sound waves are longitudinal waves.

A **transverse wave** is a wave in which particles of the medium move in a direction *perpendicular* to the direction along which the wave moves. A wave, in which the direction of the displacement at each point of the medium is at *right angles* to the direction of wave propagation, is said to be *transverse*. The particles are displaced in a vertical direction, while the wave propagates in a horizontal direction. Note that transverse waves are always characterized by particle motion being perpendicular to wave motion (Fig. 1.21). Transverse can exist on the surface a liquid, the restoring force being supplied by the surface tension of the liquid. Therefore, ripples on water surfaces are transverse waves.

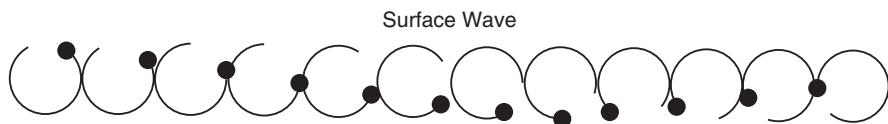


Fig. 1.23

Some waves are neither transverse nor longitudinal, but combinations of the two. For example, waves travelling through a solid medium can be either transverse waves or longitudinal waves. But waves travelling through the bulk of a fluid (such as a liquid or a gas) are *always* longitudinal waves. Transverse waves require a relatively rigid medium in order to transmit their energy. As one particle begins to move it must be able to exert a pull on its nearest neighbor. If the medium is not rigid as is the case with fluids, the particles will slide past each other. This sliding action, which is characteristic of liquids and gases, prevents one particle from displacing its neighbor in a direction perpendicular to the energy transport. It is for this reason that only longitudinal waves are observed moving through the bulk of liquids. The waves, which travel along the surface of the oceans, are referred to as surface waves. A **surface wave** is a wave in which particles of the medium undergo a circular motion. When

a water wave travels on the surface of deep water, water molecules at the surface move in nearly circular paths as shown in Fig. 1.23.

Another example of waves with both longitudinal and transverse motion may be found in solids as **Rayleigh surface waves**. The particles in a solid, through which a Rayleigh surface wave passes, move in elliptical paths, with the major axis of the ellipse perpendicular to the surface of the solid. As the depth into the solid increases the “width” of the elliptical path decreases. Rayleigh waves are different from water waves in one important way. In water wave all particles travel in clockwise circles. However, in a Rayleigh surface wave, particles at the surface trace out a *counter-clockwise* ellipse, while particles at a depth of more than 1/5th of a wavelength trace out *clockwise ellipses*.

1.10.1 Categories of Waves

Waves can also be classified according to the source that generates them. We group them mainly as mechanical waves, electromagnetic waves, matter waves, and gravitational waves.

Mechanical waves: Mechanical waves or elastic waves are governed by Newton’s laws and require a material medium for their propagation. Sound waves, seismic waves, water waves in bodies of water such as an ocean, river, and ponds are examples of mechanical waves.

Electromagnetic waves: Visible light, radio waves, microwaves, x-rays and γ -rays belong to this category. Electromagnetic waves consist of oscillating electric and magnetic fields and do not require material medium for their propagation. They all travel in free space with the same speed ‘ c ’.

Matter waves: Atomic particles exhibit wave properties under certain conditions. The laws of quantum mechanics govern such *matter waves*.

Gravitational waves: It is suggested that the cosmic bodies such as galaxies, stars produce gravitational waves and interact with each other through these waves. The gravitational waves are believed to propagate with the velocity of light.

1.11 REFLECTION AND TRANSMISSION OF WAVES AT A BOUNDARY

Whenever a travelling wave propagates through a medium, it may reach the end of the medium and encounter an obstacle or perhaps another medium through which it could travel.

When one medium ends, another

medium begins; the interface of the two media is referred to as the boundary. At the boundary, part or the entire travelling wave will be reflected. Let us consider a rope securely attached to a pole while the other end is held in hand. The end of the rope attached to the pole is referred to as a fixed end.

When the rope is given a jerk, a wave pulse is produced and it will travel through the rope towards the pole (Fig. 1.24 a). This pulse is called the incident pulse since it is incident towards the pole. When the incident pulse reaches the boundary (the pole), it will be reflected (See Fig. 1.24 b). Since the pole is assumed to be rigid, it does not transmit any part of the disturbance to the pole. The pulse, which returns to the left after bouncing off the pole, is known as the reflected pulse. It is observed that the reflected pulse is inverted. That is, if a crest is incident towards a fixed end boundary, it will reflect and return as a trough. Similarly, if a trough is incident towards a fixed end boundary, it will reflect and return as a crest. We

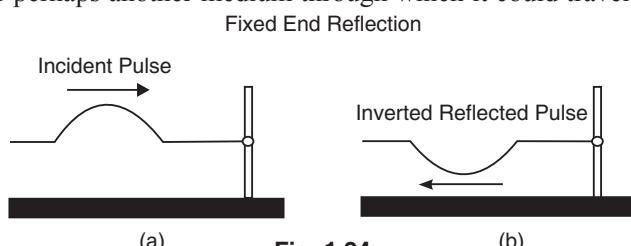


Fig. 1.24

can explain the inversion of the reflected pulse as follows. When the pulse reaches the end of the rope attached to the pole, the rope produces an upward force on the pole. By Newton's third law, the pole must then exert an equal and opposite reaction force on the rope. This force causes the pulse to invert after reflection.

The speed of the incident and reflected pulses are identical since the two pulses are travelling in the same medium. Secondly, every particle within the rope will have the same frequency. Being

connected to one another, they must vibrate at the same frequency. Since the wavelength of a wave depends upon the frequency and the speed, two waves having the same frequency and the same speed must also have the same wavelength.

Let us now consider the case where the other end of the rope was free to move. Instead of being securely attached to a pole, suppose it is attached to a ring, which is loosely fit around the pole. Because the right end of the rope is no longer secured to the pole, the last particle of the rope will be able to move when a disturbance reaches it. This end of the rope is referred to as a free end. When a pulse generated at the left end of the rope reaches the free end, it will be reflected, but the reflected pulse is not inverted (see Fig. 1.25). When a crest is incident upon a free end, it returns as a crest after reflection; and when a trough is incident upon a free end, it returns as a trough after reflection. As the pulse reaches the pole, it exerts a force on the free end, causing the ring to accelerate upward. In the process, the ring overshoots the height of the incoming pulse and is then returned to its original position. This produces a reflected pulse that is not inverted. Thus, *inversion is not observed in free end reflection*.

Let us next consider a boundary, which is neither rigid nor free. For instance, let us take the case of a light rope (thin rope) attached to a heavier rope, with each rope held at opposite ends (Fig. 1.26). When a pulse travelling on the thin rope reaches the boundary with a more dense medium (thick rope), a part of the incident pulse is reflected and returns towards the left end of the thin rope.

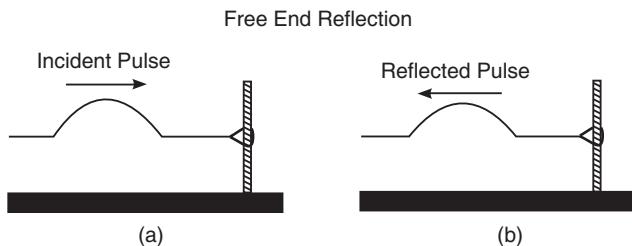


Fig. 1.25

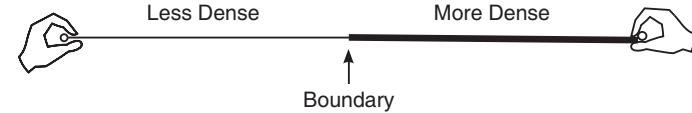


Fig. 1.26

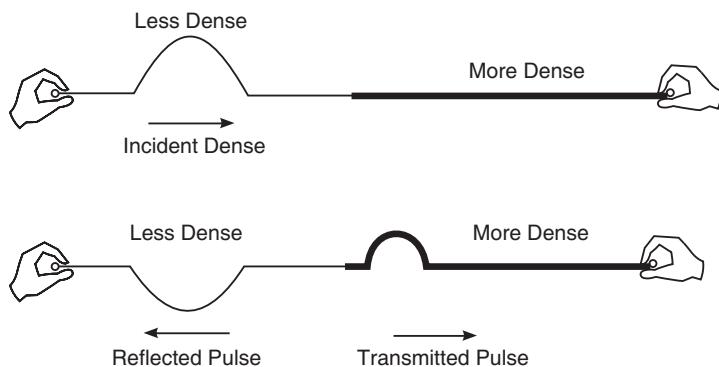


Fig. 1.27

A part of the incident pulse is transmitted into the thick rope as the transmitted pulse. It will be found that the reflected pulse is inverted whereas the transmitted pulse is not inverted (see Fig. 1.27).

Further, the transmitted pulse (in the denser medium) travels slower than the reflected pulse (in the less dense medium) and the transmitted pulse (in the denser medium) has a smaller wavelength than the reflected pulse (in the less dense medium). It is obvious that the speed and the wavelength of the reflected pulse are the same as the speed and the wavelength of the incident pulse.

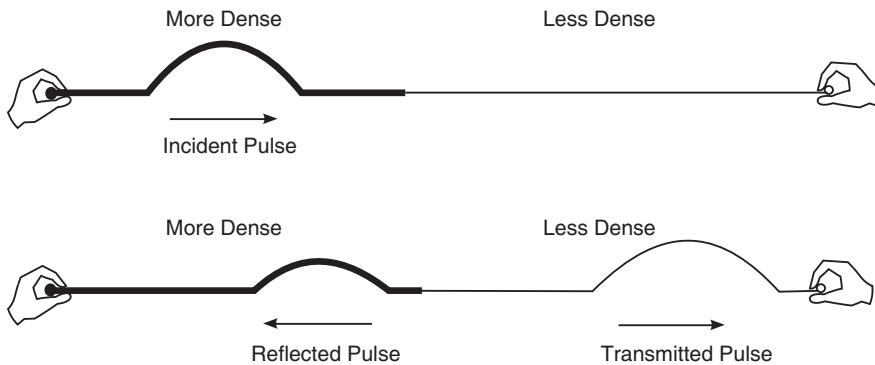


Fig. 1.28

Finally, let us consider a thick rope attached to a thin rope, with the incident pulse originating in the thick rope. If this is the case, there will be an incident pulse travelling in the denser medium (thick rope) towards the boundary with a less dense medium (thin rope). Once more, there will be partial reflection and partial transmission at the boundary. The reflected pulse in this situation will not be inverted. Similarly, the transmitted pulse is not inverted (as is always the case). Since the incident pulse is in a heavier medium, when it reaches the boundary, the first particle of the less dense medium does not have sufficient mass to overpower the last particle of the denser medium. The result is that a crest incident towards the boundary will reflect as a crest; for the same reasons, a trough incident towards the boundary will reflect as a trough.

The transmitted pulse (in the less dense medium) is travelling faster than the reflected pulse (in the denser medium) and the transmitted pulse (in the less dense medium) has a larger wavelength than the reflected pulse (in the denser medium). Further, the speed and the wavelength of the reflected pulse are the same as the speed and the wavelength of the incident pulse.

We draw the following conclusions regarding the boundary behavior of waves:

- The wave speed is always greatest in the least dense medium,
- The wavelength is always greatest in the least dense medium,
- The frequency of a wave is not altered by crossing a boundary,
- The reflected pulse becomes inverted when a wave gets reflected at a denser medium,

The amplitudes of the reflected and transmitted waves may be expressed in terms of the amplitudes of the incident wave and the densities of the string as

$$\frac{A_r}{A_i} = \frac{\sqrt{\mu_1} - \sqrt{\mu_2}}{\sqrt{\mu_1} + \sqrt{\mu_2}} \quad \text{and} \quad \frac{A_t}{A_i} = \frac{2\sqrt{\mu_1}}{\sqrt{\mu_1} + \sqrt{\mu_2}} \quad (1.50)$$

The above relations are valid for *any* type of harmonic wave that meets a discontinuity in a medium.

Note the following special cases concerning the propagation of the wave.

- (i) If $\mu_2 > \mu_1$ in the equation (1.48), the ratio A_r / A_i is negative and the displacement of the reflected pulse is *opposite* to that of the incident pulse. It means that the reflected pulse has a phase difference of π with respect to the incident pulse (see Fig. 1.27).
- (ii) If $\mu_2 < \mu_1$, then the ratio A_r / A_i is always positive and the displacement of the reflected pulse is the *same* as that of the incident pulse.
- (iii) Suppose that the string is fastened to a fixed support at $x = 0$. This is equivalent to letting μ_2 become indefinitely large. Then, $A_r / A_i = -1$ and the entire wave is reflected with a phase change of π (see Fig. 1.24).

1.12 PRINCIPLE OF SUPERPOSITION

It often happens that two or more waves propagate simultaneously through the same region in the same direction. They pass through one another as if the other wave is not present. When two pebbles are dropped at different points in a pond, the expanding water waves cross each other without either one producing any change in the other. Similarly, sound waves from different instruments in an orchestra propagate in space independent of each other and can be distinguished separately. There occur many such instances in which a number of waves meet and pass through each other without mutual effect. However, the waves act simultaneously on the particles of the medium, in the region in which they are overlapping on each other. In the region of overlap, the waves will simply add to (or subtract from) one another without permanently disrupting each other (see Figs. 1.30 & 1.31). Once having passed through the region, each wave will move out and away without getting affected by the overlap. The resultant displacement of the medium at the location of the overlap will be different from the sum of the displacements caused by the waves individually. The resultant displacement at any point and at any instant of time can be found using the **principle of superposition**. According to this principle the instantaneous displacement of the medium at any point in space or time, is simply the linear sum of the individual displacements that would have occurred for each wave alone. The principle of superposition states that

when a number of waves pass through a medium simultaneously, the instantaneous resultant displacement of the medium at every instant is the algebraic sum of the displacements of the medium due to individual waves in the absence of others.

If y_1, y_2, y_3, \dots are the displacement vectors due to waves 1, 2, 3, ... acting separately, then the resultant displacement y is given by

$$y = y_1 + y_2 + y_3 + \dots \quad (1.51)$$

As an example we consider two waves travelling simultaneously along the same path. Let $y_1(x, t)$ and $y_2(x, t)$ be the displacements that the medium would experience if each wave acted alone. When both the waves act, then the displacement of the medium is

$$y(x, t) = y_1(x, t) + y_2(x, t)$$

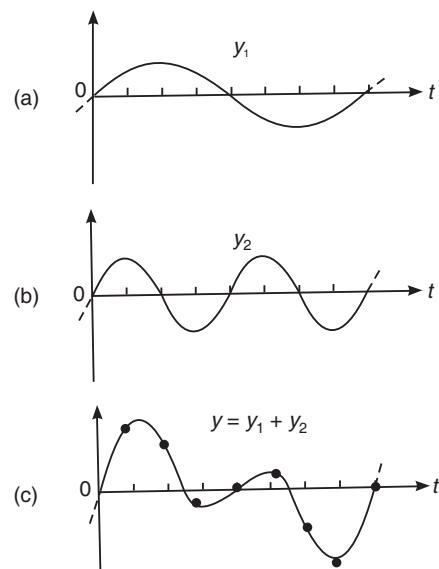


Fig. 1.29

the sum being an algebraic sum (see Fig. 1.29). The superposition principle holds as long as the amplitudes of the waves are not very large. This principle is applicable to all kind of waves and is very useful in the study of sound waves, electromagnetic waves and quantum physics also. This is true of waves which are finite in length (wave pulses) or which are continuous sine waves.

The superposition of the waves may result in the following cases;

- (i) The superposition of two waves of the same frequency moving in the same direction leads to **interference**.
- (ii) The superposition of two waves of slightly different frequencies moving in the same direction leads to **beats**.
- (iii) The superposition of two waves of the same frequency moving in the opposite direction leads to **stationary waves**.

1.12.1 Interference

The superposition of two or more pulses (waves) in a given region may give rise to interference. When the two interfering wave pulses have a displacement in the same direction, the resultant displacement is greater than the displacement of either wave. This type of interference is called **constructive interference**. Constructive interference is observed when a crest meets a crest; and when a trough meets a trough as shown in Fig. 1.30.

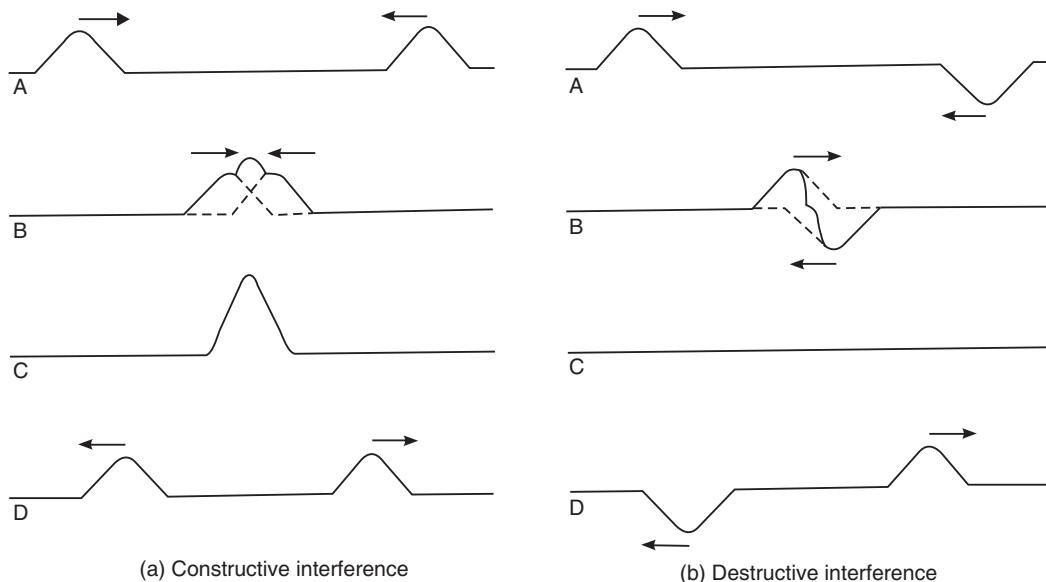


Fig. 1.30

Fig. 1.31

On the other hand, when the pulses have displacements in opposite directions, the resultant displacement is smaller than that of either pulse. This type of interference is called **destructive interference**.

In Fig. 1.31, the interfering pulses have the same maximum displacement but in opposite directions. They completely destroy each other when they have completely overlapped. At the instant of complete overlap, there is no resulting disturbance in the medium. This “destruction” is *not a permanent condition*. Destructive interference leads to only a momentary condition in which the displacement of the medium is zero. At the point of total destructive interference,

when the net wave shape and hence potential energy are zero, the wave energy is stored in the medium completely in the form of kinetic energy.

1.13 STATIONARY WAVES

Standing waves are produced whenever two waves of equal frequency amplitude interfere with one another while travelling in **opposite directions** along the same medium.

Let the two component waves be represented by

$$y_1(x, t) = A \sin(kx - \omega t) \quad (1.51a)$$

and

$$y_2(x, t) = A \sin(kx + \omega t) \quad (1.51b)$$

Using the principle of superposition, the resulting string displacement may be written as:

$$y(x, t) = y_1(x, t) + y_2(x, t) = A \sin(kx - \omega t) + A \sin(kx + \omega t) \quad (1.52)$$

By using the identity $\sin A + \sin B = 2 \sin[(A + B)/2] \cos[(A - B)/2]$, we simplify the above equation to obtain

$$y(x, t) = 2A \cos(\omega t) \sin kx \quad (1.53)$$

This wave is no longer a travelling wave because the position and time dependence have been separated. The displacement of the string as a function of position has an amplitude of $2A \sin kx$. This amplitude does not travel along the string, but stands still and oscillates up and down according to $\cos \omega t$.

The formation of a stationary wave can be represented graphically as follows.

Consider two waves *A* and *B* of the same amplitude, and frequency travelling in opposite directions. At an instant of time $t = 0$, the waves are as shown in Fig. 1.32 (a). The resultant displacement is a straight line. All the particles of the medium (depicted 1, 2, 3, 4, 5, 6 and 7) are at their equilibrium position.

Consider the waves after a time $t = T/4$. During this time, the wave *A* will advance through a distance $\lambda/4$ towards right, and the wave *B* will advance through a distance $\lambda/4$ towards the left. The resultant displacement pattern is shown in Fig. 1.32 (b).

The particles at 1, 3, 5, and 7 undergo maximum displacement and the particles at 2, 4, and 6 are at their equilibrium positions.

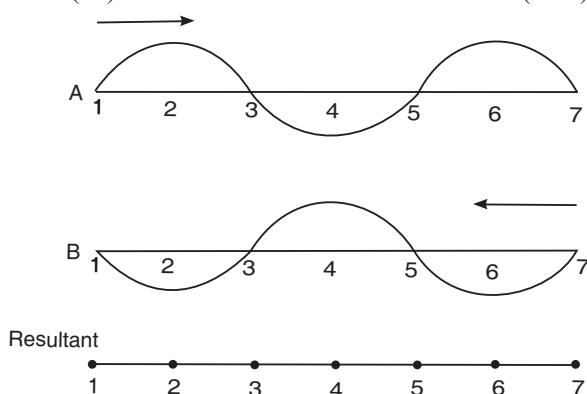


Fig. 1.32 (a)

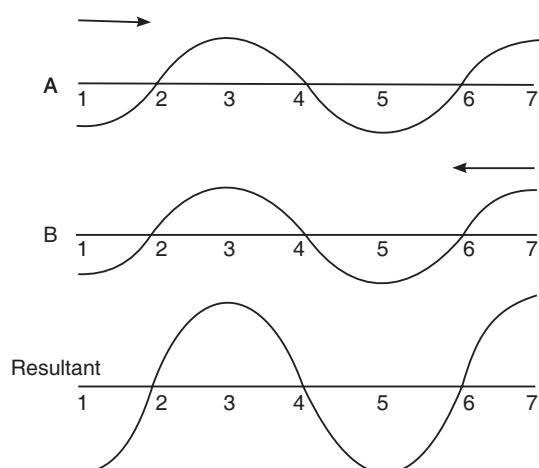


Fig. 1.32 (b)

At time $t = T/2$, the wave A will advance through a distance $\lambda/2$ towards right while the wave B advances through a distance $\lambda/2$ towards the left. The resultant displacement pattern is shown in Fig. 1.32 (c).

All the particles of the medium are at their equilibrium positions.

At time $t = 3T/4$, the wave A will advance through a distance $3\lambda/4$ towards right and the wave B will advance through a distance $3\lambda/4$ towards left. The resultant displacement pattern is shown in Fig. 1.32(d).

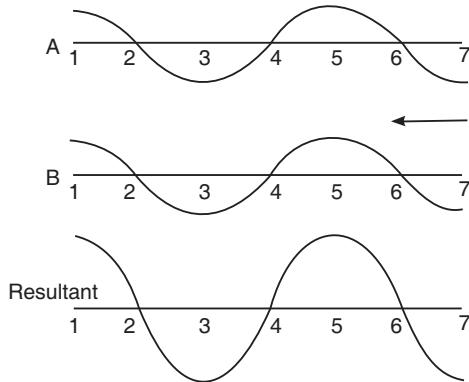


Fig. 1.32 (d)

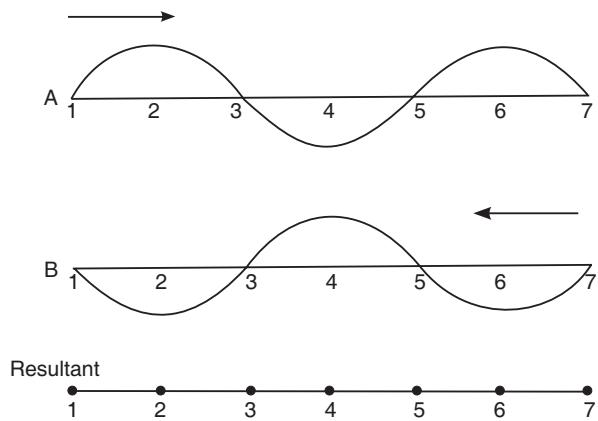


Fig. 1.32 (c)

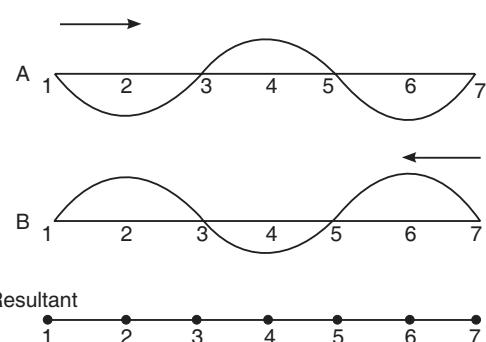


Fig. 1.32 (e)

The particles 1, 3, 5 and 7 have undergone maximum displacement and particles at 2, 4, and 6 are at their mean positions.

At time $t = T$, the wave A will advance through a distance λ towards right and the wave B will advance through a distance λ towards left. The waves are as shown in Fig. 1.32 (e). The resultant displacement is a straight line. All the particles of the medium (depicted 1, 2, 3, 4, 5, 6 and 7) are at their equilibrium position.

From the figures, it is clear that the particles of the medium such as 2, 4, and 6 etc. always remain at their equilibrium positions. The particles such as 1, 3, 5, 7 etc. continue to vibrate simple harmonically about their equilibrium positions with double the amplitude of each wave, as shown in Fig. 1.32 (f). It appears as though the wave pattern is stationary in space.

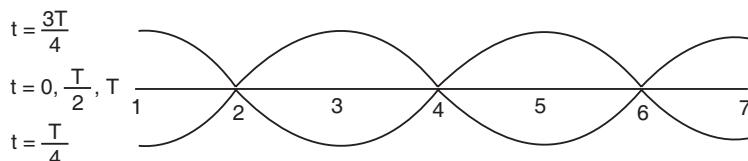


Fig. 1.32 (f)

The positions of the particles 2, 4, 6, etc. which always remain at their mean positions, are called **nodes**. Node is a position of zero displacement and maximum strain. The positions of the particles 1, 3, 5, 7, etc. which vibrate simple harmonically with maximum amplitude are called **antinodes**. From equ. (1.59), it is easy to see that the nodes are produced where $\sin kx = 0$, that is, where $kx = 0, \pi, 2\pi$ and the antinodes occur at points where $\sin kx = \pm 1$, that is, where $kx = \pi/2, 3\pi/2, 5\pi/2$, etc. The distance between any two consecutive nodes or antinodes is equal to $\lambda/2$. Between a node and an antinode, the amplitude gradually increases from zero to maximum.

A standing wave pattern is not actually a wave; rather it is the pattern resulting from the presence of two waves of the same frequency with different directions of travel within the same medium. Standing wave patterns are characterized by certain fixed points along the medium, which undergo no displacement. These points of no displacement are called *nodes* (nodes can be remembered as points of no displacement). The nodes are always located at the same location along the medium, giving the entire pattern an appearance of standing still. There are other points along the medium, which undergo vibrations between a large *positive* and large *negative* displacement. Midway between every consecutive nodal point are points which undergo maximum displacement. These points are called anti-nodes. Anti-nodes are points along the medium, which oscillate between a large *positive* displacement and a large *negative* displacement during each vibrational cycle of the standing wave. In a sense, these points are the opposite of nodes, and so they are called antinodes. A standing wave pattern always consists of an alternating pattern of nodes and antinodes. When a standing wave pattern is established in a medium, the nodes and the antinodes are always located at the same position along the medium; they are "standing still." It is this characteristic which has earned the name "standing wave."

The nodes are produced at locations where destructive interference occurs. Antinodes, on the other hand, are produced at locations where constructive interference occurs. Antinodes are always vibrating back and forth between these points of large positive and large negative displacement; this is because during a complete cycle of vibration, a crest will meet a crest; and then one-half cycle later, a trough will meet a trough. Because antinodes are vibrating back and forth between positive and negative displacements, a diagram of a standing wave is sometimes depicted by drawing the shape of the medium at an instant in time and at an instant one-half vibrational cycle later. This is shown in Fig. 1.33.

Nodes and antinodes should not be confused with crests and troughs.

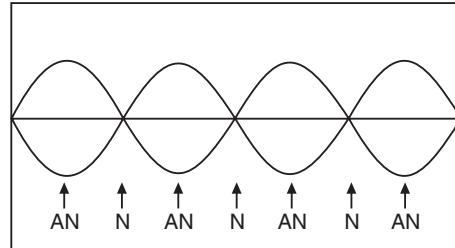


Fig. 1.33

Distinction between Travelling Waves and Standing Waves

	Travelling waves	Standing waves
1.	A travelling wave propagates in a medium continuously with a finite velocity.	A standing wave is stationary and does not move in the medium.
2.	A travelling wave transports energy from one location to the other. Hence there is energy flow across every plane in the direction of wave propagation.	There is no energy transfer in a standing wave. There is no flow of energy across any plane. The energy of oscillations periodically transform from kinetic energy to potential energy of the elastically deformed medium and vice versa.
3.	No particle on the wave is permanently at rest.	Nodes are permanently at rest in a standing wave.

4.	In a travelling wave all the points oscillate with the same amplitude regardless of their location.	All the points of a standing wave between two adjacent nodes oscillate with different amplitudes.
5.	In a travelling wave, different points oscillate with different phases.	In a standing wave, all the points between any pair of nodes oscillate in the same phase.
6.	In a travelling wave, all the particles do not pass through their mean positions or reach the extreme positions simultaneously.	In a standing wave, all the particles pass through their mean positions and reach their extreme positions simultaneously twice in each cycle.

Example 1.10: Standing waves are produced by the superposition of two waves, $y_1 = 10 \sin(3\pi t - 4x)$ and $y_2 = 10 \sin(3\pi t + 4x)$. Find the amplitude of the motion, at $x = 18$.

Solution: The resultant amplitude is given by

$$\begin{aligned} y &= y_1 + y_2 \\ &= 10 \sin(3\pi t - 4x) + 10 \sin(3\pi t + 4x) \\ &= 10[\sin 3\pi t \cos 4x - \cos 3\pi t \sin 4x + \sin 3\pi t \cos 4x + \cos 3\pi t \sin 4x] \\ &= 10[2 \sin 3\pi t \cos 4x] \\ &= 20 \cos 4x \sin 3\pi t \end{aligned}$$

The amplitude of motion is $(20 \cos 4x)$.

When $x = 18$, then $4x = 72 = [72 \times \pi/3.14]$ radians = 22.9π radians

$$\therefore \text{Amplitude} = 20[\cos(22.9 \pi)] = 20(0.9673) = 19.35 \text{ units of length.}$$

1.13.1 Harmonics

In general, standing waves form in a bounded medium. For instance, when a string is tied at both ends, standing waves set up, but set up only for a certain discrete set of frequencies. We then say that the system resonates at these frequencies. The standing wave patterns are called **oscillation modes**. Because the ends of the string cannot move, a node of the standing wave pattern must exist at each end of the string. Therefore, the length L of the string must be an integral multiple of $\lambda/2$. The allowed frequencies are then given by

$$v = \frac{\nu}{\lambda} = n \frac{\nu}{2L}, \quad n = 1, 2, 3, \dots \quad (1.54)$$

Each frequency is associated with a different standing wave pattern. These frequencies and their associated wave patterns are referred to as **harmonics**. The pattern with two nodes and one antinode is referred to as the first harmonic, that with three nodes and two antinodes is the second harmonic and are depicted in the Fig. 1.34.

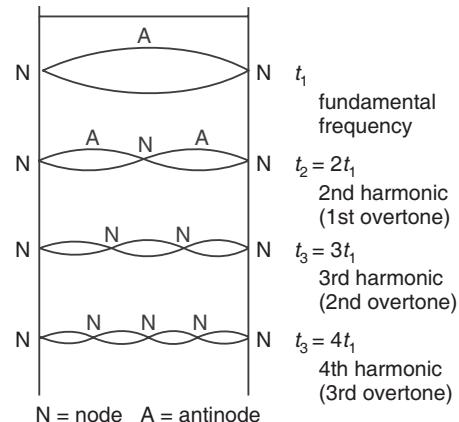


Fig. 1.34

1.14 SUPERPOSITION OF TWO PERPENDICULAR SHMs

Waves that are characterized by a **scalar wave function** always interfere when they are superposed. Thus, sound waves, which are characterized by pressure fluctuations, always interfere when they are superposed. On the other hand, the displacement of a rope and the electric field in a light wave are examples of a **vector wave function**. For such waves, interference of two waves can occur only if the oscillations lie along the same line. For instance, if a rope is along the x -axis, then a wave with displacements along the y -axis would not interfere with waves having displacements along the z -axis.

In the earlier discussions on superposition, we tacitly assumed that the oscillations of the two superposing vibrations occur in the same plane. In case of two transverse waves described by vector functions, the oscillations cannot lead to interference, as the component of the oscillations in one plane onto the oscillations in the perpendicular plane is zero. However, the oscillations can mix to produce elliptical vibrations, as shown in the following discussion.

Let us consider two waves vibrating in mutually perpendicular directions with the *same* frequency and travelling along the same direction, i.e., z -direction.

Let $x = A \cos(\omega t - kz)$ (1.55a)
and $y = B \cos(\omega t - kz + \delta)$ (1.55b)

where δ is the phase difference between the x and y -oscillations and A and B are the amplitudes of oscillations.

(a) If the two waves are in phase, then $\delta = 0$, and

$$x = A \cos(\omega t - kz) \quad \text{and} \quad y = B \cos(\omega t - kz)$$

By eliminating $\cos(\omega t - kz)$ from the above two equations, we get

$$y = \frac{B}{A}x \quad (1.56)$$

This is the *equation of a straight line*.

(b) If the two motions are in opposite phase, then $\delta = \pi$, and it is easy to see that in this case

$$y = -\frac{B}{A}x \quad (1.57)$$

Equ. (1.57) also represents a straight line.

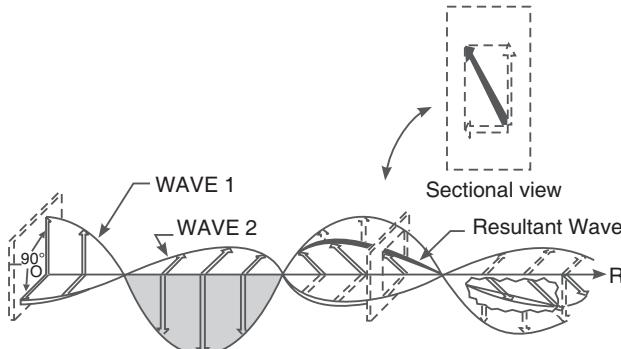


Fig. 1.35

(c) If the phase difference δ remains constant, the tip of the resultant vector E describes a certain closed curve in the xy -plane. When $\delta = \pi/2$, the oscillations differ by a phase of 90° (See Fig. 1.35). The equations now become

$$x = A \cos(\omega t - kz)$$

and

$$y = B \cos(\omega t - kz + \pi/2) = -B \sin(\omega t - kz)$$

By squaring and combining these relations, we get

$$\frac{x^2}{A^2} + \frac{y^2}{B^2} = 1 \quad (1.58)$$

Equation (1.58) is the general equation of an *ellipse*. Thus, at any particular time we find that the tip of the resultant vector traces out an ellipse. The same ellipse is obtained if $\delta = 3\pi/2$

or $-\pi/2$, but then the motion is in a counter-clockwise direction. Thus, we may say that when the phase difference δ is $\pm \pi/2$, the superposition of the two simple harmonic motions of the same frequency results in an **elliptical motion**.

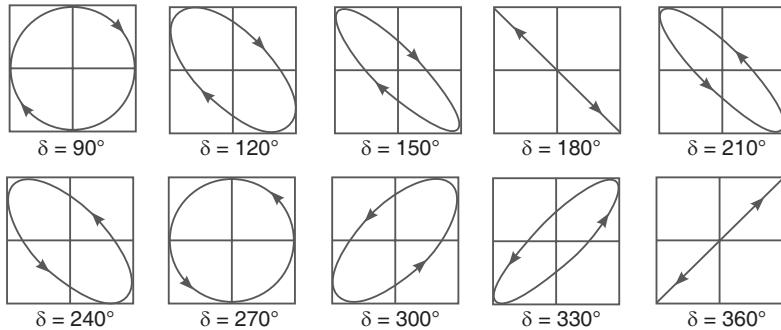


Fig. 1.36

(d) In the particular case, when $\delta = \pi/2$ and $A = B$, equ. (1.58) reduces to

$$x^2 + y^2 = A^2 \quad (1.59)$$

This is the equation of a *circle*. When $A = B$, the axes of the ellipse transforms into a circle and we have **circular motion**. That is, circular motion can be generated by combining two oscillatory motions of the same frequency and amplitude along perpendicular direction but with a phase difference of $\pm \pi/2$.

If the waves are of different frequencies,

$$x = A \cos(\omega_1 t - kz) \quad \text{and} \quad y = B \cos(\omega_2 t - kz)$$

The resultant path depends on the ratio $\omega_2 : \omega_1$ and on the phase difference δ . These paths are called **Lissajous figures** and some typical figures are shown in Fig. 1.37.

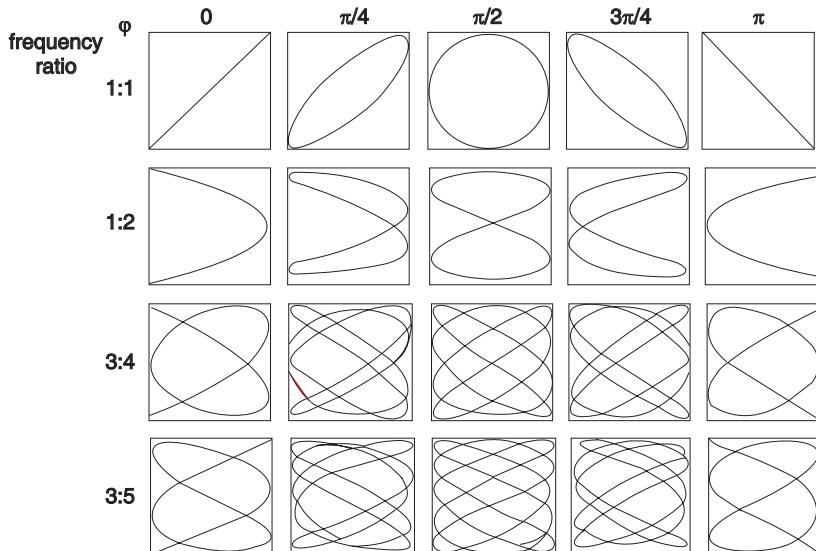


Fig. 1.37

1.15 DISPERSION

A truly sinusoidal wave has no beginning or end, either in space or time. In practice, waves are a mixture of waves of different frequencies, which travel with different speeds in a medium. For example, white light consists of waves of different frequencies. Similarly, a pulse is not a pure sinusoidal wave but consists of sinusoidal waves of different frequencies. A pulse travelling through a medium spreads out as it travels through a medium and undergoes a change in its shape. This phenomenon is known as **dispersion**. It is because of dispersion that sunlight spreads out into a spectrum of colours as it passes through a prism. The motion of waves in water is **dispersive**. On the other hand, waves on a stretched string travel at the same speed irrespective of their frequency. Similarly, pulses of sound waves have a single speed for all frequencies. Thus, these waves are *dispersion less*. In a non-dispersive wave medium, waves can propagate without deformation. Electromagnetic waves in unbounded free space are non-dispersive as well as non-dissipative and thus can propagate over astronomical distances. Sound waves in air are also nearly non-dispersive even in the ultrasonic frequency range. If not, that is, if high frequency notes (e.g., piccolo) and low frequency notes (e.g., base) propagate at different velocities, they would reach our ears at different times, and music played by an orchestra would not be harmonious. Most waves in material media are dispersive, however, and wave forms originally set up are bound to change in a manner that the wave energy is more spatially spread out or dispersed.

1.15.1 Phase Velocity

In our discussion on wave motion, a strictly single frequency sinusoidal wave train, usually called a monochromatic wave, is used to represent the characteristics of wave propagation. A monochromatic wave train is an infinite sequence of waves in time and space of crests and troughs. Following equation (1.38), the equation of a harmonic wave propagating along the x -axis has the following form

$$y = A \sin [(kx - \omega t) + \phi] \quad (1.60)$$

where ϕ is the initial phase of the wave which is determined by our choice of the beginning of counting x and t . Let us fix a value of the phase by assuming that

$$[(kx - \omega t) + \phi] = \text{constant} \quad (1.61)$$

This expression determines the relation between the time t and the place x where the phase has a fixed value. The value of $\frac{dx}{dt}$ calculated from (1.61) gives the velocity with which the given value of the phase propagates.

$$k \frac{dx}{dt} - \omega = 0$$

$$\frac{dx}{dt} = \frac{\omega}{k} = \frac{2\pi\nu}{2\pi/\lambda} = v\lambda = v \quad (1.62)$$

Thus, the velocity of wave propagation v is the velocity of phase propagation and it is therefore called the **phase velocity**. The phase velocity may be defined as the velocity of propagation of the wave front. When the waves are travelling through a non-dispersive medium, the common velocity of the waves is the phase velocity.

1.15.2 Wave Packet

A harmonic wave is characterized by a precise wavelength λ and constant amplitude. It is non-localized and extends over a volume of space. In general real waves are of complex forms. In practice waves are far from monochromatic and can be regarded as the result of superposition of waves of a number of frequencies, each component wave having its own propagation velocity in a medium.

The propagation velocity of a wave varies with frequency. The theory of Fourier analysis shows that a combination of harmonic waves with wave numbers spread over a range Δk will produce a **wave group** or **wave packet**. The superposition of a very large number of harmonic waves differing infinitesimally in frequency will produce a single wave packet. The waves cancel each other everywhere except in a small region. The wave packet is spread out in space over a length Δx .

1.15.3 Group Velocity

If a wave packet travels through a medium without changing its shape over a long distance, then the medium is said to be a non-dispersive medium. Most media in nature are dispersive and no waveform can preserve its shape over a reasonable propagation distance. The wave group generally has the maximum amplitude at a particular value of x and the velocity of this maximum amplitude point is called the **group velocity**. Thus, the velocity at which a wave group (or a pulse) travels is the group velocity of the wave group. This velocity also represents the velocity with which energy of the wave group is transmitted.

1.15.4 Relation between Group Velocity and Phase Velocity

A wave packet contains several harmonic waves of differing wavelengths. Each component wave has its own phase velocity, $v = v\lambda$. The wave packet has amplitude that is large in a small region and very small outside it. The amplitude of the wave packet varies with x and t . Such variation of amplitude is called the *modulation* of the wave. The velocity of propagation of the modulation is known as the *group velocity*, v_g . It is given by

$$v_g = \frac{d\omega}{dk} \quad (1.63)$$

$$v_g = \frac{d(vk)}{dk} = v + k \frac{dv}{dk}$$

We further write

$$\frac{dv}{dk} = \frac{dv}{d\lambda} \frac{d\lambda}{dk}$$

$$\text{But } \lambda = \frac{2\pi}{k}$$

Differentiating the above expression, we get

$$\frac{d\lambda}{dk} = -\frac{2\pi}{k^2} = -\frac{\lambda}{k}$$

$$\therefore k \frac{dv}{dk} = -\lambda \frac{dv}{d\lambda}$$

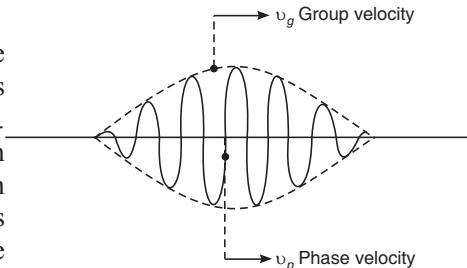


Fig. 1.38

$$v_g = v - \lambda \frac{dv}{d\lambda} \quad (1.64)$$

This is the relation that connects phase velocity and group velocity.

The *phase velocity* v of a wave is the velocity with which the wave front moves forward. It is the same as the velocity of propagation of the wave. When a number of plane waves of slightly different wavelengths travel in the same direction, they form wave groups or wave packets. The velocity with which the wave group advances in the medium is known as the *group velocity* v_g . Phase velocity is a characteristic of an individual wave whereas group velocity characterizes a group of waves. Group velocity will be the same as phase velocity if the entire constituent waves travel with the same velocity. It means in a non-dispersive medium, $v_g = v$. However, the waves of different wavelengths travel in a medium with different velocities. Therefore, the group velocity is in general less than the phase velocity.

QUESTIONS

1. What do you understand by periodic and simple harmonic motion? What is the criterion for the motion to be simple harmonic?
2. Write down the differential equation for the simple harmonic motion and formulae for the angular frequency and time period.
3. Discuss diagrammatically the phase relationship of the velocity and acceleration of a particle executing simple harmonic motion.
4. Show that for a simple harmonic oscillator, mechanical energy remains constant and it is proportional to the square of the amplitude.
5. How is the period of SHM changed when
 - (a) The mass of the particle is increased without changing the elastic constant?
 - (b) When the elastic constant is increased without changing the mass?
 - (c) When the mass and the elastic constant are changed by the same ratio?
6. Give some examples of simple harmonic oscillation.
7. Explain what is meant by natural frequency? Give the expression for the natural frequency of a simple pendulum.
8. Assuming the damping to be proportional to the velocity, write down the differential equation for a damped harmonic oscillator.
9. Give reason for the energy dissipation in the case of a damped harmonic oscillator.
10. A damped oscillator is subjected to a damping force proportional to its velocity. Set up the differential equation of the oscillation. Discuss the under-damped, over-damped and critical damped motions of the oscillator. (B.P.U.T. 2003)
11. Graphically show the displacement-time curve for oscillatory, over damped and critically damped motion of a damped oscillator. Mention the conditions of their occurrence. (B.P.U.T. 2004)
12. Why are damping devices often used on machinery? Give an example.
13. What are forced vibrations? Give two examples.
14. A forced oscillator is at resonance with the external periodic force. What is the phase difference between the driving force and the velocity of the oscillator? (B.P.U.T. 2004)
15. An oscillator is subjected to an external sinusoidal periodic force and a damping force proportional to its velocity. Set up a differential equation of the oscillator. Mention the condition under which velocity resonance occurs. (B.P.U.T. 2004)
16. Mention any two physical phenomena where energy resonance occurs.
17. Why are the forced oscillations of a damped oscillator not damped?
18. Why must the force and the velocity be in phase at energy resonance?

19. Give some examples of common phenomena in which resonance plays an important role.
20. Buildings of different heights sustain different amounts of damage in earthquake. Explain, why?
21. Give three examples of coupled oscillators.
22. What are normal modes?
23. The normal modes are independent of each other. Comment.
24. What are the main features of the normal modes of coupled oscillators?
25. Two simple pendulums of mass m and length l each, are coupled by a spring of force constant k . Write the expression for angular frequency of normal modes of vibration of the coupled system. **(B.P.U.T. 2003)**
26. Give the characteristic of the in-phase mode of the motion of two coupled oscillators.
27. In the in-phase mode, the frequency of oscillation is the same as of uncoupled oscillators whereas in the out-of-phase mode, the frequency of oscillation gets raised. Comment.
28. How do the atoms in a crystal constitute coupled system?
29. Distinguish between progressive and stationary waves.
30. Define the transverse and longitudinal waves.
31. State the wave equation for one-dimensional motion and explain the terms.
32. Mention the characteristics of wave motion. **(B.P.U.T. 2004)**
33. A wave transmits energy. Does it transfer momentum?
34. If two waves differ only in amplitude and propagate in opposite directions through a medium, will they produce standing waves?
35. In a stationary wave, the wavelength is 3.6 m. What is the distance between a node and the nearest antinode? **(B.P.U.T. 2004)**
36. Identify the terms in the wave function given by $\Psi(r, t) = A \sin(\mathbf{k} \cdot \mathbf{r} - \omega t + \phi)$ **(B.P.U.T. 2004)**
37. A wave packet propagates in a medium, which exhibits normal dispersion. Write the relationship between phase velocity and group velocity. Which is greater and why? **(B.P.U.T. 2004)**

PROBLEMS

1. A particle in SHM has velocities v_1 and v_2 when its displacement from the mean position is x_1 and x_2 respectively. Calculate its period, amplitude and maximum speed.

$$\left\{ \begin{aligned} \text{Ans : } T &= 2\pi \sqrt{\left(\frac{(x_2^2 - x_1^2)}{(v_1^2 - v_2^2)} \right)}, a = \sqrt{\left(\frac{(x_2^2 v_1^2 - x_1^2 v_2^2)}{(v_1^2 - v_2^2)} \right)}, v_{\max} = \left[\frac{(x_2^2 v_1^2 - x_1^2 v_2^2)}{(x_2^2 - x_1^2)} \right]^{1/2} \end{aligned} \right.$$

2. A particle executes SHM with a period of 0.002 s and amplitude 10 cm. Find its acceleration when it is 4 cm away from its mean position and also obtain its maximum velocity.

$$(\text{Ans: } a = -3.9 \times 10^7 \text{ cm/s}^2, v_{\max} = 3.14 \times 10^4 \text{ cm/s})$$

3. The length of a weightless spring increases by 2 cm when a weight of 1.0 kg is suspended from it. The weight is pulled down by 10 cm and released. Determine (i) Period of oscillation of spring and (ii) Kinetic energy of oscillation of the spring. **(Ans: $T = 0.29$ s, K.E. = 2.45 J)**

4. A body of mass 0.2 kg is hung from a spring of constant 80 N/m. The body is subjected to a resistive force given by ' bv ' where v is the velocity in m/s. Calculate the value of the undamped frequency, and the value of τ if the damped frequency is $\sqrt{3}/2$ of the undamped frequency.

$$(\text{Ans: } \tau = 510 \text{ sec}^{-1})$$

CHAPTER

2

Electrostatics

2.1 INTRODUCTION

If a glass rod is rubbed with silk, the rod acquires positive charge. If a hard rubber rod is rubbed with fur, the rod will acquire negative charge. This process is known as charging by friction. It means that the glass rod and rubber rods are *electrified*. Charged bodies attract each other if they have unlike charges or repel each other if they have like charges. It is subsequently discovered that electric charge is fundamentally associated with atomic particles, the electron and the proton. Electrons carry negative charge and protons carry a positive charge. Matter in its neutral state contains equal amounts of positive and negative charges. Accordingly, we interpret the electrification of the bodies as occurring due to transfer of charge. When two bodies are rubbed together, a redistribution of electrons takes place. The body which receives electrons becomes negatively charged while the body which loses electrons becomes positively charged. The study of electric forces between charged objects **at rest** is called **electrostatics**.

2.2 ELECTRIC CHARGES

The charges either positive or negative are always built up as a collection of elementary charges, carried by fundamental particles protons and electrons. They are always an integral multiple of the smallest unit of charge, that of an electron. When a body is said to be charged, it contains either an excess of electrons or a shortage of electrons. The charge residing on a charged body is

$$Q = \pm ne \quad (2.1)$$

where n is an integer taking values 1,2,3,....

The SI unit of charge is the Coulomb denoted by C. The value of e is

$$1e = 1.602 \times 10^{-19} \text{ C}$$

2.3 COULOMB'S LAW

If two charges are brought nearer, they exert forces on each other. We say that the charges are **interacting**. If the charges are **at rest**, their interaction is known as **electrostatic interaction**. The electrostatic interaction for two charged particles is given by Coulomb's law.

The Law

The electrostatic interaction between two charge particles is proportional to the square of the distance between them and its direction is along the line joining the two charges.

If q_1 and q_2 are two **point charges at rest**, and are separated by a distance r , they exert a force on each other which is given by

$$\mathbf{F} = k \frac{q_1 q_2}{r^2} \quad (2.2)$$

where k is a constant which has the following value in the SI system.

$$k = \frac{1}{4\pi\epsilon_0} \quad (2.3)$$

Here ϵ_0 is called the **permittivity of free space** and has the value

$$\begin{aligned}\epsilon_0 &= 8.85 \times 10^{-12} \text{ C}^2 / \text{Nm}^2 \\ \therefore k &= 9 \times 10^9 \text{ Nm}^2 / \text{C}^2\end{aligned}\quad (2.4)$$

When the charges are located in a dielectric medium having a **relative permittivity** of ϵ_r , the force is given by

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0\epsilon_r} \frac{q_1 q_2}{r^2} \quad (2.5)$$

The Coulomb's force is maximum when the charges in a vacuum and reduces when the charges are placed in a medium.

Vector form of the law

The Coulomb's law (2.2) may be expressed in vector form as

$$\mathbf{F} = k \frac{q_1 q_2}{r^2} \hat{\mathbf{r}} = k \frac{q_1 q_2 \mathbf{r}}{r^3} \quad (2.6)$$

In the above equation, $\hat{\mathbf{r}}$ is the unit vector along \mathbf{r} . It is given by $\hat{\mathbf{r}} = \frac{\mathbf{r}}{r}$.

- In equ. (2.6), \mathbf{F} is the force produced by the charge q_1 on the charge q_2 . The force produced by the charge q_2 on the charge q_1 is therefore $-\mathbf{F}$.
- The charges are assumed to be at rest; otherwise we need to take into account the magnetic forces.
- The size of the charges must be very small compared to their separation. Hence, the charges are assumed to be point charges.

Example 2.1: Two equal and similar charges 3 cm apart in air repel each other with a force equivalent to 1.5 kg wt. Find the magnitude of the charges.

Solution. Now,

$$F = \frac{q_1 q_2}{4\pi\epsilon_0 r^2}$$

$$\therefore 1.5 = 9 \times 10^9 \times \frac{q_1^2}{(3 \times 10^{-2})^2} \quad \text{as} \quad \frac{1}{4\pi\epsilon_0} = 9 \times 10^9 \text{ Nm}^2 \text{ C}^2$$

$$\therefore q_1^2 = \frac{1.5 \times 9 \times 10^{-4}}{9 \times 10^9} = 1.5 \times 10^{-13}$$

$$\therefore q_1 = 3.87 \times 10^{-7} \text{ C}$$

Example 2.2: Two point charges of 1 C each are separated from each other by a distance of 1 m in a vacuum.

(a) What is the force of their interaction?

(b) What will be the force if the medium between the charges is water?

Solution: $F = \frac{q_1 q_2}{4\pi\epsilon_0 \epsilon_r r^2}$; For air $\epsilon_r = 1$ and for water $\epsilon_r = 80$.

$$(a) \therefore F_{\text{air}} = 9 \times 10^9 \times \frac{1 \times 1}{1} = 9 \times 10^9 \text{ N}$$

$$(b) F_{\text{water}} = 9 \times 10^9 \times \frac{1 \times 1}{80 \times 1} = 1.1 \times 10^8 \text{ N}$$

2.4 PRINCIPLE OF SUPERPOSITION

If a number of point charges q_1, q_2, q_3, \dots are present in a region, then the total force on any particular charge is the vector sum of forces it experiences due to all other charges. This is called the **principle of superposition**. For the sake of simplicity, let there be only three charges. The force on q_3 is

$$\begin{aligned} \mathbf{F} &= \mathbf{F}_{13} + \mathbf{F}_{23} \\ &= \frac{1}{4\pi\epsilon_0} \frac{q_1 q_3}{|r_{13}|^3} \mathbf{r}_{13} + \frac{1}{4\pi\epsilon_0} \frac{q_2 q_3}{|r_{23}|^3} \mathbf{r}_{23} \end{aligned} \quad (2.7)$$

Generalizing, one finds the force acting on a charge q_j due to a number of other charges present in the region is

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0} \sum_{i \neq j} \frac{q_i q_j}{|r_{ij}|^3} \mathbf{r}_{ij} \quad (2.8)$$

2.5 ELECTRIC FIELD

Any region where an electric charge experiences a force is called an **electric field**. The force is due to the presence of other charged body in that region. For example, a charge q is placed in a region where a charged body Q is present. According to the standard convention, we take always Q to be positive. The charge q experiences a force \mathbf{F} and we say that it is in an electric field produced by the charge Q . The force that the charge Q produces is proportional to q . Thus, the force on a charged particle placed in an electric field is proportional to the charge of the particle.

$$\mathbf{F} \propto q$$

or

$$\mathbf{F} = q\mathbf{E} \quad (2.9)$$

where E is the proportionality constant and is known as **electric field strength or intensity**. Note that electric field \mathbf{E} is a vector quantity since force \mathbf{F} is a vector quantity. The direction of \mathbf{E} is along the direction of force, that is along the line joining Q to q .

Electric Field Intensity

The intensity of electric field at a point in the electric field is equal to the force per unit test charge placed at that point. Thus,

$$\mathbf{E} = \frac{\mathbf{F}}{q} \quad (2.10)$$

If q is positive, the force \mathbf{F} acting on the charge has the same direction as that of the electric field \mathbf{E} . If q is negative, the force \mathbf{F} acting on the charge has the direction opposite to the electric field \mathbf{E} .

Ideally, q must be as small as possible in order to avoid possible disturbance of the original field \mathbf{E} . For this reason, the following definition of \mathbf{E} is more commonly used.

$$E = \lim_{q \rightarrow 0} \frac{F}{q} \quad (2.11)$$

In SI system of units, the unit of electric field is Newton/Coulomb or N/C.

In the expression (2.5), let $q_1 = Q$, $q_2 = q$ and the medium be air, $\epsilon_r = 1$. Then the force produced by the charge Q on the charge q placed at a distance r from Q is given by

$$\mathbf{F} = q \left[\frac{Q}{4\pi\epsilon_0 r^2} \hat{\mathbf{r}} \right] \quad (2.12)$$

Comparing the above expression (2.12) with (2.9), we may say that the electric field E at the point where q is placed is

$$\mathbf{E} = \left[\frac{Q}{4\pi\epsilon_0 r^2} \hat{\mathbf{r}} \right]$$

Therefore, the field produced by the charge Q is

$$\mathbf{E} = \left[\frac{Q}{4\pi\epsilon_0 r^2} \hat{\mathbf{r}} \right] \quad (2.13)$$

From the above equation, we see that the electric field decreases as one moves away from the charged body Q . It reduces to zero at infinity.

2.5.1 Electric Field Due to a Group of Point Charges

The net electric field at a point due to a group of point charges can be found by applying the superposition principle. Since the Coulomb force obeys the superposition principle, the electric field intensity (force per unit charge) obeys the superposition principle. The electric field at a point P due to n point charges $q_1, q_2, q_3, \dots, q_n$ is equal to the vector sum of electric fields due to $q_1, q_2, q_3, \dots, q_n$ at point P . Thus, the resultant electric field is given by

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 + \dots + \mathbf{E}_n \quad (2.14)$$

where \mathbf{E}_1 is the electric field intensity at P due to q_1 ,

\mathbf{E}_2 is the electric field intensity at P due to q_2 ,

\mathbf{E}_3 is the electric field intensity at P due to q_3 , and so on.

Example 2.3: A force of 0.015 N acts upon a charge of $2 \times 10^{-7} \text{ C}$ at a point in an electric field. What is the strength of electric field at that point?

Solution: Electric field E , is given by $E = F/q$

$$\therefore E = \frac{0.015 \text{ N}}{2 \times 10^{-7} \text{ C}} = 7.5 \times 10^4 \text{ N/C}$$

2.6 COMPUTATION OF ELECTRIC FIELD IN SOME SPECIFIC CASES

1. Field due to a linear charge

Let us consider an infinitely long charged wire of negligible thickness and having a constant linear charge density λ . Let a point P be at a distance y from the wire, as shown in Fig. 2.1.

It is required to find the electric field intensity at P . Let us assume that the wire is made up of a number of infinitely small elements of length, dx . Let one of such elements be at a distance x , as shown in Fig. 2.1. Let the small charge on element be dq .

$$dq = \lambda dx$$

The field due to this charge dq at point is

$$dE = \frac{1}{4\pi\epsilon_0} \cdot \frac{dq}{(NP)^2} = \frac{dq}{4\pi\epsilon_0 r^2}$$

The x and y components of dE are:

$$dE_x = -dE \sin \theta \quad \text{and} \quad dE_y = -dE \cos \theta$$

The x -components of field at P cancel out each others effect. Therefore, the net field will be due to y -components only and is directed along y -axis.

The resultant field is

$$\begin{aligned} E &= \int_{x=-\infty}^{x=+\infty} dE_y = \int_{x=-\infty}^{x=+\infty} dE \cos \theta = \int_0^{\infty} 2 dE \cos \theta = \int_0^{\infty} \frac{2 dq}{4\pi\epsilon_0 r^2} \cos \theta \\ &= \int_0^{\infty} \frac{2 \lambda dx}{4\pi\epsilon_0 r^2} \cos \theta = \frac{\lambda}{2\pi\epsilon_0} \int_0^{\infty} \frac{dx}{r^2} \cos \theta \end{aligned}$$

From Fig. 2.1, we have $\frac{x}{y} = \tan \theta$

$$\therefore \frac{dx}{y} = \sec^2 \theta d\theta \quad \text{or} \quad dx = y \sec^2 \theta d\theta. \quad \text{Also, } x^2 + y^2 = r^2.$$

$$\begin{aligned} \therefore E &= \frac{\lambda}{2\pi\epsilon_0} \int_0^{\theta=\pi/2} \frac{y \sec^2 \theta d\theta}{x^2 + y^2} \cos \theta = \frac{\lambda}{2\pi\epsilon_0} \int_0^{\theta=\pi/2} \frac{y \sec^2 \theta d\theta}{y^2 \tan^2 \theta + y^2} \cos \theta \\ &= \frac{\lambda}{2\pi\epsilon_0} \int_0^{\theta=\pi/2} \frac{y \sec^2 \theta d\theta}{y^2 \sec^2 \theta} \cos \theta = \frac{\lambda}{2\pi\epsilon_0} \int_0^{\theta=\pi/2} \frac{\cos \theta}{y} d\theta \\ &= \frac{\lambda}{2\pi\epsilon_0 y} [\sin \theta]_0^{\pi/2} \end{aligned}$$

$$\text{or} \quad E = \frac{\lambda}{2\pi\epsilon_0 y} \quad (2.15)$$

2. Field due to a uniformly charged ring at an axial point

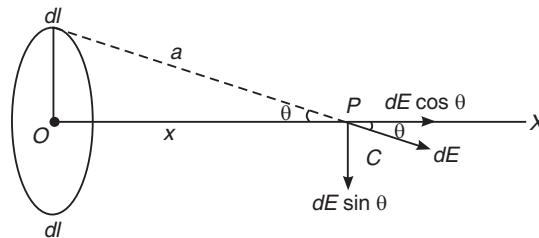


Fig. 2.2

Let us consider a uniformly charged ring of charge q and radius a is shown in Fig. 2.2. Let P be a point on the axis of the ring at a distance x from its centre. If a positive charge q is on the

ring, then the charge per unit length of the ring will be $\lambda = q / 2\pi a$. Now let us consider a small element of the ring of length dl . The charge on the element dl is $\frac{q dl}{2\pi a}$.

$$\text{Electric field due to this element is given by } dE = \frac{1}{4\pi\epsilon_0} \frac{\lambda dl}{r^2} = \frac{1}{4\pi\epsilon_0} \frac{q dl}{2\pi a r^2}$$

This field due to small element can be resolved into two components along x -axis (axis of the ring) and y -axis (perpendicular to the axis). Owing to symmetry, the perpendicular components cancel out. Hence, the resultant field will be due to the components along the axis of the ring. Therefore, the resultant field is given by

$$\begin{aligned} E &= \int dE = \int dE \cos \theta \\ \therefore E &= \int \frac{1}{4\pi\epsilon_0} \frac{q dl}{2\pi a} \frac{1}{r^2} \cos \theta \end{aligned}$$

$$\text{From Fig. 2.2, we have } a^2 + x^2 = r^2 \quad \text{and} \quad \cos \theta = \frac{x}{\sqrt{a^2 + x^2}}.$$

$$\begin{aligned} \therefore E &= \int \frac{1}{4\pi\epsilon_0} \frac{q dl}{2\pi a} \frac{1}{(a^2 + x^2)} \cdot \frac{x}{\sqrt{a^2 + x^2}} \\ &= \frac{1}{4\pi\epsilon_0} \frac{qx}{2\pi a} \frac{1}{(a^2 + x^2)^{3/2}} \int dl \\ &= \frac{1}{4\pi\epsilon_0} \frac{qx}{2\pi a} \frac{1}{(a^2 + x^2)^{3/2}} \times 2\pi a \end{aligned}$$

or

$$E = \frac{1}{4\pi\epsilon_0} \frac{qx}{(a^2 + x^2)^{3/2}} \quad (2.16)$$

3. Field due to a uniformly charged disc

A disc of radius ' a ' units is charged uniformly with a charge density $\sigma \text{ C/m}^2$. Let P be at a distance r from the centre of the disc as shown in Fig. 2.3. The disc may be regarded as formed by several annular rings of increasing radius. Let us consider a ring of radius x and the area of the hatched ring be ds .

Electric field due to the hatched ring is

$$dE = \frac{1}{4\pi\epsilon_0} \frac{\sigma ds}{r^2} \cos \theta$$

$$\text{The area of the annular ring } ds = \pi \left[(x + dx)^2 - x^2 \right] = \pi(2x dx + dx^2) \approx 2\pi x dx.$$

The term dx^2 is negligible compared to x .

$$\therefore dE = \frac{1}{4\pi\epsilon_0} \frac{\sigma 2\pi x dx}{r^2} \cos \theta = \frac{\sigma x dx}{2\epsilon_0 r^2} \cos \theta$$

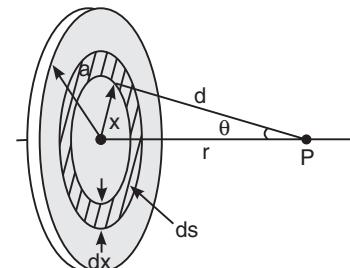


Fig. 2.3

From Fig. 2.3, it is seen that $\cos \theta = d/r \therefore r = d \sec \theta$

$$\tan \theta = x/d \therefore x = d \tan \theta \text{ and } dx = d \sec^2 \theta d\theta$$

Using these expressions

$$dE = \frac{\sigma(d \tan \theta) \cdot (d \sec^2 \theta d\theta)}{2\epsilon_0 d^2 \sec^2 \theta} \cos \theta = \frac{\sigma \sin \theta d\theta}{2\epsilon_0}$$

Electric field due to the entire disc is obtained by integrating the above expression between the limits of θ from 0 to α .

$$\therefore E = \frac{\sigma}{2\epsilon_0} \int_0^\alpha \sin \theta d\theta = \frac{\sigma}{2\epsilon_0} [-\cos \theta]_0^\alpha = \frac{\sigma}{2\epsilon_0} [1 - \cos \alpha]$$

From Fig. 2.3, we have $\cos \alpha = \frac{d}{\sqrt{a^2 + d^2}}$.

$$\therefore E = \frac{\sigma}{2\epsilon_0} \left[1 - \frac{d}{\sqrt{a^2 + d^2}} \right] \quad (2.17)$$

$$\text{If } d \gg a, \quad E = \frac{\sigma}{2\epsilon_0} \quad (2.18)$$

4. Field due to an electric dipole

A system of two equal and opposite charges separated by a small distance is called an **electric dipole**. Fig. 2.4 shows two charges q separated by a distance $2a$.

(i) Field at an axial point of the dipole

Let P be a point on the axis of the dipole at a distance x from the centre of the dipole. The field due to the negative charge $-q$ at P is given by

$$E_1 = \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{(x-a)^2}$$

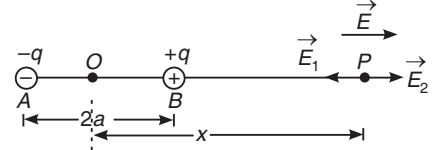


Fig. 2.4

The field due to the positive charge q at P is given by

$$E_2 = \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{(x+a)^2}$$

The resultant intensity $E = E_1 - E_2$

$$\begin{aligned} &= \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{(x-a)^2} - \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{(x+a)^2} \\ &= \frac{q}{4\pi\epsilon_0} \left[\frac{1}{(x-a)^2} - \frac{1}{(x+a)^2} \right] = \frac{q}{4\pi\epsilon_0} \left[\frac{(x+a)^2 - (x-a)^2}{(x-a)^2(x+a)^2} \right] \\ &= \frac{q}{4\pi\epsilon_0} \left[\frac{4ax}{(x-a)^2(x+a)^2} \right] \end{aligned}$$

or

$$E = \frac{qax}{\pi\epsilon_0(x^2 - a^2)^2} \quad (2.19)$$

If $x \gg a$,

$$E = \frac{qa}{\pi\epsilon_0 x^3} = \frac{2aq}{2\pi\epsilon_0 x^3}$$

or

$$E = \frac{\mu}{2\pi\epsilon_0 x^3} \quad (2.20)$$

In the above $\mu = 2a q$ is known as the electric **dipole moment**.

- The resultant field \mathbf{E} is along the axis of the dipole and is directed from $-q$ to q .
- The electric field due to a dipole is inversely proportional to cube of the distance.

(ii) Field at a point on the perpendicular bisector of the dipole

Let P be a point on the perpendicular bisector of the dipole as shown in Fig. 2.5.

Let E_1 and E_2 be the fields due to $-q$ and q respectively.

The fields due to the charges at P are given by

$$E_1 = E_2 = \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{a^2 + x^2}$$

The resultant intensity $E = E_1 + E_2 = 2E_1 \cos \theta$.

From the Fig. 2.5, we have $\cos \theta = \frac{a}{\sqrt{a^2 + x^2}}$

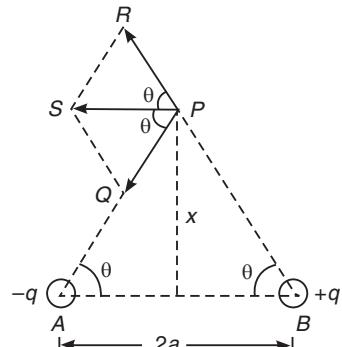


Fig. 2.5

$$E = 2 \cdot \frac{1}{4\pi\epsilon_0} \cdot \frac{q}{a^2 + x^2} \cdot \frac{a}{\sqrt{a^2 + x^2}} = \frac{2aq}{4\pi\epsilon_0(a^2 + x^2)^{3/2}}$$

or

$$E = \frac{\mu}{4\pi\epsilon_0(a^2 + x^2)^{3/2}} \quad (2.21)$$

If $x \gg a$,

$$E = \frac{\mu}{2\pi\epsilon_0 x^3} \quad (2.22)$$

- The resultant field \mathbf{E} is parallel the axis of the dipole and is directed from q to $-q$.
- The electric field due to a dipole is inversely proportional to cube of the distance.

Example 2.4: A very large sheet of charge has density of $5 \mu C/m^2$. Determine the electric field at a distance of 25 cm. Take medium as air.

Solution: Now, $\phi = \frac{q}{4\pi r^2}$

$$\therefore q = 4\pi\phi r^2 = 4\pi \times 5 \times 10^{-6} \times (0.25)^2 = 1.57 \times 10^{-5} \text{ C}$$

$$\text{But } E = \frac{q}{4\pi\epsilon_0 r^2} = \frac{1.57 \times 10^{-5}}{4\pi \times 8.854 \times 10^{-12} \times (0.25)^2} = 2.258 \times 10^6 \text{ N/C}$$

Example 2.5: A charged sphere of $80 \mu C$ is placed in air. Find the electric field intensity at a point 20 cm from the centre of the sphere.

Solution: Now,

$$E = \frac{q}{4\pi\epsilon_0 \epsilon_r r^2} = 9 \times 10^9 \times \frac{80 \times 10^{-6}}{1 \times (0.2)^2} = 1.8 \times 10^5 \text{ N/C}$$

Example 2.6: The charge per unit length on a long straight filament is $-70 \mu\text{C/m}$. Find the electric field at a distance 30 cm from the filament.

Solution: From Gauss' law, the electric field at a distance r from a long wire, is given by,

$$E = \frac{\lambda}{2\pi\epsilon_0 r} = 2 \times \frac{1}{4\pi\epsilon_0} \times \frac{\lambda}{r}$$

$$\therefore E = 2 \times (9 \times 10^9) \times \frac{-70 \times 10^{-6}}{0.3} = -4.2 \times 10^6 \text{ N/C.}$$

2.7 ELECTROSTATIC POTENTIAL

When an electric charge is moved towards a like charge or away from an unlike charge, work is done against the electric forces by the external agency that moves the charge. As a result, the electrical charge acquires **potential energy**. If the charge is released, work is done by the field and the charge accelerates. It means that its potential energy is converted into kinetic energy. In mechanics work done on a particle is expressed in terms of changes in potential energy. In case of electric field also, the work done on a charge can be expressed in terms of the potential energy of the charge.

Work done by the electric field on a charge

Let us consider a positive point test charge q placed at point A in a uniform electric field, E . The force on the charge due to electric field is qE . If we wish to move the charge with constant velocity *against* the electric field from point A to point B, we must exert a force of qE . The work done by the external force is positive and is equal to

$$W_{\text{ext}} = F_{\text{ext}}x \cos 0^\circ = +F_{\text{ext}}x = +qEx$$

At the same time, the work done by the electric field on the charge is negative since force and displacement are in opposite direction. The work done by the field is

$$W_E = F_{\text{ext}}x \cos 180^\circ = -F_{\text{ext}}x = -qEx$$

By moving the charge from point A to point B, the external force increases the **electric potential energy**, U by an amount equal to the work done on the charge.

The change in potential energy of the charge is $\Delta U = U_B - U_A$

$$\therefore \Delta U = U_B - U_A = qEx$$

where U_A and U_B are the potential energies of the charge at location A and location B respectively.

In terms of the work done *by the electric field*, the change in potential energy of the charge may be expressed as

$$\Delta U = U_B - U_A = -(-qEx) = -W_E \quad (2.23)$$

Thus, if a charge moves from one point to another in an electric field, the difference in the electric potential energy of the charge between the points is the negative of the work done by the electric field on that charge.

Line Integral

As the charge moves a distance dx along the path from A to B, the electric field does an element of work dW on it.

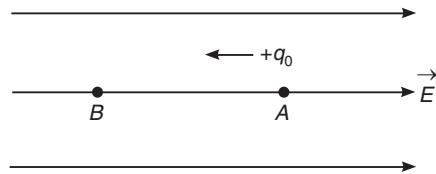


Fig. 2.6

$$dW = -\mathbf{F} \cdot d\mathbf{x} = -q\mathbf{E} \cdot d\mathbf{x}$$

∴ The total work done by the field is

$$W_{AB} = -q \int_A^B \mathbf{E} \cdot d\mathbf{x} \quad (2.24)$$

The integral in the above equation is called the **line integral**.

Electric potential difference

The electric field may be characterized by eliminating the dependence on the test charge q by defining **electric potential difference** between any two points in an electric field as the change in potential energy per unit positive test charge. We have

$$\begin{aligned} W_{AB} &= - \int_A^B \mathbf{F} \cdot d\mathbf{x} \\ \therefore W_{AB} &= - \frac{Qq}{4\pi\epsilon_0} \int_{x_1}^{x_2} \frac{1}{x^2} dx \\ &= - \frac{Qq}{4\pi\epsilon_0} \left[\frac{1}{x_1} - \frac{1}{x_2} \right] \\ &= (U_B - U_A) \end{aligned} \quad (2.25)$$

Therefore, the work done per unit charge is

$$\frac{W_{AB}}{q} = \frac{U_B}{q} - \frac{U_A}{q} = V_B - V_A \quad (2.26)$$

where $V_A = U_A/q$ is the potential energy per unit charge at location A and similarly $V_B = U_B/q$ is the potential energy per unit charge at location B. Then,

$$\Delta V = \frac{\Delta U}{q} = \frac{W}{q} \quad (2.27)$$

From now onwards, we denote $W_{AB} = W_E$ by W .

Electric potential

A charged particle placed in an electric field has potential energy because of its interaction with the field. The potential energy U of a charge at any point is equal to the negative of the work done on the charge by the electric field as the charge moves from infinity to that point in the field. Usually the point A is chosen at infinity and hence $V_A = 0$. Then we denote V_B by V . Using equ. (2.26), we define the electric potential at a point as the potential energy per unit charge placed at that point. Thus,

$$V = \frac{U}{q} \quad (2.28)$$

where U is the potential energy.

The electric potential is measured in unit of volts, which is equal to Joules/Coulomb or J/C.

$$1 \text{ V} = \frac{1 \text{ J}}{1 \text{ C}}$$

Electric potential is a scalar quantity and hence it is often referred to as **electrostatic scalar potential**.

We can write from equ. (2.25) that

$$V = \frac{Q}{4\pi\epsilon_0 r} \quad (2.29)$$

2.7.1 Calculating the Potential from the Field

Let a charge move from an initial point 1 to final point 2 in an electric field along the path shown. As the charge moves a distance along this path, the electric field does an element of work dW on it.

$$\begin{aligned} dW &= -\mathbf{F} \cdot d\mathbf{x} = -q\mathbf{E} \cdot d\mathbf{x} \\ \therefore \text{The total work done by the field } W_{12} &= -q \int_1^2 \mathbf{E} \cdot d\mathbf{x} \quad (2.30) \\ \therefore \text{The electric potential difference } V_2 - V_1 &= - \int_1^2 \mathbf{E} \cdot d\mathbf{x} \end{aligned}$$

If the initial point 1 is taken to be at infinity, $V_1 = 0$ and writing $V_2 = V$, the above equation yields

$$V = - \int_1^2 \mathbf{E} \cdot d\mathbf{x}$$

In a more general way, we write the above expression as $V = - \int_1^2 \mathbf{E} \cdot d\mathbf{r}$ (2.31)



Fig. 2.7

2.7.2 Calculating the Field from the Potential

Let us consider a set of closely spaced equipotential surfaces perpendicular to the plane of the page and passing through it. Let the potential difference between each pair of adjacent surfaces be dV . Suppose a charge moves through a small distance dx from one equipotential surface to the adjacent equipotential surface. The work that the electric field does on the charge is given by equ. (2.26) as

$$dW = q dV$$

We can also express the work done is

$$dW = -\mathbf{F} \cdot d\mathbf{x} = -q\mathbf{E} \cdot d\mathbf{x} = -qE \cos \theta (dx)$$

Equating these two equations, we obtain

$$q dV = -q E \cos \theta (dx)$$

$$\text{or } E \cos \theta = -\frac{dV}{dx} \quad (2.32)$$

As $E \cos \theta$ is the component of E in the x direction, we have to use partial derivative in the above equation. Thus, $E_x = -\frac{\partial V}{\partial x}$ (2.33)

That is, the electric field is equal to the negative of the derivative of the electric potential with respect to some coordinate.

In general, the electric potential is a function of all three spatial coordinates. If V is given in terms of rectangular coordinates, and then

$$E_x = -\frac{\partial V}{\partial x}$$

$$E_y = -\frac{\partial V}{\partial y} \quad (2.34)$$

$$E_z = -\frac{\partial V}{\partial z}$$

2.7.3 Potential Gradient

The rate of change of potential with distance is called the **potential gradient**. If the electric field is homogeneous and uniform, the potential gradient is given by

$$\text{Potential gradient} = \frac{dV}{dx} \quad (2.35)$$

where dV is the change in potential between two points separated by a distance x .

Example 2.7: Two positive charges of $12 \times 10^{-10} \text{ C}$ and $8 \times 10^{-10} \text{ C}$ are placed 10 cm apart. Find the work done in bringing the charges 4 cm closer.

Solution: The electrostatic force between the charges separated by a distance x , is given by

$$F = \frac{q_1 q_2}{4\pi\epsilon_0 x^2}$$

$$\therefore F = 9 \times 10^9 \times \frac{12 \times 10^{-10} \times 8 \times 10^{-10}}{x^2} = \frac{8.64 \times 10^{-9}}{x^2} \text{ N}$$

If the charge is moved through a small distance dx , the work done dW , in bringing the charges closer by distance dx , will be,

$$\begin{aligned} dW &= F \cdot dx \\ \therefore dW &= \frac{8.64 \times 10^{-9}}{x^2} \times dx \end{aligned}$$

Now total work done in moving the charges from 0.10 m to 0.06 m apart will be,

$$\begin{aligned} W &= - \int_{0.1}^{0.06} dW = -8.64 \times 10^{-9} \int_{0.1}^{0.06} \frac{dx}{x^2} \\ &= -8.64 \times 10^{-9} \left[\frac{1}{x} \right]_{0.1}^{0.06} \\ &= -8.64 \times 10^{-9} \left[\frac{1}{0.06} - \frac{1}{0.1} \right] \end{aligned}$$

$$\therefore W = 5.76 \times 10^{-8} \text{ J.}$$

2.8 EQUIPOTENTIAL SURFACES

An **equipotential surface** is a surface on which the potential has the same value at all points. In other words, the potential difference between any two points on an equipotential surface is zero.

Properties of equipotential surfaces

- (i) *Work done in moving a charged particle over an equipotential surface is zero:* Since the potential energy of a charged particle is the same at all points of a given equipotential surface, work done in moving a charged particle over an equipotential surface is zero.

- (ii) *Electric field is always perpendicular to an equipotential surface:* The equipotential surface through any point will be perpendicular to the direction of electric field at that point, as shown in Fig. 2.8.
- (iii) *The spacing between equipotential surfaces gives us indication of regions of strong and weak fields:* The region where the equipotential surfaces are crowded is the region of stronger field and the region where the surfaces are separated by larger distance is the region of weaker field.
- (iv) *Equipotential surfaces never intersect each other:* If two equipotential surfaces could intersect, then at the point of intersection there would be two values of electric potential which is impossible.

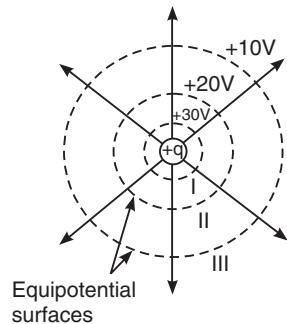


Fig. 2.8

2.9 ELECTRIC FIELD IS A CONSERVATIVE FIELD

A force is said to be **conservative**, if we can associate potential energy with it. If potential energy cannot be associated with a force, then the force is **nonconservative**. The gravitational force is conservative whereas frictional force is nonconservative.

We know that gravitational force does negative work on a body while the body is rising and an equal amount of positive work is done on its return trip. The total work done is zero for the round trip.

Definition: A field is conservative if the work done on a particle that moves through a round trip in the field is zero.

Or equivalently, a field is conservative if the work done by it on a particle between two points is the same for all paths connecting the two points.

Mathematically, a force field is said to be conservative if the line integral of the field along any closed path is zero. Therefore, an electric field is said to be conservative if

$$\oint \mathbf{E} \cdot d\mathbf{l} = 0 \quad (2.35)$$

Let a charge move from an initial point A to final point B in an electric field along the path shown in Fig. 2.9. As the charge moves a distance $d\mathbf{l}$ along this path, the electric field does an element of work dW on it.

$$dW = -\mathbf{F} \cdot d\mathbf{l} = -q\mathbf{E} \cdot d\mathbf{l}$$

∴ The total work done by the field in moving the charge from A to B is

$$W_{AB} = -q \int_A^B \mathbf{E} \cdot d\mathbf{l} \quad (2.36)$$

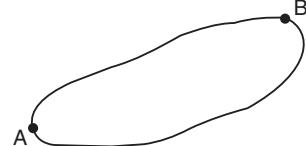


Fig. 2.9

The integral in the above equation is called the **line integral**.

If the curve C forms a closed path, then the integral along the closed path is denoted by

$$\oint_C \mathbf{E} \cdot d\mathbf{l} \quad (2.37)$$

When the path is a closed curve, the line integral is referred to as “net circulation integral” for \mathbf{E} around the chosen path. It is a measure of a vector field property called the curling up of field lines.

If the line integral around any closed path vanishes, that is $\oint_C \mathbf{E} \cdot d\mathbf{l} = 0$ then, the field is said to be **conservative**.

$$W_{AB} = -\frac{Qq}{4\pi\epsilon_0} \int_{x_A}^{x_B} \frac{1}{x^2} dx = -\frac{Qq}{4\pi\epsilon_0} \left[\frac{1}{x_A} - \frac{1}{x_B} \right]$$

It is seen from the above result that the work done depends only on the starting point x_A and the final point x_B and not on the path chosen to go from A to B. If now the charge returns from B to A through the same path or any other path the work done would be

$$W_{BA} = \frac{Qq}{4\pi\epsilon_0} \left[\frac{1}{x_A} - \frac{1}{x_B} \right]$$

Therefore, the total work done on a point charge in the electric field over any closed path is

$$W_{\text{total}} = \oint dW = W_{AB} + W_{BA} = 0$$

$$\therefore \oint_C \mathbf{E} \cdot d\mathbf{l} = 0$$

2.10 POTENTIAL AT A POINT DUE TO A GROUP OF POINT CHARGES

The potential due to a number of charges at a point can be found by taking the algebraic sum of potentials due to all charges. If point charges q_1, q_2, q_3, \dots are at distances r_1, r_2, r_3, \dots from a point P, then the resultant potential at P is given by

$$V = \frac{q_1}{4\pi\epsilon_0 r_1} + \frac{q_2}{4\pi\epsilon_0 r_2} + \frac{q_3}{4\pi\epsilon_0 r_3} + \dots$$

$$= \frac{1}{4\pi\epsilon_0} \left[\frac{q_1}{r_1} + \frac{q_2}{r_2} + \frac{q_3}{r_3} + \dots \right]$$

$$\text{or } V = \frac{1}{4\pi\epsilon_0} \sum \frac{q_n}{r_n} \quad (2.38)$$

2.11 COMPUTATION OF ELECTRIC POTENTIAL IN SOME SPECIFIC CASES

1. Potential due to a charged sphere

Let us consider a charged conducting sphere of radius R and carrying a charge of Q which is uniformly distributed throughout the sphere. The entire charge may be assumed to be concentrated at the centre of the sphere. Then, the charged sphere can be treated as identical to a point charge Q located at the centre of the sphere O.

Case 1 : P lies outside the sphere: Let P be at a distance r from the centre of the sphere. As the entire charge is concentrated at O, the potential at P will be

$$V = \frac{Q}{4\pi\epsilon_0 r}$$

Case 2 : P lies on the surface of the sphere: Let P be on the surface of the sphere. Therefore, $r = R$. Hence the potential at a point on the surface of sphere is

$$V = \frac{Q}{4\pi\epsilon_0 R}$$

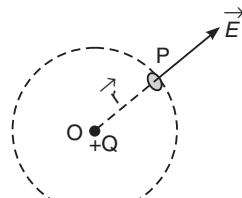


Fig. 2.10

Case 3 : P lies inside the sphere: The potential due to a charged sphere at any point inside it is the same as on the surface of the sphere.

2. Potential due to an electric dipole

An electric dipole consists of two equal and opposite charges very close to each other. Let AB be an electric dipole of length $d = 2a$. Let P be a point where we would like to determine the potential due to the dipole. Let $OP = r$ and let θ be the angle between r and the dipole axis, AB (see Fig. 2.11).

From the $\Delta^{le} OAC$, we get

$$\cos \theta = \frac{OC}{AO} = \frac{OC}{a}$$

\therefore

$$OC = a \cos \theta$$

Similarly,

$$OD = a \cos \theta$$

If $r \gg a$, we can write

$$PA = PC = PO + OC = r + a \cos \theta$$

and

$$PB = PD = PO - OD = r - a \cos \theta$$

The electric potential at the point P due to the electric dipole is

$$\begin{aligned} V &= \frac{1}{4\pi\epsilon_0} \left[\frac{q}{PB} - \frac{q}{PA} \right] \\ &= \frac{q}{4\pi\epsilon_0} \left[\frac{1}{r - a \cos \theta} - \frac{1}{r + a \cos \theta} \right] \\ &= \frac{q}{4\pi\epsilon_0} \left[\frac{2a \cos \theta}{r^2 - a^2 \cos^2 \theta} \right] \\ &= \frac{1}{4\pi\epsilon_0} \cdot \frac{2qa \cos \theta}{r^2 - a^2 \cos^2 \theta} \end{aligned}$$

But $2qa = qd = \mu$, the electric dipole moment of the dipole.

\therefore

$$V = \frac{1}{4\pi\epsilon_0} \cdot \frac{\mu \cos \theta}{r^2 - a^2 \cos^2 \theta} \quad (2.39)$$

At distances r far larger than the values of a , $a^2 \cos^2 \theta / r^2 \ll 1$ and the term can be neglected. The equation (2.39) then reduces to

$$V = \frac{1}{4\pi\epsilon_0} \cdot \frac{\mu \cos \theta}{r^2} \quad (2.40)$$

- (i) When the point lies on the axial line of the dipole on the side of the positive charge, then $\theta = 0$ and $\cos \theta = 1$. Therefore,

$$V = \frac{1}{4\pi\epsilon_0} \cdot \frac{\mu}{r^2}$$

- (ii) If the point lies on the axial line of the dipole on the side of the negative charge, then $\theta = 180^\circ$ and $\cos \theta = -1$. Therefore,

$$V = -\frac{1}{4\pi\epsilon_0} \cdot \frac{\mu}{r^2}$$

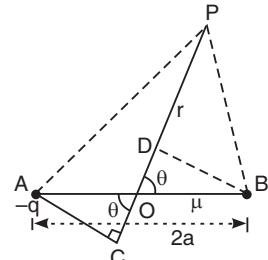


Fig. 2.11

- (iii) If the point lies on the equatorial line of the dipole, then $\theta = 90^\circ$ and $\cos \theta = 0$.
Therefore,

$$V = 0$$

Example 2.8: A hollow sphere of radius 20 cm is charged with a charge of 30×10^{-9} C. Find the potential at a distance of 50 cm from the sphere centre.

Solution: The potential at the surface of charged sphere of radius r is given by

$$V = \frac{q}{4\pi\epsilon_0 r}.$$

$$\text{So, the potential at a distance } x \text{ will be } V = \frac{q}{4\pi\epsilon_0 x} = 9 \times 10^9 \times \frac{30 \times 10^{-9}}{0.5} = 0.539 \text{ kV.}$$

Example 2.9: The potential due to an isolated point charge at a point 20 cm from the charge is 400 volts. Calculate magnitude of charge.

Solution: The potential at any point at a distance x from the point charge will be

$$V = \frac{q}{4\pi\epsilon_0 x}$$

$$\therefore 400 = 9 \times 10^9 \times \frac{q}{0.2} \quad \therefore q = 8.9 \times 10^{-9} \text{ C}$$

Example 2.10: Calculate electrostatic potential at a point due to charge of $50 \mu\text{C}$ at a distance of 15 cm from it.

Solution:

$$V = \frac{q}{4\pi\epsilon_0 r}$$

$$\therefore V = 9 \times 10^9 \times \frac{50 \times 10^{-6}}{0.15} = 3 \times 10^6 \text{ V}$$

2.12 FLUX

Electric field is a vector field. The first important property that characterizes a vector field is *flux*. Michael Faraday made use of field lines for visualizing electric and magnetic fields. Gauss introduced the concept of flux to express the relation between a field and its source.

Let us consider a uniform electric field, as shown in Fig. 2.12. Let the field lines (lines of force) penetrate a plane rectangular surface of area, A , which is perpendicular to the field. The number of field lines per unit area is proportional to the magnitude of the electric field. Therefore, the number of field lines penetrating the area A is proportional to the product EA . **The electric flux** is defined as the product of the magnitude of the electric field and surface area, A , perpendicular to the field.

$$\Phi = EA$$

When the surface is not perpendicular to the field lines, then the component of E along the normal to the surface is to be multiplied by the area. Thus,

$$\Phi = (E \cos \theta) A$$

We may express the above relation as the scalar product of vectors \mathbf{E} and \mathbf{A} , as

$$\Phi = \mathbf{E} \cdot \mathbf{A} \quad (2.41)$$

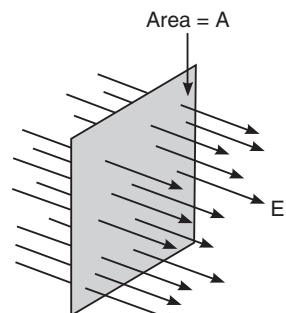


Fig. 2.12

In more general situations, the surface is of arbitrary shape. To know the flux passing through the surface, we have to divide the surface into a large number of small elements, each of area dA . The area dA of surface element is defined as a vector whose magnitude represents the area of the element and the direction is indicated by the outward normal to the elemental surface. The electric flux through this small element is

$$\Delta\Phi = \mathbf{E}_i \cdot \Delta\mathbf{A}_i$$

By adding the contributions of all elements, we obtain the total flux through the entire surface. Thus,

$$\Phi = \int_{\text{Surface}} \mathbf{E} \cdot d\mathbf{A}$$

Therefore, the flux of an electric field \mathbf{E} through an *open surface S* is given by

$$\Phi = \int_S \mathbf{E} \cdot d\mathbf{A} \quad (2.42)$$

If the surface is closed, we call it a *closed surface integral* and denote it by a circle on the integration symbol. Thus,

$$\Phi = \oint_S \mathbf{E} \cdot d\mathbf{A} \quad (2.43)$$

2.13 SOLID ANGLE

The concept of solid angle is an extension of the concept of a two-dimensional angle that is measured in radians or degrees. Let us consider the small area element dS at a distance r from the point O. Let \mathbf{n} be the unit vector perpendicular to the surface element dS . When every point of the boundary of dS is joined to O, a cone is formed.

The vector $\mathbf{OA} = \mathbf{r}$ makes an angle θ with the unit vector \mathbf{n} . The projection of the area dS on a plane normal to \mathbf{r} is given by

$$\mathbf{r} \cdot \mathbf{n} dS = \mathbf{r} \cdot dS \cos \theta$$

The solid angle $d\Omega$ subtended by the area element dS at a point O is defined as

$$d\Omega = \frac{\text{Projection of } dS \text{ perpendicular to } \mathbf{r}}{r^2} = \frac{dS \cos \theta}{r^2} = \frac{\mathbf{r} \cdot dS}{r^2} \quad (2.44)$$

Solid angle has no dimensions but is represented by the unit *steradian*.

The total solid angle subtended by the surface of a sphere at O is given by

$$\frac{\text{Total surface area}}{r^2} = \frac{4\pi r^2}{r^2} = 4\pi \quad (2.45)$$

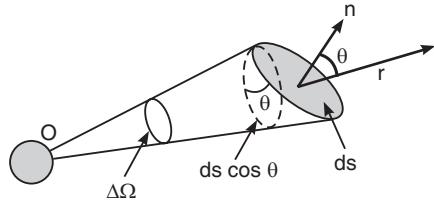


Fig. 2.13

2.14 GAUSS' LAW OF ELECTROSTATICS IN FREE SPACE

Integral laws are called *global laws* since they indicate what happens over a wide range. Gauss law is one of such powerful global laws. It gives the relationship between the integral component of the electric field over a closed surface and the total charge enclosed by the surface. This law relates the electric flux through any closed surface to the net amount of charge within the surface.

Statement

Gauss law states that the total electric flux through a closed surface enclosing a charge is equal to $\frac{1}{\epsilon_0}$ times the magnitude of the charge enclosed.

Proof

Let us consider a closed surface S surrounding a charge q , as shown in Fig. 2.14. Let dS be an element of area around the point P on the surface and \hat{n} an outward unit vector normal to it. Let θ be the angle between the electric field at P and the unit vector \hat{n} . The electric flux through the element of area dS is

$$\begin{aligned} d\phi &= \mathbf{E} \cdot d\mathbf{S} = E \cos \theta \, dS \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \cos \theta \, dS \end{aligned}$$

where r is the distance of the surface element dS from charge q .

Now $\frac{dS \cos \theta}{r^2} = d\Omega$ is the solid angle subtended by dS at q . Therefore,

$$d\phi = \frac{1}{4\pi\epsilon_0} q \, d\Omega$$

\therefore Total flux through the entire surface S is

$$\phi = \oint d\phi = \oint \mathbf{E} \cdot d\mathbf{S} = \frac{1}{4\pi\epsilon_0} q \oint d\Omega$$

The total solid angle subtended by S at O is 4π .

$$\therefore \phi = \oint \mathbf{E} \cdot d\mathbf{S} = \frac{q}{\epsilon_0} \quad (2.46)$$

Gauss' law becomes very useful in calculation of electric field in cases where Coulomb's law or principle of superposition becomes tedious.

2.15 DIVERGENCE OF ELECTRIC FIELD

The **divergence** of vector field \mathbf{E} is defined as the limiting value of the ratio of the closed surface integral and the volume enclosed by the surface over which integration is carried out, when the volume element tends to zero.

$$\text{div } \mathbf{E} = \lim_{\Delta V \rightarrow 0} \frac{\oint \mathbf{E} \cdot d\mathbf{S}}{\Delta V} \quad (2.47)$$

It is common practice to denote $\text{div } \mathbf{E}$ as $\nabla \cdot \mathbf{E}$. The divergence in rectangular coordinates is found to be given by

$$\text{div } \mathbf{E} = \nabla \cdot \mathbf{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \quad (2.48)$$

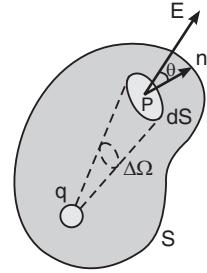


Fig. 2.13

$\nabla \cdot \mathbf{E}$ can be taken as a measure of the spreading out of the field \mathbf{E} . If a vector function \mathbf{E} spreads out from a point, then it has a *positive* divergence at that point and the point acts as a **source** of the field \mathbf{E} . On the other hand, if the field converges to a point, then $\nabla \cdot \mathbf{E}$ would be negative at that point and the point acts as a **sink** for the field \mathbf{E} . If the vector field neither converges nor diverges, then $\nabla \cdot \mathbf{E} = 0$. For a field at a point to have finite divergence means that an equivalent to a source must exist there. If the divergence of a field is zero for a small volume, it means that all the flux that enters the volume also leaves it; this region of space is then free of a source. The divergence of a vector field can be considered to be a measure of scalar sources of the field. Further, at the points of the field where the divergence of \mathbf{E} is positive, we have the sources of the field (positive charges), while at the points where it is negative, we have sinks (negative charges). The field lines emerge from the field sources and terminate at the sinks.

2.16 DIFFERENTIAL FORM OF GAUSS'S LAW

The integral form of Gauss's law relates the net flux out of a finite volume to the net amount of charge enclosed in that volume. In contrast to (2.46), the differential form of the Gauss theorem establishes the relation between the volume charge density and the changes in the field intensity \mathbf{E} in the vicinity of a given point in space. Thus, the differential law is a *local law*, which tells us what happens at a given point. The differential form of Gauss's law can be found by applying Gauss's integral law to an infinitesimally small volume surrounding a point. The integrals will then transform to differentials in the limit as the volume goes to zero, and we will obtain a point relationship of Gauss's law involving derivatives only. The differential form of Gauss's law is more general and will be very useful since derivatives are easier to calculate compared to integrals.

Let us represent the charge q in the volume V enclosed by a closed surface S as $q_{\text{int}} = \langle \rho \rangle V$, where $\langle \rho \rangle$ is the volume charge density, averaged over the volume V . Using this into equ.(2.46) and dividing the equation with V , we obtain

$$\frac{1}{V} \oint_S \mathbf{E} \cdot d\mathbf{s} = \frac{\langle \rho \rangle}{\epsilon_0} \quad (2.49)$$

We now make the volume V to tend to zero by contracting it to the point we are interested in. Then, $\langle \rho \rangle$ will tend to the value of ρ at the given point of the field and hence the ratio on the R.H.S. of equ. (2.49) will tend to ρ/ϵ_0 . When V tends to zero, the quantity on the L.H.S. is called the **divergence of the field \mathbf{E}** . By definition

$$\nabla \cdot \mathbf{E} = \lim_{V \rightarrow 0} \frac{1}{V} \oint_S \mathbf{E} \cdot d\mathbf{s}$$

Consequently, the above relation transforms into

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \quad (2.50)$$

Equ. (2.50) is the Gauss theorem expressed in the differential form. It shows that the divergence of the field \mathbf{E} at a given point depends only on the electric charge density ρ at this point.

2.17 DERIVATION OF COULOMB'S LAW FROM GAUSS LAW

Coulomb's law can be deduced from Gauss law. Let us consider an isolated point charge q , as shown in Fig. 2.15. Let us consider any imaginary spherical surface r centered on the

charge. Such an imaginary closed surface enclosing a charge is called a *Gaussian surface*. The advantage of the spherical surface is that \mathbf{E} is normal to it at all points and has the same magnitude.

In Fig. 2.15 both \mathbf{E} and $d\mathbf{S}$ are directed radially outward at any point on the surface. The angle between them is zero.

$$\therefore \mathbf{E} \cdot d\mathbf{S} = \mathbf{E} d\mathbf{S}$$

Gauss law therefore reduces to

$$\epsilon_0 \oint \mathbf{E} \cdot d\mathbf{S} = \epsilon_0 \oint \mathbf{E} d\mathbf{S} = q$$

Since \mathbf{E} is constant for all points on the sphere, it can be taken out of the integral.

$$\therefore \epsilon_0 \mathbf{E} \oint d\mathbf{S} = q$$

$$\oint d\mathbf{S} = \text{Area of the sphere, } 4\pi r^2$$

$$\therefore \epsilon_0 \mathbf{E} (4\pi r^2) = q$$

Or
$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2}$$

The above equation gives the magnitude of the electric field strength \mathbf{E} at any point at a distance r from the isolated charge q . If a second charge q_1 is kept at any point on the spherical surface, it experiences a force

$$\mathbf{F} = q_1 \mathbf{E}$$

Using the expression for \mathbf{E} into the above equation, we obtain

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0} \frac{qq_1}{r^2} \quad (2.51)$$

which is the Coulomb's law.

2.18 APPLICATIONS OF GAUSS'S LAW

1. Electric field due to a solid charged sphere

Let us consider an isolated sphere of charge Q having radius R . Let us consider a point P at a distance r from the centre O of the sphere.

Case 1: *Point P lies outside the charged sphere:* Let us draw the Gaussian surface through point P so that it encloses the sphere of charge. In this the Gaussian surface is a spherical surface of radius r and is centered on O , as shown in Fig. 2.16.

Let E be electric field at point P due to the sphere of charge Q . The field is spherically symmetrical. Therefore, E is along the normal to the spherical surface and has the same magnitude at all points on the surface of the sphere.

Total flux through the Gaussian surface is

$$\oint \mathbf{E} \cdot d\mathbf{S} = \oint \mathbf{E} d\mathbf{S} = E \oint d\mathbf{S} = E(4\pi r^2)$$

According to Gauss's theorem,

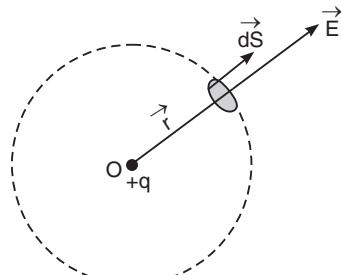


Fig. 2.15

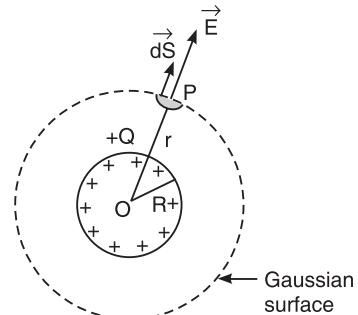


Fig. 2.16

$$\mathbf{E}(4\pi r^2) = \frac{Q}{\epsilon_0}$$

or

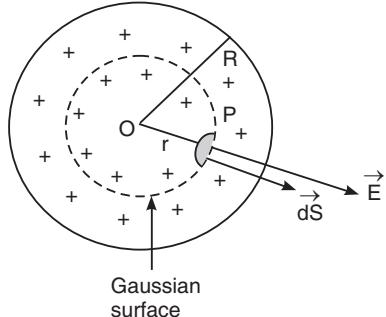
$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^2} \quad (2.52)$$

Case 2: Point P lies on the charged sphere: In this case, the Gaussian surface is a spherical surface of radius R and is centered on O . The area of the sphere is $4\pi R^2$. According to Gauss's law

$$\mathbf{E}(4\pi R^2) = \frac{Q}{\epsilon_0}$$

or

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} \frac{Q}{R^2} \quad (2.53)$$



Case 3: Point P lies inside the charged sphere:

In this case, the Gaussian surface is a spherical surface of radius $OP = r$ and is centered on O , as shown in Fig. 2.17. If Q' is the charge enclosed by the Gaussian surface, then according to Gauss's theorem

$$\mathbf{E}(4\pi r^2) = \frac{Q'}{\epsilon_0} \quad (r < R)$$

or

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} \frac{Q'}{r^2}$$

$$\text{Now, } Q' = \frac{Q}{\frac{4}{3}\pi R^3} \times \frac{4}{3}\pi r^3 = Q \frac{r^3}{R^3}$$

$$\therefore \mathbf{E} = \frac{1}{4\pi\epsilon_0} \frac{1}{r^2} \times Q \frac{r^3}{R^3}$$

or

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} \frac{Qr}{R^3} \quad (2.54)$$

Fig. 2.18 shows the variation of electric field intensity with distance r from the centre of a sphere of charge. At the centre of the sphere, $E = 0$ and within the sphere, $E \propto r$. The value is maximum at the surface of the sphere. Outside the sphere, $E \propto 1/r^2$ and the curve follows a parabolic path.

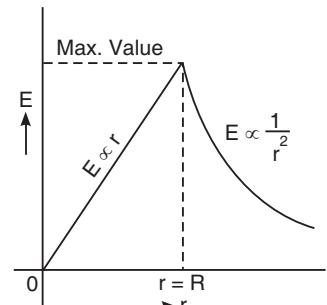


Fig. 2.18

2. Electric field due to a uniformly charged spherical shell

Let us consider a thin spherical shell of radius R . Let O be the centre of the shell and q be the charge on the shell. The electric field due to the charged spherical shell is radial and spherically symmetric. Let us consider a point P at a distance r from the centre O of the shell.

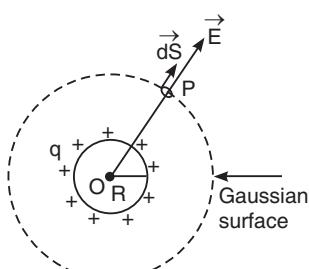


Fig. 2.19

Case 1: Point P lies outside the spherical shell: Let us draw the Gaussian surface through point P. The Gaussian surface is a spherical surface of radius r and is centered on O, as shown in Fig. 2.19.

Let E be electric field at point P due to the sphere of charge q . The field is spherically symmetrical. Therefore, E is along the normal to the spherical surface and has the same magnitude at all points on the surface of the sphere.

Total flux through the Gaussian surface is

$$\oint \mathbf{E} \cdot d\mathbf{S} = \oint \mathbf{E} dS = E \oint dS = E(4\pi r^2)$$

According to Gauss's theorem,

$$E(4\pi r^2) = \frac{q}{\epsilon_0} \quad (r > R)$$

or $E = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2}$ (2.55)

Case 2: Point P lies on the surface of spherical shell: In this case, the Gaussian surface through P just encloses the spherical shell of radius R and is centered on O. The area of the sphere is $4\pi R^2$. According to Gauss's law

$$E(4\pi R^2) = \frac{q}{\epsilon_0}$$

or $E = \frac{1}{4\pi\epsilon_0} \frac{q}{R^2}$ (2.56)

Case 3: Point P lies inside the charged sphere: In this case, the Gaussian surface through the point does not enclose any charge (Fig. 2.20).

$$E(4\pi r^2) = \frac{0}{\epsilon_0} = 0$$

∴ $E = 0$ (2.57)

Fig. 2.21 shows the variation of electric field intensity with distance from the centre O for a uniformly charged spherical shell. For r varying from 0 to R , $E = 0$. In other words, $E = 0$ inside the shell. The magnitude of E is maximum at the surface of the shell, i.e., at $r = R$. However, outside the shell $E \propto 1/r^2$ and the curve follows a parabolic path.

3. Electric field due to a line charge

Let us consider an infinitely long wire or a thin rod having a uniform linear positive charge density λ (Fig. 2.22). Let P be a point at a distance r from the wire.

Let us draw a cylindrical Gaussian surface of radius r and length l around the wire. As the electric field \mathbf{E} is parallel to the two ends of the cylinder, their contribution to the electric flux is zero. The entire contribution to the electric flux is from the surface, S , of the cylinder.

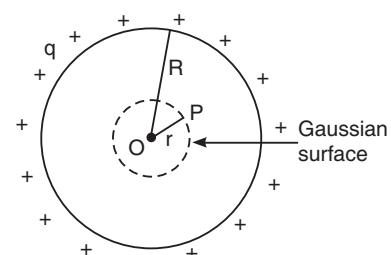


Fig. 2.20

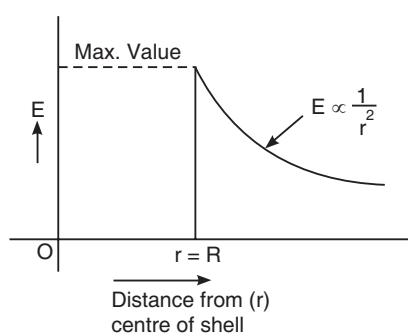


Fig. 2.21

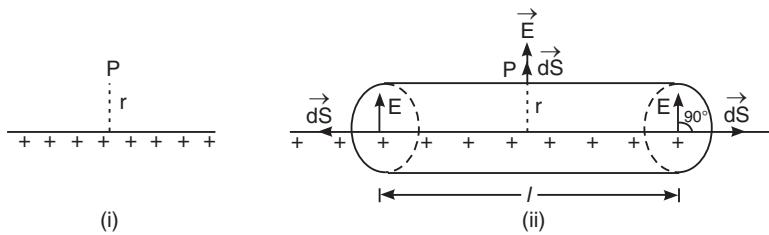


Fig. 2.22

Total flux through the Gaussian surface is

$$\oint \mathbf{E} \cdot d\mathbf{S} = \oint \mathbf{E} dS = E \oint dS = E(2\pi rl)$$

The charge enclosed inside the Gaussian surface is $= \lambda \cdot l$

According to Gauss's law

$$E \oint dS = \frac{q}{\epsilon_0}$$

$$\therefore E(2\pi rl) = \frac{\lambda l}{\epsilon_0}$$

$$\text{or } E = \frac{\lambda}{2\pi\epsilon_0 r} \quad (2.58)$$

Thus, the intensity of electric field of a line charge is inversely proportional to the distance r from the wire.

4. Electric field due to a thin sheet of charge

Let us consider a thin plane sheet of infinite extension. Let it be positively charged and have a uniform surface charge density σ on both sides of the sheet. The electric field acts perpendicular to the surfaces of the plane sheet and is directed outward, as shown in Fig. 2.23.

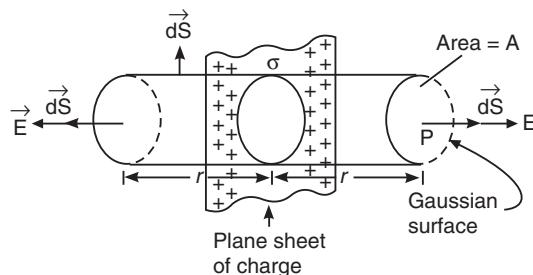


Fig. 2.23

In order to find the electric field strength due to the sheet of charge at any point P at a distance r from it, we draw a cylindrical Gaussian surface, as shown in Fig. 2.22. Since field lines are parallel to the curved surface of the cylinder, the contribution to the net electric flux from the curved surface is zero. The contribution to the flux comes from the two circular faces of the cylinder. The electric flux crossing through the Gaussian surface is

$$\oint \mathbf{E} \cdot d\mathbf{S} = \oint_{S_1} \mathbf{E} \cdot d\mathbf{S} + \oint_{S_2} \mathbf{E} \cdot d\mathbf{S} = 2E \oint dS = 2ES$$

The charge enclosed inside the Gaussian surface is $= \sigma S$

According to Gauss's law

$$\mathbf{E} \oint d\mathbf{S} = \frac{q}{\epsilon_0}$$

$$\therefore 2 \mathbf{E} S = \frac{\sigma S}{\epsilon_0}$$

or

$$\mathbf{E} = \frac{\sigma}{2\epsilon_0} \quad (2.59)$$

The above result indicates that the magnitude of the electric field due to a thin charged sheet is independent of the distance from the sheet.

Example 2.11: A large planar sheet of charge has a charge per unit area of $7.5 \mu \text{ C/m}^2$. Find electric field intensity just above the surface of sheet, measured from its midpoint.

Solution: From Gauss' law, the electric field intensity due to a non conducting planar sheet with uniform charge per unit area σ , is given by,

$$\mathbf{E} = \frac{\sigma}{2\epsilon_0} \quad \therefore \quad \mathbf{E} = \frac{7.5 \times 10^{-6}}{2 \times 8.85 \times 10^{-12}} = 4.24 \times 10^5 \text{ N/C}$$

2.19 GAUSS' LAW OF ELECTROSTATICS IN A (DIELECTRIC) MEDIUM

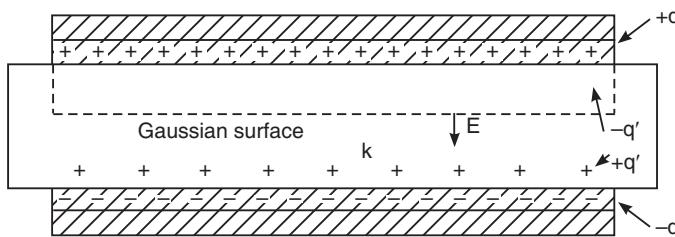


Fig. 2.24

Dielectrics (i.e., insulators) do not contain *free* electrons that can move over considerable distances. Therefore, they do not conduct electric current. However, when an external electric field is applied to the dielectric (say, held between capacitor plates, as in Fig. 2.24), the positive nuclei in each molecule are displaced along the field direction and the *bound* electrons are displaced in the opposite direction. The bound electrons are not mobile but are elastically bound to the molecule. Therefore, they can move only with in the electrically neutral molecules through a very small distance. This displacement of charges is known as **polarization** of the dielectric. As a result of polarization, uncompensated charges appear on the dielectric surface as well as in its bulk. Such charges are called **polarization charges** or **bound charges**. On the other hand, the charges that are on the plates of the capacitor are called **extraneous charges** or **free charges**. They do not belong to dielectric molecules. The electric field \mathbf{E} acting in a dielectric is the superposition of the field \mathbf{E}_0 of the extraneous charges and the field \mathbf{E}' of bound charges. Thus,

$$\mathbf{E} = \mathbf{E}_0 + \mathbf{E}'$$

Since the sources of an electric field \mathbf{E} are all electric charges—free and bound, we can write the Gauss theorem (2.46) for the field \mathbf{E} as

$$\oint \epsilon_0 \mathbf{E} ds = (q + q')_{\text{int}} \quad (2.60)$$

where q and q' are free and bound charges enclosed by the surface S . We now express q' in terms of \mathbf{P} , the **polarization vector**, as

$$\oint \mathbf{P} ds = -q'_{\text{int}} \quad (2.61)$$

Using (2.61) into (2.60), we get

$$\oint (\epsilon_0 \mathbf{E} + \mathbf{P}) ds = q_{\text{int}} \quad (2.62)$$

Let us denote the quantity in the parenthesis in the integrand of the above equation by \mathbf{D} . We thus define an auxiliary vector \mathbf{D} as

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \quad (2.63)$$

whose flux through an arbitrary enclosed surface is equal to the algebraic sum of extraneous charges enclosed by this surface.

$$\oint \mathbf{D} ds = -q_{\text{int}} \quad (2.64)$$

This is the *Gauss theorem for field \mathbf{D}* . The quantity \mathbf{D} is called **dielectric displacement**.

2.20 ELECTRIC DISPLACEMENT VECTOR

Electric field intensity, \mathbf{E} , measures the influence of a charge distribution at any point in space. This influence can also be measured in terms of the quantity **electric displacement** or **electric flux density** \mathbf{D} . The electric field \mathbf{E} depends in general upon the *permittivity* ϵ of the medium in which the charge is placed. For example, the electric field from a point charge Q is given by

$$\mathbf{E} = \frac{Q}{4\pi\epsilon r}$$

Electric flux density, \mathbf{D} , is independent of the medium and is a quantity determined by the relation

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} \quad \text{C/m}^2$$

\mathbf{D} is also called the **displacement vector**.

The vector \mathbf{D} is the sum of two completely different quantities $\epsilon_0 \mathbf{E}$ and \mathbf{P} . For this reason, \mathbf{D} is regarded an *auxiliary vector* which does not have any deep physical meaning. In isotropic dielectrics vector \mathbf{D} is collinear to \mathbf{E} and in case of anisotropic dielectrics these vectors are generally non-collinear. Its magnitude is given by the product of the electric intensity and the dielectric constant of the medium. Its magnitude is also equal to the surface density of free charges: $D = \sigma$.

The field \mathbf{D} can be depicted with the aid of electric displacement lines. Their direction and density are determined in exactly the same way as for the lines of the vector \mathbf{E} . The lines of \mathbf{E} begin and terminate on *extraneous* charges (see Fig. 2.25). The sources of the field \mathbf{D} are *all* charges. Hence, displacement lines can begin or terminate on both *extraneous* and *bound* charges. The lines of \mathbf{D} pass without interruption through the regions of the field containing bound charges.

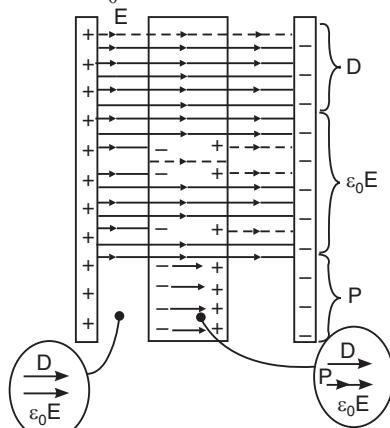


Fig. 2.25

QUESTIONS

1. State Coulomb's law in electrostatics. What do you mean by the flux of an electric field?
2. Define and explain electric intensity. What are its units? Deduce an expression for electric intensity due to a point charge.
3. State and prove Gauss's law in electrostatics.
4. State Gauss's theorem. Making use of it determine the intensity near a long charged conductor.
5. Deduce the Gauss's law in differential form $\nabla \cdot \mathbf{E} = \rho / \epsilon_0$.
6. Derive Coulomb's law from Gauss's law.
7. Determine the intensity of electric field due to a dipole at an equatorial and axial point.
8. Determine the intensity of electric field near a charged sphere and a plane sheet of charge.
9. Define and explain electric potential. What are its units?
10. Obtain an expression for the potential due to a uniformly charged sphere at an external and internal point.
11. Derive an expression for electric field intensity due to a uniformly charged spherical shell when point lies (i) outside the shell and (ii) inside the shell.
12. Does electric potential obey superposition principle? Explain.
13. What do you mean by line integral of electric field?
14. Show that the line integral of electric field over a closed path is zero.
15. Electrostatic force is a conservative force. Discuss.
16. What is an equipotential surface? Mention important properties of equipotential surfaces.
17. Apply Gauss's theorem to a dielectric medium and obtain the relationship between \mathbf{E} , \mathbf{D} , and \mathbf{P} vectors.
18. Define electric displacement and explain the significance of electric displacement vector.

PROBLEMS

1. Two spheres charged with equal but opposite charges experience a force of 2.5×10^5 N when they are placed at 2 cm apart in a medium of relative permittivity 5. Determine the charge on each sphere.
[Ans: $q = 745 \times 10^{-6}$ C]
2. A charge of 5×10^{-5} C is distributed between two small spheres which are placed with their centers 3 m apart. It is found that the spheres repel each other with a force of 0.6 N. Find the charges on two spheres.
[Ans: $q_1 = 3 \times 10^{-5}$ C, $q_2 = 2 \times 10^{-5}$ C]
3. A charge of $0.52 \mu\text{C}$ is placed in an electric field of 4.5×10^5 N/C. What is the magnitude of the force acting on the charge?
[Ans: $F = 0.23$ N]
4. Calculate the electric field intensity due to an electric dipole of length 10 cm and consisting of two charges $\pm 2 \mu\text{C}$ at a distance of 50 cm from each charge.
[Ans: $E = 1.44 \times 10^4$ N/C]
5. Two positive point charges of 16×10^{-10} C and 12×10^{-10} C are placed 10 cm apart. Find the work done in bringing the two charges 4 cm closer.
[Ans: $W = 11.52 \times 10^{-8}$ J]
6. What is the strength of electric field produced by the nucleus of a hydrogen atom at the site of electron in 1s orbit? The radius of 1s orbit is 0.53 Å.
[Ans: $E = 5.12 \times 10^{11}$ N/C]
7. Two charges $+1\mu\text{C}$ are placed at the corners of the base of an equilateral triangle. The length of the side of triangle is 0.5 m. Find electric field intensity at apex of the triangle.
[Ans: $E = 36$ kN/C]
8. An inflated balloon in the shape of a sphere of radius 14 cm has a total charge of $8 \mu\text{C}$ uniformly distributed on its surface. Calculate the electric field intensity at 50 cm and at 13 cm from the centre of the balloon.
[Ans: $E_{50} = 2.88 \times 10^5$ N/C, $E_{13} = 0$]

9. At what distance from a point charge of $6 \mu\text{C}$ would the potential equal to $2.7 \times 10^4 \text{ V}$?
[Ans: $\mathbf{r} = 2 \text{ m}$]
10. Find the work done in bringing a charge of $10 \times 10^{-4} \mu\text{C}$ from infinity to a point 25 cm from charge of $3 \times 10^{-2} \mu\text{C}$.
[Ans: $\mathbf{W} = 10.79 \times 10^{-7} \text{ J}$]
11. Determine the electric field intensity and potential in air at a distance of 3 cm from a charge $6.5 \times 10^{-2} \text{ C}$.
[Ans: $\mathbf{E} = 5 \times 10^6 \text{ N/C}$, $\mathbf{V} = 1.5 \times 10^4 \text{ V}$]
12. Two parallel plates are separated by a dielectric of 4 cm thickness and relative permittivity 2.5. If the potential difference between the plates is 3000 volts. Calculate
 (a) Potential gradient
 (b) Electric flux density between
[Ans: $\mathbf{V} = 7.5 \times 10^5 \text{ V}$, $\mathbf{D} = 1.66 \times 10^{-5} \text{ C/m}^2$]
13. A charge of $3 \times 10^{-9} \text{ C}$ moves through a distance of 0.5 m in a uniform field of 200 N/C. determine,
 (a) The electric force on the charge
 (b) Work done on the charge
 (c) Potential difference between initial and final positions
[Ans: $\mathbf{F} = 600 \times 10^{-9} \text{ N}$, $\mathbf{W} = 300 \times 10^{-9} \text{ J}$, $\mathbf{V} = 100 \text{ V}$]
14. Calculate the electric potential at the surface of the nucleus of the gold atom. Given – atomic number of gold = 79, charge on proton = $1.6 \times 10^{-19} \text{ C}$, radius of nucleus = $6.6 \times 10^{-15} \text{ \AA}$.
[Ans: $\mathbf{V} = 1.72 \times 10^5 \text{ V}$]

CHAPTER

3

Magnetostatics and Electrodynamics

Static magnetic fields are produced by permanent magnets and steady currents flowing in conductors. Magnetostatics deals with magnetic fields produced by steady currents. Faraday discovered the method of generation of electric and magnetic fields which vary with time. Electrodynamics deals with the study of varying electric and magnetic fields. Maxwell unified the important laws of electricity and magnetism and formulated a unified theory in 1861. He formulated four equations that are regarded as the basis of all electric and magnetic phenomena. The consequences of Maxwell's equations are very far reaching. Maxwell predicted the existence of electromagnetic waves and that light is a form of electromagnetic radiation.

3.1 MAGNETIC FIELD

In 1820 Oersted discovered that electric currents create magnetic fields. A steady current I flowing in a straight conductor produces a magnetic field around it. The *magnetic lines of force* exist in the form of a series of concentric circles with conductor as the centre. The direction of the field can be found by the right hand rule. If the current carrying conductor is gripped with the right hand so that the thumb points the direction of current flow, then the curled fingers point in the direction of the magnetic field. Fig. 3.1 shows the method of finding the direction of magnetic field.

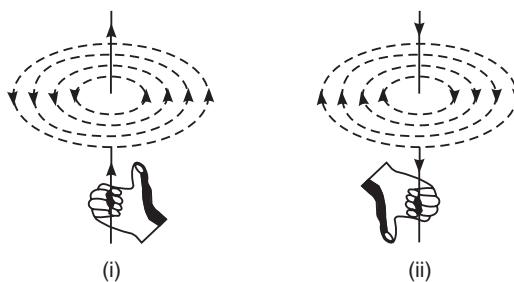


Fig. 3.1

The region around a current carrying conductor or a permanent magnet where magnetic effects are experienced is called a *magnetic field*. A magnetic field is schematically represented by magnetic lines of force, which are also known as *field lines* or lines of *magnetic induction*. A magnetic field is described either by magnetic field strength \mathbf{H} or by the *magnetic induction* (or *magnetic flux density*), \mathbf{B} .

Relation between \mathbf{B} & \mathbf{H} :

3.2 MAGNETIC FLUX DENSITY

Magnetic flux

The lines of induction are collectively called **flux**. The magnetic flux through a region is the number of lines of induction passing normally through the region. The concentration of lines of induction is an indication of magnetic field strength. It is defined in terms of the flux density.

Magnetic induction

The number of field lines passing through a unit area of cross-section is called the **magnetic flux density** or **magnetic induction**. It is denoted by magnetic induction vector, \mathbf{B} . Thus,

$$\mathbf{B} = \frac{\text{Magnetic flux}}{\text{area}} = \frac{\phi}{A} \quad (3.1)$$

Therefore, magnetic flux is given by $\phi = BA$. In a more general way, let the area be inclined at an angle to the magnetic field. Let θ be the angle between the normal to the area and the direction of magnetic field. Then,

$$\phi = B A \cos \theta = \mathbf{B} \cdot \mathbf{A} \quad (3.2)$$

Thus, magnetic flux through an area is equal to the dot product of magnetic field \mathbf{B} and area \mathbf{A} .

The unit of magnetic flux is weber (Wb) and the unit of magnetic induction is weber per square metre (Wb/m^2) or tesla (T).

The magnetic flux through any surface may also be given by the surface integral of the normal component of \mathbf{B} . Thus,

$$\phi = \int_S \mathbf{B} \cdot d\mathbf{s} \quad (3.3)$$

where $d\mathbf{s}$ is the elemental surface.

3.3 BIOT-SAVART LAW

Let a conductor of an arbitrary shape carry a steady current I . Let P be a point in the magnetic field produced by the current. Let a small element AB of length dl produce magnetic field dB at P . Let r be the distance of P from the current element $I dl$ and θ be the angle between dl and r . According to Biot-Savart law, the magnitude of magnetic field dB is directly proportional to the product $I dl \sin \theta$ and is inversely proportional to the square of distance between current element and the point P . Thus,

$$dB \propto I dl \sin \theta \quad \text{and} \quad dB \propto \frac{1}{r^2}$$

Combining these relations, $dB \propto \frac{I dl \sin \theta}{r^2}$

$$\text{or} \quad dB = k \frac{I dl \sin \theta}{r^2}$$

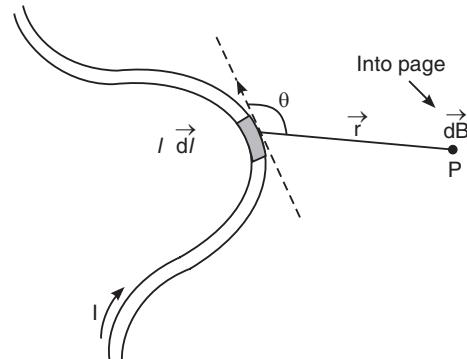


Fig. 3.2

where k is a constant proportionality. The value of k depends on the medium in which the conductor is situated and the system of units adopted. In SI units, its value for free space is

$$k = \frac{\mu_0}{4\pi} \quad \text{where } \mu_0 = 4\pi \times 10^{-7} \text{ Tm / A}$$

$$\therefore \mathbf{dB} = \frac{\mu_0}{4\pi} \frac{I dl \sin \theta}{r^2} \quad \text{Biot-Savart law} \quad (3.4)$$

The Biot-Savart law holds only for steady currents. The current element $I dl$ is the source of static magnetic field, just as a charge q is the source of static electric field. The above law is written in the vector form as

$$\mathbf{dB} = \frac{\mu_0}{4\pi} \frac{(dl \times r)}{r^3} \quad (3.5)$$

The direction of the magnetic field is given by the right hand thumb rule (see Fig. 3.1). The direction of $d\mathbf{B}$ is into the plane of the paper.

The total magnetic field at P due to the conductor is obtained by summing up the contributions of all current elements.

$$\therefore \mathbf{B} = \int d\mathbf{B} = \int \frac{\mu_0}{4\pi} \frac{I (dl \times r)}{r^3} \quad (3.6)$$

3.4 AMPERE'S LAW

Ampere's law states that *the line integral of the tangential component of the magnetic field over any closed path is equal to the amount of the current enclosed by the loop*. Thus,

$$\oint \mathbf{B} \cdot dl = \mu_0 I \quad (3.7)$$

Both Ampere's law and the Biot-Savart law are relations between a current distribution and the magnetic field that it generates. We can apply Biot-Savart law to calculate the magnetic field caused by any current distribution. On the other hand, Ampere's law allows us to calculate magnetic field with ease in case of symmetry.

Let us consider an infinitely long wire along the z-axis carrying a current I amp. The magnetic flux density due to this wire is directed everywhere circular to the wire and its magnitude is dependent only on the distance from the wire. Let us consider a circular path C of radius r in the plane normal to the wire and centered at the wire. The current enclosed by an arbitrarily closed path C is given by the surface integral of the current density over any surface S bounded by the closed path C.

The total current flowing through the surface area S is given by

$$I = \int \mathbf{J} \cdot ds \quad (3.8)$$

Therefore, we can write

$$\oint \mathbf{H} \cdot dl = \int \mathbf{J} \cdot ds$$

$$\therefore \oint_C \mathbf{B} \cdot dl = \mu_0 \int_S \mathbf{J} \cdot ds \quad (3.9)$$

This is known as *Ampere's circuital law*.

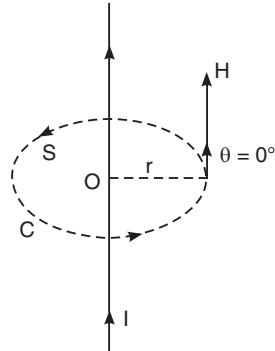


Fig. 3.3

3.4.1 Ampere's Circuital Law in Differential Form

If we now shrink the path C to a very small size ΔC so that the surface area bounded by it becomes very small, ΔS , we can write equ. (3.6) as

$$\oint_{\Delta C} \mathbf{B} \cdot d\mathbf{l} = \mu_0 \int_{\Delta S} \mathbf{J} \cdot ds \quad (3.10)$$

Since the surface area ΔS is very small we can consider the current density to be uniform over the surface so that

$$\int_{\Delta S} \mathbf{J} \cdot ds \approx \mathbf{J} \cdot \Delta S$$

The relation becomes exact in the limit $\Delta S \rightarrow 0$. Dividing both the sides of equ. (3.10) by ΔS and letting $\Delta S \rightarrow 0$, we have

$$\begin{aligned} \lim_{\Delta S \rightarrow 0} \frac{\oint_{\Delta C} \mathbf{B} \cdot d\mathbf{l}}{\Delta S} &= \lim_{\Delta S \rightarrow 0} \frac{\mu_0 \int_{\Delta S} \mathbf{J} \cdot ds}{\Delta S} \\ &= \mu_0 \lim_{\Delta S \rightarrow 0} \frac{\mathbf{J} \cdot \Delta S}{\Delta S} \\ &= \mu_0 \mathbf{J} \cdot \mathbf{n} \end{aligned} \quad (3.11)$$

Now, the curl of \mathbf{B} is defined as the vector having the magnitude given by the maximum value of the quantity on the left side of equ. (3.11). We note that this maximum value occurs for an orientation of ΔS for which the direction of its normal coincides with the direction of \mathbf{J} and it is equal to μ_0 times the magnitude of \mathbf{J} . Thus

$$|\nabla \times \mathbf{B}| = \text{maximum value of } \left(\lim_{\Delta S \rightarrow 0} \frac{\oint_{\Delta C} \mathbf{B} \cdot d\mathbf{l}}{\Delta S} \right) = \mu_0 |\mathbf{J}| \quad (3.12)$$

or

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} \quad (3.13)$$

Equ. (3.13) is Ampere's circuital law in *differential form*.

3.5 GAUSS'S LAW FOR MAGNETISM

Just as in the case of electrostatics, the magnetic flux through an element of area ds is given by the dot product of \mathbf{B} with ds . For an arbitrary surface ds bounded by a closed contour S , total magnetic flux passing through the surface is given by

$$\phi = \oint_S \mathbf{B} \cdot ds \quad (3.14)$$

The lines of vector \mathbf{B} have neither beginning nor ending. The number of lines emerging from any volume bounded by a closed surface S is always equal to the number of lines entering the volume. Hence, the flux of \mathbf{B} through any closed surface is equal to zero. Thus,

$$\oint_S \mathbf{B} \cdot ds = 0$$

Dividing both sides of the above equation by an incremental volume Δv over which the surface is to be considered closed, we get

$$\frac{\oint_S \mathbf{B} \cdot ds}{\Delta v} = 0$$

The limit of the left side of the equation, as $\Delta v \rightarrow 0$, is the divergence of the vector \mathbf{B} .

$$\therefore \nabla \cdot \mathbf{B} = 0 \quad (3.15)$$

3.6 MAGNETIC SCALAR POTENTIAL

In electrostatics, the potential V is a **scalar**. It is related to the magnetic field \mathbf{E} as

$$\mathbf{E} = -\nabla V$$

The scalar potential is related to the sources, i.e. charge distribution and can be easily calculated. In a similar way, we may define magnetic scalar potential by the following relation.

$$\mathbf{B} = -\mu_0 \nabla \phi$$

where ϕ is the **magnetic scalar potential** due to the sources.

According to Ampere's law

$$\oint \mathbf{B} \cdot d\mathbf{l} = \mu_0 i \quad (3.16)$$

In general the right hand side of the above equation is not zero and \mathbf{B} is a **solenoidal** (not a conservative) **field**. In regions outside the sources of \mathbf{B} (see Fig. 3.4), a closed path is not linked with a current and therefore, the line integral of \mathbf{B} over such a closed path vanishes. In such cases, we may consider \mathbf{B} to be a *conservative field* and therefore, express it as

$$\mathbf{B} = -\mu_0 \nabla \phi \quad (3.17)$$

This does not mean that \mathbf{B} changes its character from a solenoidal field to a conservative field. All that we say is that in the regions outside the sources, we can simplify the mathematical formulation by expressing \mathbf{B} in terms of the gradient of ϕ . Let us obtain now an expression for ϕ .

The magnetic scalar potential at P due to a current loop is the sum of the potentials due to the individual small loops. A current loop is equivalent to a magnetic dipole. The magnetic moment of the small current loop is given by

$$dm = i dA$$

where dA is the area of the small loop. The potential due to the dipole is given by an expression similar to equ. (1.40) derived in case of an electric dipole. Thus,

$$\text{The scalar potential } d\phi = \frac{1}{4\pi} \frac{i dA \cos \theta}{r^2}$$

where r is the distance of the elemental loop dA from P and θ is the angle between r and the vector $d\mathbf{A}$. But $dA \cos \theta = r^2 d\Omega$, where $d\Omega$ is the solid angle subtended at P by the boundary of the given current loop.

$$\therefore \phi = \int d\phi = \frac{i}{4\pi} \int d\Omega = \frac{i\Omega}{4\pi} \quad (3.18)$$

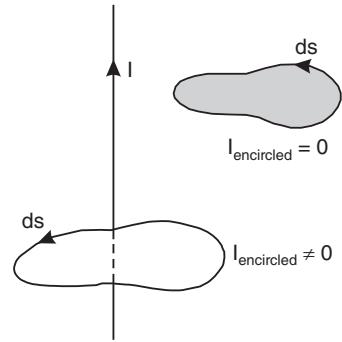


Fig. 3.4

3.7 MAGNETIC VECTOR POTENTIAL

Unlike the electric field, the magnetic field is a **solenoidal** field. In spite of that we can define the magnetic induction \mathbf{B} in terms of some potential function. Two space derivatives are possible with a vector field. They are divergence and curl. We know from equ. (3.15) that

$$\nabla \cdot \mathbf{B} = 0$$

Since the divergence of a curl is always zero, the second possibility is of expressing \mathbf{B} as the curl of some vector potential function. Thus, we write

$$\mathbf{B} = \nabla \times \mathbf{A} \quad (3.19)$$

A is called the **magnetic vector potential**. The potentials are a convenient way to relate a field to its scalar and vector sources. The vector sources for magnetic induction **B** are obtained by taking the curl of **B**. Thus,

$$\nabla \times \mathbf{B} = \nabla \times \nabla \times \mathbf{A} = \mu_0 \mathbf{J} \quad (3.20)$$

where the relationship between **B** and the current density **J** is given by Ampere's law. Unlike the case of the scalar potential, it is not apparent how the use of **A** will simplify the calculation of **B**.

According to vector identity

$$\begin{aligned} \nabla \times \nabla \times \mathbf{A} &= \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} \\ \therefore \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} &= \mu_0 \mathbf{J} \end{aligned} \quad (3.21)$$

The specification of the curl of a vector does not uniquely specify the vector. It requires that the curl of vector has the same value regardless of the value of its divergence. The equations of the magnetostatic field, $\nabla \cdot \mathbf{B} = 0$ and $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$ do not require that $\nabla \cdot \mathbf{A}$ be specified in any particular way and we can choose it at our convenience. Therefore, we choose that

$$\nabla \cdot \mathbf{A} = 0 \quad (3.22)$$

This is known as a **gauge condition**, and the above equ. (3.22) is known as the **Coulomb gauge**. Other choices of gauge (i.e. other choices of $\nabla \cdot A$) are useful in other circumstances.

Setting $\nabla \cdot \mathbf{A} = 0$ in equ. (3.21), we get

$$\nabla^2 \mathbf{A} = -\mu_0 \mathbf{J} \quad (3.23)$$

Thus, the **magnetic vector potential A** is defined here through the equations

$$\nabla \times \mathbf{A} = \mathbf{B}$$

and

$$\nabla \cdot \mathbf{A} = 0$$

We now obtain an expression for **A**. Let $d\mathbf{A}$ be the magnetic potential due to a current element $I d\mathbf{l}$. By analogy with electrostatic potential, the magnetic potential may be written as

$$d\mathbf{A} \propto \frac{I d\mathbf{l}}{r}$$

$$d\mathbf{A} = k \frac{I d\mathbf{l}}{r} = \frac{\mu_0}{4\pi} \frac{I d\mathbf{l}}{r} \quad (3.24)$$

$$\therefore \mathbf{A} = \frac{\mu_0 I}{4\pi} \oint \frac{d\mathbf{l}}{r} \quad (3.25)$$

For a current flowing in a circuit *C*, we have

$$\mathbf{A} = \frac{\mu_0 I}{4\pi} \oint_C \frac{d\mathbf{l}}{r} \quad (3.26)$$

ELECTRODYNAMICS

3.8 FARADAY'S LAWS OF INDUCTION

In 1831, Michael Faraday discovered that current is induced in a conducting loop whenever a magnet is moved toward or away from the loop. The current is induced, even when the magnet is held stationary and the loop is moved either toward or away from the magnet. However, when both the loop and the magnet are stationary, current is not induced in the loop. These observations imply that **current is induced in the loop as long as relative motion occurs between the magnet and the loop**. This phenomenon is known as **electromagnetic**

induction. The current produced in the loop is called an **induced current** and it is produced by an **induced emf**.

Faraday summed up the above into two laws known as **Faraday's laws of electromagnetic induction.**

First Law: Whenever the magnetic flux linked with a circuit changes, an emf is always induced in it.

Second Law: The magnitude of the induced emf is equal to the time rate of change of the flux linkage.

Thus,

$$E = -\frac{d\phi}{dt} \quad (3.27)$$

where ϕ is the magnetic flux through the surface bounded by the loop.

If a coil consisting of N identical and concentric loops is used and if the field lines pass through all loops, the induced emf is

$$E = -N \frac{d\phi}{dt} \quad (3.28)$$

Faraday laws state that an emf is induced if the magnetic flux changes for any reason. It implies that **an electric current is induced by a time-varying magnetic field.**

3.9 LENZ'S LAW

An induced emf drives current around a circuit just as the emf of a battery does. The conventional current produced by a battery flows in the circuits from positive terminal to the negative terminal. The same is true of the direction of conventional current flow in case of induced emf. However, the identification of positive and negative terminals in this case is not obvious. The polarity of the induced emf and direction of induced current can be determined with the help of Lenz's law. We have to note that the net magnetic field penetrating a coil results from two contributions. (i) The original magnetic field that produces the changing flux and hence produces induced emf and (ii) the **induced magnetic field** created by the induced current.

In 1834, Lenz gave the rule for determining the direction of induced current in a closed conducting loop. It is known as Lenz's law. The Lenz's law states that

An induced current in a closed conducting loop will appear in such a direction that it opposes the original flux change.

The Lenz's law is reflected mathematically in the minus sign on the R.H.S. of Faraday's law (3.27). Note that Lenz's law applies to induced currents and not to induced emfs.

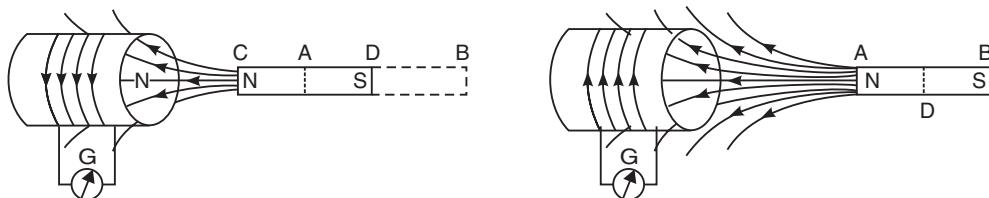


Fig. 3.5

Let us consider a coil being approached by a magnet. As the N-pole of the magnet moves towards the coil, the magnetic flux linking the coil increases. As a result, current is induced in the coil. According to Lenz's law, the direction of the induced current will be such that it opposes the increasing magnetic flux linking the coil. Therefore, the induced current will set

up magnetic flux that opposes the increase in flux through the coil. This is possible only when the right hand face of the coil becomes N-pole. Once we know the magnetic polarity of the coil face, the direction of the induced current can be determined with the help of right-hand rule.

When the magnet moves away from the coil, the flux through the coil decreases. It results in an induced current flowing in a direction such that the coil's magnetic field opposes the motion. It means that the right hand face of the coil becomes S-pole. Thus, the induced current in the loop is always in such a direction to oppose the change that produces it.

Lenz's law is a consequence of the law of conservation of energy. In the cases discussed above, the motion of the magnet is opposed. The mechanical energy spent in overcoming the opposition is converted into electrical energy which appears in the coil as current. Thus, Lenz's law follows directly from the law of conservation of energy.

3.10 INTEGRAL FORM OF FARADAY'S LAW

Let us consider a closed conducting circuit (see Fig. 3.6) through which a magnetic flux of uniform flux density \mathbf{B} exists. The conducting circuit is the contour of a surface, an element ds of which a direction specified by the unit normal, \mathbf{n} . If the magnetic flux linking the circuit is decreasing, an emf E will be induced having the direction as shown by the arrow in the Fig. 3.6. However, an emf implies the existence of an electric field. There is thus an induced electric field set up in the circuit, which is given by

$$\mathbf{E} = \oint \mathbf{E}_i \cdot d\mathbf{l} \quad (3.29)$$

the integration being performed around the path in the direction of \mathbf{E} .

Combining equations (3.27) and (3.29), we get

$$\mathbf{E} = \oint \mathbf{E}_i \cdot d\mathbf{l} = -\frac{d\phi}{dt}$$

where

$$\phi = \iint_S \mathbf{B} \cdot d\mathbf{s}$$

Therefore

$$\frac{d\phi}{dt} = \frac{d}{dt} \iint_S \mathbf{B} \cdot d\mathbf{s}$$

$$\therefore \oint \mathbf{E}_i \cdot d\mathbf{l} = -\frac{d}{dt} \iint_S \mathbf{B} \cdot d\mathbf{s} \quad (3.30)$$

The induced electric field exists in space regardless of whether a conducting wire is present or not. When a conducting wire is present an induced current will flow. Equ.(3.30) indicates that the electric field set up by a changing magnetic field is not a conservative field.

3.10.1 Faraday's law in differential form

According to Stoke's theorem, the surface integral of the curl of a vector field \mathbf{F} taken over any surface S is equal to the line integral of \mathbf{F} around the periphery C of the surface. Hence,

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = \iint_S (\nabla \times \mathbf{E}) \cdot d\mathbf{s}$$

It follows from (3.30) that

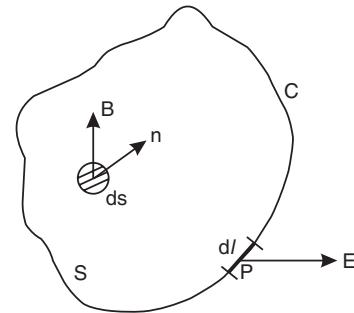


Fig. 3.6

$$\oint_S (\nabla \times \mathbf{E}) \cdot d\mathbf{s} = -\frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{s}$$

If the surface S is stationary, i.e. independent of time, then

$$\begin{aligned}\frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{s} &= \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s} \\ \therefore \int_S (\nabla \times \mathbf{E}) \cdot d\mathbf{s} &= \int_S -\frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s}\end{aligned}$$

Comparing the integrals on the two sides, we have

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad (3.31)$$

3.11 EQUATION OF CONTINUITY

Electric charge can neither be created nor destroyed. Therefore, the net charge in an isolated system remains constant. This is known as the **principle of conservation of charge**. This principle implies that *the time rate of increase (decrease) of charge within a closed volume equals the net rate of flow of charge into (out of) the volume*. This statement of conservation of charge is expressed by the equation of continuity.

Let us consider a surface S enclosing a volume. Let $d\mathbf{S}$ be a small element of this surface. Further, let \mathbf{J} be the current density at a point on the surface element. Then the current leaving the volume V bounded by the surface dS is given by

$$I = \oint_S \mathbf{J} \cdot d\mathbf{S}$$

Suppose the current is not a steady current. Then, J is a function of t as well as x, y, z and $\oint_S \mathbf{J} \cdot d\mathbf{S}$ represents the instantaneous rate at which charge is leaving the enclosed volume.

Using divergence theorem, we can write the above equation as

$$I = \oint_S \mathbf{J} \cdot d\mathbf{S} = \int_V \nabla \cdot \mathbf{J} dV \quad (3.32)$$

As some charge is leaving the volume, correspondingly the same amount of charge diminishes with in that volume. We express this fact as

$$\text{But} \quad I = -\frac{dq}{dt} = -\frac{d}{dt} \int_V \rho dV$$

where ρ is the charge density. Since volume V is a fixed volume, the time derivative operates only on the function ρ , when it is moved inside the integral. Further, the time derivative is a partial derivative since ρ is not only a function of time but also a function of position.

$$\therefore I = -\int_V \frac{\partial \rho}{\partial t} dV \quad (3.33)$$

Equating (3.32) and (3.33), we obtain

$$\int_V \nabla \cdot \mathbf{J} dV = -\int_V \frac{\partial \rho}{\partial t} dV$$

$$\text{or} \quad \int_V \left[\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} \right] dV = 0$$

Since the volume is completely arbitrary, the integrand in the above equation must be zero. Thus,

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0 \quad (3.34)$$

Equ. (3.34) is known as the *equation of continuity*. It follows from equ. (3.34) that

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t} \quad (3.35)$$

It expresses the principle of conservation of charge. Charge cannot flow away from a given volume without diminishing the amount of charge existing within the volume.

3.12 DISPLACEMENT CURRENT

Ampere's law implies that a magnetic field can be produced only by a flow of charges. Ampere's law was established as a result of large number of careful experiments done on steady situations. Maxwell showed that we run into difficulty when apply the Ampere's law for time-varying situations such as charge building up on the plates of a capacitor. He showed that we have to include another current, called *displacement current*, which also can produce a time-varying magnetic field. The need for displacement current can be well understood when current flow through a capacitor is considered.

Let us consider the circuit shown in Fig. 3.7, which consists of a parallel plate capacitor being charged (or discharged) through a certain external resistance R . If we apply Ampere's law to the contour C and the surface S_1 , we find that

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_{S_1} \mathbf{J} \cdot d\mathbf{s} = I \quad (3.36)$$

since the current is passing right through surface S_1 .

If on the other hand, Ampere's law is applied to the contour C and the surface S_2 , then J is zero at all points on S_2 . As current is not flowing through the surface S_2 ,

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \int_{S_2} \mathbf{J} \cdot d\mathbf{s} = 0 \quad (3.37)$$

Note that the two surfaces S_1 and S_2 are bounded by the same path length l and hence the contour integrals must be equal. It is easy to see that the equ. (3.36) and (3.37) contradict each other. Further, equ. (3.37) cannot be wrong. Therefore, it appears that Ampere's equ. (3.36) requires modification.

It is seen that the difficulty arises when we choose a contour between the plates of a capacitor. We know that negative charge flows from the battery up to the plate P_1 of the capacitor. Similarly, negative charge flows from plate P_2 to the battery. But there is no charge flow between the plates of the charging capacitor. However, there is a continuous current I flowing in the circuit. Although there is no current crossing the surface S_2 , there is certainly a

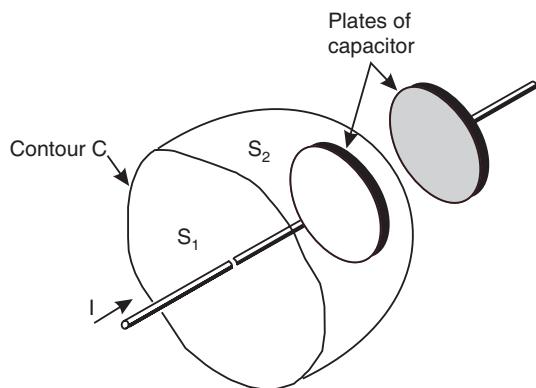


Fig. 3.7

changing electric field, because the capacitor is charging up as the current I flows in. Maxwell argued that this changing electric field constitutes an effective current.

First, we note that surface S_2 “cuts” only the electric field. In accordance with the Gauss's theorem, the flux of vector \mathbf{D} through a closed surface is $\oint \mathbf{D} \cdot d\mathbf{s} = q$. Therefore, the current I is

$$I = \frac{dq}{dt} = \frac{\partial}{\partial t} \oint \mathbf{D} \cdot d\mathbf{s} = \oint \frac{\partial \mathbf{D}}{\partial t} \cdot d\mathbf{s} \quad (3.38)$$

On the other hand, according to the continuity equation (3.35), we have

$$\oint \mathbf{J} \cdot d\mathbf{s} = -\frac{dq}{dt} \quad (3.39)$$

where \mathbf{J} is the conduction current density.

Summing up the left and the right hand sides of Eqns. (3.38) and (3.39) separately, we obtain

$$\oint \left(\mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \right) \cdot d\mathbf{s} = 0 \quad (3.40)$$

This equation is similar to the continuity equation for direct current. There is one more term $\frac{\partial \mathbf{D}}{\partial t}$ whose dimensions are the same as for current density. Maxwell termed this term as the density of displacement current. Thus,

$$\mathbf{J}_d = \frac{\partial \mathbf{D}}{\partial t} \quad (3.41)$$

The sum of the conduction and the displacement currents is called the **total current**. Its density is given by

$$\mathbf{J}_T = \mathbf{J} + \mathbf{J}_d = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad (3.42)$$

Thus, the theorem on circulation of vector \mathbf{H} , which was established for direct currents, can be generalized for an arbitrary case in the following form:

$$\oint \mathbf{H} \cdot d\mathbf{l} = \int \left(\mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \right) \cdot d\mathbf{s} \quad (3.43)$$

In this form, the theorem on circulation of vector \mathbf{H} is always valid, which is confirmed by the agreement of this equation with the results of experiments in all cases without any exception. The equation is known as Ampere-Maxwell equation. Thus, Maxwell showed that a changing electric field causes a current, which generates a magnetic field in just the same way as an actual current. Maxwell visualized that space itself is a medium, the *ether*, which had dielectric properties. If a dielectric is placed in an increasing electric field, the charges will be displaced by a continuously increasing distance. As long as the field increases in strength, the charges go on moving, thus giving rise to a displacement current. This is why Maxwell called the changing electric field term as displacement current.

It is possible to convert the line integral in equ.(3.43) into a surface integral using Stoke's theorem. Thus, we get

$$\int_S (\nabla \times \mathbf{H}) \cdot d\mathbf{s} = \int_S \left[\mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \right] \cdot d\mathbf{s}$$

Since the above result is true for any S , the integrands can be equated and we get

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad (3.44)$$

This is the general form of Ampere's law.

Displacement current is equivalent to conduction current only from the point of view of its ability of creating a magnetic field. Displacement currents exist only when an electric field varies with time. In dielectrics, displacement current consists of two essentially different components. Since vector $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$, it follows that the density $d\mathbf{D}/dt$ of displacement current is the sum of the densities of the "true" displacement current $\epsilon_0 d\mathbf{E}/dt$, and of the polarization current $d\mathbf{P}/dt$. The latter quantity appears due to the motion of bound charges. There is nothing unexpected in the fact that polarization currents excite a magnetic field, since these currents do not differ in nature from conduction currents. The new concept here is that the other component of the displacement current, $\epsilon_0 d\mathbf{E}/dt$, which is *not* connected with any motion of charges and is only due to the variation of the electric fields, also excites a magnetic field. Note that even in a vacuum, any temporal change of an electric field excites in the surrounding space a magnetic field.

3.13 MAXWELL'S EQUATIONS

The field equations which govern the time-varying electric and magnetic fields are now written as:

$$(i) \text{ Gauss's law} \quad \nabla \cdot \mathbf{D} = \rho \quad (3.45 \text{ a})$$

$$(ii) \text{ Gauss's law for magnetism} \quad \nabla \cdot \mathbf{B} = 0 \quad (3.45 \text{ b})$$

$$(iii) \text{ Faraday's law} \quad \nabla \times \mathbf{E} = - \frac{\partial \mathbf{B}}{\partial t} \quad (3.45 \text{ c})$$

$$(iv) \text{ Ampere's law} \quad \nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad (3.45 \text{ d})$$

Equations (3.45 a) to (3.45 d) are known as the **Maxwell's equations**. The first two are the divergence equations and the last two are the curl equations. The above equations can also be written in the following form:

$$(i) \text{ Gauss's law} \quad \nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon} \quad (3.46 \text{ a})$$

$$(ii) \text{ Gauss's law for magnetism} \quad \nabla \cdot \mathbf{H} = 0 \quad (3.46 \text{ b})$$

$$(iii) \text{ Faraday's law} \quad \nabla \times \mathbf{E} = - \mu \frac{\partial \mathbf{H}}{\partial t} \quad (3.46 \text{ c})$$

$$(iv) \text{ Ampere's law} \quad \nabla \times \mathbf{H} = \mathbf{J} + \epsilon \frac{\partial \mathbf{E}}{\partial t} \quad (3.46 \text{ d})$$

Maxwell's equations contain only the first derivatives of fields \mathbf{E} and \mathbf{B} with respect to time and space coordinates and the first powers of densities ρ and J of electric charges and currents. Therefore, these equations are *linear* and the fields obey superposition principle.

Physical significance of Maxwell's equations

The physical significance of Maxwell's equations can be readily interpreted from their mathematical statements.

- Maxwell's first equation (3.46 a) shows that the total electric flux density \mathbf{D} through the surface enclosing a volume is equal to the charge density ρ within the volume. It means that a charge distribution generates a steady electric field.
- Maxwell's second equation tells us that the net magnetic flux through a closed surface is zero. It implies that magnetic poles do not exist separately in the way as electric charges do. Thus, in other words, magnetic monopoles do not exist.
- The third equation shows that the emf around a closed path is equal to the time derivative of the magnetic flux density through the surface bounded by the path. It means that an electric field can also be generated by a time-varying magnetic field.
- Maxwell's fourth equation shows that the magneto motive force around a closed path is equal to the conduction current plus the time derivative of the electric flux density \mathbf{D} through any surface bounded by the path. The time derivative of the electric flux density $\partial\mathbf{D}/\partial t$ is called displacement current. Thus this equation means that a magnetic field is generated by a time-varying electric field.

3.14 MAXWELL'S EQUATIONS IN INTEGRAL FORM

Maxwell's equations (3.46 a) to (3.46 d) are in differential form. They can be converted into integral form by integrating them over an area and applying Stoke's theorem or by integrating throughout a volume and applying Divergence theorem. The integral forms of Maxwell's equations are given in the following table along with their differential forms:

	<i>Law</i>	<i>Differential form</i>	<i>Integral form</i>
1.	Gauss's law	$\nabla \cdot \mathbf{D} = \rho$	$\oint_S \mathbf{D} \cdot d\mathbf{s} = \int_V \rho dV$
2.	Gauss's law for magnetism	$\nabla \cdot \mathbf{B} = 0$	$\oint_S \mathbf{B} \cdot d\mathbf{s} = 0$
3.	Faraday's law	$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t}$	$\oint_C \mathbf{E} \cdot dl = - \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot ds$
4.	Ampere's law	$\nabla \times \mathbf{H} = \mathbf{J} + \epsilon \frac{\partial \mathbf{E}}{\partial t}$	$\oint_C \mathbf{H} \cdot dl = \oint_S \left(\mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \right) \cdot ds$

QUESTIONS

- Explain the terms magnetic induction and magnetic flux.
- State and explain Biot-Savart law.
- State and explain Ampere's circuital law.
- Write the integral form of the Ampere's circuital law.
- Obtain the differential form of Ampere's circuital law.
- Write the integral and differential forms of Gauss's law in electrostatics in vacuum.
- Apply Gauss's law and obtain an expression for magnetic field. What is the conclusion one can draw from the expression?
- Explain magnetic scalar potential and derive an expression for the same.
- Express the magnetic field in terms of vector potential.

10. State and explain Faraday's laws of electromagnetic induction.
11. What is Lenz's law of electromagnetic induction? Explain how the direction of current in a circular loop can be established with the help of Lenz's law.
12. Deduce the integral form and differential forms of Faraday's law.
13. How was the concept of displacement current helpful in removing discrepancy in Ampere's law?
14. Distinguish between conduction current and displacement current.
15. What is the physical significance of equations (i) $\nabla \cdot \mathbf{B} = 0$ and (ii) $\nabla \cdot \mathbf{D} = \rho$?
16. What is displacement current? Show that the conduction current in a lead wire is identical with the displacement in gap of the capacitor.

CHAPTER

4

Electromagnetic Waves

4.1 INTRODUCTION

Maxwell unified the theories of electricity and magnetism by way of deducing four very important equations, which combined the experimental observations reported by Gauss, Ampere, and Faraday with his concept of displacement current. The equations encapsulate the connection between the electric field and electric charge and between the magnetic field and electric current. The Maxwell's equations also define the bilateral coupling between the electric and magnetic field quantities. They along with some auxiliary equations form the fundamental tenets of electromagnetic theory. When the charge and current sources vary with time, the electric and magnetic fields become interconnected and the coupling between them produces electromagnetic waves capable of traveling through free space and in material media. In all there are four Maxwell's equations. These equations cannot be derived since they are the fundamental axioms or postulates of electrodynamics, obtained with the help of generalization of experimental results.

4.2 ELECTROMAGNETIC WAVES

Maxwell showed that by combining the four electrodynamics equations a *wave equation* was obtained which described the propagation of waves. The time variation of a magnetic field induces an electric field, while a variation of an electric field, in its turn, induces a magnetic field and electromagnetic fields can exist independently, without electric charges and currents. The continuous inter-conversion of the fields preserves them and an electromagnetic perturbation propagates in space. Such fields are called **electromagnetic waves**. The generation of electromagnetic wave does not require any medium. **The electromagnetic waves propagate through space entirely on their own.** Maxwell's theory placed no restriction on possible

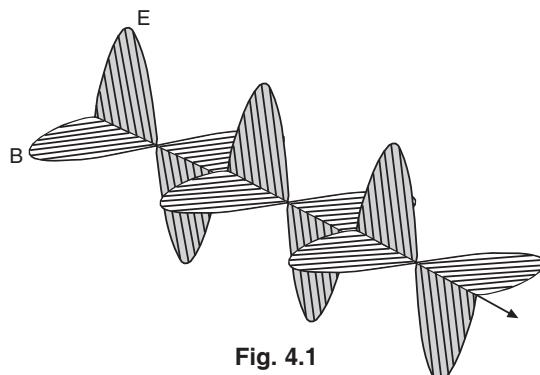


Fig. 4.1

wavelengths for the electromagnetic radiation. The vector cross products in Maxwell's third and fourth equations imply that the two fields, \mathbf{E} and \mathbf{H} are normal to each other and also normal to the direction of propagation. In a vacuum, these waves always propagate at a velocity equal to the velocity of light c .

The vectors \mathbf{E} and \mathbf{B} in an electromagnetic wave always oscillate in phase (see Fig. 4.1). The instantaneous values of \mathbf{E} and \mathbf{B} at any point are connected through the relation

$$\sqrt{\epsilon} E = \sqrt{\mu} H$$

This means that \mathbf{E} and \mathbf{H} (or \mathbf{B}) simultaneously attain their maximum values, vanish etc.

Note that electromagnetic waves are, in general, neither longitudinal nor transverse. In some types of electromagnetic waves, the electric vector \mathbf{E} is at right angles to the ray while the magnetic vector \mathbf{H} is not. Waves of this type are called **transverse electric waves** or TE waves. Since \mathbf{H} vector is not normal to the ray, it must have a component along the ray direction. In another type of waves the magnetic vector is at right angles to the ray while the electric vector \mathbf{E} is not. Waves of this type are called **transverse magnetic waves** or TM waves. If both the vectors \mathbf{E} and \mathbf{H} are at right angles to the ray, the waves are called **transverse electromagnetic waves** or TEM waves.

4.3 ELECTROMAGNETIC WAVE EQUATIONS

Maxwell equations are coupled partial differential equations in \mathbf{E} and \mathbf{H} , which cannot be solved in general. In order to simplify the equations, we should uncouple the set and obtain differential equations in \mathbf{E} or \mathbf{H} alone.

Maxwell's fourth equation is

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad (4.1)$$

By substituting $\mathbf{J} = \sigma \mathbf{E}$ and $\mathbf{D} = \epsilon \mathbf{E}$, the above equation (4.1) may be rewritten as

$$\nabla \times \mathbf{E} = \sigma \mathbf{E} + \epsilon \frac{\partial \mathbf{E}}{\partial t} \quad (4.2)$$

Maxwell's third equation is

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\mu \frac{\partial \mathbf{H}}{\partial t} \quad (4.3)$$

Taking curl of equ. (4.3), we get

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu \frac{\partial}{\partial t} (\nabla \times \mathbf{H})$$

Using equ. (4.2) into the above equation, we obtain

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu \frac{\partial}{\partial t} \left[\sigma \mathbf{E} + \epsilon \frac{\partial \mathbf{E}}{\partial t} \right]$$

or
$$\nabla \times (\nabla \times \mathbf{E}) = -\mu \sigma \frac{\partial \mathbf{E}}{\partial t} - \mu \epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2}$$

But
$$\nabla \times (\nabla \times \mathbf{E}) = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}$$

If the charge density is zero, $\nabla \cdot \mathbf{E} = 0$

$$\therefore \nabla^2 \mathbf{E} = \mu \sigma \frac{\partial \mathbf{E}}{\partial t} + \mu \epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} \quad (4.4)$$

or
$$\nabla^2 \mathbf{E} - \mu \epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} - \mu \sigma \frac{\partial \mathbf{E}}{\partial t} = 0 \quad (4.5 \text{ a})$$

This is the three-dimensional wave equation for the electric field in a conducting medium which is homogeneous and isotropic. Similar equation for magnetic field can be obtained following a similar procedure. Thus,

$$\nabla^2 \mathbf{H} - \mu \epsilon \frac{\partial^2 \mathbf{H}}{\partial t^2} - \mu \sigma \frac{\partial \mathbf{H}}{\partial t} = 0 \quad (4.5 \text{ b})$$

4.4 MAXWELL'S WAVE EQUATIONS FOR FREE SPACE

In order to understand the nature of waves, we consider *free space*, which is a large empty volume of space. Free space is a perfect dielectric and does not absorb ($\sigma = 0$) waves. As $\mu = \mu_0$ and $\epsilon = \epsilon_0$ for free space, equation (4.5a) for free space (or a dielectric medium) becomes

$$\begin{aligned} \nabla^2 \mathbf{E} - \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} &= 0 \\ \therefore \quad \nabla^2 \mathbf{E} &= \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \end{aligned} \quad (4.6)$$

This is the law that \mathbf{E} must obey.

A similar procedure for \mathbf{H} gives us

$$\therefore \quad \nabla^2 \mathbf{H} = \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{H}}{\partial t^2} \quad (4.7)$$

This is the law that \mathbf{H} must obey.

Equations (4.6) and (4.7) are very much similar to the general wave equation and constitute **wave equations**. They are *three-dimensional vector wave equations* and describe the propagation of an electromagnetic wave through a uniform medium. Since the wave equations for \mathbf{E} and \mathbf{H} are of the same form, their solutions will also have the same form. The solutions to equ. (4.6) and (4.7) lead to waves that can exist in free space. Even though the electric and magnetic fields of the waves start out on charges and currents, they detach themselves from them and move through free space as independent entities.

4.4.1 Velocity of the Electromagnetic Wave

The propagation characteristics of the electromagnetic wave are contained in the solution of equ. (4.6). To bring out the characteristics, we compare it with the general wave equation

$$\nabla^2 \xi = \frac{1}{v^2} \frac{\partial^2 \xi}{\partial t^2} \quad (4.8)$$

The comparison of equ. (4.6) with (4.8) gives

$$\begin{aligned} v^2 &= \frac{1}{\mu_0 \epsilon_0} \\ \text{or} \quad v &= \frac{1}{\sqrt{\mu_0 \epsilon_0}} \end{aligned} \quad (4.9)$$

Substituting the values of μ_0 and ϵ_0 , we find that

$$\frac{1}{\sqrt{\mu_0 \epsilon_0}} = \frac{1}{\sqrt{(4\pi \times 10^{-7} \text{ wb.A.m}^2)(8.9 \times 10^{-12} \text{ C}^2 / \text{N.m}^2)}} = 3.0 \times 10^8 \text{ m/s.}$$

$$\therefore \quad \frac{1}{\sqrt{\mu_0 \epsilon_0}} = c, \text{ the velocity of light} \quad (4.10)$$

Obviously, electromagnetic waves travel with the velocity of light in free space.

4.4.2 Relation between the Refractive Index and Relative Permittivity of a Medium

In case of a medium other than vacuum, we have to use $\epsilon_0 \epsilon_r (= \epsilon)$ and $\mu_0 \mu_r (= \mu)$, instead of ϵ_0 and μ_0 in equ. (4.10). We then get

$$v = \frac{1}{\sqrt{\mu_0 \mu_r \epsilon_0 \epsilon_r}} = \frac{1}{\sqrt{\mu_0 \epsilon_0}} \cdot \frac{1}{\sqrt{\mu_r \epsilon_r}} = \frac{c}{\sqrt{\mu_r \epsilon_r}}$$

or $\frac{c}{v} = \sqrt{\mu_r \epsilon_r}$

But $\frac{c}{v} = n$ (refractive index of the medium)
 $\therefore n = \sqrt{\mu_r \epsilon_r}$

For a non-magnetic medium $\mu_r = 1$. Therefore,

$$n = \sqrt{\epsilon_r} \quad (4.11)$$

or $n^2 = \epsilon_r$

4.5 UNIFORM PLANE WAVES

When energy is emitted by a source, it expands outwardly from the source in the form of spherical waves. The spherical wave travels at the same speed in all directions and therefore expands at the same rate. To an observer very far away from the source, the wave front of the spherical wave appears approximately planar. A plane wave is the simplest example of wave motion. In a plane wave, the electric and magnetic intensities are of constant value over any plane perpendicular to the direction of propagation. Such a plane is a surface of equal phase. A **plane wave** is thus a wave for which the phase has the same value at all points on an infinite plane. As the amplitude is also constant over the plane surface, it is called a **uniform plane wave**. Plane waves vary only in the direction of propagation and are uniform in planes normal to the direction of propagation. In this chapter we confine ourselves to the plane wave propagation in unbounded media. Plane wave propagation can be described by Cartesian coordinates, which are easier to work with mathematically than the spherical coordinates needed for describing propagation of a spherical wave.

4.5.1 The Transverse Nature of Plane Waves

Let us assume that the plane waves are traveling along the z -direction and hence \mathbf{E} is constant over any given plane parallel to the xy -plane. Similarly \mathbf{H} is constant over the xy -plane. We then have

$$\frac{\partial \mathbf{E}}{\partial x} = \frac{\partial \mathbf{E}}{\partial y} = 0 \quad \text{and} \quad \frac{\partial \mathbf{H}}{\partial x} = \frac{\partial \mathbf{H}}{\partial y} = 0 \quad (4.12)$$

These relations imply that

$$\frac{\partial E_x}{\partial x} = 0, \frac{\partial E_y}{\partial x} = 0, \frac{\partial E_z}{\partial x} = 0 \quad \text{and} \quad \frac{\partial E_x}{\partial y} = 0, \frac{\partial E_y}{\partial y} = 0, \frac{\partial E_z}{\partial y} = 0 \quad (4.13 \text{ a})$$

$$\frac{\partial H_x}{\partial x} = 0, \frac{\partial H_y}{\partial x} = 0, \frac{\partial H_z}{\partial x} = 0 \quad \text{and} \quad \frac{\partial H_x}{\partial y} = 0, \frac{\partial H_y}{\partial y} = 0, \frac{\partial H_z}{\partial y} = 0 \quad (4.13 \text{ b})$$

According to Maxwell' equation, we have

$$\nabla \cdot \mathbf{E} = 0$$

which means that

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0$$

$$\text{Since } \frac{\partial E_x}{\partial x} = 0 \quad \text{and} \quad \frac{\partial E_y}{\partial y} = 0, \quad \text{then} \quad \frac{\partial E_z}{\partial z} = 0. \quad (4.14)$$

It means that E_z is independent of z and has the same value all along the z -axis. Wave motion consists of changing values of \mathbf{E} along the direction of propagation. As E_z is constant along z -direction, it does not contribute to the wave motion and therefore E_z must be zero. $E_z = 0$ implies that \mathbf{E} lies in a plane perpendicular to the direction of propagation of the wave. Hence, the electric wave is a *transverse wave*.

Again, according to Maxwell's equation, we have

$$\nabla \cdot \mathbf{B} = 0 \quad \text{and hence} \quad \nabla \cdot \mathbf{H} = 0$$

which means that

$$\frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} + \frac{\partial H_z}{\partial z} = 0$$

$$\text{Since } \frac{\partial H_x}{\partial x} = 0 \quad \text{and} \quad \frac{\partial H_y}{\partial y} = 0, \quad \text{then} \quad \frac{\partial H_z}{\partial z} = 0. \quad (4.15)$$

Thus, H_z is independent of z and has the same value all along the z -axis. As H_z is constant along z -direction, it does not contribute to the wave motion and therefore H_z must be zero. $H_z = 0$ implies that \mathbf{H} lies in a plane perpendicular to the direction of propagation of the wave. Hence, the magnetic wave also is a *transverse wave*.

4.5.2 Relation between \mathbf{E} and \mathbf{H} in a Uniform Plane Wave

Now we assume that the variations of \mathbf{E} are simple harmonic and that \mathbf{E} is parallel to the y -axis. Then $E_z = 0$ and from equ. (4.6), we may write for a wave traveling in the positive z -direction,

$$E_y = E_1 e^{-i(\omega t - kz)} \quad (4.16)$$

$$\text{or in a simpler form as} \quad E_y = E_1 \cos[\omega(t - z/c)] \quad (4.17)$$

Maxwell's equation (4.3) can be written into the following three scalar equations.

$$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = -\frac{\partial B_x}{\partial t} \quad \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} = -\frac{\partial B_y}{\partial t} \quad \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -\frac{\partial B_z}{\partial t} \quad (4.18)$$

Using the results of equ. (4.13a) into the first relation of (4.18), we get

$$\frac{\partial E_y}{\partial z} = \frac{\partial B_x}{\partial t} \quad (4.19)$$

The magnetic flux density B_x associated with the electric field E_y can be found by integrating equ. (4.19). Thus,

$$\begin{aligned} B_x &= \int \frac{\partial E_y}{\partial z} dt = -\frac{E_1 \omega}{c} \int \sin [\omega(t - z/c)] dt \\ &= \frac{1}{c} E_1 \cos[\omega(t - z/c)] \end{aligned}$$

The constant of integration is omitted since we are not interested in a constant component of the field.

$$\therefore B_x = \frac{1}{c} E_y \quad (4.20)$$

or

$$H_x = \frac{1}{\mu c} E_y = \frac{\sqrt{\mu_0 \epsilon_0}}{\mu} E_y$$

$$\therefore H_x = \sqrt{\frac{\epsilon_0}{\mu_0}} E_y \quad (4.21)$$

Since E_y and H_x differ only by a scalar and have the same time dependence, \mathbf{E} and \mathbf{H} are in phase at all points in space and are mutually perpendicular. Their cross product $\mathbf{E} \times \mathbf{H}$ points in the direction of propagation denoted by the vector \mathbf{k} .

4.5.3 Characteristic Impedance

Equ. (4.21) may be rewritten as

$$\frac{E_y}{H_x} = \sqrt{\frac{\mu_0}{\epsilon_0}}$$

The above equation states that the ratio of the amplitudes of the vectors \mathbf{E} and \mathbf{H} always equal to the square root of the ratio of μ_0 and ϵ_0 . The ratio has the dimensions of impedance and hence it is called the **characteristic impedance** or **intrinsic impedance** of free space. It is denoted by η_0 . We generalize equ. (4.21) by writing it as

$$\frac{\mathbf{E}}{\mathbf{H}} = \eta_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} \quad (4.22)$$

Using the values of μ_0 and ϵ_0 into the above equation, we get

$$\eta_0 = \sqrt{\frac{4\pi \times 10^{-7} \text{ H/m}}{10^{-19} / 36\pi \text{ F/m}}} = 120\pi = 337 \text{ ohms} \quad (4.23)$$

In a dielectric material, $\eta = 337 / \epsilon_r^{1/2} = 337/n$, where n is the index of refraction of the material.

If \mathbf{E} and \mathbf{H} are in phase, η is a pure resistance. It is the case for free space and lossless dielectric media. If \mathbf{E} and \mathbf{H} are not in phase, as is the case in conducting media, the ratio of \mathbf{E} to \mathbf{H} is complex and hence η is called *intrinsic impedance*.

An electromagnetic wave traveling in the z -direction will have no E_z component. It may have E_y or E_x component. Such a wave is called a *linearly polarized* wave. If the field is directed in x -direction, the wave is said to be x -polarized or if it is directed in y -direction, it is said to be y -polarized. The electromagnetic wave is said to be a *plane wave* since both \mathbf{E} and \mathbf{H} lie in a plane; E_x and H_y lie in xy -plane. The wave is also uniform, since neither E_x or H_y vary with distance. Since the direction of propagation of the wave is perpendicular to E_x and H_y , the wave is called **transverse electromagnetic wave** or simply **TEM wave**.

4.6 ELECTROMAGNETIC ENERGY DENSITY

One of the important characteristics of electromagnetic waves is that it transports energy from one region to another region. The *energy density*, u , of a wave is the radiant energy per unit volume. The electromagnetic wave consists of electric field and magnetic field which independently can store energy. The energy density stored in the electric field \mathbf{E} , say existing between the plates of a charged capacitor, is given by

$$u_E = \frac{\epsilon_0}{2} E^2 \quad (4.24a)$$

Similarly the energy density stored in a magnetic field, say produced by a current carrying loop, is given by

$$u_B = \frac{1}{2\mu_0} B^2 \quad (4.24b)$$

The vectors E and B are related through

$$\mathbf{E} = c \mathbf{B}$$

$$\therefore u_E = \frac{\epsilon_0}{2} E^2 = \frac{\epsilon_0}{2} c^2 B^2 = \frac{1}{2\mu_0} B^2 = u_B$$

The above expression means that the energy in the electromagnetic waves is stored equally between the electric and magnetic fields. Thus,

$$u = u_E + u_B$$

Therefore,

$$u = \epsilon_0 E^2 \quad (4.25a)$$

or it can also be expressed as

$$u = \frac{1}{\mu_0} B^2. \quad (4.25b)$$

4.7 THE POYNTING THEOREM

An electromagnetic wave carries energy with it as it propagates through space. There exists a simple and direct relation between the rate of the energy flow and amplitudes of electric and magnetic intensities of the electromagnetic wave. We can describe the energy transfer in terms of the rate of energy flow per unit area or power per unit area, by a vector, \mathbf{S} , called the **Poynting vector**, which was introduced by the British physicist John H. Poynting.

Let us consider the Maxwell's curl equation

$$\begin{aligned} \nabla \times \mathbf{H} &= \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \\ \mathbf{E} \cdot \nabla \times \mathbf{H} &= \mathbf{E} \cdot \mathbf{J} + \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} \end{aligned} \quad (4.26)$$

Now making use of the vector identity

$$\nabla \cdot (\mathbf{E} \times \mathbf{H}) = \mathbf{H} \cdot \nabla \times \mathbf{E} - \mathbf{E} \cdot \nabla \times \mathbf{H}$$

or

$$\mathbf{E} \cdot \nabla \times \mathbf{H} = \mathbf{H} \cdot \nabla \times \mathbf{E} - \nabla \cdot (\mathbf{E} \times \mathbf{H})$$

Putting the value of $\mathbf{E} \cdot \nabla \times \mathbf{H}$ into equ. (4.26), we get

$$\begin{aligned} \mathbf{H} \cdot \nabla \times \mathbf{E} - \nabla \cdot (\mathbf{E} \times \mathbf{H}) &= \mathbf{E} \cdot \mathbf{J} + \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} \\ \text{or } -\mathbf{H} \frac{\partial \mathbf{B}}{\partial t} - \nabla \cdot (\mathbf{E} \times \mathbf{H}) &= \mathbf{E} \cdot \mathbf{J} + \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} \\ -\mu \mathbf{H} \frac{\partial \mathbf{H}}{\partial t} - \nabla \cdot (\mathbf{E} \times \mathbf{H}) &= \mathbf{E} \cdot \mathbf{J} + \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} \\ -\nabla \cdot (\mathbf{E} \times \mathbf{H}) &= \mathbf{E} \cdot \mathbf{J} + \epsilon \mathbf{E} \cdot \frac{\partial \mathbf{E}}{\partial t} + \mu \mathbf{H} \frac{\partial \mathbf{H}}{\partial t} \end{aligned}$$

As

$$\epsilon \mathbf{E} \frac{\partial \mathbf{E}}{\partial t} = \frac{\epsilon}{2} \frac{\partial \mathbf{E}^2}{\partial t} = \frac{\partial}{\partial t} \left(\frac{\epsilon \mathbf{E}^2}{2} \right)$$

and

$$\mu \mathbf{H} \frac{\partial \mathbf{H}}{\partial t} = \frac{\mu}{2} \frac{\partial \mathbf{H}^2}{\partial t} = \frac{\partial}{\partial t} \left(\frac{\mu \mathbf{H}^2}{2} \right)$$

$$-\nabla \cdot (\mathbf{E} \times \mathbf{H}) = \mathbf{E} \cdot \mathbf{J} + \frac{\partial}{\partial t} \left(\frac{\epsilon \mathbf{E}^2}{2} + \frac{\mu \mathbf{H}^2}{2} \right)$$

Rearranging the terms in the above equation and integrating it over a volume V, we obtain

$$\int_V (\mathbf{E} \cdot \mathbf{J}) dV = -\frac{\partial}{\partial t} \int_V \left(\frac{\epsilon \mathbf{E}^2}{2} + \frac{\mu \mathbf{H}^2}{2} \right) dV - \int_V \nabla \cdot (\mathbf{E} \times \mathbf{H}) \cdot dV$$

Using the divergence theorem the last term can be changed from volume integral to a surface integral. Thus,

$$\begin{aligned} \int_V \nabla \cdot (\mathbf{E} \times \mathbf{H}) dV &= \int_S (\mathbf{E} \times \mathbf{H}) \cdot d\mathbf{s} \\ \int_V (\mathbf{E} \cdot \mathbf{J}) dV &= -\frac{\partial}{\partial t} \int_V \left(\frac{\epsilon \mathbf{E}^2}{2} + \frac{\mu \mathbf{H}^2}{2} \right) dV - \int_S (\mathbf{E} \times \mathbf{H}) \cdot d\mathbf{s} \end{aligned} \quad (4.27)$$

The term on the left hand side represents the instantaneous power dissipated in volume V, which is a generalization of Joule's law. The first term on the right hand side is a negative time derivative of the stored electric and magnetic energy in the volume and hence represents the rate at which the stored energy in the volume is decreasing. The interpretation of the second term on the right hand side follows from the principle of conservation of energy. The rate of dissipation of energy in the volume V must equal the rate at which the stored energy in V is decreasing plus the rate at which energy is entering the volume from outside. Therefore, the second term represents the rate of flow of energy inward through the surface of the volume.

The term without the negative sign must hence represent the rate of flow of energy outward through the surface enclosing the volume.

4.8 THE POYNTING VECTOR

The integral of $\mathbf{E} \times \mathbf{H}$ over any surface gives the rate of energy flow through that surface. The vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$ is defined as the Poynting vector and has the dimensions of watts/sq.m. It is Poynting's theorem that the vector product $\mathbf{S} = \mathbf{E} \times \mathbf{H}$ at any point is a measure of the rate of energy flow per unit area at that point. The vector \mathbf{S} is perpendicular to both \mathbf{E} and \mathbf{H} and is in the direction of $\mathbf{E} \times \mathbf{H}$. The direction of \mathbf{S} indicates the direction of the energy density flow at that point.

$$\begin{aligned} \mathbf{S} &= \mathbf{E} \times \mathbf{H} \\ \mathbf{S} &= \frac{1}{\mu_0} \mathbf{E} \times \mathbf{B} \end{aligned}$$

or

$$\mathbf{S} = c^2 \epsilon_0 (\mathbf{E} \times \mathbf{B}) \quad (4.28)$$

Let us calculate energy dU passing during time dt through a unit area perpendicular to the direction of propagation of the wave. In a time dt , the wave front moves a distance $dz = c dt$.

$$dU = u \mathbf{c} dt$$

where u is the energy density and is given by $u = \epsilon_0 \mathbf{E}^2$.

$$\therefore du = \epsilon_0 \mathbf{E}^2 c dt$$

For an electromagnetic wave

$$\epsilon_0 \mathbf{E}^2 = \mu_0 \mathbf{H}^2 \quad \text{or} \quad \sqrt{\epsilon_0} \mathbf{E} = \sqrt{\mu_0} \mathbf{H}$$

which implies that the electric energy density in the electromagnetic wave at any instant is equal to the magnetic energy density at the same point.

$$du = \sqrt{\epsilon_0 \mu_0} EH c dt$$

or

$$du = EH dt$$

The term EH represents the magnitude of energy flux density vector \mathbf{S} . The Poynting vector \mathbf{S} gives the *instantaneous* power density. When E and H are changing with time, we are often interested in the average power. It is obtained by integrating the instantaneous Poynting vector \mathbf{S} over one period and dividing by one period.

4.9 WAVE PROPAGATION IN A LOSSY MEDIUM

When a medium has net conductivity σ , energy is dissipated in the medium and the medium is said to be a lossy medium. We assume that the medium is homogeneous, linear and isotropic. We further assume that it does not contain free volume charge. The Maxwell's equations in such a medium is given by

$$\nabla \cdot \mathbf{E} = 0 \quad (4.29)$$

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \quad (4.30)$$

$$\nabla \cdot \mathbf{H} = 0 \quad (4.31)$$

$$\nabla \times \mathbf{H} = \sigma \mathbf{E} + \epsilon \frac{\partial \mathbf{E}}{\partial t} \quad (4.32)$$

Taking curl of equ. (4.30), we get

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu \frac{\partial}{\partial t} (\nabla \times \mathbf{H})$$

Using equ. (4.32) into the above equation, we obtain

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu \frac{\partial}{\partial t} \left[\sigma \mathbf{E} + \epsilon \frac{\partial \mathbf{E}}{\partial t} \right]$$

or

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu \sigma \frac{\partial \mathbf{E}}{\partial t} - \mu \epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2}$$

But

$$\nabla \times (\nabla \times \mathbf{E}) = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}$$

Since the charge density is zero,

$$\nabla \cdot \mathbf{E} = 0$$

$$\nabla^2 \mathbf{E} = \mu \sigma \frac{\partial \mathbf{E}}{\partial t} + \mu \epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} \quad (4.33)$$

Similar equation for magnetic field is given by

$$\nabla^2 \mathbf{H} = \mu \epsilon \frac{\partial^2 \mathbf{H}}{\partial t^2} - \mu \sigma \frac{\partial \mathbf{H}}{\partial t} \quad (4.34)$$

Equ. (4.33) may be written as

$$\nabla^2 \mathbf{E} = j\omega \mu [\sigma + j\omega \epsilon] \mathbf{E}$$

or

$$\nabla^2 \mathbf{E} = [j\omega \mu \sigma - \omega^2 \mu \epsilon] \mathbf{E} \quad (4.35)$$

For a plane wave traveling in x -direction and considering only the E_y component, equ. (4.35) becomes

$$\frac{\partial^2 E_y}{\partial x^2} - [j\omega \mu \sigma - \omega^2 \mu \epsilon] E_y = 0 \quad (4.36)$$

Putting $\gamma^2 = [j\omega \mu \sigma - \omega^2 \mu \epsilon]$

$$\frac{\partial^2 E_y}{\partial x^2} - \gamma^2 E_y = 0$$

where $\gamma = \sqrt{j\omega\mu[\sigma + j\omega\epsilon]}$ and can be expressed as

$$\gamma = \alpha + j\beta = \sqrt{-\omega^2\mu\epsilon} \left[1 + \frac{\sigma}{j\omega\epsilon} \right]^{\frac{1}{2}} \quad (4.37)$$

where α represents the *attenuation constant* and measures the rate at which the amplitude of the wave gets attenuated while propagating in the medium. β is the phase *shift constant* and indicates the rate of change in phase per unit length in the direction of propagation. It is also called the *phase factor*. Since γ includes α and β , which together characterize the propagation of the wave, the factor γ is called the *propagation constant*.

4.9.1 Expressions for α and β

$$\gamma = \alpha + j\beta = \sqrt{-\omega^2\mu\epsilon} \left[1 + \frac{\sigma}{j\omega\epsilon} \right]^{\frac{1}{2}}$$

Squaring the above equation, we get

$$\alpha^2 - \beta^2 + j2\alpha\beta = -\omega^2\mu\epsilon + \omega\mu\sigma$$

Equating the real and imaginary parts on both sides, we get

$$\alpha^2 - \beta^2 = -\omega^2\mu\epsilon$$

and

$$2\alpha\beta = \omega\mu\sigma$$

Mathematical manipulations yield the following expressions for α and β .

$$\alpha = \omega \sqrt{\frac{\mu\epsilon}{2}} \left[\sqrt{1 + \left(\frac{\sigma}{\omega\epsilon} \right)^2} - 1 \right]^{1/2} \quad (4.38)$$

$$\beta = \omega \sqrt{\frac{\mu\epsilon}{2}} \left[\sqrt{1 + \left(\frac{\sigma}{\omega\epsilon} \right)^2} + 1 \right]^{1/2} \quad (4.39)$$

4.10 CONDUCTORS AND DIELECTRICS

According to the Maxwell's fourth equation we have,

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}$$

Recalling that $\mathbf{J} = \sigma\mathbf{E}$, the above equation becomes

$$\nabla \times \mathbf{H} = \sigma \mathbf{E} + \frac{\partial \mathbf{D}}{\partial t}$$

For a linearly polarized plane wave traveling in the x -direction with \mathbf{E} in the y -direction, the above equation will reduce to the following scalar equation,

$$\frac{\partial H_x}{\partial x} = \sigma E_y + \frac{\partial E_y}{\partial t} \quad (4.40)$$

Assuming that E_y is a harmonic function of time we have

$$E_y = E_0 e^{j\omega t}$$

$$\frac{\partial E_y}{\partial t} = j\omega E_0 e^{j\omega t}$$

Putting these values in equation (4.40), we get

$$-\frac{\partial H_x}{\partial x} = \sigma E_y + j\omega\epsilon E_y \quad (4.41)$$

Both terms of the right hand side of equation (4.41) have the dimensions of current density. The term σE_y represents the conduction current density, while the terms $j\omega\epsilon E_y$ represents the displacement current density. The ratio of the magnitude of the conduction current density to that of the displacement current density is given by $\frac{\sigma}{\omega\epsilon}$.

- (i) When $\omega\epsilon \gg \sigma$, that is, when the displacement current density is much greater than the conduction current density, the medium is a *loss-less or good dielectric*.
- (ii) When $\omega\epsilon \ll \sigma$, that is, when the conduction current density is much lesser than the displacement current density, the medium is classified as a *good conductor*.

$\sigma = \omega\epsilon$ can be considered to be the dividing line between dielectrics and conductors.

To be more specific the media can also be classified according to the value of ratio $\sigma/\omega\epsilon$. For dielectrics this ratio is less than 1/100; and for conductors this ratio is greater than 100.

It may be mentioned here that frequency is an important factor in determining whether a medium acts like a conductor or dielectric. For example a medium at 1kHz behaves like a conductor, while at 30 GHz its act like a dielectric. At 10MHz its behavior is that of a quasi-conductor.

4.10.1 Wave Propagation in a Good Dielectric

For a good dielectric $\sigma \ll \omega\epsilon$ i.e. conduction current is very small compared to displacement current. We have

$$\gamma = \alpha + j\beta = \sqrt{-\omega^2\mu\epsilon} \left[1 + \frac{\sigma}{j\omega\epsilon} \right]^{\frac{1}{2}} = j\omega\sqrt{\mu\epsilon} \left[1 + \frac{\sigma}{j\omega\epsilon} \right]^{\frac{1}{2}}$$

We expand the factor $\left[1 + \frac{\sigma}{j\omega\epsilon} \right]^{\frac{1}{2}}$ using the binomial expansion and retain the first two terms. Thus, we get

$$\gamma = \alpha + j\beta = j\omega\sqrt{\mu\epsilon} \left[1 + \frac{\sigma}{2j\omega\epsilon} + \frac{\sigma^2}{8\omega^2\epsilon^2} \right]$$

Equating the real and imaginary parts on both sides of the above equation, we get

$$\alpha = \frac{\sigma}{2} \sqrt{\frac{\mu}{\epsilon}} \left(1 - \frac{\sigma^2}{8\omega^2\epsilon^2} \right) \quad (4.42)$$

The expression for β is given by

$$\beta = \omega\sqrt{\mu\epsilon} \left(1 + \frac{\sigma^2}{8\omega^2\epsilon^2} \right) \quad (4.43)$$

$\omega\sqrt{\mu\epsilon}$ is the phase shift for a perfect dielectric. The effect of a small amount of loss is to add the second term as a small correction factor. The velocity of the wave in the dielectric is given by

$$v = \frac{\omega}{\beta} = \frac{1}{\sqrt{\mu\epsilon} \left(1 + \frac{\sigma^2}{8\omega^2\epsilon^2} \right)}$$

i.e.

$$v = \frac{1}{\sqrt{\mu\epsilon}} \left(1 - \frac{\sigma^2}{8\omega^2\epsilon^2} \right) \quad (4.44)$$

where $\left(\frac{1}{\sqrt{\mu\epsilon}} \right)$ is the velocity of the wave in the dielectric when the conductivity is zero i.e., in the perfect dielectric. The effect of a small amount of loss is to reduce slightly the velocity of propagation of the wave.

4.10.2 Wave Propagation in a Good Conductor

For a good conductor, we have $\omega\epsilon \ll \sigma$. We write the factor γ in the following form.

$$\gamma = \alpha + j\beta = \sqrt{j\omega\mu\sigma} \left[1 + \frac{j\omega\epsilon}{\sigma} \right]^{\frac{1}{2}}$$

We can express

$$\sqrt{j} = \sqrt{\frac{2j}{2}} = \sqrt{\frac{1+2j-1}{2}} = \sqrt{\frac{(1+j)^2}{2}} = \frac{1+j}{\sqrt{2}}$$

Since $\omega\epsilon/\sigma \ll 1$, we expand the factor $\left[1 + \frac{j\omega\epsilon}{\sigma} \right]^{\frac{1}{2}}$ as a power series and retain the first two terms. Then we obtain

$$\text{Therefore, } \gamma = \alpha + j\beta = \sqrt{\frac{\omega\mu\sigma}{2}} (1+j) \left[1 + \frac{j\omega\epsilon}{2\sigma} \right]$$

Equating the real and imaginary parts in the above equation, we get

$$\alpha = \sqrt{\frac{\omega\mu\sigma}{2}} \left[1 - \frac{\omega\epsilon}{2\sigma} \right] \quad (4.45)$$

and

$$\beta = \sqrt{\frac{\omega\mu\sigma}{2}} \left[1 + \frac{\omega\epsilon}{2\sigma} \right] \quad (4.46)$$

$$\text{It may be seen that } \alpha \approx \sqrt{\frac{\omega\mu\sigma}{2}} \approx \beta \quad (4.47)$$

The velocity of propagation of the wave in the conductor will be

$$v = \frac{\omega}{\beta} = \sqrt{\frac{2\omega}{\mu\sigma}} \quad (4.48)$$

4.10.2.1 Depth of Penetration in a Good Conductor

Using the values of α and β , the wave equation $E_y = Ae^{-(\alpha + j\beta)x}$ can be written as

$$E_y = Ae^{-\sqrt{\frac{\omega\mu\sigma}{2}}x} e^{-j\sqrt{\frac{\omega\mu\sigma}{2}}x} \quad (4.49)$$

The above equation may be rewritten as

$$E_y = Ae^{-\frac{x}{\delta}} e^{-j\frac{x}{\delta}}$$

$$\text{At } x = 0, E_y = A \text{ and at } x = \delta, E_y = \frac{A}{e}.$$

It means that E_y decreases to $1/e$ of its initial value while the wave penetrates to a distance δ in the conducting medium. It is called the depth of penetration or **skin depth**. The skin depth is

$$\delta = \sqrt{\frac{2}{\omega \mu \sigma}} = \sqrt{\frac{1}{\pi v \mu \sigma}} \quad (4.50)$$

The skin depth is inversely proportional to σ , μ and v . Copper has a conductivity of $\sigma = 5.8 \times 10^7$ mhos/m and its permeability is approximately equal to that of free space. The depth of penetration of 1 MHz wave in copper is about 0.0667 mm, whereas it is 8.67 mm at 60 Hz.

Example 4.1: The wave function of a light wave is $E(z, t) = 10^3 \sin \pi(3 \times 10^6 x - 9 \times 10^{14} t)$

- (a) determine the speed, wavelength, frequency and period of the wave,
- (b) determine the magnetic field associated with the wave.

Solution: (i)
$$E(z, t) = 10^3 \sin \pi(3 \times 10^6 x - 9 \times 10^{14} t)$$

$$= 10^3 \sin 3 \times 10^6 \pi(x - 3 \times 10^8 t) \quad (1)$$

The equation is similar to the general wave equation,

$$E(z, t) = E_0 \sin k(x - vt) \quad (2)$$

Comparing (1) with (2), we get,

$$v = 3 \times 10^8 \text{ m/s, and } k = 3 \times 10^6 \pi / \text{m}$$

$$\therefore \lambda = \frac{2\pi}{k} = \frac{2\pi}{3 \times 10^6 \pi} = 6.666 \times 10^{-7} \text{ m} = 6666 \text{ Å}$$

$$v = \frac{\nu}{\lambda} = \frac{3 \times 10^8 \text{ m/s}}{6.666 \times 10^{-7} \text{ m}} = 4.5 \times 10^{14} \text{ Hz}$$

$$T = \frac{1}{v} = \frac{1}{4.5 \times 10^{14} \text{ Hz}} = 2.2 \times 10^{-15} \text{ s}$$

(ii) The light wave propagates in the z -direction while the E-vector oscillates along x -direction in xz -plane. Since in an EM wave, the magnetic field B is normal to both E and wave propagation directions, it should be in the yz -plane.

Thus, $B_x = 0$; $B_z = 0$

$$B = B_y(z, t) = B(z, t)$$

$$\text{As } E = cB, \quad B = \frac{E}{c}$$

$$\therefore B(z, t) = \frac{10^3 \sin \pi(3 \times 10^6 y - 9 \times 10^{14} t)}{c} = \frac{10^3 \sin \pi(3 \times 10^6 y - 9 \times 10^{14} t)}{3 \times 10^8}$$

$$= 3.33 \times 10^{-6} \sin \pi(3 \times 10^6 y - 9 \times 10^{14} t)$$

Example 4.2: An electromagnetic wave moving through free space has an electric field given by, $E = 100 \sin \left[8\pi \times 10^{14} \left(t - \frac{z}{3 \times 10^8} \right) \right]$. Calculate the corresponding intensity.

Solution: $E = 100 \sin \left[8\pi \times 10^{14} \left(t - \frac{z}{3 \times 10^8} \right) \right]; \quad \therefore E_0 = 100 \text{ V/m}$

$$\text{Intensity, } I = \left(\frac{c \epsilon_0}{2} \right) E_0^2 = \frac{(3 \times 10^8) \left(8.85 \times 10^{-12} \frac{\text{C}^2}{\text{Nm}^2} \right) (100 \text{ V/m})^2}{2} = 13.3 \text{ W/m}^2$$

QUESTIONS

1. State and explain Maxwell's equations for electromagnetic field. Starting from Maxwell's equations, deduce the wave equation for a plane wave in free space.
2. Show that the ratio of the electric and magnetic fields of a uniform plane wave is constant depending upon the medium.
3. What is meant by the Poynting vector? Derive Poynting vector from Maxwell's equations and explain its physical significance.
4. Solve the Maxwell's equations for perfect dielectric containing no charge and no conduction current.
5. State Maxwell's equations in their general /integral form. Derive their form for harmonically varying fields.
6. Show that the speed of propagation of electromagnetic wave in free space is given by $c = 1/\sqrt{\mu_0 \epsilon_0}$.
7. Establish Maxwell's equations for electromagnetic fields and obtain an expression for poynting vector.
8. Write down Maxwell's field equations in differential and explain their physical meaning. Prove poynting theorem relating, the flow of energy at a point in space in an electromagnetic field.
9. Obtain Maxwell's equations and deduce an expression for the velocity of propagation of plane electromagnetic wave in a medium of dielectric constant ϵ and permeability μ .
10. Obtain an expression for: (i) the electric vector, (ii) the poynting vector, (iii) the field energy density and (iv) the momentum density in the field for a plane wave given by $B = B_0 \cos \frac{2\pi}{\lambda} (z - ct) y$ moving along the z-direction in free space. [Andhra 2001]
11. Write the wave equation for the electric field in an ionized medium. (B.P.U.T. 2004)
12. Starting from Maxwell's electromagnetic equations in free space, in absence of charges and currents, obtain the wave equation for electric field. (B.P.U.T. 2004)
13. State Poynting theorem. Explain how the Poynting vector explains the energy flow. (B.P.U.T. 2004)
14. Write Maxwell's electromagnetic equations in free space in the presence of charges and currents. Name each symbol used in the equations. (B.P.U.T. 2004)
15. Obtain the wave equation for electric field in a vacuum from appropriate Maxwell equations. (B.P.U.T. 2004)
16. Define Poynting vector. Mention its dimension and SI unit. (B.P.U.T. 2004)
17. Mention the boundary conditions satisfied by electric field and electric displacement at the boundary of two media. (B.P.U.T. 2004)

18. Write the Maxwell's electromagnetic equations in differential form in a medium, in the presence of charges and currents. Identify and state the laws of electromagnetism with which these equations are associated. **(B.P.U.T. 2003)**
19. Mention the boundary conditions satisfied by E, D, B and H at the interface between two non-conducting media. **(B.P.U.T. 2003)**
20. Starting from Maxwell's electromagnetic equations in free space, obtain the wave equations in terms of scalar and vector potentials. Mention the gauge conditions used. **(B.P.U.T. 2003)**

PROBLEMS

1. A plane electromagnetic wave in free space has an average poynting vector of 1 watt/m². Find the average energy density. **[Ans: 3.33×10^{-7} J/m³]**
2. Find the velocity of a plane wave in a lossless medium having a relative permittivity of 5 and relative permeability of 1. **[Ans: 1.34×10^8 m/sec]**
3. A plane sinusoidal linearly polarized electromagnetic wave of wavelength $\lambda = 5 \times 10^{-7}$ m travels in vacuum in the direction of the X-axis. The average intensity of the wave per unit area is 0.1 W/m² and plane of the vibration of the electric field is parallel to the Y-axis. Write the equations describing the electric and magnetic fields of the wave.
[Ans: $E_y = \sqrt{24\pi} \cos [4\pi \times 10^6 (x - ct)]$ N/coulomb, $B_z = (E_y / c)$ Tesla]
4. What must be the strength of uniform electric field if it is to have the same energy density as that possessed by a 400 gauss magnetic field [1 W/m² = 10^4 gauss, $\epsilon_0 = 8.85 \times 10^{-12}$ F/m and $\mu_0 = 4\pi \times 10^{-7}$ H/m].
5. Find the Brewster angle for a parallel-polarized wave traveling from air into glass for which $\epsilon_r = 5.0$ **(Ans: 65.91°)**
6. In free space, $E(z, t) = 1.0 \sin(\omega t - \beta z) a_x$ V/m. Show that the average power crossing a circular disk of radius 15.5 m in a $z = \text{const}$ plane is 1 W.
7. In free space, $H(z, t) = 1.33 \times 10^{-1} \cos(4 \times 10^7 t - \beta z) a_x$ A/m. Obtain an expression for $E(z, t)$. Find β and λ . **($E_0 = 50$ V/m, $4/3$ rad/m, 15π m.)**
8. A plane electromagnetic wave is traveling in an unbounded lossless dielectric medium having $\mu_r = 1$ and $\epsilon_r = 3$ and has peak electric field intensity of 6 V/m. Find (a) velocity of the wave (b) the intrinsic impedance of the medium (c) the peak value of the magnetic intensity. **(Ans: 1.73×10^8 m/s; 217.66 ohm; 2.76×10^{-2} A/m)**
9. A laser beam from 100 watt source is focused on an area of 10^{-8} m². Evaluate the magnitude of the Poynting vector on the area. **(B.P.U.T. 2004)**

CHAPTER

5

Light

5.1 INTRODUCTION

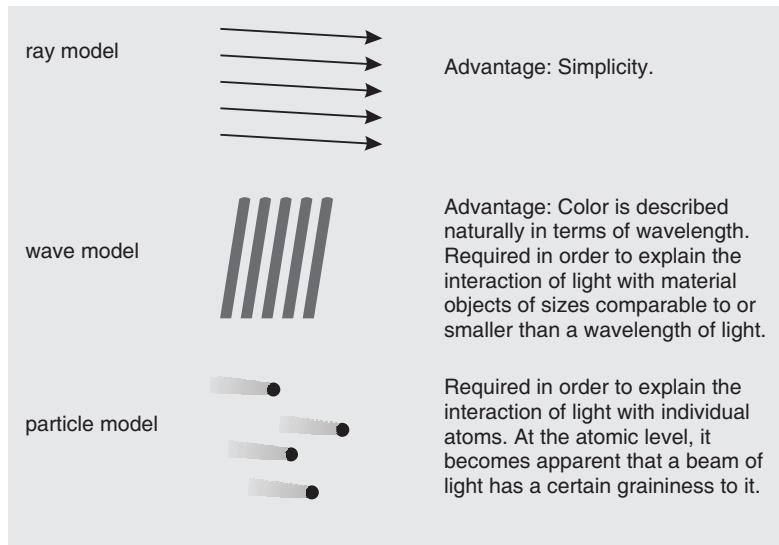
The application of optics started with the manufacture of mirrors, lenses and prisms. Lenses help mankind to overcome the natural barriers of vision. In the form of spectacles, they help us overcome our defects in vision. In the form of a telescope and a microscope, they help us expand the frontiers of vision to cover macrocosmos and microcosmos. Prisms revealed the composition of light and played a central role in disclosing the structure of atoms. Lenses also led to the development of still camera, followed by movie camera. All these applications are developments in traditional optics.

The advent of laser in 1960 opened up an entirely novel application of light. Conventional communication methods based on r.f. and microwaves have a limited capacity to carry messages. The manufacture of low loss glass fibre cables in 1970 enabled the implementation of light waves in communication in place of r.f. and microwaves. The art and science of optical communications have been rapidly perfected since then. The modern technology in its latest phase employs photons to go round optical circuits. This new technology is known as **photonics**. Photonics is used in optical computing, information processing etc vital areas. Learning optics has become much more important and relevant now than ever before.

5.2 NATURE OF LIGHT

Light is a form of energy that stimulates our vision. Energy can be transported in space in two ways—either by the actual motion of matter or by a wave disturbance. Accordingly, two independent hypotheses were proposed to explain light propagation. Isaac Newton (1642-1727) assumed that light consists of a stream of tiny particles (corpuscles) emitted by a light source. He used the corpuscular hypothesis of light to explain reflection and refraction of light. The Dutch physicist Christian Huygens (1629-1695), a contemporary of Newton assumed that light propagates in the form of waves and explained the optical phenomena known at that time. The corpuscular theory met with a limited success. The works of Thomas Young (1773-1829) and Augustin Fresnel (1788-1829) furnished convincing experimental support to the Huygens wave theory. The propagation of light, the phenomena of interference, diffraction and polarization are successfully explained by the wave theory of light. Further, Fresnel established that the light waves are transverse waves. In 1873 James Clerk Maxwell (1831-1879) developed the electromagnetic theory and propounded that light is a form of high frequency electromagnetic wave. However, Maxwell's electromagnetic theory failed to account for phenomena, such as photoelectric effect, which are associated with emission and absorption of light. Max Planck (1858-1947) put forward the quantum theory of light in 1900

which was used by Albert Einstein (1879-1955) in 1905 to explain photoelectric effect. Niels Bohr (1885-1962) postulated in 1913 the model of an atom which explains the emission and absorption of light. According to the quantum theory, light is emitted in the form of photons which are bundles of electromagnetic radiation that oscillate with a definite frequency and propagate through space with the speed of light. Individual photons behave like particles but when their number is large they exhibit the properties of a continuous wave.



5.3 THE VELOCITY OF LIGHT

Maxwell showed that light is an electromagnetic wave. Electromagnetic waves travel at an enormous speed. Measurements have shown that light travels in a vacuum at a speed of 2.998×10^8 m/s. The velocity of light is denoted by the symbol 'c'. For all practical purposes, it is taken as

$$c = 3 \times 10^8 \text{ m/s} \quad (5.1)$$

It is given by

$$c = v\lambda \quad (5.2)$$

where v is the frequency and the λ is the vacuum wavelength of light.

5.4 OPTICAL MEDIUM

A material through which light propagates is called an **optical medium**. Some materials such as air, water and glass readily transmit light and are said to be **transparent**. We can see through transparent materials. Some materials transmit part of the light but scatter most of it so that objects seen through them are not clearly visible. Such materials are called **translucent**. Materials which do not transmit light at all said to be **opaque**. We cannot see through opaque materials. Our study is related to the propagation of light through transparent media.

When light travels through any medium, its velocity reduces. The dependence of velocity of light on the medium is characterized by the quantity called **optical density**, which is expressed in terms of the absolute refractive index of the medium. The word absolute is not usually mentioned but it is implied. The **absolute refractive index** μ of an optical medium is defined as *the ratio of the velocity of light in a vacuum to the velocity of the light in the medium*.

$$\mu = \frac{\text{Velocity of light in a vacuum}}{\text{Velocity of light in the medium}} = \frac{c}{v} \quad (5.3)$$

The refractive index is a dimensionless number greater than unity since v is always less than c . The refractive index is 1 for a vacuum and it is equal to 1.0003 for air. Therefore, the velocity of light in air is considered to be equal to c for all practical purposes. An optical medium with a relatively high refractive index is said to have a high optical density and one with a low refractive index has a low optical density.

According to the relation (5.2) the decrease in the velocity of light in a medium may be attributed to either a decrease in a wavelength or frequency or both. However, in view of the law of conservation of energy, the frequency must remain constant. The frequency equals the number of wave crests per second. The number of wave crests per second approaching the boundary in one medium must equal the number of wave crests traveling away from the boundary in the second medium, as illustrated in Fig. 5.1 (a). If this were not the case, either wave crests would accumulate at the boundary or would be destroyed or created at the boundary. Such a loss or gain would be contrary to the law of conservation of energy. Hence, the frequency must be constant when a wave crosses a boundary. It implies that the decrease in velocity of light in a medium must lead to a decrease in the wavelength of light, as shown in Fig. 5.1 (b).

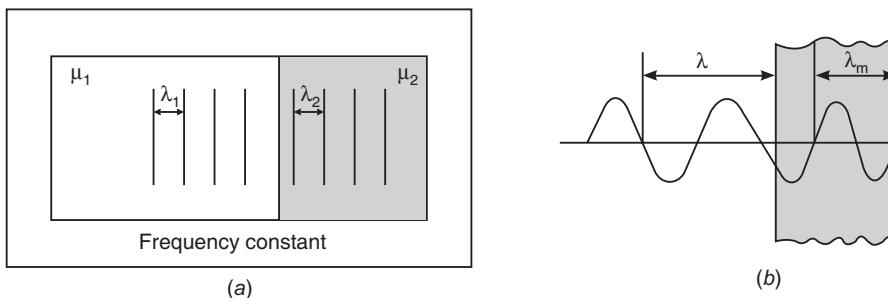


Fig. 5.1: (a) When a light wave propagates from medium 1 to another medium 2, the frequency remains constant. (b) Reduction in wavelength when a light wave travels from a rarer medium into a denser medium.

Table 1: Refractive indices for some materials ($\lambda = 5893 \text{ \AA}$)

Material	μ	Material	μ
Gases (at 0° C and 1 atm)		Solids	
Hydrogen	1.000132	Ice	1.310
Air	1.000293	Silica Glass	1.458
Carbon dioxide	1.000541	Rock salt	1.544
Liquids		Quartz	1.55
Methyl alcohol	1.329	Spinel	1.72
Water	1.333	Corundum	1.76
Ethyl alcohol	1.362	Zircon	1.923
Glycerine	1.473	Fabulite	2.409
Benzene	1.501	Diamond	2.417
Carbon disulphide	1.628	Plastics	
		Lucite	1.491
		Polystyrene	1.595

In tune with the relation (5.2) the velocity of light in an optical medium may be expressed as

$$v = \nu \lambda_m \quad (5.4)$$

where λ_m is the wavelength of light in the medium. Using equations (5.2) and (5.3) into equ. (5.4), we obtain

$$\mu = \frac{\lambda}{\lambda_m} \quad (5.5)$$

Table 1 shows the values of refractive index for a few materials.

5.5 HOMOGENEOUS ISOTROPIC MEDIUM

An optical medium can be classified into a *homogeneous* or an *inhomogeneous* medium depending on the spatial variation of its refractive index. In a **homogeneous medium** refractive index does not vary from point to point and light propagates along straight line paths in it. On the other hand, if the refractive index varies from point to point in the medium, it is said to be optically **inhomogeneous**. In such a medium light propagates along curved paths. The Earth's atmosphere is an optically inhomogeneous medium and light travels along curved paths in it. The passage of light through the atmosphere gives rise to many interesting optical effects. The distorted sun disk during the sunset, the mirages, the looming effect (Fig. 5.2), the twinkle of stars at night etc., are all due to the inhomogeneous nature of atmosphere. Such effects will not be observed on the Moon as there is no atmosphere around it.

A material is said to be **isotropic**, if it exhibits the same values of physical properties in all directions. Glass, water, air, etc., are isotropic substances. They are also homogeneous. Hence we call them **homogeneous isotropic media**. Such materials have the same value for refractive index in the three principal directions. Materials showing different values of physical properties in different directions are said to be **anisotropic**. Quartz, calcite, ice etc., crystals are examples of anisotropic medium. Such materials exhibit different refractive indices in different directions as a result of which an incident light ray splits into two rays and they propagate along two different directions with different velocities. This phenomenon is known as **double refraction or birefringence** (Fig. 5.3).

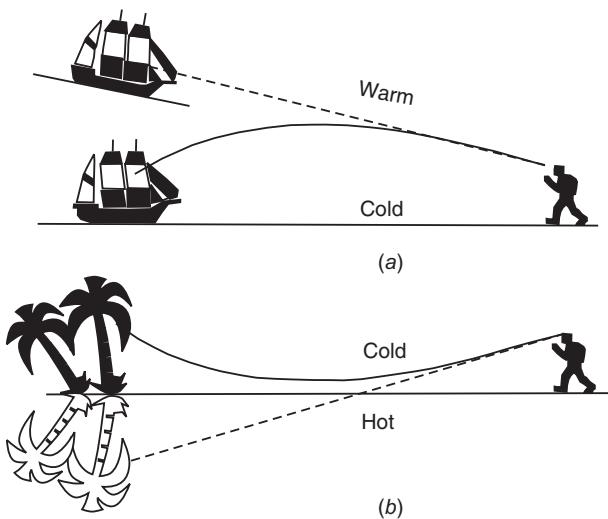


Fig. 5.2: Optical effects due to inhomogeneous nature of atmosphere (a) looming (b) mirages



Fig. 5.3: A calcite crystal laid upon a paper with some letters showing the double refraction

5.6 REFLECTION AND REFRACTION

Reflection

We see most of things with the help of light reflected from them. Unless light shines on objects, we do not see them. When a light wave encounters the boundary separating two optical media, the wave is partly reflected into the first medium and partly transmitted into the second medium. For simplicity, we consider the waves in terms of rays. If the surface irregularities are small compared to the wavelength of the incident light, the boundary is said to be *smooth*; otherwise, it is a *rough surface*. If a surface is smooth, it reflects light *specularly*. **Specular reflection** means that the angle of reflection of light is equal to the angle of incidence. Polished metals, mirrors, liquid surfaces etc., smooth surfaces reflect light specularly (Fig. 5.4a). However, many surfaces have microscopic irregularities and the reflection is *diffuse*. A rough surface reflects the light at many angles as shown in Fig. 5.4 (b). This is why glare is not experienced when light is reflected from a wall, whereas light specularly reflected from water surface causes glare. In physics, the term reflection is used to mean specular reflection.

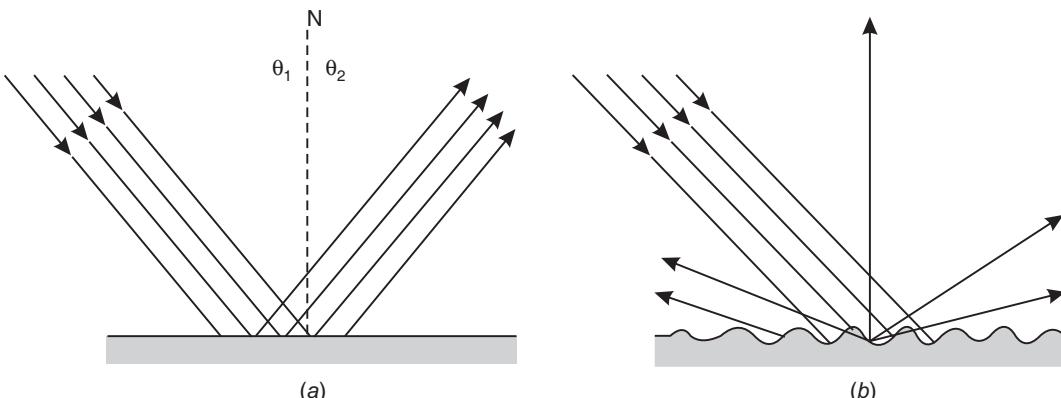


Fig. 5.4: Phenomenon of reflection (a) specular reflection occurs at smooth surfaces. The reflected rays are parallel; (b) diffuse reflection occurs at a rough surface. The rays are reflected in random directions.

Specular reflection obeys the following two laws known as **laws of reflection**:

- The incident ray, reflected ray and the normal to the surface all lie in the same plane, as shown in Fig. 5.4 (a).
- The angle of reflection θ_2 is equal to the angle of incidence θ_1 .

$$\theta_1 = \theta_2 \quad (5.6)$$

Further, the angle of reflection is not a function of colour and properties of the reflecting surface.

Refraction

The light ray that enters a transparent medium from another transparent medium is *bent* at the boundary and is said to be **refracted**. The boundary is known as a **refracting surface**. The *angle of refraction* r depends on the properties of the two media and on the *angle of incidence* i .

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2} = \frac{c/\mu_1}{c/\mu_2} = \frac{\mu_2}{\mu_1} = \mu_{21} \quad (5.7)$$

where v_1 is the velocity of light in medium 1 having a refractive index μ_1 and v_2 is the velocity of light in medium 2 having a refractive index μ_2 . The ratio $\frac{v_1}{v_2}$ or $\frac{\mu_2}{\mu_1}$ is called the

relative refractive index. The relation (5.7) was discovered by Willebrord Snell (1591-1627), the Dutch scientist and independently by Rene Descartes (1596-1660), the French mathematician. The relation is more popularly known as **Snell's law**. It follows from Snell's law that an increase in the angle of incidence causes an increase in the angle of refraction.

Further, for a given angle of incidence, the angle of refraction depends on the wavelength of the incident light and varies from colour to colour.

The experimental result (5.7) and the fact that the incident ray, refracted ray and the normal to the refracting surface all lie in the same plane are known as **laws of refraction**. The process of refraction is shown in Fig. 5.5.

5.7 TOTAL INTERNAL REFLECTION

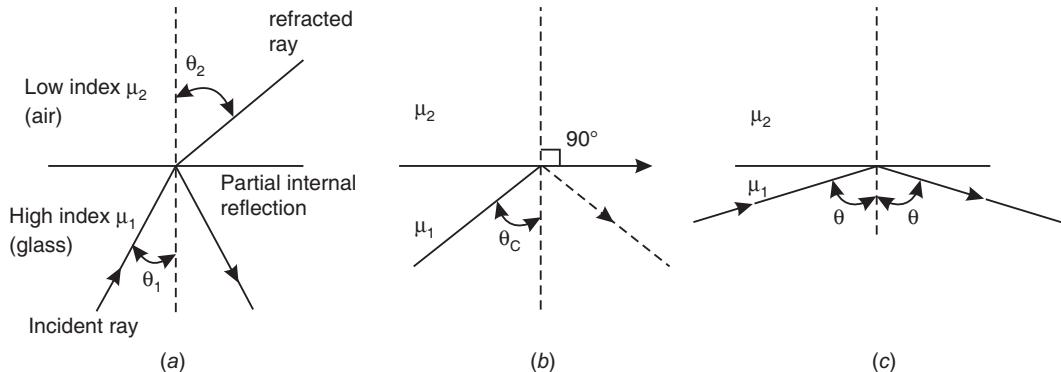


Fig. 5.6: Phenomenon of total internal reflection - (a) when light rays travel from a denser medium into a rarer medium, they bend away from the normal in the rarer medium. As the angle of incidence θ_1 increases, the angle of refraction θ_2 increases. (b) The angle of incidence θ_c which produces an angle of refraction 90° is called the critical angle. (c) For larger angles of incidence, the ray does not emerge out from denser medium but gets internally reflected.

A medium having a lower refractive index is said to be an optically **rarer medium** while a medium having a higher refractive index is known as an optically **denser medium**. When a ray of light passes from a denser medium to a rarer medium, it is bent away from the normal in the rarer medium (see Fig. 5.6a). Snell's law for this case may be written as

$$\sin \theta_2 = \left(\frac{\mu_1}{\mu_2} \right) \sin \theta_1 \quad (5.8)$$

where θ_1 is the angle of incidence of light ray in the denser medium and θ_2 is the angle of refraction in the rarer medium. Also $\mu_1 > \mu_2$. When the angle of incidence, θ_1 in the denser medium is increased, the transmission angle, θ_2 increases and the refracted rays bend more and more away from the normal. At some particular angle θ_c the refracted ray glides along the boundary surface so that $\theta_2 = 90^\circ$, as seen in Fig. 5.6(b). At angles greater than θ_c there are no refracted rays at all. The rays are reflected back into the denser medium as though they encountered a specular reflecting surface (Fig. 5.6c). Thus,

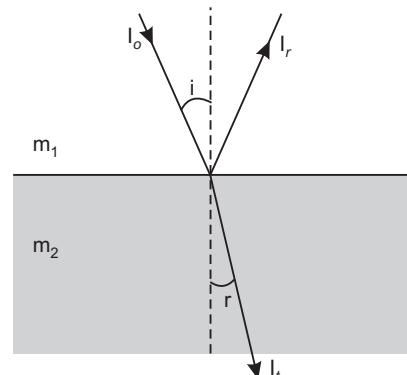


Fig. 5.5: Phenomenon of refraction
– A ray obliquely incident on air-glass interface bends toward the normal in glass.

- If $\theta_1 < \theta_c$, the ray refracts into the rarer medium
- If $\theta_1 = \theta_c$, the ray just grazes the interface of rarer-to-denser media
- If $\theta_1 > \theta_c$, the ray is reflected back into the denser medium.

The phenomenon in which light is totally reflected from a denser-to-rarer medium boundary is known as **total internal reflection**. The rays that experience total internal reflection obey the laws of reflection. Therefore, the critical angle can be determined from Snell's law.

When $\theta_1 = \theta_c$, $\theta_2 = 90^\circ$.

Therefore, from equ. (5.8), we get

$$\begin{aligned} \mu_1 \sin \theta_c &= \mu_2 \sin 90^\circ = \mu_2 \\ \therefore \quad \sin \theta_c &= \frac{\mu_2}{\mu_1} \end{aligned} \quad (5.9)$$

When the rarer medium is air, $\mu_2 = 1$ and writing $\mu_1 = \mu$, we obtain

$$\sin \theta_c = \frac{1}{\mu} \quad (5.10)$$

Total internal reflection does not take place when light propagates from a rarer to a denser medium. The critical angle is small for substances having high refractive index. For example, $\theta_c = 24^\circ$ for diamond – air interface, while it is about 42° for glass-to-air interface. This phenomenon is exploited in obtaining sparkle in crystal glass and diamonds. The phenomenon of total internal reflection has made possible to guide light through optical fibres.

Example 5.1: Determine the critical angle for a light ray traveling from water ($\mu = 1.333$) to air.

Solution: $\sin \theta_c = \frac{1}{\mu} = \frac{1}{1.333} = 0.75 \quad \therefore \quad \theta_c = \sin^{-1}(0.75) = 48.61^\circ$

5.8 REFLECTIVITY AND TRANSMISSIVITY

The laws of reflection and refraction only describe the relations between the directions of incident, reflected and refracted rays, but do not speak any thing about the intensities of the rays. If an energy E is incident on a transparent medium, it is divided into two parts—one part E_r goes with the reflected ray and the other E_t with the refracted ray. Following the law of conservation of energy,

$$E_i = E_r + E_t \quad (5.11)$$

The quantity characterizing the reflectivity of a surface is called the **reflection coefficient** ρ . It is the ratio of the reflected electric energy to the incident energy.

$$\rho = \frac{E_r}{E_i} \quad (5.12)$$

The reflection coefficient depends on the surface, the wavelength composition in the incident light, the angle of incidence and many other factors. However, if all the other factors remain constant, an increase in the angle of incidence i causes an increase in the reflection coefficient. For normal incidence it is shown that

$$\rho = \left[\frac{\mu_1 - \mu_2}{\mu_1 + \mu_2} \right] \quad (5.13)$$

It can be seen from equation (5.13), that the reflection coefficient ρ is positive when $\mu_2 < \mu_1$. It implies that the oscillation in the incident and reflected waves occur in the same phase.

Therefore, there is no change of phase of the wave upon reflection. On the other hand, if $\mu_2 > \mu_1$, ρ is negative signifying that the oscillation in the incident and reflected waves are 180° out of phase. Thus the phase of the reflected wave changes by π rad.

The **reflectivity** r of a surface is the ratio of the reflected beam intensity to the incident beam intensity. Thus

$$r = \frac{I_r}{I_i} = \left| \frac{E_r}{E_i} \right|^2 = |\rho|^2$$

$$\therefore r = \left[\frac{\mu_1 - \mu_2}{\mu_1 + \mu_2} \right]^2 \quad (5.14)$$

The **transmissivity** t is given by

$$t = \frac{4\mu_1\mu_2}{(\mu_1 + \mu_2)^2} \quad (5.15)$$

Because energy is conserved, $r + t = 1$. (5.16)

Example 5.2: A solar cell is made of silicon having a refractive index 3.5. If it is exposed to sunlight, what is the percentage of light that is reflected back by it.

Solution: $r = \left[\frac{\mu_1 - \mu_2}{\mu_1 + \mu_2} \right]^2 = \left[\frac{1 - 3.5}{1 + 3.5} \right]^2 = 0.31 \quad \therefore \text{Percentage of reflected light} = 31\%.$

5.9 ABSORPTION

The light propagation through an optical medium is attended mainly by two effects. First, the velocity of light decreases in the medium. Secondly, the intensity of light goes on decreasing with distance in the medium. It happens even in the so called transparent media, since no material is perfectly transparent. The reduction in intensity of light with increasing length of propagation in a medium is called **absorption or attenuation**.

When a beam of light passes through a thin transparent material of thickness dx (Fig. 5.7), the decrease in its intensity dI is found to be proportional to the initial intensity I and to the thickness dx .

$$dI \propto I dx$$

or

$$dI = -\alpha I dx \quad (5.17)$$

where α is the constant of proportionality and is called the **linear absorption coefficient** of the medium. It is characteristic of the medium and also is a function of wavelength. Rearranging the terms in (5.17), we get

$$\frac{dI}{I} = -\alpha dx \quad (5.18)$$

The total loss in light intensity after passage through a slab of finite thickness x is obtained by integrating the above equation (5.18). Thus, if I_0 is the intensity at $x = 0$, then

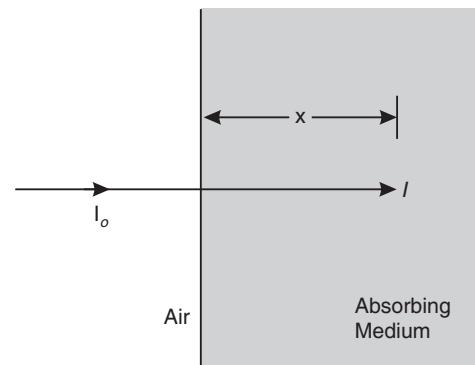


Fig. 5.7: Absorption. The intensity of light at a distance x in a medium reduces to I from I_0 as a result of absorption.

$$\begin{aligned} \int_{I_0}^I \frac{dI}{I} &= -\alpha \int_0^x dx \\ \ln \frac{I}{I_0} &= -\alpha x \\ \text{or } I(x) &= I_0 e^{-\alpha x} \end{aligned} \quad (5.19)$$

The relation (5.19) shows that if a beam of light propagates in a medium, its intensity decreases exponentially with distance. α is a measure of loss of light from the direct beam. This exponential **law of absorption** was discovered independently by Pierre Bouguer (1698–1758), the French oceanographer and by J.H.Lambert (1728–1777), the German physicist. The law is, therefore, called **Lambert-Bouguer law**.

5.10 WAVE FRONT AND THE RAY

According to Huygens theory, light is considered to propagate in the form of waves. *A wave is any disturbance, from an equilibrium condition, that propagates with time from one region of space to another.*

We are all familiar with what happens when a pebble is dropped into the still water of a well or a pond. The pebble generates ripples which expand in the form of circles. If the motion of a floating object such as a leaf is watched, we observe that the leaf moves up and down but does not ravel outward with the wave. We, therefore, conclude that it is the energy supplied by the agent of disturbance which moves forward as successive waves. We identify the wave motion with the help of crests and troughs traveling away from the centre of disturbance. It is apparent from the motion of the floating leaf that every particle of the medium (water) oscillates about its mean position at right angles to the direction of wave propagation. Such wave motion is known as *transverse wave* motion. Water waves and light waves are examples of transverse waves.

Waves start from a source and spread out into new and new regions of space. In case of waves on a water surface, the ripples start from the point of disturbance and expand in the form of circles. The circles are in fact the crests of the waves. All the particles located at the crest will be in the same phase. Therefore, a ripple is the locus of points having the *same phase*. The propagation of ripples is carried out by circles formed one after another at a distance of λ from each other, as shown in Fig. 5.8 (a).

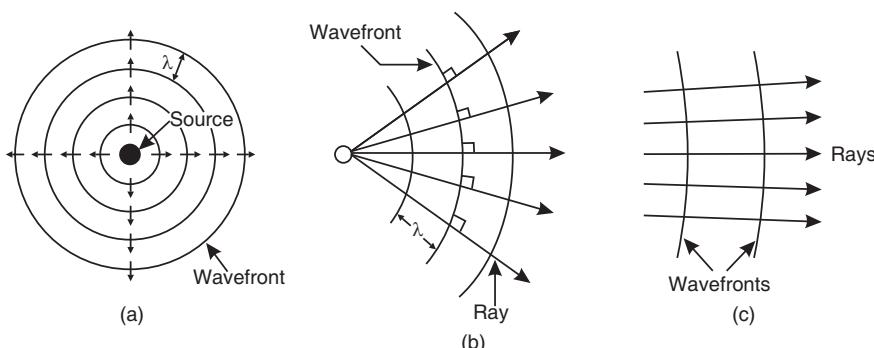


Fig. 5.8: Wave front and rays (a) a point source produces spherical waves (b) the ray direction is perpendicular to the wave front, (c) at large distance from the source, the wave fronts tend to be parallel planes.

In a similar manner, the propagation of a three dimensional wave may be visualized as being carried out by innumerable wave surfaces formed one after another at a distance of λ from each other. A **wave surface** is a spherical surface over which the phase of the wave is constant. Generally, the wave surfaces are drawn as passing through either the crest or trough of a sine wave. Therefore, the wave surfaces are separated by one wavelength. The furthermost wave surface from the source is called **wave front**. The wave front is dynamic and separates that region of space into which the wave is already spread from the region into which the wave is yet to enter. The propagation of the wave is visualized by the advancing wave front and stationary wave surfaces behind it. Usually, all the wave surfaces are referred to as wave fronts.

The shape of the wave front is determined by the source generating the waves and by the size of the object which the waves encounter. If the source is a point source, it radiates waves in all directions uniformly and the wave fronts will be a family of concentric spheres. On the other hand, a linear source produces a family of coaxial cylindrical wave fronts (Fig. 5.9). At distances far from the source, both the spherical and cylindrical wave fronts may be treated as plane wave fronts. When the size of the slit, through which a plane wave passes, is much larger than the wavelength, the wave front remains plane. On the other hand, if the slit is of the order of wavelength, the wave front becomes cylindrical.

It is convenient sometimes to describe the wave propagation in terms of rays instead of wave fronts. Rays are lines drawn normal to the wave fronts and represents the direction of energy flow. Rays are parallel lines in case of plane waves and radial lines in case of spherical waves. They are illustrated in Fig. 5.8.

5.11 MATHEMATICAL REPRESENTATION OF A PLANE WAVE

As a wave travels forward in a medium, a sequence of particles is set into identical oscillations in succession. The oscillations are communicated from particle to particle and they will be in different states of oscillation at different times. Therefore, the displacement of a particle is a function of *space coordinates* as well a function of *time* and it may be described by the function

$$y = f(x, t) \quad (5.20)$$

The displacement y is called the **wave function**. The function f may be any function. For the sake of simplicity, we assume the function to be simple harmonic. We assume the simple harmonic waveform because an understanding of sinusoidal waves helps us understanding waves of any shape. All kinds of waveforms including pulses can be constructed by adding up sinusoidal waves of appropriately selected wavelengths and amplitudes.

We can describe the displacement of any point on a harmonic wave in terms of both space and time as

$$y = A \sin \left[\frac{2\pi}{\lambda} (x - vt) \right]$$

But $v = \nu\lambda$ and $v = 1/T$.

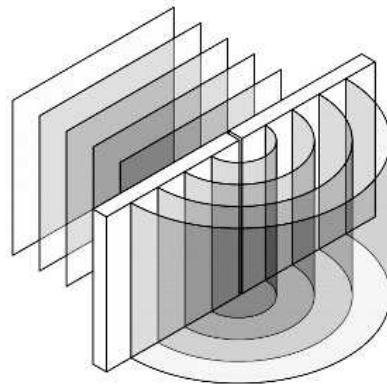


Fig. 5.9: A linear source produces a family of coaxial cylindrical wave fronts

Hence,

$$y = A \sin\left(\frac{2\pi x}{\lambda} - \frac{2\pi t}{T}\right)$$

or

$$y = A \sin(kx - \omega t) \quad (5.21)$$

where $k = \frac{2\pi}{\lambda}$. The quantity k is called the **wave number** or **propagation number**. It gives the number of wavelengths contained in a distance equal to 2π metres. In equ. (5.21) there are no limits imposed on the variables x and t . Mathematically, the equation describes a wave of infinite length, existing for all times, from the remote past to the distant future. Using slightly different arguments, the wave equation may be shown to have the form

$$y = A \sin(\omega t - kx) \quad (5.22)$$

Both the equations (5.21) and (5.22) represent basically the same progressive harmonic wave traveling along the positive x -direction. A wave traveling in the negative x -direction merely reverses the direction of velocity, i.e., v changes to $-v$. Therefore, the equation for a backward traveling wave is given by

$$y = A \sin(kx + \omega t) \quad (5.23)$$

In the above discussion, the wave equations are expressed in terms of the sine function. The wave can be represented equally well by a cosine function. The change merely implies a different choice of initial position and time. Thus,

$$y = A \cos(kx - \omega t) \quad (5.24)$$

also represents a forward moving harmonic wave.

If we regard k as a vector quantity, having a magnitude equal to the wave number k and lying in a direction parallel to the positive direction of x -axis, then

$$\mathbf{i} k = \mathbf{k} \quad \text{and} \quad kx = \mathbf{k} \cdot \mathbf{r}.$$

Therefore, we may write equ. (5.21) as

$$y = A \sin(\mathbf{k} \cdot \mathbf{r} - \omega t) \quad (5.25)$$

The above equation is independent of the system of coordinates. \mathbf{k} is called the **wave vector**.

5.11.1 Complex Representation of a Wave

It is seen that a progressive harmonic wave may be represented by either of the following forms.

$$y = A \sin(\mathbf{k} \cdot \mathbf{r} - \omega t)$$

or

$$y = A \cos(\mathbf{k} \cdot \mathbf{r} - \omega t)$$

A linear combination of the above two equations also represents a harmonic wave.

Using Euler's formula $e^{i\theta} = \cos \theta + i \sin \theta$, we can write

$$y = A [\cos(\mathbf{k} \cdot \mathbf{r} - \omega t) + i \sin(\mathbf{k} \cdot \mathbf{r} - \omega t)]$$

or

$$y = Ae^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \quad (5.26)$$

Equ. (5.26) gives the complex representation of a harmonic wave.

5.11.2 Intensity

The intensity of a wave is equal to the energy transferred on an average by the wave in a unit time through a unit area of a surface perpendicular to the direction of propagation of the wave. It is directly proportional to the square of the amplitude of the wave. Thus,

$$I \propto |A|^2 \quad (5.27)$$

5.11.3 The General Wave Equation

A wave equation is an expression that shows the variation of displacement of a particle as a function of the spatial coordinates and time. The space derivative of the equ. (5.21) is given by

$$\frac{\partial y}{\partial x} = kA \cos(kx - \omega t) \quad (5.28)$$

Similarly, the time derivative of the equ. (5.21) is

$$\frac{\partial y}{\partial t} = -\omega A \cos(kx - \omega t) \quad (5.29)$$

Combining equ. (5.28) and (5.29), we get

$$\frac{\partial y}{\partial x} = -\frac{1}{v} \frac{\partial y}{\partial t} \quad (5.30)$$

Equ. (5.30) represents differential equation for the wave moving to the right. Taking the second derivatives we can make the equation independent of the direction of x . Thus,

$$\frac{\partial^2 y}{\partial x^2} = k^2 A \sin(kx - \omega t) = k^2 y$$

and

$$\frac{\partial^2 y}{\partial t^2} = \omega^2 A \sin(kx - \omega t) = \omega^2 y$$

∴

$$\frac{\partial^2 y}{\partial x^2} = \frac{k^2}{\omega^2} \frac{\partial^2 y}{\partial t^2}$$

or

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2} \quad (5.31)$$

Equ. (5.31) represents any wave shape, propagating towards right or left with a velocity v . It is a one-dimensional wave equation. In three dimensions, we can write it as

$$\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2} + \frac{\partial^2 \xi}{\partial z^2} = \frac{1}{v} \frac{\partial^2 \xi}{\partial t^2}$$

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \xi = \frac{1}{v^2} \frac{\partial^2 \xi}{\partial t^2}$$

$$\nabla^2 \xi = \frac{1}{v^2} \frac{\partial^2 \xi}{\partial t^2} \quad (5.32)$$

where $\nabla^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right)$ is known as the Laplacian operator.

Equ. (5.32) is known as the *general wave equation* and it represents a three-dimensional traveling wave.

5.12 LIGHT IS AN ELECTROMAGNETIC WAVE

In contrast to the water waves etc, light waves do not require a material medium for their propagation. For example, light reaches to the earth from the sun travelling a long way in a vacuum. Maxwell showed that *light is a self-sustaining electromagnetic wave characterized by frequency and polarization*. A light wave consists of electric and magnetic fields, oscillating in mutually perpendicular planes and which are also perpendicular to the direction of propagation of the wave (See Fig. 5.10).

Maxwell's equations for free space can be manipulated into the form of two concise vector equations. Assuming the direction of propagation to be z -axis, the **wave equation** can be written as

$$\nabla^2 \mathbf{E} = \epsilon_0 \mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \quad (5.33)$$

and

$$\nabla^2 \mathbf{B} = \epsilon_0 \mu_0 \frac{\partial^2 \mathbf{B}}{\partial t^2} \quad (5.34)$$

These equations are identical to the general *wave equation* (5.32) provided that

$$v = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \quad (5.35)$$

Using the values of ϵ_0 and μ_0 into the above equation, it is found that $v \approx 3 \times 10^8$ m/s. Hence

$$v = c \quad \text{and} \quad c = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \quad (5.36)$$

The simplest solution for equ. (5.33) is

$$E_y = E_0 \sin(kx - \omega t) \quad (5.37)$$

Similarly, the solution for equ. (5.34) is

$$B_x = B_0 \sin(kx - \omega t) \quad (5.38)$$

E_y and B_x are the **wave functions**, which describe the size of the electric and magnetic disturbances respectively at any time.

The wave represented by equ. (5.37) has a constant amplitude E_0 . Therefore, it is a *plane wave*. The term *plane* means that the field vector \mathbf{E} (similarly \mathbf{B}) lies in a plane at each point in space, and the planes at any two different points are parallel to each other, as shown in Fig. 5.11.

It was shown that the amplitudes E_0 and B_0 are not independent but are related through

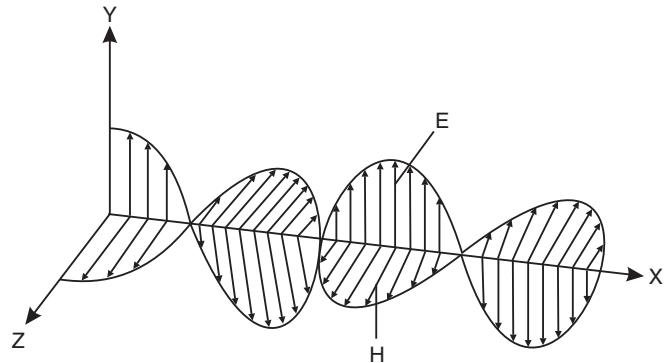


Fig. 5.10: An electromagnetic wave

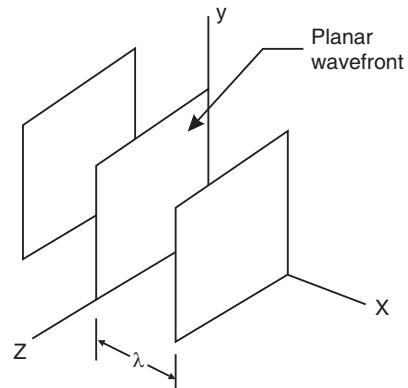


Fig. 5.11: Schematic representation of planar wave fronts

$$E_0 = cB_0 \quad \therefore \quad B_0 = \frac{1}{c}E_0 \quad (5.39)$$

The above relation between the amplitudes means that the instantaneous values are also related through

$$E = cB \quad \therefore \quad B = \frac{1}{c}E \quad (5.40)$$

Equ. (5.40) indicates that E and B fields are in phase, reaching their zero and maximum values at the same time.

The electric field vector \mathbf{E} of the light wave has greater influence on matter than the magnetic field vector \mathbf{B} . When a light wave interacts with matter, the electric and magnetic fields of light wave act upon the electrons in the material. The electric force F_E experienced by an electron due to the electric field of the light wave is given by

$$F_E = eE \quad (5.41)$$

The maximum magnetic force that can be experienced by an electron is given by

$$F_B = evB \quad (5.42)$$

where v is the velocity of electron. Using the relation (5.40) into (5.42), we find that

$$F_B = (v/c)eE = (v/c)F_E \quad (5.43)$$

As v is always far smaller than c , F_B is insignificant in comparison to F_E . We could describe optical phenomena equally well in terms of either the electric or magnetic field components, but not equally simply. We generally deal with the electric field component since it is much easier to picture. Electrons move in the direction of the electric field. In contrast, electrons move perpendicular to magnetic fields, making the description way more complicated. Thus, the electric field of the light wave has greater effect than the magnetic field of the wave. Further, it was established that the electric field of the light wave plays a major role in the physiological, photochemical and other actions of light radiation. Therefore, in the pictorial representation of the light wave, the vibrations of an electromagnetic wave are shown only by the direction of the electric field, \mathbf{E} (Fig. 5.12) and the magnetic field vector \mathbf{B} is generally omitted. The magnetic field is still present, perpendicular to the electric field, but will not be shown in the diagrams, in order to avoid confusion.

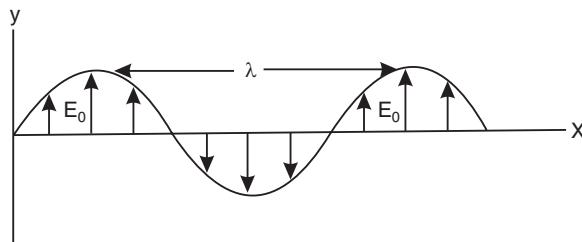


Fig. 5.12: A light wave is represented by electric field, \mathbf{E} , variations in space

5.12.1 Energy Density

One of the most significant properties of the electromagnetic wave is that it transports energy and momentum. Both the electric field and magnetic fields of the wave store energy. The *energy density* (energy per unit volume), u_E associated with the \mathbf{E} -field in free space is given by

$$u_E = \frac{1}{2}\epsilon_0 E^2 \quad (5.44)$$

Similarly, the energy density u_B associated with the \mathbf{B} -field in free space is given by

$$u_B = \frac{1}{2\mu_0} B^2 \quad (5.45)$$

Using the relation (5.36) into (5.45), we get

$$u_B = \frac{1}{2\mu_0} \left[\frac{E}{c} \right]^2 = \left[\frac{\epsilon_0 \mu_0}{2\mu_0} \right] E^2 = \frac{1}{2} \epsilon_0 E^2$$

That is

$$u_E = u_B$$

It means that the energy of the wave is shared equally between the electric and magnetic fields of the wave. The total energy density of the electromagnetic wave is

$$u = u_E + u_B = \epsilon_0 E^2 \quad (5.46)$$

5.12.2 Poynting Vector

An electromagnetic wave transports energy from one region of space to another. We can describe the energy transfer in terms of the rate of energy flow per unit area or power per unit area. It is denoted by the vector \mathbf{S} known as the **Poynting vector**.

Let us consider a plane wave front moving to the right (Fig. 5.13). In a time dt , the wave front moves a distance $dx = cdt$.

Therefore, the energy dW passing during time dt through a *unit area* perpendicular to the direction of the propagation of the wave is

$$dW = uc dt$$

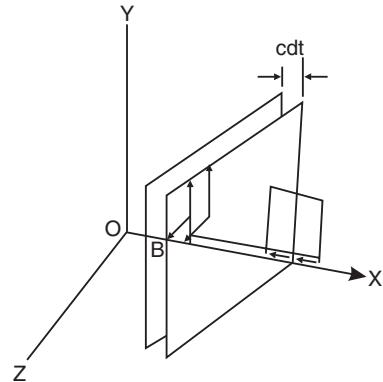


Fig. 5.13

where u is the energy density. The rate at which the energy is transported by the wave is its power. Therefore, the power transferred per unit area, S , is

$$S = \frac{\text{energy}}{dt} = uc$$

Making use of the expressions (5.47) and (5.48) into the above equation, we write

$$S = \sqrt{u} \sqrt{u} c = \left(\sqrt{\epsilon_0} E \right) \left(\frac{B}{\sqrt{\mu_0}} \right) = \frac{\epsilon_0}{\sqrt{\epsilon_0 \mu_0}} EBc = \epsilon_0 c^2 EB$$

When S is assigned the direction of propagation, it is called the **Poynting vector**. The direction of \mathbf{S} is the same as that of the cross product of \mathbf{E} and \mathbf{B} . The vectors \mathbf{E} and \mathbf{B} are mutually perpendicular and form a right-handed system with the direction of propagation of the wave. Therefore,

$$\mathbf{S} = c^2 \epsilon_0 (\mathbf{E} \times \mathbf{B}) \quad (5.47)$$

or

$$\mathbf{S} = \frac{1}{\mu_0} (\mathbf{E} \times \mathbf{B}) \quad (5.48)$$

5.12.3 Intensity

A plane electromagnetic wave traveling in the positive x-direction is described by the equ. (5.37) and (5.38).

$$\therefore S = \frac{1}{\mu} E_0 B_0 \sin^2(kx - \omega t) \quad (5.49)$$

$$S = c^2 \epsilon_0 E_0 B_0 \sin^2(kx - \omega t) \quad (5.50)$$

The electromagnetic power P that the plane wave delivers to a receiver of area 'A' oriented perpendicular to x -axis is given by

$$P = SA = \frac{1}{\mu_0} E_0 B_0 A \sin^2(kx - \omega t)$$

or

$$P = \frac{1}{\mu_0 c} E_0^2 A \sin^2(kx - \omega t) \quad (5.51)$$

If T is the time of observation, then the average power is given by

$$\begin{aligned} P_{av} &= \frac{1}{T} \int_0^T P dt \\ &= \frac{1}{\mu_0 c} \cdot E_0^2 A \cdot \frac{1}{T} \int_0^T [\sin^2(kx - \omega t)] dt \\ &= \frac{1}{2\mu_0 c} E_0^2 A \end{aligned} \quad (5.52)$$

or

$$P_{ave} = \frac{c\epsilon_0}{2} E_0^2 A \quad (5.53)$$

The **intensity** of the electromagnetic wave is defined as the average power per unit area. It is also called the **irradiance**.

∴ Intensity of the wave is

$$I = \frac{P_{ave}}{A} = \frac{1}{2\mu_0 c} E_0^2 \quad (5.54)$$

or

$$I = \frac{c\epsilon_0}{2} E_0^2 \quad (5.55)$$

Thus, it is found that

$$I \propto |E_0|^2 \quad (5.56)$$

5.12.4 Wave Characteristics of Light

- Equation (5.37) represents a plane wave propagating in the direction of $k = 2\pi / \lambda$ where λ is the *spatial period* of the wave and is known as the **wavelength**.
- The sine varies from +1 to -1 so that the maximum value of E_y is E_0 . Therefore, electric field vector E_y of the wave oscillates parallel to the y -axis with values between $+E_0$ and $-E_0$. The maximum value E_0 is the **amplitude** of the wave.
- The wave described by (5.37) has only one frequency, v ($\omega = 2\pi v$) and hence it is a **monochromatic wave**.
- The wave is a simple harmonic wave extending from $-\infty$ to $+\infty$ and has no beginning and an end.
- It is a plane wave, the wave front being normal to x -axis.
- The electric vector E of the wave always oscillates parallel to a fixed direction in space, i.e. y direction. In other words, the E vibrations are confined to xy plane. A wave having its vibrations confined to a single plane is called a **plane polarized** or a **linearly polarized wave**. The wave depicted in Fig. 5.9 is an ideally plane polarized wave.
- The wave has zero vergence and does not get attenuated as it propagates through space.
- The variation of phase along the wave extension is predictable. Therefore, if two or more such waves travel along the same direction in space, their phase variations are synchronized and the phase difference stays constant. Thus the waves will be highly coherent.

We will soon learn that actual light waves are far from ideal. It will be noted that they are wave trains of limited extension, have a certain spread in frequency around a central value, totally unpolarized and incoherent. All the natural light sources such as the sun, a lamp or a flame emit only such wave trains.

Example 5.3: The wave function for a light wave is

$$E(z, t) = 10^3 \sin \pi (3 \times 10^6 z - 9 \times 10^{14} t)$$

(i) Determine the speed, wavelength, frequency and period of the wave.

(ii) Determine the magnetic field associated with the wave.

Solution: (i)

$$\begin{aligned} E(z, t) &= 10^3 \sin \pi (3 \times 10^6 z - 9 \times 10^{14} t) \\ &= 10^3 \sin 3 \times 10^6 \pi (z - 3 \times 10^8 t) \end{aligned} \quad (a)$$

This equation is similar to the general wave equation

$$E(z, t) = E_0 \sin k (z - v t) \quad (b)$$

where E_0 is the amplitude of the wave, k is the wave number $k = \frac{2\pi}{\lambda}$ and v is the velocity of the wave.

Comparing (a) with (b), we get $v = 3 \times 10^8 \text{ m/s}$, $k = 3 \times 10^6 \pi/\text{m}$

$$\therefore \lambda = \frac{2\pi}{k} = \frac{2\pi}{3 \times 10^6 \pi/\text{m}} = 6.666 \times 10^{-7} \text{ m} = 6666 \text{ \AA}$$

$$v = \frac{\lambda}{T} = \frac{3 \times 10^8 \text{ m/s}}{6.666 \times 10^{-7} \text{ m}} = 4.5 \times 10^{14} \text{ Hz}$$

$$T = \frac{1}{v} = \frac{1}{4.5 \times 10^{14} \text{ Hz}} = 2.2 \times 10^{-15} \text{ s.}$$

(ii) The light wave propagates in z -direction while the E-vector oscillates along x -direction in the xz -plane. Since in an EM wave, magnetic field B is normal to both E and wave propagation directions, it should be in the yz -plane.

Thus,

$$B_x = 0; \quad B_z = 0$$

$$B = B_y (z, t) = B(z, t)$$

As,

$$E = cB, \quad B = \frac{E}{c}$$

$$\therefore B(z, t) = \frac{10^3 \sin \pi (3 \times 10^6 z - 9 \times 10^{14} t)}{c}$$

$$B(z, t) = 3.33 \times 10^{-6} \sin \pi (3 \times 10^6 z - 9 \times 10^{14} t)$$

Example 5.4: A plane electromagnetic wave moving through free space has an E field given by,

$$E = 100 \sin \left[8\pi \times 10^{14} \left(t - \frac{z}{3 \times 10^8} \right) \right]$$

Calculate the corresponding intensity.

$$\text{Solution:} \quad E = 100 \sin \left[8\pi \times 10^{14} \left(t - \frac{z}{3 \times 10^8} \right) \right] \quad \therefore E_0 = 100 \text{ V/m}$$

$$\text{Intensity, } I = \left(\frac{c \epsilon_0}{2} \right) E_0^2 = \frac{(3 \times 10^8 \text{ m/s}) \left(8.85 \times 10^{-12} \frac{\text{C}^2}{\text{N.m}^2} \right) \left(100 \frac{\text{V}}{\text{m}} \right)^2}{2}$$

$$= 13.3 \frac{C^2}{s.N.m} \cdot \frac{V^2}{m^2}$$

$$I = 13.3 \text{ W/m}^2$$

$$\text{Units: } 1 \frac{C^2 V^2}{s.N.m.m^2} = 1 \frac{J^2}{s.N.m.m^2} = 1 \frac{J.N.m}{s.N.m.m^2} = 1 \frac{J}{s.m^2} = 1 \frac{W}{m^2}$$

5.13 VISIBLE RANGE

The arrangement of the various electromagnetic waves in a continuous sequence of frequencies and wavelengths, as in Fig. 5.14, is called the **electromagnetic spectrum**. The spectrum includes waves covering a broad range of wavelengths. It is bounded at one end by gigantic radio waves having wavelengths of a few hundred kilometers and by γ -rays having wavelengths of 10^{-12} m at the other end. **Visible range** is that part of the spectrum constituted by waves which can be detected by the human eye. It extends from the deepest violet to the deepest red. The limiting range of these waves depends on the individual properties of the eye and varies approximately in the interval $\lambda = 4000 \text{ \AA}$ to $\lambda = 7800 \text{ \AA}$. The regions flanking the visible range are known as **infrared** on the longer wavelength side and **ultraviolet** on the shorter wavelength side. The infrared (IR) region lies in the wavelength range $7.8 \times 10^{-7} \text{ m}$ to 10^{-3} m and the ultraviolet (UV) region lies between 4000 \AA to 10 \AA . Radiation in these three regions namely IR, visible and UV together is called **optical radiation**.

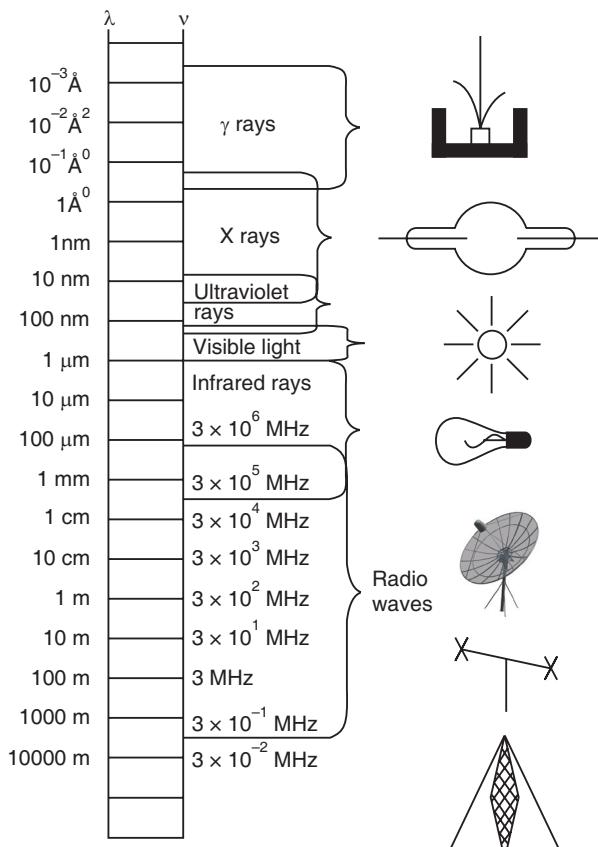


Fig. 5.14: Electromagnetic spectrum

Table 2 lists the wavelengths and the frequencies of each of the coloured region in the visible range.

Table 2: Wavelengths and frequencies of different colours

Colour	Vacuum Wavelength (\AA)	Frequency (Hertz)
Red	7800 – 6220	$3.84 \times 10^{14} – 4.82 \times 10^{14}$
Orange	6220 – 5970	$4.82 \times 10^{14} – 5.03 \times 10^{14}$
Yellow	5970 – 5770	$5.03 \times 10^{14} – 5.20 \times 10^{14}$
Green	5770 – 4920	$3.84 \times 10^{14} – 6.10 \times 10^{14}$
Blue	4920 – 4550	$3.84 \times 10^{14} – 6.59 \times 10^{14}$
Violet	4550 – 3990	$3.84 \times 10^{14} – 7.69 \times 10^{14}$

5.14 OPTICAL PATH LENGTH

The shortest distance, L between two points A and B is called the *geometric path*. The length of geometric path is independent of the medium that surrounds the path AB. When a light ray travels from the point A to point B, it travels with the velocity ‘ c ’ if the medium is air and with a lesser velocity if the medium is other than air. Therefore the light ray takes more time to go from A to B located in a medium.

From equation (5.3)

$$\mu = \frac{c}{v} = \frac{AB/t}{AB/T} = \frac{T}{t}$$

where t and T are the time taken by the light ray in air and in a medium respectively.

$$\therefore T = \mu t$$

The above relation means that a light ray takes μ times more time to cover the distance AB in a medium. To take into account the delay, we use another distance called the *optical path length*. If a ray of light travels a distance L in a medium of refractive index μ in a certain interval of time, then it would travel a greater distance Δ in air during the same interval of time. Therefore,

$$\frac{\Delta}{L} = \frac{ct}{vt} = \mu$$

or

$$\Delta = \mu L \quad (5.57)$$

i.e., Optical path length = (Refractive index) \times (Geometric path length)

Thus, *the optical path length is defined as the product of refractive index and the geometric path length*.

Note that if a ray travels a distance l in a medium of refractive index μ , the optical path length is equal to μl . In a given time light travels the same optical path length in different media. Suppose light travels a distance l_1 in a medium of refractive index μ_1 and a distance l_2 in a medium of refractive index μ_2 in time t . Then

$$\mu_1 l_1 = \mu_2 l_2$$

Example 5.5: Light of wavelength 6200 \AA travels through a film of water $0.72 \times 10^{-6} \text{ m}$ thick. If its refractive index is 1.333, find the optical path.

Solution: Optical path = (Refractive index) (geometric path)

$$\therefore \Delta = \mu L = (1.333)(0.72 \times 10^{-6} \text{ m}) = 0.96 \times 10^{-6} \text{ m} = 0.96 \mu\text{m}$$

5.15 PHASE CHANGE AND PATH DIFFERENCE

When any wave advances in space, its phase changes continuously. At a fixed time, the points at x_1 and x_2 on the wave differ in phase by an amount

$$\delta = \phi_2 - \phi_1$$

Taking help of equ. (5.21), we can write the above relation as

$$\delta = (kx_2 - \omega t) - (kx_1 - \omega t) = k(x_2 - x_1) = \frac{2\pi}{\lambda} (x_2 - x_1)$$

or

$$\delta = \frac{2\pi}{\lambda} L \quad (5.58)$$

where $L = (x_2 - x_1)$ is the spatial separation of the points, i.e. the *geometric path*. It follows from (5.58) that $\delta = 2\pi$ if $L = \lambda$, $\delta = \pi$ if $L = \lambda/2$ and so on. The waveform repeats over a distance of one λ . Therefore, the phase difference 2π is equivalent to zero phase difference. It implies that a displacement of wave by one complete wavelength leaves the waveform unchanged.

- (i) Path differences between two waves may arise if two waves propagate in a medium along two different paths and overlap at some point P (Fig. 5.15). The result of their superposition at P depends on their relative disposition with respect to each other. The relationship between the waves is defined with the help of their phase difference which may be expressed in terms of the path difference. The quantity $L = (x_2 - x_1)$ in (5.58) also represents the path difference when two waves are considered. The phase difference can then be computed using (5.58) as

$$\delta = \frac{2\pi}{\lambda} \mu L$$

or

$$\delta = \frac{2\pi}{\lambda} \Delta \quad (5.59)$$

- (ii) A path difference or phase difference between two ways may arise during the process of reflection, also. A light wave traveling in a rarer medium undergoes a phase change of π rad when it gets reflected at the boundary of denser medium as illustrated in Fig. 5.16 (a). However, a light wave does not suffer a change in phase when it gets reflected at denser – to – rarer medium boundary. For example, let us consider a light wave incident on a glass block. The wave reflected from the air-to-glass boundary experiences a phase change of π rad relative to the incident wave because $\mu_g > \mu_a$, as illustrated in Fig. 5.16 (b). The wave

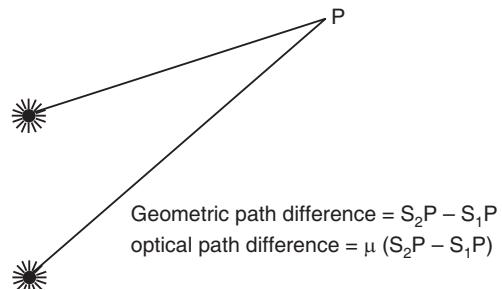


Fig. 5.15

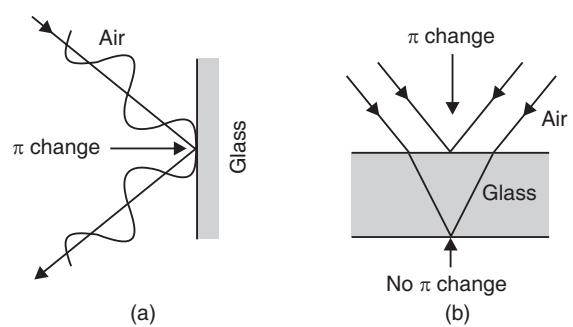


Fig. 5.16

transmitted through glass gets again reflected at the bottom surface which is glass-to-air boundary. The reflection at glass – air-boundary does not produce a phase change. A phase change of π rad is equivalent to a path difference of $\lambda/2$ as seen from the relation (5.58). Thus, a light wave gets inverted or loses a half-wave on reflection from a rarer-denser boundary.

Example 5.6: A light wave propagates from a point A to another point B. A glass plate ($\mu_g = 1.5$) of 1 mm thick is introduced in its path. If the wavelength of light is 5000 \AA , What is the change of phase of the wave at B?

Solution: When the light travels from A to B, a distance 'l' (say) in air, the optical path

$$\Delta_l = \mu_a l = l$$

When a glass plate of thickness 't' is introduced in the path AB, light travels through a length 't' in the glass and through a length $(l - t)$ in air. Therefore the optical path is now

$$\begin{aligned}\Delta_2 &= \mu_a(l - t) + \mu_g t \\ &= (l - t) + \mu_g t = l + (\mu_g - 1)t\end{aligned}$$

Additional path due to glass plate is

$$\Delta = \Delta_2 - \Delta_1 = l + (\mu_g - 1)t - l = (\mu_g - 1)t$$

∴ The phase difference

$$\begin{aligned}\delta &= \frac{2\pi}{\lambda} \Delta = \frac{2\pi}{\lambda} (\mu_g - 1)t \\ &= \frac{2\pi(1.5 - 1) \times 10^{-3} \text{ m}}{5 \times 10^{-7} \text{ m}} = 2\pi \times 10^3 \text{ radians}\end{aligned}$$

Phase of $1000 \times 2\pi$ radians indicates that waves arriving at B are in phase.

5.16 THE PRINCIPLE OF SUPERPOSITION

When two pebbles are dropped at different points in a pond, the expanding ripples pass through each other without mutual effect. Likewise, sound waves from different instruments in an orchestra propagate in space independent of each other and can be distinguished separately. There occur many such instances in which a number of waves meet and proceed unaffected. Since waves do not interact, each region of space where two or more waves meet, undergo the vibrations set up by each wave separately. The resultant displacement at a given point in space is determined by the principle of superposition.

The *principle of superposition* states that the net displacement of a given point in space at any time due to two or more waves is the algebraic sum of the displacements produced at that point by all waves.

Mathematically speaking, the principle of superposition states that if two or more waves are propagating through the space, the resultant is given by the sum of wave functions of the individual waves. Thus, if $y_1(x, t)$ and $y_2(x, t)$ are the wave functions characterizing two waves traveling in space, their resultant is given by

$$y(x, t) = y_1(x, t) + y_2(x, t) \quad (5.60)$$

Wave equations that obey the superposition principle are said to be *linear*. The superposition principle applies only to waves of small amplitude.

5.17 INTERFERENCE OF LIGHT WAVES

When harmonic waves of identical frequency propagating in a medium meet each other, they give rise to the phenomenon of interference. Let us now understand what happens when two or more light waves overlap in some region of space. Let us assume for the sake of simplicity that two sinusoidal waves of the same frequency propagate through different paths x_1 and x_2 and meet at P in the region of observation, as shown in Fig. 5.17. Let the waves be represented by

$$E_A = E_1 \sin \omega t \quad (5.61)$$

and

$$E_B = E_2 \sin (\omega t + \delta) \quad (5.62)$$

where δ is the phase difference between them.

According to the Young's principle of superposition, the resultant electric field at a given place due to the simultaneous action of two or more harmonic waves is the algebraic sum of the electric fields of the separate constituent waves.

Thus, the resultant electric field at the point P due to the simultaneous action of the two waves is given by

$$E_R = E_A + E_B \quad (5.63)$$

$$= E_1 \sin \omega t + E_2 \sin (\omega t + \delta)$$

$$= E_1 \sin \omega t + E_2 (\sin \omega t \cos \delta + \cos \omega t \sin \delta)$$

$$= (E_1 + E_2 \cos \delta) \sin \omega t + E_2 \sin \delta \cos \omega t \quad (5.64)$$

Eq. (5.64) shows that *the superposition of two sinusoidal waves having the same frequency but with a phase difference produces a sinusoidal wave with the same frequency but with a different amplitude E .*

$$\text{Let } E_1 + E_2 \cos \delta = E \cos \phi \quad (5.65)$$

$$\text{and } E_2 \sin \delta = E \sin \phi \quad (5.66)$$

where E is the amplitude of the resultant wave and ϕ is the new initial phase angle. In order to solve for E and ϕ , we square the eqn. (5.65) and (5.66) and add them.

$$(E_1 + E_2 \cos \delta)^2 + E_2^2 \sin^2 \delta = E^2 (\cos^2 \phi + \sin^2 \phi)$$

$$\text{or } E^2 = E_1^2 + E_2^2 \cos^2 \delta + 2E_1 E_2 \cos \delta + E_2^2 \sin^2 \delta$$

$$\text{or } E^2 = E_1^2 + E_2^2 + 2E_1 E_2 \cos \delta \quad (5.67)$$

Thus, it is seen that the square of the amplitude of the resultant wave is not a simple sum of the squares of the amplitudes of the superposing waves, there is an additional term which is known as the interference term.

5.17.1 Intensity Distribution

The intensity of a light wave is given by the square of its amplitude.

$$I = \frac{1}{2} \epsilon_0 c E^2 \propto E^2$$

Using this relation into (5.67), we get

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \delta \quad (5.68)$$

We see that the resultant intensity at P on the screen is not just the sum of the intensities due to the separate waves. The term $2\sqrt{I_1 I_2} \cos \delta$ is known as the **interference term**. Whenever the phase difference between the waves is zero, i.e. $\delta = 0$, we have maximum amount of light. Thus,

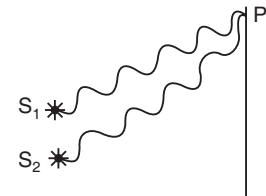


Fig. 5.17: Two light waves overlapping at point P .

$$I_{\max} = I_1 + I_2 + 2\sqrt{I_1 I_2}$$

When $I_1 = I_2 = I_0 \quad I_{\max} = 4I_0 \quad (5.69)$

It means that the resultant intensity I will be *more than the sum* of the intensities due to the two sources.

When the phase difference is $\delta = 180^\circ$, $\cos 180^\circ = -1$ and we have a minimum amount of light.

$$I_{\min} = I_1 + I_2 - 2\sqrt{I_1 I_2}$$

which, when $I_1 = I_2$, becomes

$$I_{\min} = 0 \quad (5.70)$$

It means that the resultant intensity I will be *less than the sum* of the intensities due to the two sources./

At points that lie between the maxima and minima, /when $I_1 = I_2 = I_0$, we get

$$\begin{aligned} I &= I_0 + I_0 + 2I_0 \cos \delta \\ &= 2I_0 (1 + \cos \delta) \end{aligned}$$

Then using the identity, $1 + \cos \delta = 2 \cos^2 \left(\frac{1}{2} \delta \right)$, we get

$$I = 4I_0 \cos^2 \left(\frac{1}{2} \delta \right) \quad (5.71)$$

Eqn. (5.71) shows that the intensity varies along the screen in accordance with the *law of cosine square*. Fig. 5.18 shows the variation of intensity as function of phase angle δ .

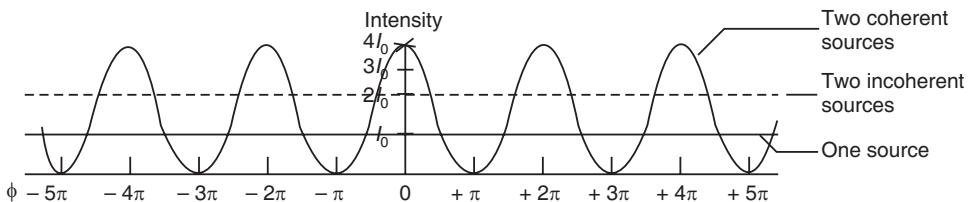


Fig. 5.18

It is seen from the plot that the intensity varies from zero at the minima to $4I_0$ at the maxima.

5.17.2 Superposition of Incoherent Waves

Incoherent waves are the waves that do not maintain a constant phase difference. Then the phase of the waves fluctuates irregularly with time and independently of each other. In case of light waves the phase fluctuates randomly at a rate of about 10^8 per second. Light detectors such as human eye, photographic film etc cannot respond to such rapid changes. The detected intensity is always the average intensity, averaged over a time interval which is very much larger than the time of fluctuation. Thus,

$$I_{\text{ave}} = I_1 + I_2 + 2\sqrt{I_1 I_2} <\cos \delta>$$

The average value of the cosine over a large time interval will be zero and hence the interference term becomes zero. Therefore, the average intensity of the resultant wave is

$$I_{\text{ave}} = I_1 + I_2$$

$$\text{If } I_1 = I_2, \text{ then } I_{\text{ave}} = 2I \quad (5.72)$$

It implies that the superposition of incoherent waves does not produce interference but gives a uniform illumination. The average intensity at any point is simply equal to the sum of the intensities of the component waves. For example, when the headlights of a car illuminate the same area, their combined intensity is simply the sum of the two separate intensities. The superposition of light from the headlights does not produce dark regions of zero intensity. The result (5.72) further implies that the superposition of incoherent light waves can never result in zero intensity.

5.17.3 Superposition of Many Coherent Waves

The result (5.69) may be written as

$$I_{\text{max}} = 2^2 I_0$$

which gives the resultant intensity when two coherent waves superpose. The resultant maximum intensity due to N coherent waves will be therefore

$$I_{\text{max}} = N^2 I_0 \quad (5.73)$$

and the minimum intensity $I_{\text{min}} = 0$ (5.74)

where N represents the number of coherent waves superposing at a point.

5.17.4 Conditions of Maximum Intensity and Zero Intensity

We may represent in a more general way the two waves meeting at P (Fig. 5.14) as follows:

$$E_A = E_1 \sin(\omega t - kx_1 + \phi_1)$$

and

$$E_B = E_2 \sin(\omega t - kx_2 + \phi_2)$$

where x_1 and x_2 represent the paths traveled by waves A and B . Therefore, the phase difference δ between the waves comes out to

$$\begin{aligned} \delta &= (-kx_1 + \phi_1) - (-kx_2 + \phi_2) = k(x_2 - x_1) + (\phi_1 - \phi_2) \\ &= \frac{2\pi}{\lambda} \mu L + (\phi_1 - \phi_2) \end{aligned}$$

or

$$\delta = \frac{2\pi}{\lambda} \Delta + (\phi_1 - \phi_2) \quad (5.75)$$

Thus, the phase difference δ between the waves is made up of two parts – one part arising from the difference in the paths traversed by them while the second part is on account of the difference in phases at the points of origin of the two waves. As a special case, let us assume that $\phi_1 = \phi_2$. Then $(\phi_1 - \phi_2) = 0$; which implies that the two waves under consideration are in phase initially. Equ. (5.75) reduces to

$$\delta = \frac{2\pi}{\lambda} \Delta$$

which indicates that the two waves have the same phase originally but subsequently developed the phase difference because they have traveled along different routes to reach the point P . At P they are displaced with respect to each other by a fraction of wavelength equal to Δ/λ .

The result of superposition of the waves at P is then determined solely by the path difference Δ . The relation (5.67) becomes

$$E^2 = E_1^2 + E_2^2 + 2E_1 E_2 \cos\left(\frac{2\pi}{\lambda} \Delta\right)$$

Assuming $E_1 = E_2$, the above relation reduces to

$$E^2 = 2E_1^2 \left(1 + \cos \frac{2\pi}{\lambda} \Delta \right)$$

$$\therefore I = 2I_1 \left(1 + \cos \frac{2\pi}{\lambda} \Delta \right)$$

Using the trigonometric relation $1 + \cos 2\theta = 2\cos^2 \theta$ into the above equation, we get

$$I = 2I_1 \left(2\cos^2 \frac{\pi}{\lambda} \Delta \right)$$

or $I = 4I_1 \left(\cos^2 \frac{\pi}{\lambda} \Delta \right) \quad (5.76)$

The relation (5.76) helps us state the conditions of maximum intensity and zero intensity in terms of the path difference. The resultant intensity

$$I = I_{\max} = 4I_1 \quad \text{if } \cos^2 \frac{\pi}{\lambda} \Delta = 1$$

This will happen when $\Delta = 0, \lambda, 2\lambda, 3\lambda, \dots, m\lambda$.

Whenever the above condition is satisfied the waves will meet in phase at P and the waves interfere constructively. Thus, *when two waves are not displaced with respect to each other or when they are displaced through an integral number of wavelengths, constructive interference takes place*. On the other hand,

$$I = I_{\max} = 0 \quad \text{if } \cos^2 \frac{\pi}{\lambda} \Delta = 0$$

which will happen if $\Delta = \frac{\lambda}{2}, 3\frac{\lambda}{2}, 5\frac{\lambda}{2}, \dots, (2m+1)\frac{\lambda}{2}$.

In this situation, the waves meet at P in inverted relation. The crests of one wave fall on the troughs of the second wave leading to cancellation of the wave displacements. Then the waves interfere destructively. Thus, *when two waves are displaced with respect to each other by an odd number of half-wavelengths, destructive interference results*. We summarize the conditions for the two types of interferences as follows:

$$\begin{array}{lll} I = I_{\max} & & \\ \text{if } \Delta = m\lambda & & \text{Constructive Interference} \\ \text{where } m = 0, 1, 2, \dots & & \end{array} \quad (5.77)$$

$$\begin{array}{lll} I = I_{\min} = 0 & & \\ \text{if } \Delta = (2m+1)\frac{\lambda}{2} & & \text{Destructive Interference} \\ \text{where } m = 0, 1, 2, \dots & & \end{array} \quad (5.78)$$

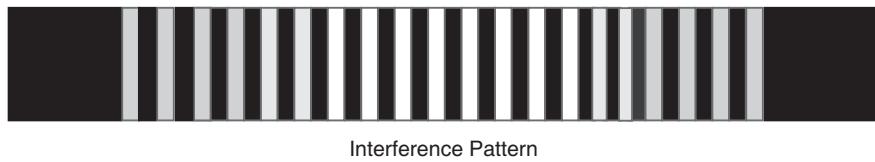
The condition (5.77) shows that the region, where constructive interference occurs, appears brighter than the surrounding region. Similarly, the region, where destructive interference occurs, appears dark. Therefore, the conditions (5.77) and (5.78) are as well the conditions for brightness and darkness. We restate the conditions (5.77) and (5.78) as

$$\Delta = m\lambda; \quad (m = 0, 1, 2, \dots) \quad \text{Brightness} \quad (5.79)$$

$$\Delta = (2m+1)\frac{\lambda}{2}; \quad (m = 0, 1, 2, \dots) \quad \text{Darkness} \quad (5.80)$$

Since the conditions for brightness and darkness repeat after a distance of one wavelength again and again, the region around point P is divided into alternate bright and dark regions.

The stationary pattern of bright and dark bands is called *interference pattern*. The bands are known as **interference fringes** (Fig. 5.19).



Interference Pattern

Fig. 5.19: Interference pattern consisting of alternate bright and dark bands.

Fig. 5.20 shows the energy distribution on a screen under different conditions.

- If only one beam of light illuminates the screen the intensity distribution is uniform throughout the area of illumination, as shown in Fig. 5.20 (a).
- If two *incoherent* beams of light of equal intensity illuminate the screen, the intensity distribution is again uniform, but twice as much light ($2I$) will reach the screen, as shown in Fig. 5.20 (b).

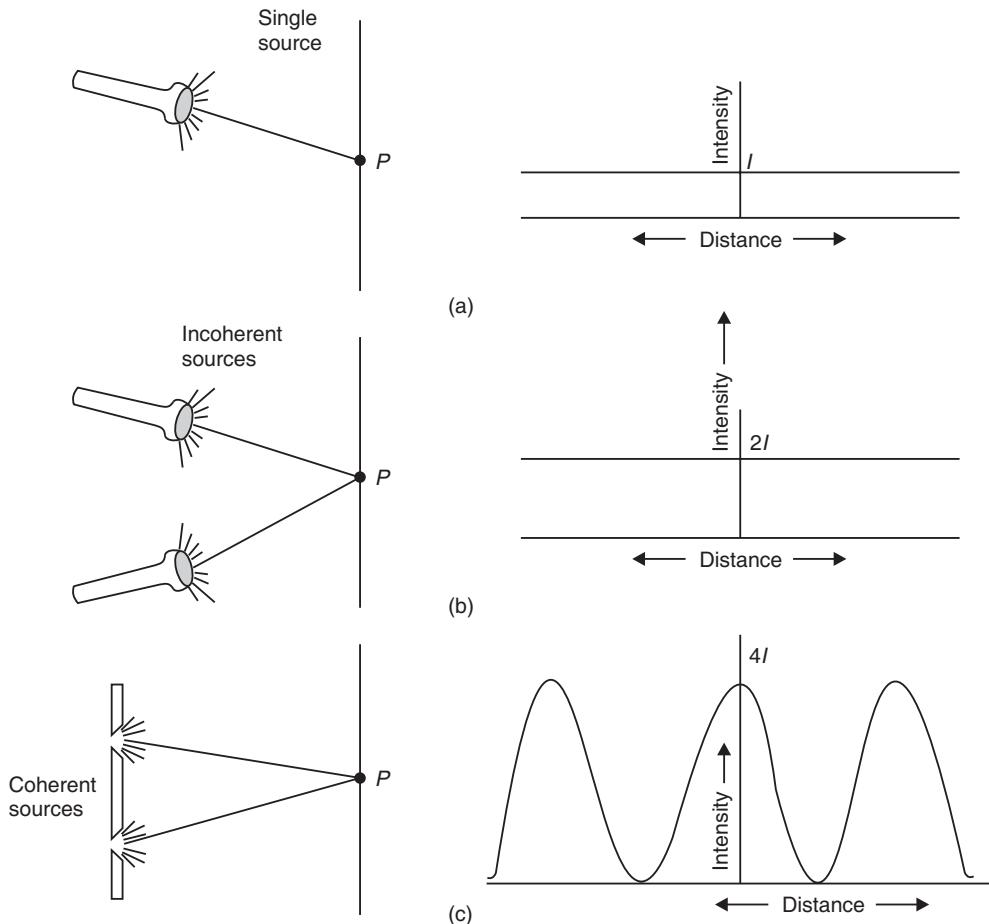


Fig. 5.20: Intensity distribution on the screen when – (a) single source illuminates the screen, (b) two incoherent sources illuminate the screen; (c) two coherent sources illuminate the screen.

(c) If two coherent beams of equal intensity illuminate the screen, the resultant intensity is not just the algebraic sum of the individual intensities of beams. Also, it is not the same at every point on the screen. It oscillates between alternate maxima and minima in accordance to the relation (5.78) and as shown in Fig. 5.20(c). The maxima contain four times the individual intensity while the minima are of zero intensity. However, the area under the curve in Fig. 5.20(c) will be equal to $2I$ signifying that the energy is neither created nor destroyed but is only redistributed during the interference.

Example 5.7: Two coherent light waves arrive at a particular point on a screen. The optical path difference between the waves is $3\mu\text{m}$. Determine the nature of interference at the point if the wavelength of the wave is 3900 \AA .

Solution: Constructive interference occurs between the waves at the point, if the optical path difference contains an even number of half-waves; otherwise the nature of interference will be destructive if it contains odd number of half-waves. Thus, if

$$\Delta = m \frac{\lambda}{2} \quad (m = 2, 4, 6, \dots) \text{ constructive}$$

$$\text{If, } \Delta = m \frac{\lambda}{2} \quad (m = 1, 3, 5, 7, \dots) \text{ destructive}$$

$$\Delta = m \frac{\lambda}{2}; m = \frac{2\Delta}{\lambda} = \frac{2 \times 3 \times 10^{-6} \text{ m}}{3.9 \times 10^{-7} \text{ m}} = 15.4 = 15$$

As m is an odd number, it means that the optical path difference is such that the waves arrive at the point with a phase difference of 180° . So the waves interfere destructively.

5.18 YOUNG'S DOUBLE SLIT EXPERIMENT

Young gave the first demonstration of the interference of light waves in 1801. Fig. 5.21 shows a plan view of the basic arrangement of his double slit experiment. The primary light source at S is a monochromatic source; it is generally a sodium lamp, which emits yellow light of wavelength at around 5893 \AA . The expanding wave front from the primary light source S falls on two narrow closely spaced slits, S_1 and S_2 as shown in Fig. 5.21. The slits at S_1 and S_2 are very narrow and partition the incident wave front. If the slits are equidistant from S , the phase of the wave at S_1 will be the same as the phase at S_2 because parts of the same wave front emerging from S pass through S_1 and S_2 . Further, waves leaving S_1 and S_2 have the same frequency as the primary source. Hence, sources S_1 and S_2 act as secondary coherent sources. The waves leaving from S_1 and S_2 interfere and produce alternate bright and dark bands on the screen at T .

Thomas Young used this experiment to make the first measurement of the wavelength of the light.

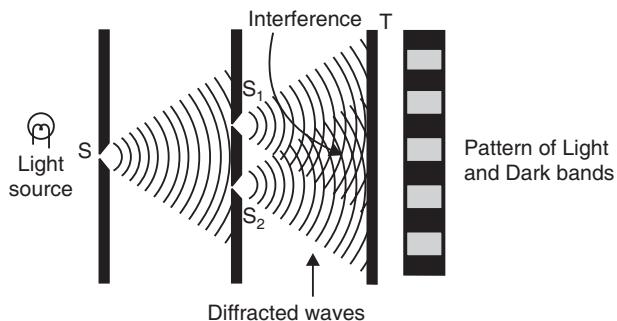


Fig. 5.21: Young's double slit arrangement -The narrow slit S acts as a source of cylindrical waves which illuminate the slits S_1 and S_2 . S_1 and S_2 behave as coherent sources and produce interference.

Now, if light is allowed to illuminate the slits S_1 and S_2 directly, instead of through slit S , interference will not be produced and the observation screen will be uniformly illuminated. We have learnt that interference is the manifestation of coherence. The absence of interference emphasizes the fact that coherence is lacking. To gain an understanding of the role of slit S in the double slit interference, let us look more intimately at the structure of real light waves.

5.19 WAVE TRAINS—LIGHT FROM COMMON SOURCES

In practice, light is emitted from a light source when excited atoms in it pass from upper excited states to a lower energy state. An atom leaving an excited state gives up the excess energy in the form of a burst of light, *photon*, and jumps to the lower normal state. The process of transition of the atom from an upper state to a lower state lasts for a brief time of about 10^{-9} sec. Therefore, the light emitted by an atom is not a continuous harmonic wave of infinite extension but is a *wave train* of finite length having a certain limited number of oscillations. Such a light burst is also called a **wave train** or a **wave packet**. After some time the atom again receives energy and jumps into excited state and subsequently emits another wave train. These emission events occur quite randomly. Other atoms in the source behave similarly but with different emission times. Adding together the wave trains, generated by all atoms in the light source, produces a succession of wave trains, which have their phases, distributed randomly (see Fig. 5.22).

To sum up, the light emitted by an ordinary light source is not an infinitely long, simple harmonic wave but is composed of a jumble of finite wave trains. We therefore call a real monochromatic source as a *quasi-monochromatic source*. The wave trains issuing out of a quasi-monochromatic source are as shown in Fig. 5.22.

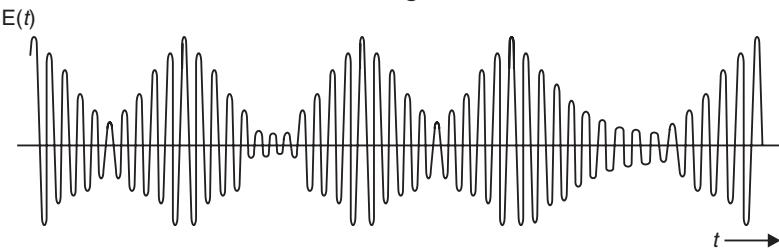


Fig. 5.22: Light is emitted in the form of wave trains

If a wave train lasts for a time interval Δt , then the length of the wave train in a vacuum is

$$l = c\Delta t \quad (5.81)$$

where c is the velocity of light in a vacuum.

$$\text{If } t = 10^{-9} \text{ s, } l = (3 \times 10^8 \text{ m/s}) (10^{-9} \text{ s}) = 0.3 \text{ m}$$

∴ The number of oscillations present in the wave train is

$$N = \frac{l}{\lambda} \quad (5.82)$$

where λ is the wavelength.

If the wave length λ is of the order of 5×10^{-7} m (5000 Å), then the number of wave oscillations present in a wave train is

$$N = \frac{0.3 \text{ m}}{5 \times 10^{-7} \text{ m}} = 0.6 \times 10^6.$$

Thus, a wave train contains about a million wave oscillations.

5.19.1 Bandwidth

A wave train (or a wave packet) is not a harmonic wave. Therefore, it cannot be represented mathematically by the simple sine function such as (5.37). The mathematical representation of a wave packet is done in terms of Fourier integrals and is complex. It is not taken up here. Instead, we describe some of the important features of a wave packet.

If light emitted from a source is analyzed with the help of a spectrograph, it is known to be made up of discrete **spectral lines**. These spectral lines are formed by wave packets emitted by atoms. Therefore, a spectral line and a wave packet are equivalent descriptions. In the first place, the wavelength of a *wave packet* or a *spectral line* is not precisely defined. There is a continuous spread of wavelengths over a finite range, $\Delta\lambda$, centered around a wavelength λ_0 . The wave packet may be considered as consisting of a number of harmonic waves which differ by infinitesimal increments of wavelengths. The maximum intensity of the wave packet will occur at λ_0 and the intensity rapidly falls off on either side of λ_0 , as illustrated in Fig. 5.23. The spread of wavelength is called *line width* or *bandwidth*. The *band width* is the wavelength interval from $\lambda_0 - \Delta\lambda/2$ to $\lambda_0 + \Delta\lambda/2$ which contains the major portion of the energy of the wave packet.

According to the Fourier analysis, the bandwidth expressed in terms of frequency is governed by the relation,

$$\Delta\nu = \frac{1}{\Delta t} \quad (5.83)$$

where Δt is the average life time of the light emitting atom in its excited state.

5.20 COHERENCE

Coherent light means that the light waves maintain constant phase difference over a period of time. A constant phase difference could be maintained easily if the waves are harmonic waves. The real light waves are not harmonic waves but a jumble of wave trains.

In passing from one wave train to the next wave train in a light wave there occurs an abrupt change in phase (see Fig. 5.24). Therefore, it is not possible to relate the phase at a point in one wave train to a point in another wave train. The phase relationship fluctuates irregularly from one wave train to another wave train. As a result, interference of waves does not occur and the resultant illumination produced by conventional sources is not so high.

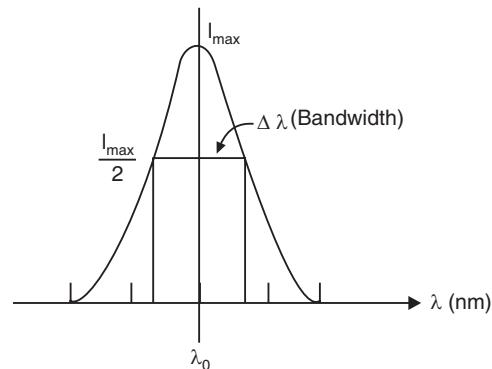


Fig. 5.23: Band width of a spectral line or wave packet

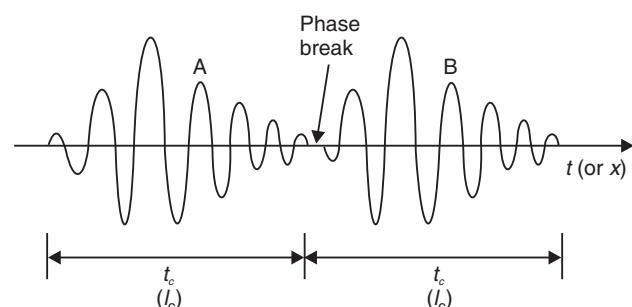


Fig. 5.24: A light wave is made up short wave trains A, B etc. In passing from one wave train to the next wave train there is an abrupt change in phase.

Coherence effects are mainly divided into two categories: *temporal* and *spatial*. The temporal coherence is related directly to the finite bandwidth of the source. The spatial coherence is related to the finite size of the source.

5.20.1 Temporal Coherence

If two waves maintain a definite relationship between their phases at a given time and at a certain time later, then the waves are said to be *temporally coherent*.

Let a point source of quasimonochromatic light S (Fig. 5.25) emit light in all directions. Let us consider light travelling along the line SP_1P_2 . The phase relationship between the points P_1 and P_2 depends on the distance P_1P_2 and the coherence length of the light beam. The electric fields at P_1 and P_2 will be correlated in phase when a single wave train extends over greater length than the distance P_1P_2 ; that is if the distance P_1P_2 is less than the coherence length l_{coh} . The waves are correlated in their rising and falling and they will preserve a constant phase difference.

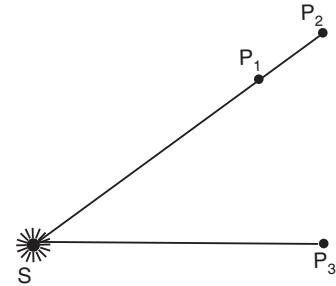


Fig. 5.25

The points P_1 and P_2 would not have any phase relationship if the longitudinal distance P_1P_2 is greater than l_{coh} , since in such a case many wave trains would span the distance. It means different independent wave trains would be at P_1 and P_2 at any instant and therefore the phase at the two points would be independent of each other. The degree to which a correlation exists is known as the amount of **longitudinal coherence**. As monochromaticity is related to the coherence length, temporal coherence is regarded as a measure of **monochromaticity**.

Temporal coherence is characterized by two parameters, namely *coherence length* and *coherence time*.

Coherence time: In case of conventional sources of light, light is emitted in the form of short wave trains and the phase of one wave train would remain constant with respect to the phase of another wave train for only about 10^{-9} sec. It implies that the two wave trains are *temporally coherent* for a maximum period of about 10^{-9} sec. This time is called the **coherence time** (t_{coh}). *Coherence time is defined as the longest time interval over which the phase undergoes change in a regular way.*

Coherence length: The distance for which each wave train remains sinusoidal is called **coherence length** (l_{coh}). *Coherence length is defined as the spatial extent over which the wave train has predictable phase.*

Expressions for Coherence length and coherence time

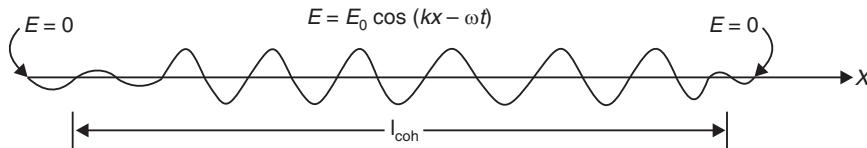


Fig. 5.26: A typical wave train

Let us consider a wave train generated by an atom (Fig. 5.26) at a particular instant. The middle of the wave train appears fairly sinusoidal for some number of oscillations whereas abrupt changes of frequency and phase occur at its ends. The length of the wave train over which it may be assumed to have a fairly sinusoidal character and predictable phase is known

as **coherence length**. We may consider coherence length as approximately equal to the length of the wave train, $c\Delta t$, over which its phase is not randomized. The time interval during which the phase of the wave train can be predicted reliably is called **coherence time**. It is the time, Δt , during which the phase of the wave train does not become randomized but undergoes change in a regular systematic way. We can therefore write

$$l_{coh} = c\Delta t \quad (5.84)$$

and

$$t_{coh} = \Delta t \quad (5.85)$$

\therefore

$$l_{coh} = c t_{coh} \quad (5.86)$$

Relation between coherence length and frequency bandwidth

A wave train consists of a group of waves, which have a continuous spread of wavelengths over a finite range $\Delta\lambda$ centered on a wavelength λ_0 . According to Fourier analysis the **frequency bandwidth** $\Delta\nu$ is given by

$$\Delta\nu \approx \frac{1}{\Delta t} \quad (5.87)$$

where Δt is the average lifetime of the excited state of the atom.

Δt in equ (5.87) is the time during which a wave train is radiated by the atom and corresponds to the coherence time, t_{coh} , of the wave packet.

$$\therefore \Delta\nu = \frac{1}{\Delta t} = \frac{1}{t_{coh}}$$

Using the relation (5.86) into the above relation, we get

$$\Delta\nu = \frac{c}{l_{coh}} \quad (5.88)$$

or

$$l_{coh} = \frac{c}{\Delta\nu} \quad (5.89)$$

Relation between coherence length and wavelength bandwidth

The frequency and wavelength of a light wave are related through the equation

$$\nu = \frac{c}{\lambda_0} \quad (5.90)$$

where λ_0 is the vacuum wavelength.

Differentiating equ. (5.90) on both sides, we get

$$\Delta\nu = -\frac{c}{\lambda_0^2} \Delta\lambda \quad (5.91)$$

Using the relation (5.88) into equ. (5.91), we obtain

$$\frac{c}{l_{coh}} = -\frac{c}{\lambda_0^2} \Delta\lambda \quad (5.92)$$

Rearranging the terms, we get

$$l_{coh} = \frac{\lambda_0^2}{\Delta\lambda} \quad (5.93)$$

The minus sign has no significance and hence is ignored. $\Delta\lambda$ is the **natural line width** or **bandwidth**.

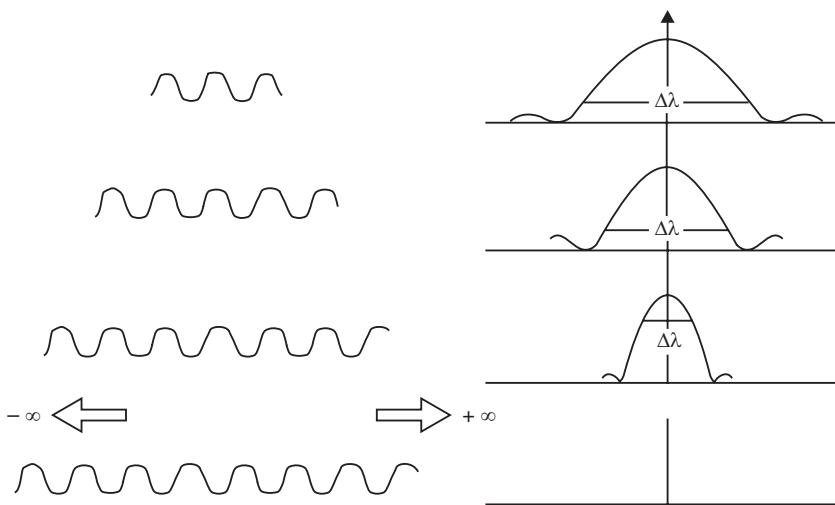


Fig. 5.27: Dependence of bandwidth on the length of the wave train

Equ. (5.93) implies that the longer is the length of the wave trains emitted by a light source, more monochromatic will be its light (Fig. 5.27). Thus, we conclude that **temporal coherence is a measure of the monochromaticity** of the light source.

5.20.2 Spatial Coherence

Spatial coherence refers to the continuity and uniformity of a wave in a direction perpendicular to the direction of propagation. *If the phase difference for any two fixed points in a plane normal to the wave propagation does not vary with time, then the wave is said to exhibit spatial coherence.* Thus, two waves from two different points of an extended source are said to be spatially coherent if they maintain a constant phase difference over a period of time. The spatial coherence is also known as **transverse coherence**.

Again looking at the point source S (Fig. 5.25), $SP_1 = SP_3$ and therefore, the fields at points P_1 and P_3 would have the same phase. Thus, an ideal point source exhibits spatial coherence, as the waves produced by it are likely to have the same phase at points in space, which are equidistant from the source. On the other hand, an extended source is bound to exhibit lesser lateral spatial coherence. Two points on the source separated by a lateral distance greater than one wavelength will behave quite independently. Therefore, correlation is absent between the phases of the waves emitted by them. **The degree of contrast of the interference fringes produced by a source is a measure of the degree of the spatial coherence of its waves.** The higher the contrast, the better is the spatial coherence.

In the Young's double slit experiment, a single slit S is placed between the light source and the double slit so that the same group of wave trains is incident on slits S_1 and S_2 . When the phase of the wave changes at S , this change is communicated simultaneously to S_1 and S_2 . Therefore, the waves emerging from S_1 and S_2 will be coherent with respect to each other. Thus, the purpose of keeping the single slit in the arrangement is to obtain spatial coherence between the waves passing through slits S_1 and S_2 .

Fig. 5.26 illustrates the above concepts clearly. A, B, and C are three wave trains emitted from different points on an extended source of light. Waves in Fig. 5.28(a) have good spatial coherence but have poor temporal coherence (note that the wave trains are shorter and hence coherence time and coherence length are smaller). Waves in Fig. 5.28(b) have good temporal coherence but have poor spatial coherence. In contrast, waves in Fig. 5.28(c) possess good temporal coherence as well as good spatial coherence.

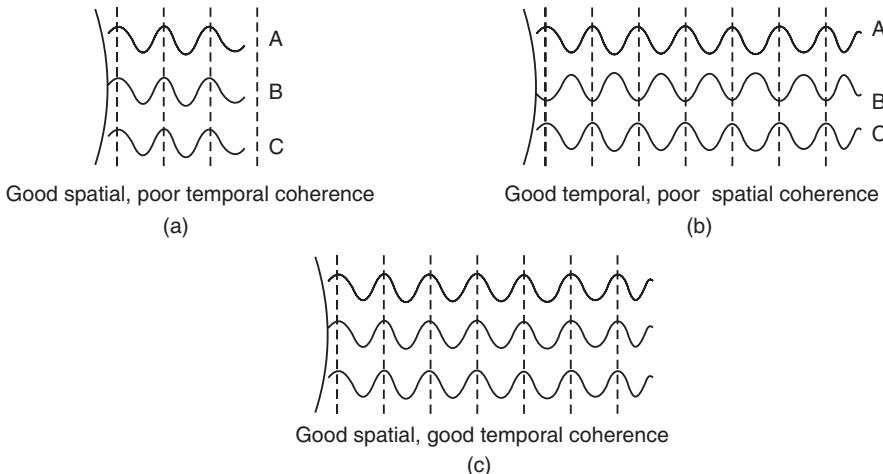


Fig. 5.28: Illustration of poor and good coherence

Example 5.8: Calculate the coherence length for CO_2 laser whose line width is 1×10^{-5} nm at IR emission wavelength of $0.6 \mu\text{m}$. (RTMNU, S-94)

$$\text{Solution: } l_{\text{coh}} = \frac{\lambda^2}{\Delta\lambda} = \frac{(10.6 \mu\text{m})^2}{1 \times 10^{-5} \text{ nm}} = \frac{(10.6 \times 10^{-6})^2}{10^{-5} \times 10^{-9}} = 11.2 \text{ km}$$

5.21 DOUBLE SLIT EXPERIMENT AGAIN

We now illustrate the concept of coherence with the help of Young's double slit experiment (Fig. 5.21).

- In this experiment the wave train coming from the slit S is divided physically into two segments by the slits S_1 and S_2 . The two sets of wave trains resulting through division of a parent wave train travel along different paths and ultimately overlap on each other. If the paths of overlapping wave trains differ by more than one coherence length, interference does not take place because the superposed waves are not parts of the same group of wave trains and are incoherent. It is therefore essential that the **optical path difference Δ must be smaller than coherence length for interference to occur**. Thus,

$$\Delta \leq l_{\text{coh}} \quad (5.94)$$

The number of fringes observed in the experiment is limited because of the above condition. Constructive interference takes place whenever

$$\Delta = \pm m\lambda \quad (m = 0, 1, 2, 3, \dots)$$

For the extreme bright fringes in the pattern, we may assume that

$$\Delta_{\text{max}} \approx l_{\text{coh}}$$

$$\therefore (m)_{\max} \lambda \approx l_{\text{coh}} \approx \frac{\lambda^2}{\Delta\lambda}$$

$$\therefore (m)_{\max} \approx \frac{\lambda}{\Delta\lambda} \quad (5.95)$$

The above relation indicates that the number of fringes observed is inversely proportional to the bandwidth of the incident light. The number will be large if the bandwidth is small or temporal coherence is high.

- The degree of spatial coherence of a beam of light is elicited from the contrast of the fringes produced by it. The broader the source of the beam, the lesser is the degree of spatial coherence. In the double slit experiment if the slits S_1 and S_2 are directly illuminated by a light source, interference fringes are not observed and the screen is uniformly illuminated. The non-production of fringes indicates that the light issuing from the slits lacks spatial coherence. If a slit S is placed between the source and the double slits, the light passing through S illuminates the slits S_1 and S_2 . Fringes are not observed as long as S is wider. On gradually reducing the width of S , formation of fringes would be observed and the fringes will become distinctly clear when the slit is reduced to pin-hole size. It means that the slit S increases the spatial coherence of waves by limiting the size of the wave front from the source. It ensures that the wave trains incident on the slits S_1 and S_2 originate from a narrow region of the source. It is found that the beam will be spatially coherent when the lateral distance, d which is the slit separation, is of the order of $\lambda/2\theta$, where θ is the angle subtended by the source at the slit. Thus, the condition is that

$$d \approx \frac{\lambda}{2\theta} \quad (5.96)$$

The light produced by lasers is highly monochromatic and coherent. When they are used in double slit experiment, interference fringes are obtained without the necessity of slit S .

5.22 DISPERSION

Light waves of different wavelengths travel with different velocities in a transparent medium and will be bent at different angles at the refracting surface. Therefore, the refractive index of a medium varies with wavelength of incident light. Thus,

$$\mu = f(\lambda) \quad (5.97)$$

Fig. 5.29 shows the variation of refractive index with wavelength. Generally, the refractive index decreases with increasing wavelength. Any material in which the refractive index varies with wavelength is said to exhibit **dispersion**. White light passing through a medium will be separated according to wavelengths. It is referred to as **chromatic dispersion**.

When white light is incident at normal on a rectangular block of a medium, violet light would merely lag behind red light.

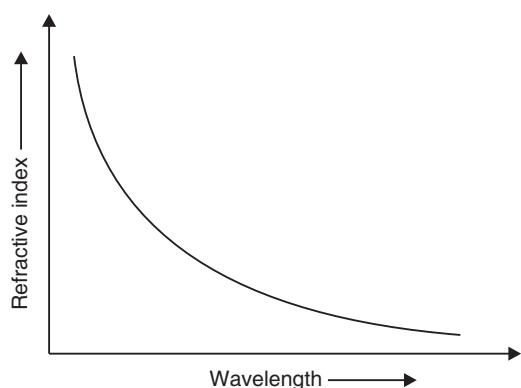


Fig. 5.29

When it is incident obliquely the colours become separated in space. When the medium is taken in a prism form, the separation of colours is much more. The angle between rays of two colours is called the **angular dispersion**.

Dispersion is expressed in terms of three refractive indices. A simple measure of the dispersions is provided by the angular separation of the red and violet rays. It is not necessary that materials of high refractive index exhibit high dispersion, as dispersion is not proportional to refractive index.

Dispersion varies as the third inverse power of wavelength as per the following relation

$$\frac{d\mu}{d\lambda} = -\frac{2B}{\lambda^3} \quad (5.98)$$

Dispersion of light in water droplets suspended in the atmosphere is responsible for the celestial spectacle called *rainbow*. Optical radiation emitted by bodies is studied with the spectrographs which use prisms or gratings to disperse colours.

5.23 SCATTERING

When a light beam encounters an obstacle of size comparable to its wavelength, diffraction occurs. If the size of the obstacle is much smaller than a wavelength (i.e. $d \ll \lambda$), light spreads in the form of spherical waves after striking the obstacle. Then light is said to have been **scattered** by the obstacle. Scattering redirects part of the incident radiation into directions other than the direction of the incident beam. Scattering is thus a special case of diffraction. The strength of the original beam decreases during the process of scattering, causing attenuation of energy.

The first quantitative study of the scattering of light was carried out by Rayleigh. In 1871, he established that the intensity of scattered light is inversely proportional to the fourth power of wavelength. Thus,

$$I_s \propto \frac{1}{\lambda^4} \quad (5.99)$$

This is known as the **Rayleigh's law of scattering**. It is familiar to us that sunlight scattered from very fine particles of smoke appears bluish in colour. The Rayleigh's law of scattering explains the reason for the bluish colour of sky on a cloudless day.

QUESTIONS

- What is meant by an optical medium?
- Explain the terms: homogeneous optical medium, and inhomogeneous optical medium.
- Explain the terms: isotropic medium and anisotropic medium.
- Define reflectivity and transmissivity. Give the mathematical expressions.
- What is total internal reflection? Derive the condition for total reflection.
- Derive Lambert-Bouguer law.
- What are the electromagnetic waves? Why light is classified in the category of electromagnetic waves?
- What are the characteristics of a wave?
- Define intensity of an electromagnetic wave and obtain an expression for it.
- What is the difference between optical and geometrical paths?
- Can the optical path between two points ever be less than the geometrical path length between these points?
- Can two 25-watt electric bulbs with point like filament of the same material and lying close to each other produce interference?

13. What becomes to the energy of light waves in destructive interference?
14. Interference can be observed only when the two sources have some common characteristics. What are they?
15. What are coherent waves?
16. State the conditions for formation of bright and dark fringes.
17. What is a periodic wave?
18. What are the parameters that characterize a wave motion?
19. Define time period, frequency and wavelength of a wave. How are they related to each other?
20. Why should the wavelength of light change, but not its frequency, in passing from one material to another?
21. Discuss the phases of oscillation at two points on a light ray if the separation between them is $\frac{3}{2}\lambda, 5\lambda, 2n\frac{\lambda}{2}$ and $(2n+a)\frac{\lambda}{2}$ where λ is the wavelength and n is an integer.
22. What are traveling waves? When a travelling wave propagates in a medium, what is that progresses forward?
23. What do you understand by the term 'Phase Velocity'?
24. Explain the following :
 - (a) Wave surface
 - (b) Wave front and
 - (c) ray
25. How are the following produced:
 - (a) Spherical waves
 - (b) Cylindrical waves and
 - (c) Plane waves?
26. A plane wave of wavelength λ and amplitude A propagates in a medium with a speed v . Write the equation of the plane wave.
27. When are two waves said to be in-phase and out-of-phase?
28. Two sinusoidal waves of equal amplitude are $\frac{1}{4}$ of a wavelength out of phase. What is the amplitude of the resultant wave?
29. Distinguish between the geometrical path length of a light ray in a medium and its optical equivalent. How are they related? **(RTMNU., 1990)**
30. What is the difference between optical and geometrical path difference? **(R.T.M.N.U., 2006)**
31. Can the optical path length between two points ever be less than the geometrical path length between those points?
32. Why light is represented by E vector? Explain monochromaticity and polarization of an e.m. wave.
33. What is the path change a light wave undergoes due to reflection at the interface of two media when
 - (i) It is travelling from a rarer medium to a denser medium and
 - (ii) It is travelling from denser medium to a rarer medium?
 What is the corresponding phase change? **(RTMNU 1993)**
34. Consider the superposition of two harmonic disturbances. Show that the resultant intensity due to them is not just the sum of their individual intensities.
35. Explain why two independent identical sources of light of the same frequency cannot give rise to interference fringe pattern.
36. What is meant by interference of light? State the fundamental conditions for obtaining sustained interference pattern.
37. Describe the interference pattern obtained due to superposition of the coherent waves. Write an expression for the intensity distribution over the plane of observation. Plot the resultant intensity as a function of distance on the screen.
38. Each of N atomic radiators is giving out radiation of same wavelength and intensity. Show under which conditions their resultant at a point can be NI or N^2I where I is intensity due to each radiator at point. **(RTMNU 1993)**
39. Explain the terms coherence length and coherence time for a light wave. Derive an expression for the coherence length of a wave train that has frequency band width $\Delta\nu$. Express the answer in terms of line width $\Delta\lambda_0$ and mean wavelength λ_0 of the wave train. **(RTMNU 1988)**

40. What do you understand by coherence? Explain temporal coherence and spatial coherence. Derive the condition for spatial coherence. **(C.S.V.T.U., 2005)**
41. Explain in brief:
(i) Coherence length (ii) Spatial coherence (iii) Temporal coherence.
Obtain an expression for coherence length. **(RTMNU 2001)**
42. Explain spatial coherence, temporal coherence and coherence and coherence length. **(Cochin Univ., 2005)**
43. What are the processes that are responsible for emission of light from a source? What can we conclude about the character of light emitted from any real source?
44. Explain the meaning of temporal coherence and explain how duration or length of a single wave train can be a measure of temporal coherence. **(RTMNU 2002)**
45. How can you experimentally distinguish between coherent and non-coherent light? **(RTMNU 2002)**
46. What is meant by dispersion of light?
47. Explain the phenomenon of light scattering. State Rayleigh's law of scattering.

PROBLEMS

- The frequency of green light is 5.5×10^{14} Hz. What is the wavelength in meters, microns, nanometers and angstrom units ? **[Ans: $\lambda = 5455 \text{ \AA}$]**
- What is the speed of light of wavelength 5000 Å in glass whose index of refraction is 1.50? **[Ans: $v = 2 \times 10^8 \text{ m/s}$]**
- The average wavelength of sodium light is 5893 Å. Find out the number of waves in 1 cm. **[Ans: $N = 1.7 \times 10^4 \text{ waves}$]**
- Monochromatic light of wavelength 4400 Å traverse from glass of refractive index 1.5 to a vacuum. Determine the change in wavelength. **[Ans : 2200 Å]**
- A soap bubble is 3×10^{-7} m thick. Red light is incident normally on the surface of the soap bubble. It gets partially reflected and partially transmitted at the top surface of the bubble. The transmitted ray travels through the film and again gets reflected partially at the bottom surface. The reflected component travels back and emerges out of the top surface. What is the geometrical path travelled by the ray in the film? What is the optical path? Refractive index is given as 1.333. **[Ans: 0.6 μm, 0.7998 μm.]**
- A sodium atom radiates for 4×10^{-12} s. What is the coherence length of light from a sodium lamp? **(N.U., W-97) [Ans: 1.2 mm]**
- If light of 6600 Å wavelength has wave trains 20λ long, what are its coherence length and coherence time? **[Ans: $1.32 \times 10^{-5} \text{ m}, 4.4 \times 10^{-14} \text{ s}$]**
- A He-Ne laser giving light at 6330 Å has a coherence length of 20 km. Determine (a) its coherence time; and (b) the number of waves per wave train. **[Ans: $6.7 \times 10^{-5} \text{ s}, 3.2 \times 10^{10} \text{ waves}$]**
- A mercury lamp has a band width $\Delta v = 1000 \text{ MHz}$. Calculate the coherence time and coherence length of its light. **[Ans: $10^{-9} \text{ s}, 30 \text{ cm}$]**
- Calculate the coherence length for CO₂ laser whose line width is $1 \times 10^{-5} \text{ nm}$ at IR emission wavelength of 10.6 μm. **(N.U., S-94) [Ans: 11.2 km]**
- White light has frequency range from $0.4 \times 10^{15} \text{ Hz}$ to $0.7 \times 10^{15} \text{ Hz}$. Find the coherence time and coherence length. **(N.U., S-96) [Ans : $3.3 \times 10^{-15} \text{ s}, 1 \mu\text{m}$]**

CHAPTER

6

Interference

6.1 INTRODUCTION

In 1678 Christian Huygens propounded the wave theory of light. He suggested that light propagates in the form of waves. Huygens did not suggest anything about the nature of light wave; whether it is a transverse wave or longitudinal wave. He had no knowledge about the speed of light or its wavelength. In 1801 Thomas Young provided the first experimental evidence for the wave theory of light from the double slit interference experiment and determined the wavelength of light waves. The interference is not easy to observe in case of light waves. One cannot observe interference fringes on a screen by placing two light sources in front of it. Sustained interference would be obtained only when coherent sources of light are used. Special devices, such as Lloyd's mirror, Fresnel biprism are designed to create two coherent sources from a single source of light and to determine accurately the wavelength of light. The striking colour effects caused by white light on a soap bubble and on a slick of oil spreading over small puddles of water are familiar to us from our childhood. Less noticed are the colours exhibited by coatings of oxide on heated metal pans in kitchen and thin layers of mica, cellophane etc. Thomas Young explained the origin of colours in all these cases on the basis of interference. The interference of certain colours constructively and others destructively causes the exhibition of iridescent colours by the thin films in reflected light. An understanding of the thin film interference phenomenon has led to a number of practical applications such as non-reflecting coatings, interference filters, and interference mirrors.

6.2 INTERFERENCE

Interference is an important consequence of superposition of waves. Let us consider two (or more) light waves of same frequency and having a constant phase difference travel in the same region of a medium simultaneously and cross each other (see Fig.6.1). Waves having the same frequency and a constant phase difference are known as **coherent waves**. Waves having fluctuating phase difference are known as **incoherent waves**. At the point of crossing, waves overlap or superpose on each other.

According to the **principle of superposition**, the combined effect of coherent waves at each point of the region of superposition is obtained by adding algebraically the disturbances due to individual waves. The resultant intensity at any point in the region of superposition depends upon the amplitudes and the phase relationships of the component waves. Let us assume here that the component waves are of the same amplitude, say A.

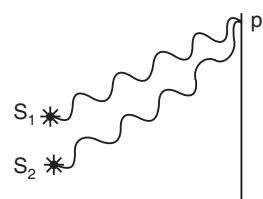


Fig. 6.1

We focus our attention mainly on two specific types of disposition of the waves (see Fig.6.2) at the point of observation, P.

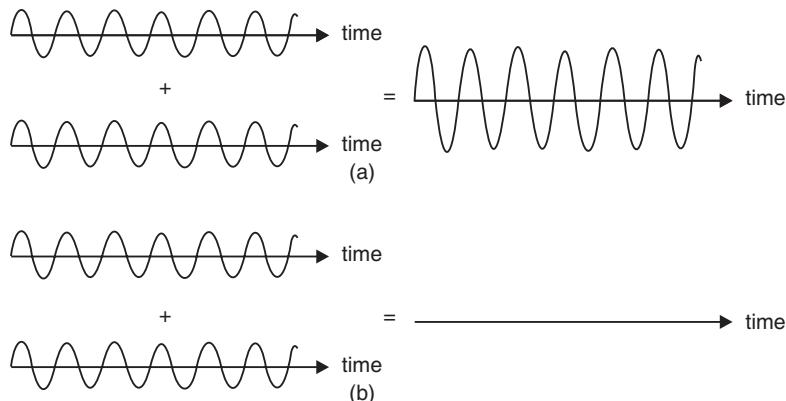


Fig. 6.2

- (i) If the two waves, meeting at P, reach their maxima, zeros and minima at the same instant of time, then their phase difference is zero (or an integral multiple of 2π). Such waves will have a crest-to-crest and trough-to-trough correspondence, as shown in Fig.6.2 (a). Then the waves are said to be **in phase**. The amplitude of the resultant wave at the point will then be equal to the sum of the amplitudes of the two waves, as shown in Fig.6.2 (a).
 - (a) The amplitude of the resultant wave = $A + A = 2A$. Hence, the intensity of the resultant wave is $I_R \propto 2^2 A^2 = 2^2 I$. It is obvious that the resultant intensity is greater than the sum of the intensities ($I + I = 2I$) due to individual waves. The interference produced at such points is known as *constructive interference*. *When two waves are not displaced with respect to each other or when they are displaced through an integral number of wavelengths, constructive interference takes place. Bright bands of light are observed at those points.* As S_1 and S_2 are coherent sources, the bright bands are **stationary**.
- (ii) If one of the waves reaches its maximum at the same time when the other reaches its minimum, then their phase difference is π radians. Such waves have a crest-to-trough correspondence, as shown in Fig. 6.2(b). The waves are said to be in **opposite phase** or **180° out-of-phase**. The amplitude of the resultant wave at the point will also be equal to the sum of the amplitudes of the two waves, as shown in Fig. 6.2(b). As the amplitude of one of the waves is negative, the amplitude of the resultant wave = $A - A = 0$. Hence, the intensity of the resultant wave is $I_R \propto 0^2 = 0$. It is obvious that the resultant intensity is less than the sum of the intensities ($I + I = 2I$) due to individual waves. The interference produced at such points is known as *destructive interference*. *When two waves are displaced with respect to each other by an odd number of half-wavelengths, destructive interference results. Dark bands of light are observed at those points.* As S_1 and S_2 are coherent sources, the dark bands are **stationary**.

Thus, when two or more coherent waves of light are superposed, the resultant effect is that at certain points **brightness** is produced while at other points **darkness** is produced in the medium.

The phenomenon of redistribution of light energy due to the superposition of light waves from two or more coherent sources is known as interference. The stationary bands of alternate darkness and brightness are known as **fringes**. The fringe pattern is obtained only when the interfering waves are coherent.

When incoherent waves overlap on each other, the resultant intensity is a simple addition of the intensities in the region of overlap. Therefore, the resultant intensity due to two incoherent waves of equal intensity is $I + I = 2I$ and the region of overlap is uniformly bright without the formation of fringes. It means that interference does not take place when incoherent waves superpose on each other.

6.3 CONDITIONS FOR OBSERVING SUSTAINED INTERFERENCE

We may now summarize the conditions that are to be fulfilled in order to observe sustained interference.

- (i) The waves from the two light sources must be of the same frequency.
- (ii) The waves from the two light sources must maintain a constant phase difference.
- (iii) For obtaining distinct bright and dark fringes, the vector sum of the overlapping electric field vectors should be zero in the dark regions. The sum will be zero only if the vectors are anti-parallel and have the same magnitude.
- (iv) If the two sets of waves are plane polarized, their planes of polarization must be the same. Waves polarized in perpendicular planes cannot produce interference effects.
- (v) The path difference between the overlapping waves must be less than the coherence length of the waves (see Art.5.20.1). If we consider two interfering wave trains, having constant phase difference, as in Fig.6.3, the interference effects occur due to parts QR of wave 1 and ST of wave 2. For the parts PQ and TU interference will not occur. Therefore, the interference pattern does not appear distinctly. When the entire wave train PR overlaps on the wave train SU, interference pattern will be distinct. On the other hand, when the path difference between the waves 1 and 2 becomes very large, the wave trains arrive at different times and do not overlap on each other. Therefore, in such cases interference does not take place. The interference pattern completely vanishes if the path difference is equal to the coherence length. It is hence required that

$$\Delta < l_{coh} \quad (6.1)$$

- (vi) The two coherent sources must lie close to each other in order to discern the fringe pattern.

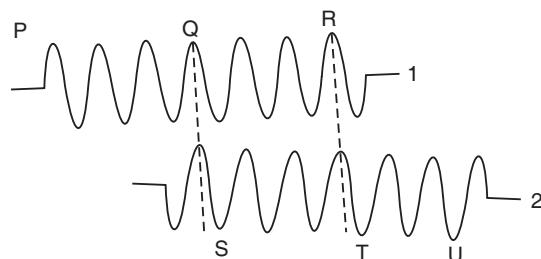


Fig. 6.3

6.4 TECHNIQUES OF OBTAINING INTERFERENCE

The phase relation between the waves emitted by two conventional light sources fluctuates rapidly and therefore they can never be coherent, though they are identical in all respects. However, two coherent sources are derived from a single source by techniques, which can be divided into two broad classes.

1. **Division of wave front:** One of the methods consists in using a narrow slit as the source and subsequently, the wave front is divided. For example, in the Young's double slit experiment, a wave front emerging from the slit S is divided into two parts by the double slit S_1S_2 . Fresnel's biprism, Lloyd's mirror, etc are the other examples where the **division of wave front** method is used.
2. **Division of amplitude:** In this method, amplitude of the light beam is divided by partial reflection into two or more beams. Thin films (wedge, Newton's rings etc), interferometers such as Michelson's interferometer etc utilize this method in producing interference.

6.5 REVIEW OF IMPORTANT CONCEPTS

When studying the topic of interference, we frequently come across the terms optical path and path difference. We should clearly understand these terms and know how to calculate them.

6.5.1 Geometrical Path

Light travels along a straight line path from a point A to another point B and it is known as the *path of the light*. The shortest path between any two points A and B is called the *geometrical path length* (GPL). GPL remains the same whether it is measured in a vacuum or in any medium.

6.5.2 Optical Path

Light travels μ times slower in a medium. Therefore, it takes μ times more time to cover the distance AB in the medium than it takes to cover the same distance in a vacuum. This time delay is accounted for by introducing another distance called *optical path length* (OPL). It is defined as

$$\text{O.P.L.} = \mu \times \text{G.P.L.}$$

or

$$\Delta = \mu L \quad (6.2)$$

The optical path length, Δ signifies the number of wavelengths that are accommodated in a given medium over the corresponding geometrical path length.

6.5.3 Path Difference

Light rays travel along different paths, which may lie in the same medium or in different media. The difference between optical paths of two rays travelling in different directions is known as the optical path difference.

6.5.4 Phase Difference

The phase of a wave arriving at a point depends on the optical path length it traversed. We know that if a wave covers in air a distance of one wavelength, 1λ , its phase changes by 2π radians. Therefore, we compute that if a wave travels a distance L in air, its phase change is given by

$$\delta = \frac{2\pi L}{\lambda} \quad (6.3)$$

When the wave travels the distance L in a medium, then

$$\delta = \frac{2\pi\Delta}{\lambda} = \frac{2\pi\mu L}{\lambda} \quad (6.4)$$

Comparing equs. (6.3) and (6.4), we find that a light path of geometric length L in a medium of refractive index μ produces the same phase change as a light path of length μL in a vacuum. Therefore, ***in the study of optics we always must calculate the optical paths travelled by light rays.***

The path difference between two in-phase waves may be zero or an integral multiple of a wavelength, λ and the path difference between two opposite-phase waves will be $\lambda / 2$ or an odd integral multiple of $\lambda / 2$.

Optical path difference and the consequent phase difference may arise due to two reasons. One reason is the difference in the optical paths (See Art.5.15) and the other is due to reflections at optical interfaces.

(a) Phase difference due to optical path difference:

Let us consider two sources of light S_1 and S_2 , as shown in Fig.6.4. Let us assume that the sources are identical and produce waves of same wavelength and that their vibrations are in the same phase at S_1 and S_2 . Light from these sources travel in air along different paths, S_1P and S_2P ; and meet at a point P . The path lengths S_1P and S_2P are different and contain different number of waves. The geometric path difference between the waves at P is $(S_2P - S_1P)$ and the optical path difference is $\mu(S_2P - S_1P)$. Since the paths contain different number of waves, the optical path difference will be equal either to a few integral number of waves or an integral number of wavelengths plus a fraction of one wavelength. This optical path difference leads to a phase difference between the waves meeting at P . It means that though the waves started with the same phase, they arrive at P with different phases because they travelled along different path lengths. Using eqn.(6.4), the phase difference between the waves at P may be expressed as

$$\delta = \frac{2\pi}{\lambda} \mu(S_2P - S_1P) \quad (6.5)$$

(b) Phase difference due to reflection at boundaries of optical interfaces:

Light waves may also undergo phase change due to reflection at some point in their path. If the waves are reflected at a rarer-to-denser medium boundary, the reflected waves suffer a phase change of π rad or 180° compared to the incident waves (See Fig. 6.5 a & b). It is seen from eqn.(6.3) that a phase change of π rad is equal to a path change of $\lambda/2$. Therefore, we must add (or subtract) $\lambda/2$ in the calculation of true optical path difference whenever a reflection occurs at a denser medium.

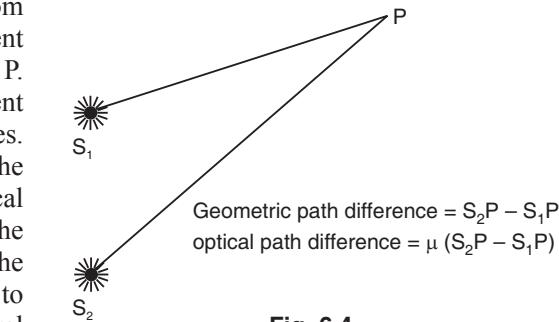


Fig. 6.4

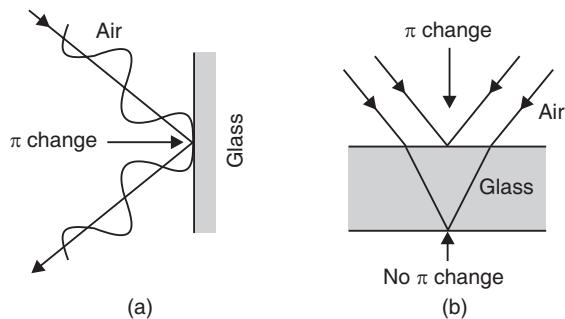


Fig. 6.5

INTERFERENCE BY DIVISION OF WAVE FRONT

6.6 FRESNEL BIPRISM

Fresnel used a biprism to show interference phenomenon. The biprism consists of two prisms of very small refracting angles joined base to base. In practice, a thin glass plate is taken and one of its faces is ground and polished till a prism (Fig.6.6 a) is formed with an obtuse angle

of about 179° and two side angles of the order of $30'$.

When a light ray is incident on an ordinary prism, the ray is bent through an angle called the *angle of deviation*. As a result, the ray emerging out of the prism appears to have emanated from a source S' located at a small distance above the real source, as shown in Fig. 6.6 (b). We say that the prism produced a *virtual image* of the source. A biprism, in the same way, creates two virtual sources S_1 and S_2 , as seen in Fig. 6.6 (c). These two virtual sources are images of the same source S produced by refraction and are hence *coherent*.

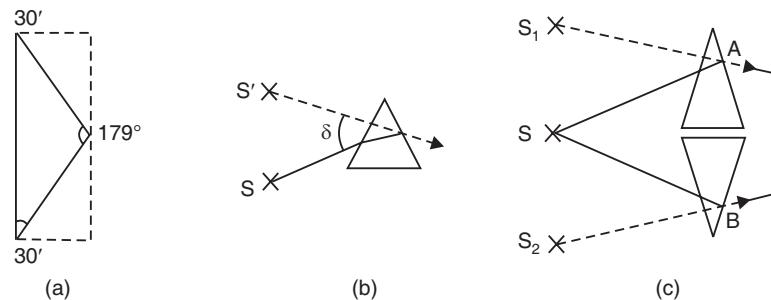


Fig. 6.6. Fresnel biprism and formation of virtual sources

6.6.1 Experimental Arrangement

The biprism is mounted suitably on an optical bench. An optical bench consists of two horizontal long rods, which are kept strictly parallel to each other and at the same level. The rods carry uprights on which the optical components are positioned. A monochromatic light source such as sodium vapour lamp illuminates a vertical slit S . Therefore, the slit S acts as a narrow linear monochromatic light source. The biprism is placed in such a way that its refracting edge is parallel to the length of the slit S . A single cylindrical wavefront impinges on both prisms. The top portion of the wavefront is refracted downward and appears to have emanated from the virtual image S_1 . The lower segment, falling on the lower part of the biprism, is refracted upward and appears to have emanated from the virtual source S_2 . The virtual sources S_1 and S_2 are coherent (see Fig. 6.7), and hence the light waves are in a position to interfere in the region beyond the biprism. If a screen is held there, interference fringes are seen. In order to observe fringes, a micrometer eyepiece is used.

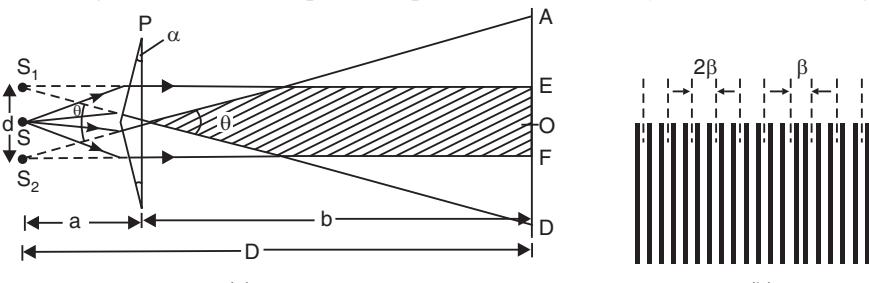


Fig. 6.7. Fresnel biprism and fringe formation

is refracted downward and appears to have emanated from the virtual image S_1 . The lower segment, falling on the lower part of the biprism, is refracted upward and appears to have emanated from the virtual source S_2 . The virtual sources S_1 and S_2 are coherent (see Fig. 6.7), and hence the light waves are in a position to interfere in the region beyond the biprism. If a screen is held there, interference fringes are seen. In order to observe fringes, a micrometer eyepiece is used.

Theory

Let S_1 and S_2 be the two virtual images of the source S . Let “ d ” be the distance between S_1 and S_2 . The fringes are formed on a screen T kept at a distance D from the biprism. The point O on the screen is equidistant from S_1 and S_2 . Hence, the waves arrive from S_1 and S_2 arrive at O simultaneously and the point O is always bright. The point O corresponds to the position of central bright fringe. On both sides of O , alternate bright and dark fringes, as shown in Fig. 6.7 (b), are produced.

Let P be an arbitrary point on screen (Fig. 6.8). Let θ be the angle that MP makes with the horizontal line MO. Let S_1N be a normal on to the line S_2P . The distances PS_1 and PN are equal. The waves emitted at the slits, S_1 and S_2 , are initially in phase with each other. The difference in the path lengths of these two waves is S_2N . We assume that the experiment is carried out in air. Therefore, the optical paths are identical with geometrical paths. The nature of the interference of the two waves at P depends simply on how

many waves are contained in the length of the path difference S_2N . If S_2N contains an integral number of wavelengths, the two waves interfere constructively, producing a maximum in the intensity of light on the screen at P. If it contains an odd number of half-wavelengths, the waves interfere destructively and produce a minimum intensity at P.

6.6.2 Optical Path Difference between the Waves at P

Let the point P be at a distance x from O (Fig. 6.8). Then

$$\begin{aligned} PE = x - \frac{d}{2} \text{ and } PF = x + \frac{d}{2}. \\ (S_2P)^2 - (S_1P)^2 &= \left[D^2 + \left(x + \frac{d}{2} \right)^2 \right] - \left[D^2 + \left(x - \frac{d}{2} \right)^2 \right] \\ \therefore (S_2P)^2 - (S_1P)^2 &= 2xd \\ \text{or } S_2P - S_1P &= \frac{2xd}{S_2P + S_1P} \end{aligned}$$

We can approximate that $S_2P \approx S_1P \approx D$.

$$\therefore \text{Path difference} = S_2P - S_1P = \frac{xd}{D} \quad (6.6)$$

We now find out the conditions for observing bright and dark fringes on the screen.

6.6.3 Bright Fringes

Bright fringes occur wherever the waves from S_1 and S_2 interfere constructively. The first place this occurs is at O, the axial point. There, the waves from S_1 and S_2 travel the same optical path length to O and arrive in phase. The next bright fringe occurs when the wave from S_2 travels one complete wavelength further than the wave from S_1 . In general constructive interference occurs if S_1P and S_2P differ by a whole number of wavelengths.

The condition for finding a bright fringe at P is that

$$S_2P - S_1P = m\lambda$$

Using the equation (6.6), it means that

$$\frac{xd}{D} = m\lambda \quad (6.7)$$

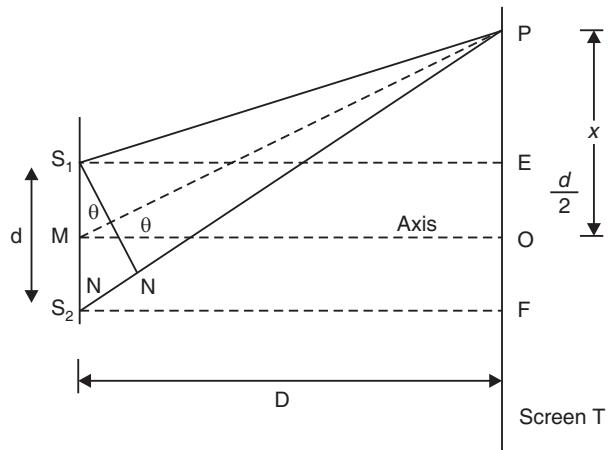


Fig. 6.8

where m is called the **order of the fringe**. The bright fringe at O, corresponding to $m = 0$, is called the *zero-order* fringe. The first-order bright fringe from the axis corresponds to $m = 1$ and the second order bright fringe to $m = 2$ and so on.

6.6.4 Dark Fringes

The first dark fringe occurs when $(S_2P - S_1P)$ is equal to $\lambda / 2$. The waves are now in opposite phase at P. The second dark fringe occurs when $(S_2P - S_1P)$ equals $3\lambda / 2$. The m^{th} dark fringe occurs when

$$(S_2P - S_1P) = (2m + 1) \lambda / 2$$

The condition for finding a dark fringe is

$$\frac{xd}{D} = (2m + 1) \frac{\lambda}{2} \quad (6.8)$$

The first-order dark fringe from the axis corresponds to $m = 1$ and the second order dark fringe to $m = 2$ and so on.

6.6.5 Separation between Neighbouring Bright Fringes

The m^{th} order bright fringe occurs when

$$x_m = \frac{m\lambda D}{d}$$

and the $(m + 1)^{\text{th}}$ order bright fringe occurs when

$$x_{m+1} = \frac{(m + 1)\lambda D}{d}$$

The bright fringe separation, β is given by

$$\beta = x_{m+1} - x_m = \frac{\lambda D}{d} \quad (6.9)$$

The same result will be obtained for dark fringes. Thus, neighbouring bright and dark fringes are separated by the same amount everywhere on the screen. The separation β is called the **fringe width**.

The width of the dark or bright fringe is given by equ.(6.9).

$$\beta = \frac{\lambda D}{d}$$

where $D (= a + b)$ is the distance of the sources from the eye-piece.

Example 6.1: A biprism is placed 5 cm from a slit illuminated by sodium light ($\lambda = 5890 \text{ \AA}$). The fringe width obtained on the screen is 0.9424 mm. The screen is at a distance of 75 cm from the biprism. Find the distance between the two coherent sources.

Solution:

$$\beta = \frac{\lambda D}{d}$$

$$\therefore 9.424 \times 10^{-2} \text{ cm} = \frac{5890 \times 10^{-8} \text{ cm} (5 + 75) \text{ cm}}{d}$$

$$\text{or } d = \frac{5890 \times 10^{-8} \times 80}{9.424 \times 10^{-2}} \text{ cm} = 0.05 \text{ cm.}$$

6.6.6 Determination of Wavelength of Light

The wavelength of the light can be determined using the equ.(6.9). For using the relation, the values of β , D and d are to be measured. These measurements are done as follows.

Adjustments

A narrow adjustable slit S , the biprism, and a micrometer eyepiece are mounted on the uprights and are adjusted to be at the same height and in a straight line. The slit is made vertical and parallel to the refracting edge of the biprism by rotating it in its own plane. It is illuminated with the light from the monochromatic source. The biprism is moved along the optical bench till, on looking through it along the axis of the optical bench, two equally bright vertical slit images are seen. Then the eyepiece is moved till the fringes appear in the focal plane of the eyepiece.

- (i) **Determination of fringe width β :** When the fringes are observed in the field view of the eyepiece, the vertical cross-wire is made to coincide with the centre of one of the bright fringes. The position of the eyepiece is read on the scale. The micrometer screw of the eyepiece is moved slowly and the number of the bright fringes, N , that pass across the cross-wire is counted. The position of the cross-wire is again read. The fringe width is then given by

$$\beta = \frac{x_{m+N} - x_m}{N}$$

- (ii) **Determination of ' d ':** A convex lens of short focal length is placed between the slit and the eyepiece without disturbing their positions. The lens is moved back and forth near the biprism till a sharp pair of images of the slit is obtained in the field of view of the eyepiece (Fig. 6.9 a). The distance between the images is measured. Let it be denoted by d_1 .

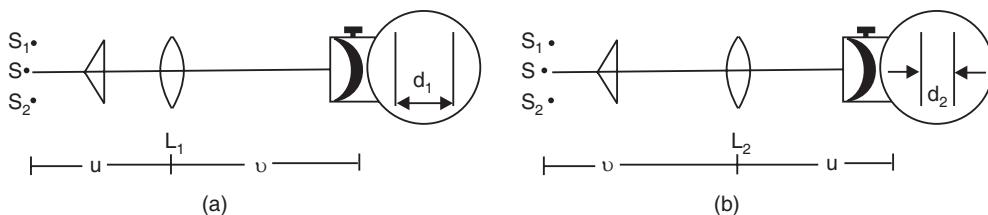


Fig. 6.9

If u is the distance of the slit and v that of the eyepiece from the lens (Fig. 6.9), then the magnification is

$$\frac{v}{u} = \frac{d_1}{d} \quad (6.10)$$

The lens is then moved to a position nearer to the eyepiece, where again a pair of images of the slit is seen (Fig. 6.9b). The distance between the two sharp images is again measured. Let it be d_2 . Again magnification is given by

$$\frac{u}{v} = \frac{d_2}{d} \quad (6.11)$$

Note that the magnification in one position is the reciprocal of the magnification in the other position.

Multiplying the equations (6.10) and (6.11), we obtain

$$\begin{aligned} \frac{d_1 d_2}{d^2} &= 1 \\ d &= \sqrt{d_1 d_2} \end{aligned} \quad (6.12)$$

Using the values of β , d and D in the equation (6.9), the wavelength λ can be computed.

Example 6.2: Fresnel biprism is used to form interference fringes using sodium light of $\lambda = 5890 \text{ \AA}$. A convex lens interposed gives two images separated by distances 0.6 mm and 0.27 mm of the coherent sources corresponding to two positions. Calculate the fringe width. The slit to screen distance is 80 cm.

Solution: Let d_1 and d_2 be the separation between the images in two positions. Then, the distance between the coherent sources is

$$d = \sqrt{d_1 d_2} = \sqrt{6 \times 10^{-4} \text{ m} \times 2.7 \times 10^{-4} \text{ m}} = 4.025 \times 10^{-4} \text{ m}$$

$$\therefore \beta = \frac{\lambda D}{d} = \frac{5890 \times 10^{-10} \text{ m} \times 0.8 \text{ m}}{4.025 \times 10^{-4} \text{ m}} = 1.171 \text{ mm.}$$

6.6.7 Interference Fringes with White Light

In the biprism experiment if the slit is illuminated by white light, the interference pattern consists of a central **white fringe** flanked on both its sides by a few coloured fringes; and general illumination beyond the fringes. The central white fringe is the *zero-order fringe*.

With monochromatic light all the bright fringes are of the same colour and it is not possible to locate the zero-order fringe. Therefore, in order to locate the zero order fringe the biprism is to be illuminated by white light.

6.6.8 Lateral Displacement of Fringes

The biprism experiment can be used to determine the thickness of a given thin sheet of transparent material such as glass or mica. If a thin transparent sheet is introduced in the path of one of the two interfering beams, the fringe system gets displaced towards the beam in whose path the sheet is introduced. By measuring the amount of displacement, the thickness of the sheet can be determined.

Suppose S_1 and S_2 are the virtual coherent monochromatic sources. The point O is equidistant from S_1 and S_2 , where we obtain the *central bright fringe*. Therefore, the optical path $S_1O = S_2O$. Let a transparent plate G of thickness t and refractive index μ be introduced in the path of one of the beams (see Fig. 6.10). The optical path lengths S_1O and S_2O are now not equal and the central bright fringe shifts to P from O. The light waves from S_1 to P travel partly in air and partly in the sheet G; the distance travelled in air is $(S_1P - t)$ and that in the sheet is t .

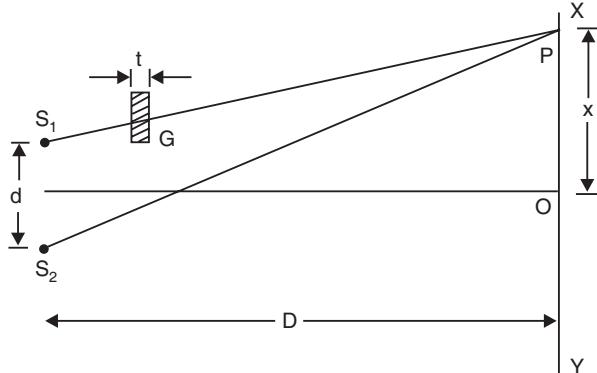


Fig. 6.10

$$\text{The optical path } \Delta_{S_1P} = (S_1P - t) + \mu t = S_1P + (\mu - 1)t$$

$$\text{The optical path } \Delta_{S_2P} = S_2P$$

The optical path difference at P is $\Delta_{S_1P} - \Delta_{S_2P} = 0$, since in the presence of the thin sheet, the optical path lengths S_1P and S_2P are equal and central zero fringe is obtained at P.

$$\therefore \Delta_{S_1P} = \Delta_{S_2P}$$

$$[S_1P + (\mu - 1)t] = S_2P$$

$$\therefore S_2P - S_1P = (\mu - 1)t$$

But according to the relation (6.7),

$$S_2 P - S_1 P = \frac{xd}{D}$$

where x is the **lateral shift** of the central fringe due to the introduction of the thin sheet.

$$\therefore (\mu - 1)t = \frac{xd}{D}$$

Hence, the thickness of the sheet is

$$t = \frac{xd}{D(\mu - 1)} \quad (6.13)$$

Example 6.3: A thin sheet of glass ($\mu = 1.5$) of thickness $6 \mu\text{m}$ is introduced in one of the paths of a biprism. It shifts the central fringe by 5 bands. Find the wavelength of light.

Solution: The displacement of fringes is given by $(\mu - 1)t = m\lambda$. Therefore,

$$\lambda = \frac{(\mu - 1)t}{m} = \frac{(1.5 - 1) \times 6 \times 10^{-6} \text{ m}}{5} = 600 \text{ nm.}$$

INTERFERENCE BY DIVISION OF WAVE AMPLITUDE

6.7 THIN FILM INTERFERENCE

An optical medium is called a **thin film** when its thickness is about the order of 1 wavelength of light in visible region. Thus, a film of thickness in the range $0.5 \mu\text{m}$ to $10 \mu\text{m}$ may be considered as a thin film. A thin film may be a thin sheet of transparent material such as glass, mica, an air film enclosed between two transparent plates or a soap bubble. When light is incident on such a film, a small part of it gets reflected from the top surface and a major part is transmitted into the film. Again, a small part of the transmitted component is reflected back into the film by the bottom surface and the rest of it emerges out of the film. A small portion of the light thus gets reflected partially several times in succession within the film (see Fig.6.11).

In transparent thin films, the two bounding surfaces strongly transmit light and only weakly reflect the incident light. Therefore, only the first reflection at the top surface and the first reflection at the bottom surface will be of appreciable strength. For example, if we consider a glass plate, having a refractive index 1.52, the reflectivity of the top surface is given by

$$r = \left[\frac{1.52 - 1}{1.52 + 1} \right]^2 = 0.042 \quad (6.14)$$

It means that about 4% of the incident light is reflected by the top surface of the glass plate, while 96% of it is transmitted into the plate. Out of the light reaching the bottom surface, again 3.8% is reflected and 92% is transmitted out of the plate. Then, again out of the 3.8% of the light 0.15% is reflected at the inner boundary of the top surface and about 3.65%

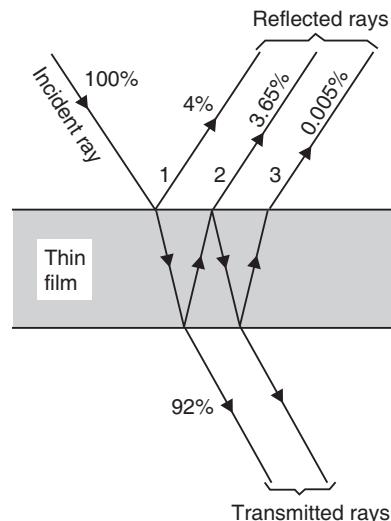


Fig. 6.11

is transmitted out into the air. After two reflections, the intensity will become insignificantly small. At each reflection, the intensity and hence the *amplitude of light wave* is divided into a reflected component and a refracted component. The reflected and refracted components travel along different paths and can be brought to overlap to produce interference. Therefore, the interference in thin films is called interference by division of amplitude. Newton and Robert Hooke first observed the thin film interference. However, Thomas Young gave the correct explanation of the phenomena. A thin film may be uniform or non-uniform in its structure. However, as long as its thickness lies within the specified limits, interference of light occurs.

6.8 PLANE PARALLEL FILM

A transparent thin film of uniform thickness bounded by two parallel surfaces is known as a *plane parallel thin film*.

When light is incident on a parallel thin film, a small portion of it gets reflected from the top surface and a major portion is transmitted into the film. Again, a small part of the transmitted component is reflected back into the film by the bottom surface and the rest of it is transmitted from the lower surface of the film. Thin films transmit incident light strongly and reflect only weakly. After two reflections, the intensities of reflected rays drop to a negligible strength. Therefore, we consider the first two reflected rays only. These two rays are derived from the same incident ray but appear to come from two sources located below the film. The sources are virtual coherent sources (see Fig.6.12). The reflected waves 1 and 2 travel along parallel paths and interfere at infinity. This is a case of *two-beam interference*.

The condition for maxima and minima can be deduced once we have calculated the optical path difference between the two rays at the point of their meeting.

6.8.1 Interference Due to Reflected Light

Let us consider a transparent film of uniform thickness ' t ' bounded by two parallel surfaces as shown in Fig.6.13. Let the refractive index of the material be μ . The film is surrounded by air on both the sides. Let us consider plane waves from a monochromatic source falling on the thin film at an angle of incidence ' i '. Part of a ray such as AB is reflected along BC, and part of it is transmitted into the film along BF. The transmitted ray BF makes an angle ' r ' with the normal to the surface at the point B. The ray BF is in turn partly reflected back into the film along FD while a major part refracts into the surrounding medium along FK. Part of the reflected ray FD is transmitted at the upper surface and travels along DE. Since the film

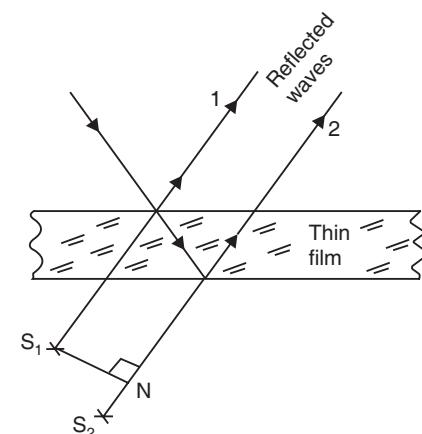


Fig. 6.12

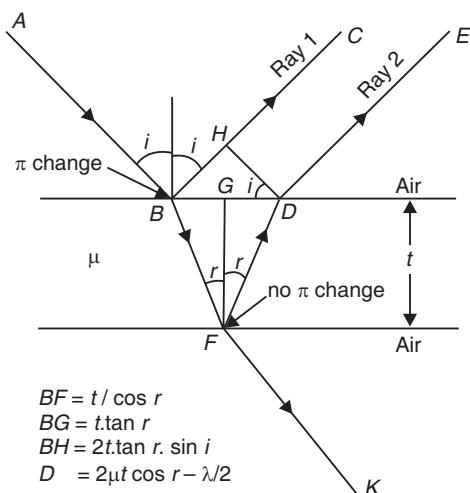


Fig. 6.13

$$\begin{aligned}BF &= t / \cos r \\ BG &= t \tan r \\ BH &= 2t \tan r \sin i \\ D &= 2\mu t \cos r - \lambda/2\end{aligned}$$

boundaries are parallel, the reflected rays BC and DE will be parallel to each other. The waves travelling along the paths BC and BFDE are derived from a single incident wave AB. Therefore they are coherent and can produce interference if they are made to overlap by a condensing lens or the eye.

(i) **Geometrical Path Difference:** Let DH be normal to BC. From points H and D onwards, the rays HC and DE travel equal path. The ray BH travels in air while the ray BD travels in the film of refractive index μ along the path BF and FD. The geometric path difference between the two rays is

$$BF + FD - BH.$$

(ii) **Optical Path Difference:**

$$\text{Optical path difference } \Delta_a = \mu L \\ \therefore \Delta_a = \mu (BF + FD) - 1(BH) \quad (6.15)$$

In the $\triangle BFD$, $\angle BFG = \angle GFD = \angle r$

$$\begin{aligned} BF &= FD \\ BF &= \frac{FG}{\cos r} = \frac{t}{\cos r} \\ \therefore BF + FD &= \frac{2t}{\cos r} \end{aligned} \quad (6.16)$$

Also,

$$\begin{aligned} BG &= GD \\ BD &= 2BG \\ BD &= 2t \tan r = t \tan r \\ \therefore BD &= 2t \tan r \\ \text{In the } \Delta^{\text{le}} BHD & \angle HBD = (90 - i) \\ \angle BHD &= 90^\circ \\ \therefore \angle BDH &= i \\ \therefore BH &= BD \sin i = 2t \tan r \sin i \end{aligned} \quad (6.17)$$

From Snell's law,

$$\begin{aligned} \sin i &= \mu \sin r \\ \therefore BH &= 2t \tan r (\mu \sin r) = \frac{2\mu t \sin^2 r}{\cos r} \end{aligned} \quad (6.18)$$

Using the equations (6.17) and (6.16) into equ.(6.15), we get

$$\begin{aligned} \Delta_a &= \mu \left[\frac{2t}{\cos r} \right] - \left[\frac{2\mu t \sin^2 r}{\cos r} \right] \\ &= \frac{2\mu t}{\cos r} [1 - \sin^2 r] \\ &= \frac{2\mu t}{\cos r} \cos^2 r \\ \therefore \Delta_a &= 2\mu t \cos r \end{aligned} \quad (6.19)$$

(iii) **Correction on account of phase change at reflection:** When a ray is reflected at the boundary of a rarer to denser medium, a path-change of $\lambda/2$ occurs for the ray BC (see Fig.6.13). There is no path difference due to transmission at D. Including the change in path difference due to reflection in eqn. (6.19), the true path difference is given by

$$\Delta_t = 2\mu t \cos r - \lambda/2 \quad (6.20)$$

6.8.2 Conditions for Maxima (Brightness) and Minima (Darkness)

Maxima occur when the optical path difference $\Delta = m \lambda$. If the difference in the optical path between the two rays is equal to an *integral number of full waves*, then the rays meet each other in phase. The crests of one wave falls on the crests of the others and the waves *interfere constructively*. Thus, when

$$2\mu t \cos r - \frac{\lambda}{2} = m\lambda \quad (6.21)$$

the reflected rays undergo constructive interference to produce brightness or maxima at the point of their meeting.

$$2\mu t \cos r = m\lambda + \lambda/2$$

or $2\mu t \cos r = (2m+1)\lambda/2 \quad \text{Condition for Brightness}$ (6.22)

Minima occur when the optical path difference is $\Delta = (2m+1)\lambda/2$. If the difference in the optical path between the two rays is equal to an *odd integral number of half-waves*, then the rays meet each other in opposite phase. The crests of one wave falls on the troughs of the others and the waves *interfere destructively*. Thus, when

$$2\mu t \cos r - \lambda/2 = (2m+1)\lambda/2 \quad (6.23)$$

the reflected rays undergo destructive interference to produce darkness. Equ.(6.23) may be rewritten as

$$2\mu t \cos r = (m+1)\lambda$$

The phase relationship of the interfering waves does not change if one full wave is added to or subtracted from any of the interfering waves. Therefore $(m+1)\lambda$ can be as well replaced by $m\lambda$ for simplicity in expression. Thus,

$$2\mu t \cos r = m\lambda \quad \text{Condition for Darkness} \quad (6.24)$$

6.8.3 Some Important Points

- (a) It is seen that the conditions of interference depend on three parameters, namely μt , λ and r . In the case of constant thickness (parallel) film, μt is constant. When a parallel beam of light is incident on such a film, r also remains constant. Then the interference conditions solely depend on the wavelength, λ .
- (b) When a parallel beam of monochromatic light is incident normal to the film, the whole film will appear uniformly dark or uniformly bright.

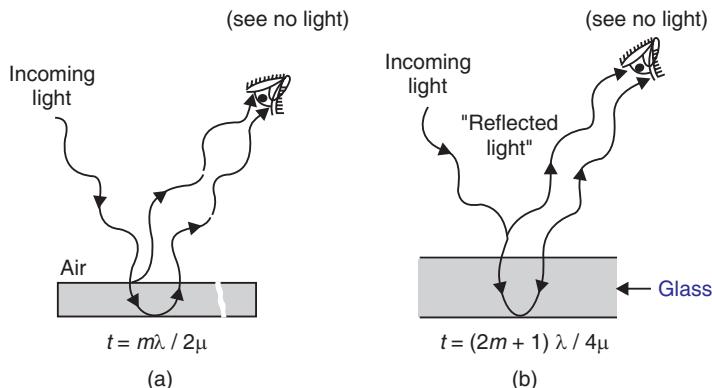


Fig. 6.14: Interference in film of constant thickness – (a) Interference is destructive and the film appears dark in reflected light for thickness satisfying $t = m\lambda/2\mu$ condition. (b) It is constructive and the film appears bright for film thickness satisfying the condition $t = (2m+1)\lambda/4\mu$.

The film will appear bright in reflected light, when the film is of $\lambda/4\mu$, $3\lambda/4\mu$, $5\lambda/4\mu$, thick and it appears dark when its thickness is $\lambda/2\mu$, λ/μ , $3\lambda/2\mu$, etc. (Fig.6.14). If the condition of constructive interference is satisfied, the film will show intense colour corresponding to the colour of the incident light.

- (c) A change in the angle of incidence of the rays leads to a change in the path difference. Consequently, if the inclination of the film with respect to the light beam is changed gradually, we find that it will appear dark and bright (or bright and dark) in succession.
- (d) If a parallel beam of white light falls on a parallel film, those wavelengths for which the path difference is $m\lambda$, will be absent from the reflected light. The other colours will be reflected. Therefore, the film will appear uniformly coloured with one colour being absent.

Example 6.4: A soap film of $5 \times 10^{-7}\text{m}$ thick is viewed at an angle of 35° to the normal. Find the wavelengths of light in the visible spectrum, which will be absent from the reflected light. Given the refractive index of the film = 1.33.

Solution: White light is incident on the film at an angle 30° . Let r be the angle of refraction of light into the film. r can be calculated from Snell's law $\mu = \frac{\sin i}{\sin r}$.

$$\therefore \sin r = \frac{\sin 30^\circ}{1.33} \quad \text{or} \quad r = 25.55^\circ \quad \text{and} \quad \cos r = 0.90.$$

The absence of certain wavelengths in reflected light is due to their undergoing destructive interference. The condition for destructive interference is

$$2\mu t \cos r = m\lambda$$

To find out the missing wavelengths, we have to use different m values into the above equation and find out which of them lie in the visible region 7000 to 4000 Å.

$$\text{For } m = 1, \text{ we get } \lambda_1 = 2 \times 1.33 \times 5 \times 10^{-7}\text{m} \times 0.90 = 12 \times 10^{-7}\text{m} = 12000 \text{ Å.}$$

$$\text{For } m = 2, \text{ we get } \lambda_2 = (2 \times 1.33 \times 5 \times 10^{-7}\text{m} \times 0.90) \div 2 = 6 \times 10^{-7}\text{m} = 6000 \text{ Å.}$$

$$\text{For } m = 3, \text{ we get } \lambda_3 = (2 \times 1.33 \times 5 \times 10^{-7}\text{m} \times 0.90) \div 3 = 4 \times 10^{-7}\text{m} = 4000 \text{ Å.}$$

$$\text{For } m = 4, \text{ we get } \lambda_4 = (2 \times 1.33 \times 5 \times 10^{-7}\text{m} \times 0.90) \div 4 = 3 \times 10^{-7}\text{m} = 3000 \text{ Å.}$$

It is clear that the first wavelength lies in the infra red and the last one lies in the UV region. The middle two wavelengths lie in the visible region. Hence, the absent wavelengths in the reflected light are 6000 Å and 4000 Å.

6.8.4 Restriction on Thickness of the Film

We know that interference colours are observed only in thin films but not in thick plates such as windowpanes or glass slabs. This is due to the fact that light waves can interfere only when both the conditions of temporal and spatial coherence are satisfied. In Fig. 6.14, we have assumed that a monochromatic wave of infinite length is incident on the film. In reality, the incident light consists of wave trains of finite length and coherence extends over the length of each wave train only. Interference can occur only when parts of the same group of wave trains overlap. Superposition of different wave trains cannot produce interference because they will be incoherent and do not maintain any constant phase relationship with each other.

Fig. 6.15 shows the real situation. Wave trains 1,2,3 of finite length are incident in succession on a thin film. Portions of each wave train are reflected by the top and bottom surfaces of the film. Each wave train is divided into two reflected wave trains (U_1, L_1, U_2, L_2 and U_3, L_3). In Fig. 6.15 (a) the film is thin and the difference in the optical path lengths of U_1 and L_1 is small compared to the length of the wave train. Their superposition produces interference, as U_1 and L_1 are parts of the same wave train 1 and hence are coherent. In Fig. 6.15 (b) the film is thicker and the optical path difference between U_1 and L_1 is large than the coherence length. Consequently, superposition takes place between parts of different wave trains, U_2 and L_1 and U_3 and L_2 . Therefore interference does not take place.

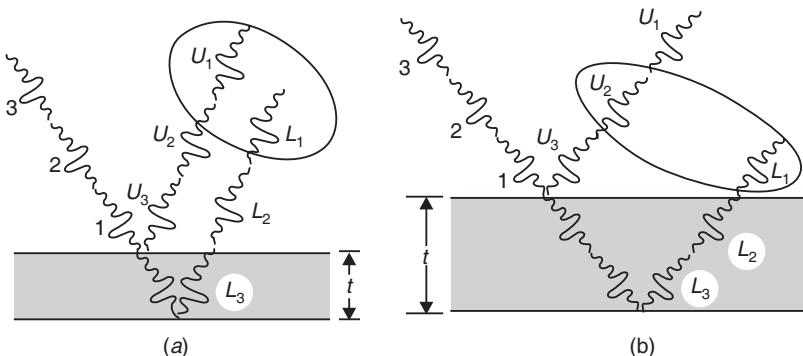


Fig. 6.15: Role of the thickness of the film—(a) when the film thickness is smaller than coherence length, superposition of reflected parts of the same wave train occurs leading to inference. (b) In thick films, different wave trains which are not coherent superpose and interference does not arise.

It implies that interference occurs *only when* the optical path difference, Δ , between the superposing waves is less than the coherence length.

$$\text{i.e.,} \quad \Delta \ll l_{\text{coh}} \quad (6.25)$$

$$\therefore \quad (2\mu t \cos r - \lambda/2) \ll l_{\text{coh}} \quad (6.26)$$

But

$$l_{\text{coh}} = \frac{\lambda^2}{\Delta\lambda}$$

$$\therefore \quad (2\mu t \cos r - \lambda/2) < \lambda^2/\Delta\lambda$$

Rearranging the terms, we obtain

$$t < \frac{\lambda \left[\frac{\lambda}{\Delta\lambda} + \frac{1}{2} \right]}{2\mu \cos r} \quad (6.27)$$

$\lambda/\Delta\lambda \gg 1/2$ and for normal incidence $\cos r = 1$.

$$\therefore \quad t < \frac{\lambda^2}{2\mu \Delta\lambda} \quad (6.28)$$

The above equation indicates that interference in thin film will be observed if the thickness of the film is less than the coherence length of the incident light waves. Normally, the coherence length of the light from ordinary sources is of the order of a fraction of a millimeter. Therefore, interference is seen with the films of thickness of the order of a few hundred microns only. It is because of this reason that thick films do not exhibit interference.

6.9 VARIABLE THICKNESS (WEDGE-SHAPED) FILM

A wedge is a thin film of varying thickness having a zero thickness at one end and progressively increasing to a particular thickness at the other end. A thin wedge of air film can be formed by two glass slides resting on each other at one edge and separated by a thin spacer at the opposite edge.

The arrangement for observing the interference pattern in a wedge shaped air-film is shown in Fig.6.16. If a parallel beam of monochromatic light illuminates the wedge from above, the rays reflected from its two bounding surfaces will not be parallel. They appear to diverge from a point near the film. The path difference between the rays reflected from the

upper and lower surfaces of the air film varies along its length due to variation in thickness. Therefore, alternate bright and dark fringes are observed on its top surface (Fig.6.17). The fringes are localized at the top surface of the film.

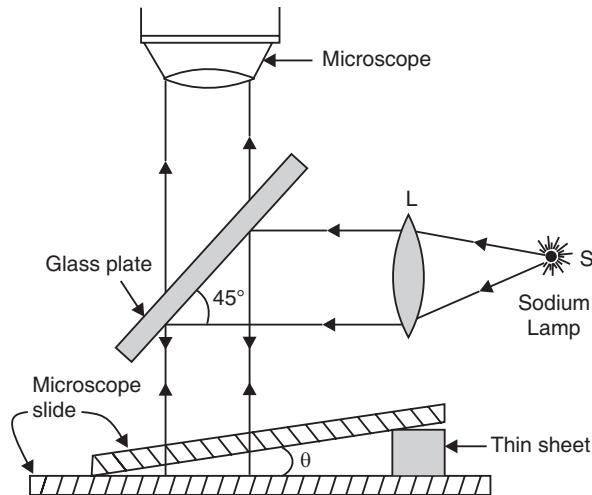


Fig. 6.16

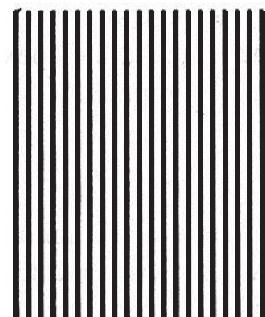


Fig. 6.17

When the light is incident on the wedge from above, it gets partly reflected from the glass-to-air boundary at the top of the air film. Part of the light is transmitted through the air film and gets reflected partly at the air-to-glass boundary, as shown in Fig.6.18. The two rays BC and FE, thus reflected from the top and bottom of the air film, are coherent as they are derived from the same ray AB through *division of amplitude*. The rays are close enough if the thickness of the film is of the order of a wavelength of light. For small film thickness the rays interfere producing darkness or brightness depending on the phase difference. The thickness of the glass plates is large compared with the wavelength of the incident light. Hence, the observed interference effects are entirely due to the wedge-shaped air film.

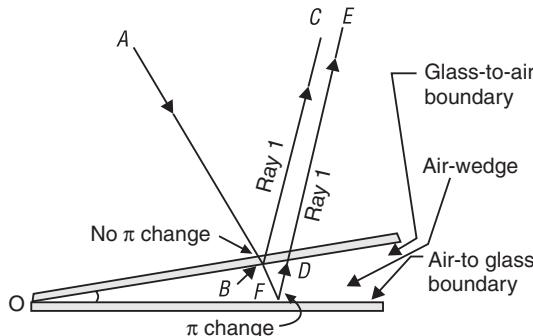


Fig. 6.18

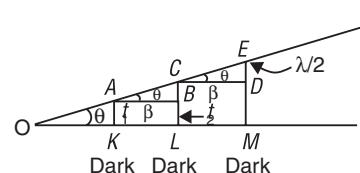


Fig. 6.19

The optical difference between the two rays BC and FE is given by

$$\Delta = 2\mu t \cos r - \lambda/2$$

where $\lambda/2$ takes account the gain of half-wave due to the abrupt jump of π radians in the phase of the wave reflected from the bottom boundary of air – to – glass.

Maxima occur when the optical path difference $\Delta = m \lambda$. If the difference in the optical path between the two rays is equal to an *integral number of full waves*, then the rays meet each other in phase. The crests of one wave falls on the crests of the others and the waves *interfere constructively*. This needs that

$$2\mu t \cos r = (2m + 1)\lambda/2$$

Minima occur when the optical path difference is $\Delta = (2m + 1)\lambda/2$. If the difference in the optical path between the two rays is equal to an *odd integral number of half-waves*, then the rays meet each other in opposite phase. The crests of one wave fall on the troughs of the others and the waves *interfere destructively*. It needs that

$$2\mu t \cos r = m\lambda$$

Referring to Fig.6.19, let us say a dark fringe occurs at A where the relation

$$2\mu t \cos r = m\lambda$$

is satisfied. If normal incidence is assumed, $\cos r = 1$ and if the thickness of air film at A is denoted by t_1 , then at A

$$2\mu t_1 = m\lambda \quad (6.29)$$

The next dark fringe will occur, say, at C where the thickness $CL = t_2$. Then at C

$$2\mu t_2 = (m + 1)\lambda \quad (6.30)$$

Subtracting equ. (6.29) from equ. (6.30), we get

$$2\mu(t_2 - t_1) = \lambda \quad (6.31)$$

But $(t_2 - t_1) = BC$

$$\therefore 2\mu(BC) = \lambda$$

or $BC = \frac{\lambda}{2\mu}$ (6.32)

From the $\Delta^{\text{le}}ABC$, $\angle CAB = \theta$ and $BC = AB \tan \theta$

$$\therefore (AB) \tan \theta = \frac{\lambda}{2\mu} \quad (6.33)$$

AB is the distance between successive dark fringes and it also equals the separation of the successive bright fringes. It is, therefore, called the **fringe width**, β . That is $AB = \beta$. We may write equ. (6.33) as

$$\therefore \beta = \frac{\lambda}{2\mu \tan \theta} \quad (6.34)$$

For small values of θ , $\tan \theta \approx \theta$.

$$\therefore \beta = \frac{\lambda}{2\mu \theta} \quad (6.35)$$

According to the relation (6.35), an increase in the angle θ makes the fringes move closer. At an angle $\theta \approx 1^\circ$, the interference pattern vanishes. On the other hand, if θ is gradually decreased, the fringe separation increases and ultimately the fringes disappear since the faces of the film become parallel at $\theta = 0^\circ$.

Example 6.5: *Fringes of equal thickness are observed in a thin glass wedge of refractive index 1.52. The fringe spacing is 0.1 mm, wavelength of light being 5893 Å. Calculate the wedge angle.*

Solution: The fringe width $\beta = \lambda/2\mu\theta$.

$$\therefore \theta = \frac{\lambda}{2\mu\beta} = \frac{5893 \times 10^{-10} \text{ m}}{2 \times 1.52 \times 10^{-4} \text{ m}} = 1.938 \text{ rad} = 0.11^\circ.$$

6.9.1 Salient Features of the Interference Pattern

- (i) Fringe at the apex is dark.
- (ii) Fringes are straight and parallel.
- (iii) Fringes are equidistant.
- (iv) Fringes of equal thickness
- (v) Fringes are localized.

(i) Fringe at the apex is dark: At the apex, the two glass slides are in contact with each other. Therefore, the thickness of the air film at the contact edge is negligible ($t \approx 0$). The optical path difference there becomes

$$\Delta = 2\mu t - \lambda/2 = 0 - \lambda/2 = -\lambda/2 \quad (6.36)$$

It implies that a path difference of $\lambda/2$ or a phase difference of π occurs between the reflected waves at the edge. The two waves interfere destructively. Therefore, the fringe at the apex is always dark (Fig. 6.20).

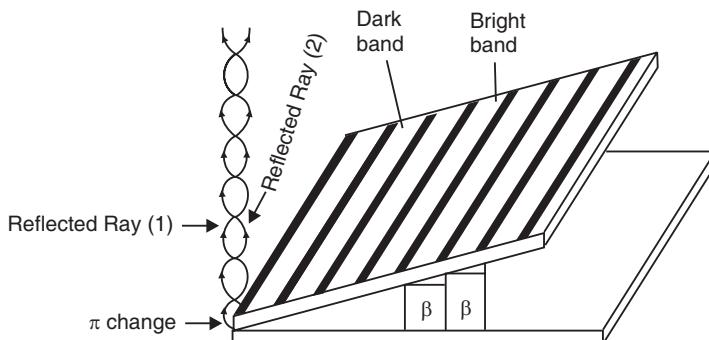


Fig. 6.20: At the contact edge, the reflected rays (1) and (2) are 180° out of phase and produce a dark band.

(ii) Straight and parallel fringes: Each fringe in the pattern is produced by the interference of rays reflected from sections of the wedge having the same thickness. The locus of points having the same thickness lies along lines parallel to the contact edge. Therefore, the fringes are straight. Since the fringes are equidistant, they will be parallel.

(iii) Equidistant fringes: The fringe width β is given by

$$\beta \approx \lambda/2\theta \quad (6.37)$$

where λ is the wavelength of the incident light and θ is the angle of the wedge. As the quantities λ and θ are constants, β is constant for a given wedge angle. Therefore, the fringes are equidistant.

(iv) Fringes of equal thickness: As each bright or dark fringe is a locus of constant film thickness, the fringes are called fringes of equal thickness.

(v) Localized fringes: The fringes are very close to the top surface of the air wedge and can be seen with a microscope.

6.9.2 Determination of the Wedge Angle

The wedge angle θ can be experimentally determined with the help of a travelling microscope. Using the microscope the positions of dark fringes at two distant points Q and R are noted (Fig. 6.21). Let the distance OQ be x_1 and OR be x_2 . Let the thickness of the wedge be t_1 at Q and t_2 at R.

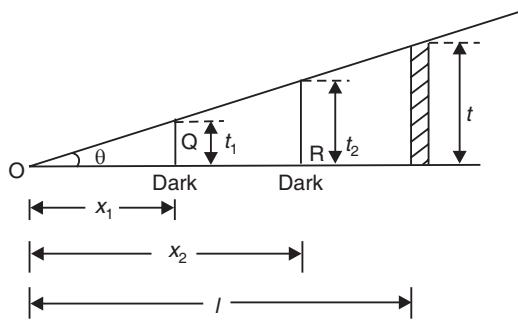


Fig. 6.21

The dark fringe at Q is given by

$$2\mu t_1 = m\lambda \quad (6.38)$$

But as θ is very small, we can write

$$\begin{aligned} t_1 &= x_1 \tan \theta \approx x_1 \theta \\ \therefore 2\mu x_1 \theta &= m\lambda \end{aligned} \quad (6.39)$$

We can write similarly for the dark fringe at R as

$$2\mu x_2 \theta = (m + N)\lambda \quad (6.40)$$

where N is the number of dark fringes lying between the positions Q and R. Subtracting equ. (6.39) from equ.(6.40), we get

$$\begin{aligned} 2\mu(x_2 - x_1)\theta &= N\lambda \\ \therefore \theta &= \frac{N\lambda}{2\mu(x_2 - x_1)} \end{aligned} \quad (6.41)$$

In case of air $\mu = 1$ and the above relation reduces to

$$\theta = \frac{N\lambda}{2(x_2 - x_1)} \quad (6.42)$$

6.9.3 Determination of the Thickness of the Spacer

The thickness of the spacer used to form the wedge shaped air film between the glass slides can be determined from the above measurements. If 't' is the thickness of the spacer (foil or wire) used, we can write

$$t = l \tan \theta \approx l \theta \quad (6.43)$$

where l is the length of the air wedge.

$$\therefore t = \frac{l N \lambda}{2(x_2 - x_1)} \quad (6.44)$$

6.9.4 Number of Dark Fringes in an Air Wedge

The fringe width in air wedge is given by $\beta = \frac{\lambda}{2\theta}$

The wedge angle is given by $\theta = \frac{t}{l}$

$$\therefore \beta = \frac{\lambda l}{2t} \quad (6.45)$$

The number of dark fringes observed, N, is related to l , as

$$l = N\beta \quad (6.46)$$

$$\therefore \beta = \frac{\lambda N \beta}{2t}$$

$$\therefore N = \frac{2t}{\lambda} \quad (6.47)$$

6.10 COLOURS IN THIN FILMS

The colours exhibited in reflection by thin films of oil, mica, soap bubbles and coatings of oxides on heated metals etc are due to interference of light from an extended source such as sky. Thomas Young explained the origin of colours in thin films. It may be understood as follows. The films are usually observed by reflected light. The eye looking at the thin film receives light waves reflected from the top and bottom surfaces of the film. The reflected rays are very close to each other and are in a position to interfere. The optical path difference between the interfering rays is $\Delta = 2\mu t \cos r - \lambda/2$. It is seen that the path difference depends upon the thickness t of the film, the wavelength λ and the angle r , which is related to the angle of incidence of light on the film. White light consists of a range of wavelengths and for specific values of t and r , waves of only certain wavelengths (colours) constructively interfere. Therefore, only those colours are present in the reflected light. The other wavelengths interfere destructively and hence are absent from the reflected light. Hence, the film at a particular point appears coloured. As the thickness and the angle of incidence vary from point to point, different colours are intensified at different places. The colours seen are not isolated colours, as at each place there is a mixture of colours. The composition of colours is different at different places and contours of impressive hues are observed over the entire surface of the film.

6.11 NEWTON'S RINGS

Newton's rings are another example of fringes of equal thickness. Newton's rings are formed when a plano-convex lens L of a large radius of curvature placed on a sheet of plane glass AB. The combination forms a thin circular air film of variable thickness in all directions around the point of contact of the lens and the glass plate. The locus of all points corresponding to specific thickness of air film falls on a circle whose centre is at O. Consequently, interference fringes are observed in the form of a series of concentric rings with their centre at O (Fig. 6.22). Newton originally observed these concentric circular fringes and hence they are called Newton's rings.

The experimental arrangement for observing Newton's rings is shown in Fig. 6.23.

Monochromatic light from an extended source S is rendered parallel by a lens L' . It is incident on a glass plate inclined at 45° to the horizontal, and is reflected normally down onto a plano-convex lens placed on a flat glass plate. Part of the light incident on the system is reflected from the glass-to-air boundary, say from point D (Fig. 6.24). The remainder of the light is transmitted through the air film. It is again reflected from the air-to-glass boundary, say from point J. The two rays reflected from the top and bottom of the air film are derived through division of amplitude from the same incident ray CD and are therefore coherent. The rays 1 and 2 are close to each other and interfere to produce darkness or brightness. The condition of brightness or darkness depends on the path difference between the two reflected light rays, which in turn depends on the thickness of the air film at the point of incidence.

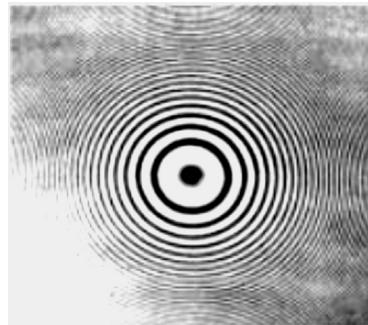


Fig. 6.22

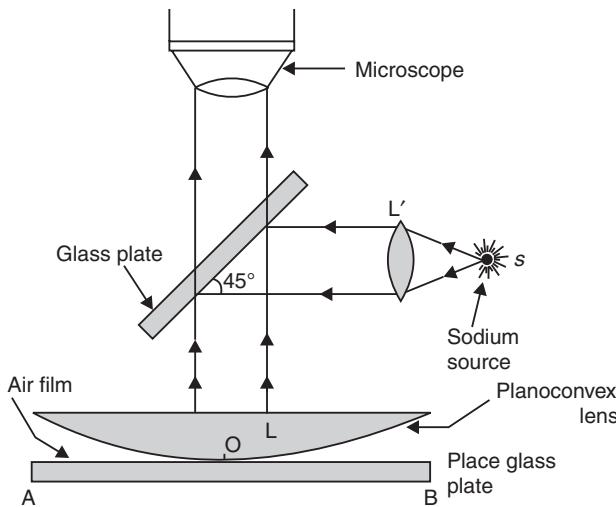


Fig. 6.23

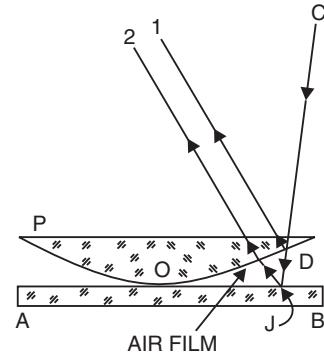


Fig. 6.24

6.11.1 Condition for Bright and Dark Rings

The optical path difference between the rays is given by $\Delta = 2\mu t \cos r - \lambda/2$. Since $\mu = 1$ for air and $\cos r = 1$ for normal incidence of light,

$$\Delta = 2t - \lambda/2 \quad (6.48)$$

Intensity maxima occur when the optical path difference $\Delta = m\lambda$. If the difference in the optical path between the two rays is equal to an *integral number of full waves*, then the rays meet each other in phase. The crests of one wave falls on the crests of the others and the waves *interfere constructively*. Thus, if $2t - \lambda/2 = m\lambda$

$$2t = (2m + 1)\lambda/2 \quad (6.49)$$

bright fringe is obtained.

Intensity minima occur when the optical path difference is $\Delta = (2m + 1)\lambda/2$. If the difference in the optical path between the two rays is equal to an *odd integral number of half-waves*, then the rays meet each other in opposite phase. The crests of one wave fall on the troughs of the other and the waves *interfere destructively*.

Hence, if

$$\begin{aligned} 2t - \lambda/2 &= (2m + 1)\lambda/2 \\ 2t &= m\lambda \end{aligned} \quad (6.50)$$

and dark fringe is produced.

6.11.2 Circular Fringes

In Newton's ring arrangement, a thin air film is enclosed between a plano-convex lens and a glass plate. The thickness of the air film at the point of contact is zero and gradually increases as we move outward. The locus of points where the air film has the same thickness then fall on a circle whose centre is the point of contact. Thus, the thickness of air film is constant at points on any circle having the point of lens-glass plate contact as the centre. The fringes are therefore circular.

6.11.3 Radii of Dark Fringes

Let R be the radius of curvature of the lens (Fig. 6.25). Let a dark fringe be located at Q. Let the thickness of the air film at Q be PQ = t. Let the radius of the circular fringe at Q be OQ = \$r_m\$. By the Pythagoras theorem,

$$PM^2 = PN^2 + MN^2$$

$$\therefore R^2 = r_m^2 + (R - t)^2$$

$$\therefore r_m^2 = 2Rt - t^2 \quad (6.51)$$

As $R \gg t$, $2Rt \gg t^2$.

$$\therefore r_m^2 \approx 2Rt \quad (6.52)$$

The condition for darkness at Q is that

$$2t = m\lambda$$

$$\therefore r_m^2 \approx m\lambda R$$

$$r_m = \sqrt{m\lambda R} \quad (6.53)$$

The radii of dark fringes can be found by inserting values 1,2,3, for m .

Thus,

$$r_1 = \sqrt{1\lambda R} \quad \text{or} \quad r_1 \propto \sqrt{1}$$

$$r_2 = \sqrt{2\lambda R} \quad \text{or} \quad r_2 \propto \sqrt{2}$$

$$r_3 = \sqrt{3\lambda R} \quad \text{or} \quad r_3 \propto \sqrt{3} \text{ and so on}$$

It means that *the radii of the dark rings are proportional to square root of the natural numbers.*

The above relation also implies that $r_m \propto \sqrt{\lambda}$

Thus, *the radius of the m^{th} dark ring is proportional to square root of wavelength.*

Ring Diameter:

$$\begin{aligned} \text{Diameter of } m^{\text{th}} \text{ dark ring} \quad D_m &= 2r_m \\ &= 2\sqrt{2Rt} \\ \text{or} \quad D_m &= 2\sqrt{m\lambda R} \end{aligned} \quad (6.54)$$

Example 6.6: In a Newton's rings experiment, the diameter of 10^{th} dark ring due to wavelength 6000 \AA in air is 0.5 cm . Find the radius of curvature of the lens.

Solution: Radius of curvature, $R = \frac{(D/2)^2}{m\lambda} = \frac{(0.5 \times 10^{-2}/2)^2 \text{ m}^2}{10 \times 6000 \times 10^{-10} \text{ m}} = 104 \text{ cm}$.

6.11.4 Spacing between Fringes is not Even

It is seen that the diameter of dark rings is given by

$$D_m = 2\sqrt{m\lambda R}$$

where $m = 1, 2, 3, \dots$

The diameters of dark rings are proportional to the square root of the natural numbers. Therefore, the diameter of the ring does not increase in the same proportion as the order of the ring, for example, if m increases as 1, 2, 3, 4, the diameters are

$$\begin{aligned} D_1 &= 2\sqrt{\lambda R} \\ D_2 &= 2(1.4)\sqrt{\lambda R} \\ D_3 &= 2(1.7)\sqrt{\lambda R} \\ D_4 &= 2(2)\sqrt{\lambda R} \text{ and so on} \end{aligned}$$

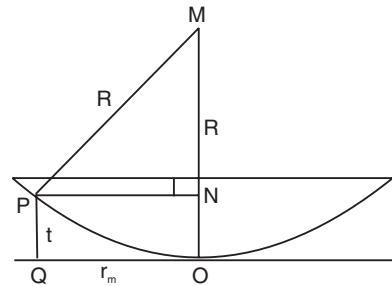


Fig. 6.25

Therefore, the rings get closer and closer, as m increases. This is why the rings are not evenly spaced.

6.11.5 Fringes of Equal Thickness

Newton's rings are formed as result of interference between light waves reflected from the top and bottom surfaces of a thin air film enclosed between a plano-convex lens and a plane glass plate. The occurrence of alternate bright and dark rings depends on the optical path difference arising between the reflected rays. If the light falls normally on the air film the optical path difference between the waves reflected from the two surfaces of the film is

$$\Delta = 2t - \lambda/2$$

It is seen that the path difference between the reflected rays arises due to the variation in the thickness ' t ' of the air film. Reflected light will be of minimum intensity for those thickness for which the path difference is $m\lambda$ and maximum intensity for those thickness for which the path difference is $(2m + 1)\lambda/2$. Thus, each maxima and minima is a locus of constant film thickness. Therefore, the fringes are known as fringes of equal thickness.

6.11.6 Dark Central Spot

The central spot is dark as seen by reflection. Newton's rings are produced due to superposition of light rays reflected from the top and bottom surfaces of a thin air film enclosed between a plano-convex lens and a plane glass plate. The occurrence of brightness or darkness depends on the optical path difference arising between the reflected rays. The optical path difference is given by $\Delta = 2t - \lambda/2$.

At the point of contact 'O' of the lens and glass plate (Fig. 6.26), the thickness of air film is negligibly small compared to a wavelength of light.

$$\begin{aligned} \therefore t &\approx 0 \\ \therefore \Delta &\approx \lambda/2 \end{aligned}$$

The wave reflected from the lower surface of the air film suffers a phase change of π while the wave reflected from the upper surface of the film does not suffer such change.

Thus, the superposing waves are out of step $\lambda/2$ which is equivalent to a phase difference of 180° (or π rad). Thus the two interfering waves at the centre are opposite in phase and produce a dark spot.

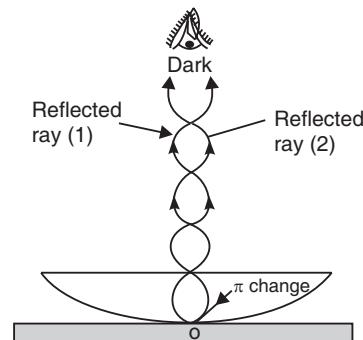


Fig. 6.26

6.11.7 Determination of Wavelength of Light

A plano-convex lens of large radius of curvature (about 100 cm) and a flat glass plate are cleaned. The lens is kept with its convex face on the glass plate and they are held in position with the help of a metal ring arrangement. The system is held under a low power travelling microscope kept before a sodium vapour lamp. It is arranged that the yellow light coming from the sodium lamp falls on a glass plate held at 45° to the light beam. The light is turned through 90° and is incident normally on the lens-plate system. The microscope is adjusted till the circular rings came into focus. The centre of the cross-wire is made to come into focus on the centre of the dark spot, which is at the centre of the circular ring system. Now, turning the screw the microscope is moved on the carriage slowly towards one side, say right side Fig. 6.27 (a). As the cross-wires move in the field of view, dark rings are counted. The movement is stopped when the 22nd dark ring is reached. Then the microscope is moved in the opposite

side and stopped at the 20th or 19th dark ring. The vertical cross-wire is made tangential to the 19th ring and the reading is noted with the help of the scale graduated on the carriage. Thus, starting from the 19th ring, the tangential positions of the 18th, 17th, 16th, ..., 5th dark rings are noted down. Now, the microscope is moved quickly to the left side of the ring system and it is stopped at the 5th dark ring. The cross-wire is again made tangential to the 5th dark ring and its position is noted. The difference between the readings on right and left sides of the 5th dark ring gives its diameter value. The procedure is repeated till 19th ring is reached and its reading is noted. From the value of the diameters the squares of the diameters are calculated. A graph is plotted between D_m^2 and between the ring number 'm'. A straight line would be obtained, as shown in Fig.6.27 (b).

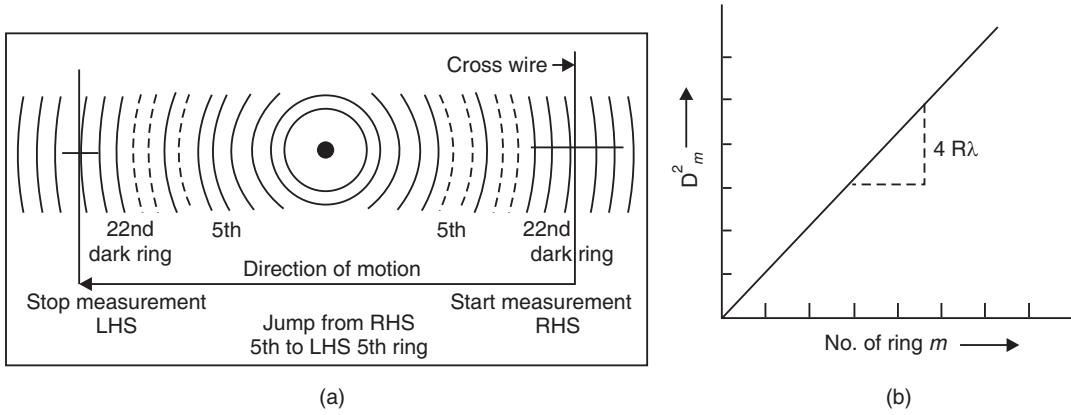


Fig. 6.27

We have

$$D_m^2 = 4m\lambda R \quad (6.55)$$

For the $(m + p)^{\text{th}}$ ring,

$$\begin{aligned} D_{m+p}^2 &= 4(m + p)\lambda R \\ D_{m+p}^2 - D_m^2 &= 4p\lambda R \end{aligned} \quad (6.56)$$

$$\lambda = \frac{D_{m+p}^2 - D_m^2}{4pR} \quad (6.57)$$

The slope of the straight line (Fig.6.27 b) gives the value of $4\lambda R$. Thus,

$$\lambda = \frac{\text{Slope}}{4R} \quad (6.58)$$

The radius of curvature R of the lens may be determined using a spherometer and λ is computed with the help of the above equation.

Example 6.7: In a Newton's rings experiment the diameter of the 15th ring was found to be 0.59 cm and that of the 5th ring was 0.336 cm. If the radius of the plano-convex lens is 100 cm, calculate the wavelength of light used.

$$\text{Solution: } \lambda = \frac{D_{m+p}^2 - D_m^2}{4pR} = \frac{D_{15}^2 - D_5^2}{4 \times 10 \times R} = \frac{(5.9 - 3.36)^2 \times 10^{-6} \text{ m}^2}{4 \times 10 \times 1 \text{ m}} = 5880 \text{ Å}$$

Example 6.8: In a Newton's rings experiment the diameter of the 4th and 12th dark rings are 0.400 cm and 0.700 cm respectively. Determine the diameter of 20th dark ring.

Solution: $\lambda = \frac{D_{m+p}^2 - D_m^2}{4pR} = \frac{D_{12}^2 - D_4^2}{4 \times 8 \times R}$

Considering 20th and 4th dark rings,

$$\lambda = \frac{D_{20}^2 - D_4^2}{4 \times 16 \times R}.$$

Dividing these two equations, we get

$$D_{20}^2 - D_4^2 = 2(D_{12}^2 - D_4^2)$$

or $D_{20}^2 = 2D_{12}^2 - D_4^2 = 2(0.700)^2 - (0.400)^2 \text{ cm}^2 = (0.98 - 0.16) \text{ cm}^2 = 0.82 \text{ cm}^2$.

$$\therefore D_{20} = 0.906 \text{ cm.}$$

6.11.8 Refractive Index of a Liquid

The liquid whose refractive index is to be determined is filled in the gap between the lens and plane glass plate. Now the liquid film substitutes the air film. The condition for interference may then be written as

$$2\mu t \cos r = m\lambda \quad \text{Darkness}$$

where μ is the refractive index of the liquid. For normal incidence the equation becomes

$$2\mu t = m\lambda$$

The diameter of m^{th} dark ring is given by

$$[D_m^2]_L = \frac{4m\lambda R}{\mu} \quad (6.59)$$

Similarly, the diameter of the $(m + p)^{\text{th}}$ ring is given by

$$[D_{m+p}^2]_L = \frac{4(m+p)\lambda R}{\mu} \quad (6.60)$$

Subtracting equ.(6.59) from equ.(6.60), we get

$$[D_{m+p}^2]_L - [D_m^2]_L = \frac{4p\lambda R}{\mu} \quad (6.61)$$

But we know that

$$(D_{m+p}^2)_{\text{air}} - (D_m^2)_{\text{air}} = 4p\lambda R \quad (6.62)$$

$$\therefore \mu = \frac{(D_{m+p}^2)_{\text{air}} - (D_m^2)_{\text{air}}}{(D_{m+p}^2)_{\text{liq}} - (D_m^2)_{\text{liq}}} \quad (6.63)$$

Example 6.9: In a Newton's Rings experiment, the diameter of the 15th ring was found to be 0.59 cm and that of the 5th ring was 0.336 cm. If the radius of the plano-convex lens is 100 cm, calculate the wavelength of light used. What happens to ring diameter if air film is replaced with liquid of refractive index 1.5?

Solution: $\lambda = \frac{D_{m+p}^2 - D_m^2}{4pR} = \frac{D_{15}^2 - D_5^2}{4 \times 10 \times R} = \frac{(0.59^2 - 0.336^2) \text{ cm}^2}{4 \times 10 \times 100 \text{ cm}} = 5902 \text{ Å.}$

The diameters of the Newton's rings will be reduced when the air film is replaced with a liquid film.

For example,

$$(D_{15}^2)_{liq} = \frac{(D_{15}^2)_{air}}{\mu} = \frac{0.59^2 \text{ cm}^2}{1.5} = 0.232 \text{ cm}^2$$

$$\therefore D_{15} = \sqrt{0.232 \text{ cm}^2} = \mathbf{0.4817 \text{ cm}}$$

Example 6.10: In a Newton's rings experiment the diameter of 10th ring changes from 1.40 to 1.27 cm when a drop of liquid is introduced between the lens and the glass plate. Calculate the refractive index of the liquid.

Solution:

$$\mu = \frac{(D_m^2)_{air}}{(D_m^2)_{liq}} = \frac{(1.40 \text{ cm})^2}{(1.27 \text{ cm})^2} = 1.215.$$

6.11.9 Newton's Rings by Transmitted Light

Newton's rings are observed also when we view the light transmitted through the lens-plate combination. The light rays passing through the system gets partly reflected at the top and bottom surfaces of the air film enclosed between the lens and glass plate (Fig.6.28). The optical path difference between two rays reflected from the top of the air film and the bottom of the air film is given by

$$\Delta = 2\mu t \cos r$$

Since $\mu = 1$ for air and $\cos r = 1$ for normal incidence of light,

$$\Delta = 2t$$

Intensity maxima occur when the optical path difference $\Delta = m\lambda$. Thus, if

$$2t = m\lambda \quad (6.64)$$

bright fringe is obtained.

Intensity minima occur when the optical path difference is $\Delta = (2m+1)\lambda/2$. Hence, if

$$2t = (2m+1)\lambda/2 \quad (6.65)$$

dark fringe is produced.

Taking the value of $t = \frac{r_m^2}{2R}$, the radii of the bright and dark rings can be calculated. We find that for **bright rings**

$$r_m = \sqrt{m\lambda R} \quad (6.66)$$

and for **dark rings**

$$r_m = \sqrt{\frac{(2m+1)\lambda R}{2}} \quad (6.67)$$

Thus, the ring system due to transmitted light is just opposite to that we observe in the reflected light. Wherever we get bright rings in the reflected light, we observe dark rings in the transmitted light and vice versa. At the center, we find bright spot in the transmitted light.

Thus, the ring system in transmitted light is complementary to that seen in reflected light. However, the rings in transmitted light are much poorer in contrast.

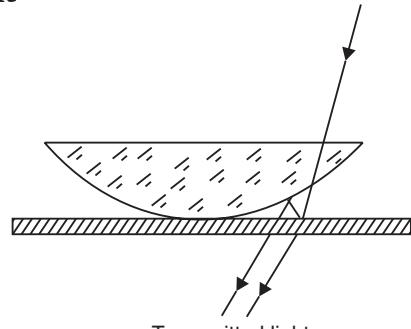


Fig. 6.28

6.12 APPLICATIONS OF INTERFERENCE

The applications of interference phenomenon are wide and varied. Interference is used for making precision measurements. For example, the wavelength of light can be measured with accuracy up to eight significant digits. Interference is used for measuring small displacements. The refractive indices of liquids and gases are measured using interference. We study here two important applications, which utilize the phenomenon of thin film interference.

6.12.1 Testing of Flatness of Surfaces

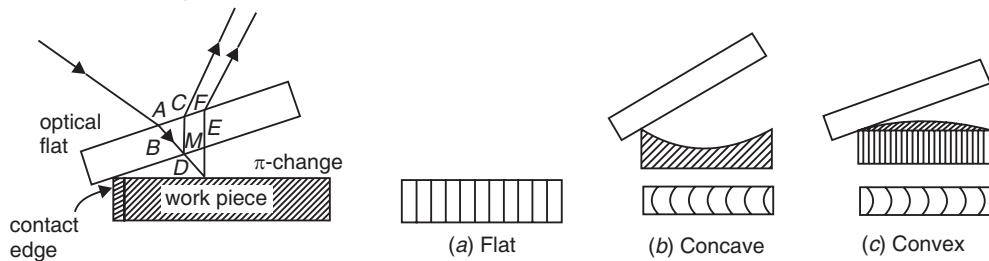


Fig. 6.29

In modern technology thin film interference is widely used. One of the applications is testing of flatness of surfaces. Machine components retain surface irregularities left after machining. The extent of suitability of the component for a particular application depends on the irregularities which act as sources of stress leading to fatigue cracks. The surfaces of components which are going to be subjected to high stress and load reversals are therefore required to have a smooth surface finish. The smoothness of a surface can be quickly inspected visually by keeping an optical flat on the component at an angle and illuminating it with a monochromatic light (Fig.6.29). The air wedge formed between the component and optical flat produces straight and equidistant fringes if the component surface is smooth. If the fringes are curved towards the contact edge, the surface is concave and if the fringes curve away, it is convex (Fig.6.29).

Testing of a lens surface: One of the important uses of Newton's rings is in the testing of the optical components manufactured for use in telescopes and other instruments. The grinding of a lens surface is tested by keeping it on a master. A master is an optical flat which is a cylindrical disc made of fused quartz. The two faces of the optical flat are perfectly parallel to each other. The departure from the flatness of each face is less than a light wavelength. If a lens is ground perfectly, a circular fringe pattern is observed. Otherwise variations are observed which give an indication of how the lens must be ground and polished to remove the imperfections. High quality lenses are ground with a precision of less than a light wavelength.

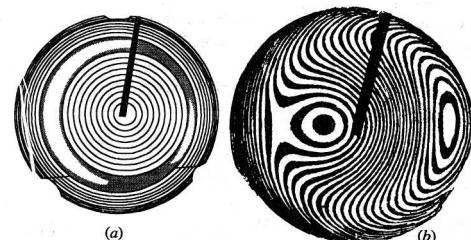


Fig. 6.30: Lens inspection using Newton's rings (a) Circular ring pattern indicates the high quality of grinding. (b) Distorted pattern indicates irregularities.

6.12.2 Thickness of a Thin Film Coating

Dielectric and metallic thin films are often coated on optical components, solar cells etc. One of the methods of determination of thickness of such thin films is based on multiple beam interference. A partially coated substrate is used for the determination. The surfaces

of the substrate and the thin film on it are coated with a transparent metallic film of uniform thickness. A glass plate is also coated on one of its surfaces with the transparent metallic film. When the substrate and the glass plate are placed in contact and examined under monochromatic light, the reflected light shows a fringe system, as shown in Fig. 6.31. A shift occurs in the fringes as we pass from the region occupied by thin film to the region where thin film is absent. The amount of displacement of one set of the fringes with respect to the second set of fringes is given by

$$s = 2t \quad \text{or} \quad t = s/2$$

where t is the thickness of the thin film. By measuring ' s ', t can be calculated.

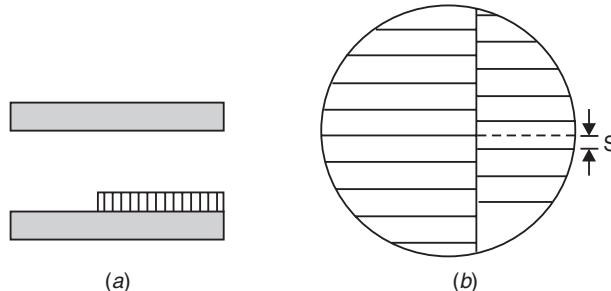


Fig. 6.31: Determination of the thickness of a thin coating.

6.12.3 Anti-Reflecting Coatings

Optical instruments such as telescopes and cameras use multicomponent glass lenses. When light is incident on the lens, part of the incident light is reflected away and that much amount of light is lost and wasted. When more surfaces are there, the number of reflections will be large and the quality of the image produced by a device will be poor. In case of solar cells, which operate on sunlight (daylight), the electrical energy produced will be less because of the loss of part of light energy due to reflection, at the cell surface. It is found that coating the surface with a thin transparent film of suitable refractive index can reduce such loss of energy due to reflections at surface. Such coatings are called **antireflection coatings**. Thus,

"Antireflection (AR) coatings are thin transparent coatings of optical thickness of one-quarter wavelength given on a surface in order to suppress reflections from the surface".

Alexander Smakula discovered in 1935 that the reflections from a surface can be reduced by coating the surface with a thin transparent dielectric film.

A thin film can act as an AR coating if it meets the following two conditions:

- (i) **Phase condition:** The waves reflected from the top and bottom surfaces of the thin film are in *opposite phase* such that their overlapping leads to destructive interference, and
- (ii) **Amplitude condition:** The waves have *equal amplitudes*.

The above conditions enable us determine respectively (a) the required thickness of the film and (b) the refractive index of the material to be used for forming the film.

(i) **Phase condition and minimum thickness of the film:** Let the thickness of the film be t and the refractive of the film-material be μ_f . The phase

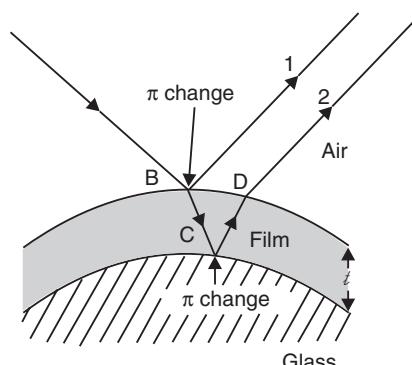


Fig. 6.32

condition requires that the waves (ray 1 and ray 2) reflected from the top and bottom surfaces of thin film be 180° out of phase. It requires that the optical path difference between the two rays must equal one half-wave or an odd number of half-waves. Referring to Fig.6.32, the optical path difference between ray 1 and ray 2 is

$$\Delta = 2\mu_f t \cos r - \lambda/2 - \lambda/2$$

the first $\lambda/2$ corresponds to the π change at the top surface of the film (air-to-film boundary) and the second $\lambda/2$ to the π change that occurs at the film-to glass boundary because $\mu_f < \mu_g$. If we assume normal incidence of light, $\cos r = 1$ and the above equation reduces to

$$\Delta = 2\mu_f t - \lambda = 2\mu_f t$$

We wrote the above equality remembering that an *addition of a full wave or subtraction of a full wave from a train of waves does not affect the original phase relation*. The ray 1 and ray 2 interfere destructively if the optical path difference satisfies the condition that $\Delta = (2m + 1)\lambda/2$.

Thus, it requires that

$$2\mu_f t = (2m + 1)\lambda/2$$

For the film to be transparent, its thickness should be a minimum, which happens when $m = 0$.

$$\begin{aligned} 2\mu_f t_{\min} &= \lambda/2 \\ \therefore t_{\min} &= \frac{\lambda}{4\mu_f} \quad (\mu_f < \mu_g) \end{aligned} \quad (6.68)$$

It means that the optical thickness of the AR coating should be of one-quarter wavelength. Such quarter-wavelength coatings suppress the reflections and cause the light to pass into the transmitted component.

(ii) Amplitude condition: The amplitude condition requires that the amplitudes of reflected rays, ray 1 and ray 2 are equal. That is,

$$E_1 = E_2 \quad (6.69)$$

It requires that

$$\left[\frac{\mu_f - \mu_a}{\mu_f + \mu_a} \right]^2 = \left[\frac{\mu_g - \mu_f}{\mu_g + \mu_f} \right]^2 \quad (6.70)$$

where μ_a , μ_f , and μ_g are the refractive indices of air, thin film and glass substrate respectively. As $\mu_a = 1$, the above expression may be rewritten as

$$\left[\frac{\mu_f - 1}{\mu_f + 1} \right]^2 = \left[\frac{\mu_g - \mu_f}{\mu_g + \mu_f} \right]^2$$

Expanding the above equation, we get

$$\begin{aligned} \frac{\mu_f^2 - 2\mu_f + 1}{\mu_f^2 + 2\mu_f + 1} &= \frac{\mu_g^2 - 2\mu_g\mu_f + \mu_f^2}{\mu_g^2 + 2\mu_g\mu_f + \mu_f^2} \\ 4\mu_f^3\mu_g + 4\mu_f\mu_g &= 4\mu_f^3 + 4\mu_f\mu_g^2 \end{aligned}$$

Dividing by $4\mu_f$ and rearranging the terms

$$\begin{aligned} \mu_f^2 - \mu_g\mu_f + \mu_g^2 - \mu_g &= 0 \\ \mu_f^2 &= \mu_g(1 + \mu_f^2 - \mu_g) \end{aligned}$$

$$\begin{aligned}\therefore \mu_f^2 &\approx \mu_g \text{ (as } \mu_f \approx \mu_g) \\ \therefore \mu_f &= \sqrt{\mu_g} \end{aligned} \quad (6.71)$$

It implies that the refractive index of thin film should be less than that of the substrate and possibly nearer to its square root.

In case of glass, if we take $\mu_g = 1.5$, $\mu_f = \sqrt{\mu_g} = 1.22$.

The materials which have refractive index nearer to this value are magnesium fluoride, MgF_2 ($\mu = 1.38$) and cryolite, $3\text{NaF}\cdot\text{AlF}_3$ ($\mu = 1.36$). Apart from the refractive index, the material should possess some more additional properties. The film should adhere well, should be durable, scratch proof and insoluble in ordinary solvents. MgF_2 and cryolite satisfy these requirements. However, among the two, magnesium fluoride is cheaper and is hence widely used as AR coating.

It may be noted that the condition (6.68) is satisfied only at one particular wavelength. The wavelength normally chosen is 5500 \AA for which the eye is most sensitive. This wavelength is located in the yellow-green portion of the spectrum. Consequently, the reflection of red and violet light will be larger when white light is incident on the component such as a camera lens. Hence, the component shows *purple hue* in reflected light.

Example 6.11: A glass microscope lens ($\mu = 1.5$) is coated with magnesium fluoride ($\mu_f = 1.38$) film to increase the transmission of normally incident light $\lambda = 5800 \text{ \AA}$. What minimum film thickness should be deposited on the lens?

Solution:

$$t_{\min} = \frac{\lambda}{4\mu_f} = \frac{5800 \times 10^{-10} \text{ m}}{4 \times 1.38} = 1051 \text{ \AA}$$

Example 6.12: Can a thin film of water ($\mu_f = 1.33$) formed on a glass window pane ($\mu_f = 1.52$) act as a non-reflecting film? If so, how thick should be the water film?

Solution: A film of refractive index μ_f can act as a non-reflecting film on a substrate having refractive index μ , if $\mu_f = \sqrt{\mu}$.

Here, $\sqrt{\mu} = \sqrt{1.52} = 1.233$.

As the refractive index of water is 1.33, it is nearer to water film $\sqrt{\mu}$ can act as a non-reflecting film on glass.

The minimum thickness of the film is given by

$$t_{\min} = \frac{\lambda}{4\mu_f}.$$

As human eye is more sensitive to green, it may be assumed that $\lambda = 5500 \text{ \AA}$.

$$\therefore t_{\min} = \frac{5500 \times 10^{-10} \text{ m}}{4 \times 1.33} = 1034 \text{ \AA}.$$

Multilayer AR coatings:

A single layer AR coating is effective only at one particular wavelength. A much wider coverage across the spectrum is possible with multiple coatings, called *multilayers*. In practice three layer coatings are widely used and are highly effective over most of the visible spectrum. The central layer is half-wave ($\lambda / 2$) thick and is of high refractive index materials

such as zirconium dioxide (ZrO_2 , $\mu = 2.1$). The outside layer is of magnesium fluoride having $\lambda / 4$ thickness and the layer adjacent to the substrate is again a $\lambda / 4$ thick coating of cesium fluoride (CeF_3 , $\mu = 1.63$) or aluminium oxide (Al_2O_3 , $\mu = 1.76$). Some of the antireflection coatings use up to 100 layers of alternating high and low refractive index materials.

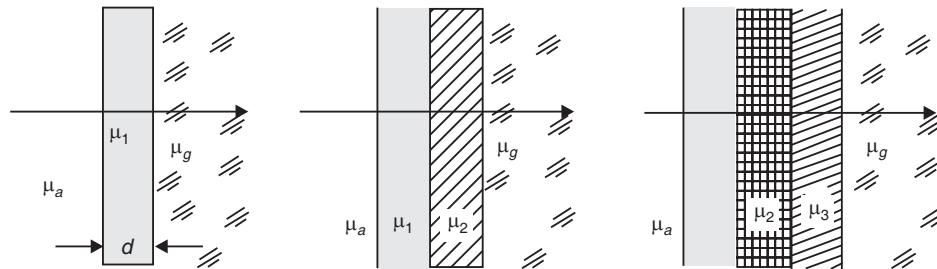


Fig. 6.33

6.12.4 Interference Filters

An interference filter is an optical system that will transmit a very narrow range of wavelengths and thus provides a monochromatic beam of light. Interference filters are fabricated earlier as follows. A thin metallic film, usually of aluminium or silver, is deposited on a glass substrate by vacuum deposition technique. Then a thin layer of cryolite is deposited over this. The structure is again covered by another metallic film. Another plate is placed over it to protect the thin film structure. By varying the thickness of the dielectric film, any particular wavelength can be filtered out. However, the filtered light will have a narrow spectrum centered on the chosen wavelength. By increasing the reflectivity of the surfaces, the transmitted spectrum can be made narrower. But it is not possible to increase the thickness of metallic films indefinitely, as they start absorbing the light.

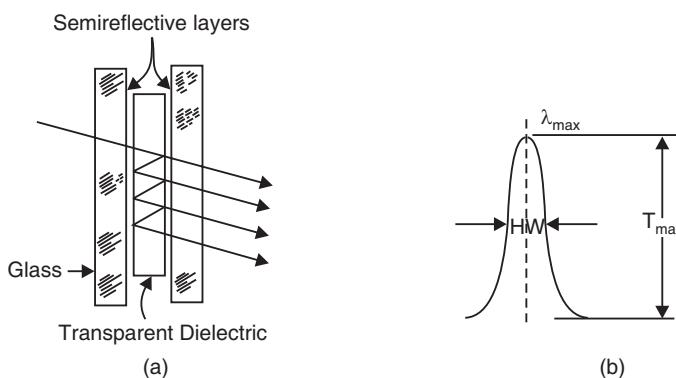


Fig. 6.34

In modern versions metallic films are not used; instead dielectric films are used. In an all dielectric interference filter, layers of dielectric materials of appropriate refractive indices are deposited. To obtain an interference filter, a $\lambda/4$ thick film of titanium oxide is deposited and then over it a film of dielectric material with lower refractive index, such as magnesium fluoride is deposited. On this, again a $\lambda/4$ thick film of titanium oxide is deposited. In this way alternately high and low refractive index materials are deposited to obtain an interference filter. With multiple coatings, it is possible to fabricate filters, which are capable of transmitting a

very narrow spectrum of a width as small as 11 \AA or even less, about a chosen wavelength in the visible region. Modern filters use up to 100 layers.

6.13 MICHELSON'S INTERFEROMETER

An **interferometer** is an instrument in which the phenomenon of interference is used to make precise measurements of wavelengths or distances. Michelson designed an ingenious interferometer which utilizes the thin film interference.

6.13.1 Principle

A beam of light from an extended source is divided into two coherent beams of equal intensities by partial reflection and refraction. These beams travel in two mutually perpendicular directions and come together after reflection from plane mirrors. The beams overlap on each other and produce interference fringes. The fringes are *circular* if the reflecting mirrors are exactly perpendicular to each other or *straight* if the mirrors are inclined at a small angle.

6.13.2 Construction

The schematic of a simple Michelson interferometer is shown in Fig. 6.35 (a). It consists of a beam splitter G_1 , a compensating plate G_2 , and two plane mirrors M_1 and M_2 . The beam splitter G_1 is a partially silvered plane parallel glass plate and it has the property that it transmits half the incident light and reflects the rest. The compensating plate G_2 is a simple plane parallel glass plate having the same thickness as G_1 . The two plates G_1 and G_2 are held parallel to each other and are inclined at an angle of 45° with respect to the mirror M_2 . The mirror M_1 is mounted on a carriage and can be moved exactly parallel to itself with the help of a micrometer screw. The distance through which the mirror M_1 is moved can be read with the help of a graduated drum D attached to the screw. Displacements of the order of 0.1 mm can be easily read. The plane mirrors M_1 and M_2 can be made perfectly perpendicular with

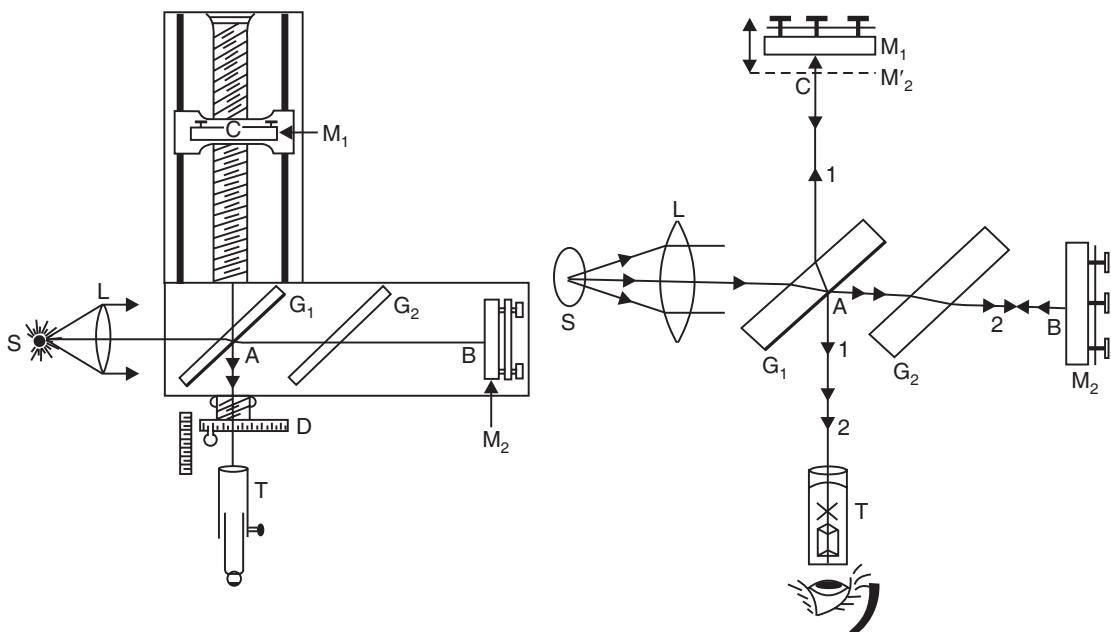


Fig. 6.35

the help of the fine screws attached to them. The interference bands are observed in the field of view of the telescope T.

6.13.3 Working

Monochromatic light from an extended source is rendered parallel by means of a collimating lens L and is made incident on the beam splitter G₁. It is partly reflected at the back surface of G₁ along AC and partly transmitted along AB. The beam AC travels normally towards the plane mirror M₁ and is reflected back along the same path and comes out along AT. The transmitted beam travels toward the mirror M₂ and is reflected along the same path. It is reflected at the back surface of G₁ and proceeds along AT. The two beams received along AT are produced from a single source through division of amplitude and are hence *coherent*. The superposition of these beams leads to interference and produces interference fringes.

From the Fig.6.35 (b) it is clearly seen that a light ray starting from the source S and undergoing reflection at the mirror M₁ passes through the glass plate G₁ three times. On the other hand, in the absence of plate G₂, the ray reflected at M₂ travels through the glass plate G₁ only once. For compensating this path difference, a compensating plate G₂ of the same thickness as that of G₁ is inserted into the path AB and is held exactly parallel to G₁.

If we look into the instrument from T, we see mirror M₁ and in addition we see a virtual image, M'₂, of mirror M₂. Depending on the positions of the mirrors, image M'₂ may be in front of, or behind, or exactly coincident with mirror M₁.

Optical path difference between the two waves at T:

The optical path of the wave that travelled along

$$LG_1M_1T = 2G_1M_1 + \lambda/2 + \lambda/2 = 2G_1M_1 + \lambda.$$

The optical path of the wave that travelled along

$$LG_1M_2T = 2G_1M_2 + \lambda/2 + \lambda/2 = 2G_1M_2 + \lambda.$$

Therefore, the optical path difference between the two waves

$$= 2(G_1M_1 - G_1M_2) = 2(x_1 - x_2) = 2d.$$

The two beams interfere constructively if $\Delta = m\lambda$ or destructively if $\Delta = (2m + 1)\lambda/2$. Anything that changes the optical path difference will cause a change in the relative phase of the two waves. As an example, if mirror M₁ is moved by a distance $\frac{1}{2}\lambda$, the path difference changes by λ and the fringe pattern will shift by one fringe.

6.13.4 Circular Fringes

Circular fringes are produced with monochromatic light when the mirrors M₁ and M₂ are exactly perpendicular to each other. The origin of the circular fringes can be understood as follows.

If we look into the instrument from T, we see mirror M₁ directly, and in addition we will see the virtual image M'₂ of mirror M₂ formed by reflection in the glass plate G₁. It means that one of the interfering beams come from M₁ and the other beam appears to come from the virtual image M'₂. The situation is similar to an air film enclosed between mirrors M₁ and M'₂ with the difference that in case of a real film between two surfaces, multiple reflections take place, whereas in this case only two reflections take place.

If one looks towards M₁ through T, one observes a virtual image of M₂ in G₁, parallel to M₁, say M'₂. M₁ and M'₂ act as two coherent sources formed by the thin film of thickness M₁M'₂.

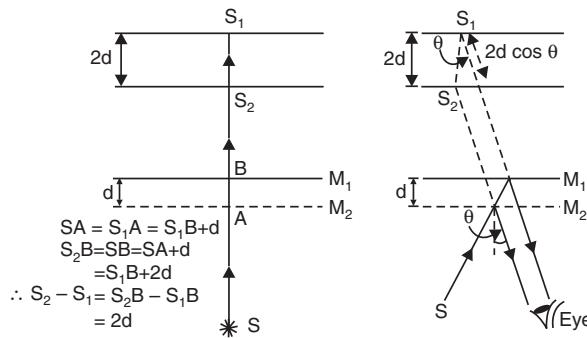


Fig. 6.36

$$M_1 M'_2 = G_1 M_1 - G_1 M'_2 = G_1 M_2 = (x_1 - x_2)$$

If the two arms of the interferometer are equal in length, and image M'_2 coincides with mirror M_1 . If M'_2 and M_1 do not coincide, the distance between them is finite, $M'_2 M_1 = (x_1 - x_2) = d$. The effect is that of light from a point source S falling on a uniformly thick film of air whose thickness is equal to d . Now, a light ray is reflected by both M'_2 and M_1 and the observer will see two virtual images- S_1 due to reflection at M'_2 and S_2 due to reflection at M_1 . The virtual images are separated by a distance $2d$.

If the observer looks into the system at an angle θ , the path difference between the two beams will be $2d \cos \theta$. The light that comes from M_2 and goes to T undergoes rare-to-dense reflection and therefore a π -phase change occurs. In view of this, the total path difference between the two beams is given by

$$\Delta = 2d \cos \theta + \lambda / 2. \quad (6.72)$$

The condition for obtaining brightness

$$2d \cos \theta + \lambda / 2 = m\lambda \quad (6.73)$$

where $m = 0, 1, 2, \dots$.

For a given mirror separation d , a given wavelength λ and order m , angle θ is constant. This means that the fringes are of circular shape. They are called *fringes of equal inclination*.

In case the mirror M_1 coincides with the virtual image M'_2 , $d = 0$, the optical path difference between the interfering beams will be $\lambda/2$ (Refer to eqn. 6.73). Consequently, we obtain a minimum at the coincidence position and the centre of the field will be dark, as shown in Fig. 6.37 (a).

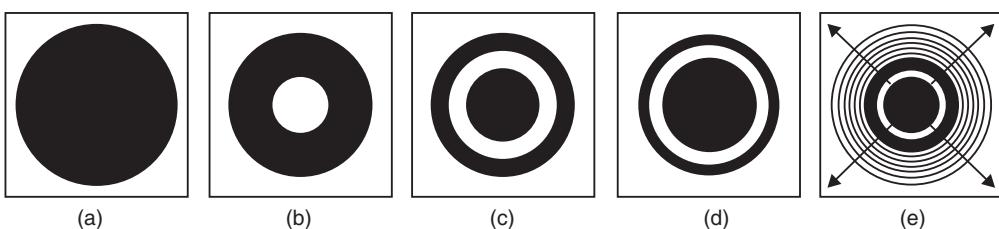


Fig. 6.37

If one of the mirrors is now moved through a distance $\lambda/4$, the path difference changes by $\lambda/2$ and therefore a maximum is obtained. By moving the mirror through another $\lambda/4$, a

minimum is obtained; moving it by another $\lambda/4$ again a maximum is obtained and so on. Therefore, a new ring appears in the centre of the field each time the mirror is moved through $\lambda/2$. As d increases new rings appear in the centre faster than rings already present disappear in the periphery; and the field becomes more crowded with thinner rings. Conversely, as d is made smaller, the rings contract and disappear in the centre.

6.13.5 Localized Fringes

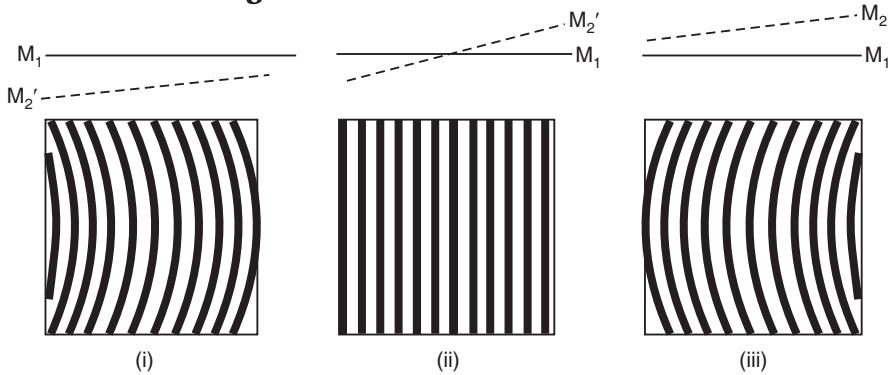


Fig. 6.38

When the two mirrors are tilted, they are not exactly perpendicular to each other and therefore the mirror M_1 and the virtual image M'_2 are not parallel. In this case the air path between them is wedge-shaped and the fringes appear to be straight. If one of the mirrors is moved, the fringes move across the field. The position of any particular bright fringe is taken up by the one next to it. The fringes can be counted as they pass a reference mark. If m fringes move across the field of view when M_1 moves through a distance d , then

$$d = m \lambda/2$$

or
$$\lambda = \frac{2d}{m} \quad (6.74)$$

Example 6.13: In a Michelson interferometer 200 fringes cross the field of view when the movable mirror is moved through 0.0589 mm. Calculate the wavelength of light used.

Solution:
$$\lambda = \frac{2d}{m} = \frac{2 \times 0.0589 \times 10^{-3} \text{ m}}{200} = 5890 \text{ \AA}$$

6.13.6 White Light Fringes

Instead of a monochromatic source, if a white light source is used, a few coloured fringes with a central dark fringe can be observed. In observing these fringes, the mirrors are slightly tilted as for localised fringes and position of M_1 is found where it intersects M'_2 . This position is often difficult to find with white light. The position can best be located with monochromatic light when the fringes become straight. Then a very slow motion of M_1 in this region using white light will bring these fringes into view, when a central dark fringe is surrounded by 8 to 10 coloured fringes on either side are observed. These fringes are useful for the determination of zero path difference.

6.14 APPLICATIONS OF MICHELSON INTERFEROMETER

Michelson interferometer can be used to determine (i) the wavelength of a given monochromatic source of light (ii) the difference between the two neighbouring wavelengths or

resolution of the spectral lines, (iii) refractive index and thickness of various thin transparent materials and (iv) for measurement of the standard metre in terms of the wavelength of light. We study here only the first three applications.

6.14.1 Measurement of Wavelength

Michelson interferometer is used to determine the wavelength of light from a monochromatic source. The monochromatic source is kept at S. If the mirrors M_1 and M_2 are exactly perpendicular, circular fringes are obtained. If the mirror M_1 is moved forward or backward, the circular fringes appear or disappear at the centre. Now, as the mirror is moved through a known distance d and the number of fringes disappearing at the centre is counted. Suppose d_1 is the initial thickness of the air film between the mirror M_1 and the image of M_2 corresponding to the bright fringe of order m_1 and d_2 is the final thickness of the air film corresponding to a bright fringe of order m_n in the same position. Then,

$$2d_1 = m_1 \lambda$$

and

$$2d_2 = m_n \lambda$$

By subtraction, we get $2(d_2 - d_1) = (m_n - m_1)\lambda$

\therefore

$$2d = N\lambda$$

\therefore

$$\lambda = \frac{2d}{N} \quad (6.75)$$

6.14.2 Determination of the Difference in the Wavelength of Two Waves

If a source of light consists of two wavelengths λ_1 and λ_2 , which differ slightly, then two sets of fringes corresponding to the two wavelengths are produced in a Michelson interferometer. By adjusting the position of the mirror M_1 of the interferometer, the position is found when the fringes are very bright. In this position, the bright fringe due to λ_1 coincides with the bright fringes due to λ_2 . When the mirror M_1 is moved, the two sets of fringes get out of step because their wavelengths are different. When the mirror M_1 has been moved through a certain distance, the bright fringe due to one set will coincide with the dark fringe due to the other set and no fringes will be seen in this case. Again by moving the mirror M_1 , a position is reached when a bright fringe of one set falls on the bright fringe of the other and the fringes are again distinct. This is possible when the m^{th} order of the longer wavelength coincides with the $(m + 1)^{\text{th}}$ order of the shorter wavelength.

Let m_1 and m_2 be the changes in the order at the centre of the field when the mirror M_1 is displaced through a distance d between two consecutive positions of maximum distinctness of the fringes.

$$\therefore 2d = m_1 \lambda_1 = m_2 \lambda_2$$

If λ_1 is greater than λ_2

$$m_2 = m_1 + 1$$

$$\therefore 2d = m_1 \lambda_1 = (m_1 + 1) \lambda_2$$

$$\therefore m_1 = \frac{\lambda_2}{\lambda_1 - \lambda_2}$$

$$\therefore 2d = \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2}$$

$$\text{or } \lambda_1 - \lambda_2 = \frac{\lambda_1 \lambda_2}{2d}$$

Taking λ as the mean wavelength of the two wavelengths λ_1 and λ_2 , the small difference $\Delta\lambda$ is given by

$$\Delta\lambda = \lambda_1 - \lambda_2 = \frac{\lambda^2}{2d} \quad (6.76)$$

6.14.3 Thickness of a Thin Transparent Sheet

Let a transparent sheet of thickness t and refractive index μ be inserted in the path of one of the interfering beams of Michelson interferometer. The optical path of that beam increases because of the sheet. It becomes μt instead of t . The increase in the optical path is $2(\mu t - t)$ or $(\mu - 1)t$. Since the beam traverses the medium twice, the extra path difference between the two interfering beams is $2(\mu - 1)t$. If m is the number of fringes by which the fringe system is displaced, then

$$2(\mu - 1)t = m\lambda$$

When monochromatic light is used, it is difficult to distinguish the sudden shift of fringes on insertion of the thin sheet. It is also not possible to count the number of fringes shifted. The difficulty is overcome by using white light first to locate the central dark fringe and it is made to coincide with the cross-wire of the telescope. The thin sheet is then introduced into the path of the beam. Position of mirror M_1 is adjusted till again a dark fringe of zero path difference coincides with the cross-wire of the telescope. The distance d through which the mirror is moved is noted. The white light is now replaced with the monochromatic light and the mirror M_1 is moved back slowly and the number of fringes contained in d is found. The thickness t is obtained from the relation

$$t = \frac{m\lambda}{2(\mu - 1)} \quad (6.77)$$

6.15 MOIRE FRINGES

Moiré patterns or fringes are the relatively thick lines produced when two patterns of thin lines overlap. For example, when two window screens are held close together, intersection of their lines produces another repetitive sequence of lines, which is a moiré pattern. The term moiré comes from the French and refers to the wavy finish of textiles made from wool. It was Lord Rayleigh that first noticed the moiré effect in optics, in 1874.

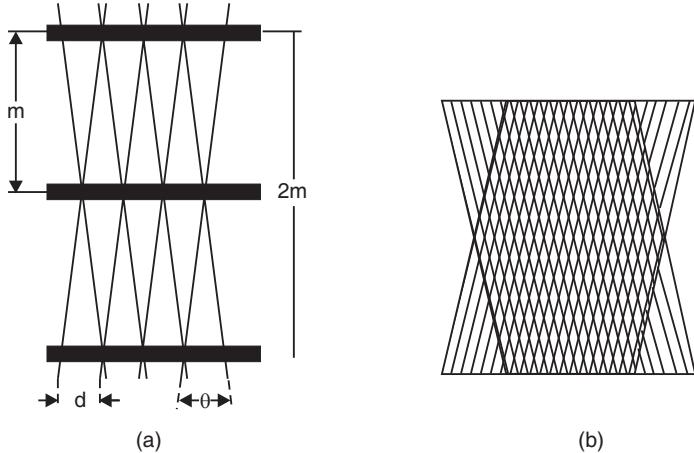


Fig. 6.39

Let us assume that the spacing of the lines in the two original grids is d and that one of the grids is turned through an angle θ with respect to the other grid (Fig. 6.39 a). Assuming that θ is small, it follows that

$$\theta \approx \frac{d}{m} \quad \text{or} \quad m \approx \frac{d}{\theta} \quad (6.78)$$

Thus, the more parallel the grids, the wider the spacing m of the moiré lines.

The sketch opposite demonstrates the formation of moiré fringes. One grid consists of equally space d straight lines sloping upwards to the left. The grid sloping upwards to the right consists of straight lines whose spacing, which is the same at the left as that of the first grid, doubles going from left to right. The pattern is relatively bright around the points of intersection of the two grids. The loci of these points specify the positions of the moiré fringes. The limiting slope, where the spacings are equal (on the left), is horizontal. Scanning from left to right the slope increases very quickly.

QUESTIONS

1. Define interference. What do you mean by coherent sources? (C.S.V.T.U., 2006)
2. Explain the phenomenon of interference of light. What are the necessary conditions to get clear and distinct interference fringes? (Calicut Univ., 2006)
3. What are the methods used for obtaining coherent sources?
4. Write the names of two classes into which the phenomenon of interference is divided. (C.S.V.T.U., 2008)
5. Describe Fresnel's biprism. Explain how the wavelength of light can be determined with its help. (RGPV, 2007)
6. How Fresnel's biprism can be used to determine the thickness of a given thin sheet of transparent material? (RGPV, 2008)
7. Give the theory of Fresnel biprism to determine the wavelength of monochromatic light source. (C.S.V.T.U., 2005, 2006)
8. State the basic conditions for the phenomenon of interference of light. Briefly discuss the effect of introducing a thin plate in the path of one of the interference beams in a biprism experiment. Deduce an expression for the displacement of the fringes. Show how this method is used for finding the thickness of a mica sheet.
9. State the conditions necessary for obtaining sustainable interference pattern using two sources. (Amaravati Univ., 2002)
10. Derive conditions for path difference for interference in thin parallel film due to reflected light.
11. Why is the concept of coherence of central importance in the study of interference? How is the interference pattern controlled by the temporal and spatial coherence of the source?
12. Obtain conditions for maxima and minima due to interference of reflected light in thin transparent film of uniform thickness. (R.T.M.N.U., 2005)
13. What is thin film? Obtain an expression for the path difference in case of interference in thin films due to reflected light. (R.T.M.N.U., 2006)
14. What do you mean by thin film? Deduce the condition of minima in case of a thin parallel film. (R.T.M.N.U., 2007)
15. Explain the phenomenon of interference in thin film of uniform thickness due to reflected light. What happens when:
 - (i) Monochromatic light is incident normally on the uniform thin film?
 - (ii) White light is incident on the film?
16. (a) Obtain the condition for a thin film to appear bright in reflected light.
 (b) Explain colours of thin films. (Calicut Univ., 2007)
17. Derive an expression for condition of maxima and minima for reflected light in case of thin transparent film of uniform thickness. (Univ. of Pune, 2007)
18. Derive the condition for path difference for interference in parallel thin film due to reflected light. (Amaravati Univ., 2007, 2008)
19. A thin film of uniform thickness is illuminated by monochromatic light. Obtain the conditions of brightness and darkness of the film as observed in reflected light. When seen in reflected light, why does an excessively thin film appear to be perfectly dark, when illuminated by white light? (Univ. of Pune, 2007)

20. A thin film illuminated by white light appears coloured when observed in reflected light. Explain why. **(Univ. of Pune, 2008)**
21. What is thin film? Obtain an expression for fringe width in the interference pattern of wedge shape film. How this phenomenon is used to determine the thickness of thin wire?
22. Give the theory and formation of wedge shaped films. How these can be used to find the wavelength of light used? **(Calicut Univ., 2005)**
23. A wedge shaped air film is illuminated by monochromatic light. Obtain an expression for the fringe width of interference pattern formed. **(R.T.M.N.U., 2006)**
24. Obtain an expression for fringe width in wedge shaped thin film. **(C.S.V.T.U., 2008)**
25. (a) How will you find the thickness of a thin wire using air wedge method?
 (b) Discuss the testing of optical planeness of surfaces. **(Calicut Univ., 2007)**
26. Explain how plane ness of a surface can be tested by air wedge. **(Calicut Univ., 2005)**
27. Explain how interference pattern can be used for testing the optical flatness of surfaces. **(Univ. of Pune, 2007), (R.T.M.N.U., 2007)**
28. Explain the formation of Newton's rings. Determine the wavelength of sodium light using Newton's rings experiment. **(Calicut Univ., 2006)**
29. Explain Newton's rings method for determining the wavelength of monochromatic light. Why is the centre of fringes dark and how can we get a bright centre? **(C.S.V.T.U., 2006)**
30. Describe Newton's rings experiment to determine the wavelength of incident monochromatic light. **(R.T.M.N.U., 2007)**
31. Describe and explain the formation of Newton's rings in reflected light. Prove that in reflected light
 (i) the diameters of the dark rings are proportional to the square roots of natural numbers and
 (ii) the diameters of the bright rings are proportional to the square roots of odd natural numbers. **(C.S.V.T.U., 2007)**
32. Give experimental set up to obtain Newton's rings. Explain how interference takes place. **(Univ. of Pune, 2007)**
33. Why are circular fringes obtained in Newton's rings arrangement? Why these fringes are called fringes of equal thickness? Why is the central fringe a dark spot when examined in reflected light?
34. Explain the formation of Newton's rings. Obtain an expression for the diameter of dark rings in reflected system. What will happen to the diameter of n^{th} dark ring if air is replaced by water film? Explain. **(Univ. of Pune, 2008)**
35. In Newton's Ring experiment, why:
 (i) The Plano convex lens has larger radius of curvature?
 (ii) The rings get closer away from centre?
 (iii) Central fringe is dark in reflected light?
 (iv) Fringes are circular.
36. What will happen if the convex lens in the Newton's ring apparatus is lifted up by $\lambda/4$ where λ is the wavelength of the light used? **(R.T.M.N.U., 2007)**
37. How are Newton's rings formed? Draw a neat diagram showing the formation of rings, as well as the experimental set up. Why are the rings circular? How are the ring diameters and film thickness related? Why the rings are not evenly spaced?
38. Explain the formation of Newton's rings. Prove that in Newton's rings by reflected light, the diameters of bright rings are directly proportional to square root of odd natural numbers. Hence explain how rings are getting closer with increase in diameter order. **(Univ. of Pune, 2007)**
39. In Newton's ring experiment, if observed in reflected light, is it possible to obtain bright spot at centre? Justify your answer. In Newton's ring experiment, why are the rings crowded away from the centre?
40. Describe and explain the formation of Newton's rings in reflected light. Prove that in reflected light (i) diameter of the dark rings are proportional to the square roots of natural numbers and (ii) diameters of bright rings are proportional to the square roots of odd numbers. **(RGPV, 2007)**

41. In a Newton's rings experiment, light of red colour is used first and then a blue light. Which set of rings would have larger diameter and therefore greater spacing between them?
42. Show that the diameter of the n th dark ring is given by $D_n = 2\sqrt{m\lambda R}$. (Amaravati Univ., 2007)
43. Derive an expression for wavelength of light in Newton's ring experiment. (Amaravati Univ., 2007)
44. Explain, how the refractive index of a liquid can be found out by Newton's rings.
(Amaravati Univ., 2006, 2008)
45. Draw appropriate diagrams illustrating the interference in thin films in the following cases. Label the interfering rays as ray 1 and 2, and write down expressions for the total path difference between them:
(i) constant thickness film;
(ii) wedge shaped film;
(iii) film enclosed between a plano-convex lens and plane glass plate in Newton's rings apparatus.
46. Why are lenses coated with thin film to improve transmission of light?
47. Explain the use of thin film as anti-reflection coating. (Univ. of Pune, 2007, 2008)
48. What is anti-reflection coating? Explain its principle and application. (R.T.M.N.U., 2005)
49. What do you understand by antireflection coating? Deduce an expression for minimum thickness of antireflection coating. Why do coated lenses look purple by reflected light? (R.T.M.N.U., 2007)
50. Draw a neat diagram of Michelson interferometer. State the conditions for obtaining
(i) circular fringes (ii) straight line fringes (iii) white light fringes.
51. Describe the construction and working of Michelson interferometer. How can it be used for measuring wavelength of monochromatic light? (RGPV, 2007, 2008)
52. Draw a neat labeled diagram of Michelson's interferometer and explain how it can be used to find the wavelength of monochromatic light. (Univ. of Pune, 2008)
53. Describe Michelson interferometer and explain how the fringes form in it. How can this be used for measuring the wavelength of monochromatic light ? Derive the formula. (Anna Univ., 2003)
54. How will you use Michelson's interferometer to determine the thickness of a thin transparent film or plate? (Anna Univ., 2005)
55. Explain the construction, types of fringes and applications of Michelson interferometer. (Anna Univ., 2006)
56. Explain how the wavelength of monochromatic source of light can be experimentally determined using a Michelson's interferometer. (Anna Univ., 2007)
57. Determine the wavelength of a monochromatic light and the resolution of spectral lines using Michelson's interferometer. (Calicut Univ., 2006)
58. Explain why circular fringes in Michelson interferometer shift in the field of view on displacing the movable mirror and deduce a relation between the wavelength and the displacement of the mirror.
59. What are the applications of Michelson interferometer? (Anna Univ., 2006)

PROBLEMS

1. A biprism forms interference fringes with monochromatic light of wavelength 546 nm. On introducing a thin glass plate of refractive index 1.5, in the path of one of the interfering beams, the central bright fringe shifts to the position previously occupied by the third bright fringe. Find the thickness of the plate. [Ans: $3.3 \mu\text{m}$]
2. In a biprism experiment a thin transparent sheet of refractive index 1.53 is placed in the path of one of the interfering beams. The central fringe is found to shift by a distance equal to the width of seven fringes. Calculate the thickness of the transparent sheet if the wavelength of the monochromatic light used is 5460 Å. [Ans: $7.2 \mu\text{m}$]
3. In a Fresnel biprism experiment 75 fringes are obtained in the field of view with wavelength 5893 Å. What will be the number of fringes in the field of view with green light of wavelength 5200 Å. [Ans: 85]
4. Interference fringes are produced by monochromatic light of wavelength 5460 Å. When a thin transparent sheet of thickness 6.3×10^{-4} cm is introduced in the path of one of the interfering

- beams, the central fringe shifts to a position occupied by 6th bright fringe. Compute the refractive index of the sheet. [Ans: 1.52]
5. Interference fringes are produced with light of wavelength 600 nm. A thin glass film of refractive index 1.5 is interposed in the path of one of the interfering beams. The central bright band is shifted to the position previously occupied by the fifth bright band. Find the thickness of the film. [Ans: 6 μm]
6. A glass wedge of angle 0.01 radian is illuminated by monochromatic light of 6000 Å falling normally on it. At what distance from the edge of the wedge will the tenth fringe be observed by reflected light? (RTMNU S -05) [Ans: $x = 5\beta = 1.5 \times 10^{-4} \text{ m}$]
7. A glass plate having parallel sides has thickness $t = 4 \times 10^{-4} \text{ mm}$ and RI = 1.5. If it is illuminated normally by white light, what wavelengths will be intensified in reflected beam in visible spectrum? (RTMNU W -04) [Ans: $\lambda = 4800 \text{ \AA}$]
8. A soap film of 5000 Å thickness is viewed at an angle of 35° to the normal. Find the wavelengths in the visible light which will be absent in the reflected light. The refractive index of the film is 1.333. (RTMNU S -04) [Ans: $\lambda_1 = 6016 \text{ \AA}$, $\lambda_2 = 4011 \text{ \AA}$]
9. In Newton rings experiment diameter of 15th ring was found to be 0.59 cm and that of 5th ring was 0.336 cm. If radius of plano-convex lens is 100 cm, calculate wavelength of light used. What happens to ring diameter if air film is replaced with liquid of refractive index 1.5? (RTMNU W-02) [Ans: $\lambda = 5880 \text{ \AA}$]
10. In Newton's rings experiment, diameter of 10th dark ring due to wavelength 6000 Å in air is 0.5 cm. Find the radius of curvature of the lens. (RTMNU, S- 03)
11. Yellow light of wavelength 5893 Å strikes a film of oil on water at an angle 30°. The 8th dark band is seen. Compute the thickness of the oil film if the refractive index of the oil is 1.44. [Ans: 1.75 μm]
12. White light is incident on a transparent film of refractive index 1.33 and thickness 1.6 μm at an angle of 45°. When the reflected light is examined, a dark band corresponding to 500 nm is seen. Find the order of the band. [Ans: 7]
13. Newton's rings are formed with sodium light in an experiment. What is the order of the dark ring, which has double the diameter of the 4th dark ring? [Ans: 16]
14. An engineer is interested in enhancing the (i) reflected (ii) transmitted portion of the light incident on a glass lens. Explain how this can be achieved using the following thin films :
 (i) Mg F₂ ($\mu = 1.38$) (ii) ZnS ($\mu = 2.37$)
 Assume $\lambda = 5500 \text{ \AA}$ and μ for material of lens = 1.5. (RTMNU W-03, S-06)
 [Ans: For enhancing transmission, Mg F₂ ($\mu_f < \mu_g$) will be used and $t_{\min} = 996.4 \text{ \AA}$; for enhancing reflection, ZnS ($\mu_f > \mu_g$) should be used and $t_{\min} = 580.16 \text{ \AA}$]
15. A camera lens is to be coated for antireflection effect. The refractive index of the lens material is 1.55. Discuss the principle involved and find out the requirement of the coating.
16. A glass microscope lens ($\mu = 1.5$) is coated with magnesium fluoride ($\mu = 1.38$) film to increase the transmission of normally incident light $\lambda = 5800 \text{ \AA}$. What minimum film thickness should be deposited on the lens? (RTMNU W-05)
17. A material having an index of refraction of 1.3 is used to coat a piece of glass. What should be the minimum thickness of this film in order to minimize the reflected light at a wavelength of 500 nm ? What should be the refractive index of the glass to get best effects? Why? (RTMNU W-06)
18. In an experiment with Michelson's interferometer the readings of two consecutive positions of the movable mirror for the maximum distinctness of fringes were found to be 1.2829 mm and 1.5774 mm. If the mean wavelength of sodium D-lines is 5893 Å, find the difference between the two wavelengths. [Ans: 5.896 Å]
19. Fringes of equal inclination are observed in a Michelson interferometer. As one of the mirrors is moved back by 1 mm, 3663 fringes move out from the centre of the pattern. Calculate the wavelength of light used. [Ans: 5460 Å]

CHAPTER

7

Diffraction

7.1 INTRODUCTION

It is a matter of common experience that the path of light entering a dark room through a hole in the window illuminated by sunlight is straight. Similarly, if an opaque obstacle is placed in the path of light, a sharp shadow is cast on the screen, indicating thereby that light travels in straight lines. But it has been observed that when a beam of light passes through a small opening (a small circular hole or a narrow slit) it spreads to some extent into the region of the geometrical shadow also. If light energy is propagated in the form of waves, then similar to sound waves, one would expect bending of a beam of light round the edges of an opaque obstacle or illumination of the geometrical shadow. Diffraction phenomena are part of our common experience. The luminous border that surrounds the profile of a mountain just before the sun rises behind it, the light streaks that one sees while looking at a strong source of light with half shut eyes and the colored spectra (arranged in the form of a cross) that one sees while viewing a distant source of light through a fine piece of cloth, the colours seen on a compact disc etc are all examples of diffraction effects. Diffraction sets a limit to the image formation ability of optical instruments.

7.2 DIFFRACTION

Diffraction phenomenon is a common characteristic of all kinds of waves. It is a matter of common experience that sound waves readily bend around walls and buildings. When waves pass near an obstacle (barrier), they tend to bend around the edges of the obstacle. *The bending of waves around an obstacle and deviation from a rectilinear path is called diffraction.*

Dependence of the phenomenon on wavelength

Fig. 7.1 illustrates the passage of waves through an opening. When the opening is large compared to the wavelength, the waves do not bend round the edges. When the opening is

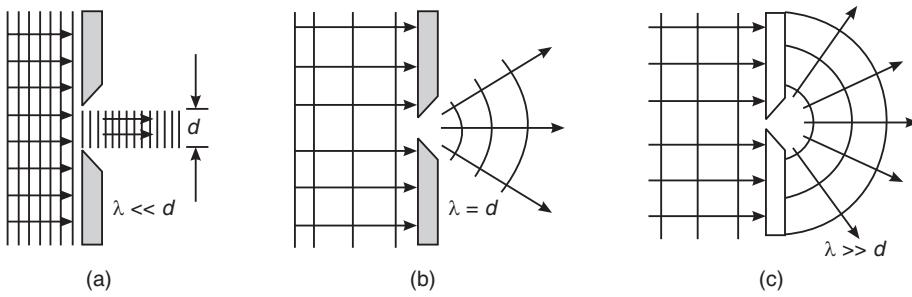


Fig. 7.1

small, the bending effect round the edges is noticeable. When the opening is very small (of the order of one wavelength), the waves spread over the entire surface behind the opening. The opening acts as an independent source of waves, which propagate in all directions. The diffraction effect is observable quite *close to the opening* when the size of the opening is very small. When the opening is large, diffraction effect is observed at *greater distances from the opening*. In general diffraction of light waves become noticeable only when *the size of the obstacle is comparable to a wavelength of light*.

Diffraction pattern

Diffraction of light waves is not readily apparent since light wavelength is very small compared to any physical obstacle. If an opaque obstacle is placed in the path of light, a sharp shadow is cast on the screen, giving the impression that light travels in straight lines. Similarly, the path of light entering a dark room through a hole in the door is seen to be straight. Thus, rectilinear propagation of light is apparent in day-to-day experience. However, careful experiments reveal that light passing through a tiny opening such as slit or aperture produces alternate regions of darkness and brightness beyond the region of geometric shadow. Such alternate bright and dark bands are known as the **diffraction pattern**. The bright central portion is known as a **central maximum** and it is bounded on either side by a series of secondary maxima separated by dark bands, called **minima**; each successive bright band becomes less intense proceeding outward and away from the central maximum. The presence of the bright region clearly indicates that light has reached there, showing thereby bending of light beam. The maxima and minima are created by interference of diffracted light waves. Diffraction phenomenon in optics is a fundamental demonstration that light behaves like a wave. But the diffraction pattern in the geometrical shadow of an obstacle is not commonly observed because the light sources are not point sources and secondly the obstacles used are of very large size compared to the wavelength of light.

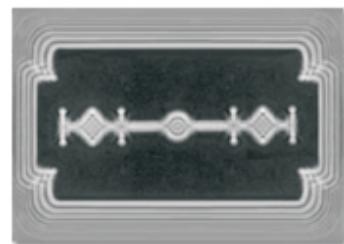


Fig. 7.2

A simple way to observe diffraction

A simple way to observe diffraction of light is by holding our hand in front of a strong light source and observing the light transmitted through the opening between any two fingers. When the fingers are brought very close to each other, one sees a series of dark lines parallel to the fingers. The diffraction bands produced by a razor blade are shown in Fig. 7.2.

Appearance of maxima and minima is due to interference of secondary wavelets

In general, *diffraction occurs whenever a portion of a wave front is obstructed in some way*. The basic idea that explains diffraction is based on Huygens theory. The behaviour of light beyond the screen with an aperture can be qualitatively explained with the aid of Huygens' Principle. The portion of the wave front that is incident on the opaque portion of the screen is obstructed while a small portion of the wave front is allowed to pass through the aperture (see Fig. 7.3). Every point on this portion of the wave front acts as a center of spherical secondary wavelets. Constructing the envelope of these secondary wavelets, we

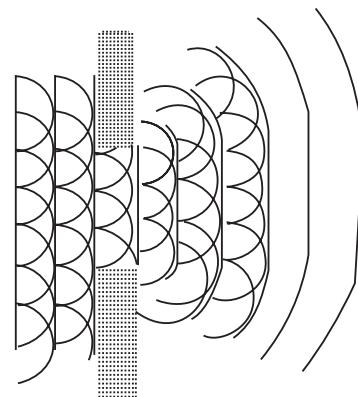


Fig. 7.3

find that the wave spreads into the region of geometric shadow bending around the edges of the aperture.

Fresnel fused the Huygens' principle with the Young's concept of interference of waves and explained the occurrence of alternate bright and dark bands as due to the superposition and interference of waves. The points on the primary wave front are mutually coherent and the secondary waves emitted by them are therefore coherent (and are also of the same frequency as that of the primary wave) and interfere.

7.3 DISTINCTION BETWEEN INTERFERENCE AND DIFFRACTION

The main differences between interference and diffraction are as follows:

<i>Interference</i>	<i>Diffraction</i>
1. Interference is the result of interaction of light coming from different wave fronts originating from the source.	1. Diffraction is the result of interaction of light coming from different parts of the same wave front.
2. Interference fringes may or may not be of the same width.	2. Diffraction fringes are not of the same width.
3. Regions of minimum intensity are perfectly dark.	3. Points of minimum intensity are not perfectly dark.
4. All bright bands are of same intensity.	4. All bright bands are not of same intensity.

7.4 THE TWO TYPES OF DIFFRACTION

The diffraction phenomena are broadly classified into two types: Fresnel diffraction and Fraunhofer diffraction.

1. **Fresnel diffraction:** In this type of diffraction, the source of light and the screen are effectively at finite distances from the obstacle (see Fig. 7.4a). Lenses are not used to make the rays parallel or convergent. The incident wave front is not planar. As a result, the phase of secondary wavelets is not the same at all points in the plane of the obstacle. The resultant amplitude at any point of the screen is obtained by the mutual interference of secondary wavelets from different elements of unblocked portions of wave front. It is experimentally simple but the analysis proves to be very complex.

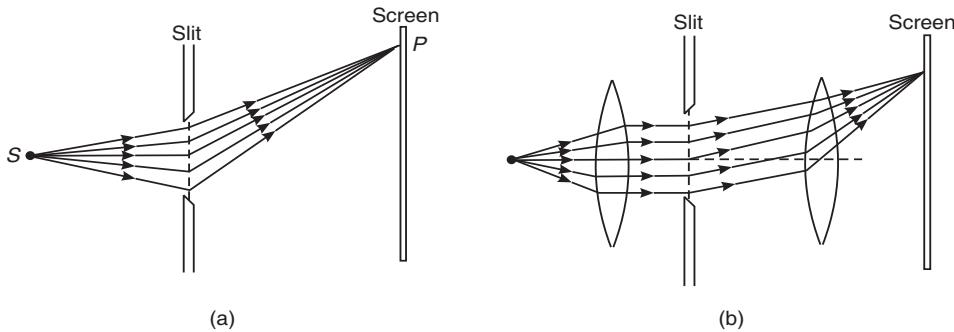


Fig. 7.4: Conditions for (a) Fresnel diffraction and (b) Fraunhofer diffraction

2. **Fraunhofer diffraction:** In this type of diffraction, the source of light and the screen are effectively at infinite distances from the obstacle (see Fig. 7.4b). The conditions required for Fraunhofer diffraction are achieved using two convex lenses, one to make the light from the source parallel and the other to focus the light

after diffraction on to the screen. The incident wave front as such is plane and the secondary wavelets, which originate from the unblocked portions of the wave front, are in the same phase at every point in the plane of the obstacle. The diffraction is produced by the interference between parallel rays, which are brought into focus with the help of a convex lens. This problem is simple to handle mathematically because the rays are parallel.

7.5 FRAUNHOFFER DIFFRACTION AT A SINGLE SLIT

Let us consider the arrangement in Fig. 7.5 to obtain Fraunhofer diffraction by a single slit. The lenses L_1 and L_2 keep the source and the screen effectively at infinity. The first lens renders the beam parallel while the second makes the screen to receive the parallel rays converging to a point, P. Let us consider a single narrow slit AB of width d perpendicular to the plane of the page.

Let it be illuminated by a parallel beam of monochromatic light of wavelength λ . According to the considerations of geometrical optics, a sharp image of the slit is expected to form at P in the focal plane of the lens. However, what we obtain on the screen is a slit image of maximum brightness at the center, followed by the secondary maxima on the either side of gradually decreasing intensities, with distance. This intensity distribution on the screen is known as the Fraunhofer diffraction pattern. Let us try to understand the reasons for this unique distribution of intensity in the image of a single slit.

7.5.1 Formation of Maxima and Minima

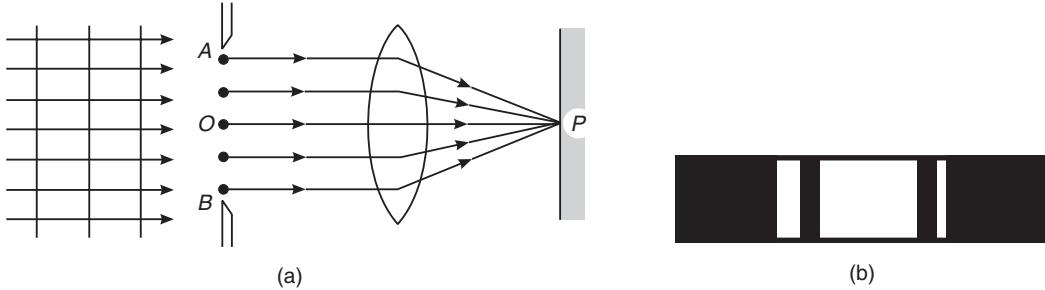


Fig. 7.6: Fraunhofer diffraction at a single slit (a) Conditions at the central maximum. The rays parallel to axis come to focus at P giving a bright band. (b) Typical diffraction pattern consisting of a central bright band flanked by weaker bright bands.

Fig. 7.6 shows a plane wave front (parallel rays) incident on the slit AB. A small part AB is sliced off from the incident wave front. According to Huygens' principle, each point on AB acts as a source of secondary wavelets. It would then be appropriate to replace the wave front AB with a string of point sources. As all points on AB are in phase, the point sources will be coherent. Hence, light from one portion of the slit can interfere with light from another portion and the resultant intensity on the screen will depend on the direction θ . The secondary wavelets travelling parallel to OP come to focus at P. The waves from points equidistant from 'O' and situated in the upper and lower halves OA and OB start in phase. They will travel the same distance in reaching P. The optical path difference is therefore zero and the waves will

be in phase at P. They reinforce each other to produce an intensity maximum at P. It is at the centre of the diffraction pattern and is called *zero order central maxima*.

For any other point like Q on the screen (Fig. 7.7), the light from different parts of the aperture travels different distances and the difference increases as we consider points at increasing distance from P. The path difference between the waves reaching the points from different parts of the aperture increases gradually.

Now consider the secondary waves travelling in a direction making an angle θ with OP. These secondary waves are brought to focus by the lens at a point Q, which will have a maximum or minimum intensity, depending on the path difference between the waves arriving at Q from different points on the wave front AB. It is convenient to divide the wave front AB into two halves AO and OB. A line AM is drawn perpendicular to the direction of the diffracted rays. Waves ON and BM are in phase at the slit. Wave BM travels farther than ON.

The path difference between these wavelets = ON = $(d/2) \sin\theta$.

If ON = $\lambda/2$, the two waves interfere destructively and produce darkness at Q. This is true for any two waves that originate at points separated by $(d/2)$, as the path difference between such points will be $\lambda/2$. For every point in the upper half OA, there is a corresponding point in the lower half OB. The path difference between the waves from these corresponding points will be $\lambda/2$. Hence, the waves from the upper half AO interfere destructively with waves from the lower half OB, if

$$\frac{d}{2} \sin \theta = \frac{\lambda}{2}$$

i.e., $\sin \theta = \lambda/d$ (7.1)

Therefore, the intensity at Q is zero and a dark band called the *first order minimum* is produced at Q. A similar dark band occurs at Q' below P at an angular distance θ governed by the equation (7.1). It is also called the *first order minimum*. We may divide the slit into four parts, six parts and so on. Arguments similar to the above show that a dark band occurs whenever

$$\sin \theta = 2\lambda/d, 3\lambda/d, 4\lambda/d, 5\lambda/d, \dots \text{etc.}$$

They are known as second order minimum, third order minimum, etc. In general minima appear if the following condition is satisfied:

$$\sin \theta_m = m\lambda/d \quad \text{condition for minimum} \quad (7.2)$$

where $m = 1, 2, 3, \dots$

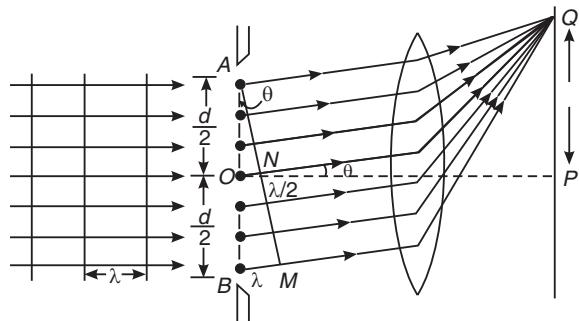


Fig. 7.7: Conditions at the first minimum of diffraction pattern-Each point on the sliced portion AB of the wave front acts as a point source. Any two waves that originate at points separated by $d/2$ distance are 180° out of phase and interfere destructively.

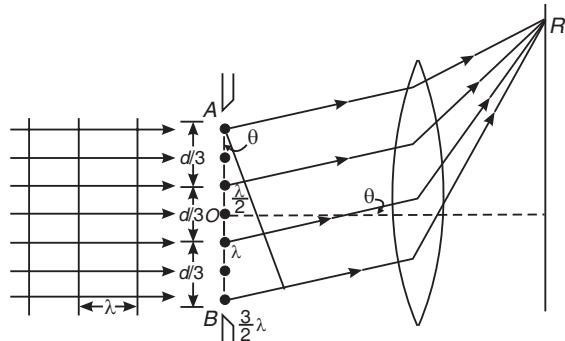


Fig. 7.8: Condition for first order maxima of diffraction pattern

More generally, the condition for minima may be expressed as

$$d \sin \theta = \pm m \lambda \quad (7.2a)$$

In addition to the central maximum there are *secondary maxima*, which lie in between the secondary minima on either side of the central maximum. These are located in a direction in which the path difference ON is an odd multiple of $\lambda/2$. Hence for secondary maxima

$$ON = d \sin \theta = (2m + 1)\lambda/2$$

In general the secondary maximum are given by

$$\sin \theta_m = \left(\frac{2m+1}{d} \right) \frac{\lambda}{2} \quad \text{condition for maxima} \quad (7.3)$$

Thus, the diffraction pattern due to a single slit consists of a central bright maximum flanked by secondary maxima and minima on both the sides.

7.5.2 Intensity Distribution in Diffraction Pattern Due to a Single Slit

Let a plane wave be incident normally on a long narrow slit of width d . Let us imagine that this slit width d is divided into N parallel strips of each of width Δx .

Each strip acts as a secondary source of radiation giving out wavelets leading to a characteristic distribution of intensity at a point Q. The position of the point Q in the Fig. 7.7 is fixed by the angle θ . The wavelets from adjacent strips have a path difference of $\Delta x \sin \theta$. The resultant of all the wavelets reaching Q can be estimated by integration.

Let the disturbance caused at Q by the unit width of the slit be

$$y_0 = A \cos \omega t$$

Amplitude of the wavelet from the width dx at A, when it reaches Q is ($A dx$).

$$\text{Phase of this wavelet at } Q = \omega t + \frac{2\pi}{\lambda} \times OQ = \omega t + \frac{2\pi}{\lambda} \times x \sin \theta$$

where x is the distance of Q from O, the mid point of the slit.

The disturbance caused at Q by the wavelet is given by

$$dy = A dx \cos \left[\omega t + \frac{2\pi x \sin \theta}{\lambda} \right]$$

The total disturbance at Q by all wavelets from the slit of width a is given by

$$y = \int_{-a/2}^{+a/2} dy = \int_{-a/2}^{+a/2} A \cos \left[\omega t + \frac{2\pi x \sin \theta}{\lambda} \right] dx$$

$$\text{Let } \frac{2\pi \sin \theta}{\lambda} = k. \text{ Therefore, } \cos \left[\omega t + \frac{2\pi x \sin \theta}{\lambda} \right] = \cos(\omega t + kx)$$

But

$$\cos(\omega t + kx) = \cos \omega t \cdot \cos kx - \sin \omega t \cdot \sin kx$$

$$\therefore y = \int_{-a/2}^{+a/2} A [\cos \omega t \cdot \cos kx - \sin \omega t \cdot \sin kx] dx$$

$$y = \int_{-a/2}^{+a/2} A [x \cos \omega t \cdot \cos kx] dx - \int_{-a/2}^{+a/2} A [\sin \omega t \cdot \sin kx] dx$$

$$y = A \cos \omega t \int_{-a/2}^{+a/2} \cos kx dx - A \sin \omega t \int_{-a/2}^{+a/2} \sin kx dx$$

$$\begin{aligned}
 y &= A \cos \omega t \left[\frac{\sin kx}{k} \right]_{-a/2}^{+a/2} - A \sin \omega t \left[\frac{-\cos kx}{k} \right]_{-a/2}^{+a/2} \\
 y &= A \cos \omega t \left[\frac{\sin \left(\frac{ka}{2} \right) + \sin \left(\frac{ka}{2} \right)}{k} \right] + A \sin \omega t \left[\frac{\cos \left(\frac{ka}{2} \right) - \cos \left(\frac{ka}{2} \right)}{k} \right] \\
 y &= A \cos \omega t \left[\frac{2 \sin \left(\frac{ka}{2} \right)}{k} \right] = 2A \cos \omega t \left[\frac{\sin \left(\frac{2\pi \cdot \sin \theta \cdot a}{\lambda} \right)}{\frac{2\pi \cdot \sin \theta}{\lambda}} \right] \\
 y &= \left[A \frac{\sin \left(\frac{\pi a \sin \theta}{\lambda} \right)}{\frac{\pi \sin \theta}{\lambda}} \right] \cos \omega t = \left[Aa \frac{\sin \left(\frac{\pi a \sin \theta}{\lambda} \right)}{\frac{\pi a \sin \theta}{\lambda}} \right] \cos \omega t \quad (7.4)
 \end{aligned}$$

or $y = A_0 \cos \omega t$

The quantity in brackets gives the amplitude A_0 of the resultant disturbance at Q.

For $\theta = 0$, $A_0 = A_0 = Aa$.

$$y = A_0 \left[\frac{\sin \left(\frac{\pi a \sin \theta}{\lambda} \right)}{\frac{\pi a \sin \theta}{\lambda}} \right] \cos \omega t \quad (7.5)$$

Let $\alpha = \frac{\pi a \sin \theta}{\lambda}$.

$$\therefore y = A_0 \cos \omega t = A_0 \left[\frac{\sin \alpha}{\alpha} \right] \cos \omega t$$

$$\therefore A_0 = A_0 \left[\frac{\sin \alpha}{\alpha} \right] \quad (7.6)$$

Intensity distribution is given by

$$I = I_0 \left[\frac{\sin \alpha}{\alpha} \right]^2 \quad (7.7)$$

where I_0 is the intensity of principal maximum at $\theta = 0$.

Thus, the intensity at any point on the screen is proportional to $\left(\frac{\sin \alpha}{\alpha} \right)^2$. A phase

difference of 2π corresponds to a path difference of λ . Therefore, a phase difference of 2α is given by

$$2\alpha = \frac{2\pi}{\lambda} a \sin \theta \quad (7.8)$$

where $a \sin \theta$ is the path difference between the secondary waves from A and B (Fig. 7.5).

$$\alpha = \frac{\pi}{\lambda} a \sin \theta \quad (7.9)$$

Thus, the value of α depends on the angle of the diffraction θ . The value of $\left(\frac{\sin^2 \alpha}{\alpha^2}\right)$ for different values of θ gives the intensity at the point under consideration. Fig. 7.9 represents the intensity distribution. It is a graph of $\left(\frac{\sin^2 \alpha}{\alpha^2}\right)$ (along the Y-axis), as a function of α or $\sin \theta$ (along the X-axis).

It is seen that most of the light is confined to the central band and its intensity is far greater than that of any other maxima. The intensity of the secondary maxima falls off rapidly as one moves away from the centre. The intensity of the first secondary maximum is about $1/22$ and that of the second is $1/61$ of the intensity of the principal maximum. The secondary maxima are too faint to be visible ordinarily.

7.5.3 Linear Width of the Principal Maximum

If x is the distance of the first secondary minimum from the center of the principal maximum, then the width of the central maximum is

$$W = 2x$$

If the lens L is very near to the slit or the screen is far away from the lens, and f is the focal length of the lens, then $OP = f$ is very large (Fig. 7.6a). As a result,

$$\sin \theta = \frac{x}{f}$$

But, for the first minimum $\sin \theta = \frac{\lambda}{d}$

$$\therefore \frac{x}{f} = \frac{\lambda}{d} \quad \text{or} \quad x = \frac{f\lambda}{d} \quad (7.10)$$

Hence, the linear width of the central maximum is given by

$$W = 2x = \frac{2f\lambda}{d} \quad (7.11)$$

Note that when the slit width $d \gg \lambda$, we see on the screen uniform illumination in the shape of the slit. As the slit width is reduced, the illumination starts to spread out and dark bands become visible. Further, the width of the central maximum increases as the slit narrows, as illustrated in Fig. 7.10.

The position of the first secondary maxima on either side of the central maxima is given by

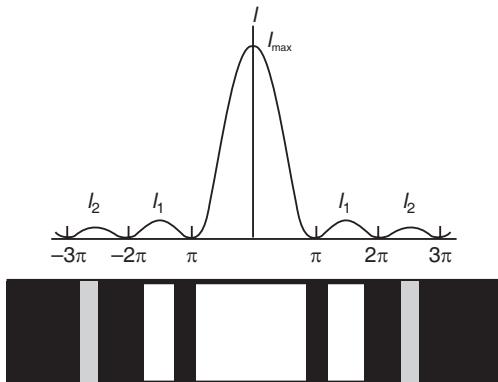


Fig. 7.9

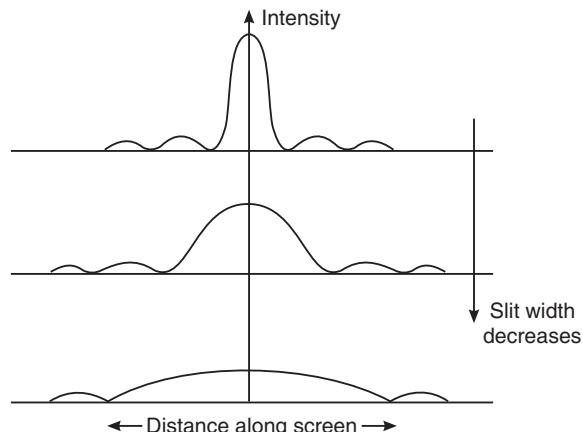


Fig. 7.10: Central maximum increases in width as the slit width decreases

$$\sin \theta_1 = \frac{3\lambda}{2d}$$

Therefore, in this case $\frac{x}{f} = \frac{3\lambda}{2d}$

and $x = \frac{3f\lambda}{2d}$ (7.12)

Example 7.1: In a single slit diffraction pattern the distance between the first minima on either side of the central zero maximum is 4.4 mm as observed on a screen at a distance of 0.7 m. The wavelength of light used is 5890 Å. Calculate the slit width.

Solution: The slit width, $d = \frac{f\lambda}{x} = \frac{(0.7\text{m})(5890 \times 10^{-10}\text{m})}{4.4 \times 10^{-3}\text{m}} = 0.09\text{ mm}$

Example 7.2: Parallel light (5000 Å) is normally incident on a single slit. The central maximum fans out at 30° on both sides of the direction of the incident light. Calculate the slit width. For what width of the slit the central maximum would spread out to 90° from the direction of the incident light?

Solution: In a Fraunhofer diffraction pattern due to single slit of width a , the directions of minima are given by

$$a \sin \theta = \pm m \lambda \quad m = 1, 2, 3, \dots$$

Therefore, the angular spread of the central maximum on either side of the incident light is given by

$$\sin \theta = \lambda/a$$

Here $\theta = 30^\circ$ so that $\sin \theta = 0.5$ and $\lambda = 5000 \text{ \AA} = 5 \times 10^{-5} \text{ cm}$.

$$\therefore a = \frac{\lambda}{\sin \theta} = \frac{5 \times 10^{-5} \text{ cm}}{0.5} = 1 \times 10^{-4} \text{ cm}$$

For $\theta = 90^\circ$

$$\therefore a = \frac{\lambda}{\sin 90^\circ} = \frac{5 \times 10^{-5} \text{ cm}}{1} = 5 \times 10^{-5} \text{ cm} = 5000 \text{ \AA}$$

Example 7.3: Calculate the angles at which the first dark band and the next bright band are formed in the Fraunhofer diffraction pattern of a slit 0.3 mm wide ($\lambda = 5890 \text{ \AA}$).

Solution: In a single slit Fraunhofer's pattern, the directions θ of the minima are given by

$$a \sin \theta = \pm m \lambda \quad m = 1, 2, 3, \dots$$

For the first dark band on either side of the central maximum, $m = 1$.

$$\therefore a \sin \theta = \lambda$$

$$\sin \theta = \frac{\lambda}{a} = \frac{5890 \times 10^{-8} \text{ cm}}{0.03 \text{ cm}} = 0.00196$$

$$\therefore \theta = 6^\circ 7'$$

The angle of diffraction θ' corresponding to the first bright band on either side of the central maximum is approximately given by

$$a \sin \theta' = 3\lambda/2$$

$$\therefore \sin \theta' = \frac{3\lambda}{2a} = \frac{3}{2} \times 0.00196 = 0.00294$$

$$\theta' = 10'$$

7.6 FRAUNHOFER DIFFRACTION AT DOUBLE SLIT

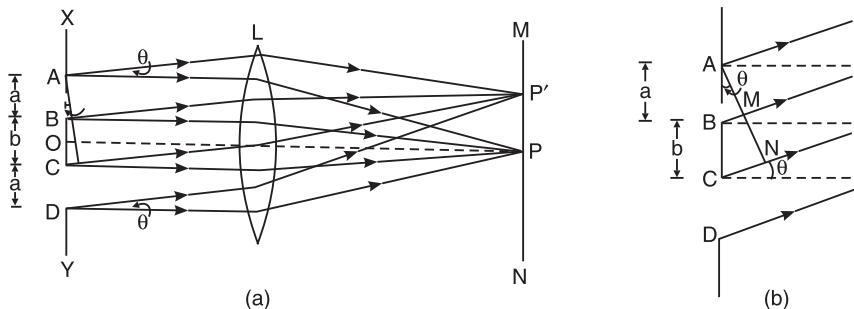


Fig. 7.11

In Fig. 7.11(a), AB and CD are two rectangular slits parallel to one another and perpendicular to the plane of the paper. The width of each slit (open portion) is a and the width of the *opaque portion* is b . L is a collecting lens and MN is a screen perpendicular to the plane of the paper. P is a point on the screen such that OP is perpendicular to the screen. Let a plane wave front be incident on the surface of XY. All the secondary waves traveling in a direction parallel to OP come to focus at P. Therefore, P corresponds to the position of the central bright maximum.

In this case, the diffraction pattern has to be considered in two parts (i) the interference phenomenon due to the secondary waves emanating from the corresponding points of the two slits and (ii) the diffraction pattern due to the secondary waves from the two slits individually. For calculating the positions of interference or maxima and minima, the angle is denoted as θ , where θ refers to the angle between the direction of the secondary waves and the initial direction of the incident light.

(i) Interference maxima and minima: Let us consider the secondary waves traveling in a direction inclined at an angle θ with the initial direction.

In the Δ^{le} CAN (Fig. 7.11 b)

$$\sin \theta = \frac{CN}{AC} = \frac{CN}{a+b}$$

or

$$CN = (a+b) \sin \theta$$

If this path difference, CN is equal to odd multiples of $\lambda/2$, θ gives the direction of minima due to interference of the secondary waves from the two slits.

$$\therefore CN = (a+b) \sin \theta_n = (2n+1) \frac{\lambda}{2} \quad (7.13)$$

Putting $n = 1, 2, 3$, etc, the values of $\theta_1, \theta_2, \theta_3$, etc, corresponding to the directions of minima can be obtained.

From equation (7.13)

$$\sin \theta_n = \frac{(2n+1)\lambda}{2(a+b)} \quad (7.14)$$

On the other hand, if the secondary waves travel in a direction θ' such that the path difference is even multiples of $\lambda/2$, then θ' gives the direction of the maxima due to interference of light waves emanating from the two slits.

$$\therefore CN = (a+b) \sin \theta'_n = 2n \frac{\lambda}{2}$$

or $\sin \theta'_n = \frac{n\lambda}{(a+b)}$ (7.15)

Putting $n = 1, 2, 3$ etc, $\theta'_1, \theta'_2, \theta'_3$ etc corresponding to the directions of the maxima can be obtained.

From equation (7.14), we get

$$\sin \theta_1 = \frac{3\lambda}{2(a+b)}$$

and

$$\sin \theta_2 = \frac{5\lambda}{2(a+b)}$$

$$\therefore \sin \theta_2 - \sin \theta_1 = \frac{\lambda}{(a+b)} \quad (7.16)$$

Similarly, it can be seen from eq. (7.15), that $\sin \theta'_2 - \sin \theta'_1 = \frac{\lambda}{(a+b)}$.

Thus, the angular separation between any two consecutive minima (or maxima) is equal to $\frac{\lambda}{(a+b)}$. The angular separation is inversely proportional to $(a+b)$, the distance between the two slits.

(ii) Diffraction maxima and Minima: Let us consider the secondary waves traveling in a direction inclined at an angle θ with the initial direction of the incident light. If the path difference BM is equal to λ the wavelength of the light used, then θ will give the direction of the diffraction minimum (Fig. 7.11 b). That is, the path difference between secondary waves emanating from the extremities of a slit (i.e., points A and B) is equal to λ . Considering the wave front on AB to be made up of the two halves, the path difference between the corresponding points of the upper and lower halves is equal to $\lambda/2$. The effect at P' due to the wave front incident on AB is zero. Similarly, for the same direction of the secondary waves, the effect at P' due to the wave front incident on the slit CD is also zero. In general,

$$a \sin \theta_n = n\lambda$$

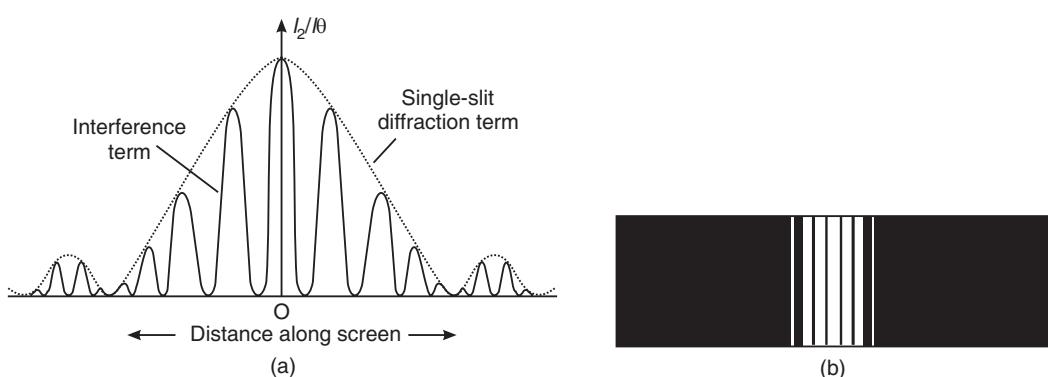


Fig. 7.12

Putting $n = 1, 2, 3$, etc, the values of $\theta_1, \theta_2, \theta_3$ etc, corresponding to the directions of diffraction minima can be obtained.

Intensity distribution due to the Fraunhofer diffraction at two parallel slits is shown in Fig. 7.12 (a). The full line represents equally spaced interference maxima and minima and the

dotted curve represents the diffraction maxima and minima. In the region originally occupied by the central maximum of the single slit diffraction pattern, equally spaced interference maxima and minima are observed. The intensity of the central interference maximum is four times the intensity of the central maximum of the single slit diffraction pattern. The intensity of other interference maxima on the two sides of the central maximum gradually decreases. In the region of the secondary maxima due to diffraction at a single slit, equally spaced interference maxima of low intensity are observed. A typical diffraction pattern obtained on the screen is shown in Fig. 7.12 (b).

7.6.1 Expression for Resultant Intensity

Let us now find an expression for the resultant intensity and its variation with the angle θ . By Huygens' principle, every point in the slits AB and CD sends out secondary wavelets in all directions. From the theory of diffraction at a single slit, the resultant amplitude due to wavelets diffracted from each slit in a direction θ is

$$A_\theta = A_0 \left[\frac{\sin \alpha}{\alpha} \right]$$

In the case of double slit, we have interference of two waves of amplitude A_0 each, having a phase difference $\delta = \frac{2\pi}{\lambda} d \sin \theta$. The

resultant amplitude R is given by (see Fig. 7.13)

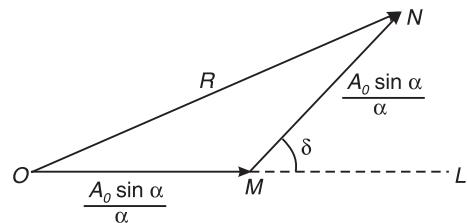


Fig. 7.13

$$ON^2 = OM^2 + MN^2 + 2(OM)(MN) \cos NML$$

$$\begin{aligned} R^2 &= \left[\frac{A_0 \sin \alpha}{\alpha} \right]^2 + \left[\frac{A_0 \sin \alpha}{\alpha} \right]^2 + 2 \left[\frac{A_0 \sin \alpha}{\alpha} \right] \left[\frac{A_0 \sin \alpha}{\alpha} \right] \cos \delta \\ &= 2 \left[\frac{A_0 \sin \alpha}{\alpha} \right]^2 (1 + \cos \delta) = 2 \left[\frac{A_0 \sin \alpha}{\alpha} \right]^2 \left(1 + 2 \cos^2 \frac{\delta}{2} - 1 \right) \end{aligned}$$

$$\text{or } R^2 = 4A_0^2 \frac{\sin^2 \alpha}{\alpha^2} \cos^2 \frac{\delta}{2}$$

$$\therefore I = 4I_0 \frac{\sin^2 \alpha}{\alpha} \cos^2 \frac{\delta}{2} = 4I_0 \frac{\sin^2 \alpha}{\alpha} \cos^2 \beta \quad (7.17)$$

where $\beta = \delta/2$.

Thus, the resultant intensity at any point depends on two factors –

- (i) The factor $I_0 \frac{\sin^2 \alpha}{\alpha}$ is the same as that derived for a single slit Fraunhofer diffraction. It represents the intensity variation in the diffraction pattern due to any individual slit.
- (ii) The factor $\cos^2 \beta$ gives the interference pattern due to waves overlapping from the two slits.

The resultant intensity at any point on the screen is given by the product of these two factors and will be zero when either of these factors is zero.

Maxima and minima:

1. The diffraction term $I_0 \left(\frac{\sin^2 \alpha}{\alpha^2} \right)$ gives the central maximum in the direction $\theta = 0$ having alternate minima and secondary maxima of decreasing intensity on either side.

The angular positions of minima are given by,

$$\sin \alpha = 0, \quad \alpha = \pm m\pi, \quad m = 1, 2, 3, \dots$$

i.e.
$$\frac{\pi a \sin \theta}{\lambda} = \pm m\pi$$

or
$$\sin \theta = \pm \frac{m\lambda}{a} \quad m = 1, 2, 3, \dots \text{ but not zero.}$$

The angular positions of secondary maxima approach to

$$\alpha = \pm \frac{3\pi}{2}, \pm \frac{5\pi}{2}, \pm \frac{7\pi}{2}, \dots$$

2. According to the second factor ($\cos^2 \beta$), the intensity will be maximum when

$$\cos^2 \beta = 1.$$

i.e.,
$$\cos^2 \beta = 1, \text{ where } n = 0, 1, 2, 3, \dots$$

$$\therefore \frac{\pi d \sin \theta}{\lambda} = \pm n\pi \quad \text{when } n = 0, \theta = 0. \quad (7.18)$$

Thus, the central maximum of interference pattern lies along the direction of incident light. This is called the **principal maximum of zero order**. The central maximum of diffraction pattern also lies along this direction. At this point all the waves arrive with the same phase. Hence, the intensity of central maximum is the highest.

The intensity will be maximum when

$$\cos^2 \beta = 0$$

i.e.,
$$\frac{\delta}{2} = \pm (2n+1) \frac{\pi}{2}$$

$$\therefore \frac{\pi d \sin \theta}{\lambda} = \pm (2n+1) \frac{\pi}{2}$$

$$d \sin \theta = \pm (2n+1) \frac{\lambda}{2} \quad (7.19)$$

It can be shown that for small values of θ , the maxima and minima are equally spaced. If a is kept constant and d is varied, the positions of maxima and minima due to diffraction remain unchanged while those due to interference undergo a change. Fig. 7.14 represent the intensity distribution determined

by the factor $\cos^2 \beta$ and $I_0 \left(\frac{\sin^2 \alpha}{\alpha^2} \right)$

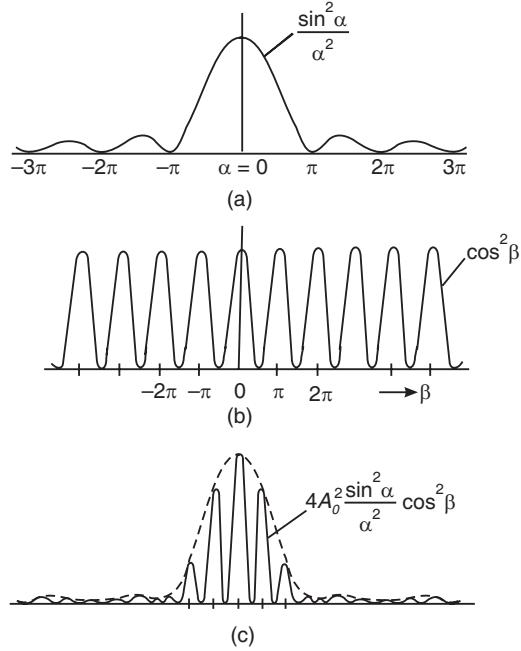


Fig. 7.14

respectively. The resultant of these curves is shown in Fig. 7.14. The resultant is obtained by multiplying the ordinates of first two curves at every point.

The entire pattern may be considered as consisting of interference fringes due to interference of light from both the slits, their intensity is being modulated by the diffraction occurring at individual slits.

7.7 DIFFRACTION DUE TO N-SLITS—DIFFRACTION GRATING (NORMAL INCIDENCE)

Let us now consider the diffraction pattern produced by N -slits, each of width a . The separation between consecutive slits is $d = a + b$, where a is the width of the open portion and b is the width of the opaque portion. Such a device consisting of a large number of parallel slits of equal width and separated from one another by equal opaque spaces is called a **diffraction grating**. The distance d between the centres of the adjacent slits is known as the **grating period**.

Rowland (1848-1901) produced **transmission gratings** by ruling extremely close, equidistant and parallel lines on optically plane glass plates with a diamond point. The rulings (diamond scratch) scatter light and are effectively opaque while the parts without ruling transmit light and act as slits.

Because of the expenses and difficulty involved in fabrication, commonly used gratings are reproduced from the original ruled gratings. The **replica gratings** are made by pouring a thin layer of collodion solution over the surface of a ruled grating and the solution is allowed to harden. The collodion film is peeled carefully afterwards from the grating. The film retains the impression of the rulings of the original grating in the form of ridges. The ruled lines, which scatter light, act as opaque spaces whereas the spaces between them which transmit incident light act as parallel slits. The film is mounted between glass plates and it acts as a *plane transmission grating*. The number of lines on a plane transmission grating is of the order of 6000 lines per cm.

7.8 PLANE DIFFRACTION GRATING - THEORY

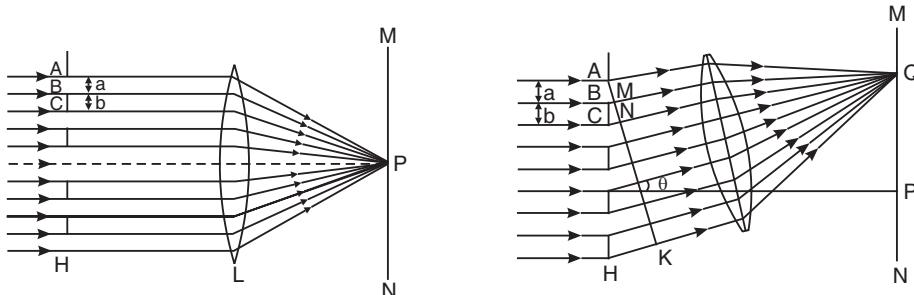


Fig. 7.15

Let us consider the plane transmission grating held normal to the plane of the page (Fig. 7.15 a) and represented by the section ABC...H. Let the width of the transparent portion AB be equal to a and opaque portion BC be equal to b . The distance $(a + b) = d$ and is called the *grating constant* or *grating period*. Let a parallel beam of monochromatic light of wavelength λ be incident normally on the grating surface. Then all the secondary waves travelling in the same direction as that of the incident light will come to focus at the point P on the screen. The screen is placed at the focal plane of the collecting lens, L. The point P where all the secondary waves reinforce one another corresponds to the position of the *central bright maximum*.

Now let us consider the secondary waves travelling in a direction inclined at an angle θ with the direction of the incident light (Fig. 7.15 b). The waves travel different distances and it is obvious that there is a path difference between the waves coming out from each slit and bending at an angle θ . These secondary waves come to focus at the point Q on the screen. The intensity at Q will depend on the path difference between the secondary waves originating from the corresponding points A and C of two neighbouring slits. In the Fig. 7.15 (b), AB = a and BC = b . The path difference between the secondary waves starting from A and C is equal to $AC \sin \theta$.

But

$$AC = AB + BC = a + b$$

$$\text{Path difference} = AC \sin \theta$$

$$= (a + b) \sin \theta$$

The point Q will be of maximum intensity if this path difference is equal to integral multiples of λ . It means that all the secondary waves originating from the corresponding points of the neighbouring slits reinforce one another and the angle θ gives the direction of maximum intensity. In general

$$(a + b) \sin \theta_m = m\lambda \quad (7.20)$$

where θ_m is the direction of the m^{th} principal maximum. If $(a + b) \sin \theta = \lambda$, we obtain maximum intensity at Q. When $(a + b) \sin \theta = 2\lambda$, there will be again a maximum and so on. Between the central maximum P and the first maximum at Q there will be minimum intensity and so on.

Similar maxima and minima are obtained on the other side of central maximum. Thus, on each side of the central maximum at P, principal maxima and minimum intensity are observed due to diffracted light. The position of m^{th} minimum is given by

$$(a + b) \sin \theta_m = (2m + 1)\lambda/2. \quad (7.21)$$

7.8.1 Intensity Distribution

When illuminated by a beam of monochromatic radiation, the system produces N wavelets at an angle θ , each of the amplitude $A_\theta = A_0 \frac{\sin \alpha}{\alpha}$. The phase difference between successive wavelets is $\delta = \frac{2\pi d \sin \theta}{\lambda}$.

The resultant amplitude can be expressed as

$$y = A_0 [\cos \omega t + \cos(\omega t + \delta) + \cos(\omega t + 2\delta) + \dots + \cos(\omega t + N\delta)]$$

Expressing the amplitude terms as real parts of complex numbers, we have

$$y = A_0 e^{j\omega t} \left[1 + e^{j\delta} + e^{2j\delta} + \dots \right] = A_0 e^{j\omega t} \left[\frac{1 - e^{jN\delta}}{1 - e^{j\delta}} \right]$$

We get the intensity by multiplying the amplitude with its complex conjugate.

$$\begin{aligned} I = A^2 &= A_0^2 \left[\frac{(1 - e^{jN\delta})(1 - e^{-jN\delta})}{(1 - e^{j\delta})(1 - e^{-j\delta})} \right] = A_0^2 \left[\frac{1 - \cos N\delta}{1 - \cos \delta} \right] \\ &= A_0^2 \frac{\sin^2 \frac{N}{2}\delta}{\sin^2 \frac{\delta}{2}} = A_0^2 \frac{\sin^2 \left[\frac{N\pi d \sin \theta}{\lambda} \right]}{\sin^2 \left[\frac{\pi d \sin \theta}{\lambda} \right]} \end{aligned}$$

or $I = I_0 \left(\frac{\sin^2 \alpha}{\alpha^2} \right) \frac{\sin^2 \left[\frac{N\pi d \sin \theta}{\lambda} \right]}{\sin^2 \left[\frac{\pi d \sin \theta}{\lambda} \right]}$

or $I = I_0 \left(\frac{\sin^2 \alpha}{\alpha^2} \right) \frac{\sin^2 N\gamma}{\sin^2 \gamma} \quad (7.22)$

where $\gamma = \frac{\pi d \sin \theta}{\lambda}$.

The expression for intensity is a product of two terms. The term $I_0 \left(\frac{\sin^2 \alpha}{\alpha^2} \right)$ represents the intensity distribution due to a single slit diffraction. The second term $\frac{\sin^2 N\gamma}{\sin^2 \gamma}$ represents the distribution of intensity due to interference produced by waves from N equally spaced point sources.

Principal Maxima

When $\sin \gamma = 0$, that is $\gamma = \pm n\pi \quad (n = 0, 1, 2, 3, \dots)$

We have $\sin N\gamma = 0$ and hence $\frac{\sin N\gamma}{\sin \gamma}$ becomes indeterminate.

According to L'Hospital's rule,

$$\lim_{\alpha \rightarrow m\pi} \frac{\sin N\gamma}{\sin \gamma} = \pm N$$

Therefore, $\lim_{\alpha \rightarrow \pm m\pi} \left[\frac{\sin N\gamma}{\sin \gamma} \right]^2 = N^2$

Substituting this value into equ.(7.22), we get $I = I_0 \left(\frac{\sin^2 \alpha}{\alpha^2} \right) N^2$, which is maximum.

Thus the resultant intensity of maxima is

$$I = I_0 \left(\frac{\sin^2 \alpha}{\alpha^2} \right) N^2. \quad (7.23)$$

Fig. 7.16 shows the intensity distribution determined by the factor

$$\left(\frac{\sin^2 \alpha}{\alpha^2} \right) \text{ and } \frac{\sin^2 N\gamma}{\sin^2 \gamma}$$

respectively. The resultant of these curves is obtained by multiplying the ordinates of first two curves at every point.

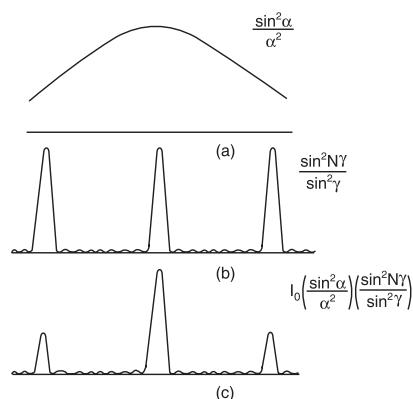


Fig. 7.16

As the maxima are very intense, they are called **principal maxima**. In order to find the resultant intensity of any of the principal maxima in the diffraction pattern, we have to multiply the square of the number of the slits (N^2) with the factor $I_0\left(\frac{\sin^2 \alpha}{\alpha^2}\right)$, which is the intensity due to a single slit.

The **direction of the principal maxima** are given by $\sin \gamma = 0$, that is $\gamma = \pm n\pi$.

$$\text{or} \quad d \sin \theta = \pm n\lambda \quad (7.24)$$

This equation is known as the **grating equation**.

If we put $n = 0$, we get $\theta = 0$. This is the direction in which waves from all slits arrive in phase and produce a bright central image. This maxima is called the **zero order principal maxima**. If we put $n = 1, 2, 3, \dots$ we obtain the directions of the first, second, third order principal maxima respectively. Therefore, the direction of occurrence of principal maxima is given by

$$\sin \theta_n = \frac{n\lambda}{d} = \frac{n\lambda}{a+b}$$

or

$$\sin \theta_n = nN\lambda \quad (7.25)$$

where $N = \frac{1}{a+b}$ is the number of ruled lines per unit width of the grating.

Minima

The intensity is zero when $\sin N\gamma = 0$, or $N\gamma = \pm m\pi$

$$\text{or} \quad N\gamma = \frac{N\pi d \sin \theta}{\lambda} = \pm m\pi$$

$$\text{or} \quad N \cdot d \sin \theta = \pm m\lambda \quad (7.26)$$

Here, m can take all integral values except 0, $N, 2N, 3N$, etc because these values give the positions of principal maxima. The positive and negative signs indicate that the minima of a given order lie symmetrically on both the sides of the central principal maxima.

It is seen from equ.(7.26) that $m = 0$ gives principal maximum of zero order while $m = 1, 2, 3, \dots, (N-1)$ give the minima. Then $m = N$ gives principal maximum of first order. Thus, between zero order, and first order principal maxima we have $(N-1)$ minima. Similarly, it can be shown that there are $(N-1)$ minima between first order and second order principal maxima and so on. Between two such consecutive minima, the intensity has to be maximum, and these maxima are known as **secondary maxima**. The secondary maxima are not visible in the grating spectrum, as the number of slits is very large.

7.8.2 Missing of Orders

Under certain conditions, it is possible that grating forms the first and third order spectra while the second order spectrum is missing. Such a situation arises when for a given angle of diffraction θ_1 the path difference between the diffracted rays from the extreme ends of one slit is equal to an integral multiple of λ . Suppose the path difference is λ . Then each slit can be considered to be made up of two halves, and the path difference between the secondary waves from the corresponding points in the two halves will be $\lambda/2$. This results in destructive interference.

Mathematically, we express this as

$$(a+b) \sin \theta = n\lambda \quad (7.27)$$

For single slit diffraction, the condition for minima is

$$a \sin \theta = m\lambda \quad (7.28)$$

If both the above conditions are satisfied simultaneously, the principal maxima will not be present in that direction. Dividing equ. (7.26) by (7.27), we obtain

$$\frac{a+b}{a} = \frac{n}{m}$$

i.e., $n = \frac{a+b}{a} m$ (7.29)

The above is the condition for the n^{th} order spectrum to be absent.

If we wish to suppress the second order spectrum, then $n = 2m = 2 (\because m = 1)$.

Then, $\frac{a+b}{a} = \frac{2m}{m} = 2$

or $a+b = 2a$
 $a = b$

Thus, if the width of each slit a is equal to the width b of the ruling, the second order spectrum will be absent.

7.8.3 Maximum Number of Orders Possible

The grating equation $d \sin \theta = \pm n\lambda$ may be rewritten as

$$n = \frac{d \sin \theta}{\lambda} = \frac{\sin \theta}{N\lambda}$$

The maximum value that θ can take is 90° and hence the maximum possible value of $\sin \theta$ is 1. It implies that

$$(n)_{\max} \leq \frac{1}{N\lambda} \quad (7.30)$$

The above relation (7.30) gives the maximum number of orders that would be seen in the spectrum produced by a plane transmission grating having N lines per unit width.

7.8.4 Determination of Wavelength of Light with a Plane Transmission Grating

The wavelength of spectral lines can be determined using a diffraction grating and a spectrometer. The slit of the collimator in the spectrometer is illuminated by an appropriate source of light (say light from sodium lamp) and the initial adjustments of the spectrometer are made

- (i) The collimator and the telescope are adjusted for parallel rays. This is done by Schuster's method using a prism.
- (ii) Then the grating is adjusted for *normal incidence*. To do this, the telescope is set in line with the collimator such that direct image of the slit is obtained at the position of the vertical cross-wire in the field of view of the telescope. Now the axes of the collimator and the telescope are in the same line. The position of the telescope is noted from the circular scale. Then the telescope is turned through 90° from this position and clamped. In this position the axis of the telescope is perpendicular to the axis of collimator (step 2 in Fig. 7.17). The given transmission grating is mounted at the centre of the prism table such that the grating surface is obtained in the centre of field of view of the telescope. This means that the parallel rays of light from the collimator are incident at an angle 45° on the grating surface because the axis of the collimator and the telescope are perpendicular to each other (step 3). The prism table is now rotated through 45° in the proper direction so that the grating surface is normal to the incident light faces the telescope (step 4). The prism table is clamped in this position. The above steps are shown in Fig. 7.17.

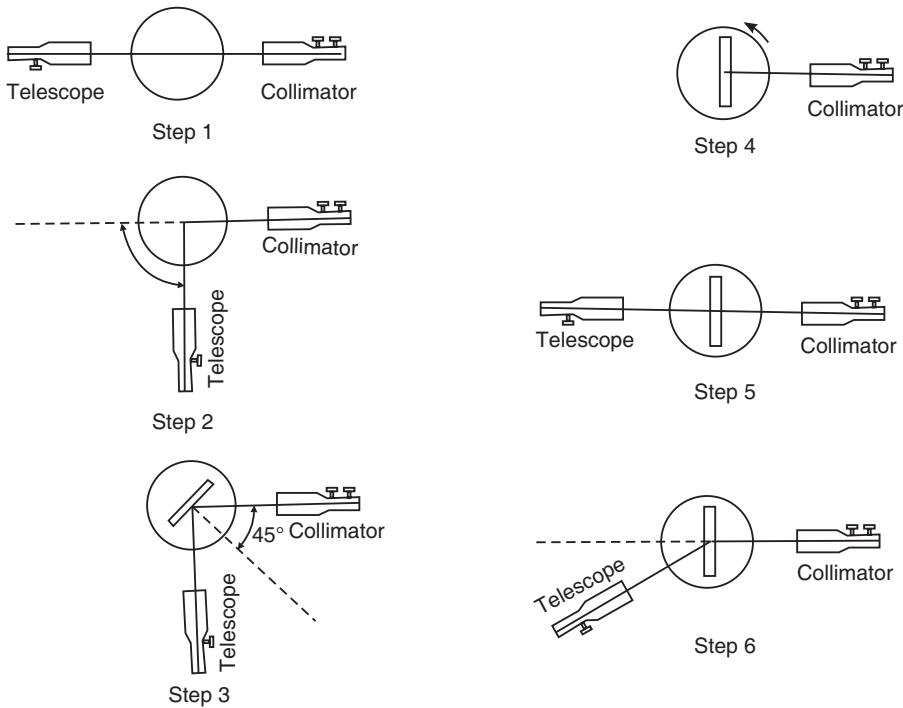


Fig. 7.17

- (iii) On viewing through the telescope (step 5), the grating spectrum is observed (Fig. 7.18). It consists of a direct image flanked by the different order images on its both sides. To measure θ for a spectral line of a given order, the telescope is focused on, say, first order image of that line (step 6). The position of the telescope is adjusted such that the line falls on the intersection of the cross-wires. Then the readings of both the verniers are noted. Then the telescope is taken to the other side of the direct image and the corresponding readings of first order image are noted. The difference between the readings of the two positions of the line gives 2θ for that line. From this the diffraction angle θ is found. The procedure is repeated for higher orders.

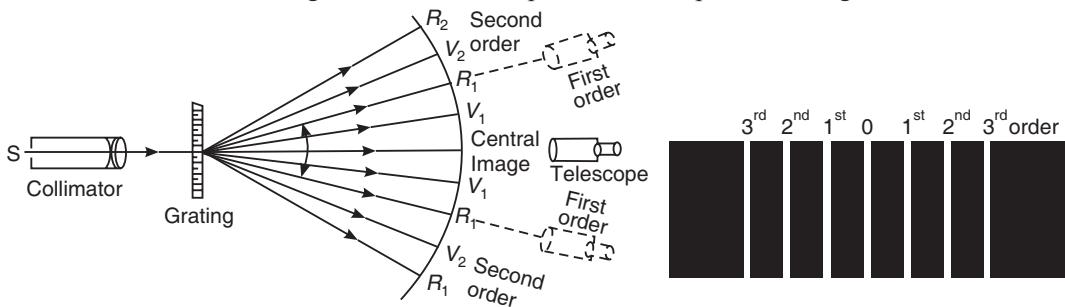


Fig. 7.18

- (iv) Knowing the value of d (or N) and n , the wavelength λ can be calculated from the grating equation $d \sin \theta = n\lambda$.

If the source of light emits radiations of different wavelengths, then the grating disperses the beam and in each order a spectrum of the constituent wavelengths is observed. To find the wavelength of any spectral line, the diffracting angles are noted in the first and second orders and using the grating equation, the wavelength of the spectral lines can be calculated.

With a diffraction grating, the wavelength of the spectral line can be determined very accurately. The method involves only the accurate measurement of the angles of diffraction.

As the angles are large they can be measured accurately with a properly calibrated spectrometer. Further, as the method does not involve measurements of very small distances (as in the case of interference experiments) an accurate value of λ can be obtained.

7.8.5 Dispersive Power of Grating

Dispersive power of a grating is defined as the ratio of the difference in the angle of diffraction of any two neighbouring spectral lines to the difference in wavelength between the two spectral lines. It can also be defined as the difference in the angle of diffraction per unit change in wavelength. The diffraction of the n^{th} order principal maximum for a wavelength, λ , is given by the equation

$$(a + b) \sin \theta = n \lambda$$

Differentiating this equation with respect to θ and λ , we get

$$(a + b) \cos \theta d\theta = n d\lambda$$

or
$$\frac{d\theta}{d\lambda} = \frac{nN'}{\cos \theta}$$

(7.31)

From equ.(7.31) it is clear that the dispersive power of the grating $d\theta/d\lambda$ is (i) directly proportional to the order of the spectrum, n (ii) directly proportional to the number of lines per cm, N' and (iii) inversely proportional to $\cos \theta$. Thus, the angular spacing of any two spectral lines is double in the second order spectrum than that in the first order. Secondly, the angular dispersion of the lines is more with a grating having a larger number of lines per cm. Thirdly, the angular dispersion is a minimum when $\theta = 0$. If the value of θ is not large, the value of $\cos \theta$ can be taken as unity and the influence of this factor can be neglected. Then it is clear that the angular dispersion of any two spectral lines is directly proportional to the difference in wavelength of the spectral lines. A spectrum of this type is called a **normal spectrum**.

If the linear spacing of two spectral lines of wavelengths λ and $\lambda + d\lambda$ is dx in the focal plane of the telescope objective or photographic plate, then

$$dx = f d\theta$$

where f is the focal length of the objective. The linear dispersion is

$$\frac{dx}{d\lambda} = f \frac{d\theta}{d\lambda} = \frac{fnN'}{\cos \theta}$$
(7.32)

or
$$dx = \frac{fnN'}{\cos \theta} \cdot d\lambda$$

The linear dispersion is useful in studying the photographs of a spectrum.

Example 7.4: A plane diffraction grating has the value of grating constant equal to 15×10^{-4} cm. Calculate the position of the third order maximum for $\lambda = 2.4 \times 10^{-4}$. (R.G.P.V.-2007)

Solution:
$$\sin \theta_3 = \frac{n\lambda}{d} = \frac{3 \times 2.4 \times 10^{-4} \text{ cm}}{15 \times 10^{-4} \text{ cm}} = 0.427$$

Hence, $\theta_3 = \sin^{-1}(0.427) = 25.28^\circ$.

Example 7.5: A grating has 15 cm of the surface ruled with 6000 lines/cm. What is the resolving power of the grating in the first order?

Solution: Total rulings on grating surface, $N = 15 \text{ cm} \times 6000 \text{ lines/cm} = 9 \times 10^4$

Resolving power, $R = m N = 1 \times 9 \times 10^4 = 9 \times 10^4$.

Example 7.6: In a grating spectrum, which spectral line in 4th order will overlap with 3rd order line of 5491Å?

Solution: The grating equation is

$$(a + b) \sin \theta = n\lambda$$

If the n^{th} order of wavelength λ_1 (say) coincides with the $(n + 1)^{\text{th}}$ order of λ_2 , then

$$(a + b) \sin \theta = n\lambda_1 = (n + 1)\lambda_2.$$

Here $n = 3$, $\lambda_1 = 5491\text{\AA}$, $(n + 1) = 4$, $\lambda_2 = ?$

$$\therefore \lambda_2 = \frac{n\lambda_1}{n+1} = \frac{3 \times 5491 \times 10^{-8} \text{ cm}}{4} = 4096 \times 10^{-8} \text{ cm} = 4096\text{\AA}.$$

7.9 RESOLVING POWER

When two objects or their images are very close to each other, they appear as one and it may not be possible for the eye to see them as separate. If the objects are not seen separately, then we say that the details are *not resolved* by the eye. Optical instruments are used to assist the eye in resolving the objects or images. The method adapted to seeing the close objects as separate objects is called **resolution**. The ability of an optical instrument to produce distinctly separate images of two objects located very close to each other is called its **resolving power**. It is defined as the reciprocal of the smallest angle subtended at the objective by two point objects, which can just be distinguished as separate.

7.9.1 Rayleigh's Criterion

The theory of optical instruments is based on the laws of geometrical optics and rectilinear propagation of light. These laws are only approximately true. When a beam of light from a point object passes through the objective of a telescope, the lens acts like a circular aperture and produces a diffraction pattern instead of a point image. This diffraction pattern is a bright disc surrounded by alternate dark and bright rings (see Fig. 7.19). It is known as **Airy's disc**. If there are two point objects lying close to each other, then two diffraction patterns are produced, which may overlap on each other and it may be difficult to distinguish them as separate (see Fig. 7.20).

To obtain the measure of the resolving power of an objective lens Rayleigh suggested that the two images of such point-objects lying close to each other may be regarded as separated if the central maximum of one falls on the first minimum of the other (Fig. 7.21 b). In other words, when the central bright image of one falls on the first dark ring of the other, the two images are said to be resolved (see Fig. 7.20 c). This is equivalent to the condition that the distance between the centers of the patterns shall be equal to the radius of the central disc. This is called the **Rayleigh criterion** for resolution and is also known as **Rayleigh's limit of resolution**. (See Fig. 7.21).

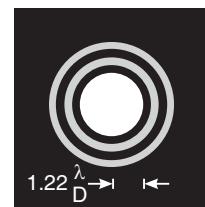


Fig. 7.19: Airy Disc

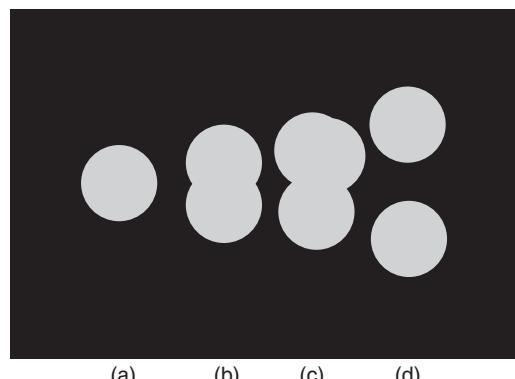


Fig. 7.20

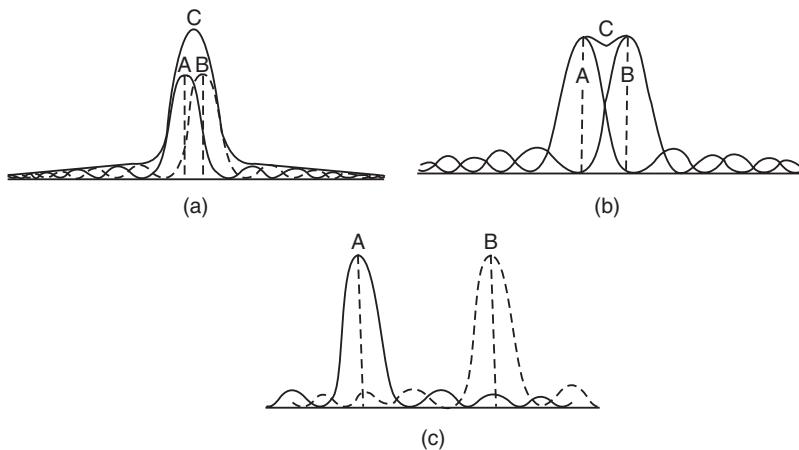


Fig. 7.21: The images of two sources are discernable when they satisfy Rayleigh criterion as seen in the third image at the top

7.10 RESOLVING POWER OF A PLANE TRANSMISSION GRATING

One of the important properties of a diffraction grating is its ability to resolve spectral lines, which have nearly the same wavelength. The spectral resolving power of a grating is defined in terms of the smallest wavelength interval ($d\lambda$) that can be detected by it. It is given by $\lambda/d\lambda$ where λ is the average of the two wavelengths and $d\lambda$ is their difference.

$$\text{Resolving Power} = \frac{\lambda}{d\lambda} = \frac{\lambda}{d\theta} \cdot \frac{d\theta}{d\lambda} \quad (7.33)$$

Let us now find the values of $\frac{d\theta}{d\lambda}$ and $\frac{\lambda}{d\theta}$.

The diffraction grating equation is

$$(a + b) \sin \theta = m\lambda.$$

Differentiating the above equation both sides, we get

$$(a + b) \cos \theta \cdot d\theta = m d\lambda$$

$$\therefore \frac{d\theta}{d\lambda} = \frac{m}{(a + b) \cos \theta} \quad (7.34)$$

where $d\theta$ is the angle between the two diffracted beams whose difference in wavelength is $d\lambda$.

The light diffracted from the grating enters the objective of a telescope in a grating spectrometer. If the diffracted beam completely fills the objective then width of the beam equals the diameter d of the objective lens. The angular limit of resolution of a telescope objective is given by

$$d\theta = \frac{\lambda}{d}$$

Now $d = AB \cos \theta = l \cos \theta$ where l is the length of the grating.

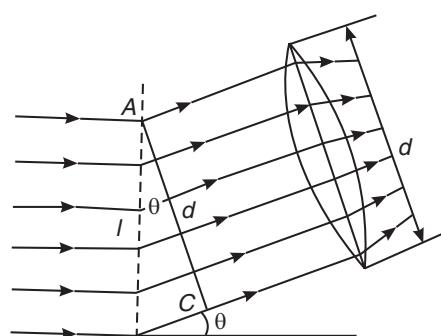


Fig. 7.22

$$\therefore d\theta = \frac{\lambda}{l \cos \theta}$$

or

$$\frac{\lambda}{d\theta} = l \cos \theta$$

$$\begin{aligned} \text{Resolving Power, } R &= \frac{\lambda}{d\lambda} = \frac{\lambda}{d\theta} \cdot \frac{d\theta}{d\lambda} \\ &= l \cos \theta \times \frac{m}{(a+b) \cos \theta} \\ &= \frac{ml}{(a+b)} \\ &= m N \end{aligned} \quad (7.35)$$

where $N = l/(a+b)$ = number of rulings on the grating surface and m is the order of the spectrum. Hence, the resolving power of a grating is given by the simple expression,

$$\text{R.P.} = mN$$

Example 7.7: The sodium yellow doublet has wavelengths 5890\AA and 5896\AA . What should be the resolving power of a grating to resolve these lines?

Solution: Mean wavelength $\lambda = (5890 + 5896)/2 = 5893\text{\AA}$.

Wavelength difference, $d\lambda = 5896 \text{\AA} - 5890 \text{\AA} = 6\text{\AA}$

Therefore, the resolving power R required for a grating to resolve these lines is given by

$$R = \frac{\lambda}{d\lambda} = \frac{5893}{6} = 982.$$

QUESTIONS

1. Explain clearly what is diffraction of light.
2. Why do radio waves diffract around buildings, although light waves do not? **(Calicut Univ.,2006)**
3. Give the difference between interference and diffraction. **(Amaravati Univ.,2003), (Calicut Univ.,2005, 2007)**
4. Distinguish between Fresnel and Fraunhofer class of diffractions. **(R.G.P.V.-2007)**
5. What is the source of the colours seen on the surface of CDs (compact discs)?
6. Deduce the conditions for maxima and minima for diffraction at a single slit.
7. Describe and explain the nature of fringes obtained with the help of a single slit placed before a parallel beam of monochromatic light. **(Amaravati Univ.,2008)**
8. A narrow slit, illuminated by monochromatic light produces Fraunhofer diffraction. Graphically show the intensity distribution in the diffraction pattern. Write the expression for intensity distribution.
9. What is diffraction and explain Fresnel's and Fraunhofer diffraction? **(Amaravati Univ.,2008)**
10. Describe and explain the phenomenon of diffraction due to a straight edge. Determine the position of maximum and minimum intensity. **(Calicut Univ.,2005)**
11. Discuss the Fraunhofer diffraction at a single slit. Obtain the condition for principal maximum and minimum. **(Univ. of Pune, 2007), (C.S.V.T.U.,2005)**
12. Describe and explain the nature of fringes obtained with the help of a single slit placed before a parallel beam of monochromatic light. **(RGPV,2007), (C.S.V.T.U.,2006)**

13. Describe and explain the Fraunhofer diffraction pattern obtained by narrow slit and illuminated by a parallel beam of monochromatic light. **(Calicut Univ.,2006)**
14. Discuss the Fraunhofer diffraction pattern shown by double slit.
15. Explain diffraction at a circular aperture. **(Amaravati Univ.,2003)**
16. What is diffraction grating? How is it obtained? **(Univ. of Pune, 2007)**
17. Discuss the Fraunhofer diffraction at a single slit. Extend this theory to the case of a plane transmission grating. **(C.S.V.T.U.,2007)**
18. What is plane transmission grating? Define grating element. Explain how it can be prepared? **(Amaravati Univ.,2005, 2008)**
19. What is plane diffraction grating? How is it used to find the wavelength of light?**(R.G.P.V.-2008)**
20. (a) What is a plane transmission grating?
(b) Obtain the expression $(a + b) \sin \theta = n\lambda$.
21. (c) What is dispersive power of a grating? Obtain an equation for dispersive power. **(Calicut Univ., 2007)**
22. Explain the theory of plane transmission grating and derive equation for maxima and minima. **(Amaravati Univ., 2007)**
23. Explain the theory of plane transmission grating with equation for maximum intensity. **(Amaravati Univ., 2007)**
24. Is it possible that no minimum is recorded in a single slit diffraction pattern? If so, under what condition?
25. Give the theory of plane diffraction grating. Obtain the condition for the formation of n^{th} order maximum. **(Univ. of Pune, 2008)**
26. State the factors on which the resolving power of grating depends. **(Univ. of Pune, 2008)**
27. (a) Explain the construction and working of a diffraction grating.
(b) Describe Rayleigh's criteria. **(Calicut Univ.,2007)**
28. What is plane diffraction grating? Explain how it is used to determine the wavelength of a spectral line of a given source of light. **(Amaravati Univ.,2006)**
29. How do you determine the wavelength of light using diffraction grating? Explain. **(Calicut Univ.,2006)**
30. How will you determine the wavelength of spectral line found in diffraction pattern by using plane transmission grating? **(Amaravati Univ.,2008)**
31. Two plane gratings A and B have the same width of ruled surface but A has greater number of lines than B. Compare intensity and width of principal maxima.
32. Why does a diffraction grating have closely spaced rulings?
33. Why does it have a large number of rulings?
34. Why are the colours in the spectrum of a light source linear in shape?
35. What are the advantages of increasing the number of rulings in a grating?
36. In a transmission grating, how spectral lines get affected if the rulings are made closer?
37. If width of transparencies and opacities of grating are equal which spectra would be absent in a transmission grating.
38. Explain the meaning of 'missing spectra' in the diffraction pattern of a plane transmission grating.
39. What is the condition, which must be satisfied if the second order spectra are to be absent from the grating spectrum?
40. In a grating what is the effect of changing (a) total number of lines, (b) the number of lines per cm and (c) the width of the grating?

41. Explain why increasing the number of slits in a grating sharpens the maxima?
42. Explain why decreasing the wavelength sharpens the maxima in a grating?
43. Explain why increasing slit spacing in a grating sharpens the maxima?
44. Derive an expression for the resolving power of a grating. **(C.S.V.T.U.,2006)**
45. Explain what you understand by the resolving power of an optical instrument? Determine the resolving power and dispersive power of the diffraction grating. **(Calicut Univ.,2005)**
46. Define resolving power of plane diffraction grating. Hence prove that it is independent of grating element. **(Bombay Univ.)**
47. Obtain an expression for the resolving power of grating. **(Univ. of Pune, 2008)**
48. Explain the resolving power of the plane diffraction grating. **(Calicut Univ.,2006)**
49. Define resolving power of an optical instrument. Find the expression for resolving power of telescope and discuss in detail. **(R.G.P.V.-2008)**
50. Discuss Rayleigh criterion of resolution. **(RGPV,2007)**
51. State and explain Rayleigh's criterion for limit of resolution. **(Calicut Univ.,2007)**
52. State Rayleigh criteria for the resolution of spectral lines. Distinguish between the resolving power and dispersive power of the diffraction grating. **(Calicut Univ.,2006)**
53. State Rayleigh criterion of resolution. Hence obtain an expression for the resolving power of a telescope. **(Univ. of Pune, 2007)**

PROBLEMS

- Light of wavelength 5500 Å falls normally on a slit of width 22×10^{-5} cm. Calculate the angular position of the first two minima on either side of the central maximum. **[Ans: $14^\circ 29'$, 30°]**
- Plane waves of wavelength 6000 Å falls normally on a straight slit of width 0.20mm. Calculate the total angular width of the central maximum and also the linear width as observed on a screen placed 2m away. **[Ans: 6×10^{-3} rad, 12 mm]**
- A screen is placed 200 cm away from a narrow slit which is illuminated with light of wavelength 6000 Å. If the first minima lie 5mm on either side of central maximum, calculate the slit width. **[Ans: 0.24 mm]**
- Calculate the possible order of spectra with a plane transmission grating having 18,000 lines per inch when light of wavelength 4500 Å is used. **[Ans: 3]**
- Light of wavelength 5000 Å is incident normally on a single slit of width 1 mm. Calculate the normalized intensity for an angle of diffraction of 30° . **[Ans: 0]**
- A transmission grating has 8000 rulings per cm. The first order principal maximum due to a monochromatic source of light occurs at an angle of 30° . Determine the wavelength of light.
- A monochromatic light of wavelength 6000 Å is incident on a plane diffraction grating with grating element 6.0×10^{-5} cm. What is the maximum order of spectrum that can be observed?
- The second order maximum for a wavelength of 6360 Å in a transmission grating coincides with third order maximum of an unknown light. Determine the wavelength of the unknown light.
- A transmission diffraction grating has 5000 lines/cm. If the slits are 10^{-4} cm wide, will there be any missing orders? If so, identify these.
- A plane transmission grating has 40,000 lines in all with grating element 12.5×10^{-5} cm. Calculate the maximum resolving power for which it can be used in the range of wavelength 5000 Å. **[Ans: 80,000]**
- A plane transmission grating has 16,000 lines per inch over a length of 5 inches. Find (a) the resolving power of the grating in the second order and (b) the smallest wavelength difference that can be resolved for light of wavelength 6000 Å. **[Ans: 1,60,000; 0.0375 Å]**

CHAPTER

8

Polarization

8.1 INTRODUCTION

Interference and diffraction phenomena proved that light is a wave motion and enabled the determination of the wavelength. However, they do not give any indication regarding the character of the waves. Whether the light waves are longitudinal or transverse, or whether the vibrations are linear or circular cannot be deduced from the above two phenomena, as all kinds of waves under suitable conditions exhibit interference and diffraction. In 1816 Arago and Fresnel showed that light waves vibrating in mutually perpendicular planes do not interfere. In 1817 Thomas Young postulated that light waves are *transverse waves* and explained the absence of interference between light waves polarized in mutually perpendicular planes. Thus, the existence of polarization property is a direct consequence of light being a transverse wave. Light coming from common light sources is unpolarized. It can be transformed into different types of polarization using optical devices. The state of polarization cannot be perceived by an unaided human eye. An understanding of polarization is essential for understanding the propagation of electromagnetic waves guided through wave-guides and optical fibres. Polarized light has many important applications in industry and engineering. One of the most important applications is in liquid crystal displays (LCDs), which are widely used in wristwatches, calculators, TV screens etc.

8.2 POLARIZATION

Waves are basically of two types: (i) longitudinal waves and (ii) transverse waves.

- (i) A wave in which particles of the medium oscillate to and fro along the direction of propagation is called a **longitudinal wave**. Waves produced on a spring and sound waves are examples of longitudinal waves. The longitudinal wave consists of alternate compressions and rarefactions, as shown in Fig. 8.1 (a & b).
- (ii) A wave in which every particle of the medium oscillates up and down at right angles to the direction of wave propagation is called a **transverse wave**. Ripples on water surface and waves on a rope are examples of transverse waves. The wave propagates in the form of alternating crests and troughs, as shown in Fig. 8.1 (c & d).

In a longitudinal wave, all directions perpendicular to the wave propagation are equivalent. On the other hand, a preferential direction normal to the wave propagation exists in a transverse wave. The preferential direction in a transverse wave is the direction of vibration of the particles and it differs from all other directions. The existence of a preferential

direction for a transverse wave leads to the characteristic phenomenon known as *polarization*. Polarization is not found with longitudinal waves as they do not possess a directional property. Thus, polarization is specific to transverse waves.

Light waves are transverse waves consisting of electric and magnetic fields vibrating perpendicular to each other and to the direction of propagation. The vibrating electric field vector and the direction of propagation of the wave constitute a plane. There is an infinite number of such planes around the direction of propagation. In an ideal light wave, the vibrations of electric vector are confined to a single plane. In practice, light sources emit a mixture of light waves whose planes of vibration are randomly oriented about the direction of propagation. Such random orientation of vibration planes gives rise to symmetry about the wave propagation direction. As a result, the transverse nature of the wave gets concealed. The process of removing the symmetry and bringing in one-sidedness in the light wave is called **polarization**.

8.3 UNPOLARIZED AND POLARIZED LIGHT

Light wave is a transverse electromagnetic wave made up of mutually perpendicular, fluctuating electric and magnetic fields. Fig. 8.2 (a) shows the electric field in the xy -plane, the magnetic field in the xz -plane and the propagation of the wave in the x -direction. The right-hand part of the diagram shows the variation of the electric field in space as the wave propagates. Traditionally, light wave is described by the **electric field vector**, \mathbf{E} , and accordingly, only the electric field vector is shown in Fig. 8.2 (b).

As the electric field is a vector, it points in a particular direction in space. The **polarization** of an electromagnetic wave refers to the orientation of its electric field vector \mathbf{E} . If we could view a light wave coming from an ordinary source towards us, then we would observe that the direction of \mathbf{E} is randomly varying with time on a very fast scale.

The light from an incandescent bulb, for example, emits a mixture of light waves with electric field components that change randomly on a scale of 10^{-14} s, almost as fast as the optical frequency itself. As a result, the **direction of oscillation of the electric field vector in an ordinary light beam occurs in all the possible planes perpendicular to the beam direction**, as illustrated in Fig. 8.3. A light wave, in which \mathbf{E} -vector oscillates

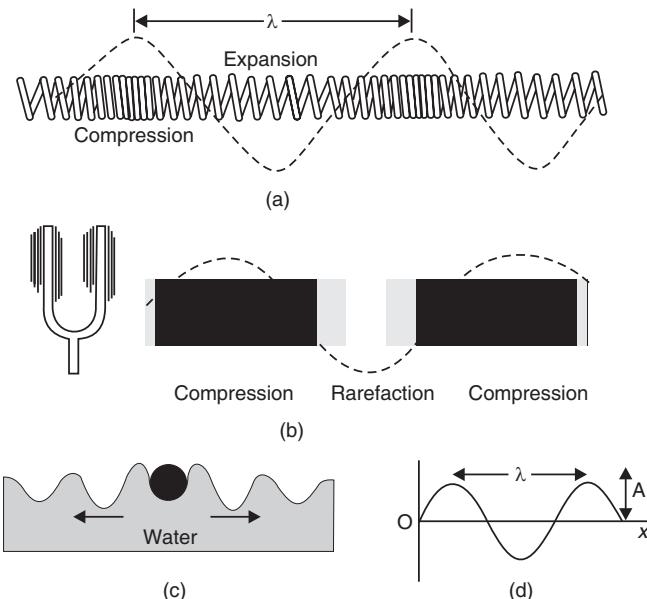


Fig. 8.1

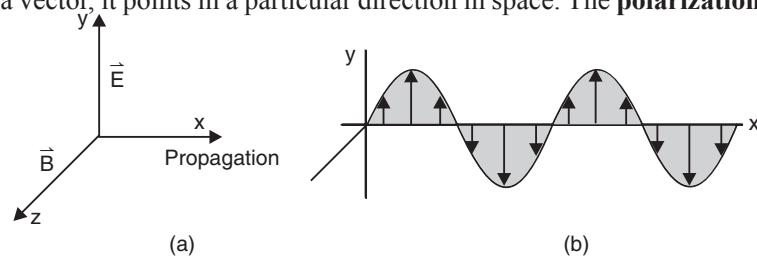


Fig. 8.2: A light wave is described by the electric field vector, \mathbf{E}

in more than one plane, is referred to as **unpolarized light**. Light emitted by the sun, by an incandescent lamp, or by a candle flame is unpolarized light.

Polarized Light

Polarized light is not produced naturally. It is obtained by converting natural light into polarized light using optical elements. The process of transforming unpolarized light into polarized light is polarization. A polarized light wave is a light wave with a definite direction of oscillation of the **E**-vector, which occurs in a *single plane* or in *some specific way*. For example, the wave in Fig. 8.2 (b) is a polarized wave. **Polarized light** is the light that contains waves that only fluctuate in one specific plane.

We designate the plane created by the direction of oscillation of the electric field vector **E** and the direction of propagation of the beam as the **plane of polarization of light wave**. Thus, the *xy*-plane is the plane of polarization in Fig. 8.2 (b).

8.4 NATURAL LIGHT IS UNPOLARIZED LIGHT

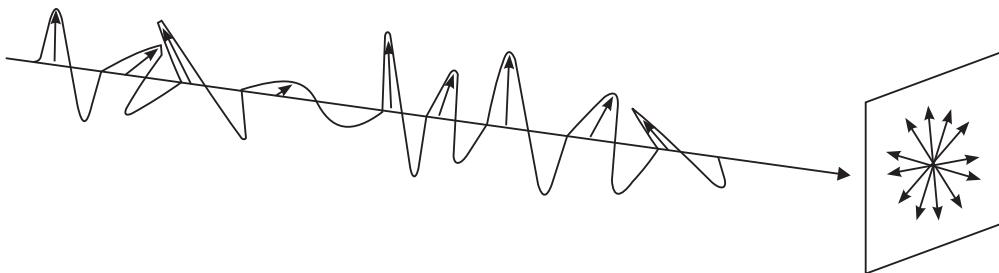


Fig. 8.3: Natural light is unpolarized

We mentioned earlier that light from the ordinary light sources is unpolarized. Let us find out the reason. We know that atoms emit light. Any light source consists of a very large number of atomic emitters. Each atom radiates, at a specific instant, a **wave packet** (also known as a **wave train**) that lasts for about 10^{-8} s. Light radiated by a source is a mixture of wave packets emitted by different atoms at different instances (see Fig. 8.3). Individual wave packets will be polarized, but each wave packet has its own polarization, and they are not correlated in any way. There is no continuity of plane of polarization and the plane of polarization varies from wave packet to wave packet in a completely random manner. The polarization of the waves determined at any particular spot will fluctuate randomly, and very rapidly, with no preferred direction. Any direction is equally likely, and the usual graphical representation of ordinary light is shown in the right-hand figure in Fig. 8.4 (a). Fig. 8.4 (a) is a fictitious diagram. It only implies that the natural light consists of electric field vectors of many possible orientations lying at different angles between 0 and 360° and hence symmetrically distributed about the direction of propagation.

The concept of unpolarized light is rather difficult to visualize. In general, it is looked upon as consisting of an average of half of its vibrations horizontally polarized and half of its vibrations vertically polarized. Due to the random distribution of the optical vectors, the

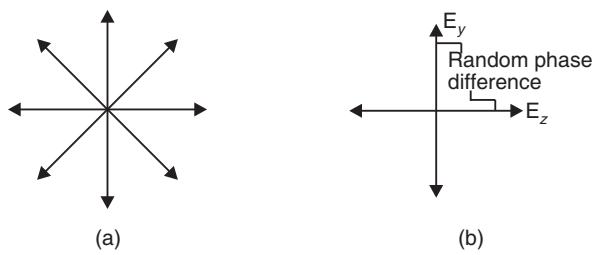


Fig. 8.4: (a) Pictorial representation of Natural light.
(b) Unpolarized light is viewed as a combination of **incoherent** vertically and horizontally polarized waves.

amplitude of the vertical and horizontal component vectors are taken as equal; however, the **two components are incoherent**, i.e. have a randomly changing phase difference. In view of this, unpolarized is pictorially represented by two electric vectors as shown in Fig. 8.4 (b).

A comparison of Unpolarized and Polarized light

	<i>Unpolarized light</i>	<i>Polarized light</i>
1.	Consists of waves with planes of vibration equally distributed in all directions about the ray direction.	Consists of waves having their electric vector vibrating in a single plane normal to ray direction.
2.	Symmetrical about the ray direction	Asymmetrical about the ray direction
3.	Produced by conventional light sources.	Is to be obtained from unpolarized light with the help of polarizers.
4.	May be regarded as the resultant of two <i>incoherent</i> waves of equal intensity but polarized in mutually perpendicular planes.	May be regarded as the resultant of two mutually perpendicular <i>coherent</i> waves having zero phase difference.

8.5 TYPES OF POLARIZATION

The polarization of a light wave describes *the shape and locus of the tip of the E vector* (in the plane perpendicular to the direction of propagation) *at a given point in space as a function of time*. Depending upon the locus of the tip of the E vector, light may exhibit three different states of polarization. They are

- (i) plane or linear polarization,
- (ii) elliptical polarization and
- (iii) circular polarization.

Apart from these, the light may also be partially polarized.

An unaided human eye cannot identify the state of polarization of light. However, some insects and animals possess polarization sensitive vision.

8.5.1 Plane Polarized Light

Plane polarized light waves are light waves in which the oscillations occur in a single plane. In a plane-polarized wave, the oscillations of electric field vector \mathbf{E} are strictly confined to a single plane *perpendicular* to the direction of propagation. As the direction of the field vector at some point in space and time lies along a line in a plane perpendicular to the direction of wave propagation, a plane-polarized wave is also known as a **linearly polarized** wave.

With linear polarization, the orientation of the E-vector stays constant at a point in space. That is, the direction of \mathbf{E} does not vary with time, but its magnitude varies sinusoidally with time. If the field is pointing either up or down, we call it **vertical polarization**, and if it is pointing either right or left, we call it **horizontal polarization**. Electric fields are not restricted to pointing exactly along vertical or horizontal axes, but can be at any arbitrary angle to those axes. Linearly polarized light, polarized at any arbitrary angle, may be regarded as a combination of horizontally and vertically polarized

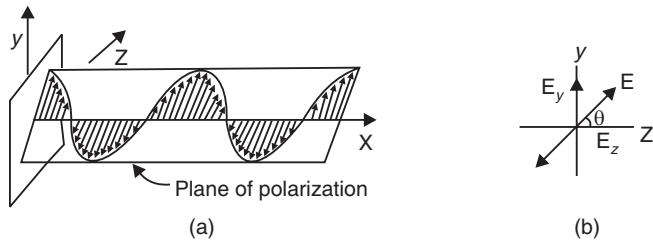
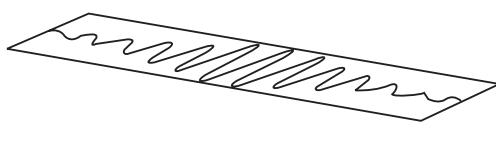


Fig. 8.5: A light wave polarized in an arbitrary direction

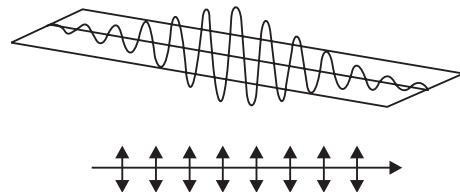
light, with appropriate amplitude, and which are oscillating **in phase** or 180° **out of phase**. The key point is that the two component waves are **coherent**.

Let x be the direction of travel of the light and, y and z be directions in the plane of the electric field. The electric field makes a constant angle θ to the z -direction, as shown in the Fig. 8.5 (a). The wave in Fig. 8.5 (a) is the resultant wave due to superposition of two **coherent** linearly polarized waves, oscillating in phase, as shown in Fig. 8.6.

Representation of linearly polarized light in diagrams



(a) Horizontal Linear Polarization



(b) Vertical Linear Polarization

Fig. 8.7: Representation of linearly polarized light in diagrams

Linearly polarized light is represented in diagrams as shown in Fig. 8.7. When the electric field vector oscillates horizontally in a direction perpendicular to the plane of the paper, the light wave is represented by dots (Fig. 8.7a). When the electric field vector oscillates vertically in the plane of the paper, the light wave is represented by arrows, as shown in Fig. 8.7(b).

8.5.2 Circularly Polarized Light

A light wave is said to be **circularly polarized**, if in the course of wave propagation, the magnitude of the electric vector \mathbf{E} stays constant but it rotates at a constant rate about the direction of propagation and sweeps a circular helix in space, as shown in Fig. 8.8. This is a picture in terms of the **space variation of**

\mathbf{E} . Alternately, if we could see the wave advancing towards our eyes, we would find that the tip of the \mathbf{E} vector tracing a circle in space, completing one revolution within one wavelength (see Fig. 8.9b). Hence, the state of polarization is called **circular polarization**. In, circularly polarized light, there is no preference to specific direction of oscillation.

A circularly polarized light wave may be regarded as the resultant wave produced due to superposition of two coherent linearly polarized waves of *equal amplitude* oscillating in mutually perpendicular planes, and are out of phase by 90° .

Let us again consider two linearly polarized waves having equal amplitude, out of which one is polarized in z -direction (horizontally polarized wave) and the other in y -direction (vertically polarized wave). Let us further assume that they are **coherent** and out of phase by 90° (see Fig. 8.9a). At some instant, the \mathbf{E} -field vector of z -polarized wave will have

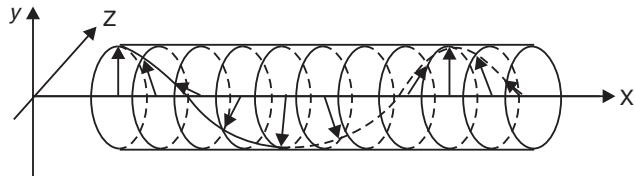


Fig. 8.8: Space variation of \mathbf{E} ; \mathbf{E} vector sweeps a circular helix in space

maximum amplitude and the E-field vector of y-polarized wave will be at zero. At that instant, the polarization is horizontal. A little time later, the z-polarized wave has decreased a little, while the y-polarized wave has started to increase. Then, the light looks like it is polarized at a slight angle-- mostly horizontal, but with a small vertical component. A little while later, the z-component has decreased some more, and the y-component has increased some more, and the angle is greater. And so on. Eventually, the y-polarized wave is at a maximum, and the z-polarized wave is zero, and we have pure vertical polarization. If we stand at one point in space, and look at the direction of the wave, we will observe that the E-vector sweeps a circle in space. Hence, it is called circularly polarized wave. Note that the *oscillations of the resultant E-vector do not take place in a single plane*.

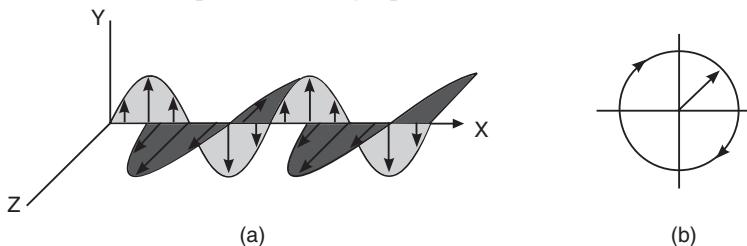


Fig. 8.9: Circularly polarized wave is a combination of horizontally polarized and vertically polarized waves that are out of phase by 90° and having equal amplitudes.

If the rotation of the tip of E is clockwise as seen by an observer looking back towards the source, then the wave is said to be **right-circularly polarized** (Fig. 8.10a). If the tip of E rotates anti-clockwise, then the wave is said to be **left-circularly polarized** (Fig. 8.10b). This is a description of circular polarization in terms of the **time variation** of E.

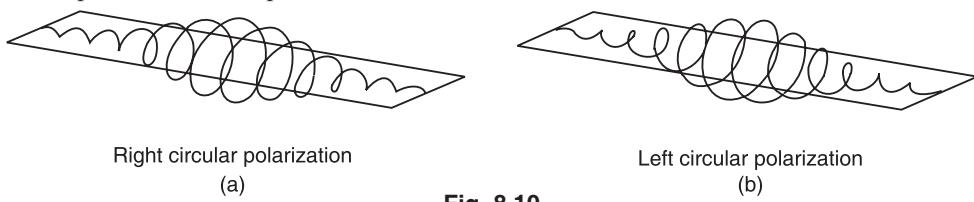


Fig. 8.10

8.5.3 Elliptically Polarized Light

A light wave is said to be **elliptically polarized**, if the magnitude of electric vector E changes with time and the vector E rotates about the direction of propagation and sweeps a flattened helix in space, as shown in Fig. 8.11. This is a description of elliptically polarized light in terms of the **space variation** of E. Alternately, if we imagine that we are looking at the light wave advancing towards us, we would observe that the tip of the E vector traces an ellipse in space. Hence, it is called elliptically polarized light. This is a description of elliptical polarization in terms of the **time variation** of E.

An elliptically polarized light wave may be regarded as the resultant wave produced due to superposition of two **coherent** linearly polarized waves of different amplitudes, oscillating in mutually perpendicular planes and are out of phase. If waves of differing amplitude are related in phase by 90° , or if the relative phase difference is other than 90° then the resultant light wave is elliptically polarized.

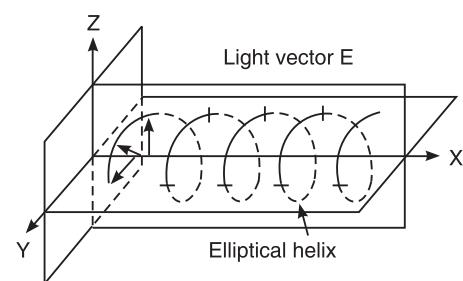


Fig. 8.11: Space variation of E; E vector sweeps a flattened helix in space

Let us consider two linearly polarized waves having different amplitudes, out of which one is polarized in y -direction and the other in z -direction. Let us further assume that they are **coherent** and out of phase by an arbitrary angle δ (see Fig. 8.12). The \mathbf{E} -field vector of y -polarized wave will have maximum amplitude at times when the \mathbf{E} -field vector of z -polarized wave is a minimum and vice versa (Fig. 8.12). Then the oscillations of the resultant \mathbf{E} -vector do not take place in a single plane. The magnitude of resultant \mathbf{E} -vector varies at each point in space and the overall rotation of the \mathbf{E} -vector has the appearance of a flattened helix.

When we are looking back towards the source, if the rotation of \mathbf{E} vector occurs clockwise, it is said to be a **right-elliptically-polarized** wave. If it rotates anti-clockwise, it is said to be a **left-elliptically polarized** wave.

8.5.4 Partially Polarized Light

Usually, light is neither totally polarized nor unpolarized but a mixture of the two types. It can be viewed as a mixture of plane polarized light and unpolarized light. Partially polarized light is represented as shown in Fig. 8.13.

Partially polarized light, like natural light, can be represented in the form of a superposition of two incoherent plane-polarized waves with mutually perpendicular planes of oscillations. In case of natural light the amplitude of these waves is the same and for partially polarized light, it is different.

Degree of Polarization: If we pass partially polarized light through a polarizer, and if we rotate the polarizer about the direction of the ray, the intensity of the transmitted light will change within the limits from I_{\max} to I_{\min} . The transition from one of these values to the other will occur upon rotation through an angle of 90° . We define the degree of polarization with the help of the following expression.

$$P = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \quad (8.1)$$

% polarisation

$$\% \text{ polarization} = \frac{(I_{\max} - I_{\min})}{(I_{\max} + I_{\min})} \times 100 \quad (8.2)$$

For plane polarized light $I_{\min} = 0$, and hence $P = 1$ and the % polarization is 100%. For natural light, $I_{\max} = I_{\min}$, and hence $P = 0$ and the % polarization is zero. If $I_{\max} = 2I_{\min}$, $P = 0.33$ and % polarization = $100/3 = 33\%$.

Note that the concept of the degree of polarization cannot be applied to elliptically and circularly polarized light.

8.6 PRODUCTION OF PLANE POLARIZED LIGHT

We now study the methods of producing plane-polarized light. Plane polarized light may be produced from unpolarized light using the following five optical phenomena:

(i) reflection, (ii) refraction, (iii) scattering, (iv) selective absorption (dichroism), and (v) double refraction. Out of these five, the phenomena of selective absorption and double refraction are helpful in practical production of plane polarized light.

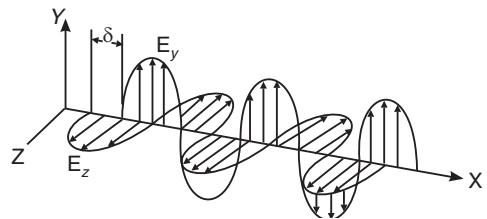


Fig. 8.12: Elliptically polarized wave is a combination of horizontally polarized and vertically polarized waves that are of different amplitudes and out of phase by θ

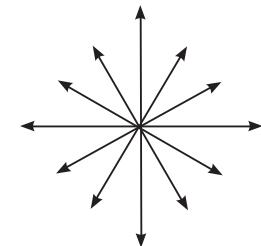


Fig. 8.13

8.6.1 Polarization by Reflection from Dielectric Surfaces

E.L. Malus, French engineer discovered in 1808 polarization of light by reflection. He noticed that when natural light is incident on a smooth surface, at a certain angle the reflected beam is plane polarized. The extent to which polarization occurs is dependent upon the angle at which the light is incident on the surface and upon the material, which the surface is made of. Metallic surfaces reflect light with a variety of vibrational directions; such reflected light is unpolarized. However, light that is specularly reflected from dielectric surfaces, such as asphalt roadways, water etc, is linearly polarized. If the extent of linear polarization is large, a person perceives **glare** from such surfaces. On bright sunny days, the glare caused by sunlight on a roadway or a field of snow, may be almost blinding to the human eye.

When light wave is incident on a boundary between two dielectric materials, part of it is reflected, and part of it is transmitted.

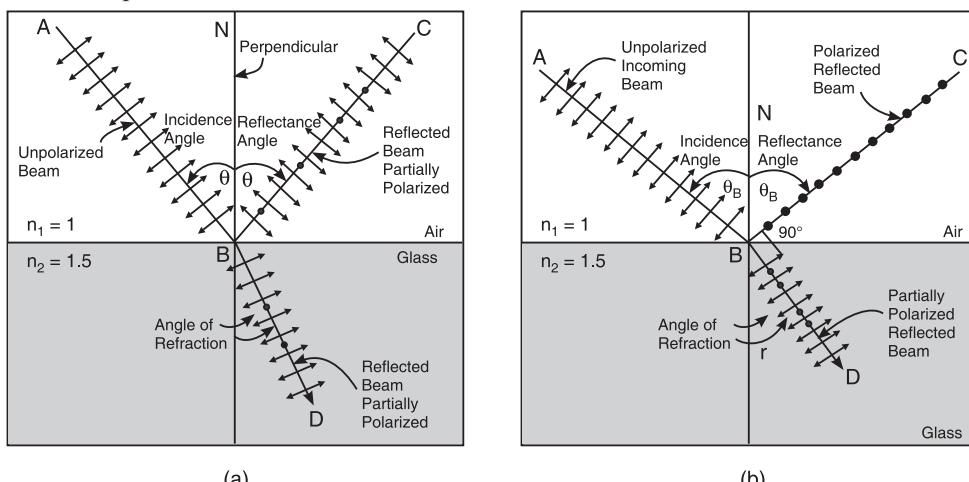


Fig. 8.14: (a) Reflection and Refraction at the surface between two media (b) Brewster Law

Fig. 8.14 (a) shows an unpolarized light beam AB incident on a glass surface. The incident ray AB and the normal NB define the plane of incidence. The electric vector E of the ray AB can be resolved into two components, one perpendicular to the plane of incidence and the other lying in the plane of incidence. The perpendicular component is represented by dots and is called the **s-component**. The parallel component is represented by the arrows and is called the **p-component**. In case of completely unpolarized light the two components are of equal magnitude. At a particular angle θ_B , the reflection coefficient for p-component goes to zero and the reflected beam does not contain any p-component (see Fig. 8.14 b). It contains only s-component and is totally *plane polarized*. The angle θ_B is called the **polarizing angle** or **Brewster's angle**.

This particular method of polarizing light is not advantageous, as the intensity of the reflected beam is very small. Only 15% of s-component is reflected. The refracted light is a mixture of 100% of p-component and the balance 85% s-component. Therefore, the refracted ray is strong but partially polarized.

8.6.1.1 Brewster's law

Sir David Brewster performed a series of experiments on the polarization of light by reflection at a number of surfaces. He found that the polarizing angle depends upon the refractive index of the medium. In 1892, Brewster proved that the **tangent of the angle at which polarization**

is obtained by reflection is numerically equal to the refractive index of the medium. If θ_B is the angle and μ is the refractive index of the medium, then

$$\mu = \tan \theta_B \quad (8.3)$$

This is known as **Brewster's law**.

If natural light is incident on a smooth surface at the polarizing angle, it is reflected along BC and refracted along BD, as shown in Fig. 8.14 (b). Brewster found that the maximum polarization of reflected ray occurs when it is at right angles to the refracted ray. It means that $\theta_B + r = 90^\circ$.

$$\therefore r = 90^\circ - \theta_B \quad (8.4)$$

According to Snell's law,

$$\frac{\sin \theta_B}{\sin r} = \frac{\mu_2}{\mu_1} \quad (8.5)$$

where μ_2 is the absolute refractive index of reflecting surface and μ_1 is the refractive index of the surrounding medium. It follows from equ.(8.4) and equ.(8.5) that

$$\frac{\sin \theta_B}{\sin (90^\circ - \theta_B)} = \frac{\mu_2}{\mu_1}$$

or

$$\frac{\sin \theta_B}{\cos \theta_B} = \frac{\mu_2}{\mu_1}$$

$$\therefore \tan \theta_B = \frac{\mu_2}{\mu_1} \quad (8.6)$$

Equ.(8.6) shows that the polarizing angle depends on the refractive index of the reflecting surface. The polarizing angle θ_B is known as **Brewster angle**. Light reflected from any angle other than Brewster angle is partially polarized.

Application of Brewster's law:

- (i) Brewster's law can be used to determine the refractive indices of opaque materials.
- (ii) It helps us in calculating the polarizing angle necessary for total polarization of reflected light for any material if its refractive index is known. However, the law is not applicable for metallic surfaces.

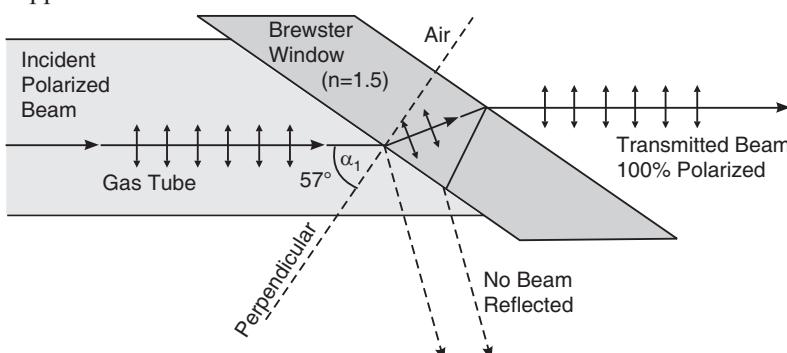


Fig. 8.15: Brewster angle window at the end of gas laser

- (iii) In gas lasers it is common to arrange two glass windows at the two ends of the laser tube. The windows are arranged at Brewster angle to the axis of the laser tube and hence they are called **Brewster windows**. The light beam traveling between the mirrors of the laser is reflected many times from these mirrors. Since the mirrors are at Brewster angle, all the light that is polarized perpendicular to the beam plane

is emitted out of the laser cavity at early stage. In the gas tube, there remains only radiation polarized in the beam plane. The advantage of this arrangement is that the beam has no reflection losses, since only the transmitted polarized beam is traveling between the mirrors. The radiation out of these lasers is polarized as can be seen in Fig. 8.15.

- (iv) Another application utilises the Brewster angle for transmitting a light beam into or out of an optical fibre without reflection losses.

Example 8.1: It is desired to use a plate of glass to obtain polarized light. If the refractive index of glass is 1.5, what is the polarizing angle?

Solution: Polarizing angle, $\theta_B = \tan^{-1} \mu = \tan^{-1}(1.5) = 56.31^\circ$.

Example 8.2: Sunlight is reflected from a calm lake. The reflected light is 100% polarized at a certain instant. What is the angle between the sun and the horizon at that instant? The refractive index of water is 1.33.

Solution: Since the reflected is 100% polarized, the angle of incidence is equal to the Brewster angle. By Brewster's law, $\mu = \tan \theta_B$. Therefore,

$$1.33 = \tan \theta_B$$

$$\therefore \theta_B = 53.06^\circ$$

The angle between the sun and the horizon $= 90^\circ - 53.06^\circ = 36.94^\circ = 36^\circ 54'$.

Example 8.3: The critical angle of incidence for total reflection in case of water is 48° . What is its polarization angle? What is the angle of refraction corresponding to the polarization angle?

Solution: The refractive index, $\mu = \frac{1}{\sin \theta_C} = \frac{1}{\sin 48^\circ} = 1.346$.

Now, $\mu = \tan \theta_B$ or $1.346 = \tan \theta_B$ $\therefore \theta_B = 53.4^\circ = 53^\circ 21'$

If r is the angle of refraction, $r = 90^\circ - 53^\circ 21' = 36^\circ 39'$

8.6.2 Polarization by Refraction - Pile of Plates

When unpolarized light is incident at Brewster angle on a smooth glass surface, the reflected light is totally polarized, while the refracted light is partially polarized. If natural light is transmitted through a single plate, the transmitted beam is only partially polarized. If a stack of glass plates is used instead of a single plate, reflections from successive surfaces occur leading to the filtering of the s-component in the transmitted ray. Ultimately, the transmitted ray consists of p-component alone. It is found that a stack of about 15 glass plates is required for this purpose. The glass plates are supported in a tube of suitable size and inclined at an angle of about 33° to the axis of the tube, as shown in Fig. 8.16. Such an arrangement is called a **pile of plates**. Unpolarized light enters the tube and is incident on the plates at Brewster angle and the transmitted light will be totally polarized parallel to the plane of incidence.

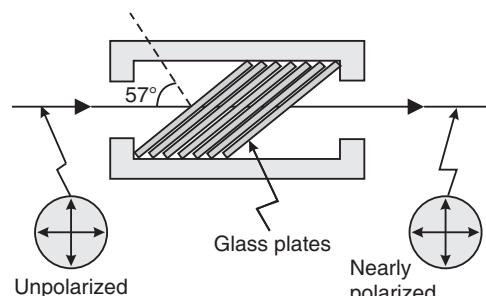


Fig. 8.16

8.6.3 Polarization by Scattering

If a narrow beam of natural light is incident on a transparent medium containing a suspension of ultramicroscopic particles, the light scattered is partially polarized. The incident light

causes electrons in the scattering medium to vibrate. A vibrating electron emits most light in a direction perpendicular to its vibration and none along the direction of its vibration. The electric field of the emitted radiation is parallel to the direction of electron vibration. Hence light scattered through about 90° with respect to the incident direction is strongly polarized. The direction of vibration of E vector in the scattered light will be perpendicular to the plane defined by the direction of propagation and the direction of observation, i.e., the plane of the paper, as illustrated in Fig. 8.17.

The light from a blue sky is quite strongly polarised, particularly at 90° from the sun. It is not completely polarised because a significant amount of sunlight has undergone multiple-scattering, i.e. has been scattered more than once. Light scattered twice through a total angle of 90° would be less polarised than light scattered once.

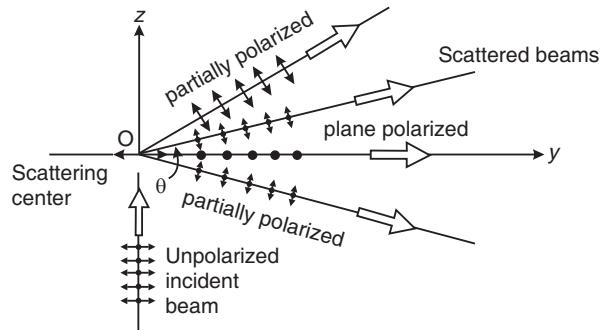


Fig. 8.17

8.6.4 Polarization by Selective Absorption

A number of crystalline materials absorb more light in one incident plane than another, so that light progressing through the material become more and more polarized as they proceed. This difference in the absorption for the light rays is known as **selective absorption** or **dichroism**. Biot discovered this phenomenon in 1815. When natural light passes through a crystal such as tourmaline, it is split into two components, which are polarized in mutually perpendicular planes. The crystal absorbs light that is polarized in a direction parallel to a particular plane in the crystal but freely transmits the light component polarized in a direction perpendicular to that plane. If the crystal is of proper thickness, one of the components is totally absorbed and the other component emerging from the crystal is linearly polarized. Selective absorption is illustrated in Fig. 8.18.

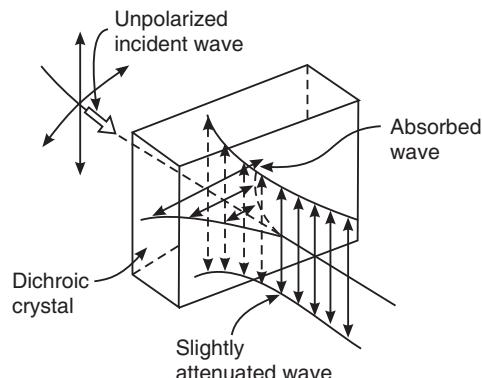


Fig. 8.18

8.6.5 Polarization by Double Refraction

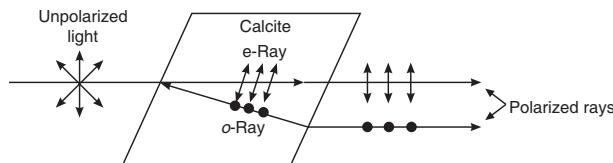


Fig. 8.19

When a beam of unpolarized light is incident on the surface of an anisotropic crystal such as calcite or quartz, it is found that it will separate into two rays (see Fig. 8.19) that travel in different directions. This phenomenon is called **birefringence** or **double refraction**. The two

rays are known as **ordinary ray (o-ray)** and **extraordinary ray (e-ray)**, which are linearly polarized in mutually perpendicular directions. A single linearly polarized ray is obtained in practice through elimination of one of the two polarized rays.

8.7 POLAROID SHEETS

In 1928 E. H. Land invented Polaroid sheets, which utilize the phenomenon of selective absorption. The sheets are fabricated as follows. A clear plastic sheet of long chain molecules of PVA (polyvinyl alcohol) is heated and then stretched in a given direction to many times its original length. During the stretching process the PVA molecules become aligned along the direction of stretching. The sheet is then laminated to a rigid sheet of plastic to stabilise its size. It is then exposed to iodine vapour. The iodine atoms attach themselves to the straight long chain PVA molecules and consequently form long parallel conducting chains. The iodine atoms provide electrons, which can move easily along the aligned chains, but not perpendicular to them. When natural light is incident on the sheet, the electromagnetic vibrations that are in a direction parallel to the alignment of the iodine atoms are strongly absorbed because of the dissipative effects of the electron motion in the chains. Consequently, only those vibrations in a direction perpendicular to the direction of molecular chains are transmitted. Thus, the light transmitted through the polaroid sheet is polarized. A sheet fabricated according to this process is known as H-sheet.

These sheet polarizers are inexpensive and can be made in large sizes. Polaroid sheets are widely used in sunglasses, camera filters etc to eliminate the unwanted glare from objects.

Polaroid sheets are extensively used for the production and detection of linearly polarized light.

8.8 POLARIZER AND ANALYZER

A **polarizer** is an optical element, which utilizes the phenomenon of selective absorption or double refraction, and transforms unpolarized light into polarized light. Plane polarized light is obtained by eliminating one of the two components in the unpolarized light. When natural light is incident on a polarizer, the **E**-field component that is parallel to the chains of iodine atoms induces current in the conducting chains and is therefore strongly absorbed. Consequently, the light transmitted contains only the component that is perpendicular to the direction of **E**-vector in the transmitted beam corresponds to the transmission axis of the Polaroid sheet.

Effect of polarizer on natural light

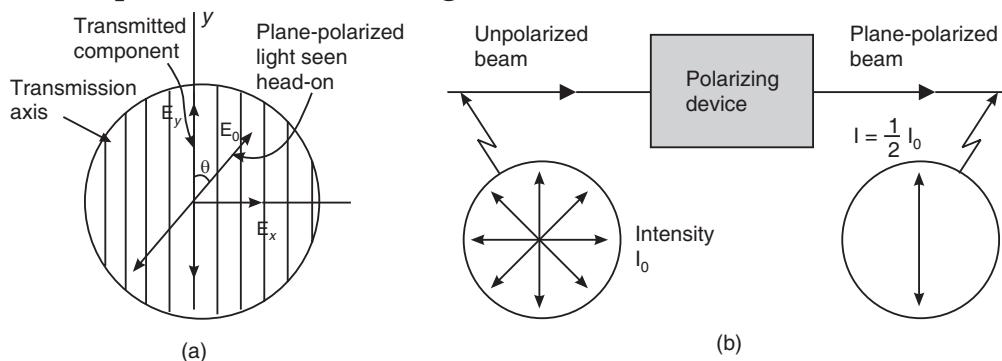


Fig. 8.20: (a) Action of polarizer on linearly polarized wave (b) The intensity of an unpolarized beam reduces to half after passing through a polarizer

Let us now understand the action of polarizer on the incident unpolarized light. Let us consider unpolarized light incident on a polarizer with the electric vector \mathbf{E}_0 making an angle θ with respect to the *transmission axis* of the polarizer. The electric vector \mathbf{E}_0 may be resolved into its component vectors lying parallel and perpendicular to the transmission axis of the polarizer (see Fig. 8.20), that is E_y , parallel to the transmission axis and E_x , perpendicular to the transmission axis of the polarizer. The polarizer transmits the parallel component while blocking the perpendicular component. Thus, it is the parallel component E_y that is transmitted by the polarizer. But

$$E_y = E_0 \cos \theta \quad (8.7)$$

and hence, the intensity of the transmitted component is given by

$$I \propto E^2 = E_0^2 \cos^2 \theta \quad (8.8)$$

In unpolarized light all the values of θ are equally probable. Therefore, the fraction of light transmitted through the polarizer equals the average value of $\cos^2 \theta$, which is equal to $\frac{1}{2}$.

Thus,

$$\therefore I = \frac{1}{2} E_0^2 = \frac{1}{2} I_0 \quad (8.9)$$

An *analyzer* is an optical element, which is used to identify the plane of vibration of plane polarized light. An analyzer is not different from a polarizer in its structure. It differs from a polarizer only in its working.

8.8.1 Production of Linearly Polarized Light Using a Polarizer

A polarizer is associated with a specific direction called the **transmission axis** of the polarizer. If natural light is incident on a polarizer, only those vibrations that are **parallel** to the transmission axis are allowed through the polarizer whereas the vibrations that are in perpendicular directions are totally blocked. Therefore, the transmitted light contains waves oscillating in the same plane, as illustrated in Fig. 8.21. Thus, the transmitted beam is linearly polarized.

According to equ.(8.9), when unpolarized light of intensity I_0 is incident on a polarizer, the intensity of light transmitted by the polarizer is $I_0/2$.

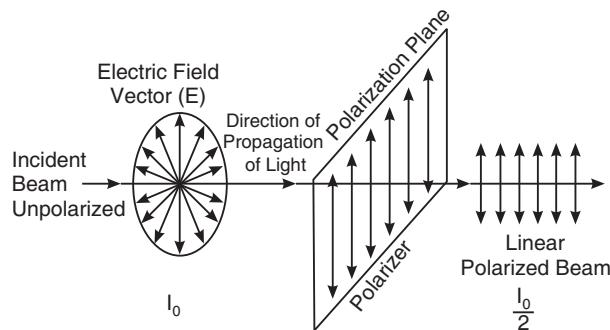


Fig. 8.21: Production of linearly polarized light

8.8.2 Detection of Linearly Polarized Light

To examine light coming from some direction either after emission or reflection etc, we use a Polaroid sheet. The Polaroid sheet used to determine the plane of polarization of light is known as an **analyzer**. There is no difference between a polarizer and analyzer in fabrication but they differ in their roles. Both the polarizer and analyzer are characterized by a transmission axis.

When the transmission axis of the analyzer A is set up parallel to that of polarizer P, light transmitted by the polarizer, passes unhindered through the analyzer (Fig. 8.22 a).

If the transmission axes are set at an angle θ , light is partially transmitted (Fig. 8.22 b). As the angle rotates from 0 to 90 degrees, the amount of light that is transmitted decreases.

When the axes are perpendicular to each other, the polarized light from P is extinguished by the analyzer A (Fig. 8.22 c). The polarizer and analyzer are said to be **crossed** in this configuration.

When we rotate the axis of the analyzer with respect to that of the polarizer, we obtain two positions of maximum intensity and two positions of zero intensity in *one full rotation*. Conversely, if we obtain two positions of maximum intensity and two positions of zero intensity in one full rotation of the analyzer, we conclude that the light incident on the analyzer is plane-polarized light.

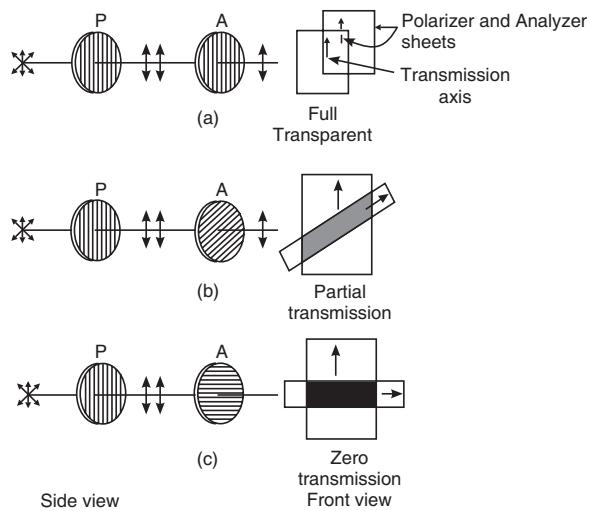


Fig. 8.22

8.9 MALUS' LAW

The amount of light transmitted through a polarizer at an arbitrary angle [Fig. 8.22 (b)] is given by **Malus's Law**.

In 1809 Malus found that the *intensity of polarized light transmitted through a polarizer is proportional to the square of cosine of the angle between the plane of polarization of the light and the transmission axis of the polarizer*.

This statement is known as Malus' law.

If unpolarized light of intensity I_0 is incident on a polarizer, plane polarized light

of intensity $I_0/2$ is transmitted by it. Let us denote $I_0/2$ by I_1 . Let this plane polarized light pass through an analyzer. The intensity of the light transmitted through the analyzer is given by

$$I = E_1^2 \cos^2 \theta = I_1 \cos^2 \theta = \frac{1}{2} I_0 \cos^2 \theta \quad (8.10)$$

Light transmitted through the analyzer at specific settings are as follows.

Case (i): If $\theta = 0^\circ$	axes parallel	$I = I_1 = \frac{1}{2} I_0$
---------------------------------	---------------	-----------------------------

Case (ii): If $\theta = 90^\circ$	axes perpendicular	$I = 0$
-----------------------------------	--------------------	---------

Case (iii): If $\theta = 180^\circ$	axes parallel	$I = I_1 = \frac{1}{2} I_0$
-------------------------------------	---------------	-----------------------------

Case (iv): If $\theta = 270^\circ$	axes perpendicular	$I = 0$
------------------------------------	--------------------	---------

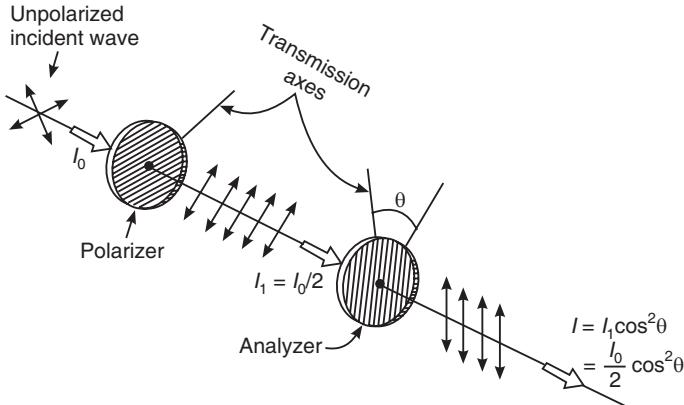


Fig. 8.23

Thus, we obtain two positions of maximum intensity and two positions of zero intensity when we rotate the axis of the analyzer with respect to that of the polarizer.

Example 8.4: Unpolarized light falls on two polarizing sheets placed one on top of the other. What must be the angle between the characteristic directions of the sheets if the intensity of the transmitted light is one-third intensity of the incident beam?

Solution:

Intensity of the light transmitted through the first polarizer $I_1 = I_0/2$, where I_0 is the intensity of the incident unpolarized light.

Intensity of the light transmitted through the second polarizer is $I_2 = I_1 \cos^2\theta$ where θ is the angle between the characteristic directions of the polarizer sheets.

But

$$I_2 = I_0 / 3 \quad (\text{given})$$

∴

$$I_2 = I_1 \cos^2\theta = I_0 \cos^2\theta/2 = I_0/3$$

$$\cos^2\theta = 2/3 \quad \text{or} \quad \cos \theta = 0.8165$$

∴

$$\theta = 35.3^\circ$$

Example 8.5: Light of intensity I_0 is incident on a polarizer. What is the intensity of the resultant beam if: (i) incident light is unpolarized? (ii) incident light is plane polarized with its electric field making an angle of 30° with the axis of the polarizer?

Solution: If incident beam is unpolarized, then the intensity of the resultant beam will be $I_0/2$.

When the incident light is plane polarized, according to Malus' law

$$I = I_0 \cos^2\theta = I_0 \cos^230^\circ = I_0(0.866)^2 = (3/4)I_0$$

8.10 ANISOTROPIC CRYSTALS

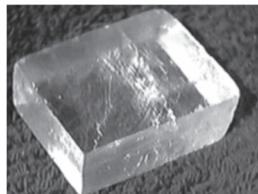
When a light beam is incident on an *isotropic medium* such as a glass slab, it refracts as a single ray. An optically isotropic material is one in which the index of refraction is the same in all directions. Glass, water and air are examples of isotropic materials. The atoms in a crystal are arranged in a regular periodic manner. If the arrangement of atoms differ in different directions within a crystal, then the physical properties vary with the direction. The thermal conductivity, electrical conductivity, velocity of light and hence refractive index etc properties depend on the crystallographic direction along which the property is measured. Then we say that the crystal is **anisotropic**. In such anisotropic crystals the force of interaction between the electron cloud and the lattice is different in different crystallographic directions. The natural frequency of the electron cloud is likewise dependent on the direction in which the electrons are caused to vibrate by the incident light wave. This results in different velocities in different directions and the index of refraction is different in different directions within the crystal.

The anisotropic crystals are divided into two classes: *uniaxial* and *biaxial* crystals. In case of **uniaxial crystals**, one of the refracted rays is an ordinary ray and the other is an extraordinary ray. In **biaxial crystals** both the refracted rays are extraordinary rays. Calcite, tourmaline and quartz are examples of uniaxial crystals whereas mica, topaz and aragonite are examples of biaxial crystals.

8.10.1 Calcite Crystal

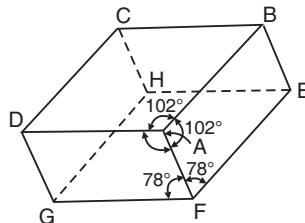
Calcite is a common naturally occurring substance. Both marble and limestone are made up of many small calcite crystals bonded together. A large crystal of calcite is colourless and transparent. It was at one time found in great quantities in Iceland and hence it is also known as **Iceland spar**. Naturally occurring calcite crystals (Fig. 8.24a) has rhombohedral cleavage which means it breaks into blocks with parallelogram-shaped faces. It is bounded by six faces (Fig. 8.24b), each of which is a parallelogram with angles equal to $101^\circ 55'$ and $78^\circ 5'$. The rhombohedron has only two corners A and H where all the face angles are obtuse ($101^\circ 55'$).

These two corners appear as the *blunt corners* of the crystal. At the rest of six corners there is one obtuse angle and two acute angles.

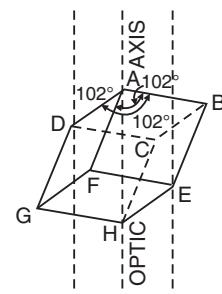


Calcite Crystal

(a)



(b)



(c)

Fig. 8.24

8.10.2 Optic Axis

A line bisecting any one of the blunt corners (A or H) and making equal angles with each of the three edges meeting there, is the **optic axis** (see Fig. 8. 24c). In fact any line parallel to this line is also an optic axis. Thus, the optic axis is a direction and not a specific line in the crystal. Hence an optic axis can be drawn through every point in the crystal, that is, any line parallel to the line above will also be the optic axis. It is to be noted that the optic axis is not obtained by joining the two blunt corners. Only in a special case, when the three edges of the crystal are equal, the line joining the two blunt corners A and H coincide with the crystallographic axis of the crystal and it gives the direction of the optic axis. The optic axis is actually the axis of symmetry of the crystal. A ray of light propagating along optic axis does not suffer double refraction, because the structure of the crystal is symmetric about that direction.

The optic axis is the direction in a uniaxial crystal along which the e-ray and the o-ray travel with the *same* speed and consequently double refraction *does not* take place along this direction. The corresponding refractive index is the refractive index for ordinary light, say μ_o .

8.10.3 Principal Section

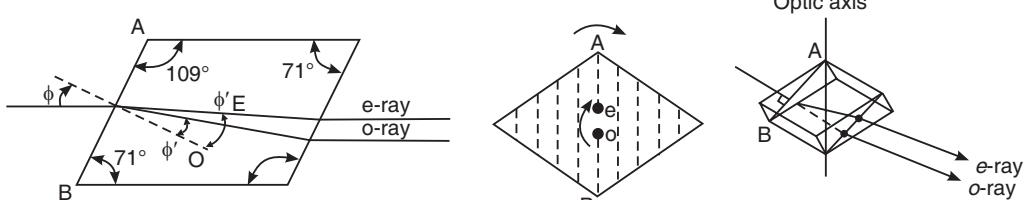


Fig. 8.25

A plane containing the optic axis and perpendicular to a pair of opposite faces of the crystal is called the principal section of the crystal for that pair of faces (Fig. 8. 25c). Thus, there are three principal sections passing through any point within the crystal, one corresponding to each pair of opposite faces. A principal section always cuts surfaces of calcite crystal in a parallelogram having angles 71° and 109° (see Fig. 8.25 a). Fig. 8.25 (b) shows a face of the crystal in which the end-view of the principal section AB is shown by the dotted line AB. The lines parallel to AB represent the end-views of other principal sections parallel to AB with in the crystal.

8.10.4 Principal Plane

Defining principal section is not enough to understand the directions of vibrations for the o-ray and e-rays. Hence, two more planes are defined as *principal plane for the o-ray* and the *principal plane for the e-ray*. The plane containing the optic axis and the o-ray is called the *principal plane of the o-ray* and the plane containing the optic axis and the e-ray is called the *principal plane of the e-ray*. The directions of vibrations in the o-ray and e-ray can be understood with reference to these planes. In general, the two principal planes do not coincide. Under the particular case, when the plane of incidence is the principal section of the crystal, then the principal planes of o- and e-rays and the principal section of the crystal coincide.

8.11 DOUBLE REFRACTION IN CALCITE CRYSTAL

Fig. 8.25 (a) shows a principal section of calcite crystal. A ray of light is incident on the face AB of the crystal and it travels along the principal section. The ray is split into two rays, namely o-ray (fast ray) and e-ray (slow ray). The o-ray travels through the crystal without deviation while the e-ray is refracted at some angle. As the opposite faces of the crystal are parallel, the rays emerge out parallel to the incident ray. Within the crystal the o-ray *always* lies in the plane of incidence whereas e-ray does not lie in the plane of incidence. e- ray lies in the plane of incidence only when the plane of incidence is a principal section.

If a mark (dot or cross) is made on a paper and then the calcite crystal (AB face) is placed on it, two images are seen through the crystal, as illustrated in Fig. 8.25 (b). The images are produced by the o-ray and e-ray. The intensities of the images are lesser than that of the original mark. The line joining them lies in the principal section. If now the crystal is rotated slowly about an axis passing through the o-image, the e-image moves round in a circle while the o-image remains stationary. It shows that the velocity of propagation of o-ray is the same in all directions, while that of e-ray changes with direction. O-ray obeys the laws of refraction and the e-ray does not follow the ordinary laws of refraction.

The e-ray and o-ray are linearly polarized. The e-ray has its vibrations (i.e., the optical vector) *parallel* to the principal section whereas the vibrations (optical vector) in o-ray are *perpendicular* to the principal section, as indicated in Fig. 8.19. The vibration directions can be established by examining the rays through a polarizer. As the polarizer is held in the path of the rays and rotated slowly, the intensity of one of the images, say the o-image, increase while that of the e-image decreases. In one position, the intensity of the o-image will be a maximum while the e-image is extinguished. Further rotation through 90° from this particular position, causes the o-image to disappear and e-image intensity to become a maximum. It proves that the e- and o-rays are linearly polarized in mutually perpendicular directions.

The o-ray travels with the same velocity in all directions in the crystal whereas the e-ray travels with different velocities in different directions. Therefore, refractive index corresponding to o-ray is a constant and is designated by μ_o . The refractive index corresponding to e-ray varies and its maximum (or minimum) value is denoted by μ_e . The difference between the refractive indices is known as the amount of double refraction or birefringence. Thus,

$$\Delta\mu = \mu_e - \mu_o \quad (8.11)$$

8.11.1 Huygens' Explanation of Double Refraction

In order to explain the double refraction exhibited by anisotropic crystals, Huygens postulated that the incident light excites two separate wavelets within the crystal, one spherical wavelet associated with the ordinary waves and one ellipsoidal wavelet associated with the extraordinary waves. For example, the plane wavefront, in Fig. 8.26, incident normally on the crystal

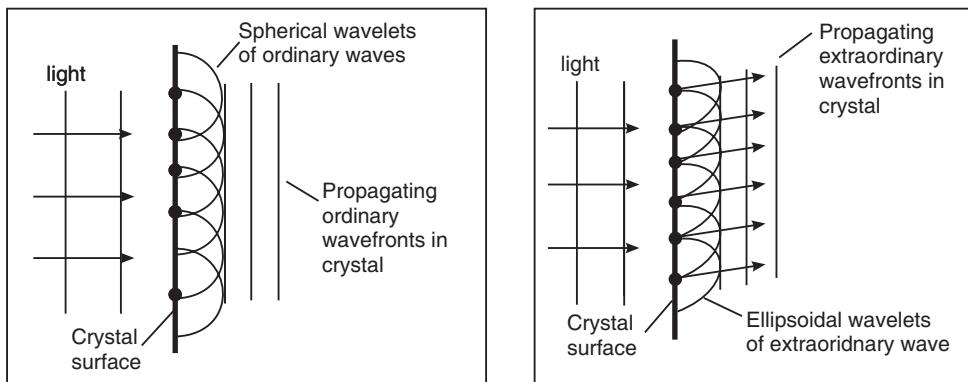


Fig. 8.26

surface generates spherical wavelets as well as ellipsoidal wavelets. The spherical wavelets propagate equally quickly in all directions (Fig. 8.26a). The wave surface corresponding to o-ray is therefore spherical. The tangent to these waves lies straight ahead and, by successive application of the principle, the plane wave propagates straight ahead with speed v_0 . The ellipsoidal wavelets propagate at different speeds in different directions. The wave surface corresponding to e-ray is therefore an ellipsoid of revolution about the optic axis. The common tangent to these ellipsoids after a little time is the new wavefront. The line from the point of generation of each ellipsoid to the tangent point on that ellipsoid is off at an angle and defines the direction of travel of the extraordinary wavefronts. The wavefronts are not perpendicular to their direction of travel (Fig. 8.26b).

The two wave surfaces touch each other at the two points where they are intersected by the optic axis. As light propagates through the crystal, the two wave surfaces travel in different directions in the crystal. Ultimately, two refracted rays emerge from the crystal.

8.11.2 Ordinary and Extra-ordinary Rays

We now compare the properties of o- and e-rays.

- (i) o-ray obeys the laws of refraction and the e-ray does not follow the ordinary laws of refraction.
- (ii) Both o-ray and e-ray are plane polarized. They are polarized in mutually perpendicular planes. The electric vector of o-ray vibrates perpendicular to the principal section of o-ray while the vibrations of e-ray take place parallel to the principal section of e-ray.
- (iii) O-ray travels with the same speed in all directions within the crystal. The e-ray travels with different speeds along different directions in the crystal. However, the speed of e-ray will be equal to that of o-ray along the optic axis direction.
- (iv) Because o-ray travels with the same velocity in all directions, the refractive index corresponding to it has a constant value. On the other hand, the refractive index for e-ray varies from direction to direction. The principal refractive index for o-ray is defined as follows:

$$\mu_0 = \frac{c}{v_0} = \frac{\text{velocity of light in a vacuum}}{\text{velocity of o-ray in the crystal}} \quad (8.12)$$

The principal refractive index for e-ray in ***positive crystals*** is defined as follows:

$$\mu_e = \frac{c}{(v_e)_{\min}} = \frac{\text{velocity of light in a vacuum}}{\text{minimum velocity of e-ray in the crystal}} \quad (8.13)$$

The principal refractive index for e-ray in ***negative crystals*** is defined as follows:

$$\mu_e = \frac{c}{(v_e)_{\max}} = \frac{\text{velocity of light in a vacuum}}{\text{maximum velocity of e-ray in the crystal}} \quad (8.14)$$

- (v) When natural light is incident on an anisotropic crystal at an angle to the optic axis, it splits into o-and e-rays, which travel in *different directions* with *different velocities* (8.27 a).

When natural light is incident a direction perpendicular to the optic axis, o-ray and e-ray propagate in the *same direction* the crystal but with *different velocities*, as shown Fig. 8.27 (b). In a negative crystal

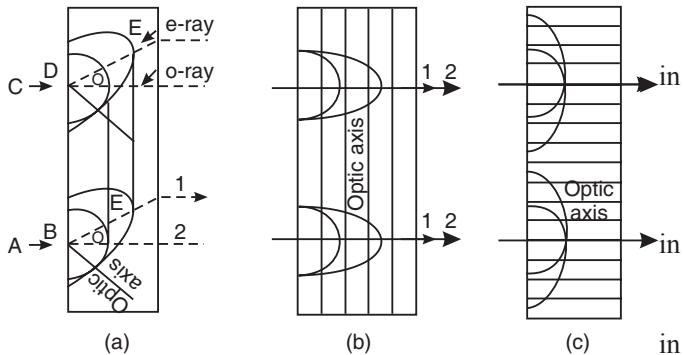


Fig. 8.27

e-ray leads o-ray and in case of a positive crystal o-ray leads e-ray.

When natural light is incident on the crystal in a direction parallel to the optic axis, it does not split into two rays. The o- and e- rays travel in the *same direction with the same velocity*, as shown in Fig. 8.27 (t).

- (vi) The distinction of o-ray and e-ray exists only within the crystal. Once they emerge from the crystal, they travel with the same velocity. The rays outside the crystal differ only in their direction of travel and plane of polarization. The designation of o- ray and e-ray has no meaning outside the crystal.

8.11.3 Positive and Negative Crystals

Because of two different wave fronts, two different types of uniaxial crystals exist. In one type of crystals, the spherical wave front of o-ray is enclosed by the ellipsoidal wave front of e-ray. Such crystals are known as ***negative crystals***. They are called negative crystals because the refractive index corresponding to the e-ray is less than that corresponding to o-ray. Calcite crystal is an example of negative type crystals. In the other case, the extraordinary wave front lies within the ordinary wave front and such crystals are called ***positive crystals***. They are positive because the refractive index for the extraordinary ray is greater than that of o-ray. Quartz crystal is an example of positive crystals.

We compare here the characteristics of the positive and negative crystals.

- (i) In positive uniaxial crystals, the ellipsoid of revolution corresponding to the e-ray is totally contained within the sphere corresponding to the o-ray.
In negative uniaxial crystals, the ellipsoid of revolution for e-ray lies completely outside the sphere corresponding to o-ray. The two cases are depicted in Fig. 8.28.
- (ii) In positive crystals the e-ray velocity has a maximum value along the optic axis and a minimum value in a direction perpendicular to the optic axis.

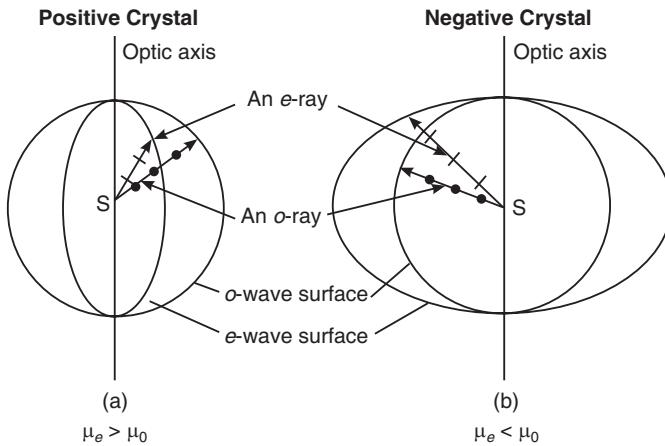


Fig. 8.28

On the other hand, in negative crystals the velocity of e-ray has a minimum value parallel to the optic axis and a maximum value in a direction perpendicular to the optic axis.

- (iii) In positive crystals, e-ray travels slower than o-ray in all directions except along the optic axis.

$$\begin{array}{ll} v_e = v_o & \text{— parallel to optic axis} \\ v_e < v_o & \text{— other directions} \end{array}$$

In negative crystals, o-ray travels slower than e-ray in all directions except along the optic axis.

$$\begin{array}{ll} v_e = v_o & \text{— parallel to optic axis} \\ v_e > v_o & \text{— other directions} \end{array}$$

- (iv) In positive crystals the principal refractive index for e-ray is larger than the principal refractive index for o-ray.

$$\mu_e > \mu_o$$

In negative crystals the principal refractive index for o-ray is larger than the principal refractive index for e-ray.

$$\mu_e < \mu_o$$

- (v) **Birefringence** of a positive crystal is given by

$$\Delta\mu = \mu_e - \mu_o \quad (8.15a)$$

As $\mu_e > \mu_o$ in these crystals, $\Delta\mu$ is a positive quantity for positive crystals. Quartz and ice are examples of positive crystals.

$\Delta\mu$ is a negative quantity for negative crystals as $\mu_e < \mu_o$ in these crystals and therefore

$$\Delta\mu = \mu_o - \mu_e \text{ is a negative quantity.}$$

- The birefringence of a *negative crystal* is given by

$$\Delta\mu = \mu_o - \mu_e \quad (8.15b)$$

Calcite is an example of negative crystals.

8.12 NICOL PRISM

Nicol prism is a polarizing device fabricated from a double refracting crystal. It is similar to a Polaroid sheet in its action. A Nicol prism is made from calcite crystal. William Nicol designed it in 1820. A rhomb of calcite crystal about three times as long as it is thick, is

obtained by cleavage from the original crystal. The ends of the rhombohedron are ground until they make an angle of 68° instead of 71° with the longitudinal edges. This piece is then cut into two along a plane perpendicular both to the principal axis and to the new end surfaces MP and QN. The two parts of the crystal are then cemented together with Canada balsam, whose refractive index lies between the refractive indices of calcite for the o-ray and e-ray. $\mu_o = 1.66$, $\mu_e = 1.486$ and $\mu_{\text{Canada balsam}} = 1.55$. The position of the optic axis AB as shown in Fig. 8.29. The refractive index for e-ray depends upon the direction in which e-ray is propagating in the crystal. The difference between the refractive index between o-ray and that for e-ray goes on increasing with the angle between the two rays in the crystal. When this angle is 90° , the difference is a maximum. Thus, for a fixed value for μ_o , the μ_e has its maximum or minimum value in perpendicular direction. In the above $\mu_e = 1.486$ represents the minimum value.

Unpolarized light is made to fall on the crystal as shown in Fig. 8.29 at an angle of about 15° . The ray after entering the crystal suffers double refraction and splits up into o-ray and e-ray. The two rays with their directions of vibrations are as shown in the Fig. 8.29. The values of the refractive indices and the angles of incidence at the Canada balsam layer are such that the e-ray is transmitted while the o-ray is internally reflected. The face where the o-ray is incident is blackened so that the o-ray is completely absorbed. Then we get only the plane-polarized e-ray coming out of the Nicol. Thus, the Nicol works as a polarizer.

For studying the optical properties of transparent substances, two Nicols are used - one as a polarizer and the other as an analyzer.

When two Nicol prisms P and A are placed adjacent to each other as shown in Fig. 8.30, one of them acts as a polarizer and the other acts as an analyser. If unpolarized ray of light is incident on the Nicol prism P, a linearly polarized e-ray emerges from P with its vibration direction lying in the principal section of P. The state of the polarization of the light emerging from polarizer P can be examined with another Nicol prism A, which for convenience is called an analyser. Let now this ray be incident on the second Nicol prism A, whose principal section is parallel to that of P. The vibration direction of the ray will be in the principal section of A and hence it is transmitted unhindered through the analyser A.

If the Nicol prism A is gradually rotated, the intensity of the e-ray decreases in accordance with Malus law. When its principal section becomes perpendicular to that of the Nicol prism P (Fig. 8.30 b), the vibrations of the ray, emerging from P and incident on A, will be

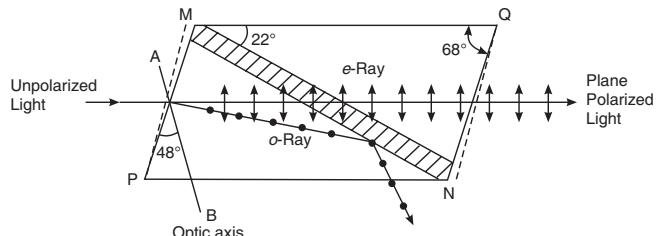


Fig. 8.29

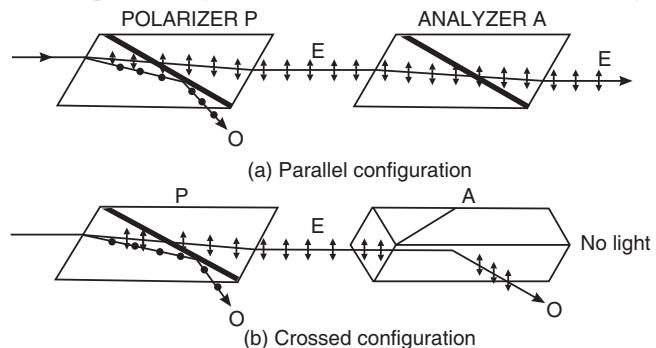


Fig. 8.30

perpendicular to the principal section of A. In this position the ray behaves as o-ray inside the prism A and is totally internally reflected by the Canada balsam layer. Hence no light is transmitted by the prism A. In this configuration, the two Nicol prisms P and A are said to be **crossed**. If the Nicol prism A is further rotated through another 90° , the intensity of light emerging from A will go on increasing. The intensity will become a maximum when its principal section is again parallel to that of the polarizer P. Thus, the Nicol P produces linearly polarised light while the prism A detects it. Hence the prism P is called a polarizer and the prism A an analyser.

8.13 EFFECT OF POLARIZER ON LIGHT OF DIFFERENT POLARIZATIONS

The action of a polarizer, whether a Nicol prism or a Polaroid sheet, on light of different types of polarization is as follows:

- (i) If **unpolarized light** is incident on a polarizer, it transmits half the intensity of light incident on it. The transmitted light stays constant in intensity on rotation of the polarizer.
- (ii) If **partially polarized light** is incident on the polarizer, the intensity of the transmitted light depends on the direction of the transmission axis of the polarizer. The intensity varies from a maximum value I_{\max} to a minimum value I_{\min} . Two positions of I_{\max} and two positions of I_{\min} occur in one full rotation of the polarizer.
- (iii) If **plane polarized light** is incident on the polarizer, the intensity of the transmitted light varies from zero to a maximum twice in one full rotation of the polarizer.
- (iv) When **circularly polarized light** is incident on the polarizer, the transmitted light remains constant in intensity on rotation of the polarizer. This is interpreted as follows: the circular vibrations may be resolved into two mutually perpendicular linear vibrations of equal amplitude. When the circularly polarized light is incident on the polarizer, the vibration parallel to its transmission axis passes through the polarizer while the perpendicular component is not allowed. This happens in all positions of the polarizer in its rotation. Therefore, the intensity of the transmitted light stays constant in one full rotation of the polarizer.
- (v) When **elliptically polarized light** is incident on the polarizer, the transmitted light varies in intensity from a maximum value I_{\max} to a minimum value I_{\min} on rotation of the polarizer (Fig. 8.31). I_{\max} occurs when the transmission axis of the polarizer coincides with the semi-major axis of the ellipse and I_{\min} occurs when the transmission axis coincides with the semi-minor axis of the ellipse.

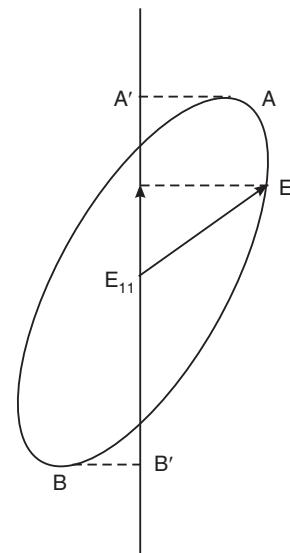


Fig. 8.31

8.14 PHASE DIFFERENCE BETWEEN e-RAY AND o-RAY

We have seen that natural light incident on the surface of an anisotropic crystal undergoes double refraction and produces two plane polarized waves (Fig. 8.32a). Let us consider the particular case of a slice of a calcite crystal (a negative uniaxial crystal) where the optic axis is parallel to refracting face of the crystal. The two waves travel along the same direction in the crystal but with different velocities. As a result, when the waves emerge from the

rear face of the crystal, an optical path difference would have developed between them (Fig. 8.32b). The optical path difference can be calculated as follows:

Let d be the thickness of the crystal.

$$\text{The optical path for o-ray within the crystal} = \mu_o d$$

$$\text{The optical path for e-ray within the crystal} = \mu_e d$$

$$\therefore \text{The optical path difference between e-ray and o-ray} = \Delta = (\mu_o - \mu_e)d \quad (8.16)$$

Consequently, a phase difference arises between the two waves. It is given by

$$\delta = \frac{2\pi}{\lambda}(\mu_o - \mu_e)d \quad (8.17)$$

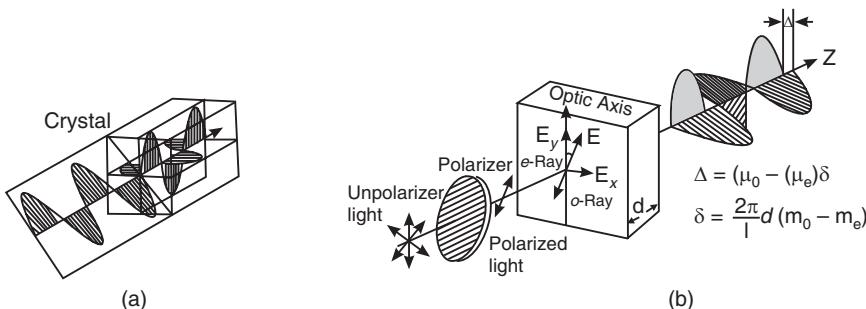


Fig. 8.32

As the two component waves are derived from the same incident wave, the two waves are in phase at the front face and have emerged from the crystal with a constant phase difference and hence it may be expected that the waves are in a position to interfere with each other. However, as the planes of polarization of e-ray and o-ray are perpendicular to each other, interference cannot take place between e- and o-rays. The waves instead combine with each other to give elliptically polarized wave.

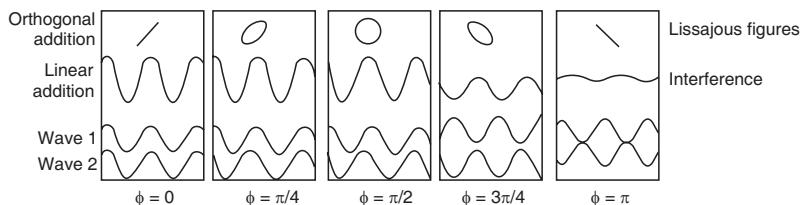


Fig. 8.33

The most important point to be noted here is as follows: The superposition of two coherent waves having a *common plane* of polarization yields a wave polarized in the same plane. The superposition leads to linear addition of waves, i.e., **interference**. The resultant vibration is *linear* and occurs in the same direction as that of the two superposing vibrations (see Fig. 8.33). On the other hand, if the two coherent waves are polarized in orthogonal planes, the resultant vibrational motion takes place in two dimensions either in the form of an ellipse, a circle or a straight line depending on the phase difference between the waves (Fig. 8.33). This is known as *orthogonal addition*.

Example 8.6: Plane-polarized light of wavelength 5400 \AA is incident perpendicularly on a quartz plate cut with faces parallel to optic axis. Find the thickness of quartz plate, which introduces phase difference of 60° between e- and o-rays.

Solution: The path difference between the waves is given by $\Delta = (\mu_e - \mu_o)d$

The phase difference between the waves is given by $\delta = \frac{2\pi}{\lambda}\Delta$

$$\Delta = \frac{60^\circ}{360^\circ} \lambda = \frac{\lambda}{6} \quad \therefore d = \frac{\lambda}{6(\mu_e - \mu_o)}$$

or $d = \frac{5400 \text{ \AA}}{6(1.553 - 1.544)} = \frac{0.54}{0.054} \mu\text{m} = 10 \mu\text{m}$.

Example 8.7: Plane-polarized light of wavelength 6000 \AA is incident perpendicularly on a calcite plate of thickness 0.04 mm . Calculate the phase retardation that it will introduce between the e-ray and o-ray. Given that $\mu_o = 1.642$ and $\mu_e = 1.478$.

Solution: The phase difference between the waves is given by

$$\delta = \frac{2\pi}{\lambda}\Delta = \frac{2\pi}{\lambda}d(\mu_e - \mu_o)$$

$$\therefore \delta = \frac{2 \times 3.143}{6000 \times 10^{-10} \text{ m}} \times 4 \times 10^{-5} \text{ m}(1.642 - 1.478) = 68.7 \text{ rad.}$$

8.15 SUPERPOSITION OF WAVES LINEARLY POLARISED AT RIGHT ANGLES

Let us now look at the result of superposition of two waves linearly polarized at right angles to each other (see Fig. 8.34 a). Let us consider two light waves travelling in the same direction, x ; one wave is polarized in the xy -plane and the other is polarized in yz -plane. We are interested to know the state of polarization of the resultant wave at the plane $x = \text{constant}$.

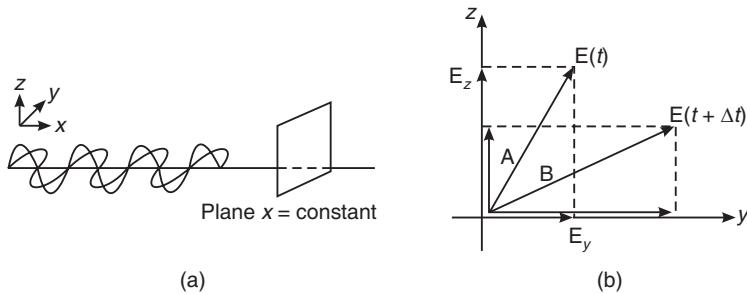


Fig. 8.34

Let the two orthogonal waves be represented by

$$E_y = E_1 \cos(kx - \omega t) \quad (8.18)$$

$$E_z = E_2 \cos(kx - \omega t + \delta) \quad (8.19)$$

The waves are of the same frequency $v = \omega/2\pi$. δ is the phase difference between the waves. At a given time t , the optical vectors E_y and E_z produce a resultant optical vector of magnitude, say A and at a slightly later time, $t + \Delta t$, they give rise to a resultant vector of magnitude B which points in a different direction (see Fig. 8.34b). With progress of time the tip of the resultant optical vector moves along a curve in the yz -plane. We apply the principle of superposition to find the equation of the curve traced by the tip of the resultant vector of the two vectors.

According to the principle of superposition

$$E = E_y + E_z$$

$$= E_1 \cos(kx - \omega t) + E_2 \cos(kx - \omega t + \delta) \quad (8.20)$$

The equation of the curve may be found by eliminating 't' from the equations. We can write the expansion of eq.(8.20) as

$$\begin{aligned} E_z &= E_2 \cos(kx - \omega t) \cos \delta - E_2 \sin(kx - \omega t) \sin \delta \\ &= E_2 \cos(kx - \omega t) \cos \delta \pm [1 - \cos^2(kx - \omega t)]^{1/2} E_2 \sin \delta \end{aligned} \quad (8.21)$$

We find from equ.(8.18) that

$$\begin{aligned} \cos(kx - \omega t) &= E_y/E_1 \\ \therefore E_z &= E_2 \frac{E_y}{E_1} \cos \delta \pm \sqrt{1 - \left(\frac{E_y}{E_1}\right)^2} E_2 \sin \delta \end{aligned} \quad (8.22)$$

Rearranging the terms, we get

$$\left[E_z - \frac{E_2}{E_1} E_y \cos \delta \right] = \pm \sqrt{1 - \left(\frac{E_y}{E_1}\right)^2} E_2 \sin \delta$$

On squaring both the sides, we obtain

$$E_z^2 + \frac{E_y^2 E_2^2}{E_1^2} \cos^2 \delta - \frac{2E_y E_z E_2}{E_1} \cos \delta = E_2^2 \sin^2 \delta - \frac{E_y^2 E_2^2}{E_1^2} \sin^2 \delta$$

Rearranging the terms, we get

$$\begin{aligned} E_z^2 + \frac{E_y^2 E_2^2}{E_1^2} (\cos^2 \delta + \sin^2 \delta) - \frac{2E_y E_z E_2}{E_1} \cos \delta &= E_2^2 \sin^2 \delta \\ E_z^2 + \frac{E_y^2 E_2^2}{E_1^2} - \frac{2E_y E_z E_2}{E_1} \cos \delta &= E_2^2 \sin^2 \delta \end{aligned} \quad (8.23)$$

Dividing both the sides by E_2^2 and rearranging the terms, we obtain

$$\frac{E_y^2}{E_1^2} + \frac{E_z^2}{E_2^2} - \frac{2E_y E_z}{E_1 E_2} \cos \delta = E_2^2 \sin^2 \delta \quad (8.24)$$

Equation (8.24) is the general equation of an *ellipse*. Hence, the tip of the resultant vector traces an ellipse in the yz -plane. The ellipse is constrained within a rectangle having sides $2E_1$ and $2E_2$. The major axis makes an angle α with the y -axis (see Fig. 8.35).

$$\tan 2\alpha = 2E_1 E_2 \cos \delta / (E_1^2 - E_2^2)$$

Special cases:

- When $\delta = 0$, or $\pm 2m\pi$, the two waves are in phase. $\cos \delta = 1$ and $\sin \delta = 0$ and the equ. (8.24) reduces to

$$\frac{E_y^2}{E_1^2} + \frac{E_z^2}{E_2^2} - \frac{2E_y E_z}{E_1 E_2} = 0$$

$$\left[\frac{E_y}{E_1} - \frac{E_z}{E_2} \right]^2 = 0$$

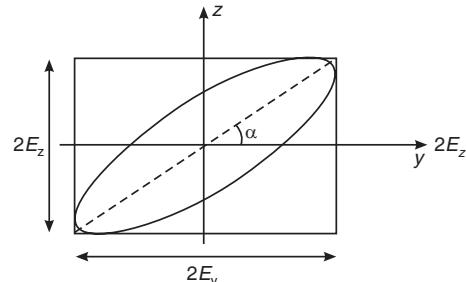


Fig. 8.35

$$\begin{aligned} \frac{E_y}{E_1} - \frac{E_z}{E_2} &= 0 \\ \therefore E_z &= \frac{E_2}{E_1} E_y \end{aligned} \quad (8.25)$$

The above equation represents a straight line, having a slope (E_2/E_1) . Therefore, the equation represents a wave having its plane of polarisation making an angle $\tan^{-1}(E_2/E_1)$ with respect to the y-axis. It means that **the resultant of two plane-polarised waves, which are in phase (i.e., coherent waves), is again a plane-polarised wave.**

2. When $\delta = \pi$, or $\pm (2m + 1)\pi$, the two waves are in opposite phase.

$\cos \delta = -1$ and $\sin \delta = 0$ and the equ.(8.24) reduces to

$$\begin{aligned} \frac{E_y^2}{E_1^2} + \frac{E_z^2}{E_2^2} + \frac{2E_y E_z}{E_1 E_2} &= 0 \\ \left[\frac{E_y}{E_1} + \frac{E_z}{E_2} \right]^2 &= 0 \\ \frac{E_y}{E_1} + \frac{E_z}{E_2} &= 0 \\ \therefore E_z &= -\frac{E_2}{E_1} E_y \end{aligned} \quad (8.26)$$

This equation represents a straight line of a slope $(-E_2/E_1)$. Therefore, the equation represents a wave having its plane of polarization making an angle $\tan^{-1}(-E_2/E_1)$ with respect to the y-axis. It means that **the resultant of two plane-polarized waves, which are in opposite phase (i.e., coherent waves), is again a plane-polarized wave.**

3. If $\delta = \pi/2$, or $\pm (2m + 1)\pi/2$, then $\cos \delta = 0$ and $\sin \delta = 1$. Equ.(8.24) reduces to

$$\frac{E_y^2}{E_1^2} + \frac{E_z^2}{E_2^2} = 1 \quad (8.27)$$

This is the equation of an ellipse whose major axis and minor axis coincide with y- and z-co-ordinate axes. Therefore, when the two plane polarized waves are out of phase by 90° , their resultant is an elliptically polarized wave.

4. In the particular case, when $\delta = \pi/2$ and $E_1 = E_2 = E_0$, equ.(8.24) reduces to

$$E_y^2 + E_z^2 = E_0^2 \quad (8.28)$$

This is the equation of a circle. Hence the resultant light is circularly polarized.

Fig. 8.36 shows more generally how the E vector changes with time for various values of δ .

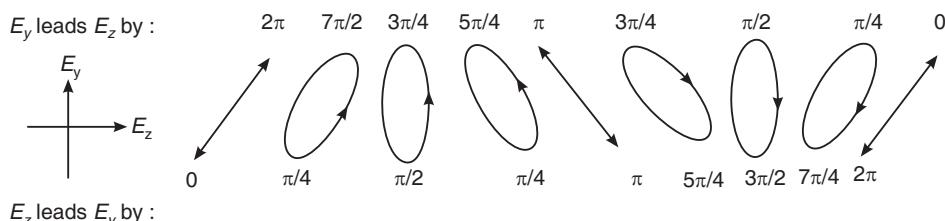


Fig. 8.36

Example 8.8: At a given point the electric fields of component waves of a polarized wave are
 $E_x = 10 \cos \omega t$ and $E_y = 20 \cos(\omega t + \pi)$

Determine the type of polarization and the direction of polarization.

Solution: At any instant, t , we have $E_x = 10 \cos \omega t$ and $E_y = -20 \cos \omega t$.

$$\therefore \cos \omega t = \frac{E_x}{10} = -\frac{E_y}{20} \quad \text{or} \quad E_y = -2E_x$$

It means that the light is plane polarized in xy-plane with slope (-2) . As $\tan^{-1}(-2) = -63.4^\circ$, the light is polarized at an angle of -63.4° with the x-axis.

Example 8.9: Find out the state of polarization represented by the following equations.

$$\begin{aligned} E_x &= E_0 \sin(\omega t - kz) \\ \text{and} \quad E_y &= E_0 \sin(\omega t - kz) \end{aligned}$$

Solution: $E_x = E_0 \sin(\omega t - kz)$
 $E_y = E_0 \sin(\omega t - kz)$

Squaring the above equations, we get

$$\begin{aligned} E_x^2 &= E_0^2 \sin^2(\omega t - kz) \\ E_y^2 &= E_0^2 \sin^2(\omega t - kz) \end{aligned}$$

Adding the above equations, we get

$$E_x^2 + E_y^2 = E_0^2$$

The equation represents a circle. Therefore, the light is circularly polarized.

8.16 RETARDERS

A **retarder** is a uniform plate of birefringent material whose optic axis lies in the plane of the plate. Retarders are called quarter-wave plates, half-wave plates and full-wave plates depending on their action. They divide the incident wave into two polarized waves that travel perpendicular to the plate at different speeds. A phase retardation of one wave relative to the other is therefore introduced as the waves cross the thickness d of the plate. They are used to produce circularly or elliptically polarised light and to analyse polarised light into its elliptical components.

8.16.1 Quarter Wave Plate

A *quarter wave plate* is a thin plate of birefringent crystal having the optic axis parallel to its refracting faces and its thickness adjusted such that it introduces a quarter-wave ($\lambda/4$) path difference (or a phase difference of 90°) between the e-ray and o-ray propagating through it.

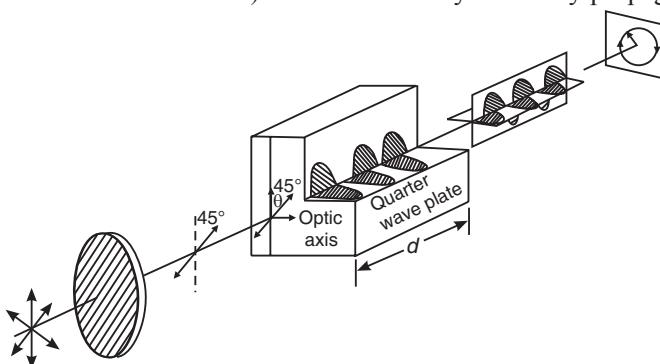


Fig. 8.37

When a plane polarized light wave is incident on a negative birefringent crystal having the optic axis parallel to its refracting face, the wave splits into e-wave and o-wave (Fig. 8.37). The two waves travel along the same direction but with different velocities. As a result, when they emerge from the rear face of the crystal, an optical path difference would be developed between them. Thus,

$$\therefore (\mu_o - \mu_e)d = \frac{\lambda}{4} \quad (\text{calcite plate is assumed}) \quad (8.29)$$

$$d = \frac{\lambda}{4[\mu_o - \mu_e]} \quad (8.30)$$

A quarter wave plate introduces a phase difference δ , between e-ray and o-ray given by $\delta = (2\pi/\lambda)\Delta = \pi/2 = 90^\circ$.

A quarter-wave plate is used for producing elliptically or circularly polarized light. It converts plane-polarized light into elliptically or circularly polarized light depending upon the angle that the incident light vector makes with the optic axis of the quarter wave plate.

Example 8.10: Plane-polarized light passes through a double refracting crystal of thickness $40 \mu\text{m}$ and emerges out as circularly polarized light. If the birefringence of the crystal is 0.00004 , find the wavelength of the incident light.

Solution: A quarter-wave plate changes plane-polarized light into circularly polarized light. Its thickness is given by

$$d = \frac{\lambda}{4(\mu_e - \mu_o)} \quad (\text{Positive crystal is assumed})$$

$$\therefore \lambda = 4d(\mu_e - \mu_o) = 4 \times 40 \times 10^{-6} \text{m} \times 0.00004 = 6400 \text{\AA.}$$

Example 8.11: A beam of plane polarized light is changed into circularly polarized light by passing it through a slice of crystal 0.003 cm thick. Calculate the birefringence of the crystal assuming this to be the minimum thickness that will produce the effect, ($\lambda = 6 \times 10^{-5} \text{ cm}$).

Solution: Plane polarized light is converted into circularly polarized light by a suitably oriented quarter wave plate. Its thickness is given by

$$d = \frac{\lambda}{4(\mu_e - \mu_o)} = \frac{\lambda}{4\Delta\mu}$$

$$\therefore \Delta\mu = \frac{\lambda}{4d} = \frac{6 \times 10^{-5} \text{cm}}{4 \times 0.003 \text{cm}} = 0.005$$

8.16.2 Half Wave Plate

A half wave plate is a thin plate of birefringent crystal having the optic axis parallel to its refracting faces and its thickness chosen such that it introduces a half-wave ($\lambda/2$) path difference (or a phase difference of 180°) between e-ray and o-ray.

When a plane polarized light wave is incident on a birefringent crystal having the optic axis parallel to its refracting faces, it splits into two waves: o- and e-waves. The two waves travel along the same direction inside the crystal but with different velocities. As a result, when they emerge from the rear face of the crystal, an optical path difference would be developed between them.

$$\therefore (\mu_o - \mu_e)d = \frac{\lambda}{2} \quad (8.31)$$

$$\therefore d = \frac{\lambda}{2(\mu_o - \mu_e)} \quad (8.32)$$

A half wave plate introduces a phase difference δ , between e-ray and o-ray given by $\delta = (2\pi/\lambda)\Delta = \pi = 180^\circ$.

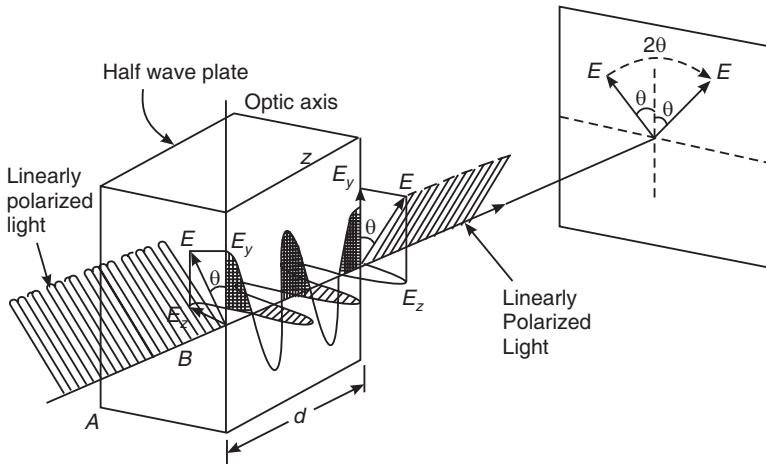


Fig. 8.38

Now let a plane polarized light be incident normally on the half-wave plate. Let the electric vector E make an angle θ with the optic axis of the half wave plate (See Fig. 8.38). The incident wave splits into two waves, e- and o-waves. The waves progressively develop path difference as they travel through the crystal and they emerge with a phase difference of 180° . Note the E_z directions at B on the front face and at the rear face of the crystal plate. When the two waves combine, they yield a plane-polarized wave, which has its plane of polarization rotated through an angle of 2θ .

Therefore, *a half-wave plate rotates the plane of polarization of the incident plane polarized light through an angle 2θ* . The half wave plate will invert the handedness of elliptical or circular polarized light, changing right to left and vice versa.

Conclusion:

Now we are in a position to understand what happens when e-ray and o-ray overlap on each other after emerging from an anisotropic crystal plate. It is obvious that they cannot produce interference fringes as in a double slit experiment. On the other hand, they combine to produce different states of polarization depending upon their optical path difference.

1. When the optical path difference is 0 or an even or odd multiple of $\lambda/2$, the resultant light wave is linearly polarized.
2. When the optical path difference is $\lambda/4$, the resultant light wave is **elliptically polarized**.
3. In the particular instance when the wave amplitudes are equal and the optical path difference is $\lambda/4$, the resultant light wave is **circularly polarized**.

Example 8.12: Plane polarized light is incident on a piece of quartz cut parallel to the axis. Find the least thickness for which the ordinary and extraordinary rays combine to form plane polarized light.

Solution: When plane polarized light is incident on a half wave plate, the emergent beam will also be plane polarized. The least thickness of the quartz plate is given by

$$\therefore d = \frac{\lambda}{2(\mu_e - \mu_o)} = \frac{5 \times 10^{-5} \text{ cm}}{2(91.5533 - 1.5442)} = 27 \text{ } \mu\text{m}$$

Example 8.13: A half-wave plate is fabricated for a wavelength of 3800 \AA . For what wavelength does it work as a quarter-wave plate?

Solution: The thickness of a half-wave plate is $d = \frac{\lambda_1}{2(\mu_e - \mu_o)}$. The same plate is required to act as a quarter-wave plate. Therefore, we can write that $d = \frac{\lambda_2}{4(\mu_e - \mu_o)}$.

$$\therefore d = \frac{\lambda_1}{2(\mu_e - \mu_o)} = \frac{\lambda_2}{4(\mu_e - \mu_o)}$$

$$\therefore \lambda_2 = 2\lambda_1 = 2 \times 3800 \text{ \AA} = 7600 \text{ \AA}.$$

8.17 PRODUCTION OF ELLIPTICALLY POLARIZED LIGHT

A quarter wave plate and a polarizer are the optical devices necessary to produce elliptically polarized light from unpolarized light.

Unpolarized light is first converted to plane polarized light by allowing it to pass through a polarizer (a polaroid sheet or a Nicol prism), as shown in Fig. 8.39. The plane-polarized light is then made incident on a quarter wave plate. The quarter wave plate or the polarizer is rotated such that the electric vector E of plane polarized light wave makes an angle θ ($\neq 45^\circ$) with the optic axis of the quarter wave plate. The incident ray divides into o-ray and e-ray of amplitudes $E \sin \theta$ and $E \cos \theta$. The rays travel along the same direction in the crystal with different velocities. The two rays are polarized in orthogonal planes. They are in phase at the front face but progressively get out of phase as they travel through the crystal. When they emerge out of the crystal they will have a path difference of $\lambda/4$ or a phase difference of 90° . When they combine, they produce elliptically polarized light.

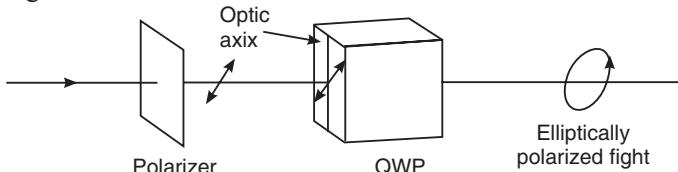


Fig. 8.39: Production of electrically polarized light

8.17.1 Detection of Elliptically Polarized Light

The light beam is allowed to pass through an analyzer (a polaroid sheet or a Nicol prism). If on rotating the analyzing Polaroid sheet or Nicol, the intensity of the emerging beam varies from a maximum

to a minimum value, but never reaching zero, then the incident light is elliptically polarized. A similar result would be obtained if the incident light were partially polarized. The two cases may be distinguished by inserting a quarter wave plate in the path of light before it falls on the analyzer.

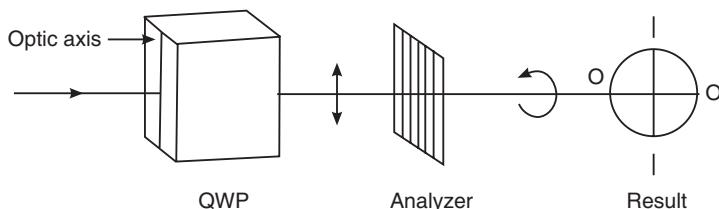


Fig. 8.40: Detection of elliptically polarized light

If the incident light is elliptically polarized, it may be considered as resultant of two coherent plane polarized waves that is e-ray and o-ray, which are out of phase by 90° . If the light passes through the quarter wave plate, an additional phase difference of 90° is introduced between the e-ray and o-ray. Therefore, the total phase difference becomes 180°

between the e-ray and o-ray. On emerging from the quarter plate, the e-and o-rays combine to produce plane-polarized light. If the light coming out of quarter wave plate is examined with an analyzer, light will be extinguished twice in one full rotation of the polarizer as shown in Fig. 8.40. In such a case, the incident light is elliptically polarized.

8.18 PRODUCTION OF CIRCULARLY POLARIZED LIGHT

A quarter wave plate and a polarizer are the optical devices required for producing circularly polarized light from unpolarized light.

Unpolarized light is first converted to plane polarized light by allowing it to pass through a polarizer (a polaroid sheet or a Nicol prism), as shown in Fig. 8.41. Plane polarized light is then made incident on a

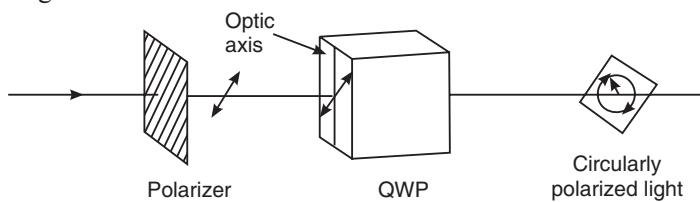


Fig. 8.41: Production of Circularly Polarized Light

quarter wave plate. The polarizer and the quarter wave plate are rotated such that the electric vector E of the plane-polarized wave makes an angle of 45° with the optic axis of the quarter wave plate. The plane polarized wave incident on the quarter wave plate splits into two rays, o-ray and e-ray of equal amplitude ($E_1 \cos 45^\circ = E_2 \sin 45^\circ$). The two rays travel in the same direction inside the crystal but with different velocities. The two rays are in phase at the front face of the crystal but progressively get out of phase as they travel through the crystal. As they emerge from the rear face of the crystal, they will have a path difference of $\lambda/4$ or phase difference of 90° . The two rays are linearly polarized in mutually perpendicular directions. When they combine, they produce circularly polarized light.

8.18.1 Detection of Circularly Polarized Light

The light beam is allowed to pass through an analyzer (a polaroid sheet or a Nicol prism).

If on rotating the analyzing polaroid sheet or Nicol, the intensity of the emerging beam remains uniform, then the incident light is circularly polarized. A similar result would be

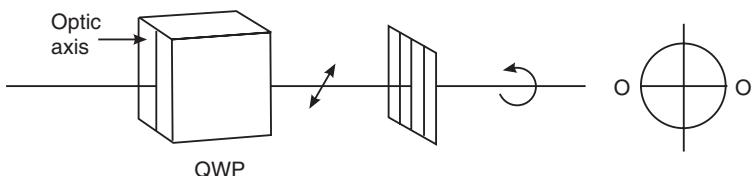


Fig. 8.42: Detection of circularly polarized light

obtained if the incident light is ordinary unpolarized light. The two cases may be distinguished by inserting a quarter wave plate in the path of light before it falls on the analyzer. If the original light is circularly polarized, it may be considered as resultant of two coherent plane-polarized waves, that is e-ray and o-ray, which are out of phase by 90° . If the light passes through the quarter wave plate, an additional phase difference of 90° is introduced between the e-ray and o-ray. Therefore, the total phase difference becomes 180° between the e-ray and o-ray. On emerging from the quarter plate, the e- and o-rays combine to produce plane-polarized light. Therefore, if the light coming out of quarter wave plate is examined with an analyzer, light will be extinguished twice in one full rotation of the polarizer as shown in Fig. 8.42 or otherwise the incident light is unpolarized. In such a case the incident light is circularly polarized.

8.19 ANALYSIS OF POLARIZED LIGHT

In practice light may exhibit any one of the three types of polarization, or may be unpolarized or a mixed type. The unaided eye cannot distinguish the different types of polarization. However, using a polarizer and a quarter wave plate, the actual type of polarization of a light beam can be ascertained. The following steps are used in the analysis of the type of polarization.

The light of unknown polarization is allowed to fall normally on a polarizer. The polarizer is slowly rotated through a full circle and the intensity of the transmitted light is observed. If the intensity of the transmitted light is extinguished twice in one full rotation of the polarizer, then the incident light is *plane polarized*.

(i) If the intensity of the transmitted light varies between a maximum and a minimum value but does not become extinguished in any position of the polarizer, then the incident light is either elliptically polarized or partially polarized (Fig. 8.43b).

(ii) If the intensity of the transmitted light remains constant on rotation of the polarizer, then the incident light is either circularly polarized or unpolarized (Fig. 8.43c).

To distinguish between elliptically polarized and partially polarized or between the circularly polarized and unpolarized light, we take the help of quarter wave plate. The light is first made to be incident on the quarter wave plate and then made to pass through the polarizer.

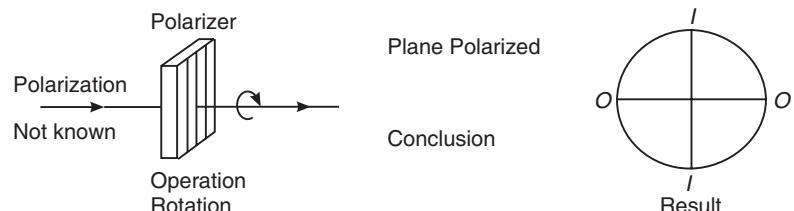


Fig. 8.43 (a)

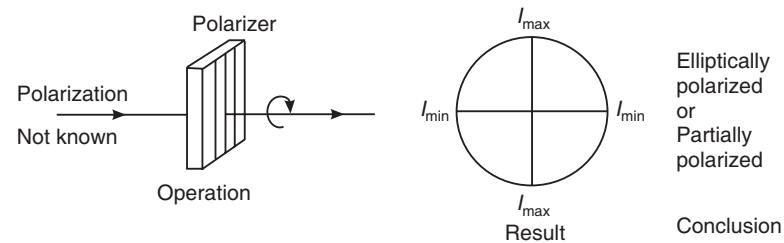


Fig. 8.43 (b)

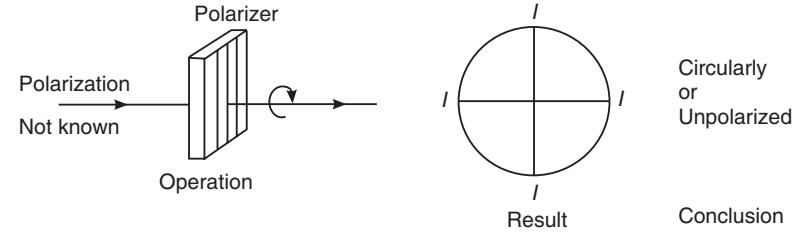


Fig. 8.43 (c)

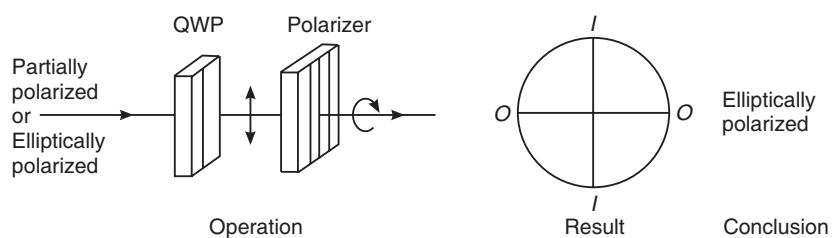


Fig. 8.43 (d)

(iii) If the incident light is elliptically polarized, the quarter wave plate converts it into a plane polarized beam. When this plane polarized light passes through the polarizer, it would be extinguished twice in one full rotation of the polarizer (Fig. 8.43d).

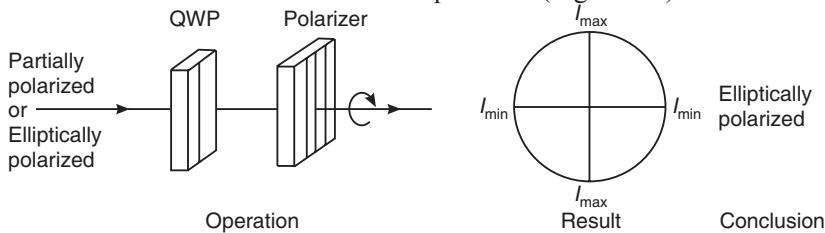


Fig. 8.43 (e)

On the other hand, if the transmitted light intensity varies between a maximum and a minimum without becoming zero, then the incident light is partially polarized (Fig. 8.43e).

(iv) If the incident light is circularly polarized, the quarter wave plate converts it into plane polarized light. When this plane polarized light passes through the polarizer, it would be completely extinguished twice in one full rotation of the polarizer (Fig. 8.43f).

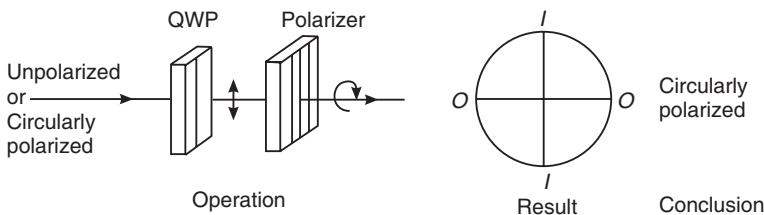


Fig. 8.43 (f)

On the other hand, if the intensity of the transmitted light stays constant, then the incident light is unpolarized (Fig. 8.43g).

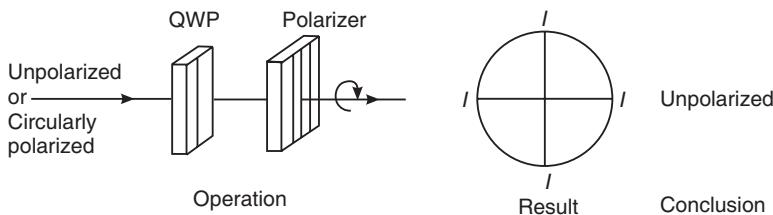


Fig. 8.43 (g)

8.20 APPLICATIONS OF POLARIZED LIGHT

The phenomenon of polarization has many practical applications in daily life. We discuss here some of the interesting applications here.

1. **Sunglasses:** Light reflected from nonmetallic surfaces such as water, snow clad mountains, asphalt roads etc is partially polarized. At angles nearer to the Brewster's angle, the reflected light contains a large concentration of vibrations in a plane parallel to the reflecting surface (Fig. 8.44). Such a highly polarized light causes **glare** in one's eyes and makes it difficult to view the objects. When the amount of glare is large, daily activities such as driving on a road etc would become very difficult to perform.

The phenomenon of polarization is utilized in making sunglasses, which will drastically reduce the glare. Polarized sunglasses contain, over their lenses, polarizing filters that are oriented vertically with respect to the frames (Fig. 8.45). As the reflected light is partially polarized, light waves having their electric field vectors oriented in the same direction as the polarizing lenses (and perpendicular to the reflecting surface) are passed through. On the other hand, light waves having their electric field vectors oriented parallel to the reflecting surface (and perpendicular to the filters in the lenses) are blocked by the lenses. Thus, polarized sunglasses eliminate the glare from an illuminated surface.

2. Photography: Polarization by scattering occurs as light passes through our atmosphere. The scattered light often produces a glare in the skies. In photography, this partial polarization of scattered light produces a washed-out sky. The problem is overcome by the use of a polarizing filter fitted to the camera. As the filter is rotated, the partially polarized light is blocked and the glare is reduced. Thus, a vivid blue sky as the backdrop of a beautiful foreground is captured using polarizing filters.

3. Stereoscopic Movies: The phenomenon of polarization is used in making and viewing stereoscopic movies. Stereoscopic movies are three-dimensional movies, which give the same effect of depth as seen on a stage. The three-dimensional impression is obtained through binocular vision. Three-dimensional movies are actually two movies being shown at the same time through two projectors. For making a stereoscopic movie, two views of the same scene are shot simultaneously from two slightly different camera locations. One view corresponds to that seen by the right eye and the other corresponds to the view seen by the left eye. Each individual movie is then projected from different sides of the audience onto a screen through a polarizing filter. The polarizing filter used for the projector on the left may have its polarization axis aligned horizontally while the polarizing filter used for the projector on the right would have its polarization axis aligned vertically. Consequently, there are two slightly different movies being projected onto a screen; each movie is cast by light, which is polarized with an orientation perpendicular to the other movie. The audience then wears glasses, which have two Polaroid filters. Each filter has a different polarization axis - one is horizontal and the other is vertical. The result of these arrangements of projectors and filters is that the left eye sees the movie, which is projected from the right projector while the right eye sees the movie that is projected from the left projector. This gives the viewer a perception of depth.

4. Optical Microscopy: Polarization of light is also very useful in many aspects of optical microscopy. Microscopes may be configured to use crossed polarizers, in which case the first polarizer, described as the **polarizer**, is placed below the sample in the light path

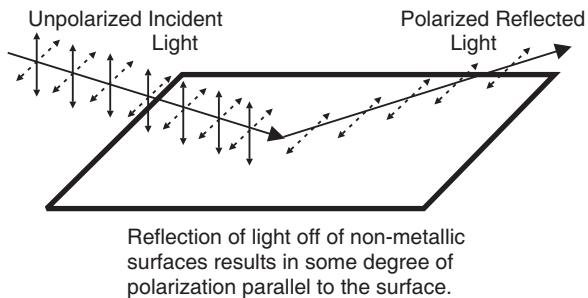


Fig. 8.44

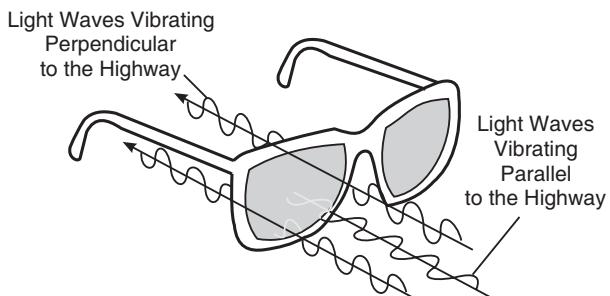


Fig. 8.45

and the second polarizer, known as the **analyzer**, is placed above the sample, between the objective and the eyepieces. If the microscope stage is left empty, the analyzer blocks the light polarized by the polarizer and no light is visible. However, when a **birefringent**, or doubly refracting, sample is placed on the stage between the crossed polarizers, the microscopist can visualize various aspects of the sample. This is because the birefringent sample rotates the light, allowing it to successfully pass through the analyzer.

5. LCDs: Another interesting use of light polarization is the liquid crystal display (**LCD**) utilized in applications such as wristwatches, computer screens, timers, and clocks. These devices are based upon the interaction of rod-like liquid crystalline molecules with an electric field and polarized light waves.

An LCD consists of a liquid crystal material, which is double refracting, of about $10 \mu\text{m}$ thick suitably supported between two thin glass plates having transparent conducting coatings on their inner surfaces (Fig. 8.46a). The conducting coating is etched in the form of a digit or character, as shown in Fig. 8.46(b). The assembly of glass plates with liquid crystal material in between is sandwiched between two crossed-polarizer sheets.

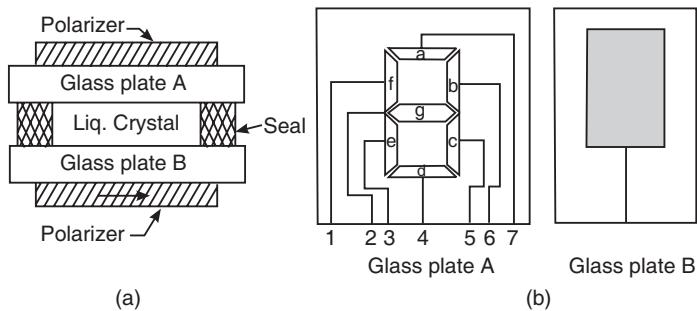


Fig. 8.46

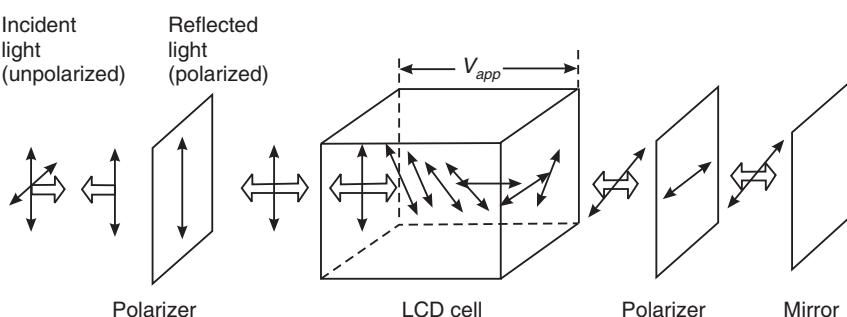


Fig. 8.47

During the fabrication of LCDs, the liquid crystal molecules are aligned in such a way that their long axes undergo a 90° rotation, as illustrated in Fig. 8.47. It is called a twisted molecular arrangement. When natural light is incident on the assembly, the front polarizer converts it into linearly polarized light. As the linearly polarized light propagates through the LCD, the optical vector is rotated through 90° by the twisted molecular arrangement. Therefore, it passes unhindered through the rear polarizer whose transmission axis is perpendicular to that of the front polarizer. A reflecting coating at the back of the rear polarizer sends back the light, which emerges unobstructed by the front polarizer. Consequently, the display appears uniformly illuminated. When a voltage is applied



Fig. 8.48

to the device, the molecules between the electrodes untwist and align along the field direction. As a result, the optical vector does not undergo rotation as it passes through that region. The rear polarizer blocks the light and therefore a dark digit or character is seen in that region, as illustrated in Fig. 8.48.

6. Enhancing visibility of digital displays: Circular polarizers are used to enhance the visibility of digital displays. They can cut out the extraneous light reflected from the face of the display, improving the contrast of the display. They use the fact that mirror reflection changes the handedness of the polarization of the light. Right circular polarization becomes left circular polarization upon reflection. Let us consider a sheet of circular polarizer placed in front of a digital display (Fig. 8.49).

External light that falls on the polarizer becomes circularly polarized before it reaches the front of the display. The reflected light gets the handedness of its polarization reversed, and its return path is blocked by the polarizer. Light generated by the display passes through the polarizer and hence is seen without the background reflecting light.

7. Photoelasticity: Photoelasticity is an experimental method to determine stress distribution in various engineering components. The method is mostly used in cases where mathematical methods become quite cumbersome. Photoelasticity is especially useful for the study of objects with irregular boundaries and stress concentrations, such as pieces of machinery with notches or curves, structural components with slits or holes, and materials with cracks.

Principle: The method is based on the property of double refraction, which is exhibited by photoelastic materials on the application of stress. Double refraction or birefringence is a property by virtue of which a ray of light passing through a birefringent material splits into two beams (e- and o-rays). The two beams travel along the same path in the material and their speed at each point in the material is directly related to the state of stress at that point. Because the velocities of light propagation are different in each direction, there occurs a phase shifting of the light waves. Therefore, light emerges out of the component as two beams vibrating out of phase with one another and when they are combined, produce interference pattern.

The stressed component is examined under monochromatic polarized light in a polariscope. The polarizer in the polariscope produces polarized light. When the analyzer in the polariscope recombines the waves, interference pattern is observed. Regions of stress where the wave phases cancel appear dark, and regions of stress where the wave phases add appear bright. Therefore, in models of complex stress distribution, bright and dark fringe patterns (isochromatic fringes) are projected from the model. As these fringes are related to the stresses, the magnitude and direction of stresses at any point can be determined by examination of the fringe pattern. When the component is unloaded, the photoelastic fringe pattern disappears.

When white light is used in place of monochromatic light, coloured fringes are observed. White light is often used for demonstration, and monochromatic light is used for precise measurements.

The above method is suitable when the component is transparent. In the case of opaque components, a thin sheet of photoelastic material is suitably bonded to the surface of the

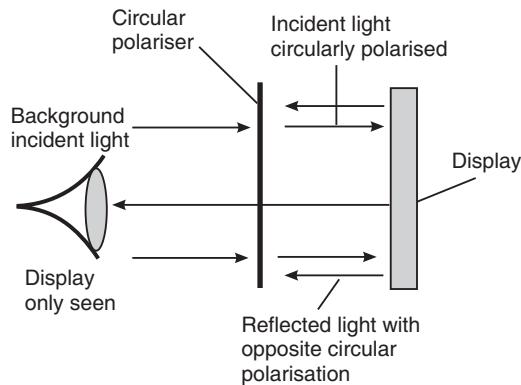


Fig. 8.49

component. When the component is loaded, the surface strain in the component is transmitted to the photoelastic sheet producing stress in it. The resulting fringe pattern is observed by illuminating the component with polarized light and viewing it through an analyzer. More commonly, a transparent scale model of the part is made out of a material, which is optically sensitive to stress such as epoxy, glyptol or polyester resins. The model is then subjected to the forces that the actual component would experience in use. The birefringence varies from point to point over the surface of the model. When viewed with crossed polarizers, a complicated fringe pattern is seen which provides a visual means of observing overall stress characteristics of an object. The patterns can be projected on a screen or photographic film.

QUESTIONS

1. Define and explain polarization. (Amaravati Univ., 2006)
2. What do you understand by polarization of light? Distinguish between polarized and unpolarized light. (Amaravati Univ., 2003, 2007)
3. Define the terms plane of vibration and plane of polarization. (Anna Univ., 2004)
4. Discuss the methods by which plane polarized light can be produced.
5. Describe the process of production of plane polarized light by reflection. (Amaravati Univ., 2006)
6. State Brewster's law and give its significance. (Amaravati Univ., 2002, 2003)
7. State Brewster's law. How can this law be used to produce plane-polarized light?
8. State and explain Brewster's law. (Amaravati Univ., 2007, 2008)
9. What is Brewster's law? Give any two applications of it. (Amaravati Univ., 2005)
10. Show that when light is incident on a transparent material at the Brewster angle, the reflected and refracted rays are at right angles.
11. Discuss some of the applications of Brewster's law.
12. State Brewster's law of polarization. Write mathematical expression for it. (Univ. of Pune, 2007)
13. How can the refractive index of a smooth opaque dielectric material be determined?
14. What is polarization? State the law of Malus. (Amaravati Univ., 2002)
15. State and explain the law of Malus.
16. Describe the fabrication of a Polaroid.
17. How can plane-polarized light be detected?
18. Unpolarized light falls on two polarizing sheets so oriented that **no** light is transmitted from the combination. If a third polarizing sheet is placed between them, can light be transmitted. Explain.
19. What is dichroism?
20. Explain polarization by double refraction. (Amaravati Univ., 2004, 2005)
21. Explain o-ray and e-ray.
22. What is an optic axis?
23. Explain the terms: (i) Double refraction (ii) optic axis (iii) positive and negative crystals.
24. Distinguish between:
 - (i) Ordinary ray and extraordinary ray.
 - (ii) Positive and negative crystals
 - (iii) Uniaxial and biaxial crystals
25. (a) Describe a Nicol prism and explain how it acts as an analyzer.
 (b) Explain the phenomenon of double refraction in uniaxial crystals. (Calicut Univ., 2006)
26. What are the differences between positive and negative crystals? (Calicut Univ., 2006)
27. Explain the phenomenon of double refraction on the basis of Huygen's wave theory of light. (Univ. of Pune, 2007)
28. Explain the principle, construction and working of a Nicol prism with a neat diagram. (Univ. of Pune, 2007)
29. Give the construction and working of Nicol prism. (Amaravati Univ., 2002, 2006)

52. Distinguish between (i) circularly polarized light and unpolarized light and (ii) elliptically polarized and partially polarized light. **(Anna Univ., 2007)**
53. Explain production of plane polarized and circularly polarized light. **(Amaravati Univ., 2008)**
54. Explain how you will distinguish between unpolarized light and circularly polarized light. **(Univ. of Pune, 2007)**
55. Explain how the analysis of the type of polarization of light beam can be ascertained using a polarizer and a quarter plate. **(Amaravati Univ., 2006)**
56. Explain production of elliptically and circularly polarized light. **(Amaravati Univ., 2003, 2004)**
57. Explain how circularly polarized and elliptically polarized light are produced and detected. **(Calicut Univ., 2006)**
58. Explain some of the applications of polarized light.

PROBLEMS

- Consider a positive crystal with refractive indices for e-ray 1.31 and for o-ray 1.309. What should be the minimum thickness of that crystal so that it can act as a quarter wave plate for light of wavelength 6000 Å? **(RTMNU S -05)**
- Ice is a positive crystal with indices of refraction of 1.309 and 1.310. What should be the minimum thickness of ice so that it can act as a quarter wave plate for light of wavelength 6000Å? **(RTMNU W -04)**
- Light of intensity I_0 is incident on a polarizer. What is the intensity of the resultant beam if:
 - Incident light is unpolarized?
 - Incident light is plane polarized with its electric field making an angle of 30° with the axis of the polarizer?**(RTMNU, S - 03)**
- Find the thickness of a quarter wave plane that can convert plane polarized light into elliptically polarized light. Use the following data: $\lambda = 5890\text{Å}$, $\mu_0 = 1.658$ and $\mu_e = 1.486$. **(RTMNU W -03)**
- Calculate the thickness of
 - Quarter wave plate and
 - Half wave plate.
Given that $\mu_e = 1.553$, $\mu_o = 1.544$, $\lambda = 5000 \text{ \AA}$ **(RTMNU W-05)**
- Calculate the thickness of double refracting plate capable of producing a path difference of $\lambda/4$ between extraordinary and ordinary waves.
Given $\lambda = 5890\text{\AA}$, $\mu_0 = 1.530$, $\mu_e = 1.540$ **(RTMNU, S - 03)**
- Calculate the thickness of a plate which would convert plane polarized light into circularly polarized light. Given: $\lambda = 5890\text{\AA}$, $\mu_0 = 1.658$, $\mu_e = 1.486$. **(RTMNU S-07)**
- A half wave plate is designed from a crystal for $\lambda = 600 \text{ nm}$. If $(\mu_0 - \mu_e) = 0.0057$, calculate the thickness of the plate. **[Ans: 52 μm]**
- Calculate the thickness of a mica sheet required for making a quarter wave plate for $\lambda = 500 \text{ nm}$. The refractive indices for o-ray and e-ray in mica are 1.586 and 1.592. **[Ans: 20 μm]**
- Two Nicol prisms are so arranged that the amount of light transmitted through them is maximum. What will be the percentage reduction in intensity of the transmitted light when the analyzer is rotated through (a) 30° (b) 90° ? **[Ans: 25%, 100%]**
- Plane polarized light of $\lambda = 500 \text{ nm}$ is incident on a quartz crystal parallel to the optic axis. Find the least thickness for which the o-ray and e-rays combine to form plane-polarized light. Given: $\mu_o = 1.5442$, $\mu_e = 1.5533$. **[Ans: 27.5 μm]**

CHAPTER

9

Optical Activity

9.1 INTRODUCTION

Certain crystals and solutions possess a natural ability to rotate the plane of polarization about the direction of propagation. It is known as **optical activity**. In case of crystals this ability arises due to the twisted arrangement of atomic layers with respect to one another. In liquids and solutions the optical activity is due to certain structural asymmetry in their molecules. The optical activity found in bigger organic molecules provides a number of clues which help us understand biological activity.

Many crystalline materials exhibit birefringence naturally, without application of any voltage. The birefringence is present all the time. Examples of such crystals are quartz and calcite. There are also a number of crystals that are not birefringent naturally but in which application of a voltage or magnetic field induces birefringence. The induced optical activity leads to the ability to control light beams in a variety of ways and is the basis of a number of applications such as light-beam modulators, Q-switches, and deflectors.

9.2 OPTICAL ROTATION

When a beam of plane polarized light propagates through a quartz crystal along the optic axis, the plane of polarization steadily turns about the direction of the beam. The optical rotation can be detected as follows. If two polaroid sheets or Nicol prisms are held in crossed configuration and if a beam of unpolarized light is viewed through them, the field of view appears to be completely dark. Now let a quartz crystal, cut with its faces perpendicular to the optic axis, be inserted between the polarizers such that light is incident normally on the crystal. The field of view now appears lit up indicating that the light is not cut off by the analyzer. In order to cut off the transmitted light, we find that the analyzer is to be rotated through a certain angle. The experiment establishes that the plane polarized light produced by the polarizer remains plane polarized while passing through the quartz crystal but the plane of polarization is rotated through an angle. This angle is the angle through which the analyzer is rotated in order to cut off the light totally. The optical rotation, i.e., rotation of the plane of polarized light is shown in Fig. 9.1.

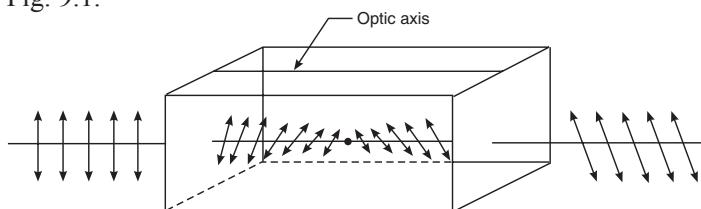


Fig. 9.1

The ability to rotate the plane of polarization of plane polarized light by certain substance is called **optical activity**. Substances, which have the ability to rotate the plane of the polarized light passing through them, are called **optically active** substances. Quartz and cinnabar are examples of optically active crystals while aqueous solutions of sugar, tartaric acid are optically active solutions.

Optically active substances are classified into two types.

- (i) **Dextrorotatory substances:** Substances which rotate the plane of polarization of the light toward the right are known as right-handed or dextrorotatory.
- (ii) **Laevorotatory substances:** Substances which rotate the plane of polarization of the light toward the left are known as left-handed or laevorotatory.

9.3 SPECIFIC ROTATION

A measure of the optically activity of a sample is the rotation produced for a 1 mm slab for a solid, or a 100 mm path length for a liquid. This measure is called the *specific rotation*. Liquids usually rotate the light much less than solids. Solutions of solids will obviously show an effect that depends on the concentration of active material and, to a small extent, both on temperature and the solvent.

If an optically active material is kept between two crossed polarizers, the field of view becomes bright. In order to get darkness once again, the analyzer has to be rotated through an angle. The angle through which the analyzer is rotated equals the angle through which the plane of polarization is rotated by the optically active substance. This angle depends on

- (a) The thickness of the substance,
- (b) Density of the material or concentration of the solution,
- (c) Wavelength of light, and
- (d) The temperature.

The amount of rotation θ caused by crystalline materials is given by

$$\theta = \alpha l \quad (9.1)$$

where α is called the **rotational constant**.

In solutions the amount of rotation θ is given by

$$\theta = s c l \quad (9.2)$$

where c is the concentration and s is called the **specific rotation**.

The specific rotation for a given wavelength of light at a given temperature is defined conventionally as the rotation produced by one decimetre long column of the solution containing 1 gm of optically active material per c.c. of solution.

$$[s]_{\lambda}^t = \frac{\theta}{l \times C} = \frac{\text{Rotation in degrees}}{\text{Length in decimetres} \times \text{conc. in gm/c.c.}} \quad (9.3)$$

9.4 FRESNEL'S EXPLANATION

A linearly polarised light can be considered as a resultant of two circularly polarised vibrations rotating in opposite directions with the same angular velocity. Fresnel assumed that plane-polarised light on entering a crystal along the optic axis is resolved into two circularly polarised vibrations rotating in opposite directions with the same angular frequency. In an optically inactive crystal like calcite, the two circularly polarised vibrations travel with the same angular velocity. On the other hand, in an optically active crystal like quartz, the two circularly polarised vibrations travel with the different angular velocities.

Fig. 9.2 (a) shows plane polarized light entering a calcite crystal along the optic axis AB and split up into two circular motions rotating in opposite directions. They are represented by OL and OR. OL is the circularly polarised vector rotating in the anticlockwise direction and

OR is the circularly polarised vector rotating in the clockwise direction. If OL and OR start simultaneously from OA and rotate with the same angular velocity, then at any subsequent time, the resultant of OL and OR will lie along OA . Hence on emerging from the crystal, the two circular waves combine to produce a linear vibration along the initial direction, OA (Fig. 9.2 b). Therefore, crystals like calcite do not rotate the plane of vibration.

In case of quartz, the linearly polarised light, the component having clockwise rotation travels faster than the anticlockwise component. When the components emerge out of the crystal, they are at an angle 2θ . The resultant of these two vectors OR and OL is now along OD (Fig. 9.2 c). Before entering the crystal, the plane of vibration is along OA

and after emerging from the crystal, it is along OD which makes an angle θ with the optic axis OA . After passing through the crystal, the circular components combine to produce a linear vibration whose direction is now along OD . Therefore, the plane of vibration has rotated through an angle θ . The angle through which the plane of vibration is rotated depends on the thickness of the crystal.

Left and right circularly polarised light travel through the crystal at different speeds. Because one travels more slowly than the other, a phase shift builds up between them. When right and left-handed light are combined together at any point, the result is always linear polarization; but the angle of the polarisation depends on the phase shift *between* the two circular polarisations. If μ_R and μ_L are the refractive indices corresponding to the right and left circularly polarized light, the angle of polarisation, θ , is given by:

$$\theta = \frac{\pi}{\lambda} (\mu_R - \mu_L) d \quad (9.4)$$

where d is the thickness of the material. As a result, the direction of polarisation spirals around x in space.

The cause of this special behaviour for circularly polarised light is that the optically active molecules are chiral, meaning that they have a helical twist in them. Any arrangement of atoms with a helical structure can form left-handed helices or right-handed helices. One is a mirror image of the other. Now there is a curious property of helices not possessed by rotating circles. If we look at a wheel rotating clockwise from one side, the same wheel appears to go around anticlockwise from the other side. The ‘wise’ description just depends on our viewpoint. Helices are different. If we turn round a right-handed helix (in which the spiral appears to go around clockwise as we look into the helix), then it is still a right-handed helix. So, helical molecules may interact with circularly polarised light differently, depending on whether their chirality is the same handedness as the circular polarisation or opposite handedness. This is the fundamental reason for the two refractive indices μ_R and μ_L . It is interesting that the cause of the phenomenon lies at a molecular level and does not have anything to do with spatial ordering of molecules. This is why the phenomenon occurs with liquids as well as solids.

9.5 POLARIMETER

A **polarimeter** is an instrument used for determining the optical rotation of solutions. When used for determining the optical rotation of sugar it is called a **saccharimeter**.

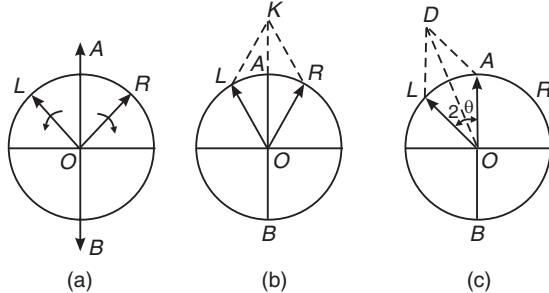


Fig. 9.2

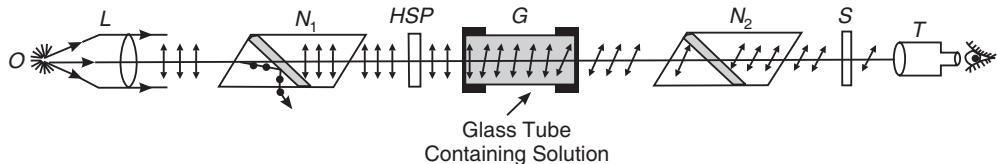


Fig. 9.3

Construction: A polarimeter consists of a glass tube for holding the solution under test held between crossed Nicol prisms. Beyond the polarizing Nicol prism a half-shade plate is located which is used for accurately adjusting the two Nicols for crossed position. Light from a monochromatic source 'O' is rendered parallel by the lens L and is incident on the polarizer, N₁. The light transmitted by the polarizer is plane polarized. The polarized beam then passes through the half-shade plate HSP and a glass tube G containing the solution. The light emerging from the solution will be incident on the analyzer N₂. The light is observed through a telescope T. The analyzing Nicol N₂ can be rotated about the axis of the tube and the rotation can be measured with the help of a graduated circular scale.

Working: To find the specific rotation of a solution, the analyzer is first adjusted such that field of view is completely dark. Then the glass tube is filled with the solution and is held in position. The field of view now becomes illuminated. The field of view can again be made dark by rotating the analyzer through a certain angle which gives the optical rotation of the solution. The practical difficulty in this method is in determination of the exact position for which complete darkness is achieved. The difficulty is overcome by using what is known as a Laurent's half-shade device.

It consists of a semicircular half wave plate ACB of quartz cemented to a semicircular plate ADB of glass. The optic axis of the wave plate is parallel to the line of separation AB. The half wave plate introduces a phase difference of 180° between e-ray and o-ray passing through it. The thickness of the glass plate is such that it transmits the same amount of light as done by the quartz half wave plate. One half of the incident light passes through the quartz plate ACB and the other half through the glass plate ADB. When the light after passing through the polarizer is incident normally on the half shade plate and has vibrations along OP. On passing through the glass, half the vibrations will remain along OP but on passing through the quartz half, the vibrations will split into e- and o-rays. The o-vibrations are along OD and e-vibrations are along OA. The half wave plate introduces a phase difference of π rad between the two vibrations. The vibrations of o-ray will occur along OC instead of OD on emerging from the plate. Therefore the resultant vibration will be along OQ whereas the vibrations of the beam emerging from glass plate will be along OP. In effect, the half wave plate turns the plane of polarization of the incident light through an angle 2θ .

If the principal plane of the Nicol N₂ is aligned parallel to OP, the plane polarized light emerging from the glass tube will pass through the glass plate of the half shade plate and that part appears brighter. On the other hand light coming out of the quartz plate is partially obstructed and the corresponding field of view appears less bright. If the principal plane of N₂ is parallel to OQ the quartz half will appear brighter than the glass half. Thus, the two halves

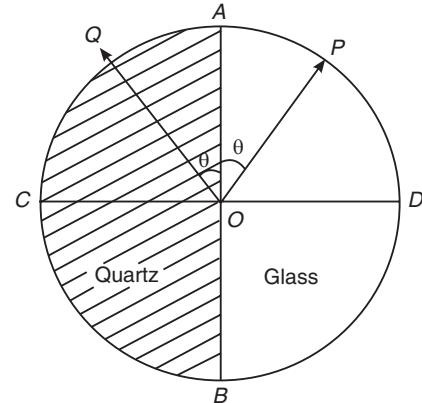


Fig. 9.4: Half shade plate

of the plate are unequally illuminated. When the principal plane of N_2 is parallel to AB, the two halves appear equally bright and when it is parallel to CD, the two halves are equally dark.

To find the specific rotation of a solution, the analyzer is first set in the position for equal darkness without solution in the tube G. The reading on the circular scale is noted. Next, the tube is filled with the optically active solution of known concentration. The field of view is now partially illuminated. The analyzer is rotated till the field of view becomes equally dark. The reading on the circular scale is noted again. The difference between the two scale readings gives the angle of rotation of the plane of polarization caused by the solution. Knowing the values of θ , l , and c , the specific rotation is obtained using the formula (9.3). Or otherwise, knowing the value of the specific rotation, the concentration of the solution can be determined with the help of the equation (9.3).

Example 9.1: The rotation in the plane of polarization in a certain substance is $10^\circ/\text{cm}$. Calculate the difference between the refractive indices for right and left circularly polarized light in the substance. Given $\lambda = 5893 \text{ \AA}$.

$$\text{Solution: } \theta = \frac{\pi}{\lambda} [\mu_R - \mu_L] d \quad \therefore [\mu_R - \mu_L] = \frac{\theta \lambda}{\pi d}$$

$$\text{It is given that } \frac{\theta}{d} = 10^\circ = \frac{10 \times \pi}{360^\circ} = \frac{\pi}{36} \text{ radian/cm} \quad \text{and} \quad \lambda = 5893 \text{ \AA} = 5893 \times 10^{-8} \text{ cm.}$$

$$\therefore \mu_R - \mu_L = \frac{\pi}{36} \cdot \frac{5893 \times 10^{-8} \text{ cm}}{\pi} = 1.6 \times 10^{-6}.$$

Example 9.2: The indices of refraction of quartz for right-handed and left-handed circularly polarized waves of wavelength 7620 \AA travelling in the direction of optic axis have the following values.

$$\mu_R = 1.53914 \quad \text{and} \quad \mu_L = 1.53920$$

Calculate the rotation of the plane of polarization of light in degrees produced by a plate of 0.5 mm thick.

$$\text{Solution: } \theta = \frac{\pi}{\lambda} [\mu_R - \mu_L] d = \frac{3.14 \times 0.5 \times 10^{-3} \text{ m}}{7620 \times 10^{-10} \text{ m}} (1.53920 - 1.53914)$$

$$= 0.1236 \text{ radian} = \frac{0.1236 \times 180^\circ}{\pi} = \frac{0.1236 \times 180^\circ}{3.14} = 705'.$$

Example 9.3: A 200 mm long tube containing 48 cm^3 of sugar solution produces an optical rotation of 11° when placed in a saccharimeter. If the specific rotation of sugar solution is 66° , calculate the quantity of sugar contained in the tube in the form of a solution.

Solution: It is given that $\theta = 11^\circ$, $l = 200 \text{ mm} = 20 \text{ cm}$, $S = 66^\circ$, and $V = 48 \text{ cm}^3$.

$$C = \frac{10 \theta}{IS} = \frac{10 \times 11^\circ}{20 \text{ cm} \times 66^\circ} = 0.0833 \text{ g/cm}^3$$

Mass of sugar in solution $M = CV = 0.0833 \text{ g/cm}^3 \times 48 \text{ cm}^3 = 4 \text{ grams}$.

Example 9.4: A 20 cm long tube containing sugar solution is placed between crossed Nicols and illuminated by light of wavelength of 6000 \AA . If the specific rotation is 60° and optical rotation is 12° , what is the strength of the solution?

Solution: The specific rotation at a given temperature and for a given wavelength is given by

$$C = \frac{10\theta}{ls} = \frac{10 \times 12^\circ}{20\text{cm} \times 60^\circ} = 0.1$$

Therefore, it is 10% solution of sugar, i.e., 1 gm sugar is dissolved in 10 gm of water.

Example 9.5: 20 cm length of a certain optically active solution causes right-handed rotation of 40° and 30 cm of another solution causes left-handed rotation of 24° . What will be the optical rotation produced by 30 cm length of the mixture of the above solutions in volume ratio 1:2. It is given that the solutions do not react chemically.

Solution: As the length of the mixtures is 30 cm and the solutions are in the volume ratio 1:2, we may assume that 10 cm length is of the first solution and 20 cm length is of the second solution.

The optical rotation produced by the first solution

$$= 40^\circ \times \frac{10\text{cm}}{20\text{cm}} = 20^\circ \text{ (right-handed)} = -20^\circ.$$

The optical rotation produced by the second solution

$$= 24^\circ \times \frac{20\text{cm}}{30\text{cm}} = 16^\circ \text{ (left-handed)} = +16^\circ$$

∴ Total optical rotation = $-20^\circ + 16^\circ = -4^\circ$.

∴ The resultant optical rotation is **4° right-handed**.

9.6 ELECTRO-OPTIC AND MAGNETO-OPTIC EFFECTS

Isotropic transparent materials such as glass do not exhibit double refraction under ordinary circumstances. However, they acquire the optical properties of a uniaxial crystal under the action of external forces. Consequently, they exhibit double refraction. The appearance of double refraction under the influence of an external agent is known as **artificial double refraction or induced birefringence**. The direction of the optical axis in such materials will be collinear with the direction of the external force. The action of the external force is to cause distortion of the molecular arrangement within the material and thereby transform the isotropic substance into an *anisotropic* substance. The induced birefringence disappears as soon as the external force ceases to act.

The materials which experience a change in their optical behaviour under the action of an electric field are called electro-optic materials and the resulting optical effects are known as **electro-optic effects**. Similarly, the materials that get influenced by a magnetic field are called magneto-optic materials and the resulting optic effects are known as **magneto-optic effects**. The electro-optic and magneto-optic materials play a very important role in modern technology.

9.7 ELECTRO-OPTIC EFFECTS

9.7.1 Kerr Effect

Optical anisotropy induced in an isotropic liquid under the influence of an electric field is known as the Kerr effect. John Kerr discovered it in 1875.

A Kerr cell is required for studying the effect. It consists of a sealed glass cell filled with a liquid comprising of asymmetric molecules. Two plane electrodes of specific length l are arranged in it with their faces strictly parallel to each other. When a voltage is applied to them a uniform electric field is produced in the cell. The Kerr cell is placed between a crossed polarizer system. When the electric field is applied, the molecules of the liquid tend to align

along the field direction. As the molecules are asymmetric, the alignment causes anisotropy and the liquid becomes double refracting. The induced birefringence is proportional to the square of the applied electric field and to the wavelength of incident light. Thus,

$$\Delta\mu = K \lambda E^2 \quad (9.5)$$

where K is known as the Kerr constant.

Among the liquids, nitrobenzene ($C_6H_5NO_2$) is found to have the highest value for the Kerr constant. Therefore, Kerr cells use nitrobenzene.

Kerr cell is used as an electro-optic shutter in high-speed photography, as a light chopper in the measurement of the speed of light.

9.7.2 Pockels Effect

F. Pockels discovered in 1893 that the application of an electric field to piezoelectric crystals makes them birefringent. Normally, piezoelectric crystals are birefringent but in certain directions do not exhibit double refraction. When an electric field is impressed along these directions, double refraction is induced along these directions also.

A Pockels cell consists of a piezoelectric crystal, for example lithium niobate placed between crossed polarisers. Transparent electrodes (thin conducting coatings of tin oxide or indium) are deposited on opposite sides of the crystal. The crystal is oriented in such a way that its optic axis lies along the direction of the electric field applied between the electrodes. The transparent electrodes ensure free propagation of light through the crystal. A Pockels set up is shown in Fig. 9.6.

The birefringence induced in the crystal is proportional to the strength of the applied field. Thus,

$$\Delta\mu = kE \quad (9.6)$$

where k is a constant characteristic of the material. Equ.(9.6) shows that Pockels effect is a linear effect.

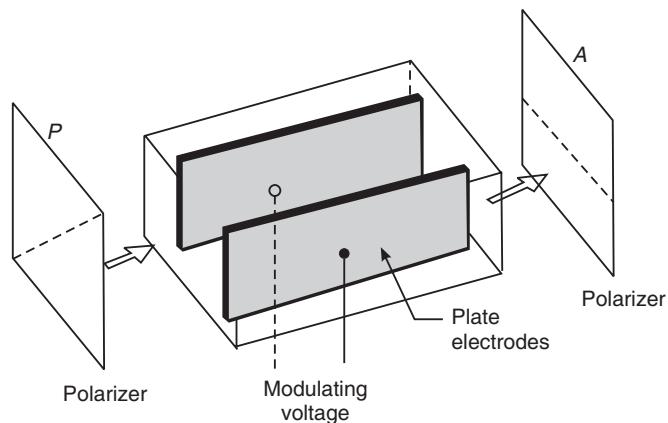


Fig. 9.5. Kerr cell

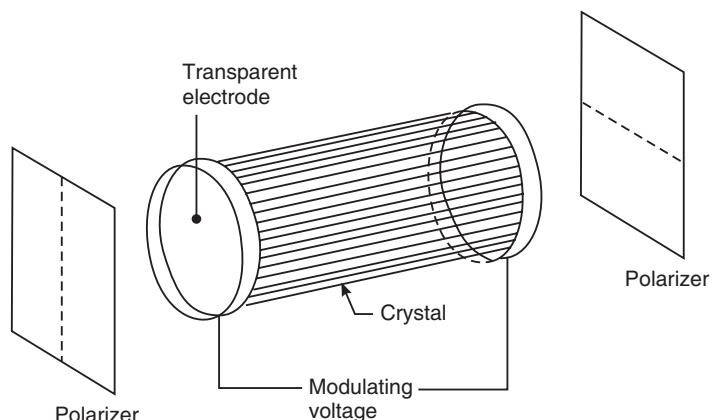


Fig. 9.6. Pockels cell

The total birefringence of the cell is initially made equal to $\lambda/2$. When the electric field is increased, the beam is transmitted or hindered depending on the phase difference between the o-ray and e-ray. The device switches on and off periodically. Pockels cells are used in fast switching applications and in fibre optics. It can be used to obtain amplitude, frequency or phase modulation.

A Pockels cell is simple in construction and requires a small voltage of the order of 1.5 kV whereas a Kerr cell is complicated in construction and requires higher voltages of the order of 15 kV. The piezoelectric crystals of ammonium dihydrophosphate (ADP) and potassium dihydrophosphate (KDP) are widely used in Pockels cell.

Kerr and Pockels cells are widely used as electro-optic shutters in Q-switching of lasers.

9.8 MAGNETO-OPTIC EFFECTS

9.8.1 Cotton-Mouton Effect

The Cotton-Mouton effect is a magneto-optic effect. An isotropic material acquires the optical behaviour of a uniaxial crystal under the action of an external magnetic field. The set up is shown in Fig. 9.7.

The induced birefringence is governed by the relation

$$\Delta\mu = C \lambda B^2 \quad (9.7)$$

where C is a constant characteristic of the material. The magnitude of the induced birefringence is usually very small.

9.8.2 Faraday Effect

Optically inactive substances acquire the ability of rotating the plane of polarisation of light when they are subjected to a magnetic field, parallel to propagation direction. Michael Faraday discovered this effect and hence it is called Faraday effect. This effect occurs in most optically transparent dielectric materials (including liquids) when they are subjected to strong magnetic fields. The set up for observing Faraday effect is shown in Fig. 9.8.

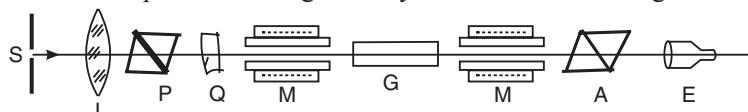


Fig. 9.8

The angle of rotation θ of the plane of polarisation is proportional to the length of the path of light in the material and to the strength of the applied magnetic field. Thus,

$$\theta = VlH \quad (9.8)$$

where V is known as **Verdet constant**.

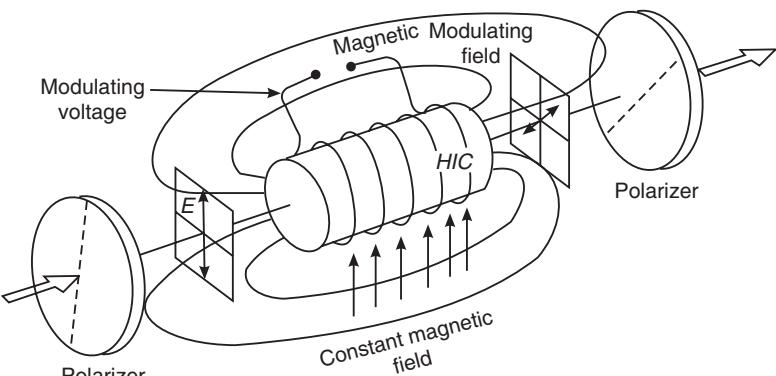


Fig. 9.7

The angle θ of rotation is not very large. For magnetic field strengths of the order of 10^6 A/m and $l = 0.1 \text{ m}$, θ is about 1° to 2° .

One of the interesting problems encountered in satellite communications is the Faraday effect. As radio waves pass through the ionosphere, their plane of polarisation is rotated by the ionised particles in conjunction with the Earth's magnetic field. A horizontally polarised wave becomes vertically polarised because of the Faraday rotation in the ionosphere. This problem is solved by using an antenna with circular polarisation, which ensures that the waves are received satisfactorily, no matter how they have been rotated.

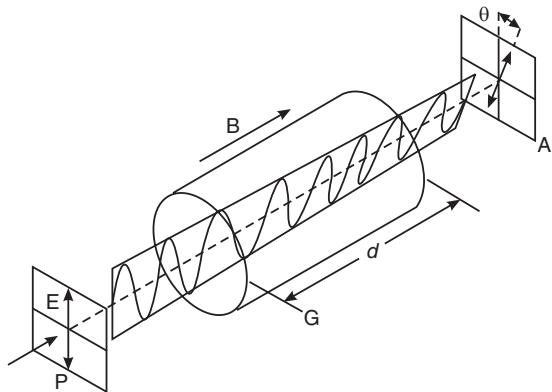


Fig. 9.9

9.9 ANISOTROPY INDUCED BY MECHANICAL STRAIN

Materials such as glass become double refracting when they are subjected to mechanical strain. Experiments show that the induced birefringence is directly proportional to the stress σ experienced at a given point of the material. Thus,

$$\Delta\mu = k\sigma \quad (9.9)$$

where k is the proportionality constant characteristic of the material. This effect was discovered by Sir David Brewster in 1816.

When the material is held between crossed polarisers P and A, as in Fig. 9.10, the field of view appears dark as long as the external force is not applied. As soon as the force is applied, coloured contours will be seen. Dark regions indicate the absence of strain in those areas. Each coloured contour shows the areas that are identically deformed. Such contours enable us assess the distribution of the stresses in the material.

The mechanically induced birefringence is used to study stresses in girders, beams etc. The model of the object under investigation is made of transparent plastic material and is then loaded. Using crossed polarizer system, the stresses produced at different positions are analysed and estimated. This method of analysis is known as photo-elastic analysis and is widely employed in civil and mechanical engineering practices.

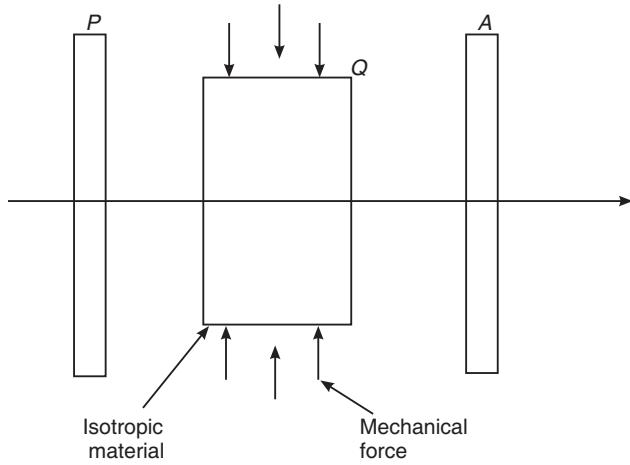


Fig. 9.10

9.10 PHOTOELASTICITY

Photoelasticity is an experimental method to determine stress distribution in various engineering components. The method is mostly used in cases where mathematical methods become quite cumbersome. Photoelasticity is especially useful for the study of objects with

irregular boundaries and stress concentrations, such as pieces of machinery with notches or curves, structural components with slits or holes, and materials with cracks.

Principle: The method is based on the property of double refraction, which is exhibited by photoelastic materials on the application of stress. Double refraction or birefringence is a property by virtue of which a ray of light passing through a birefringent material splits into two beams (e- and o-rays). The two beams travel along the same path in the material and their speed at each point in the material is directly related to the state of stress at that point. Because the velocities of light propagation are different in each direction, there occurs a phase shifting of the light waves. Therefore, light emerges out of the component as two beams vibrating out of phase with one another and when they are combined, produce interference pattern.

The stressed component is examined under monochromatic polarized light in a polariscope. The polarizer in the polariscope produces polarized light. When the analyzer in the polariscope recombines the waves, interference pattern is observed. Regions of stress where the wave phases cancel appear dark, and regions of stress where the wave phases add appear bright. Therefore, in models of complex stress distribution, bright and dark fringe patterns (isochromatic fringes) are projected from the model. As these fringes are related to the stresses, the magnitude and direction of stresses at any point can be determined by examination of the fringe pattern. When the component is unloaded, the photoelastic fringe pattern disappears.

When white light is used in place of monochromatic light, coloured fringes are observed. White light is often used for demonstration, and monochromatic light is used for precise measurements.

The above method is suitable when the component is transparent. In the case of opaque components, a thin sheet of photoelastic material is suitably bonded to the surface of the component. When the component is loaded, the surface strain in the component is transmitted to the photoelastic sheet producing stress in it. The resulting fringe pattern is observed by illuminating the component with polarized light and viewing it through an analyzer. More commonly, a transparent scale model of the part is made out of a material, which is optically sensitive to stress such as epoxy, glyptol or polyester resins. The model is then subjected to the forces that the actual component would experience in use. The birefringence varies from point to point over the surface of the model. When viewed with crossed polarizers, a complicated fringe pattern is seen which provides a visual means of observing overall stress characteristics of an object. The patterns can be projected on a screen or photographic film.

9.10.1 Stress-Optic Law

At any point in a loaded component there is stress acting in every direction. The directions in which the stresses have maximum and minimum value for the point are known as *principal directions*. The corresponding stresses are known as maximum and minimum *principal stresses*. Let us consider a model of uniform thickness made of a transparent high polymer material. Let the model be loaded such that it is in a plane state of stress (see Fig. 9.11). Then the state of stress can be characterized by σ_x , σ_y and τ_{xy} or by the principal stresses σ_1 , σ_2 and their orientation with respect to a set of axes. Let n_0 be the refractive index of the material when it was not stressed. When the model is put in a stress, the model becomes double refracting and the directions of polarization of light at the point P coincide with the direction of principal stress axis at that point. If n_1 and n_2 are the refractive indices for vibration corresponding to these two directions, then

$$n_1 - n_0 = c_1 \sigma_1 + c_2 \sigma_2 \quad (9.10)$$

$$n_2 - n_0 = c_1 \sigma_2 + c_2 \sigma_1 \quad (9.11)$$

where c_1 is called the *direct stress optic coefficient* and c_2 the *transverse stress optic coefficient*. Since the stresses vary uniformly, σ_1 , σ_2 and θ are continuously distributed functions over the model in the xy -plane. The directions of the polarizing axes as well as the values of n_1 and n_2 vary uniformly over the xy -face of the model.

If linearly polarized light is incident normally at any point P of the model, the incident light gets resolved along σ_1 and σ_2 ; and these two vibrating components travel through the thickness of the model with different velocities. When they emerge there will be a certain amount of relative phase difference between these two components. The phase difference is given by

$$\delta = \frac{2\pi d}{\lambda} (n_1 - n_2) \quad (9.12)$$

$$\delta = \frac{2\pi d}{\lambda} (c_2 - c_1)(\sigma_1 - \sigma_2)$$

$$\delta = \frac{2\pi d}{\lambda} C(\sigma_1 - \sigma_2) \quad (9.13)$$

where $C = c_2 - c_1$ is the relative or **differential stress-optic coefficient** expressed in terms of brewsters ($1 \text{ brewster} = 10^{-12} \text{ m}^2/\text{N}$).

Equ.(9.13) shows that in a transparent and isotropic model in which the stresses are two-dimensional, the phase difference between the two wave components traveling through the model is directly proportional to the difference of the principal stresses.

When the two wave components are brought together, interference takes place and we get a fringe pattern, which depends on relative retardation, given by equ.(9.13). Thus, studying the fringe pattern one can determine the state of stress at various points in the material.

The number of wavelengths of relative path difference is given by

$$N = \frac{\delta}{2\pi} = \frac{d}{\lambda} C(\sigma_1 - \sigma_2) \quad (9.14)$$

Equs.(9.13) and (9.14) are called stress-optic relations or **stress-optic law**. These equations relate the state of stress at a point to the optical behaviour of the model.

In practice one computes the values of $(\sigma_1 - \sigma_2)$ from the observed values of δ or N . Then,

$$(\sigma_1 - \sigma_2) = \frac{N\lambda}{dC} = \frac{N}{d} F \quad (9.15)$$

F is called the **material fringe value**. If $d = 1 \text{ cm}$ and $N = 1$ wavelength, then F gives the value of $(\sigma_1 - \sigma_2)$. It produces a relative phase difference of 2π radians on a model of unit thickness. This is a property of the model material and wavelength of light used. The quantity

$$\frac{F}{d} = f \quad (9.16)$$

is called the **model fringe value**.

At those points in a stressed model where $\sigma_1 = \sigma_2$, the fringe order is zero and black dots appear at these points. Such points are called *isotropic points*. If $\sigma_1 = \sigma_2 = 0$, then also

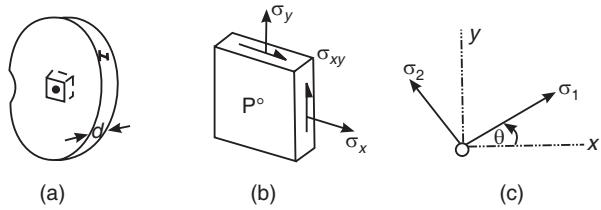


Fig. 9.11

the fringe order is zero at these points and black dots appear. Such points are called *singular points*.

9.10.2 Definition of Isoclinics and Isochromatics

Isoclinics are the locus of the points in the specimen along which the principal stresses are in the same direction.

Isochromatics are the locus of the points along which the difference in the first and second principal stress remains the same. Thus they are the lines which join the points with equal maximum shear stress magnitude.

9.10.3 Photo Elastic Bench

Photo elastic bench is an optical instrument used to analyse the stress distribution in a model subjected to load. The instrument utilizes the properties of polarized light in its operation.

Principle: When a transparent material is stressed, it becomes double refracting. On examining the stressed material between crossed polarizers using light, interference fringes are observed. The fringes are used to test and measure the stress and strain produced in the material.

Construction: A photoelastic bench mainly consists of the following parts.

- (i) Polariscope
- (ii) Loading frame
- (iii) Light source and
- (iv) Camera.

The block diagram of the arrangement of a photoelastic bench is shown in Fig. 9.12.

For photoelastic analysis, two types of polarisopes are used.

- (a) Plane polariscope
- (b) Circular polariscope.

In the plane polariscope, plane-polarized light is used and in the circular polariscope, circularly polarized light is used.

9.10.3.1 Plane polariscope

The basic arrangement of a lens type polariscope is shown in Fig. 9.12.

Working: An incandescent lamp serves as a white light source. The first field lens FL_1 gives a parallel beam of light in the field of view. The unpolarized light emerging from the field lens is then passed through the polarizer P and gets plane-polarized.

The model M made of a photoelastic material is loaded in a loading frame by which various types of loads can be applied. The plane polarized light emerging out of P passes through the model in stressed condition and splits into two beams. These beams are plane polarized in mutually perpendicular planes and are incident on the analyzer A . The

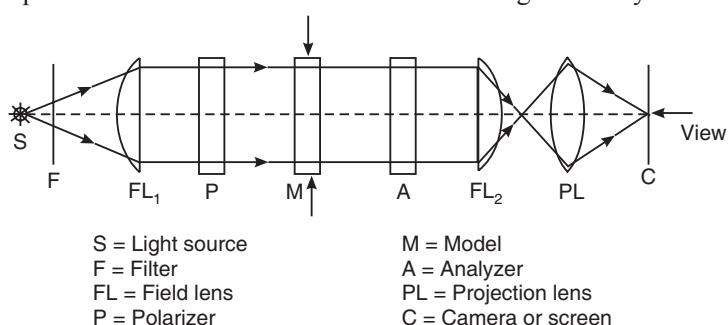


Fig. 9.12

analyzer combines these two beams coming from the model. The polarizer and analyzer are generally coupled together by a flexible coupling to achieve their simultaneous rotation.

The second field lens FL_2 makes the parallel beam of light to converge on the projection lens, PL, which finally projects the interference fringes on to the screen or camera C.

The fringe pattern in a plane polariscope setup consists of both the isochromatics and the isoclinics. The isoclinics change with the orientation of the polariscope while there is no change in the isochromatics.

9.10.3.2 Circular polariscope

The circular polariscope contains two quarter-wave plates extra in addition to all the elements of a plane polariscope.

The first quarter-wave plate is kept between the polarizer P and the model M, while the second one is held between the model and the analyzer A.

The first QWP converts plane polarized light into circularly polarized light and the second QWP converts circularly polarized light into plane-polarized light.

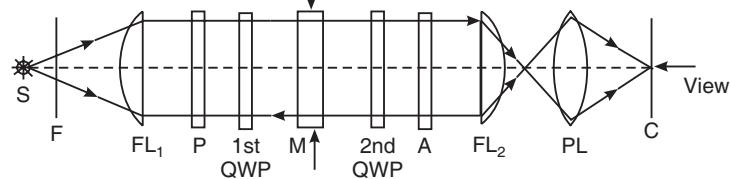


Fig. 9.13

The basic advantage of a circular polariscope over a plane polariscope is that in a circular polariscope setup we only get the isochromatics and not the isoclinics. This eliminates the problem of differentiating between the isoclinics and the isochromatics.

9.10.4 Isoclinics and Isochromatics

Isoclinics and isochromatics are the two different types of fringes observed in photoelastic studies. Isoclinic fringes occur whenever either principal stress direction coincides with the axis of polarization of the polarizer. They provide information about the *directions of the principal stresses* in the component. Isoclinic fringes can be removed by using a circular polarizer. Isochromatic fringes are lines of constant **principal stress difference** ($\sigma_1 - \sigma_2$). With monochromatic light, they appear as dark and bright fringes and with white light illumination, they appear as coloured fringes. In a plane polariscope, the two types of fringes are found superposed on each other and can be distinguished by rotating the component. Isoclinic fringes vary in intensity as they pass through the extinction positions, whereas isochromatic fringes remain unchanged.

Let us consider a model suitably held in a plane polariscope. Let the polarizer and analyzer combination be held in crossed configuration, which is called as the *dark field* set up. When the model is stressed, it becomes double refracting. At the point where the ray passes, the polarizing axes coincide with the principal stress axes σ_1, σ_2 , at that point.

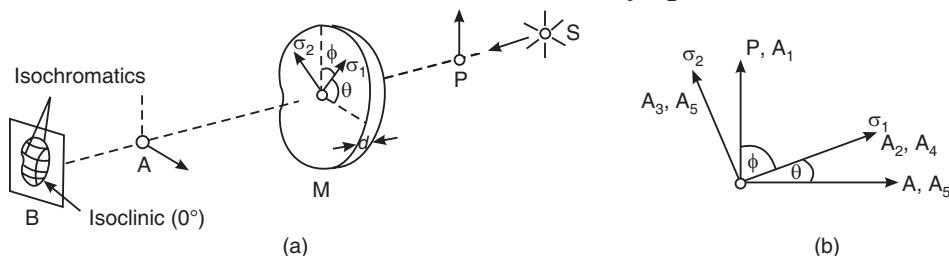


Fig. 9.14

In general, the polarizer makes an angle ϕ with the σ_1 -axis. If ϕ happens to be zero or 90° , the direction of the polarizer coincides with either σ_1 (or σ_2), then plane polarized light incident on the model at that point will emerge as plane polarized light. Since the analyzer is kept crossed with respect to the polarizer, no light comes out of the analyzer. Consequently, at all those points of the model where the directions of the principal stresses happen to coincide with the particular orientation of the polarizer-analyzer combinations, the light coming out of the analyzer is zero. If the polarizer-analyzer combination happens to coincide with the directions of σ_1, σ_2 stresses at one point of the model, then in general, there will be a locus of points in the model along which this condition is satisfied.

The locus of points where the directions of principal stresses coincide with a particular orientation of the polarizer-analyzer combination is known as the **isoclinic**. For example, if the polarizing element is kept vertical and the analyzer horizontal, then on the screen a dark band will be seen which is the locus of points where σ_1 , and σ_2 directions happen to be vertical and horizontal. If one measures angles from the vertical reference axis, this isoclinic will be called the 0° isoclinic. If now the polarizer is turned through 30° and the analyzer is also rotated through an equal amount, then the previously observed 0° isoclinic vanishes and a new dark band is observed on the screen. This is the 30° isoclinic and it represents the locus of points in the model where the principal stress axes are oriented at 30° and $30^\circ + \pi/2$ with respect to the vertical.

Let us now consider another situation. Suppose at a particular point of the model, the values of σ_1 and σ_2 are such that they cause a relative phase difference of $2 m\pi$ where m is an integer. If the phase difference is $2 m\pi$, the model behaves like a full wave plate at that particular point. Therefore, at all these points of the model where the values of σ_1 and σ_2 are such as to cause a phase difference of $2 m\pi$, the intensity of light will be zero. On the screen a series of dark bands corresponding to these locus of points are observed. These dark bands or fringes are called **isochromatics**. An isochromatic is a locus of points where the values of σ_1 and σ_2 are such as to cause a phase difference of $2 m\pi$, when the background is dark. The locus of points, where the values of $\sigma_1 - \sigma_2$ are such as to cause zero radian of phase difference, is called the *zero-order* fringe. The locus of points, where the values of $\sigma_1 - \sigma_2$ are such as to cause 2π radian of phase difference, is called the *first-order* fringe. Similarly, one can observe second order fringe, third order fringe and so on. Fig. 9.15 shows typical isoclinic and isochromatics for a stressed circular disc.

9.10.5 Mathematical Analysis

Let the plane polarized light coming from the polarizer be

$$E_1 = A \cos \omega t \quad (9.17)$$

On entering the model, the light vector gets resolved along the principal stress axes. Thus

$$E_2 = A \cos \varphi \cos \omega t \quad (9.18 a)$$

and

$$E_3 = A \sin \varphi \cos \omega t \quad (9.18 b)$$

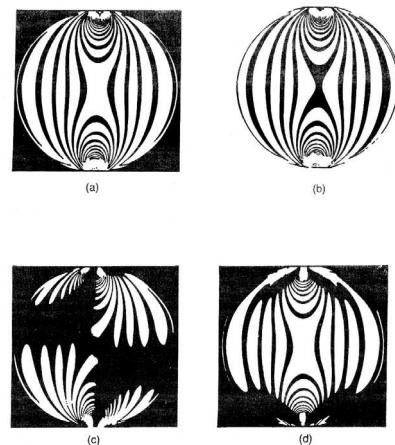


Fig. 9.15: Isoclinics and Isochromatics for a diametrically stressed circular disc
 (a) Isochromatics – in dark field set up
 (b) Isochromatics – in bright field view
 (c) 15° - Isoclinic (d) 45° - Isoclinic

and σ_2 are such as to cause a phase difference of $2 m\pi$, when the background is dark. The locus of points, where the values of $\sigma_1 - \sigma_2$ are such as to cause zero radian of phase difference, is called the *zero-order* fringe. The locus of points, where the values of $\sigma_1 - \sigma_2$ are such as to cause 2π radian of phase difference, is called the *first-order* fringe. Similarly, one can observe second order fringe, third order fringe and so on. Fig. 9.15 shows typical isoclinic and isochromatics for a stressed circular disc.

In traveling the thickness d of the model, the two components acquire a relative phase difference δ . On leaving the model, the light components are

$$E_4 = A \cos \varphi \cos (\omega t + \delta) \quad (9.19 \text{ a})$$

and

$$E_5 = A \sin \varphi \cos \omega t \quad (9.19 \text{ b})$$

On entering the analyzer, only the components along the axis of analyzer are allowed. Thus,

$$E_6 = E_4 \sin \varphi - E_5 \cos \varphi \quad (9.20)$$

$$= A \cos \varphi \sin \varphi \cos(\omega t + \delta) - A \cos \varphi \sin \varphi \cos \omega t$$

$$= \frac{A}{2} \sin 2\phi [\cos(\omega t + \delta) - \cos \omega t]$$

$$= \frac{A}{2} \sin 2\phi [\cos \omega t (\cos \delta - 1) - \sin \omega t \sin \delta]$$

$$= \frac{A}{2} \sin 2\phi \left[-2 \cos \omega t \sin^2 \frac{\delta}{2} - 2 \sin \omega t \sin \frac{\delta}{2} \cos \frac{\delta}{2} \right]$$

$$= -A \sin 2\phi \sin \frac{\delta}{2} \left[\cos \omega t \sin \frac{\delta}{2} + \sin \omega t \cos \frac{\delta}{2} \right]$$

$$= -A \sin 2\phi \sin \frac{\delta}{2} \sin \left(\omega t + \frac{\delta}{2} \right)$$

$$= -a \sin \left(\omega t + \frac{\delta}{2} \right) \quad (9.21)$$

where $a = A \sin 2\phi \sin \frac{\delta}{2}$ is the amplitude of the light emerging from the analyzer. The intensity of the light is given by

$$I = |a|^2 = A^2 \sin^2 2\phi \sin^2 \frac{\delta}{2} \quad (9.22)$$

The intensity of light is zero under two conditions:

$$\text{When } \phi = 0 \text{ or } \pi/2; \text{ or/and} \quad \dots (1)$$

$$\text{When } \delta = 2m\pi \quad (m = 0, 1, 2, 3, \dots) \quad \dots (2)$$

Condition (1) indicates that light extinction occurs at a point where the direction of the principal stresses coincides with the direction of the polarizer and analyzer. The locus of points where the direction of principal stresses has a common orientation with reference to a given axis is an isoclinic.

Condition (2) tells us that the light intensity is zero when the relative phase difference is equal to $2m\pi$. The locus of points satisfying this condition is called an isochromatic.

Example 9.6: In an experiment using a photo-elastic bench, the difference between principal stresses is $8 \times 10^9 \text{ Nm}^{-2}$. The photelastic material has relative stress optic coefficient of 2 brewsters. Calculate the difference between refractive indices along the principal stresses.

Solution: Difference between principal stresses, $\sigma_1 - \sigma_2 = 8 \times 10^9 \text{ Nm}^{-2}$

Stress optic coefficient, $C = 2 \text{ brewsters} = 2 \times 10^{-12} \text{ m}^2 \text{ N}^{-1}$

Difference between refractive indices,

$$(n_2 - n_1) = C \times (\sigma_1 - \sigma_2)$$

$$= 2 \times 10^{-12} \text{ m}^2 \text{ N}^{-1} \times 8 \times 10^9 \text{ Nm}^{-2} = 0.016.$$

QUESTIONS

1. Give the construction and working of Laurent's half shade polarimeter.
2. Explain Fresnel's theory of rotation of the plane of polarisation.
3. What is meant by artificial double refraction?
4. What are electro-optic effects?
5. What are magneto-optic effects?
6. Explain Kerr effect.
7. What is Pockels effect?
8. What is Cotton-Mouton effect?
9. Discuss Faraday effect.
10. Explain photoelasticity. What is photoelastic effect?
11. State the stress-optic law and obtain an expression for the same.
12. What are isoclinic and isochromatic fringes?
13. Describe the working of a plane polariscope
14. Explain the working of a circular polariscope.

PROBLEMS

1. A 10 cm long tube containing 10% sugar solution produces optical rotation of 13.2° . Find the specific rotation of sugar under given experimental conditions. [Ans: 66°]
2. Calculate the specific rotation if the plane of polarization is turned through 26.4° , traversing 20 cm length of 20% sugar solution. [Ans: 66°]
3. A length of 25 cm of a solution containing 50 gm of solute per litre causes a rotation of the plane of polarization of light by 5° . Find the rotation of plane of polarization by a length of 75 cm of a solution containing 100 gm of solute per litre. [Ans: 30°]
4. For quartz the refractive indices for right-handed and left-handed vibrations are 1.55810 and 1.55821 respectively for $\lambda = 4000\text{\AA}$. Find the amount of optical rotation produced at $\lambda = 4000\text{\AA}$ by a plate of quartz 2 mm thick and with its faces perpendicular to the optic axis. [Ans: 98°]
5. A solution of camphor in alcohol in a tube of 20 cm long is found to rotate the plane of vibration of light by 27° . What is the mass of the camphor in unit volume of solution? The specific rotation of camphor is + 54° . [Ans: 0.25gm/cc]
6. A plate of quartz cut with its faces perpendicular to the optic axis is found to annul exactly the rotation of plane of polarization of sodium light produced by a 30 cm length of 18% solution of lactose. Find the thickness of the quartz plate. Given that specific rotation of lactose is 52.53° and that 1 mm quartz rotates the plane of polarization of sodium light by 21.71° . [Ans: 1.31mm]
7. A sugar solution in a tube of length 20 cm produces optical rotation of 13° . The solution is then diluted to one-third of its previous concentration. Find the optical rotation produced by 30 cm long tube containing the diluted solution. [Ans: 6.5°]
8. 80 gm of impure sugar when dissolved in a litre of water gives an optical rotation of 99° when placed in a tube of length 20 cm. If the specific rotation of sugar is 66° , find the percentage purity of the sugar sample. [Ans: 93.75%]
9. The rotation in the plane of polarization in a certain substance, at 5893\AA , is $10^\circ/\text{cm}$. Calculate the difference between the refractive indices for right and left circularly polarized light in the substance. [Ans: 8.186×10^{-7}]
10. For a given wavelength 1 mm of quartz cut perpendicular to the optic axis rotates the plane of polarization by 20° . Find for what thickness will no light of this wavelength be transmitted when the quartz pieces is interposed between parallel Nicols. [Ans: 4.5 mm]

CHAPTER

10

Optical Fibres

10.1 INTRODUCTION

In 1870 John Tyndall, a British physicist demonstrated that light can be guided along the curve of a stream of water. Owing to total internal reflections light gets confined to the water stream and the stream appears luminous. A luminous water stream is the precursor of an optical fibre. In the 1950's, the transmission of images through optical fibres was realized in practice. Hopkins and Kapany developed the flexible fibrescope, which was used by the medical world in remote illumination and viewing the interior of human body. It was Kapany who coined the term fibre optics. By 1960, it had been established that light could be guided by a glass fibre. In 1966 Charles Kao and George Hockham proposed the transmission of information over glass fibre, but the fibres available at that time heavily attenuated light propagating through them. In 1970 Corning Glass Works produced low-loss glass fibres. The invention of solid state lasers in 1970 made optical communications practicable. Commercial communication systems based on optical fibres made their appearance by 1977. Apart from the use as communicational channel, optical fibres are widely used in other areas. Fibro-scopes made of optical fibres are widely used in a variety of forms in medical diagnostics. Sensors for detecting electrical, mechanical, thermal energies are made using optical fibres.

Fibre optics is a technology in which signals are converted from electrical into optical signals, transmitted through a thin glass fibre and reconverted into electrical signals.

10.2 OPTICAL FIBRE

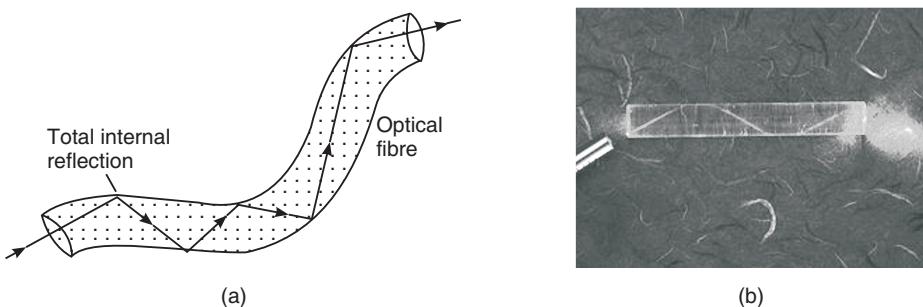


Fig. 10.1: Illustration of a transparent fibre guiding light along its length by total internal reflection.

Definition: *An optical fibre is a cylindrical wave guide made of transparent dielectric, (glass or clear plastic), which guides light waves along its length by total internal reflection. It is as*

thin as human hair, approximately $70\text{ }\mu\text{m}$ or 0.003 inch diameter. (Note that a thin strand of a metal is called a *wire* and a thin strand of dielectric materials is called a *fibre*).

Principle: The propagation of light in an optical fibre from one of its ends to the other end is based on the principle of *total internal reflection*. When light enters one end of the fibre, it undergoes successive total internal reflections from sidewalls and travels down the length of the fibre along a zigzag path, as shown in Fig. 10.1 (a). A small fraction of light may escape through sidewalls but a major fraction emerges out from the exit end of the fibre, as shown in Fig. 10.1 (b). Light can travel through fibre even if it is bent [Fig. 10.1(c)].

Structure:

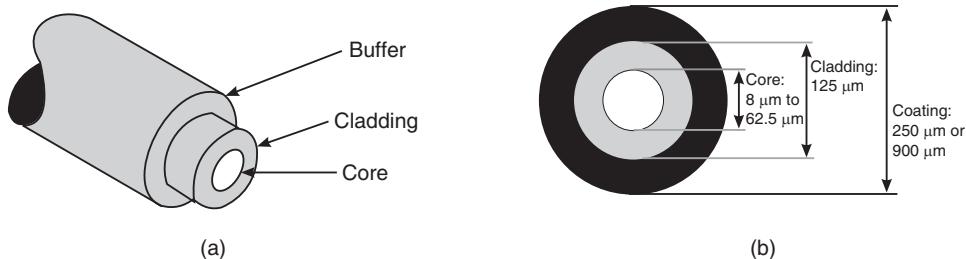


Fig. 10.2: Side view and cross sectional view of a typical optical fibre

A practical optical fibre is cylindrical in shape (Fig. 10.2a) and has in general three coaxial regions (Fig. 10.2b).

- The innermost cylindrical region is the light guiding region known as the **core**. In general, the diameter of the core is of the order of 8.5 μm to 62.5 μm .
- It is surrounded by a coaxial middle region known as the **cladding**. The diameter of the cladding is of the order of 125 μm . The refractive index of cladding (n_2) is always lower than that of the core (n_1). Light launched into the core and striking the core-to-cladding interface at an angle greater than critical angle will be reflected back into the core. Since the angles of incidence and reflection are equal, the light will continue to rebound and propagate through the fibre.
- The outermost region is called the **sheath** or a **protective buffer coating**. It is a plastic coating given to the cladding for extra protection. This coating is applied during the manufacturing process to provide physical and environmental protection for the fiber. The buffer is elastic in nature and prevents abrasions. The coating can vary in size from 250 μm or 900 μm .

To sum up

- Core is the inner light-carrying member.
- Cladding is the middle layer, which serves to confine the light to the core.

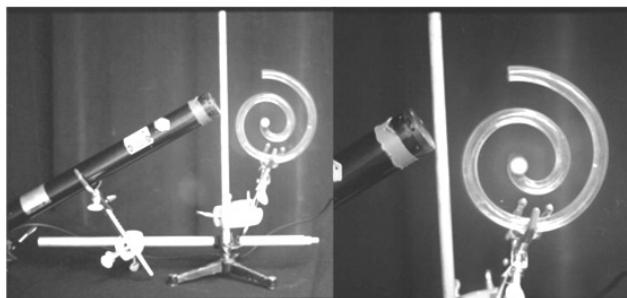


Fig. 10.1(c): A laser beam repeatedly bounces off the surface of the rod (by total internal reflection) as it makes its way around the plastic spiral and emerges out at the other end. The whole coil appears to glow red due to the scattering of light by the plastic. However, the path of light is seen clearly in the coil.

- Buffer coating surrounds the cladding, which protects the fibre from physical damage and environmental effects.

10.2.1 Necessity of Cladding

The actual fibre is very thin and light entering a bare fibre will travel along the fibre through repeated total internal reflections at the glass-air boundary. For use in communications and other applications, the optical fibre is provided with a cladding. *The cladding maintains uniform size of the fibre, protects the walls of the fibre from chipping, and reduces the size of the cone of light that will be trapped in the fibre.*

- It is necessary that the diameter of an optical fibre remains constant throughout its length and is surrounded by the same medium. Any change in the thickness of the fibre or the medium outside the fibre (when the fibre gets wet due to moisture etc) will cause loss of light energy through the walls of the fibre.
- A very large number of reflections occur through the fibre and it is necessary that the condition for total internal reflection must be accurately met over the entire length of the fibre. If the surface of the glass fibre becomes scratched or chipped, the normal to the edge will no longer be uniform. As a result, the light traveling through the fibre will get scattered and escapes from the fibre. This also causes loss of light energy.
- Part of light energy penetrates the fibre surface. The intensity of the light decreases exponentially as we move away from the surface, as the light is able to penetrate only a very small distance outside the fibre. However, anytime the fibre touches something else, the light can leak into the new medium or be scattered away from the fibre. This effect causes a significant leakage of the light energy out of the fibre. Even a small amount of dust on the surface would cause a fair amount of leakage.
- If bare optic fibres are packed closely together in a bundle, light energy traveling through the individual fibres tends to get coupled through the phenomenon of *frustrated total internal reflection*. Cladding of sufficient thickness prevents the leakage of light energy from one fibre to the other.

The fiber is provided with a cladding in order to prevent loss of light energy due to the above reasons.

- The cladding causes a reduction in the size of the cone of light that can be trapped in the fibre. Light entering the fibre at larger angles will strike the fibre walls at smaller angles (higher modes) and ultimately travel a longer distance. Such higher modes of a light signal will take longer time to reach the end of the fibre than the lower modes. Therefore, a pulse sent through optical fibre spreads out. The spreading would be larger, the larger the cone of acceptance. Such pulse spreading limits the rate of data transmission through the fibre. As fibers with a cladding have smaller cone of acceptance, they carry information at a much higher bit rate than those without a cladding.

Thus, the cladding performs the following important functions:

- Keeps the size of the fibre constant and reduces loss of light from the core into the surrounding air.
- Protects the fiber from physical damage and absorbing surface contaminants.
- Prevents leakage of light energy from the fibre through evanescent waves.
- Prevents leakage of light energy from the core through frustrated total internal reflection.
- Reduces the cone of acceptance and increases the rate of transmission of data.
- A solid cladding, instead of air, also makes it easier to add other protective layers over the fibre.

10.2.2 Optical Fibre System

An optical fibre is used to transmit **light signals** over long distances. It is essentially a **light-transmitting medium**, its role being very much similar to a coaxial cable or wave-guide used in microwave communications. Optical fibre requires a **light source** for launching light into the fibre at its input end and a **photodetector** to receive light at its output end. As the diameter of the fibre is very small, the light source has to be dimensionally compatible with the fibre core. Light emitting diodes and laser diodes, which are very small in size, serve as the light sources. The electrical input signal is in general of digital form. It is converted into an optical signal by varying the current flowing through the light source. Hence, the intensity of the light emitted by the source is modulated with the input signal and the output will be in the form of light pulses. The light pulses constitute the signal that travels through the optical fibre. At the receiver end, semiconductor photodiodes, which are very small in size, are used for detection of these light pulses. The photodetector converts the optical signal into electrical form. Thus, a basic *optical fibre system* consists of a LED/laser diode, optical fibre cable and a semiconductor photodiode.

10.2.3 Optical Fibre Cable

Optical fibre cables are designed in different ways to serve different applications. More protection is provided to the optical fibre by the “cable” which has the fibres and strength members inside an outer covering called a “jacket”. We study here two typical designs: a single fibre cable or a multifibre cable.

- **Single Fibre Cable:** Around the fibre a tight buffer jacket of Hytrel is used (see Fig. 10.3). The buffer jacket protects the fibre from moisture and abrasion. A strength member is arranged around the buffer jacket in order to provide the necessary toughness and tensile strength. The strength member may be a steel wire, polymer film, nylon yarn or Kevlar yarn. Finally, the fibre cable is covered by a Hytrel outer jacket. Because of this arrangement fibre cable will not get damaged during bending, rolling, stretching or pulling and transport and installation processes. The single fibre cable is used for indoor applications.

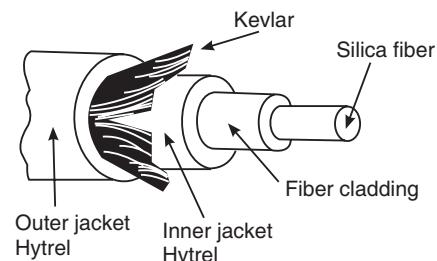


Fig. 10.3: Single fibre cable

Because of this arrangement fibre cable will not get damaged during bending, rolling, stretching or pulling and transport and installation processes. The single fibre cable is used for indoor applications.

- **Multifibre Cable:** A multifibre cable consists of a number of fibres in a single jacket. Each fibre carries light independently. The cross-sectional view of a typical telecommunication cable is shown in Fig. 10.4. It contains six insulated optical fibre strands and has an insulated steel cable at the center for providing tensile strength. Each optical fibre strand consists of a core surrounded by a cladding, which in turn is coated with insulating jacket.

The fibres are thus individually buffered and strengthened. Six insulated copper wires are distributed in the space between the fibres. They are used for electrical transmission,

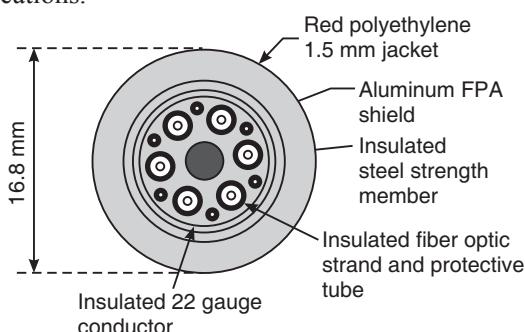


Fig. 10.4: Cross-sectional view of a typical multi fibre cable

if required. The assembly is then fitted with in a corrugated aluminium sheath, which acts as a shield. A polyethylene jacket is applied over the top.

10.3 TOTAL INTERNAL REFLECTION

A medium having a lower refractive index is said to be an optically **rarer medium** while a medium having a higher refractive index is known as an optically **denser medium**. When a ray of light passes from a denser medium to a rarer medium, it is bent away from the normal in the rarer medium (see Fig. 10.5a). Snell's law for this case may be written as

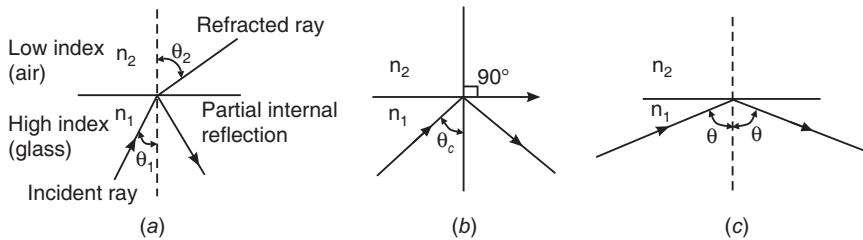


Fig. 10.5: Phenomenon of total internal reflection

$$\sin \theta_2 = \left(\frac{n_1}{n_2} \right) \sin \theta_1 \quad (10.1)$$

where θ_1 is the angle of incidence of light ray in the denser medium and θ_2 is the angle of refraction in the rarer medium. Also $n_1 > n_2$. When the angle of incidence, θ_1 in the denser medium is increased, the transmission angle, θ_2 increases and the refracted rays bend more and more away from the normal. At some particular angle θ_c the refracted ray glides along the boundary surface so that $\theta_2 = 90^\circ$, as seen in Fig. 10.5 (b). At angles greater than θ_c there are no refracted rays at all. The rays are reflected back into the denser medium as though they encountered a specular reflecting surface (Fig. 10.5 c). Thus,

- If $\theta_1 < \theta_c$, the ray refracts into the rarer medium
- If $\theta_1 = \theta_c$, the ray just grazes the interface of rarer-to-denser media
- If $\theta_1 > \theta_c$, the ray is reflected back into the denser medium.

The phenomenon in which light is totally reflected from a denser-to-rarer medium boundary is known as **total internal reflection**. The rays that experience total internal reflection obey the laws of reflection. Therefore, the critical angle can be determined from Snell's law.

$$\text{When } \theta_1 = \theta_c, \quad \theta_2 = 90^\circ.$$

Therefore, from equ. (10.1), we get

$$\begin{aligned} n_1 \sin \theta_c &= n_2 \sin 90^\circ = n_2 \\ \therefore \sin \theta_c &= \frac{n_2}{n_1} \end{aligned} \quad (10.2)$$

When the rarer medium is air, $n_2 = 1$ and writing $n_1 = n$, we obtain

$$\sin \theta_c = \frac{1}{n} \quad (10.3)$$

10.4 PROPAGATION OF LIGHT THROUGH AN OPTICAL FIBRE

The diameter of an optical fibre is very small and as such we cannot use bigger light sources for launching light beam into it. Light emitting diodes (LEDs) and laser diodes are the optical

sources used in fibre optics. Even in case of these small sized sources, a focusing lens has to be used to concentrate the beam on to the fibre core. Light propagates as an electromagnetic wave through an optical fibre.

However, light propagation through an optical fibre can as well be understood on the basis of *ray model*. According to the ray model, light rays entering the fibre strike the core-clad interface at different angles. As the refractive index of the cladding is less than that of the core, majority of the rays undergo total internal reflection at the interface and the angle of reflection is equal to the angle of incidence in each case. Due to the cylindrical symmetry in the fibre structure, the rays reflected from an interface on one side of the fibre axis will suffer total internal reflections at the interface on the opposite side also. Thus, the rays travel forward through the fibre via a series of total internal reflections and emerge out from the exit end of the fibre (Fig. 10.6). Since each reflection is a total internal reflection, there is no loss of light energy and light confines itself within the core during the course of propagation. Because of the negligible loss during the total internal reflections, optical fibre can carry the light waves over very long distances. Thus, the optical fibre acts essentially as a wave-guide and is often called a **light guide** or **light pipe**. At the exit end of the fibre, the light is received by a photo-detector.

Total internal reflection at the fibre wall can occur and light propagates down the fibre, only if the following two conditions are satisfied.

1. The refractive index of the core material, n_1 , must be slightly greater than that of the cladding, n_2 .
2. At the core-cladding interface (Fig. 10.7), the angle of incidence ϕ between the ray and the normal to the interface must be greater than the critical angle ϕ_c defined by

$$\sin \phi_c = \frac{n_2}{n_1} \quad (10.4)$$

It is to be noted here that only those rays, that are incident at the core-clad interface at angles greater than the critical angle will propagate through the fibre. Rays that are incident at smaller angles are refracted into the cladding and are lost.

10.4.1 Critical Angle of Propagation

Let us consider a step index optical fibre into which light is launched at one end. The end at which light enters the fibre is called the **launching end**. Fig. 10.7 depicts the conditions at the launching end. In a step-index fibre, the refractive index changes abruptly from the core to the cladding. Now, we consider two rays entering the fibre at two different angles of incidence.

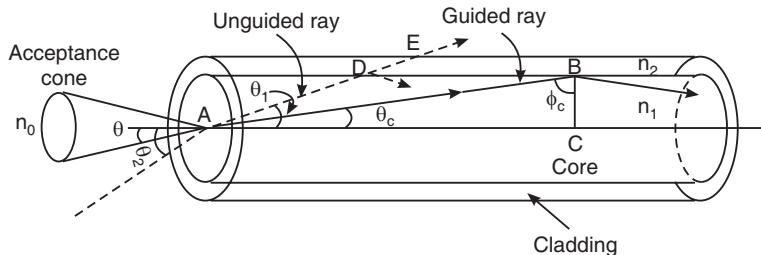


Fig. 10.7: Light rays incident at an angle smaller than critical propagation angle will propagate through the fibre.

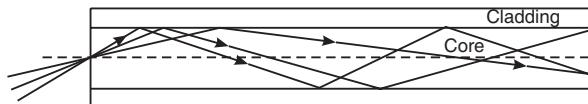


Fig. 10.6: Propagation of light rays through an optical fibre due to total internal reflection.

The ray shown by the broken line is incident at an angle θ_2 with respect to the axis of the fibre. This ray undergoes refraction at point A on the interface between air and the core. The ray refracts into the fibre at an angle θ_1 ($\theta_1 < \theta_2$). The ray reaches the core-cladding interface at point D. At point D, refraction takes place again and the ray travels in the cladding. Finally, at point E, the ray refracts once again and emerges out of fibre into the air. It means that the ray does not propagate through the fibre.

Let us next consider the ray shown by the solid line in Fig. 10.7. The ray incident at an angle θ undergoes refraction at point A on the interface and propagates at an angle θ_c in the fibre. At point B on the core-cladding interface, the ray undergoes total internal reflection, since $n_1 > n_2$. Let us assume that the angle of incidence at the core-cladding interface is the *critical angle* ϕ_c , where ϕ_c is given by

$$\phi_c = \sin^{-1} (n_2/n_1) \quad (10.4a)$$

A ray incident with an angle larger than ϕ_c will be confined to the fibre and propagate in the fibre. A ray incident, at the core-cladding boundary, at the critical angle is called a **critical ray**. The critical ray makes an angle ϕ_c with axis of the fibre. It is obvious that rays with propagation angles larger than θ_c will not propagate in the fibre. Therefore, the angle θ_c is called the **critical propagation angle**. From the Δ^{le} ABC, it is seen that

$$\frac{AC}{AB} = \sin \phi_c. \quad \text{Also, } \frac{AC}{AB} = \cos \theta_c$$

From the relation (10.4a), $\sin \phi_c = n_2 / n_1$.

$$\cos \theta_c = n_2 / n_1 \quad (10.5)$$

$$\therefore \theta_c = \cos^{-1}(n_2 / n_1) \quad (10.6)$$

Thus, only those rays which are refracted into the cable at angles $\theta_r < \theta_c$ will propagate in the optical fibre.

Example 10.1: In an optical fibre, the core material has refractive index 1.43 and refractive index of clad material is 1.4. Find the propagation angle.

Solution: $\cos \theta_c = \frac{n_2}{n_1} = \frac{1.40}{1.43} = 0.979$

Therefore, propagation angle $\theta_c = \cos^{-1}(0.979) = 11.8^\circ$

Example 10.2: In an optical fibre, the core material has refractive index 1.6 and refractive index of clad material is 1.3. What is the value of critical angle? Also calculate the value of angle of acceptance cone.

Solution: Critical angle is given by

$$\sin \phi_c = \frac{n_2}{n_1} = \frac{1.3}{1.6} = 0.8125$$

$$\therefore \phi_c = 54.3^\circ$$

$$\begin{aligned} \text{Acceptance angle } \theta_0 &= \sin^{-1} \left[\sqrt{n_1^2 - n_2^2} \right] = \sin^{-1} \left[\sqrt{1.6^2 - 1.3^2} \right] \\ &= \sin^{-1} (0.87) \\ &= 60.5^\circ \end{aligned}$$

Angle of acceptance cone = $2\theta_0 = 121^\circ$

10.4.2 Acceptance Angle

Let us again consider a step index optical fibre into which light is launched at one end, as shown in Fig. 10.8. Let the refractive index of the core be n_1 and the refractive index of the cladding be n_2 ($n_2 < n_1$). Let n_0 be the refractive index of the medium from which light is launched into the fibre. Assume that a light ray enters the fibre at an angle θ_i to the axis of the fibre. The ray refracts at an angle θ_r and strikes the core-cladding interface at an angle ϕ . If ϕ is greater than critical angle ϕ_c , the ray undergoes total internal reflection at the interface, since $n_1 > n_2$. As long as the angle ϕ is greater than ϕ_c , the light will stay within the fibre.

Applying Snell's law to the launching face of the fibre, we get

$$\frac{\sin \theta_i}{\sin \theta_r} = \frac{n_1}{n_0} \quad (10.7)$$

If θ_i is increased beyond a limit, ϕ will drop below the critical value ϕ_c and the ray escapes from the sidewalls of the fibre. The largest value of θ_i occurs when $\phi = \phi_c$.

From the Δ^{le} ABC, it is seen that

$$\sin \theta_r = \sin (90^\circ - \phi) = \cos \phi \quad (10.8)$$

Using equation (10.8) into equation (10.7), we obtain

$$\sin \theta_i = \frac{n_1}{n_0} \cos \phi$$

$$\text{When } \phi = \phi_c, \quad \sin [\theta_{i_{\max}}] = \frac{n_1}{n_0} \cos \phi_c \quad (10.9)$$

$$\begin{aligned} \text{But} \quad \sin \phi_c &= \frac{n_2}{n_1} \\ \therefore \cos \phi_c &= \frac{\sqrt{n_1^2 - n_2^2}}{n_1} \end{aligned} \quad (10.10)$$

Substituting the expression (10.10) into (10.9), we get

$$\sin [\theta_{i_{\max}}] = \frac{\sqrt{n_1^2 - n_2^2}}{n_0} \quad (10.11)$$

Quite often the incident ray is launched from air medium, for which $n_0 = 1$.

Designating $\theta_{i_{\max}} = \theta_0$, equation (10.11) may be simplified to

$$\sin \theta_0 = \sqrt{n_1^2 - n_2^2}$$

$$\therefore \theta_0 = \sin^{-1} \left[\sqrt{n_1^2 - n_2^2} \right] \quad (10.12)$$

The angle θ_0 is called the **acceptance angle** of the fibre. *Acceptance angle is the maximum angle that a light ray can have relative to the axis of the fibre and propagate down the fibre.*

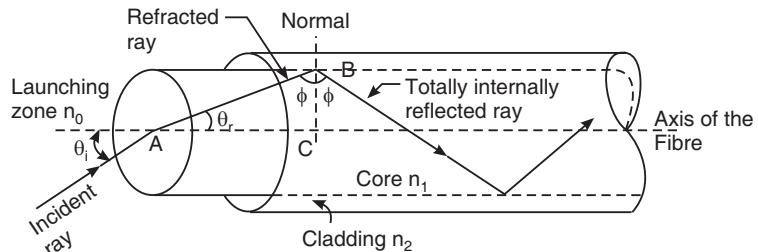


Fig. 10.8: Geometry for the calculation of acceptance angle of the fibre.

Thus, only those rays that are incident on the face of the fibre making angles less than θ_0 will undergo repeated total internal reflections and reach the other end of the fibre. Obviously, larger acceptance angles make it easier to launch light into the fibre.

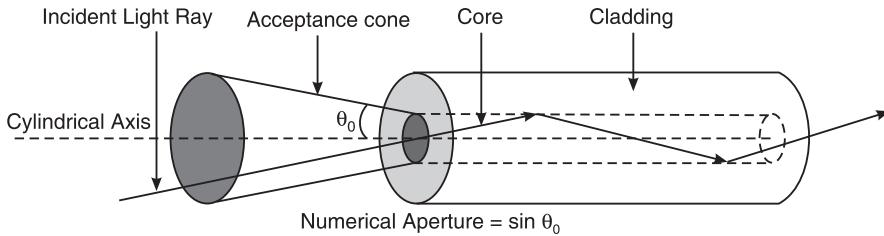


Fig. 10.9

In three dimensions, the light rays contained within the cone having a full angle $2\theta_0$ are accepted and transmitted along the fibre (see Fig. 10.9). Therefore, the cone is called the **acceptance cone**.

Light incident at an angle beyond θ_0 refracts through the cladding and the corresponding optical energy is lost.

Example 10.3: Calculate the numerical aperture and acceptance angle of an optical fibre from the following data:

$$n_1(\text{core}) = 1.55 \quad \text{and} \quad n_2(\text{cladding}) = 1.50$$

$$\text{Solution: } \text{NA} = \sqrt{n_1^2 - n_2^2} = \sqrt{1.55^2 - 1.50^2} = \sqrt{0.153} = 0.391.$$

$$\text{Acceptance angle } \theta_0 = \sin^{-1} \left[\sqrt{n_1^2 - n_2^2} \right] = \sin^{-1} \left[\sqrt{1.55^2 - 1.50^2} \right] = 23.02^\circ$$

Example 10.4: What is the numerical aperture of an optical fibre cable with a clad index of 1.378 and a core index of 1.546?

$$\text{Solution: } \text{NA} = \sqrt{n_1^2 - n_2^2} = \sqrt{1.546^2 - 1.378^2} = \sqrt{0.491} = 0.70$$

Example 10.5: A fibre cable has an acceptance angle of 30° and a core index of refraction of 1.4. Calculate the refractive index of the cladding.

$$\text{Solution: } \sin \theta_0 = \sqrt{n_1^2 - n_2^2}$$

$$\therefore \sin^2 \theta_0 = n_1^2 - n_2^2$$

$$\begin{aligned} n_2^2 &= n_1^2 - \sin^2 \theta_0 = (1.4)^2 - \sin^2 30^\circ = 1.96 - 0.25 \\ &= 1.71 \\ \therefore n_2 &= 1.308 \end{aligned}$$

Example 10.6: Calculate the angle of acceptance of a given optical fibre, if the refractive indices of the core and the cladding are 1.563 and 1.498 respectively.

$$\text{Solution: } \sin \theta_0 = \sqrt{n_1^2 - n_2^2} = \sqrt{(1.563)^2 - (1.498)^2} = 0.4461$$

$$\theta_0 = \sin^{-1}(0.4461) = 26.49^\circ$$

10.5 FRACTIONAL REFRACTIVE INDEX CHANGE

The fractional difference Δ between the refractive indices of the core and the cladding is known as *fractional refractive index change*. It is expressed as

$$\Delta = \frac{n_1 - n_2}{n_1} \quad (10.13)$$

This parameter is always positive because n_1 must be larger than n_2 for the total internal reflection condition. In order to guide light rays effectively through a fibre, $\Delta \ll 1$. Typically, Δ is of the order of 0.01.

10.6 NUMERICAL APERTURE

The main function of an optical fibre is to accept and transmit as much light from the source as possible. The light gathering ability of a fibre depends on two factors, namely core size and the numerical aperture. The acceptance angle and the fractional refractive index change determine the numerical aperture of fibre.

The numerical aperture (NA) is defined as the sine of the acceptance angle. Thus,

$$NA = \sin \theta_0$$

where θ_0 is the acceptance angle.

But

$$\sin \theta_0 = \sqrt{n_1^2 - n_2^2}$$

∴

$$NA = \sqrt{n_1^2 - n_2^2} \quad (10.14)$$

$$n_1^2 - n_2^2 = (n_1 + n_2)(n_1 - n_2) = \left(\frac{n_1 + n_2}{2} \right) \left(\frac{n_1 - n_2}{n_1} \right) 2n_1$$

Approximating $\frac{n_1 + n_2}{2} \approx n_1$, we can express the above relation as $(n_1^2 - n_2^2) = 2n_1^2 \Delta$. It gives

$$NA = \sqrt{2n_1^2 \Delta}$$

∴

$$NA = n_1 \sqrt{2\Delta} \quad (10.15)$$

Numerical aperture determines the light gathering ability of the fibre. It is a measure of the amount of light that can be accepted by a fibre. It is seen from equ. (10.14) that NA is dependent only on the refractive indices of the core and cladding materials and does not depend on the physical dimensions of the fibre. The value of NA ranges from 0.13 to 0.50. A large NA implies that a fibre will accept large amount of light from the source (see Fig. 10.10).



Fig. 10.10: Illustration of the propagation of light through low and high numerical aperture fibres.

Example 10.7: Calculate the fractional index change for a given optical fibre if the refractive indices of the core and the cladding are 1.563 and 1.498 respectively.

Solution: Fractional index change $\Delta = \frac{n_1 - n_2}{n_1} = \frac{1.563 - 1.498}{1.563} = \frac{0.065}{1.563} = 0.0415$

Example 10.8: Calculate the refractive indices of the core and the cladding material of a fiber from the following data:

$$\text{Numerical aperture (NA)} = 0.22 \text{ and } \Delta = 0.012$$

where Δ is the fractional refractive index change.

Solution:

$$NA = n_1 \sqrt{2\Delta}$$

$$0.22 = n_1 \sqrt{2 \times 0.012} = 0.155 n_1.$$

$$\therefore n_1 = \frac{0.22}{0.155} = 1.42$$

$$\Delta = \frac{n_1 - n_2}{n_1} \quad \therefore \frac{1.42 - n_2}{1.42} = 0.012$$

$$\therefore n_2 = 1.42 - 0.012 = 1.408$$

Example 10.9: Find the fractional refractive index and numerical aperture for an optical fibre with refractive indices of core and cladding as 1.5 and 1.49 respectively.

Solution:

$$\Delta = \frac{n_1 - n_2}{n_1} = \frac{1.5 - 1.49}{1.5} = 0.0067$$

$$NA = n_1 \sqrt{2\Delta} = 1.5 \sqrt{2 \times 0.0067} = 0.174$$

10.7 SKIP DISTANCE AND NUMBER OF TOTAL INTERNAL REFLECTIONS

We shall now calculate the number of total internal reflections that a light ray undergoes as it travels through an optical fibre of length L . Let a be the radius of the fibre. Let a light ray be incident on one end of the fibre at an angle θ_1 to the axis and refract into

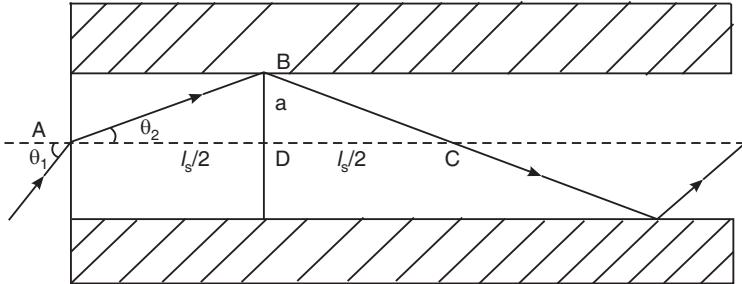


Fig. 10.11: Skip Distance l_s

the fibre at an angle θ_2 , as shown in Fig. 10.11. The ray undergoes the first reflection at B. The distance $AC = l_s$ is known as the **skip distance** and represents the distance between two successive reflections of the ray. In the $\Delta^{le}ABD$, $AD = \frac{1}{2} AC = \frac{l_s}{2}$, $BD = a$ and $\angle BAD = \theta_2$.

$$\therefore \tan \theta_2 = \frac{BD}{AD} = \frac{a}{l_s/2}$$

$$\text{or } l_s = \frac{2a}{\tan \theta_2} \quad (10.16)$$

In terms of the incidence angle θ_1 , the above equation may be rewritten as

$$l_s = 2a \left[\frac{\cos \theta_2}{\sin \theta_2} \right] \quad \text{or} \quad l_s^2 = (2a)^2 \left[\frac{1}{\sin^2 \theta_2} - 1 \right]$$

Using equ.(10.7) into the above expression, we obtain

$$l_s^2 = (2a)^2 \left[\left(\frac{n_1}{n_0 \sin \theta_1} \right)^2 - 1 \right]$$

$$\text{or } l_s = 2a \left[\left(\frac{n_1}{n_0 \sin \theta_1} \right)^2 - 1 \right]^{1/2} \quad (10.17)$$

In case of air, $n_0 = 1$ and $l_s = 2a \left[\left(\frac{n_1}{\sin \theta_1} \right)^2 - 1 \right]^{1/2}$ (10.17a)

The number of total internal reflections in the total fibre length is given by

$$\begin{aligned} N &= \frac{\text{Total length of the cable}}{\text{The distance travelled during one reflection}} = \frac{L}{l_s} \\ \therefore N &= \frac{L \tan \theta_2}{2a} \end{aligned} \quad (10.18a)$$

Also,

$$N = \frac{L}{2a \left[\left(\frac{n_1}{\sin \theta_1} \right)^2 - 1 \right]^{1/2}} \quad (10.18b)$$

For example, if $n_1 = 1.50$, $\theta_1 = 30^\circ$ and $a = 25 \mu\text{m}$, equ. (10.17a) gives $l_s = 141 \mu\text{m}$. Alternately, computing the value of θ_2 and using it in equ. (10.16), we obtain $l_s = 141 \mu\text{m}$. Therefore, if light travels through a length of 1 m of the optical fibre of the above specifications, it is reflected 7092 times. The same result is obtained using eq. (10.18b) or (10.18a).

Example 10.10: An optical fibre is 2m long and has a diameter of 20 μm . If a ray of light is incident on one end of the fibre at an angle of 40° , how many reflections does it undergo before emerging from the other end? Refractive index of fibre is 1.3.

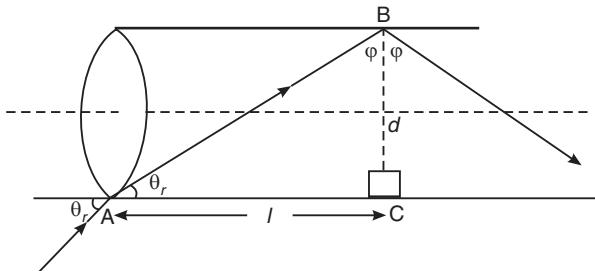


Fig. i

$$\frac{\sin \theta_r}{\sin \theta_i} = \frac{n_1}{n_2}$$

$$\therefore \sin \theta_r = \frac{n_1}{n_2} \sin \theta_i = \frac{(1)(\sin 40^\circ)}{1.3} = 0.4940$$

$$\therefore \theta_r = 29.6^\circ$$

From Fig. i, the distance l travelled along the fibre during one internal reflection is

$$l = \frac{d}{\tan \theta_r} = \frac{20 \mu\text{m}}{\tan 29.6^\circ} = \frac{2 \times 10^{-5} \text{m}}{0.568} = 3.52 \times 10^{-5} \text{m}$$

Number of reflections in the cable is

$$N = \frac{\text{Total length of the cable}}{l} = \frac{2 \text{m}}{3.52 \times 10^{-5} \text{m}} = 56818$$

10.8 MODES OF PROPAGATION

Light propagates as an electromagnetic wave through an optical fibre and its propagation is governed by Maxwell's equations. Complete understanding of propagation of light waves through optical fibres requires a thorough understanding of solution of these equations in the context of optical fibres. When a plane electromagnetic wave propagates in free space, it

travels as a transverse electromagnetic wave. The electric field and magnetic field components associated with the wave are perpendicular to each other and also perpendicular to the direction of propagation. It is known as a TEM wave. When the light ray is guided through an optical fibre, it propagates in different types of modes. Each of these guided modes consists of a variety of electromagnetic field configurations, such as transverse electric (TE), transverse magnetic (TM) and hybrid modes. Hybrid modes are combination of transverse electric and magnetic modes.

In simple terms these *modes can be visualised as the possible number of allowed paths of light* in an optical fibre (see Fig. 10.6). The paths are all zigzag paths excepting the axial direction. Though the rays having propagation angles between $\theta = 0^\circ$ and $\theta = \theta_c$ will be in a position to undergo total internal reflections, all of them will not however propagate along the optical fibre. Only a certain ray directions are allowed. As a zigzag ray gets repeatedly reflected at the walls of the fibre, phase shift occurs. Consequently, the waves travelling along certain zigzag paths will be in phase and undergo constructive interference, while the waves coursing along certain other paths will be out of phase and diminish due to destructive interference. The light ray paths along which the waves are in phase inside the fibre are known as **modes**. Each mode is a pattern of electric and magnetic field distributions that is repeated along the fibre at equal intervals. The number of modes propagating in a fibre increases as θ_c or Δ increases. Increasing the core refractive index increases the number of propagating modes. On the other hand, increasing the clad refractive index decreases the number of propagating modes. The number of modes that a fibre will support depends on the ratio d/λ , where d is the diameter of the core and λ is the wavelength of the wave being transmitted. The zero order ray travels along the axis is known as the *axial ray*.

Note that each mode carries a portion of the light from the input signal.

Types of modes:

In a fibre of fixed thickness, the modes that propagate at angles close to the critical angle ϕ_c (i.e., critical propagation angle θ_c) are **higher order modes**, and modes that propagate with angles larger than the critical angle (i.e., lower than the critical propagation angle) are **lower order modes** (see Fig. 10.12). In case of lower order modes, the fields are concentrated near the center of the fibre. In case of higher order modes, the fields are distributed more towards the edge of the wave-guide and tend to send light energy into the cladding. This energy is lost ultimately. The higher order modes have to traverse longer paths and hence take larger time than the lower order modes to cover a given length of the fibre. Thus, the higher order modes arrive at the output end of the fiber later than the lower order modes.

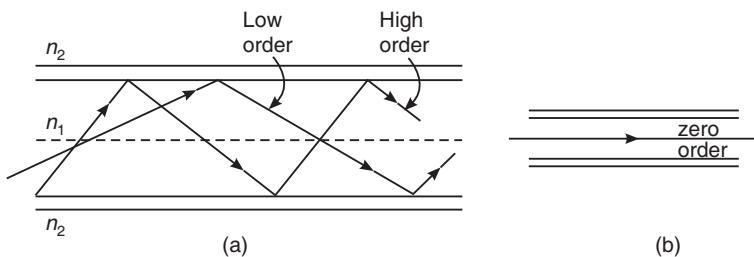


Fig. 10.12: (a) Low and High-order ray paths in a multimode fibre.
(b) Axial ray in a single mode fibre

10.9 TYPES OF RAYS

The rays that propagate through an optical fibre can be classified into two categories: (i) meridional rays and (ii) skew rays.

1. **Meridional ray:** A ray that propagates through the fibre undergoing total internal reflection is called meridional ray. It passes through the longitudinal axis of the fibre core (Fig. 10.13 a). The propagation of meridional rays is possible only in the TM or TE modes.
2. **Skew ray:** The ray that describes angular helical path along the fibre is called a skew ray (see Fig. 10.13 b). These rays do not pass through the axis of the core. These rays are propagated in either hybrid EH or HE modes. Some of these modes produce losses through leakage of radiation. In real situations, the skew rays constitute a substantial portion of the total number of guided rays. They tend to propagate only in the annular region near the outer surface of the core and do not fully utilize the core as the medium. However, they are complementary to the meridional rays and increase the light gathering capacity of the fibre.

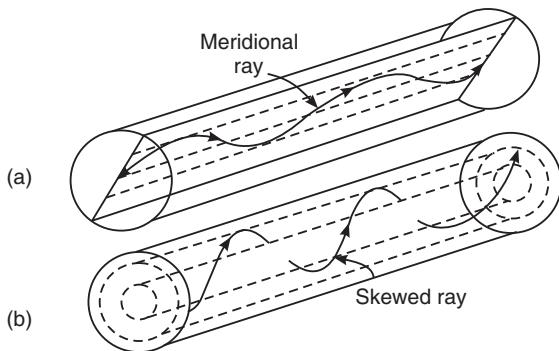
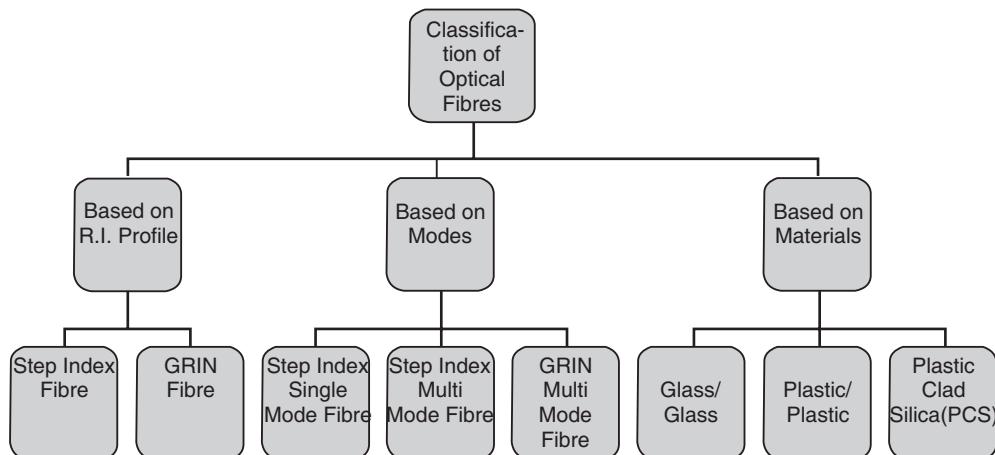


Fig. 10.13

10.10 CLASSIFICATION OF OPTICAL FIBRES

Optical fibres are classified as follows into various types basing on different parameters.



A. Classification basing on refractive index profile:

Refractive index profile of an optical fibre is a plot of refractive index drawn on one of the axes and the distance from the core axis drawn on the other axis (see Fig. 10.14). Optical fibres are classified into the following two categories on the basis of refractive index profile.

1. Step index fibres and 2. Graded index (GRIN) fibres.

Step index refers to the fact that the refractive index of the core is constant along the radial direction and abruptly falls to a lower value at the cladding and core boundary (see Fig. 10.14a). In case of GRIN fibres, the refractive index of the core is not constant but varies smoothly over the diameter of the core (see Fig. 10.14b). It has a maximum value at the center

and decreases gradually towards the outer edge of the core. At the core-cladding interface the refractive index of the core matches with the refractive index of the cladding. The refractive index of the cladding is constant.

B. Classification basing on the modes of light propagation:

On the basis of the modes of light propagation, optical fibres are classified into two categories as

1. Single mode fibres (SMF) and 2. Multimode fibres (MMF).

A **single mode fibre** (SMF) has a smaller core diameter and can support only one mode of propagation. On the other hand, a **multimode fibre** (MMF) has a larger core diameter and supports a number of modes.

Thus, on the whole, the optical fibres are classified into three types:

- Single mode step-index fibre (SMF)
- Multimode step-index fibre (MMF)
- Graded index (multimode) fibre (GRIN).

C. Classification basing on materials:

On the basis of materials used for core and cladding, optical fibres are classified into three categories.

1. Glass/glass fibres (glass core with glass cladding)
2. Plastic/plastic fibres (plastic core with plastic cladding)
3. PCS fibres (polymer clad silica)

10.11 THE THREE TYPES OF FIBRES

We now study the detailed structure and characteristics of the three types of optical fibres, mentioned above in Art. 10.10.B.

10.11.1 Single Mode Step Index Fibre

Structure

A single mode step index fibre has a very fine thin core of diameter of 8 μm to 12 μm (see Fig. 10.15 c). It is usually made of germanium doped silicon. The core is surrounded by a thick cladding of lower refractive index. The

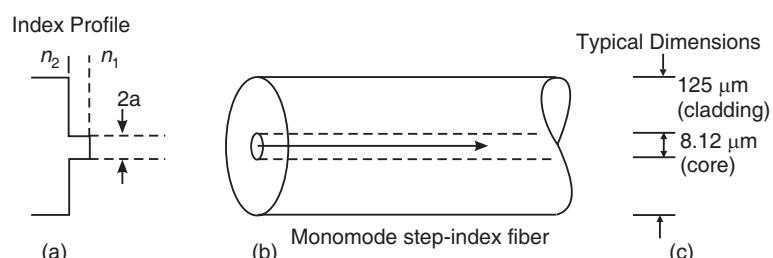


Fig. 10.15: Single mode step index fibre (a) R.I. profile (b) ray paths (c) typical dimensions

cladding is composed of silica lightly doped with phosphorous oxide. The external diameter of the cladding is of the order of 125 μm . The fibre is surrounded by an opaque protective

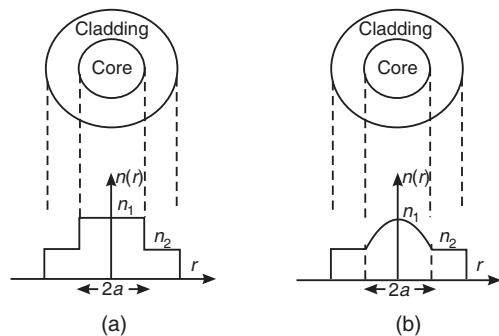


Fig. 10.14: Classification of optical fibres based on R.I. profile (a) Step index fibre (b) GRIN fibre

sheath. The refractive index of the fibre changes abruptly at the core-cladding boundary, as shown in Fig. 10.15 (a). The variation of the refractive index of a step index fibre as a function of radial distance can be mathematically represented as

$$\begin{aligned} n(r) &= n_1 [r < a \text{ inside core}] \\ &= n_2 [r > a \text{ in cladding}] \end{aligned} \quad (10.19)$$

Propagation of light in SMF

Light travels in SMF along a single path that is along the axis (Fig. 10.15 b). Obviously, it is the zero order mode that is supported by a SMF. Both Δ and NA are very small for single mode fibres. This relatively small value is obtained by reducing the fibre radius and by making Δ , the relative refractive index change, to be small. The low NA means a low acceptance angle. Therefore, light coupling into the fibre becomes difficult. Costly laser diodes are needed to launch light into the SMF.

10.11.2 Multimode Step Index Fibre

Structure

A multimode step index fibre is very much similar to the single mode step index fibre except that its core is of larger diameter. The core diameter is of the order of 50 to 100 μm , which is very large compared to the wavelength of light. The external diameter of cladding is about 150 to 250 μm (Fig. 10.16c).

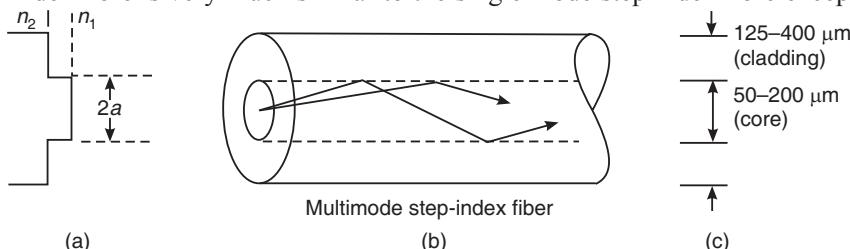


Fig. 10.16: Multimode step index fibre (a) R.I. Profile (b) Ray paths (c) Typical dimensions

Propagation of light in MMF

Multimode step index fibres allow finite number of guided modes. The direction of polarization, alignment of electric and magnetic fields will be different in rays of different modes. In other words, many zigzag paths of propagation are permitted in a MMF. The path length along the axis of the fibre is shorter while the other zigzag paths are longer. Because of this difference, the lower order modes reach the end of the fibre earlier while the high order modes reach after some time delay (Fig. 10.16 b).

10.11.3 Graded Index (GRIN) Fibre

A graded index fibre is a multimode fibre with a core consisting of concentric layers of different refractive indices.

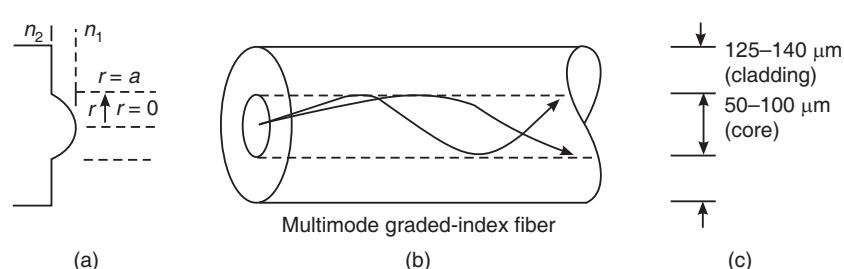


Fig. 10.17: GRIN fibre (a) R.I. Profile (b) Ray paths (c) Typical dimensions

Therefore, the refractive index of the core varies with distance from the fibre axis. It has a high value at the centre and falls off with increasing radial distance from the axis. A typical structure and its index profile are shown in Fig. 10.17 (a). Such a profile causes a periodic focussing of light propagating through the fibre. The size of the graded index fibre is about the same as the step index fibre. The variation of the refractive index of the core with radius measured from the center is given by

$$n(r) = \begin{cases} n_1 \sqrt{1 - \left[2\Delta \left(\frac{r}{a} \right)^\alpha \right]}, & r < a \text{ inside core} \\ n_2, & r > a \text{ in cladding} \end{cases} \quad (10.20)$$

where n_1 is maximum refractive index at the core axis, a the core radius, and α the grading profile index number which varies from 1 to ∞ . When $\alpha = 2$, the index profile is parabolic and is preferred for different applications.

Propagation of light in GRIN fibre

As a light ray goes from a region of higher refractive index to a region of lower refractive index, it is bent away from the normal. The process continues till the condition for total internal reflection is met. Then the ray travels back towards the core axis, again being continuously refracted (Fig. 10.18 a).

The turning around may take place even before reaching the core-cladding interface. Thus, continuous refraction is followed by total internal reflection and again continuous refraction towards the axis. In the graded index

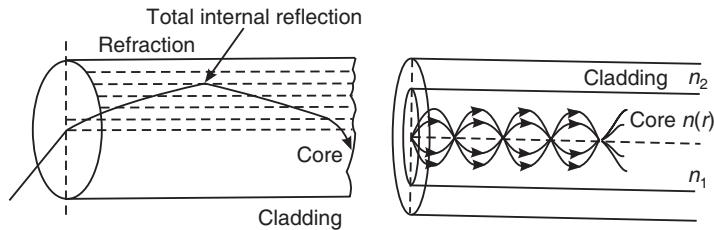


Fig. 10.18: (a) An expanded ray diagram showing refraction at the various high to low index interfaces within graded index fibre, giving an overall curved ray path. (b) Light transmission in a graded index fibre.

fibre, rays making larger angles with the axis traverse longer path but they travel in a region of lower refractive index and hence at a higher speed of propagation. Consequently, all rays traveling through the fibre, irrespective of their modes of travel, will have almost the same optical path length and reach the output end of the fibre at the same time (see Fig. 10.18b).

In case of GRIN fibres, the acceptance angle and numerical aperture decrease with radial distance from the axis. The numerical aperture of a graded index fibre is given by

$$\begin{aligned} NA &= \sqrt{n^2(r) - n_2^2} \approx n_1(2\Delta)^{\frac{1}{2}} \sqrt{1 - \left(\frac{r}{a} \right)^2} \\ &= n_1 \sqrt{2\Delta \left[1 - \left(\frac{r}{a} \right)^2 \right]} \end{aligned} \quad (10.21)$$

10.12 MATERIALS

Optical fibres are fabricated from glass or plastic which are transparent to optical frequencies. Step index fibres are produced in three common forms – (i) a glass core cladded with a glass having a slightly lower refractive index, (ii) a silica glass core cladded with plastic and (iii) a plastic core cladded with another plastic. Generally, the refractive index step is the smallest

for all glass fibres, a little larger for the plastic clad silica (PCS) fibres and the largest for all plastic construction.

10.12.1 All glass fibres

The basic material for fabrication of optical fibres is silica (SiO_2). It has a refractive index of 1.458 at $\lambda = 850 \text{ nm}$. Materials having slightly different refractive index are obtained by doping the basic silica material with small quantities of various oxides. If the basic silica material is doped with germania (GeO_2) or phosphorous pentoxide (P_2O_5), the refractive index of the material increases. Such materials are used as core materials and pure silica is used as cladding material in these cases. When pure silica is doped with boria (B_2O_3) or fluorine, its refractive index decreases. These materials are used for cladding when pure silica is used as core material. Examples of fiber compositions are

- SiO_2 core – $\text{B}_2\text{O}_3.\text{SiO}_2$ cladding
- $\text{GeO}_2.\text{SiO}_2$ core – SiO_2 cladding

The glass optical fibres exhibit very low losses and are used in long distance communications.

10.12.2 All plastic fibres

In these fibres, perspex (PMMA) and polystyrene are used for core. Their refractive indices are 1.49 and 1.59 respectively. A fluorocarbon polymer or a silicone resin is used as a cladding material. A high refractive index difference is achieved between the core and the cladding materials. Therefore, plastic fibres have large NA of the order of 0.6 and large acceptance angles up to 70° . The main advantages of the plastic fibres are low cost and higher mechanical flexibility. The mechanical flexibility allows the plastic fibres to have large cores, of diameters ranging from 110 to $1400 \mu\text{m}$. However, they are temperature sensitive and exhibit very high loss. Therefore, they are used in low cost applications and at ordinary temperatures (below 80°C). Examples of plastic fiber compositions are

- | | | |
|----------------------------------|--------------|-----------|
| • Polystyrene core | $n_1 = 1.60$ | NA = 0.60 |
| - Methyl methacrylate cladding | $n_2 = 1.49$ | |
| • Polymethyl methacrylate core | $n_1 = 1.49$ | NA = 0.50 |
| - cladding made of its copolymer | $n_2 = 1.40$ | |

10.12.3 PCS fibres

The plastic clad silica (PCS) fibres are composed of silica cores surrounded by a low refractive index transparent polymer as cladding. The core is made from high purity quartz. The cladding is made of a silicone resin having a refractive index of 1.405 or of perfluorinated ethylene propylene (Teflon) having a refractive index of 1.338. Plastic claddings are used for step-index fibres only. The PCS fibres are less expensive but have high losses. Therefore, they are mainly used in short distance applications.

10.13 V-NUMBER

Let us consider a narrow beam of monochromatic light launched on the front end of a step-index fibre, at an angle less than the acceptance angle of the fibre. Let the wavelength of the light be λ_0 and the diameter of the fiber be d . It appears to us from the ray concept that all the rays contained in the beam propagate along the fibre, such that there can be infinite modes of propagation. However, in practice, only a limited number of modes of propagation are possible in an optical fibre. To understand the reason for this behaviour, we have to recall that phase changes occur as the light waves travel forward. The phase shift takes place due to two reasons – (i) due to optical path length traversed and (ii) due to total internal reflection at the core-cladding interface.

- (i) When a wave travels a distance l in a medium of refractive index n_1 , it undergoes a phase change δ_1 given by

$$\delta_1 = k n_1 l = \frac{2\pi l n_1}{\lambda} \quad (10.22)$$

where k is the propagation constant.

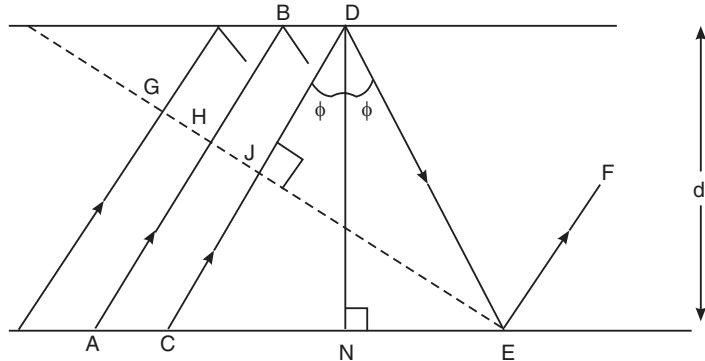


Fig. 10.19

- (ii) Whenever a wave with component normal to the reflecting surface undergoes total internal reflection, the phase shift, δ_2 , is given by

$$\delta_2 = 2 \tan^{-1} \frac{\sqrt{n_1^2 \cos^2 \phi - n_2^2}}{n_1 \sin \phi} \quad (10.23)$$

In Fig. 10.19, AB and CD are parallel rays in an incident beam. The line GJ is perpendicular to the propagation path of the rays AB, CD and hence represents a plane wavefront. The points G and J lying on the same wavefront will be in phase with each other. As the point E, which is on the reflected ray DE, lies on the wavefront GJ, the points J and E must be in phase with each other. However, moving from the point J to E along the ray, we find that there occurs a phase shift given by

$$\delta = (JD + DE) \frac{2\pi n_1}{\lambda_0} - 2\delta_2 \quad (10.24)$$

The factor 2 in the above equation takes into account the two total internal reflections at D and E. In the Δ^{le} DNE

$$\frac{DN}{DE} = \cos \phi. \text{ Therefore, } DE = \frac{DN}{\cos \phi} = \frac{d}{\cos \phi}$$

Further, in the Δ^{le} JDE

$$\frac{JD}{DE} = \cos 2\phi. \text{ Therefore, } JD = DE \cos 2\phi$$

$$JD + DE = DE (1 + \cos 2\phi) = 2 DE \cos^2 \phi.$$

$$\text{Therefore, } JD + DE = 2 \frac{d}{\cos \phi} \cos^2 \phi = 2d \cos \phi$$

Using the above expression into equ. (10.24), we obtain

$$\delta = \frac{4d \pi n_1 \cos \phi}{\lambda_0} - 2\delta_2$$

Now the condition for the wave associated with the ray CD to propagate along the optical fibre is that the phase of the twice reflected wave must be the same as that of the incident wave. That is, the wave must interfere constructively with itself. If this phase condition is not satisfied, the wave would interfere destructively with itself and just die out. It means that the total phase shift must be equal to an integer multiple of 2π radians. Thus,

$$\frac{4d\pi n_1 \cos \phi}{\lambda_0} - 2\delta_2 = 2\pi m$$

or

$$m = \frac{2d n_1 \cos \phi_m}{\lambda_0} - \frac{\delta_2}{\pi} \quad (10.25)$$

where m is an integer that determines the allowed ray angles for propagation of the wave and ϕ_m is the value of ϕ corresponding to a particular value of m . In order to sustain total internal reflection,

$$\begin{aligned} \sin \phi_m &\geq \frac{n_2}{n_1} \\ \therefore \cos \phi_m &\leq \frac{\sqrt{n_1^2 - n_2^2}}{n_1} \\ \therefore m &\leq \frac{2d \sqrt{n_1^2 - n_2^2}}{\lambda_0} - \frac{\delta_2}{\pi} \end{aligned} \quad (10.26)$$

or

$$m \leq \frac{2V}{\pi} - \frac{\delta_2}{\pi} \quad (10.27)$$

where V is given by

$$V = \frac{\pi d}{\lambda_0} \sqrt{n_1^2 - n_2^2} \quad (10.28)$$

V -number is more generally called normalized **frequency** of the fibre. Each mode has a definite value of V -number below which the mode is cut off. Equ. (10.28) can be written as

$$V = \frac{\pi d}{\lambda_0} (NA) \quad (10.29)$$

or

$$V = \frac{\pi d}{\lambda_0} n_1 \sqrt{2\Delta} \quad (10.30)$$

The maximum number of modes N_m supported by an SI fibre is given by

$$N_m = \frac{1}{2} V^2 \quad (10.31)$$

Thus, for $V=10$, N_m is 50. When the normalized frequency V is less than 2.405, the fibre can support only one mode, which propagates along the axial length of the fibre, and the fibre becomes a single mode fibre. It means that for single mode transmission in a MMF, V must be less than 2.405. The wavelength at which the fibre becomes single mode is called **cutoff wavelength**, λ_c of the fibre. Using equ. (10.29), we can write

$$\lambda_c = \frac{\pi d}{2.405} (NA) \quad (10.32)$$

It is seen from the above equation that single mode property can be realized in a multimode fibre by decreasing the core diameter and/or decreasing Δ such that $V < 2.405$.

In case of GRIN fibres, for larger values of V ,

$$N_m \equiv \frac{V^2}{4} \quad (10.33)$$

Example 10.11: A step-index fibre is made with a core of refractive index 1.52, a diameter of 29 μm and a fractional difference index of 0.0007. It is operated at a wavelength of 1.3 μm . Find the V-number and the number of modes that the fibre will support.

Solution: $V = \frac{\pi d}{\lambda_0} n_1 \sqrt{2\Delta} = \frac{3.143 \times 29 \times 10^{-6} \text{ m}}{1.3 \times 10^{-6} \text{ m}} \times 1.52 \sqrt{2 \times 0.0007} = 4.049$

$$\therefore \text{Number of modes, } N = \frac{1}{2} V^2 = \frac{1}{2} (4.049)^2 = 8 \text{ modes}$$

10.14 FABRICATION

A number of techniques are available to produce all glass fibres. In one of the methods, known as the double crucible method, fibres are directly produced from the melt.

Double Crucible Technique

The double crucible consists of two concentric platinum crucibles having thin orifices at the bottom. Raw material for the core-glass is placed in the inner crucible and the raw material for cladding is fed to the outer crucible. The double crucible arrangement is mounted vertically (Fig. 10.20) in a furnace. The furnace is maintained at a suitable temperature to take the raw material into molten state. The

fibres are drawn through the thin orifices at the bottom of the crucibles. As both the materials are drawn simultaneously, a filament of core glass surrounded by a tube of cladding glass is obtained in the process. The thickness of the fibre is monitored and the fibre is then coated with a polymer. Subsequently, it is passed through a plastic extrusion die to form a plastic sheath over the fibre.

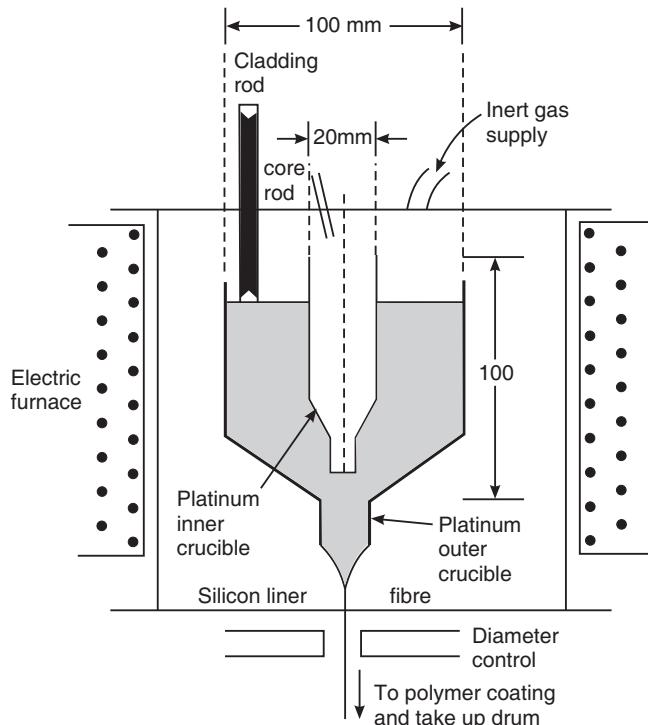


Fig. 10.20

10.15 SPLICING

It is often required to join two optical fibres together to form a continuous optical waveguide. The method and technique for connecting the fibres depends on whether a permanent joint is required or easily disconnected joint is required. The permanent bonding technique is called **splice** technique and easily disconnected joint techniques are called **connectors**. Splicing is analogous to soldering in metal wires. It consists of fusing of two fibre ends and bonding

them together in an alignment structure. The generally accepted splicing method is arc fusion splicing, which melts the fiber ends together with an electric arc. For quicker fastening jobs, a “mechanical splice” is used.

1. Fusion splicing

Fusion splicing is the act of joining two optical fibres end-to-end using heat. The goal is to fuse the two fibers together in such a way that light passing through the fibers is not scattered or reflected back by the splice, and so that the splice and the region surrounding it are almost as strong as the virgin fiber itself.

Fusion splicing is done with a specialized instrument that typically operates as follows: The two cable ends are fastened inside a splice enclosure that will protect the splices, and the fiber ends are stripped of their protective polymer coating (as well as the sturdier outer jacket, if present). The ends are *cleaved* (cut) with a precision cleaver to make them perpendicular, and are placed into special holders in the splicer. The splice is usually inspected via a magnified viewing screen to check the cleaves before and after the splice. The splicer uses small motors to align the end faces together, and emits a small spark between electrodes at the gap to burn off dust and moisture. Then the splicer generates a larger spark that raises the temperature above the melting point of the glass, fusing the ends together permanently. The splices offer controlled alignment of fiber optic cables to achieve losses as low as 0.05 dB.

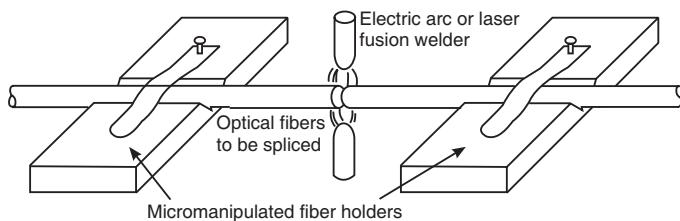


Fig. 10.21

2. Mechanical splicing

A **mechanical splice** is a junction of two or more optical fibres that are aligned and held in place by a self-contained assembly. The fibers are not permanently joined, just precisely held together so that light can pass from one to another. They are easily applied in the field, require little or no tooling and offer losses of about 0.2 dB.

Mechanical fiber splices are designed to be quicker and easier to install, but there is still the need for stripping, careful cleaning and precision cleaving. The fiber ends are aligned and held together by a precision-made sleeve, often using a clear index-matching gel that enhances the transmission of light across the joint. Such joints typically have higher optical loss and are less robust than fusion splices, especially if the gel is used. All splicing techniques involve the use of an enclosure into which the splice is placed for protection afterward.

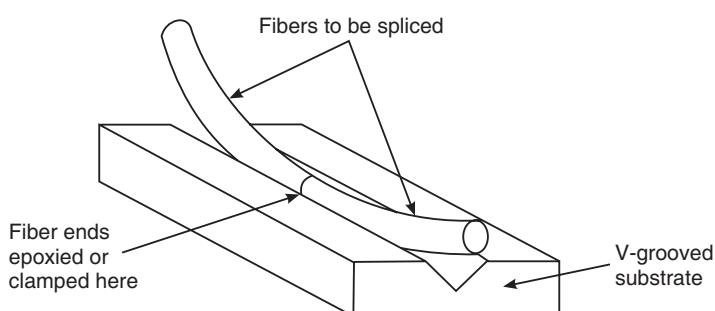


Fig. 10.22

(i) V-groove splice technique

The V-block is the simplest mechanical splice. The bared fibres to be joined are placed in the groove. Angular alignment is particularly well controlled. The two fibres can slide in the groove until they touch. They are then epoxied permanently into position, so end-separation errors are minimal. If the epoxy is index matched to the fibre, even small gaps can be tolerated with little loss. Lateral misalignment would be negligible in the groove if both fibres had the same core and cladding diameters. A cover plate can be placed over the V-block to protect the splice further.

(ii) Elastomer splice technique

Another splice is essentially a precision sleeve made with elastomeric materials. The elastomer is an elastic material usually made into a cylinder with an opening along its axis. The groove is a

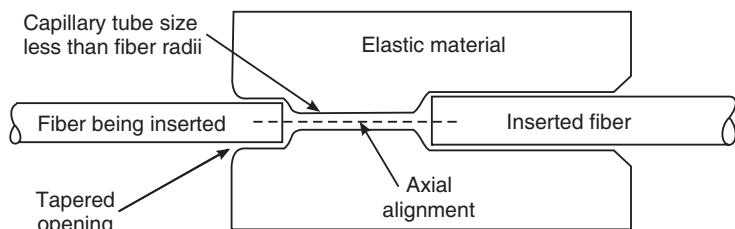


Fig. 10.23

little smaller than the fibre but accepts and centers it by expanding slightly when the fibre is inserted. The fibres are inserted from both the ends of the cylinder and touch near its midpoint. The slice can be epoxied for permanent connection. An external splice holder is used for full protection of the splice.

10.16 LOSSES IN OPTICAL FIBRE

As a light signal propagates through a fibre, it suffers loss of amplitude and change in shape. The loss of amplitude is referred to as *attenuation* and the change in shape as *distortion*.

10.16.1 Attenuation

When an optical signal propagates through a fibre its power decreases exponentially with distance. The loss of optical power as light travels down a fiber is known as **attenuation**. The attenuation of optical signal is defined as *the ratio of the optical output power from a fibre of length L to the input optical power*. If P_i is the optical power launched at the input end of the fibre, then the power P_o at a distance L down the fibre is given by

$$P_o = P_i e^{-\alpha L} \quad (10.34)$$

where α is called the **fibre attenuation coefficient** expressed in units of km^{-1} . Taking logarithms on both the sides of the above equation, we obtain

$$\alpha = \frac{1}{L} \ln \frac{P_i}{P_o} \quad (10.35)$$

In units of dB / km , α is defined through the equation

$$\therefore \alpha_{\text{dB/km}} = \frac{10}{L} \log \frac{P_i}{P_o} \quad (10.36)$$

In case of an ideal fibre, $P_o = P_i$ and the attenuation would be zero.

Different Mechanisms of Attenuation

There are several loss mechanisms responsible for attenuation in optical fibres. They are broadly divided into two categories: *intrinsic* and *extrinsic* attenuation. Intrinsic attenuation

is caused by substances inherently present in the fiber, whereas extrinsic attenuation is caused by external forces such as bending.

A. Intrinsic Attenuation

Intrinsic attenuation results from materials inherent to the fiber. It is caused by impurities present in the glass. During manufacturing, there is no way to eliminate all impurities. When a light signal hits an impurity in the fiber, either it is scattered or it is absorbed. Intrinsic attenuation can be further characterized by two components:

- **Material absorption**
- **Rayleigh scattering**

Absorption by material: Material absorption occurs as a result of the imperfection and impurities in the fiber and accounts for 3-5% of fiber attenuation. The most common impurity is the hydroxyl (OH^-) molecule, which remains as a residue despite stringent manufacturing techniques. These radicals result from the presence of water remnants that enter the fiber-optic cable material through either a chemical reaction in the manufacturing process or as humidity in the environment. The natural impurities in the glass absorb light signal, and convert it into vibrational energy or some other form of energy. Hydroxyl radical ions (OH), and transition metals such as copper, nickel, chromium, vanadium and manganese have electronic absorption in and near visible part of the spectrum. Their presence causes heavy losses.

Even a highly pure glass absorbs light in specific wavelength regions. Strong electronic absorption occurs at UV wavelengths, while vibrational absorption occurs at IR wavelengths.

Losses due to impurities can be reduced by better manufacturing processes. In improved fibres, metal ions are practically negligible. The largest loss is caused by OH ions. These cannot be sufficiently reduced. The absorption of light either through intrinsic or impurity process constitutes a transmission loss because that much energy is subtracted from the light propagating through the fibre. The absorption losses are found to be at minimum at around $1.3 \mu\text{m}$.

Unlike scattering, absorption can be limited by controlling the amount of impurities during the manufacturing process.

Rayleigh Scattering: Rayleigh scattering accounts for the majority (about 96%) of attenuation in optical fiber. The local microscopic density variations in glass cause local variations in refractive index. These variations, which are inherent in the manufacturing process and cannot be eliminated, act as obstructions and scatter light in all directions (Fig. 10.24). This is known as *Rayleigh scattering*. The Rayleigh scattering loss greatly depends on the wavelength. It varies as $1/\lambda^4$ and becomes important at lower wavelengths. Thus, Rayleigh scattering sets a lower limit, on the wavelengths that can be transmitted by a glass fibre at $0.8 \mu\text{m}$, below which the scattering loss is very high.

Any wavelength that is below 800 nm is unusable for optical communication because attenuation due to Rayleigh scattering is high. At the same time, propagation above 1700 nm is not possible due to high losses resulting from infrared absorption.

Fig. 10.25 shows the variation of attenuation with wavelength measured for a typical fiber-optic cable.

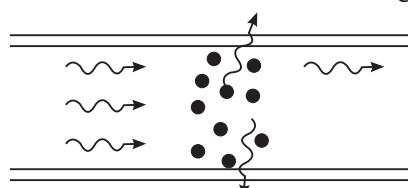


Fig. 10.24: Rayleigh scattering, showing attenuation of an incident stream of photons due to localized variations in refractive index.

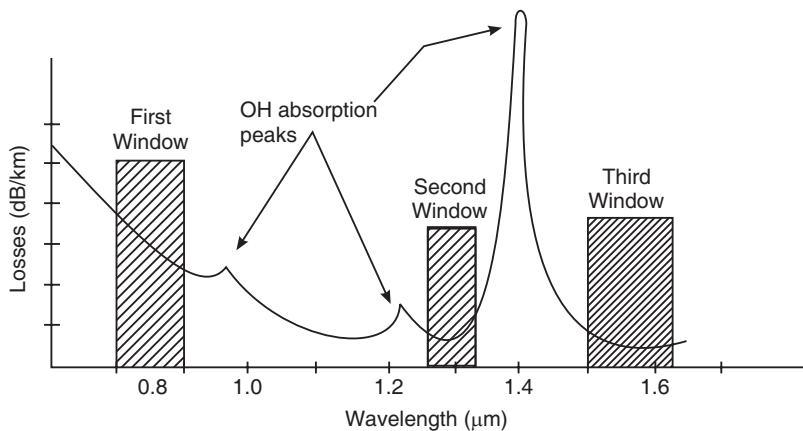


Fig. 10.25: A typical plot of fibre attenuation versus wavelength for a silica based optical fibre.

For better performance, the choice of wavelength must be based on minimizing loss and minimizing dispersion. Such windows are selected for communication purposes. It is seen from the attenuation curve that it has a minimum at around a particular band of optical wavelengths. The band of wavelengths at which the attenuation is a minimum is called **optical window** or **transmission window** or **low-loss window**. There are three **principal windows**. These correspond to wavelength regions in which attenuation is low and matched to the capability of a transmitter to generate light efficiently and a receiver to carry out detection.

λ (nm)	Approx. loss (dB/km)
820–880	2.2
1200–1320	0.6
1550–1610	0.2

From the above data it is seen that the range 1550 to 1610 nm is most preferable. From the point of view of dispersion, the low intramodal dispersion wavelength of about 1300nm is most suitable.

B. Extrinsic Attenuation or Bending losses

Extrinsic attenuation is caused by two external mechanisms: **macrobending** or **microbending**. Both of them cause a reduction of optical power. If a bend is imposed on an optical fiber, strain is placed on the fiber along the region that is bent. The bending strain affects the refractive index and the critical angle of the light ray in that specific area. As a result, the condition for total internal reflection is no longer satisfied. Hence, light traveling in the core can refract out, and loss occurs.

Macrobend losses: A **macrobend** is a large-scale bend that is visible. When a fibre is bent through a large angle, strain is placed on the fiber along the region that is bent. The bending strain will affect the refractive index and the critical angle of the light ray in that specific area. As a result, light traveling in the core can refract

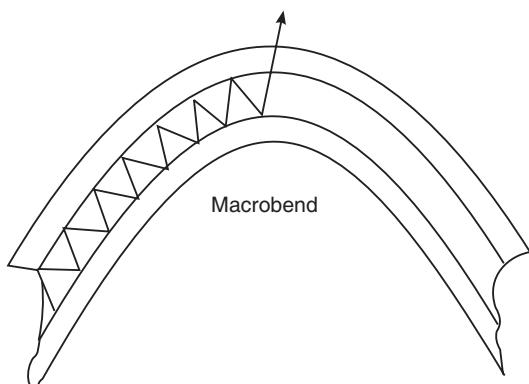


Fig. 10.26: Macrobend loss

out, and loss occurs. (Fig. 10.26). To prevent macrobends, optical fiber has a *minimum bend radius* specification that should not be exceeded. This is a restriction on how much bend a fiber can withstand before experiencing problems in optical performance or mechanical reliability.

Microbend losses

Microbend is a small-scale distortion. It is localized and generally indicative of pressure on the fiber. Microbending might be related to temperature, tensile stress, or crushing force.

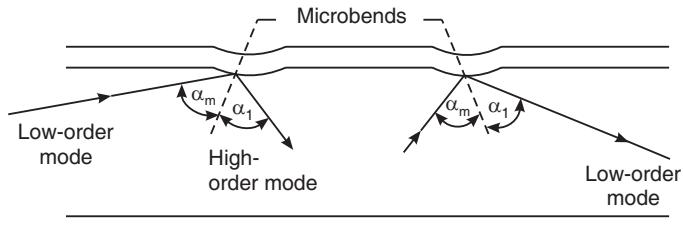


Fig. 10.27: Microbend losses

Microbending is caused by imperfections in the cylindrical geometry of fiber during the manufacturing process or installation processes. The bend may not be clearly visible upon inspection. Structural variations in the fibre, or fibre deformation, cause radiation of light away from the fibre (Fig. 10.27). Microbending may occur, for example, due to winding of optical fibre cable over spools. Light rays get scattered at the small bends and escape into the cladding. Such losses are known as microbend losses.

Example 10.12: Optical power of 1 mW is launched into an optical fibre of length 100 m. If the power emerging from the other end is 0.3 mW, calculate the fibre attenuation.

$$\text{Solution: Attenuation, } \alpha = \frac{10}{L} \log \frac{P_i}{P_o} = \frac{10}{0.1 \text{ km}} \log \frac{1 \text{ mW}}{0.3 \text{ mW}} = 52.3 \text{ dB/km}$$

Example 10.13: What is the attenuation in dB/km, if 15% of the power fed at the launching end of a $\frac{1}{2}$ km fibre is lost during propagation?

$$\text{Solution: Attenuation, } \alpha = \frac{10}{L} \log \frac{P_i}{P_o} = \frac{10}{0.5 \text{ km}} \log \frac{1}{0.15} = 16.48 \text{ dB/km}$$

10.16.2 Distortion

In an optical fibre communication system, the information (signal) is coded in the form of discrete pulses of light, which are transmitted through the fibre. The light pulses are of a given width, amplitude and interval. The number of pulses that can be sent per unit time will determine the information capacity of the fibre. More information can be sent by optical cable when distinct pulses can be transmitted in more rapid succession. The pulses travel through the transmitting medium (i.e., optical fibre) and reach the detector at the receiving end. For the information to be retrieved at the detector, it is necessary that the optical pulses are well resolved in time. However, the light pulses broaden and spread into a wider time interval because of the different times taken by different rays propagating through the fibre. This phenomenon is known as **distortion** or **pulse dispersion**. Hence, even though two pulses may be well resolved at the input end, they may overlap on each other at the output end, as shown in Fig. 10.28. It is obvious that the pulse broadening depends on the length of the travel of the pulses through the fibre. Hence, dispersion is expressed in units of **ns/km** (time/distance).

The following three different dispersion mechanisms determine the distortion of the signal in an optical fibre. They are

- Intermodal dispersion and
- Intramodal dispersion.

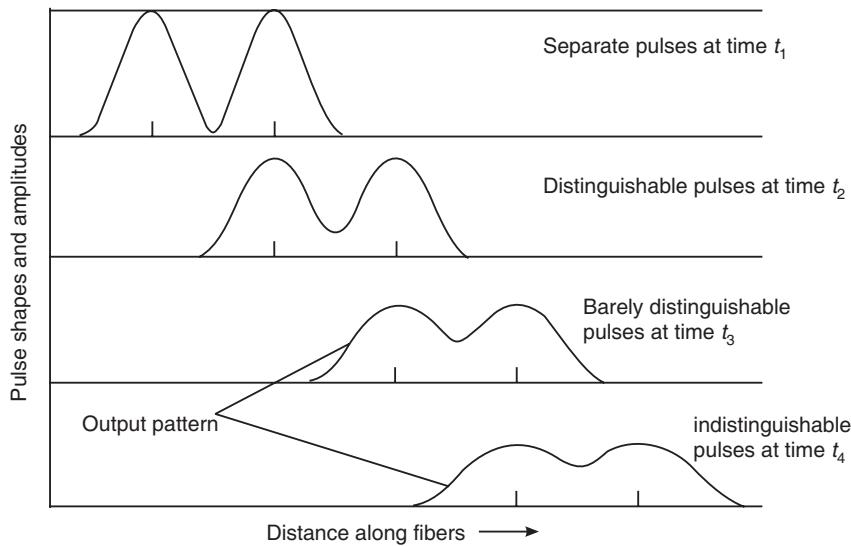


Fig. 10.28: Distortion of the pulses traveling along a fibre

Intramodal dispersion is again divided into the following two types.

- Material dispersion
- Waveguide dispersion

Intermodal Dispersion: Intermodal dispersion occurs as a result of the differences in the group velocities of the modes. For example, let us consider the propagation of a pulse through a multimode fibre. The power associated with the single pulse gets distributed into the various modes or paths guided by the fibre. The lower order modes (rays reflected at larger angles) travel a greater distance than the higher order modes (lower angle rays). The path length along the axis of the fibre is shorter while the other zigzag paths are longer. Because of this difference, the lower order modes reach the end of the fibre earlier while the high order modes reach after some time delay (Fig. 10.19). As a result, light pulses broaden as they travel down the fibre, causing signal distortion. The output pulses no longer resemble the input pulses (see Fig. 10.29). This type of distortion is known as **intermodal** or simply **modal dispersion**. This imposes limitation on the separation between successive pulses and thereby reduces the transmission rate and capacity.

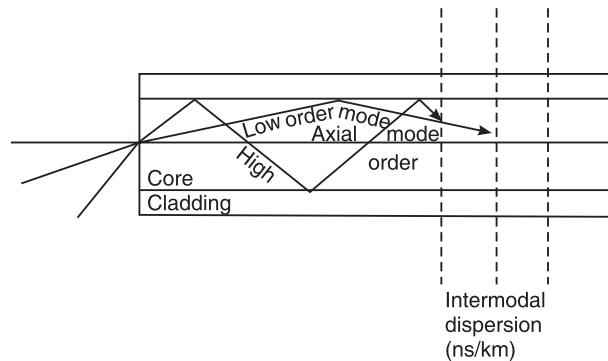


Fig. 10.29: Lower order modes reach the end of the fibre earlier while the high order modes reach after some time delay

Expression for total time delay due to modal dispersion in Step-Index fibre:

The total time delay between the arrival of the axial ray and the slowest ray, the one traveling the longest distance is

$$\Delta t = t_{\max} - t_{\min} \quad (10.37)$$

Referring to the Fig. 10.11, the time taken by a refracted ray to traverse the distance ABC of the fibre would be

$$t' = \frac{AB + BC}{v} = \frac{n_1 AC}{c \cos \theta_r} \quad (\theta_2 \text{ in Fig. 10.11 is indicated as } \theta_r \text{ here})$$

where $v = c/n_1$ is the speed of the light in the core. Since the ray path will repeat itself, the time taken by a ray to traverse a length L of the fibre is

$$t = \frac{n_1 L}{c \cos \theta_r} \quad (10.38)$$

The above relation shows that the time taken by a ray in the fibre core is a function of the angle θ_r . For the axial ray $\theta_r = 0$ and hence

$$t_{\min} = \frac{n_1 L}{c} \quad (10.39)$$

In case of the ray that travels the longest path, $\theta_r = \theta_C$. Therefore,

$$t_{\max} = \frac{n_1 L}{c \cos \theta_C}$$

Using equ. (10.5) into the above expression, we get

$$t_{\max} = \frac{n_1^2 L}{n_2 c} \quad (10.40)$$

Therefore, making use of equations (10.40) and (10.39) into (10.37), we obtain

$$\Delta t = \frac{n_1 L}{c} \left[\frac{n_1}{n_2} - 1 \right] \quad (10.41)$$

Using the equ.(10.13) for fractional refractive index change into the above equ. (10.41), we get

$$\Delta t = \frac{n_1 L}{c} \left[\frac{\Delta}{1-\Delta} \right] \quad (10.42)$$

We can also express the relation (10.41) in the following form

$$\begin{aligned} \Delta t &= \frac{n_1 L}{c} \left[\frac{n_1 - n_2}{n_2} \right] = \frac{n_1 L}{c} \left[\frac{n_1 - n_2}{n_2} \right] \left[\frac{n_1 + n_2}{n_1 + n_2} \right] \\ &= \frac{n_1 L}{c} \left[\frac{(n_1^2 - n_2^2)}{n_2(n_1 + n_2)} \right] = \frac{n_1 L}{c} \frac{(n_1^2 - n_2^2)}{2n_1 n_2} \quad (\text{as } n_1 \approx n_2) \end{aligned}$$

$$\text{or} \quad \Delta t = \frac{L}{2n_2 c} (NA)^2 \quad (10.43)$$

It is seen from the equ. (10.43) that the time delay is proportional to the square of the value of NA. Therefore, a large NA fibre allows more modes of propagation of light, which will result in greater modal dispersion. A smaller NA limits the number of modes, hence reduces dispersion. It is further seen that the intermodal dispersion does not depend upon the spectral width of the source. It follows that a light pulse from an ideal monochromatic source would still get broadened.

Example 10.13: A step-index fibre is with a core of refractive index 1.55 and cladding of refractive index 1.51. Compute the intermodal dispersion per kilometer of length of the fibre and the total dispersion in a 15 km length of the fibre.

Solution: $\Delta t = \frac{n_1 L}{c} \left[\frac{n_1}{n_2} - 1 \right] = \frac{1.55 \times 10^3 \text{ m}}{3 \times 10^8 \text{ m/s}} \left[\frac{1.55}{1.51} - 1 \right] = 138 \text{ ns/km.}$

Total dispersion for 15 km length = $\Delta t \times 15 \text{ km} = (138 \text{ ns/km}) \times 15 \text{ km} = 2.07 \mu\text{s.}$

Intramodal Dispersion:

Intramodal dispersion is the spreading of light pulse within a single mode. The two main causes of Intramodal dispersion are (a) material dispersion and (b) waveguide dispersion.

(a) **Material Dispersion:** Glass is a dispersive medium. A light pulse is a wave packet, composed of a group of components of different wavelengths. The different wavelength components will propagate at different speeds along the fibre (Fig. 10.30). The short wavelength components travel slower than long wavelength components, eventually causing the light pulse to broaden. This type of distortion is known as **material dispersion**. It is often called the **chromatic dispersion**. Obviously, the spectral width of the source determines the extent of material dispersion.

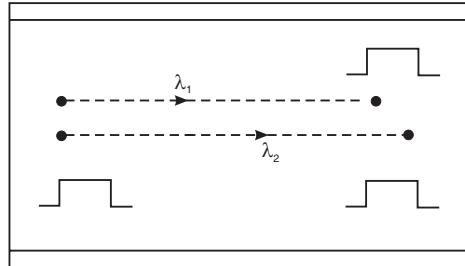


Fig. 10.30: The different wavelength components propagate at different speeds along the fibre

Expression for time delay due to material dispersion:

Let us consider a plane wave propagating in fiber core. It is represented by $\Psi \propto \exp(kx - \omega t)$. The wave number k is given by

$$k = \frac{2\pi}{\lambda} = \frac{2\pi}{\lambda_0} \cdot \frac{\lambda_0}{\lambda} = \frac{2\pi}{\lambda_0} \cdot n = \frac{2\pi v}{c} \cdot n$$

or

$$k = \frac{\omega n}{c} \quad (10.44)$$

and

$$\omega = 2\pi v = \frac{2\pi}{\lambda} c \quad (10.45)$$

A wave packet of finite spread of wavelengths travels with group velocity v_g is given by

$$v_g = \frac{d\omega}{dk}$$

∴

$$\frac{1}{v_g} = \frac{dk}{d\omega} = \frac{d}{d\omega} \left(\frac{\omega n}{c} \right) = \frac{n}{c} + \frac{\omega}{c} \frac{dn}{d\omega} = \frac{1}{c} \left[n + \omega \frac{dn}{d\omega} \right]$$

But

$$\frac{dn}{d\omega} = \frac{dn}{d\lambda} \cdot \frac{d\lambda}{d\omega} = -\frac{\lambda^2}{2\pi c} \cdot \frac{dn}{d\lambda}$$

∴

$$\frac{1}{v_g} = \frac{1}{c} \left[n - \frac{\omega \lambda^2}{2\pi c} \frac{dn}{d\lambda} \right] = \frac{1}{c} \left[n - \lambda \frac{dn}{d\lambda} \right]$$

As the signal propagates through the fibre, each spectral component can be assumed to travel independently and to undergo a time delay per unit length in the direction of propagation, which is given by

$$t_{\text{mat}} = \frac{L}{v_g} = \frac{L}{c} \left[n - \lambda \frac{dn}{d\lambda} \right] \quad (10.46)$$

The pulse spread Δt_{mat} for a source of spectral width $\Delta\lambda$ is found by differentiating the equ.(10.46) with respect to λ and then multiplying by $\Delta\lambda$. Thus,

$$\Delta t_{\text{mat}} = \frac{dt_{\text{mat}}}{d\lambda} \Delta\lambda = -\frac{L\lambda}{c} \frac{d^2n}{d\lambda^2} \Delta\lambda = D_{\text{mat}}(\lambda) L \Delta\lambda \quad (10.47)$$

where $D_{\text{mat}}(\lambda)$ is the material dispersion.

$$D_{\text{mat}}(\lambda) = -\frac{\lambda}{c} \frac{d^2n}{d\lambda^2} \quad (10.48)$$

From the equ.(10.48) it is seen that the material dispersion can be reduced either by choosing sources with narrower spectral range or by operating at longer wavelengths. To cite an example, an LED operating at 820 nm and having a spectral width of 38 nm results in dispersion of about 3 ns/km in a certain fibre. In the same fibre, dispersion can be reduced to 0.3 ns/km using a laser diode operating at 1140 nm and having a spectral width of 3 nm. Thus, using a more and more monochromatic source operating at higher wavelength, the material dispersion is reduced.

(b) **Wave-guide Dispersion:** Waveguide dispersion arises from the guiding properties of the fibre. The group velocities of modes depend on the wavelength. Hence, the effective refractive index for any mode varies with wavelength. It is equivalent to the angle between the ray and the fibre axis varying with wavelength which subsequently leads to a variation in the transmission times for the rays and hence dispersion (see Fig. 10.31). Waveguide dispersion is generally small in MMF, but it is important in SMF.

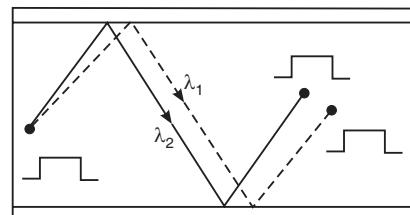


Fig. 10.31: Wave guide dispersion

The intermodal distortion can be reduced if graded index fibre is used. In case of a graded index fibre, the refractive index is larger at the center and it gradually decreases away from the center. A pulse traveling along the axis of the fibre, travels along a shorter path but it takes longer time to reach the end of the fibre since it is traveling through a medium of higher refractive index. On the other hand, the pulse traveling away from the axis travels a longer distance but takes lesser time since it is traveling through a medium of lower refractive index. As a result both the pulses reach the end of the fibre simultaneously. Thus, using a GRIN fibre can reduce the problem of intermodal dispersion. Low NA fibres exhibit smaller dispersion. Dispersion may be restricted by a careful selection of low NA fibre and a narrow spectral width fibre.

In a MMF, all three pulse spreading mechanism exist simultaneously. In case of SMF, only material and wave-guide dispersion exist.

Total Dispersion

All the above three dispersions contribute pulse spreading during signal transmission through an optical fibre. The total dispersion introduced by an optical fibre is given by the root mean square value of all the three dispersions. Thus,

$$(\Delta t)_T = \sqrt{(\Delta t)_{\text{inter modal}}^2 + (\Delta t)_{\text{mat}}^2 + (\Delta t)_{\text{wg}}^2} \quad (10.49)$$

10.17 BANDWIDTH

It is learnt in the above section that various dispersion mechanisms cause broadening of the information signal in time domain. If the pulses spread more, they can interfere with the adjacent pulses resulting in *Inter Symbol Interference* or ISI in short and there can be so much of ISI that it becomes impossible to distinguish between the individual pulses. Therefore, for a given broadening, the pulses have to be separated by a minimum time interval in order to avoid overlapping of the pulses. This would determine the ultimate information-carrying capacity of the system. When the pulse separation is increased, the data transfer rate decreases. Thus, broadening of pulses puts an upper limit on the rate of pulse transmission. To a first approximation, it may be taken that the bandwidth in hertz is equal to the digital bit rate.

Thus, $B_T = \frac{1}{\tau} = B$, where τ is the input pulse duration. In other words, the maximum allowable transmission rate is called **bandwidth**. In practice, the fibre bandwidth is expressed in terms of MHz.km, a product of frequency and distance. This is known as **bandwidth-distance** product, which specifies the usable bandwidth over a definite distance. With the increase in distance, different dispersion effects would increase in the optical fibre and as a result the usable bandwidth reduces. The attenuation per kilometer and the bandwidth-kilometer product are the important performance parameters of optical fibres.

10.18 CHARACTERISTICS OF THE FIBRES

A. Step-index single-mode fibre

- It has a very small core diameter, typically of about $10\mu\text{m}$.
- Its numerical aperture is very small.
- It supports only one mode in which the entire light energy is concentrated.
- A single mode step index fibre is designed to have a V number between 0 and 2.4.
- Because of a single mode of propagation, loss due to intermodal dispersion does not exist. With careful choice of material, dimensions, and wavelength, the total dispersion can be made extremely small.
- The attenuation is least.
- The single mode fibres carry higher bandwidth than multimode fiber.
- It requires a monochromatic and coherent light source. Therefore, laser diodes are used along with single mode fibres.

Advantages

- No degradation of signal
- Low dispersion makes the fibre suitable for use with high data rates. Single-mode fiber gives higher transmission rate and up to 50 times more distance than multimode.
- Highly suited for communications.

Disadvantages

- Manufacturing and handling of SMF is more difficult.
- The fibre is costlier.
- Launching of light into fibre is difficult.
- Coupling is difficult.

Application

- Used as under water cables

B. Step-index multi-mode fibre

- It has larger core diameter, typically ranging between 50-100 μm .
- The numerical aperture is larger and it is of the order of 0.3.
- Larger numerical aperture allows more number of modes, which causes larger dispersion. The dispersion is mostly intermodal.
- Attenuation is high.
- Incoherent sources like LEDs can be used as light sources with multimode fibres.

Advantages

- The multimode step index fibre is relatively easy to manufacture and is less expensive.
- LED or laser source can be used.
- Launching of light into fibre is easier.
- It is easier to couple multi-mode fibres with other fibres.

Disadvantages

- Has smaller bandwidth.
- Due to higher dispersion data rate is lower and transmission is less efficient.

Table 1: Comparison of Different Types of Fibres

S.No.	Feature	SMF	MMF	GRIN
1.	Typical core diametre	10 μm	50 to 100 μm	50 to 100 μm
2.	Δn	Very small	Large	
3.	Numerical Aperture	Small	Large	Smaller than that of MMF
4.	Number of modes	Only one	Many	Many
5.	Attenuation	Least	High	Lower
6.	Dispersion	Zero Intermodal dispersion	Large	Intermodal dispersion is zero. Material dispersion is present.
7.	Bandwidth	>3 GHz-km	<200 MHz-km	200 MHz-km to 3 GHz-km
8.	Advantages	No degradation of signal, High data transfer rate, Highly suitable for communications	Less expensive, LED or laser source can be used, Launching of light is easier, Coupling of fibres is easier.	LED or Laser light source can be used.
9.	Drawbacks	Costly, Requires a laser source, Coupling is difficult, Launching of light into fibre difficult, Intensity gets reduced.	Degrades signal, less suitable for communications.	
10.	Applications	Under water cables	Data Links	Telephone Lines

- It is less suitable for long distance communications.

Application

- Used in data links.

C. Graded-index multi-mode fibre

- Core diameter is in the range of 50-100 μm .
- Numerical aperture is smaller than that of step-index multimode fibre.
- The number of modes in a graded index fibre is about half that in a similar multimode step-index fibre.
- Has medium attenuation.
- Intermodal dispersion is zero, but material dispersion is present.
- Has better bandwidth than multimode step-index fibre.

Advantage

- Either an LED or a laser can be used as the source of light with GRIN fibres.

Disadvantages

- The manufacture of graded index fibre is more complex. Hence, it is the most expensive fibre.
- Coupling fibre to the light source is difficult.

Application

- Used in telephone links.

Table-1 gives a comparison of the different optical fibres.

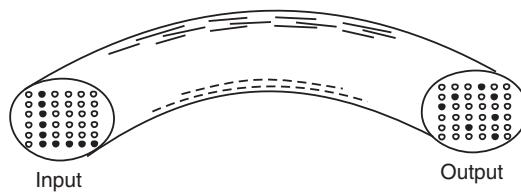
10.19 APPLICATIONS

Transmission of light via an optical fibre has a wide variety of applications. We discuss here some of the applications. Broadly, optical fibres have three different applications, apart from other miscellaneous applications.

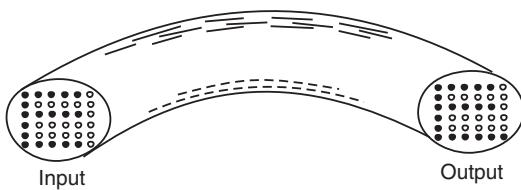
- a. They are used for illumination and short distance transmission of images.
- b. They are used as wave-guides in telecommunications.
- c. They are used in fabricating a new family of sensors.

10.19.1 Illumination and Image Transmission

A large number of fibres whose ends are bound together, ground and polished, form flexible bundles. One of the ends of the bundle acts as an input end while the other acts as an output end. If the relative positions of the fibre terminations at both the ends are not the same, and if no attempt is made to align the fibres in an orderly array, the bundle is said to be an *incoherent bundle*. In such a case, there would not be any correlation in the positions of the fibre terminations at one end of the bundle with that at the other end of the bundle.



(a) Incoherent bundle-image of letter L scrambled as dark spots



(a) Coherent bundle - image of letter E clearly seen

Fig. 10.32: Fibre Optic bundles

The primary function of such bundles is simply to conduct light from one region to another. Such **flexible light carriers** are relatively easy to make and inexpensive. They are used for illumination purpose.

When the fibres are carefully arranged so that their terminations occupy the same relative positions in both of the bound ends of the bundle, the bundle is said to be *coherent*. Such a bundle is capable of transmitting undistorted images to a distant place. When one end of such a **flexible image carrier** is placed face down flat on an illuminated surface, a point-by-point image of the surface appears at the other end.

Endoscopes

The most important application of the coherent bundles is in diagnostic field as an optical endoscope. An endoscope is an optical instrument which facilitates visual inspection of internal parts of a human body. It is also called a fiberscope. It requires about 10,000 fibres forming a bundle of 1 mm diameter and it can resolve objects with a separation of 70 μm . By allowing direct viewing of what was formerly hidden, a fiberscope has become a vital diagnostic tool for industry and medicine. The broncho-fiberscope, gastrointestinal fiberscope, laparoscope etc are the endoscopes used in medical diagnosis.

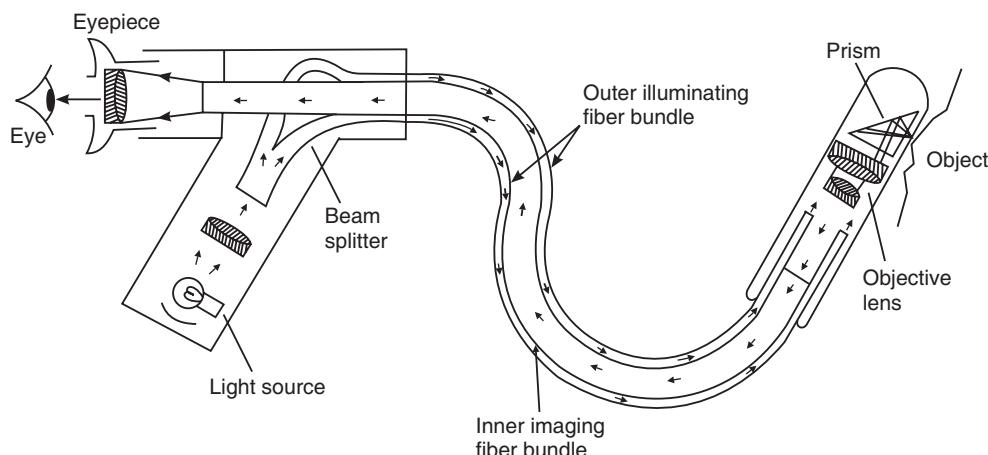


Fig. 10.33

Fig. 10.33 shows the schematic diagram of a flexible endoscope. The endoscopes are designed using low quality, large diameter and short silica fibres. There are two fiber bundles in an endoscope. One of them is used to illuminate the interior of the body and the other is used to collect the reflected light from the illuminated area. A telescope system is added in the internal part of endoscope for obtaining a wider field of view and better image quality. At the object end, there is an assembly of objective lens and prism which are kept in a transparent glass cover and at the viewing end, there is an eye lens. The input end of the endoscope contains a powerful light source. The light rays are focused and coupled to the illuminating fiber bundle. The light rays are finally incident on the surface of the object under study. The light rays reflected from the object surface are received by the objective lens through a prism and are transmitted through the imaging fibre bundle to the viewing end of the scope. Here the eye piece reconstructs the image of the object and one can view the image of the surface of the object. Endoscope pictures can be recorded on a videotape recorder.

10.19.2 Optical Communications

Traditionally, electronic communications were carried out by sending electrical signals through copper cables, coaxial cables or waveguides. In recent years optical fibres are being

used, where light signals replace electrical signals. A basic communications system consists of a transmitter, a receiver and an information pathway. Normally, the information to be communicated is a non-electrical message, which is to be converted first into an electrical form. The conversion is done by a transducer. For example, a microphone converts sound waves into currents. Similarly, a video camera converts images into currents. These electrical messages are of low frequency and cannot be transmitted directly. Therefore, they are superposed on a carrier wave of very high frequency. The process of imposing a message signal on a carrier wave is called *modulation*. Two different techniques of modulation are available. In analog modulation a continuous wave carries the message. In digital modulation message is transmitted in discrete form using binary digits. The message travels along the transmission channel and is received at the *receiver*. The receiver demodulates the modulated wave and separates out the message and feeds to a transducer such a loud speaker. The bandwidth requirement of the message and the bandwidth of the carrier determine the number of messages that can be simultaneously transmitted on an information channel. For example, a bandwidth of 4 kHz is required for voice transmission while 6 MHz bandwidth is required for TV signal transmission. When signals are transmitted in analog form the carrier should have double the above bandwidth. The normal TV communications has a bandwidth of 250 MHz and therefore, it can simultaneously transmit 20 TV programmes. However, instead of microwaves if light waves are used as carrier wave, the bandwidth will be about 10^8 MHz and can therefore transmit about 10^6 TV programs at a time. Thus, the use of the light waves expands our communication capabilities tremendously.

10.19.3 Medical Applications

Fibre optic technology is used in medical diagnostics as well as in medical procedures. The fibre optic endoscope is used to inspect internal organs for diagnostic purposes.

In ophthalmology, a laser beam guided by optical fibres is used to reattach detached retina and to correct defective vision.

In cardiology, optical energy transmitted through an optical fibre is used to evaporate built-up plaque that is blocking an artery.

In the treatment of cancer also the optical fibre technology is used. The process involves injection of special chemicals that penetrate only the cancerous cells. Infra red energy is transmitted via the fibre illuminates the affected area and is absorbed by the special chemical in the cancerous cells. The heat generated destroys the cancerous cells.

10.19.4 Military Applications

An aircraft, a ship or a tank needs tons of copper wire for wiring of the communication equipment, control mechanisms, instrument panel illumination etc. Use of optical fibre in place of copper reduces weight and further maintains true communication silence to the enemy. For example, a shipboard radar system requires about 250 m of coaxial cable, with a weight of 7 tons and a diameter of 45 cm. These cables can be replaced by optical fibre weighing 20 kg and measuring 2.5 cm in diameter.

Fibre guided missiles are used in recent wars. Sensors mounted on the missile transmit video information through the optical fibre to a ground control van and receive commands from the van again. The control van continuously monitors the course of the missile and if necessary corrects its course to ensure that the missile precisely hits the target.

10.20 FIBRE OPTIC COMMUNICATION SYSTEM

A fibre optic communication system is very much similar to a traditional communications system and has three major components. A **transmitter** converts electrical signal to light

signals, an **optical fibre** transmits the signals and a **receiver** captures the signals at the other end of the fibre and converts them to electrical signals.

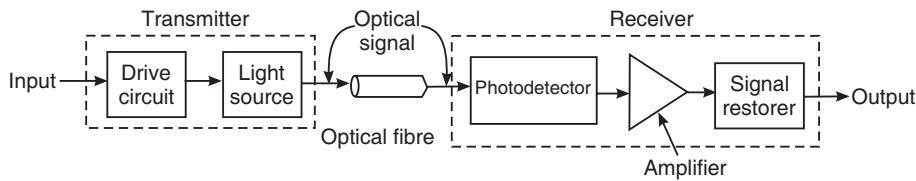


Fig. 10.34: Illustration of a typical fibre optic communication link.

The block diagram Fig. 10.34 illustrates a typical communications system. The transmitter consists of a light source supported by necessary drive circuits. A transducer converts a non-electrical message into an electrical signal and is fed to a light source. The light source is a miniature source, either a light emitting diode or a semiconductor laser. In either case, light is emitted in the IR range with a wavelength of 850 nm (0.85 μm), 1300 nm (1.3 μm) or 1550 nm (1.55 μm). The light waves are modulated with the signal. By varying the intensity of the light beam from the laser diode or LED, analog modulation is achieved. By flashing the laser diode or LED on and off at an extremely fast rate, digital modulation is achieved. A pulse of light represents the number 1 and the absence of light at a specified time represents zero. A message can be transmitted by a particular sequence of these 1s or 0s. If the receiver is programmed to recognize such digital patterns, it can reconstruct the original message. Though the digital modulation requires more complicated equipment such as encoders and decoders and also more bandwidth than analog modulation, it allows greater transmission distance with the same power. This is a great advantage and hence digital modulation has become popular and widely used nowadays. The transmitter feeds the analog or digitally modulated light wave to the transmission channel, namely optical fibre link. The optical signal travelling through the fibre will get attenuated progressively and distorted due to dispersion effects. Therefore, repeaters are to be used at specific intervals to regenerate the signal. At the end of the fibre, an output coupler directs the light from the fibre onto a semiconductor photodiode, which converts the light signals to electrical signals. The photodetector converts the light waves into electrical signals which are then amplified and decoded to obtain the message. The output is fed to a suitable transducer to convert it into an audio or video form.

Applications

Optical fibre communications systems can be broadly classified into two groups: (i) local and intermediate range systems where the distances involved are small and (ii) long-haul systems where cables span large distances.

(i) **Local area networks:** The local area network (LAN) is a computer oriented communication system. LAN operates over short distances of about 1 to 2 km. It is multiuser oriented system. In LAN, a number of computer terminals are interconnected over a common channel allowing each computer to use data and programs from any other. An optical data bus offers a great reduction in cost and increases enormously the information handling capacity.

(ii) **Long-haul communication:** One of the most important applications of fibre optic communication is long-haul communication. Long-haul communication systems are used for long distances, 10 km or more. Telephone cables connecting various countries come under this category. A rather sophisticated long-haul network is the NSFNET which links six supercomputer centres throughout U.S.A.

10.21 MERITS OF OPTICAL FIBRES

Optical fibres have many advantageous features that are not found in conducting wires. Some of the important advantages are given here.

1. Cheaper: Optical fibres are made from silica (SiO_2) which is one of the most abundant materials on the earth. The overall cost of a fibre optic communication is lower than that of an equivalent cable communication system.

2. Smaller in size, lighter in weight, flexible yet strong: The cross section of an optical fibre is about a few hundred microns. Hence, the fibres are less bulky. Typically, a RG-19/U coaxial cable weighs about 1100 kg/km whereas a PCS fibre cable weighs 6 kg/km only. Optical fibres are quite flexible and strong.

3. Not hazardous: A wire communication link could accidentally short circuit high voltage lines and the sparking occurring thereby could ignite combustible gases in the area leading to a great damage. Such accidents cannot occur with fibre links since fibres are made of insulating materials.

5. Immune to EMI and RFI: In optical fibres, information is carried by photons. Photons are electrically neutral and cannot be disturbed by high voltage fields, lightening, etc. Therefore, fibres are immune to externally caused background noise generated through electromagnetic interference (EMI) and radiofrequency interference (RFI).

6. No cross talk: The light waves propagating along the optical fibre are completely trapped within the fibre and cannot leak out. Further, light cannot couple into the fibre from sides. In view of these features, possibility of cross talk is minimized when optical fibre is used. Therefore, transmission is more secure and private.

7. Wider bandwidth: Optical fibres have ability to carry large amounts of information. While a telephone cable composed of 900 pairs of wire can handle 10,000 calls, a 1mm optical fibre can transmit 50,000 calls.

8. Low loss per unit length: The transmission loss per unit length of an optical fibre is about 4 dB/km. Therefore, longer cable-runs between repeaters are feasible. If copper cables are used, the repeaters are to be spaced at intervals of about 2 km. In case of optical fibres, the interval can be as large as 100 km and above.

10.21.1 Disadvantages

Installation and maintenance of optical fibres require a new set of skills. They require specialized and costly equipment like optical time domain reflectometers etc. All this means heavy investment.

10.22 FIBRE OPTIC SENSORS

Fibre optic sensors are transducers, which generally consist of a light source coupled with an optical fibre and a light detector held at the receiver-end. The fibres used could be either multimode or single mode type. The sensors can be used to measure pressure, temperature, strain, the acoustic field, magnetic field, etc physical parameters. The advantages of these sensors are that they are lighter, occupy lesser volume and are cheaper.

The optical fibre merely carries the light beam in some of the sensors and in others the fibre itself acts as the sensor. We study here a few typical examples of the sensors.

10.22.1 Temperature Sensors

(a) Intensity modulated sensor

Principle: In this type of sensor, temperature is measured by the modulation of intensity of the reflected light from a target, a silicon layer. The operation of the temperature sensor

is based on the $1 \mu\text{m}$ wavelength light-absorption characteristics of silicon as a function of temperature. Depending on the temperature, the amount of light absorbed by the silicon layer varies. The change in intensity of the reflected light is proportional to the change in temperature.

Construction: Fig. 10.35 illustrates a temperature sensor with a multimode fibre. The fibre is coated at one end with a thin silicon layer. The silicon layer is in turn coated with a reflective coating at the back. The silicon layer acts as the sensing element.

Working: The light from a light source is launched into the fibre from one of the ends of one of its branches (see Fig. 10.35). It passes first through the fibre and then through the silicon layer. The mirror coating at the other end of the silicon layer reflects the light back which again travels through the silicon layer. The reflected light emerges out through another branch of multimode fibre and is collected by a photodetector. The amount of the reflected light is converted into voltage by the photodetector. The absorption of light by the silicon layer varies with temperature and the variation modulates the intensity of the light received at the detector. Temperature measurements can be made with a sensitivity of 0.001°C .

(b) Phase modulated sensor

Principle: This temperature sensor is based on phase variation resulting due to the variation of refractive index of the optical fibre under the influence of temperature.

Construction: Fig. 10.36 shows a single mode fibre sensor arranged in what is known as the Mach-Zehnder arrangement. A light source such as a laser produces light. A beam splitter divides the light into two parts and sends light through the sensing fibre and the reference fibre. Light passing out of the two fibre elements is fed to a detector, which measures the difference in phase of the two light waves. Accurate measurements of the temperature may be obtained from these patterns.

Working: The light from the source is divided into two parts by the beam splitter. One part is allowed through sensor fibre, and the other part is passed through the reference fibre. Light rays entering the fibres are coherent and have the same phase. Prior to heating, the optical path lengths of the two fibre elements are same and hence both the outputs will be in phase. When the sensor fibre is subjected to heating, the temperature causes a change in the refractive index of the optical fibre. Therefore, the light coming out of the two fibres at the other end will have phase difference due to difference in optical path difference caused by the heating. When the rays are superposed, they interfere and interference pattern will be observed. As temperature increases, the phase difference between the two outputs increases and is observed as a displacement of the

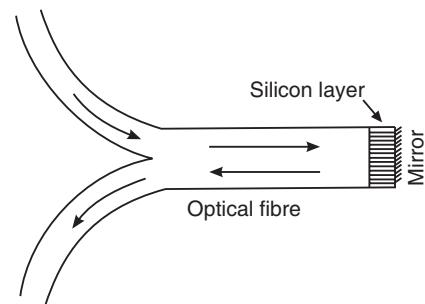


Fig. 10.35: A typical temperature sensor

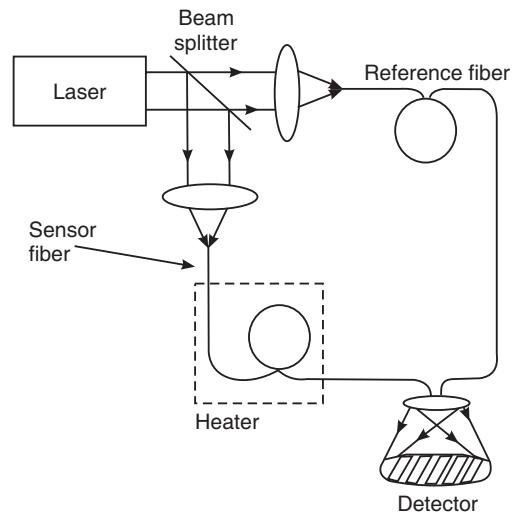


Fig. 10.36: Temperature sensor using phase variations

fringe pattern. By determining the fringe displacement, we can determine the magnitude of temperature.

10.22.2 Displacement Sensor

Principle: The basic principle employed in displacement sensor consists of using an adjacent pair of fibre optic elements, one to carry light from a remote source to an object whose displacement or motion is to be measured and the other to receive the light reflected from the object and carry it back to a remote photodetector.

Construction: Fig. 10.37 shows the arrangement of a displacement sensor. Two separate optical fibres are positioned adjacent to each other. One of them transmits light coming from a light source. The other fibre receives light reflected from the object under study and passes it on to a photodetector.

Working: Light from the transmitting fibre element is incident on the object under study. The light receiver fibre element is positioned adjacent to the transmitting fibre. If the gap between the object and the fibre elements is zero, the light from the transmit fibre would be directly reflected back into itself and little or no light would go into the receive fibre. When the object moves away, the gap increases and some of the reflected is captured by the receive fibre which in turn is carried to the photodetector. As the gap increases, a distance will be reached at which a maximum reflected light is received by the photodetector. Further increase in the gap will result in a decrease in the light at the receiver fibre face and corresponding drop in the signal output from the photodetector. By proper calibration, we can obtain the displacement of the object in terms of the strength of the output signal of the photodetector.

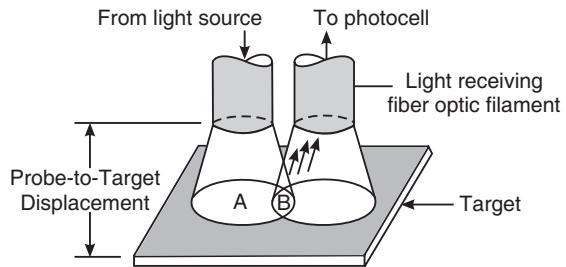


Fig. 10.37: Object moving away from probe, causes increase in reflected light intensity.

10.22.3 Force Sensor

Principle: This sensor is based on variations of light intensity. When an optical fibre is pressed, a small change occurs in light propagation direction due to microbending of the fibre. As a result, energy from one mode is transferred to another mode through mode coupling. In addition, higher order modes are likely to change into radiation modes. All these effects cause

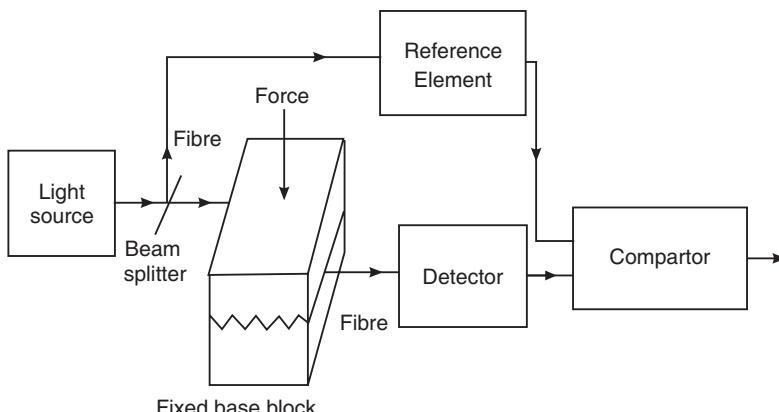


Fig. 10.38: Force sensor using microbend losses

a loss in intensity of the light transmitted through the fibre. Therefore, the change in intensity of the transmitted light is proportional to the force applied on the optical fibre.

Construction: An optical fibre without jacket is placed held between two corrugated blocks, as shown in Fig. 10.38. Light from a source is divided into two parts by a beam splitter. One part is allowed through the fibre that is held between the blocks, which acts as a sensor element, and the other part is passed through an exactly identical fibre, which acts as a reference element. Photodetectors measure the intensity of transmitted light. A comparator detects the difference between the light intensities.

Working: When a force is applied on the upper corrugated block, the fibre is pressed and microbend losses are introduced in the fibre. The microbendings produce mode coupling such that energy of one mode is transferred to other higher modes. Also, higher modes are converted into leaky modes which reduce the amount of energy transmitted through the fibre. The changes in the light intensity due to these losses are detected by a photodetector and compared with that of the light coming out of the reference element. The change in intensity is related to the force and hence is a measure of the applied force.

10.22.4 Liquid Level Detector

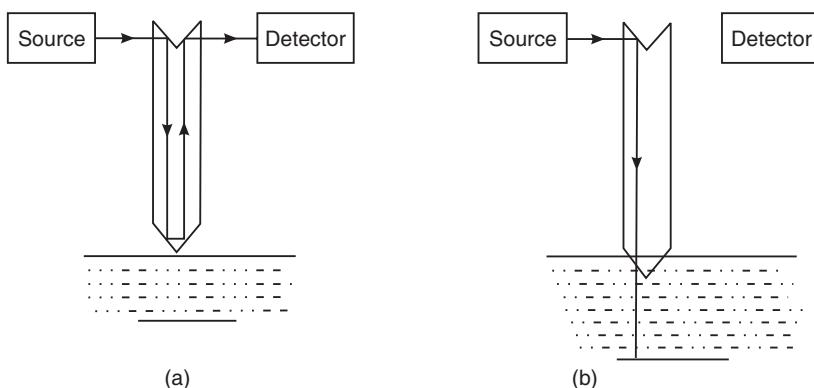


Fig. 10.39: Liquid Level Detector

Principle: The liquid level detector described here is based on the principle of total internal reflection.

Construction: A simple liquid level detector is shown in Fig. 10.39. A notch is made at one end of a multimode optical fibre and its other end is chamfered as shown in Fig. 10.39. A light source sends light on to the fibre and a photodetector on the other side registers light emerging out from the fibre.

Working: The optical fibre is arranged at the desired height in a vessel. The refractive index of the fibre is chosen to be less than that of the liquid whose level is to be detected. Light from the light source is made to be incident on one of the inclined faces of the notch. The light turns through 90° and travels through the fibre. On reaching the chamfered end of the fibre, it gets internally reflected at the fibre-air boundary, if the liquid is below the desired level. Then, it is again turned through 90° at the opposite face, travels back through the fibre to be turned once again through 90° and is detected at the detector (Fig. 10.39a).

When the liquid rises and touches the fibre end, total internal reflection ceases and the light is transmitted into the liquid. Hence, the photodetector does not receive any light (Fig. 10.39b). Thus, an indication of the liquid level is obtained at the detector.

QUESTIONS

1. What is an optical fibre? What is the principle involved in its working? **(R.T.M.N.U.,2006)**
2. What are optical fibres? Give its applications. **(Amaravati Univ.,2007)**
3. Explain core, cladding and sheath. **(Amaravati Univ.,2008)**
4. With a neat diagram, explain the structure of an optical fibre. **(M.G.Univ.,2006), (Calicut Univ.,2005)**
5. Explain the phenomenon of total internal reflection of light. How is it used in fiber optic communications?
6. Explain the principle of optical fibre as a waveguide for light. **(M.G.Univ.,2005)**
7. What is meant by critical propagation angle of an optical fibre? Obtain an expression for the critical propagation angle.
8. What is meant by critical angle of an optical fibre? Obtain an expression for the critical angle.
9. Describe the propagation of light in optical fibre and obtain expression for numerical aperture. **(Cochin Univ., 2004)**
10. Explain the following terms:

<i>(i)</i> Critical angle	<i>(ii)</i> Acceptance cone	<i>(iii)</i> Numerical aperture
		(R.T.M.N.U.,2006)
11. Deduce an expression for acceptance angle of an optical fibre. **(R.T.M.N.U.,2006)**
12. Explain the following terms:

<i>(i)</i> Numerical aperture	<i>(ii)</i> Acceptance angle	<i>(iii)</i> Acceptance cone in case of optical fibres.
		(Cochin Univ., 2005)
13. For an optical fibre define the following terms:

<i>(i)</i> Acceptance angle and acceptance cone	<i>(ii)</i> Numerical aperture	<i>(iii)</i> V-number.
		(RGPV,2004)
14. Using ray theory, derive the condition for transmission of light within an optical fibre.
15. Explain how light is propagated through an optical fibre and determine the numerical aperture and acceptance angle. **(Calicut Univ.,2005)**
16. Explain with necessary theory, the propagation of light in optical fibres. Derive an expression for numerical aperture. **(M.G.Univ.,2006)**
17. Derive an expression for angle of acceptance of fibre in terms of refractive index of the core and the cladding of an optical fibre. What is meant by acceptance cone?
18. What is meant by acceptance of angle for an optical fibre? Show how it is related to numerical aperture.
19. What do you understand by the terms acceptance angle and acceptance cone? Derive an expression for acceptance angle in terms of refractive indices of the core and the cladding. **(R.T.M.N.U.,2007)**
20. Derive an expression for acceptance angle and numerical aperture for an optical fibre. **(G.T.U.,2009), (Anna Univ., 2005)**
21. Derive the relation for numerical aperture of an optical fibre? How is it useful on optical fibres. **(M.G.Univ.,2005)**
22. Explain what you understand by acceptance angle and numerical aperture.
23. Derive an expression for numerical aperture of a step index optical fiber. **(RGPV,2007), (Calicut Univ.,2007)**
24. Explain the terms acceptance angle and acceptance cone. **(Calicut Univ.,2007)**
25. Derive an expression for N.A. for S.I. fibre in terms of refractive index of the core and relative refractive index difference between the core and the cladding. **(R.T.M.N.U.,2007)**
26. Derive an expression for numerical aperture of a step-index fibre in terms of Δ .
27. Define the relative refractive index difference of an optical fibre. Show how it is related to numerical aperture.

28. Explain the mechanism of light propagation in optical fibre. Discuss the different types of optical fibres with suitable diagrams. **(V.T.U.,2008)**
29. Classify the fibres on the basis of refractive index profile, on the basis of modes and on the basis of materials.
30. What is meant by multimode step index fibre? **(Amaravati Univ.,2008)**
31. Explain step index and graded index fibres. **(Cochin Univ., 2005)**
32. What is step index and graded index optical fibre? **(Amaravati Univ.,2005)**
33. Explain the different types of optical fibre, along with the refractive index profile and mode propagation sketches. **(V.T.U.,2008)**
34. Explain what is step-index, graded index, monomode and multimode fibre. Draw relevant sketches. **(R.T.M.N.U.,2007)**
35. Differentiate between the step-index fibre and graded-index fibre. **(RGPV,2007), (M.G.Univ.,2005), (Kerala Univ., 2004)**
36. Distinguish between single mode and multimode fibres with suitable diagram. **(Calicut Univ.,2007)**
37. What are step-index fibre and graded-index fibre? How light propagates through step index and graded index fibre? Explain how pulse dispersion is minimized in graded index fibre. **(RGPV, 2008)**
38. Describe the light propagation and modal dispersion in different types of fibres with neat ray diagrams indicating refractive index profile and the core diameter. **(V.T.U.,2008)**
39. Describe various mechanisms of attenuation in optical fibres. **(R.T.M.N.U.,2007)**
40. What is attenuation in an optical fibre? Explain the attenuation mechanisms. **(V.T.U.,2007)**
41. Mention few applications of optical fibre. **(Kerala Univ., 2004)**
42. List the main components of optical communication system. Describe the basic optical communication system.
43. Explain with neat block diagram the principle of optic fibre communication. Also explain the signal distortion and optical transmission losses in optical fibres. **(Calicut Univ.,2005)**
44. Explain optical communication through block diagram. For long distance communication whether (i) mono-mode or multimode and
(ii) step index or graded index fibre, which are preferable and why?
45. State the main components of optical fibre communication system. **(G.T.U., 2009)**
46. Discuss the advantages of optical fibre communication system over the conventional coaxial communication system. **(G.T.U., 2009)**
47. Give the block diagram of optic fibre communication system explaining the functions of different blocks. Compare its merits over conventional communication system. **(M.G.Univ.,2005), (Calicut Univ., 2007)**
48. What is the principle behind the functioning of an optic fibre? What are the advantages of the optical fibres over coaxial cables? **(M.G.Univ., 2005)**
49. Explain the principle of optic fibre communication. Mention advantages of fibre optic communication. **(Kerala Univ., 2004)**
50. Give the various advantages of optical fibres over conventional cables. **(Cochin Univ., 2005)**
51. Explain the important applications of optical fibre. **(M.G.Univ.,2005), (Calicut Univ., 2005)**
52. Explain with basic principle, the construction and working of any one type of optical fibre sensor.
53. Discuss any one application of an optical fibre as a sensor. **(R.T.M.N.U., 2006)**
54. Describe various types of optical fibres. Explain optical fibre sensors. **(M.G.Univ., 2006)**
55. Write a short note on fibre optic sensor. **(M.G.Univ., 2005)**

PROBLEMS

1. An optical fibre has a core material of refractive index of 1.55 and cladding material of refractive index 1.50. The light is launched into the fibre from air. Calculate its numerical aperture.
2. The numerical aperture of an optical fibre is 0.39. If the difference in the refractive indices of the material of its core and the cladding is 0.05, calculate the refractive index of material of the core.
3. An optical fibre has an acceptance angle 26.80° . Calculate its numerical aperture. (Ans: 0.4508)
4. An optical fibre refractive indices of core and cladding are 1.53 and 1.42 respectively. Calculate its critical angle. (Ans: 68.14°)
5. Consider a fibre having a core of index 1.48, a cladding of index 1.46 and has a core diameter of 30 μm . Show that all rays making an angle less than 9.43° with the axis will propagate through the fibre.
6. A step-index fibre is made with a core of index 1.54, a cladding of index 1.50 and has a core diameter of 50 μm . It is operated at a wavelength of 1.3 μm . Find the V-number and the number of modes that the fibre will support. (Ans: 42.15, 888)
7. Using a step index fibre with $n_1 = 1.48$ and $n_2 = 1.46$ and the core radius $a = 30\mu\text{m}$. Calculate the number of total internal reflections that will occur on its propagation in a length of 1 km fibre.
8. A step-index fibre has a core refractive index of 1.44 and the cladding refractive index of 1.41. Find (i) the numerical aperture, (ii) the relative refractive index difference, and the acceptance angle. (Ans: 0.292, 0.021, 33.96°)
9. An optical fibre has a numerical aperture of 0.20 and a cladding refractive index of 1.59. Find the acceptance angle for the fibre in water which has a refractive index of 1.33. (Ans: 8°39')
10. Compute the cut-off parameter and the number of modes supported by a fibre which has a core refractive index of 1.54 and the cladding refractive index of 1.50. The radius of the core is 25 μm and operating wavelength is 1300 nm. (Ans: 42.15, 888)
11. Find the numerical aperture and acceptance angle of a fibre of core index 1.4 and $\Delta = 0.02$. (Ans: 0.28, 32.52°)
12. Compute the total dispersion in 10 km length of a step index fibre, which has a core refractive index of 1.55 and the relative refractive index difference of 0.026. (Ans: 138ns)
13. Consider a bare step index fibre having a refractive index of 1.46. The radius of the fibre is 50 μm . Compute the pulse dispersion per km. (Ans: 2238ns)
14. Compute the cut-off parameter and the number of modes supported by a fibre, which has a core refractive index of 1.47 and the cladding refractive index of 1.45. The radius of the core is 50 μm and operating wavelength is 850 nm. (Ans: 44.64, 996)
15. A step-index fibre has a normalized frequency $V = 26.6$ at 1300 nm wavelength. If the core radius is 25 μm , calculate the numerical aperture. (Ans: 0.22)
16. Find the core radius necessary for single mode operation at 820 nm of a step index fibre, which has a core refractive index of 1.480 and the cladding refractive index of 1.478. (Ans: 4.08 μm)
17. A signal of 100mW is injected into a fibre. The outcoming signal from the other end is 40 mW. What is the loss in dB? (Ans: 3.98 dB)
18. A communication system uses a 10 km fibre having a fibre loss of 2.5 dB/km. Find the input power if the output power is 1.265 μW . (Ans: 400 μW)
19. A fibre length 100m has power input 10 μW and power output 8.8 μW . Find the power loss in dB/km. (Ans: 5.55 dB/km)
20. When the mean optical power launched into a 8 km length fibre is 120 μW , the mean optical power at the fibre output is 3 μW . Determine
 - (i) the overall signal attenuation in dB through the fibre,
 - (ii) the signal attenuation per km for the fibre,
 - (iii) the overall signal attenuation for a 10 km optical link using the same fibre,
 - (iv) the numerical input/output ratio. (Ans: 16 dB, 2.0 dB/km, 20 dB, 100)

CHAPTER

11

Architectural Acoustics

11.1 INTRODUCTION

Acoustics is the science of sound and deals with the origin, propagation and auditory sensation of sound. The study of sound plays a very important role in various branches of engineering and has developed to such a level that it has become an independent branch of engineering known as **acoustic engineering** or **sound engineering**. The area of study of design of musical instruments is known as **musical acoustics**, the technology of sound production and recording as **electro-acoustics**, use of sound in medical diagnosis and therapy as **bio-acoustics**, the design of buildings, auditoriums, musical halls, lecture halls, recording rooms, etc. as **architectural acoustics**.

Architectural acoustics deals in general with the behaviour of sound waves in closed spaces and their design to give the best sound effects. The acoustic properties of buildings were not studied on a scientific basis till about 1900. Buildings designed to screen movies, to stage dramas or for music concerts often lacked the proper acoustic quality and were found unfit for such activities. The Fogg Art Museum hall in Harvard University, U.S.A. turned out to be highly defective when it was built. The lectures given in it were not intelligible to audience. Prof. Wallace C. Sabine, Professor of Physics in Harvard University was entrusted with the responsibility of eliminating the acoustical defects of the hall. Sabine undertook a systematic study of the problem and evolved conditions for a satisfactory acoustic quality of a hall. He found that quite often reverberation was the main cause for a defective quality of a hall. Addition of absorbent materials at appropriate surface enhances the quality of sound in the halls. Other precautions are to be taken about the shape of walls, ceiling and the hall in total so that acoustic defects do not arise. Thus, Prof. Sabine laid the foundations of acoustic engineering.

11.2 SOUND

Sound is always produced by some vibrating body. The vibrating body excites mechanical waves in the surrounding medium. These mechanical waves propagate in air in the form of a series of compressions and rarefactions of air molecules (Fig. 11.1). On reaching the ear, they cause the eardrum to vibrate, leading to the sensation of hearing. Sound cannot travel in a vacuum and requires the presence of an elastic medium for its propagation. Sound waves are longitudinal waves. The compressions and rarefactions caused by the vibrating body modulate the normal atmospheric pressure with small pressure changes occurring regularly above and below it. Thus, a sound wave is one complete cycle of pressure variation.

The wave motion of sound does not change the mean position of the vibrating particles (molecules) and the average maximum distance of a particle from its mean position is called **amplitude**. Even a sound of 0.01 mm amplitude is enough to be audible, while a sound of 0.1 mm amplitude is quite a loud sound.

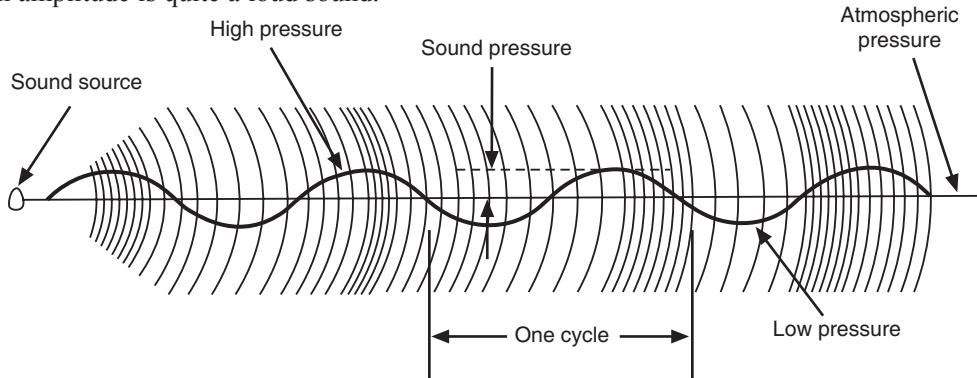


Fig. 11.1

11.2.1 Sound Velocity

The velocity of sound is not a constant and depends on the nature and temperature of the medium through which it travels. In general, the velocity of sound in a gaseous medium is governed by the relation

$$v = \sqrt{\frac{B}{\rho}} \quad (11.1)$$

where B is the bulk modulus of the medium and ρ its density. The speed of sound in air is commonly taken as 344 m/s for normal conditions. This is very less compared to the velocity of light. Table-1 lists the speed of sound in some materials. It may be noted that sound travels faster in liquid media than in gaseous media and much faster in solid media.

Table 1: Speed of sound in some materials at 20°C

Medium	Speed (m/s)	Medium	Speed (m/s)
Air	344	Brick	4300
Hydrogen	1305	Mild steel	5050
Pure water	1480	Aluminium	5150
Plexiglass	1800	Glass	5200
Soft wood	3350	Granite	6400
Concrete	3400	Gypsum Board	6800

11.2.2 Sound Wavelength

The velocity wave may be expressed as the product of frequency f and wavelength, λ . Thus,

$$v = f\lambda \quad (11.2)$$

Therefore, the wavelength may be written as

$$\lambda = \frac{v}{f} \quad (11.3)$$

It follows that the wavelength of sound will be larger in a medium having a higher velocity. When the medium is air, we can write

$$\lambda = \frac{344 \text{ m/s}}{f} \quad (11.4)$$

11.3 CLASSIFICATION OF SOUND

Sound waves are classified based on their *frequency* into three groups, namely *sonic* (audible), *infrasonic* and *ultrasonic* waves.

- The waves that produce a sense of sound on a human ear are called **audible** waves. Sound waves with frequencies lying in the range of 16 Hz to 20,000 Hz are audible waves.
- Sound waves with frequencies below 16 Hz are called **infrasonic** waves.
- Sound waves with frequencies above 20 kHz are called **ultrasonic** waves.

Sometimes elastic waves with frequencies of 10^{10} Hz and higher are called **hypersonic** waves. They correspond to thermal waves in liquids or solids.

Audible sound waves can be further classified according to their frequency spectrum as

- (i) Musical sounds and
- (ii) Noise

Musical sounds produce a pleasing sensation on the ear. They have the following characteristics.

- Musical sound has a line spectrum containing multiple frequencies.
- Musical sounds are *periodic vibrations*.
- Sudden changes in amplitude do not occur.

On the other hand, **noise** causes irritation and strain to our ears. If very loud, noise may cause permanent or temporary deafness. Noise has the following characteristics.

- Noise is a jumble of irregularly timed, *non-periodic vibrations*.
- It consists of a complex spectrum of frequencies.
- It undergoes erratic changes in amplitude and frequency.

11.4 CHARACTERISTICS OF MUSICAL SOUND

A sound wave, which has a single well-defined frequency, is called a **tone**. Thus, every single sinusoidal sound wave is a *pure tone*. In general, sound sources do not produce a single frequency. Normally, a *musical note* consists of several tones (or a series of harmonic waves) of different frequencies of varying intensity and the pitch of that note corresponds to the lowest tone it contains. The lowest-pitch tone (of frequency, f) is the loudest and is called the **fundamental** tone. This frequency is dominant and defines the pitch of a note. The additional frequencies $2f$, $3f$, $4f$,..... accompanying the fundamental tone are called **overtones** or **harmonics**. The intensities of overtones diminish with increase of their frequencies. Each source (musical instrument) produces different overtones and the overtones so present are characteristic of that source.

Musical sound has the following three characteristics:

- Pitch or frequency
- Timbre or Quality
- Loudness or Intensity

11.4.1 Pitch

The first characteristic of a musical sound is its pitch. Pitch is a subjective sensation perceived when a tone of a given frequency is sounded. It enables us to classify a musical note as high or low and to distinguish a shrill sound from a flat sound of the same intensity sounded on the same musical instrument. Pitch of a musical sound is determined by its frequency but it

is also function of its intensity and wave form. Greater is the frequency of a musical note, higher is the pitch and vice versa. The frequency and pitch are two different characteristics. The frequency is a physical quantity and can be measured accurately, while pitch of a note is a physiological quantity which is merely a sensation experienced by a listener. The change in pitch with loudness is most pronounced at a frequency of about 100 Hz. In the 100 Hz range the pitch increases with increasing loudness. For frequencies between 1 kHz and 5 kHz which is the range for which the ear is most sensitive, the pitch of a tone is relatively independent of its loudness. In general, the pitch varies in a parabolic manner with frequency in the range 20 Hz to 10 kHz.

11.4.2 Timbre

The second characteristic of a musical sound is its timbre or quality. Timbre is the name given to subjective sensation, which enables us to distinguish the same note played on different instruments or sung by different singers. Memory of the timbre helps us identify different sounds such as the instrument being played or the person who is speaking or singing. The notes produced by musical instruments are complex and consist of a large of tones of different frequencies. Even if two different musical instruments produce notes of the same pitch, the overtones will be different. This difference defines the quality of the note produced by one instrument and distinguishes the note produced by it from that of the other instrument. If two musical notes have the same fundamental pitch (the same fundamental frequency) but differ in overtones (that is in harmonics), they are said to differ in **quality**. The frequencies and amplitudes of the overtones present in the musical notes can be analyzed with the help of Fourier analysis. The analysis showed that the quality of a note depends upon the presence or absence of particular overtones and their relative intensities.

11.4.3 Loudness

Loudness signifies how far and to what extent, sound is audible. Loudness of sound is again a subjective perception. Because of the varying sensitivity of the ear, different people perceive the same sound differently. What is loud to one person may be soft for another. An objective measure of loudness is sound *intensity*, which is applicable equally to every one. Loudness is found to vary with frequency also. Intensity of sound is a physical quantity and does not depend on the listener.

Intensity: *The intensity of the sound wave is defined as the rate of flow of sound energy through a unit area normal to the direction of propagation.* Since the rate of flow of energy is power, the intensity of sound wave is measured in units of power per unit area, i.e., in watts/sq.m. The sound intensity is proportional to the square of the wave amplitude. Thus,

$$I \propto P^2 \quad (11.5)$$

where P is the pressure amplitude. The intensity of the faintest sound wave that can be heard is about 10^{-12} W/m². The range of intensity of sound that an ear can hear ranges from 10^{-12} W/m² to about 1 W/m. Sounds of same intensity but of different frequencies may differ in loudness.

11.5 WEBER-FECHNER LAW

Weber-Fechner law states that *the degree of sensation of sound is proportional to the logarithm of the stimulus producing the sound.* If L is the degree of loudness due to intensity I , then

$$L = k \log I \quad (11.6)$$

where k is the proportionality constant depending on the sensitivity of the ear, quality of the sound and other factors.

Equ.(11.6) is the mathematical statement of the Weber-Fechner law. Because of the logarithmic nature, loudness is not doubled when the intensity is doubled. However, we can show that loudness increases by the same amount each time when the intensity is doubled.

Let I_1 be the initial intensity, which produces loudness L_1 . Then,

$$L_1 = k \log I_1$$

Now, if the intensity is doubled, I_1 becomes $2I_1$ and the loudness due to doubled intensity will be

$$L_2 = k \log 2 I_1 = k \log 2 + k \log I_1$$

or

$$L_2 = k \log 2 + L_1$$

∴

$$L_2 - L_1 = k \log 2 \quad (11.7)$$

As $k \log 2$ is a constant, it follows from equ.(11.7) that the loudness increases by the same amount whenever the intensity is doubled, irrespective of the initial intensity.

11.6 SOUND INTENSITY LEVEL - DECIBEL

The range of variation of sound intensity is very large. The loudness of a sound as judged by the ear is proportional to the logarithm of intensity. It means that our ear is a logarithmic instrument. Therefore, the absolute intensity of sound wave is not of practical significance. Instead, the relative intensity is of more concern for us. The lowest intensity of sound at 1 kHz to which a normal human ear can respond is $I_0 = 10^{-12} \text{ W/m}^2$. This is known as the **threshold of hearing** and is chosen as the “zero” or “standard” intensity. Intensity of a sound is measured with reference to this standard frequency. *The ratio of the intensity of sound wave to the threshold intensity of hearing is defined as the intensity level of sound.*

If I and I_0 represent the intensities of two sounds of a particular frequency (chosen normally to be 1 kHz), L_1 and L_0 are their corresponding measures of loudness, then according to Weber-Fechner law

$$L_1 = k \log I$$

and

$$L_0 = k \log I_0$$

The difference in the loudness of the two sounds is given by

$$L = \log \frac{I}{I_0} \quad \text{bel} \quad (11.8)$$

L is called the intensity level and is expressed in bels, a unit named after Alexander Graham Bell, the inventor of telephone.

Def: 1 bel is defined as the relative intensity between two sound notes if one is 10 times more intense than the other.

The unit of bel is large and in practice a smaller unit *decibel* is used.

$$1 \text{ decibel} = \frac{1}{10} \text{ bel}$$

Accordingly, the intensity level of a sound wave is defined as

$$L = 10 \log \frac{I}{I_0} \quad \text{decibels (dB)} \quad (11.9)$$

Threshold of audibility

The intensity level corresponding to the intensity I_0 will be 0 dB, since from equ.(11.9), we get

$$L = 10 \log \frac{I_0}{I_0} = 10 \log 1 = 0$$

0 dB level represents the **threshold of audibility**.

Physical significance of a decibel

The smallest change in intensity level that the human ear can detect is 1 dB. Let us find the corresponding change in intensity or loudness. From equ.(11.9), when $L = 1\text{dB}$,

$$1 \text{ dB} = 10 \log \frac{I}{I_0}$$

or $\log \frac{I}{I_0} = \frac{1}{10} = 0.1$

$\therefore \frac{I}{I_0} = 10^{0.1} = 1.26$

or $I = 1.26I_0$

It means that a change in intensity by 26% increases the intensity level by one decibel. Further,

if $I = 100 I_0$, $L = 10 \log 10^2 = 20 \text{ dB}$,

$I = 1000 I_0$, $L = 10 \log 10^3 = 30 \text{ dB}$,

$I = 10000 I_0$, $L = 10 \log 10^4 = 40 \text{ dB}$ and so on.

Thus, when two sounds differ by 20 dB, the louder of them is 100 times more intense and when they differ by 40 dB, the louder one is 10,000 times more intense. The loudest sound that can be heard without pain is about 120 dB. This is known as the **threshold of feeling** or **pain threshold**. Some of the typical sound levels are shown in Table-2.

Table 2: Some typical sound levels

Source	Sound level (dB)
Rustling of tree leaves	0 - 20
Quiet living room	20 - 40
Average office	40 - 60
Average street noise	60 - 80
Train sound	80 - 100
Thunder	100 - 120
Aeroplane noise at a distance of 3 m	130 (painful)

11.6.1 Sound Pressure Level (SPL)

In acoustical problems, sound levels are generally dealt in terms of pressure rather than intensity. In fact, sound measuring devices respond to pressure exerted by sound. Equ.(11.5) shows that the sound intensity is proportional to the square of its pressure. Using equ.(11.5) into equ.(11.9), an expression for the SPL may be obtained as follows.

$$\text{SPL} = 10 \log \frac{I}{I_0} = 10 \log \left(\frac{P}{P_0} \right)^2 = 20 \log \frac{P}{P_0} \text{ decibels} \quad (11.10)$$

The reference pressure P_0 is usually taken as $P_0 = 2 \times 10^{-5} \text{ N/m}^2$.

The SPL can be directly measured on a sound level meter.

11.7 HUMAN AUDIOGRAM

An audiogram for the normal human ear is shown in Fig. 11.2. The human ear exhibits basically a nonlinear response. The lower curve represents the faintest sounds that can be heard and the upper curve the loudest sounds that can be heard without pain. Further, the sensitivity of the ear varies with frequency. Thus, threshold of hearing of sound depends upon both the intensity and frequency of sound. A sound is audible above a certain minimum intensity and a certain minimum frequency. For a person with normal hearing, the threshold of audibility at 1 kHz is 0 dB; at 200 Hz and 15 kHz it is about 20 dB and at 50 Hz and at 18 kHz it is about 50 dB. When the intensity of sound exceeds a maximum limit, it produces a sensation of pain on the ear. Similarly, there is a maximum frequency limit beyond which the sound is not heard. Thus, the maximum audible intensity of sound also depends upon both the intensity and frequency of sound. In Fig. 11.2, the two curves, the threshold of hearing and the threshold of pain enclose an area called auditory area. A sound whose frequency and intensity are not within the limits set by this area, is not heard by humans.

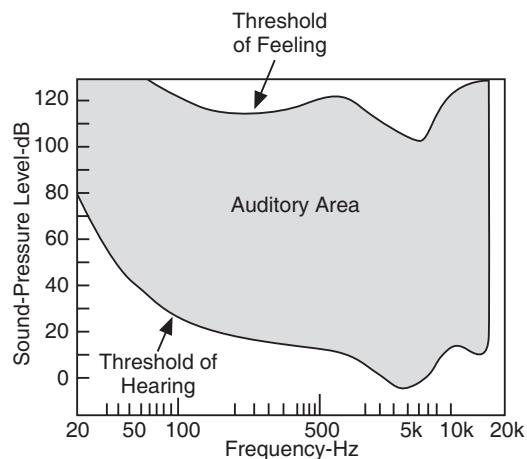


Fig.11.2

11.8 PHON

Phon is a measure of loudness and it is equal to the loudness of an equally loud 1kHz frequency note expressed in decibels.

Saying that two sounds have equal intensity is not the same thing as saying that they have equal loudness. The loudness of a sound depends on the frequency of the sound and the sensitivity of the ear. Therefore, two different 60-decibel sounds will not in general have the same loudness. The loudness cannot be measured directly with a meter. However, the phon scale is determined by the results of experiments in which volunteers were asked to adjust the loudness of a sound at a given frequency until they judged its loudness to equal that of a 1 kHz signal. A sound of known intensity level is produced at the frequency 1 kHz (1 kHz being taken as the standard frequency) and then a sound at some other frequency of interest is generated. The intensity of the second sound is varied in intensity till the volunteers judge it to be of the same loudness as that of 1 kHz sound. If the intensity level of both sounds is N decibels, then the equivalent loudness is said to be N phons. If a given sound is perceived to be as loud as a 60 dB sound at 1000 Hz, then it is said to have a loudness of 60 phons.

60 phons means “as loud as a 60 dB, 1000 Hz tone”

Thus, the phon is a unit that is related to dB by the *psychophysically measured* frequency response of the ear. The loudness of complex sounds can be measured by comparison to 1000 Hz test tones. At 1 kHz, readings in phons and dB are, by definition, the same. Therefore, for a 1 kHz note,

$$\begin{aligned} \text{Loudness in Phons (LP)} &= 10 \log(I/I_0) \\ &= 10 \log I + 10 \log (1/I_0) \\ &= 10 \log I + 10 \log 10^{12} \end{aligned} \quad (11.11)$$

$$\begin{aligned}
 \text{or} \quad & LP = 10 \log I + 120 \\
 \therefore \quad & \log L = 0.033(10 \log I + 120 - 40) \\
 & = 0.33 \log I + 2.64 \\
 \text{which reduces to} \quad & L = 445I^{0.33} \\
 \text{that is,} \quad & L \propto I^{1/3} \quad \text{for a 1 kHz tone.}
 \end{aligned} \tag{11.12}$$

11.8.1 SONE

The use of the phon as a unit of loudness is an improvement over just quoting the level in decibels, but it is still not a measurement, which is directly proportional to loudness. The *sone* is derived from psychophysical measurements, which involved volunteers adjusting sounds until they judge them to be twice as loud. This allows one to relate perceived loudness to phons. A **sone** is defined to be equal to 40 phons. In other words, **a sone is the loudness of a 1kHz tone of 40-db intensity level**. Experimentally it was found that a 10 dB increase in sound level corresponds approximately to a perceived doubling of loudness. So that approximation is used in the definition of the phon: 0.5 sone = 30 phon, 1 sone = 40 phon, 2 sone = 50 phon, 4 sone = 60 phon, etc.

11.9 SOUND REFLECTION

Sound waves are longitudinal waves and require an elastic medium for their propagation. They exhibit all the wave properties such as reflection, diffraction, interference etc. When sound encounters an obstacle it undergoes reflection as well as diffraction. Sound waves are reflected when the dimensions of the obstacle are large in comparison to the wavelength. Generally, in enclosed spaces such as auditoriums etc, the walls and ceiling are quite large. Therefore, reflection of sound plays a very important role in enclosed spaces. The diffraction of sound takes place when the size of the obstacle or opening is smaller compared to the wavelength. It occurs in enclosed spaces because of uneven surfaces, windows, doors etc. The diffraction tends to diffuse the sound uniformly in the enclosed spaces. Therefore, sound reflection alone becomes most important in the study of acoustics of halls and buildings.

The reflection of sound in an enclosed space leads to two important effects, namely *echo* and *reverberation*. Echoes and reverberation are both reflections of sound. A reflection is called an **echo** if the time between the original sound and its reflection is long enough that both sounds can be heard distinctly. If a room has lots of echoes and they are closely spaced in time so that they are not discernible, then this large number of echoes is known as **reverberation**.

11.9.1 Echoes

An **echo** is produced when the sound reflected from an obstacle reaches the ear after the sound from the source has already been heard. Thus, there is a repetition of the sound in this case. The sensation of sound persists for about 100 ms after the source stopped giving sound. Hence, in order that an echo may be heard as distinctly separate, it must reach the ear 100 ms later than the direct sound. When the distance of the obstacle from the source is 17 m or more, echo will be heard distinctly. When the reflection arrives within 60 ms or less after the original sound, the listener will not hear the reflection as distinct echo. When sound is reflected from a number of reflecting surfaces, multiple echoes are heard. Thus, the heavy rolling sound of a thunder is a result of successive reflections of sound from clouds, mountains, rocks etc reflecting surfaces.

11.9.2 Reverberation

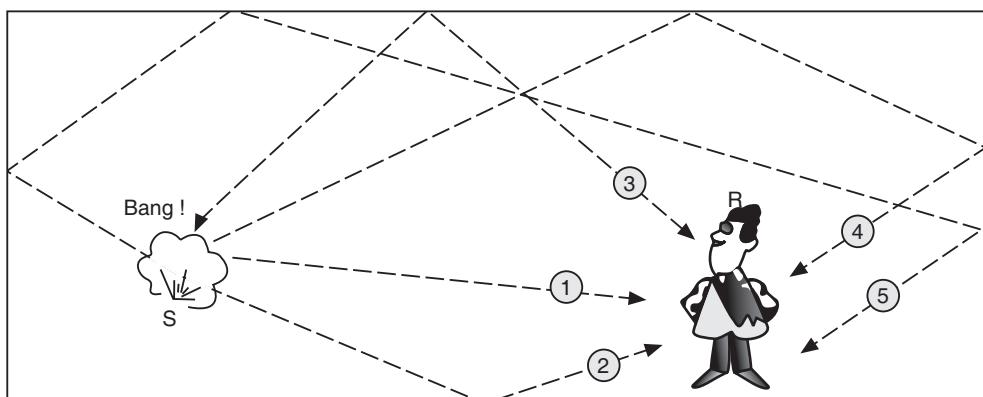


Fig. 11.3

Sound produced in an enclosure does not die out immediately after the source has ceased to produce it. A sound produced in a hall undergoes multiple reflections from the walls, floor and ceiling before it becomes inaudible. A person in the hall continues to receive successive reflections of progressively diminishing intensity (Fig. 11.3). This prolongation of sound before it decays to a negligible intensity is called **reverberation**.

Some reverberation is often desirable, especially in a hall used for musical performance. A small amount of reverberation improves the original sound. However, too much reverberation causes boomy sound quality in a musical performance. Speeches given in such a hall would be unintelligible. Reverberation is a familiar phenomenon experienced in halls without furniture.

Note that the reverberation of sound pertains to enclosed spaces only. In open air the sound spreads out in all directions without repeated reflections.

11.10 REVERBERATION TIME

The time taken by the sound in a room to fall from its average intensity to inaudibility level is called the **reverberation time** of the room. Reverberation time is defined as the time during which the sound energy density falls from its steady state value to its one-millionth (10^{-6})

value after the source is shut off. That is $\frac{I_{\text{final}}}{I_{\text{initial}}} = 10^{-6}$. We can also express reverberation

time in terms of sound energy level in dB as follows. If initial sound level is L_i and the final level is L_f , then we can write

$$L_i = 10 \log \frac{I_i}{I} \quad \text{and} \quad L_f = 10 \log \frac{I_f}{I}$$

$$\therefore L_i - L_f = 10 \log \frac{I_i}{I_f} \quad (11.13)$$

$$\text{As } \frac{I_{\text{initial}}}{I_{\text{final}}} = 10^6, \quad L_i - L_f = 10 \log 10^6 = 60 \text{ dB}$$

Thus, the reverberation time is the period of time in seconds, which is required for sound energy to diminish by 60 dB after the sound source is stopped.

11.11 SOUND ABSORPTION

When sound is incident on the surface of any medium, it splits into three parts. One part is reflected from the surface; another part gets absorbed in the medium, while the remaining part is transmitted through the medium and emerges on the other side. The property of a surface by which sound energy is converted into other form of energy is known as **absorption**. In the process of absorption sound energy is converted into heat due to frictional resistance inside the pores of the material. The fibrous and porous materials absorb sound energy more, than other solid materials.

11.11.1 Absorption Coefficient

Different surfaces absorb sound to different extents. The effectiveness of a surface in absorbing sound energy is expressed with the help of absorption coefficient. *The coefficient of absorption ‘ α ’ of a material is defined as the ratio of sound energy absorbed by its surface to that of the total sound energy incident on the surface.* Thus,

$$\alpha = \frac{\text{Sound energy absorbed by the surface}}{\text{Total sound energy incident on the surface}} \quad (11.14)$$

In order to compare the relative efficiency of different absorbing surfaces, it is essential to select a standard in terms of which all surfaces can be described. A unit area of open window is selected as the standard. The entire sound incident on an open window is fully transmitted and none is reflected. Therefore, it is considered as an **ideal absorber of sound**. Thus the unit of absorption is the open window unit (O.W.U.), which is named a “**sabin**” after the scientist who established the unit. A 1 m² sabin is the amount of sound absorbed by one square metre area of fully open window. Table-3 lists the absorption coefficients of various materials.

The value of ‘ α ’ depends on the nature of the material as well as the frequency of sound. The greater the frequency the larger is the value of ‘ α ’ for the same material. Therefore, the values of ‘ α ’ for a material are determined for a wide range of frequencies. It is a common practice to use the value of ‘ α ’ at 500 Hz in acoustic designs.

Table 3: Absorption coefficients of some materials

Material	Absorption coefficient per m ² at 500 Hz
Open window	1.00
Ventilators	0.10 to 0.50
Stage curtain	0.20
Curtains in heavy folds	0.40 to 0.75
Carpet	0.40
Audience (One adult in upholstered seat)	0.46
Fibrous plaster, Straw board	0.30
Perforated compressed fibre board	0.55
Woodwool board	0.20
Concrete	0.17
Marble	0.01

If a material has the value of “ α ” as 0.5, it means that 50% of the incident sound energy will be absorbed per unit area. If the material has a surface area of S sq.m., then the total absorption provided by that material is

$$a = \alpha \cdot S \quad \text{sabins} \quad (11.15)$$

If there are different materials in a hall, then the total sound absorption by the different materials is given by

$$\begin{aligned} A &= a_1 + a_2 + a_3 + \dots \\ A &= \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \dots \end{aligned} \quad (11.16)$$

$$\text{or } A = \sum_{n=1}^n \alpha_n S_n \quad (11.17)$$

where $\alpha_1, \alpha_2, \alpha_3, \dots$ are absorption coefficients of materials with areas S_1, S_2, S_3, \dots

11.12 SABINE'S FORMULA FOR REVERBERATION TIME

Prof. Wallace C. Sabine (1868-1919) determined the reverberation times of empty halls and furnished halls of different sizes and arrived at the following conclusions.

- (i) The reverberation time depends on the reflecting properties of the walls, floor and ceiling of the hall. If they are good reflectors of sound, then sound would take longer time to die away and the reverberation time of the hall would be long.
- (ii) The reverberation time depends directly upon the physical volume V of the hall.
- (iii) The reverberation time depends on the absorption coefficient of various surfaces such as carpets, cushions, curtains etc present in the hall.
- (iv) The reverberation time depends on the frequency of the sound wave because absorption coefficient of most of the materials increases with frequency. Hence high frequency sounds would have shorter reverberation time.

Prof. Sabine summarized his results in the form of the following equation.

$$\text{Reverberation Time, } T \propto \frac{\text{Volume of the Hall, } V}{\text{Absorption, } A}$$

$$\text{or } T = k \frac{V}{A}$$

where k is a proportionality constant. It is found to have a value of 0.161 when the dimensions are measured in metric units. Thus,

$$T = \frac{0.161 V}{A} \quad (11.18)$$

where A is given by the relation (11.16). Equation (11.18) is known as Sabine's formula for reverberation time. It may be rewritten as

$$T = \frac{0.161 V}{\sum_{n=1}^n \alpha_n S_n} \quad (11.19)$$

$$\text{or } T = \frac{0.161 V}{\alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \dots + \alpha_n S_n} \quad (11.20)$$

The Sabine equation works well for large enclosures.

11.12.1 Optimum Reverberation Time

Sabine determined the time of reverberation for halls of various sizes. In these measurements, he used an organ pipe as the source, which was blown at a definite frequency and under a constant pressure. The instant of cutting off of the sound and the instant at which the observer

ceased to hear the sound were recorded. And from the results, he deduced the reverberation time that is likely to be most satisfactory for the purpose for which a hall is built. Such satisfactory value is known as the **optimum reverberation time**. The optimum reverberation time for music or speech lies between 0.6 to 0.75 sec whereas it lies between 0.8 to 1.00 for orchestra. The reverberation time of a hall can be adjusted to a desired value by arranging absorbent materials for the various surfaces of the hall.

Example 11.1: A hall has a volume of 1200 m^3 . Its total absorption is equivalent to 480 m^2 of open window. What will be the effect on the reverberation time if audience fill the hall and thereby increase the absorption by another 480 m^2 of open window?

Solution: Reverberation time

$$T_1 = \frac{0.161 V}{\sum \alpha S} = \frac{0.161 \times 1200 \text{ m}^3}{480 \text{ m}^2} = 0.40 \text{ s}$$

When the audience are present in the hall, the reverberation time is

$$T_2 = \frac{0.161 V}{\sum \alpha S} = \frac{0.161 \times 1200 \text{ m}^3}{(480 + 480) \text{ m}^2} = 0.20 \text{ s.}$$

Example 11.2: A classroom has dimensions $20 \times 15 \times 5 \text{ m}^3$. The reverberation time is 3.5 sec. Calculate the total absorption of its surfaces and the average absorption coefficient.

Solution:

$$T = \frac{0.161 V}{\sum \alpha S}$$

$$\therefore \sum \alpha S = \frac{0.161 (20 \times 15 \times 5) \text{ m}^3}{3.5 \text{ s}} = 69$$

The surface areas of the walls, ceiling and floor of the room = $2(20 \times 15 + 15 \times 5 + 5 \times 20) \text{ m}^2$

$$\therefore \alpha_{\text{average}} = \frac{69}{2(20 \times 15 + 15 \times 5 + 5 \times 20)} = \frac{69}{950} = 0.07$$

11.13 REVERBERATION THEORY

The Sabine formula (11.18) was originally an empirical formula. It was later derived from reverberation theory. The reverberation theory explains the nature of growth and decay of sound energy in an enclosure. The following assumptions are made in the theory.

- (i) The enclosure (i.e. room) is large enough such that the sound energy is uniformly distributed in it.
- (ii) Sound travels uniformly in all directions.
- (iii) Absorption of sound by the air is neglected.

Let us consider a sufficiently big hall. Let a source produce sound in the hall. We assume that the sound fills the hall uniformly and hence each elemental portion of the volume of hall can be regarded as a source of sound energy. We first calculate the radiation of energy from the volume

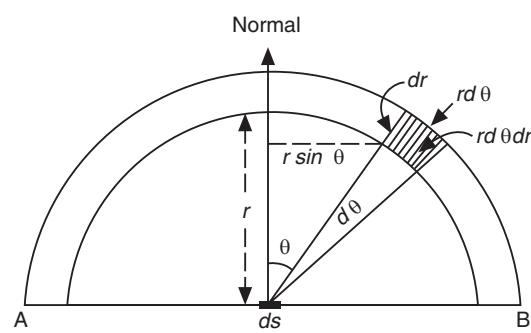


Fig. 11.4

element dV in the hall and then the rate at which it is incident on the surface element ds of a plane wall, AB (see Fig. 11.4).

Let us draw a normal from the center of the element ds . From the same center let us draw two circles with radii r and $r + dr$ in the plane containing the normal. Then two radii are drawn w.r.t. the normal at angles θ and $\theta + d\theta$. Consider the portion of the area (shaded area in Fig. 11.4) lying between the two radii and between the circles.

The arc length of this shaded area = $r d\theta$.

Radial length = dr

\therefore Surface area of this element = $r d\theta dr$.

If the shaded elemental area is rotated through a small angle $d\phi$, this elemental area sweeps through a distance $r \sin\theta d\phi$.

Volume traced out by the elemental area, $dV = (rd\theta dr)(r \sin\theta d\phi) = r^2 \sin\theta d\theta dr d\phi$

The sound energy present in this volume element at any moment = $E r^2 \sin\theta d\theta dr d\phi$ where E is the sound energy per unit volume. Since the energy from the volume element travels uniformly in all directions, the energy traveling per unit solid angle along any direction is

$$dW = \frac{E dV}{4\pi}$$

The solid angle subtended by ds at dV is

$$d\Omega = \frac{ds \cos\theta}{r^2}$$

The amount of energy that reaches ds from dV is equal to the product of the energy traveling per unit solid angle and the solid angle subtended by ds at the volume element dV . Thus,

The amount of energy that reaches ds from dV

$$\begin{aligned} &= \frac{E dV}{4\pi} \cdot \frac{ds \cos\theta}{r^2} \\ &= \frac{E ds \sin\theta \cos\theta d\theta d\phi dr}{4\pi} \end{aligned}$$

The total energy received by ds in one second from the entire volume is obtained by integrating the above expression. Thus,

$$\begin{aligned} \text{Total energy} &= \frac{E ds}{4\pi} \int_{\varphi=0}^{2\pi} \int_{\theta=0}^{\pi/2} \int_{r=0}^v \sin\theta \cos\theta d\theta d\phi dr \\ &= \frac{E ds}{4\pi} v \times 2\pi \times \int_{\theta=0}^{\pi/2} \sin\theta \cos\theta d\theta \\ &= \frac{Evds}{4} \end{aligned} \quad (11.21)$$

If α is the absorption coefficient of the surface of the wall AB, then the sound energy absorbed per second by the surface element ds is

$$dW_A = \frac{Ev\alpha dS}{4}$$

Therefore, the total energy absorbed by all wall surfaces in the hall is

$$W_A = \frac{Ev}{4} \sum \alpha ds$$

or

$$W_A = \frac{EvA}{4} \quad (11.22)$$

where A is the total absorption of all surfaces.

11.13.1 Build-up of Sound in a Hall

Let P be the power of sound source and V the total volume of the hall. The total energy in the hall at a particular instant will be EV where E is the energy density at that instant.

$$\text{Rate of growth of energy in the hall} = \frac{d}{dt}(EV) = V \frac{dE}{dt}$$

At any instant,

$$\begin{aligned} \left(\begin{array}{l} \text{Rate of growth of} \\ \text{energy in the hall} \end{array} \right) &= \left(\begin{array}{l} \text{Rate of supply of} \\ \text{energy from the source} \end{array} \right) - \left(\begin{array}{l} \text{Rate of absorption of} \\ \text{all surfaces in the hall} \end{array} \right) \\ \therefore V \frac{dE}{dt} &= P - \frac{EvA}{4} \\ \frac{dE}{dt} + \frac{vA}{4V} E &= \frac{P}{V} \end{aligned} \quad (11.23)$$

Putting $\frac{vA}{4V} = \alpha$, the above equation may be rewritten as

$$\frac{dE}{dt} + \alpha E = \frac{4P}{vA} \alpha$$

Multiplying with $e^{\alpha t}$ on both sides of the above equation, we get

$$\begin{aligned} \left[\frac{dE}{dt} + \alpha E \right] e^{\alpha t} &= \frac{4P}{vA} \alpha e^{\alpha t} \\ \frac{d}{dt} [E e^{\alpha t}] &= \frac{4P}{vA} \alpha e^{\alpha t} \end{aligned}$$

Integrating on both the sides, we get

$$E e^{\alpha t} = \frac{4P}{vA} e^{\alpha t} + K \quad (11.24)$$

where K is the constant of integration. The value of K may be found using the boundary conditions.

11.13.2 Growth of the Energy Density

If t is measured from the instant the sound source emits sound, the initial condition becomes $E = 0$ at $t = 0$. Using these initial conditions into equ. (11.24), we obtain the value of K .

$$K = -\frac{4P}{vA}$$

Using the value of K into equ.(11.24), we get

$$E = \frac{4P}{vA} (1 - e^{-\alpha t})$$

or

$$E = E_m (1 - e^{-\alpha t}) \quad (11.25)$$

where

$$E_m = 4P/vA$$

The equ.(11.25) indicates that the sound energy density grows in an exponential manner with time, t . It increases till it attains the steady state value E_m , at $t = \infty$.

11.13.3 Decay of Sound Energy in the Hall

Let us assume that after a certain time that the energy attained the steady state value, the source of the sound is switched off. Then, P becomes zero and let that instant be taken as $t = 0$. Then, the initial conditions would be $P = 0$ at $t = 0$ and $E = E_m$. Using these conditions into (11.24), we get

$$K = E e^{\alpha t} \quad (11.26a)$$

As $|K| = \frac{4P}{\nu A} = E_m$, we can write the above equation (11.26a) as

$$E = E_m e^{-\alpha t} \quad (11.26)$$

Equ. (11.26) indicates that the sound energy decays exponentially after the source of sound is switched off. Fig. 11.5 shows the growth and decay of sound energy in the hall.

11.13.4 Deduction of Sabine's Formula

The reverberation time T is defined as the time taken for the sound energy density to fall from its steady value to its one-millionth value. It means that

$$\frac{E}{E_m} = 10^{-6}, \text{ at } t = T.$$

It follows from equ. (11.26) that $e^{-\alpha t} = 10^{-6}$

$$\therefore e^{\alpha T} = 10^6 \quad (\text{using } t = T)$$

Taking logarithm on both the sides, we get

$$\alpha T = 6 \log_{10} 10 = 6 \times 2.3026$$

Using the expression for α , we get

$$\frac{\nu A}{4V} T = 6 \times 2.3026$$

or

$$T = \frac{4 \times 6 \times 2.3026 \times V}{\nu A}$$

Taking $\nu = 344$ m/s

$$T = \frac{4 \times 6 \times 2.3026 \times V}{(344) A} = \frac{0.161 V}{A}$$

The above equation is identical to the empirical formula (11.18) that Sabine proposed for reverberation time.

Limitations of Sabine's formula

- (i) Sabine's formula does not give correct result for absorption coefficient more than 0.2.
- (ii) Sabine's formula gives contradictory result in case of a dead room. In case of complete absorption, $\alpha = 1$ and therefore, reverberation time T should be zero. But according to Sabine's formula, we have $T = 0.161 V$ seconds.

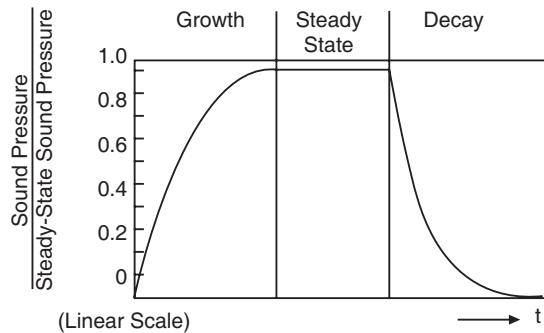


Fig. 11.5

- (iii) In the derivation of Sabine's formula it is assumed that sound energy distributes uniformly in the room and that there is no loss of energy in the air. However, this is not a practical reality.

11.14 DETERMINATION OF ABSORPTION COEFFICIENT

There are two methods which are used for the measurement of absorption coefficients of acoustic materials.

Method 1: One of the methods is based on the determination of reverberation times of a room without any material and with the material under test. If T_1 is the reverberation time of in the empty room, then

$$T_1 = \frac{0.161V}{\sum_{n=1}^{\infty} \alpha_n S_n} = \frac{0.161V}{A}$$

where $A = \sum \alpha_n S_n$ denotes the absorption due to the walls, flooring and ceiling of the empty room.

Then a certain amount of absorbing material of area S and absorption coefficient α' is added in the room and again the reverberation time is measured. Let it be T_2 .

$$\begin{aligned} T_2 &= \frac{0.161V}{A + \alpha'S}. \\ \text{Then } \frac{1}{T_2} - \frac{1}{T_1} &= \frac{\alpha'S}{0.161V} \\ \therefore \alpha' &= \frac{0.161V}{S} \left[\frac{1}{T_2} - \frac{1}{T_1} \right] \end{aligned} \quad (11.27)$$

Knowing the quantities on the right hand side of the above equation, the absorption coefficient α' of the material under test can be calculated.

Method 2: The above method cannot be used if the absorbing material is already fixed to the walls and ceiling in the room. The method adopted for such cases, consists in finding the reverberation times for two sources of differing emitting powers.

Let the powers of the two sources be P_1 and P_2 respectively. The steady state energy densities of the sources will be

$$E_1 = \frac{4P_1}{\nu A} \quad \text{and} \quad E_2 = \frac{4P_2}{\nu A}$$

Let T_1 and T_2 be the respective times of decay of energy density to the inaudibility level E_0 . Then,

$$E_0 = \frac{4P_1}{\nu A} e^{-\alpha T_1}$$

$$\text{and } E_0 = \frac{4P_2}{\nu A} e^{-\alpha T_2}$$

$$\therefore \frac{P_2}{P_1} = e^{\alpha(T_2 - T_1)}$$

$$\text{or } \alpha = \frac{\log_e P_2 - \log_e P_1}{(T_2 - T_1)}$$

But $\alpha = \frac{vA}{4V}$

$$\therefore \frac{\log_e P_2 - \log_e P_1}{(T_2 - T_1)} = \frac{vA}{4V}$$

$$A = \frac{4V \log_e(P_2 / P_1)}{v(T_2 - T_1)}$$

$$\alpha S = \frac{4V \log_e(P_2 / P_1)}{v(T_2 - T_1)}$$

$$\alpha = \frac{4V \log_e(P_2 / P_1)}{vS(T_2 - T_1)} \quad (11.28)$$

Knowing the quantities on the right hand side of the above equation, we can calculate the absorption coefficient of the material fixed in the room.

Example 11.3: A hall has a volume of 5000 m^3 . It is required to have reverberation time of 1.5 second. What should be the total absorption in the hall?

Solution:

$$T = \frac{0.161V}{\sum \alpha S}$$

$$\therefore \sum \alpha S = \frac{0.161 \times 5000 \text{ m}^3}{1.5 \text{ s}} = 537 \text{ O.W.U.m}^2$$

Example 11.4: The reverberation time is found to be 1.5 sec for an empty hall and it is found to be 1 sec when a curtain cloth of 20 m^2 is suspended at the center of the hall. If the dimensions of the hall are $10 \times 8 \times 6 \text{ m}^3$, calculate the coefficient of absorption of curtain cloth.

Solution: Absorption of empty hall

$$A = \frac{0.161(10 \times 8 \times 6) \text{ m}^3}{1.5 \text{ s}} = 51.52$$

Reverberation time of the hall with curtain is

$$1 \text{ s} = \frac{0.161(10 \times 8 \times 6)}{51.52 + \alpha(2 \times 20)}$$

The factor 2 in the denominator takes into account the two sides of the curtain.

\therefore The coefficient of absorption of curtain cloth

$$\alpha = \frac{77.28 - 51.52}{40} = 0.64$$

11.15 FACTORS AFFECTING ACOUSTICS OF BUILDINGS AND THEIR REMEDIES

A hall or auditorium designed for lectures or concerts is regarded to have right acoustical quality when the following conditions prevail in it.

- (i) The initial sound from the source should be of adequate intensity.
- (ii) The sound should be evenly spread over the whole area covered by the audience.

- (iii) The successive sounds in the speech or music should be clear and distinct; and the tonal quality of music is not altered.
- (iv) All undesired sound should be reduced to the extent that it will not interfere with the normal hearing or speech.

There are several factors that affect the acoustical quality of a hall. We discuss here seven common acoustical defects.

1. Reverberation Time: If a hall is to be acoustically satisfactory, it is essential that it should have the right reverberation time. The reverberation time should be neither too long nor too short. A very short reverberation time makes a room 'dead'. On the other hand, a long reverberation time renders speech unintelligible. The optimum value for reverberation time depends on the purpose for which a hall is designed. A reverberation time of 0.6 sec is acceptable for speeches and lectures, while a reverberation time of 1 to 2 secs is satisfactory for concerts. In case of theatres the optimum value varies with the volume. For small theatres 1.1 to 1.5 secs is suitable whereas for large theatres it may go up to 2.3 secs.

Remedies: The reverberation time can be controlled by the suitable choice of building materials and furnishing materials. If the reverberation time of a hall is too long, it can be cut down by increasing the absorption or reducing volume and if it is too short, it can be increased by changing high absorption materials to materials of low absorption or increasing volume.

Since open windows allow the sound energy to flow out of the hall, there should be a limited number of windows. They may be opened or closed to obtain optimum reverberation time.

Cardboard sheets, perforated sheets, felt, heavy curtains, thick carpets etc are used to increase wall and floor surface absorption. Therefore, the walls are to be provided with absorptive materials to the required extent and at suitable places. Heavy fold curtains may be used to increase the absorption. Covering the floor with carpet also increases the absorption.

Audience also contribute to absorption of sound. The absorption coefficient of an individual is about 0.45 sabins. In order to compensate for an increase in the reverberation time due to an unexpected decrease in audience strength, upholstered seats are to be provided in the hall. Absorption due to an upholstered chair is equivalent to that of an individual. In the absence of audience the upholstered chair absorbs the sound energy and it does not contribute to absorption when it is occupied.

2. Loudness: Sufficient loudness at every point in the hall is an important factor for satisfactory hearing. Excessive absorption in the hall or lack of reflecting surfaces near the sound source may lead to decrease in the loudness of the sound.

Remedies: A hard reflecting surface positioned near the sound source improve the loudness. Polished wooden reflecting boards kept behind the speaker and sometimes above the speaker will be helpful.

Low ceilings are also of help in reflecting the sound energy towards the audience. Adjusting the absorptive material in the hall will improve the situation.

When the hall is large and audience more, loud speakers are to be installed to obtain the desired level of loudness.

3. Focussing: Reflecting concave surfaces cause concentration of reflected sound, creating a sound of larger intensity at the focal point (O in Fig. 11.6). These spots are known

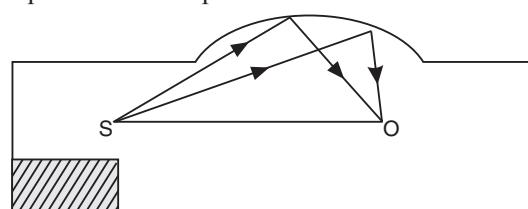


Fig. 11.6

as **sound foci**. Such concentrations of sound intensity at some points lead to deficiency of reflected sound at other points. The spots of sound deficiency are known as *dead spots*. The sound intensity will be low at dead spots and inadequate hearing. Further, if there are highly reflecting parallel surfaces in the hall, the reflected and direct sound waves may form standing waves which leads to uneven distribution of sound in the hall.

Remedies: The sound foci and dead spots may be eliminated if curvilinear interiors are avoided. If such surfaces are present, they should be covered with highly absorptive materials. Suitable sound diffusers are to be installed in the hall to cause even distribution of sound in the hall.

A paraboloidal reflecting surface arranged with the speaker at its focus is helpful in directing a uniform reflected beam of sound in the hall (Fig. 11.7).

4. Echoes: When the walls of the hall are parallel, hard and separated by about 17m distance, echoes are formed. Curved smooth surfaces of walls also produce echoes.

Remedies: This defect is avoided by selecting proper shape for the auditorium. Use of splayed side walls (see Fig. 11.9) instead of parallel walls greatly reduces the problem and enhance the acoustical quality of the hall.

Echoes may be avoided by covering the opposite walls and high ceiling with absorptive material.

5. Echelon effect: If a hall has a flight of steps, with equal width, the sound waves reflected from them will consist of echoes with regular phase difference (Fig. 11.8). These echoes combine to produce a musical note which will be heard along with the direct sound. This is called **echelon effect**. It makes the original sound unintelligible or confusing.

Remedies: It may be remedied by having steps of unequal width.

The steps may be covered with proper sound absorbing materials, for example with a carpet.

6. Resonance: Sound waves are capable of setting physical vibration in surrounding objects, such as window panes, walls, enclosed air etc. The vibrating objects in turn produce sound waves. The frequency of the forced vibration may match some frequency of the sound produced and hence results in resonance phenomenon. Due to the resonance, certain tones of the original music may get reinforced and may result in distortion of the original sound.

In a hall the whole air mass vibrates if sound is continuously produced from a source. The vibration of air in turn adds to the resonant frequencies of the hall depending on its dimensions. If lower modes of resonant frequencies are excited by the source, the sound distribution in the hall will be erratic.

Remedies: The vibrating bodies may be suitably damped to eliminate resonance due to them.

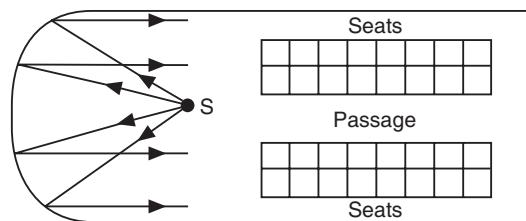


Fig. 11.7

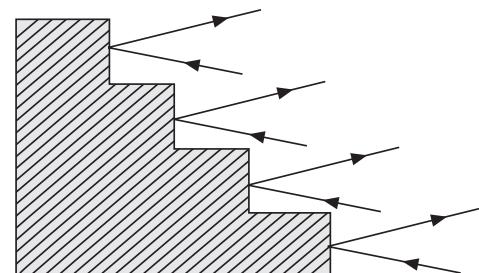


Fig. 11.8

In larger halls, the resonant frequencies are quite low. Hence by selecting larger halls resonance defect can be eliminated.

7. Noise: Noise is unwanted sound which masks the satisfactory hearing of speech and music. There are mainly three types of noises that are to be minimized. They are (i) *air-borne noise*, (ii) *structure-borne noise* and (iii) *internal noise*. Internal noise may be reduced by adding absorbing materials on walls, ceiling etc. But the other two types of noises get transmitted through the structural links with neighbouring structures. In these cases, noise attenuation becomes important. Acoustical (absorptive) materials cannot help in these cases. Acoustical materials only minimize reflection but cannot minimize transmission of sound. It is therefore, essential that the structural barriers of the building must be such that they transmit minimum energy. Thus, to increase the transmission loss of sound energy, cavity walls, compound walls, double-pane window construction, floating floors etc are used.

- (i) The noise that comes into building through air from distant sources is called **air-borne noise**. A part of it directly enters the hall through the open windows, doors or other openings while another part enters by transmission through walls and floors.

Remedies: The building may be located on quite sites away from heavy traffic, market places, railway stations, airports etc. They may be shaded from noise by interposing a buffer zone of trees, gardens etc.

- (ii) The noise which comes from impact sources on the structural extents of the building is known as the **structure-borne noise**. It is directly transmitted to the building by vibrations in the structure. The common sources of this type of noise are foot-steps, moving of furniture, operating machinery etc.

Remedies: The problem due to machinery and domestic appliances can be overcome by placing vibration isolators between machines and their supports.

Cavity walls, compound walls may be used to increase the noise transmission loss and keep the noise in the building at desired level.

- (iii) **Internal noise** is the noise produced in a hall or office etc. They are produced by air conditioners, movement of people etc.

Remedies: The walls, floors and ceilings may be provided with enough sound absorbing materials.

The gadgets or machinery should be placed on sound absorbent material.

Split-type air conditioners etc are to be used.

11.16 ACOUSTIC DESIGN OF A HALL

There are several factors that determine the acoustical quality of a hall. We discuss here six major factors that are to be considered in the acoustic design of a hall.

1. Site Selection: A proper site with quite surroundings is to be selected for an auditorium. It should be away from the busy highway vehicular traffic, rail traffic, airport or any other noisy location. Otherwise, the vibrations produced by the traffic will be conveyed into the hall through the structures, which contribute to the noise in the hall. Elaborate and costly arrangements will have to be made to reduce the noise level. For auditoriums, without air-conditioning, and hence require doors and windows to be kept open during performance, the orientation of the hall should be such that the external noise is maintained at low level. When air-conditioning is provided, care should be taken to reduce the plant noise and grill noise. Through an appropriate orientation, layout and structural design, the background noise level in the hall should be kept at around 45 dB.

2. Volume: The hall should be big enough so that sound intensity spreads uniformly over its entire area. Smaller rooms lead to irregular distribution of sound because of formation of standing waves. When the length of the hall, L is very large in comparison to the longest wavelength of sound, the room is considered to be large in the acoustical sense and the sound within such a hall may be regarded as spread uniformly.

The floor area of the hall is computed, excluding the stage, based on the requirement of 0.6 to 0.9 m²/person. The height of the hall is determined by the presence or absence of the balcony, ventilation requirement etc. An average height of 6 m for small halls and 7.5 m for large halls are usually adopted. It is desirable to provide slight increase in the height of ceiling near the center of the hall.

The recommended volumes for different types of auditoriums are as follows;

- (a) Concert halls 4.0 to 5.5 m³/person
- (b) Theatres 4.0 to 5.0 m³/person
- (c) Public lecture halls 3.5 to 4.5 m³/person

3. Shape: The shape of the hall plays a very important role in determining its acoustical quality. The side walls and ceiling are potentially useful reflecting surfaces and should be carefully designed to maximize their usefulness. The rear walls and floors are potential sources of useless and harmful reflections which are to be avoided. Parallel hard walls create echo problems. Use of splayed side walls greatly reduce the problem and enhance the acoustical quality of the room. In view of this a fan-shaped floor plan (See Fig. 11.9) is preferred. The side walls are arranged to have an angle of not more than 100° with the curtain line. The fan shaped plan provides favourable reflection of sound from sides.

A concave surface within the hall is not desirable because it focuses sound reflections. Such surface must be broken up with smaller convex surfaces so that sound is diffused in all directions.

4. Reverberation: If a hall is to be acoustically satisfactory, it is essential that it should have the right reverberation time. The reverberation time should not be either too long or too short. A very short reverberation time makes a room ‘dead’. On the other hand, a long reverberation time renders speech unintelligible. The optimum value for reverberation time

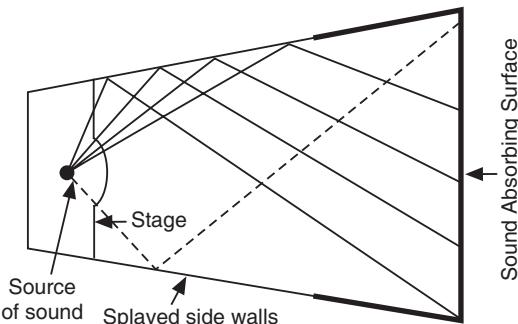


Fig. 11.9

Table 4: Reverberation Time and Acoustics

Reverberation time (seconds)	Acoustics
0.50 to 1.50	Excellent
1.50 to 2.00	Good
2.00 to 3.00	Satisfactory
3.00 to 5.00	Bad
> 5.00	Very bad

depends on the purpose for which the hall is designed. A reverberation time of 0.5 secs is acceptable for speeches and lectures, while a reverberation time of 1 to 2 secs is satisfactory for concerts. In case of theatres, the optimum value varies with volume. 1.1 to 1.5 secs is suitable for small theatres, whereas for large theatres it may go up to 2.3 secs.

Sabine equation can be applied to the acoustical design as well as for acoustical corrections of the halls. For a hall of a given volume V, the value of $\sum aS$ which would give the optimum reverberation time can be determined. Even before the hall is built, the surface area of absorbent material can be computed right from the architect's plan. The reverberation time can be controlled by suitable choice of building materials and furnishing materials. If the reverberation time is too long, it can be cut down by increasing the absorption and if it is too short, it can be increased by replacing high absorption materials with materials of low absorption. Since open windows allow the sound energy to flow out of the hall, there should be a limited number of windows. Cardboard sheets, perforated sheets, felt, heavy curtains, thick carpets etc are used to increase wall and floor surface absorption. As already mentioned earlier, audience also contribute to absorption of sound. The audience absorb sound more in high frequency range than in the middle or in low frequency range. It is therefore, desirable to add special low frequency absorbers on ceiling and walls in order to achieve optimum reverberation time over as wide a frequency range as possible.

Table 5: Optimum Reverberation Time for Halls

Activity in Hall	Optimum Reverberation Time (sec)	Audience Factor
Conference halls	1 to 1.5	One-third
Cinema theatre	1.3	Two-thirds
Assembly halls	1 to 1.5	Quorum
Public lecture halls	1.5 to 2	One-third
Music concert halls	1.5 to 2	Full
Churches	1.8 to 3	Two-thirds
Large halls	2 to 3	Full

5. Seating Arrangement: The seats should be arranged in concentric arcs of the circles. Flat floor seating of more than a few rows is deprived of good visibility and good hearing. Sloped floor seating is essential for a large audience to have good visibility and good acoustics. The successive rows of seats have to be raised over the preceding ones, with the result that the floor level rises towards the rear end. The rise in level may be about 8 to 12 cms per row. Further, the seats in each row should be staggered sideways in relation to those in front so that the line of sight of a person in any row is not obstructed by the person sitting in front of him. The back to back distance of chairs in successive rows should be at least 75 cms and this may be increased up to 106 cms for extra comfort.

The angle subtended with the horizontal at the front-most observer, by the highest object to be seen on the stage, should not exceed 30°. On this basis, the distance of the first row should be about 4.5 m for movie watching and 3.6 m for theatres. In case of movies, synchronization of sound with lip movements is most essential; and in case of dramas, a person with normal vision should be able to discern facial expressions of the performers. In order to satisfy these conditions, it is recommended that the distance of the farthest seat from the curtain line should not exceed 23 m.

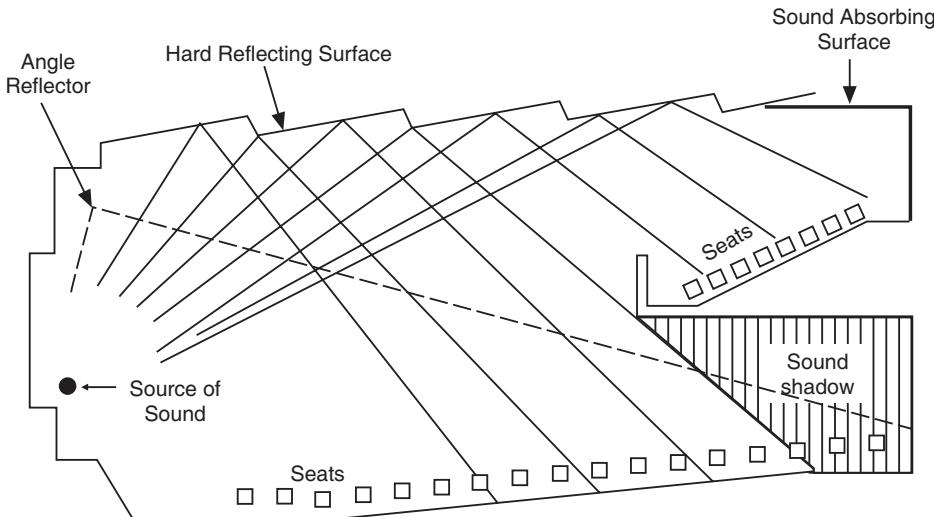


Fig. 11.10

When balcony is provided, its projection L_1 into the hall should not be more than twice the free height H_1 of opening of balcony recess ($L_1 \leq 2H_1$). If balconies are too deep, sound shadow forms and the persons in the seats below the balcony do not receive ceiling reflections (see Fig. 11.10). Suitable sound reflectors should be positioned at appropriate places to get rid of this defect.

6. Acoustic Treatment of Interior Surfaces: The interior surface of the hall should be given utmost attention to make the hall acoustically satisfactory. If the side walls are parallel, they are to be covered with absorbent materials from a length of about 7.5 m from the proscenium end. As the reflections from the near wall are of no use, the rear wall should be covered with absorbents. In large halls, a false ceiling is usually provided. The false ceiling positioned near the proscenium should be constructed of reflective material and inclined in a proper way to help reflections of sound from the stage to reach the rear seats in the hall. Concave shaped ceilings in the form of dome should be avoided. The rear portion of the ceiling may be treated with sound absorbing material so that build-up of audience noise is prevented. The floor should be covered with a carpet. Carpet on the floor not only covers a useless reflecting surface but also greatly reduces audience noise.

Example 11.5: For an empty assembly hall of size $20 \times 15 \times 10 \text{ m}^3$ the reverberation time is 3.5 sec. Calculate the average absorption coefficient of the hall. What area of the wall should be covered by the curtain so as to reduce the reverberation time to 2.5 sec. Given the absorption coefficient of curtain cloth is 0.5.

Solution: Total absorption of the empty hall

$$\begin{aligned}
 A &= \frac{0.161V}{T_1} \\
 &= \frac{0.161(20 \times 15 \times 10)}{3.5} = 138
 \end{aligned} \tag{i}$$

Average absorption coefficient

$$a_{av} = \frac{138}{2(20 \times 15 + 15 \times 10 + 20 \times 10)} = 0.106$$

When the walls are covered with curtain cloth of surface area S_1 , the reverberation time T_2 becomes

$$T_2 = \frac{0.161V}{A + a_m S_1 - a_{av} S_1} \quad (ii)$$

where a_m is the absorption coefficient of curtain cloth.

Using (i) and (ii), we get

$$(a_m - a) = \frac{0.161V}{S_1} \left[\frac{1}{T_2} - \frac{1}{T_1} \right]$$

or

$$S_1 = \frac{0.161V}{(a_m - a)} \left[\frac{1}{T_2} - \frac{1}{T_1} \right]$$

$$S_1 = \frac{0.161(20 \times 15 \times 10) m^3}{0.5 - 0.106} \left[\frac{1}{3.5} - \frac{1}{2.5} \right]$$

The area of the wall to be covered with curtain

$$S_1 = 140 \text{ m}^2$$

Example 11.6: A hall has a volume of 2265 m^3 and its total absorption is equivalent to 92.9 m^2 of open window. What will be the effect on reverberation time if audience fill the hall and thereby increase the absorption by another 92.9 m^2 ?

Solution: Reverberation time, $T = \frac{0.161V}{\sum \alpha S}$

Initial reverberation time, $T_i = \frac{0.161 \times 2265 \text{ m}^3}{92.9 \text{ m}^2} = 3.9 \text{ s.}$

With the audience of the total absorption

$$= 92.9 \text{ m}^2 + 92.9 \text{ m}^2 = 185.8 \text{ m}^2.$$

Final reverberation time, $T_f = \frac{0.161 \times 2265 \text{ m}^3}{185.8 \text{ m}^2} = 1.95 \text{ s.}$

QUESTIONS

1. What is the nature of sound waves?
2. How do we classify sound? (Anna Univ., 2005)
3. What is the frequency range of sonic waves?
4. What is the difference between musical sounds and noise?
5. What are the characteristics of musical sounds?
6. Define noise.
7. Define intensity of sound. What is its unit? (Anna Univ., 2006)
8. How does loudness differ from intensity of a sound?
9. What are the units of measurement of loudness and of intensity?
10. Define decibel. (Anna Univ., 2005)

11. What is standard intensity? Give its value. (G.T.U., 2009)
12. Define phon and sone.
13. What is meant by reverberation? Is it desirable to have in a building?
14. What does sound absorption coefficient signify?
15. Why does acoustics play an important role in building designs?
16. State Weber-Fechner law. (Anna Univ., 2006)
17. What is reverberation time? Using Sabine's formula, explain how the sound absorption coefficient of a material is determined.
18. (a) Define reverberation time of a hall. Explain clearly what causes reverberation and how it can be minimized.
 (b) Explain the various requirements of a good auditorium.
 (c) Define absorption coefficient of a material and describe a method for its determination. (Bombay Univ.)
19. Explain the terms reverberation and reverberation time. Deduce the Sabine's formula for the reverberation time. (Andhra Univ.)
20. Explain various factors affecting architectural acoustics and their remedies.
21. Discuss the factors reverberation, resonance, echelon effect, focusing and reflection that affect the acoustics in hall and the remedies for them.
22. State and explain Sabine's formula for reverberation time of a hall. Derive Sabine's formula for reverberation time. (Anna Univ., 2005, 2007)
23. Explain how the reverberation time of a hall is affected by (a) its size (b) nature of its wall surfaces and (c) audience.
24. Define absorption coefficient of a material and describe a method for its determination. (C.S.V.T.U., 2008)
25. Explain: Reverberation. Describe a method for determination of the sound absorption coefficient of material. (Bombay Univ.)
26. State and explain Sabine's formula for reverberation time and describe briefly how one can determine the sound absorption coefficient for materials like curtain cloth. (Bombay Univ.)
27. (a) A carpet is to be used to cove the floor in a music hall for absorption of sound. Describe the method to determine absorption coefficient of the carpet.
 (b) Explain reflection, reverberation and echo in case of sound in a hall. (Bombay Univ.)
28. Explain the terms: (i) reflection (ii) reverberation and (iii) echo of sound energy and then show graphically, the nature of growth and decay of sound energy in a hall due to reverberation (Given rate of absorption of energy by walls = $\frac{1}{4}$ ECA, where E is energy density, C is velocity of sound and A is total absorption.) (C.S.V.T.U.,2006)
29. What is meant by reverberation? Discuss Sabine's formula. (G.T.U.,2009)
30. Define reverberation. Prove that total absorption at all the surfaces of the wall where the sound is falling is equal to ECA where E is energy density, C is velocity of sound and A is total absorption. (C.S.V.T.U.,2007)
31. Write down the Sabine's formula. Explain the terms involved in it and describe the units of each of them.
32. What are the limitations of Sabine's formula? Discuss.
33. Explain how the absorption coefficient of an acoustic material is determined.
34. State the acoustic requirements of a good auditorium. Explain how these requirements can be achieved.

35. Write an essay on the factors affecting architectural acoustics. Give remedies. (Anna Univ., 2006)
36. Explain how the reverberation time of a hall is affected by
(i) size (ii) nature of its wall surfaces (iii) audience.
37. What is echelon effect?
38. State any five factors affecting the acoustics of the building and give atleast two remedies for each. (G.T.U.,2009)

PROBLEMS

1. The volume of a room is 1500 m^3 . The wall area of the room is 260 m^2 , the floor area is 1400 m^2 and the ceiling area is 140 m^2 . The average sound absorption coefficient for wall is 0.03, for the ceiling is 0.80 and the floor is 0.06. Calculate the average absorption coefficient and the reverberation time.
[Ans: 0.237 sabins; 1.93 s]
2. A hall has a volume of 12500 m^3 and reverberation time of 1.5 sec. If 200 cushioned chairs are additionally placed in the hall, what will be the new reverberation time of the hall. The absorption of each chair is 1.0 O.W.U. [Ans: 1.31 s]
3. The volume of a room is 600 m^3 , the wall area of the room is 220 m^2 , the floor area is 120 m^2 and the ceiling area is 120 m^2 . The average sound absorption coefficient for the walls is 0.03, for the ceiling 0.8 and for the floor it is 0.06. Calculate the average sound absorption coefficient and the reverberation time. [Ans: 0.24, 0.875s]
4. A cinema hall has a volume of 7500 m^3 . What should be the total absorption in the hall if the reverberation time of 1.5 sec is to be maintained? [Ans: 825 sabins]
5. Reverberation time is found to be 3 sec for an empty reverberation chamber and an acoustic sheet of 15 m^2 is suspended at the center of the reverberation hall. Calculate the coefficient of sound absorption of acoustic sheet if the volume of the chamber is 600 m^3 .
6. A hall of volume 5500 m^3 is found to have a reverberation time of 2.3s. The sound absorbing surface of the hall has an area of 750 m^2 . Calculate the average absorption coefficient. [Ans: 0.513]
7. Reverberation time for a cubical chamber of 10m width is 2.68 sec. Calculate its average absorption coefficient. If one of the walls is covered with acoustic tiles the reverberation time will decrease to 2 sec. Calculate the sound absorption coefficient of acoustic tiles.
8. The average reverberation time of a hall is 1.5 sec and the area of the interior surface is 3340 m^2 . If the volume of the hall is 1200 m^3 , find the absorption coefficient.
[Ans: 0.2 sabins]
9. The volume of a hall is 475 m^3 , the area of the wall is 200 m^2 , area of the floor and ceiling each is 100 m^2 . If the absorption coefficients of the wall, ceiling and floor are 0.025, 0.020 and 0.550 respectively, find the reverberation time for the hall. [Ans: 1.264s]
10. A hall has a volume of $1,20,000 \text{ m}^3$. It has a reverberation time of 1.5 s. What is the average absorbing power of the surface if the total absorbing surface area is $25,000 \text{ m}^3$?
[Ans: 0.524 O.W.U/m²]
11. A hall has a volume of 2265 m^3 and its total absorption is equivalent to 92.9 m^2 . How many persons should be seated in the hall so that the reverberation time becomes 2 s? Given that the absorption area of one person is equivalent to 18.6 m^2 of open window. Calculate the reverberation time of the empty hall also.
[Ans: 500 persons, 4s]
12. A hall of volume 1586 m^3 is found to have a reverberation time of 2s. If the area of the sound absorbing surface is 650 m^2 , calculate the average absorption coefficient. [Ans: 0.195]

CHAPTER

12

Ultrasonics

12.1 INTRODUCTION

The human ear is sensitive to sound waves of frequencies ranging from 16 Hz to 20 kHz. Waves of frequencies beyond the upper audible limit ($f > 20\text{kHz}$) are called *ultrasonic waves*. Human ear cannot sense ultrasonic sounds but dogs and other animals are endowed with an ability to hear the high frequency sounds. Propagation of these waves through material media depends on the elastic properties and the density of the medium. In fluids they are propagated as longitudinal waves whereas in solids they travel as both longitudinal as well as transverse waves. *The wavelengths of ultrasonic waves are very small and this smallness in wavelength makes them useful in many of their applications.* They travel in straight lines and can be concentrated at a given location. Ultrasonic waves are widely used in medical diagnostics, marine applications, nondestructive testing of finished products and so on. Bats and dolphins are known to generate ultrasonic waves and use the reflections of the waves to find their way. The bat emits a series of short ultrasonic pulses at a frequency of 20 kHz to 60 kHz; the bat's large ears are specialized to detect these sounds. The waves reflected from the surrounding objects are perceived by the bat and from the time elapsed between the generation and reflection of the pulses, the direction and distance of the objects are determined. Some of the marine animals use ultrasonic pulses to locate fish, to avoid obstacles etc. The usability of ultrasonic waves by sea animals is due to the fact that light is strongly absorbed by seawater and the radius of visibility is limited whereas ultrasonic waves are less absorbed by seawater.

12.2 PRODUCTION OF ULTRASONIC WAVES

There are mainly two important methods for generating ultrasonic waves, which are based on two different phenomena, namely *magnetostriiction* and *piezoelectric effects*. Magnetostriiction method is used to produce waves in the frequency range of 20 kHz to 100 kHz and the piezoelectric method is used for the production of waves of frequencies greater than 1 MHz.

12.2.1 Magnetostriiction Effect

Joule discovered the phenomenon of magnetostriiction in 1847. When a rod of ferromagnetic materials such as iron or nickel is kept in a magnetic field parallel to its length, the rod suffers a change in its length. The change in length is of the order of 1 ppm. This change in length is independent of the direction of the magnetic field and depends only on the magnitude of the field and nature of the material. This phenomenon is known as **magnetostriiction**. Nickel exhibits a large magnetostriiction effect compared to other ferromagnetic materials.

A simple method of producing longitudinal vibrations is to apply an ac magnetic field parallel to the axis of the rod of a ferromagnetic material. An ac magnetic field is produced by wounding coil of wire around the rod and by passing an alternating current through the wire. If the alternating magnetic field oscillates at frequency f , the rod changes in length once in each half cycle. It results in setting up vibrations in the rod whose frequency is twice the frequency of the magnetic field. If the rod is not magnetized initially, the resulting changes in its length are independent of the direction of the field. The change may be either an elongation or contraction depending upon the material. Normally, the amplitude of the vibrations is small. But when the frequency of the alternating fields is equal to the natural frequency of the rod, resonance occurs and the amplitude of the vibrations will be considerably larger. Further, if the frequency of the alternating field lies in ultrasonic range, an ultrasound of frequency $2f$ will be generated in the medium that is in contact with the ends of the rod.

As the rod vibrates longitudinally, the following relation governs the frequency of oscillations.

$$f = \frac{m}{2L} \sqrt{\frac{Y}{\rho}} \quad (12.1)$$

where L is the length of the rod, Y the Young's modulus, ρ the density of the rod and $m = 1, 2, 3, \dots$.

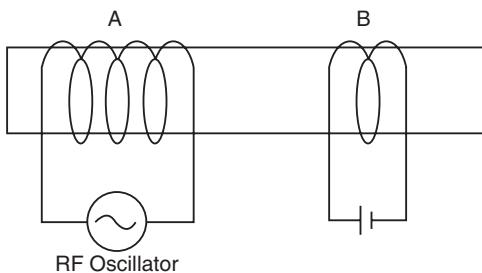


Fig. 12.1. Magnetostrictive Vibrator

If it is desired that the frequency of vibration of the bar be the same as that of the *ac* current, a steady **polarizing magnetic field** must be applied to the bar. The polarizing magnetic field can be produced by passing *dc* current through a second coil, as shown in Fig. 12.1. If the magnitude of the polarizing magnetic field is greater than that of the *ac* field, the frequency of the vibration of the bar will be equal to that of the *ac* magnetic field.

12.2.2 Magnetostriiction Ultrasonic Generator

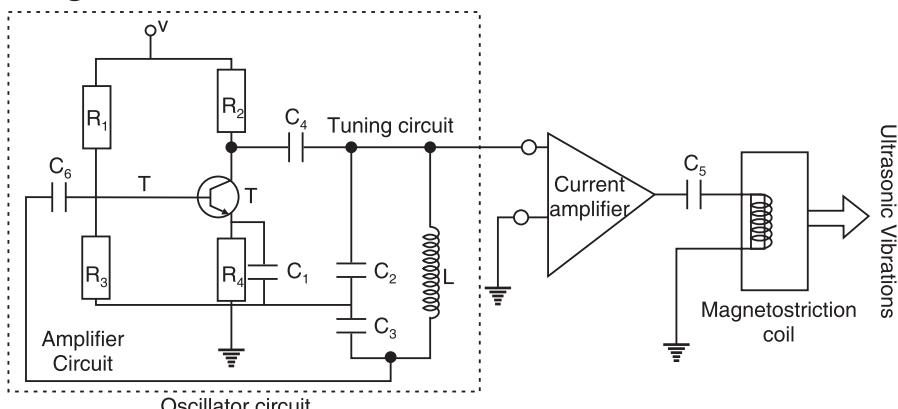


Fig. 12.2. The circuit diagram of a magnetostriiction ultrasonic generator

The circuit diagram of a magnetostriction ultrasonic generator using transistors is shown in Fig. 12.2. It is basically a Colpitt's oscillator. The transistor T is biased with the help of the resistances R_1 , R_2 , R_3 , and R_4 . The inductance L and capacitors C_2 and C_3 constitute the tank circuit. When the circuit is switched on, oscillations build up in the tank (resonant) circuit. The oscillations are fed back to the transistor base through the feed back capacitor C_6 . The appropriate frequencies are amplified and the oscillations corresponding to them are sustained. The oscillations appearing at the output terminals of the oscillator circuit are fed to a current amplifier, which raises the level of the oscillations. The output of the current amplifier is fed to the magnetostriction coil through a coupling capacitor C_5 . Under the action of the high frequency electrical signal passing through the coil, the magnetization of the nickel bar within the coil varies and as a result, it produces ultrasonic waves.

Advantages

- Magnetostrictive materials are inexpensive.
- Large output power can be produced.

Limitations

- Frequencies higher than 300 kHz cannot be generated.
- Single frequency oscillations cannot be generated.

Example 12.1: Calculate the natural frequency of 40 mm length of a pure iron rod. Given the density of pure iron is $7.25 \times 10^3 \text{ kg/m}^3$ and its Young's modulus is $115 \times 10^9 \text{ N/m}^2$. Can you use it in magnetostriction oscillator to produce ultrasonic waves?

$$\text{Solution: } f = \frac{n}{2L} \sqrt{\frac{Y}{\rho}} = \frac{1}{2 \times 40 \times 10^{-3} \text{ m}} \left[\frac{115 \times 10^9 \text{ N/m}^2}{7.25 \times 10^3 \text{ kg/m}^3} \right]^{\frac{1}{2}} = 49.75 \text{ kHz.}$$

It can be used in magnetostriction oscillator to produce ultrasonic waves.

Example 12.2: Calculate the length of an iron rod which can be used to produce ultrasonic waves of 20 kHz. Given that

$$\text{Young's modulus of iron} = 11.6 \times 10^{10} \text{ N/m}^2$$

$$\text{Density of iron} = 7.23 \times 10^3 \text{ kg/m}^3.$$

$$\text{Solution: } f = \frac{n}{2L} \sqrt{\frac{Y}{\rho}} \quad \therefore \quad L = \frac{n}{2f} \sqrt{\frac{Y}{\rho}} = \frac{1}{2 \times 20 \times 10^3 \text{ Hz}} \left[\frac{11.6 \times 10^{10} \text{ N/m}^2}{7.23 \times 10^3 \text{ kg/m}^3} \right]^{\frac{1}{2}} = 1 \text{ m.}$$

12.3 PIEZOELECTRIC EFFECT

The French physicists Pierre Curie and Paul-Jean Curie discovered the piezoelectric effect in 1880. When one pair of opposite faces of certain asymmetric crystals such as quartz is compressed, opposite electric charges appear on the other pair of opposite faces of the crystal. If the crystals are subjected to tension, the polarities of the charges are reversed. This development of charges as a result of the mechanical deformation is known as the **direct piezoelectric effect**. Crystals that exhibit piezoelectric effect are called piezoelectric crystals. Ammonium phosphate, quartz, PZT (lead zirconate titanate) are examples of piezoelectric materials.

The converse effect can also occur. If an electric field is applied across one pair of faces of a piezoelectric crystal, it gets deformed along the direction of the other opposite pair of faces. If an alternating voltage is applied between the two opposite faces of the crystal, it vibrates with the frequency of the field. The mechanical deformation of piezoelectric materials caused by an external electric field is known as the **inverse piezoelectric effect**.

12.3.1 Piezoelectric Crystal

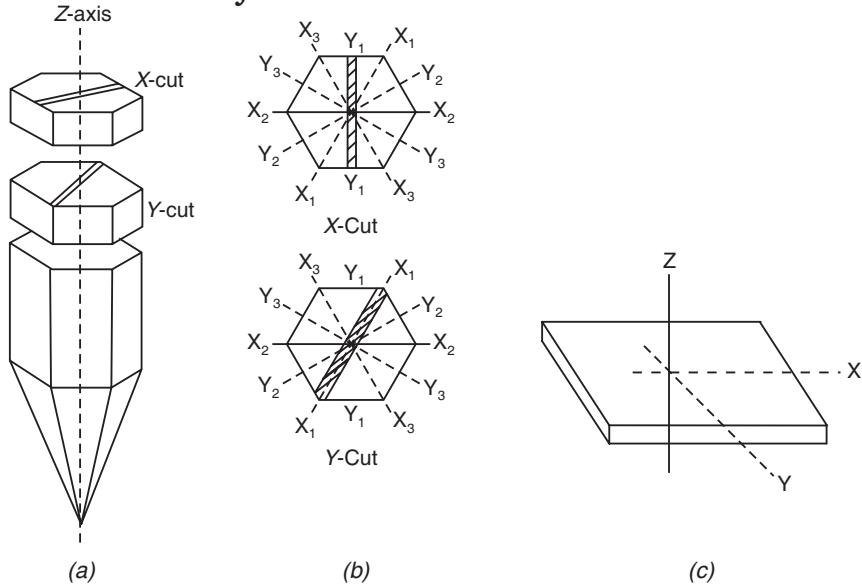


Fig. 12.3. (a) Cutting a thin slice from a quartz crystal. (b) x-cut and y-cut
(c) A slice of a quartz crystal

Quartz crystal is the most popular choice in fabrication of piezoelectric vibrators (transducers). The natural quartz crystal has the shape of a hexagonal prism with a pyramid attached to each end. It has to be cut in a particular direction and taken as a thin slab to fabricate the transducer. The axis along the longest dimension of the natural crystal is called *optic axis* or z-axis (see Fig. 12.3). The three lines, which pass through the opposite corners of the crystal, constitute its three x-axes or electrical axes. Similarly, the three lines which are perpendicular to the sides of the hexagon form the three y-axes, which are known as *mechanical axes*. Thin plates of the quartz crystal cut perpendicular to one of its x-axis are known as x-cut plates. Similarly, thin plates cut perpendicular to one of its y-axis are known as y-cut plates. x-cut plates generate longitudinal mode of ultrasonic vibrations of frequencies up to several hundred kHz, while y-cut plates generate transverse mode of vibrations of frequencies ranging from 1 MHz to 10 MHz.

The frequency of the length vibrations of x-cut crystal is given by

$$f = \frac{m}{2l} \sqrt{\frac{Y}{\rho}} \quad (12.2)$$

where m is an integer, Y is Young's modulus along the appropriate direction and ρ is the density of the crystal plate.

12.3.2 Piezoelectric Ultrasonic Generator

The circuit diagram of a piezoelectric ultrasonic generator using transistors is shown in Fig. 12.4. It is basically a Hartley oscillator. The transistor is biased using the network of resistances R_1 , R_2 , R_3 , and R_4 . The coils L_1 , L_2 and capacitor C_4 constitute the tuning (resonant) circuit. The tuning circuit is coupled to the transistor T through the coupling capacitor C_2 . Capacitor C_3 provides the positive feed back to the amplifier T. The oscillations generated by the tank circuit are sustained and the electrical signal obtained at the output is applied to the electrodes of the piezoelectric crystal through the coupling capacitor C_5 .

Because of high frequency electrical signal applied to it, the piezoelectric crystal produces ultrasonic waves. The frequency of these ultrasonic waves can be varied by varying the values of the components of the tuning circuit.

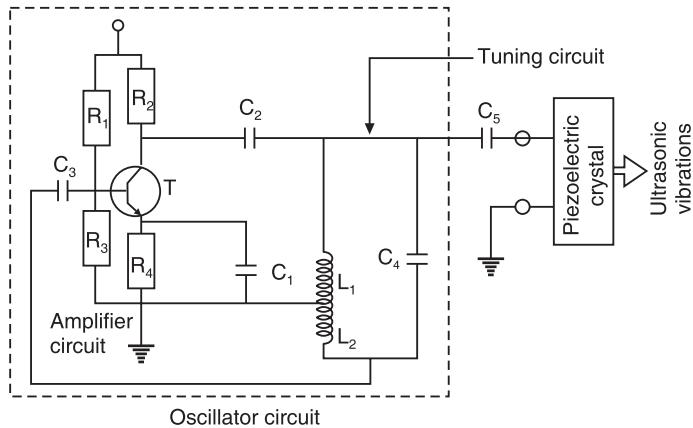


Fig. 12.4. Piezoelectric ultrasonic generator

Advantages

- High frequency waves of frequencies up to 500 MHz can be generated.
- Single frequency output can be obtained
- A range of frequencies can be covered using different transducers.

Example 12.3: Calculate the natural frequency of ultrasonic waves using the following data.

$$\text{Thickness of quartz plate} = 5.5 \times 10^{-3} \text{ m}$$

$$\text{Young's modulus of quartz} = 8.0 \times 10^{10} \text{ N/m}^2$$

$$\text{Density} = 2.65 \times 10^3 \text{ kg/m}^3.$$

$$\text{Solution: } f = \frac{n}{2l} \sqrt{\frac{Y}{\rho}} = \frac{1}{2 \times 5.5 \times 10^{-3} \text{ m}} \left[\frac{8.0 \times 10^{10} \text{ N/m}^2}{2.65 \times 10^3 \text{ kg/m}^3} \right]^{\frac{1}{2}} = 49.9 \text{ kHz.}$$

12.4 DETECTION OF ULTRASONIC WAVES

1. Kundt's tube Method: A Kundt's tube can be used to detect ultrasonic waves of relatively longer wavelengths. Stationary ultrasonic waves are produced in air contained in a long tube supported horizontally. Lycopodium powder sprinkled along the inner surface of the tube collects into small heaps at the nodes and is blown off at the antinodes. The appearance of heaps indicates the presence of waves and the mean distance between two successive heaps is equal to $\lambda/2$.

2. Piezoelectric detector: In this method, ultrasonic waves are applied to one pair of faces of a quartz crystal. As a result, varying electric charges are produced on the other pair of faces of the crystal. These charges, being small, are amplified and detected.

3. Thermal detection: A fine platinum wire probe is used in this method of detection of ultrasonic waves. Due to alternate compressions in the medium resulting from ultrasonic waves, there occurs a change of temperature at nodes. As the platinum probe moves through the medium, its resistance changes at nodes. The change in the resistance of platinum wire is detected by using a sensitive bridge.

12.5 PROPERTIES OF ULTRASONIC WAVES

- (i) The speed of propagation of ultrasonic waves depends upon their frequency. It increases with increase in frequency.
- (ii) The wavelength of the waves is very small and the waves exhibit negligible diffraction effects.
- (iii) They can travel over long distances as a highly directional beam and without appreciable loss of energy.
- (iv) They are highly energetic. Owing to the high frequencies involved, ultrasonic waves may have intensities up to 10 kW/m^2 . Normally, 1 to 2 kW/m^2 intensities are used.
- (v) They produce cavitation effects in liquids.

12.6 CAVITATIONS

Microscopic bubbles of about 10^{-9} to 10^{-8} m sizes are always present in a liquid. A decrease in pressure above the liquid causes an intense evaporation in the bubbles and leads to their growth. The growth of bubbles leads to their collapse. The entire process of growth and collapse of bubbles occurs within one millisecond. During the collapse of a bubble, a shock wave is formed causing an abrupt increase in the temperature of the gas within the bubble.

When ultrasonic waves propagate thorough liquid media, they induce alternate regions of rarefaction and compression. A negative local pressure at the spot of rarefaction causes local boiling of the liquid accompanied by the bubble growth and collapse. This phenomenon is known as **cavitation**. When the minute bubbles collapse, the local pressure increases up to thousands of atmospheres and consequently local temperature increases by about as much as $10,000^\circ\text{C}$. The numerous shock waves combine to act as liquid hammer. In such conditions the liquid displays high crushing power.

12.7 TYPES OF ULTRASONIC WAVES

We classify ultrasonic waves into four types based on the mode of vibration of the particles of the medium with respect to the direction of propagation of the waves. The types of waves are

- Longitudinal or compressional waves
- Transverse or shear waves
- Surface or Rayleigh waves

12.7.1 Longitudinal or Compressional Waves

In fluids, gases and liquids, and ultrasonic waves propagate in the form longitudinal waves. The molecules of the medium move back and forth in the direction of propagation of the wave and produce alternate regions of compression and rarefaction. The velocity of ultrasonic waves in fluids is given by

$$v = \sqrt{\frac{\gamma P}{\rho}} \quad (12.3)$$

where P is pressure, ρ is the density.

Longitudinal ultrasonic waves are widely used in ultrasonic testing of materials because of their easy generation and detection.

12.7.2 Transverse or Shear Waves

In a transverse wave, particles of the medium vibrate in a direction perpendicular to the direction of propagation of the wave. For a transverse wave to travel through a material, it is necessary that each particle of the medium is strongly bound to its neighbours so that as one particle moves, it pulls its neighbours. Therefore, transverse or shear waves can only

propagate in solids. The velocity of a transverse wave is therefore about 50% of the velocity of the longitudinal wave in the same material.

It is to be noted that longitudinal as well as transverse waves travel through solids. The velocity of propagation of longitudinal waves in solids is determined by Young's modulus and is given by

$$v_l = \sqrt{\frac{Y}{\rho}} \quad (12.4)$$

where Y is Young's modulus.

The velocity of shear waves in solids is determined by the rigidity modulus and is given by

$$v_t = \sqrt{\frac{\eta}{\rho}} \quad (12.5)$$

where η is the rigidity modulus.

12.7.3 Surface or Rayleigh Waves

Surface waves were first described by Rayleigh and are therefore called *Rayleigh waves*. This type of waves travels along flat or curved surfaces, without going into the bulk of the medium. These waves have a velocity of about 90% of the shear wave velocity in the same material. They enable detection of surface or near surface cracks or defects. Further, these waves can bend around corners and hence they can be used for testing complicated shapes.

12.8 DETERMINATION OF VELOCITY OF ULTRASONIC WAVES

Determination of velocity of ultrasonic waves in a medium basically consists of determining the wavelength of the ultrasonic wave in that medium. We describe here two of the methods used for this purpose.

- Interferometer method
- Acoustic diffraction method

A. Interferometer Method

The velocity of ultrasonic waves is determined using an ultrasonic interferometer.

Experimental setup

The interferometer consists of an ultrasonic generator having a liquid cell connected to its tank circuit. The cell, T , is a vertical cylindrical tube filled with the liquid medium under test (Fig. 12.5a). A piezoelectric crystal C is mounted at the bottom of the cell. Reflector, R is a metallic plate, mounted at the top end of the cell and it can be moved parallel to itself with the help of a micrometer screw. The surface of crystal C and the reflector are made exactly parallel to each other.

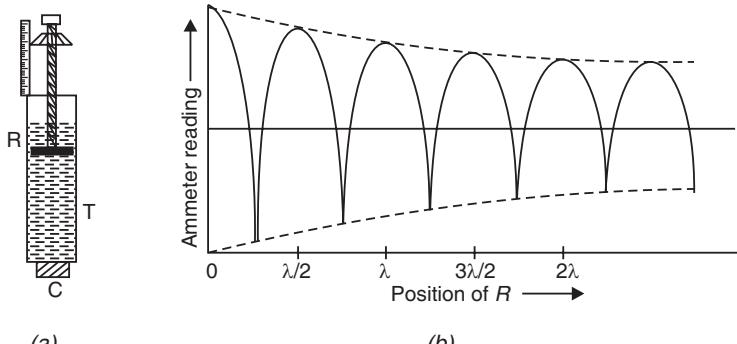


Fig. 12.5

Working

When the piezoelectric crystal, C is excited ultrasonic waves of known frequency are produced and propagate through the liquid. These waves are reflected at the reflector R and travel back to C. The position of the reflector can be adjusted such that the forward and backward waves form a standing wave pattern in the medium. As the reflector moves, the reading of the microammeter in the circuit fluctuates. The readings of microammeter are recorded and plotted against the position of R. The plot will pass through maxima and minima as shown in Fig. 12.5 (b). The distance between two consecutive minima or maxima is $\lambda_u/2$. From this, we get the value of the wavelength λ_u of the ultrasonic wave in the medium.

Determination of velocity

Then, knowing the value of the frequency of the ultrasonic waves used, the velocity of the ultrasonic wave in the liquid v_u is computed from the relation

$$v_u = f\lambda_u \quad (12.6)$$

B. Acoustic diffraction method

When ultrasonic waves propagate in a liquid medium, the alternating compressions and rarefactions change the density of the medium. It leads to a periodic variation of refractive index of the liquid. Thus, a liquid column subjected to ultrasonic waves act as grating called **acoustic grating**. If monochromatic light is passed through the liquid column at right angles to the column, the liquid causes diffraction of light. The diffraction pattern can be used to determine the wavelength and velocity of the waves.

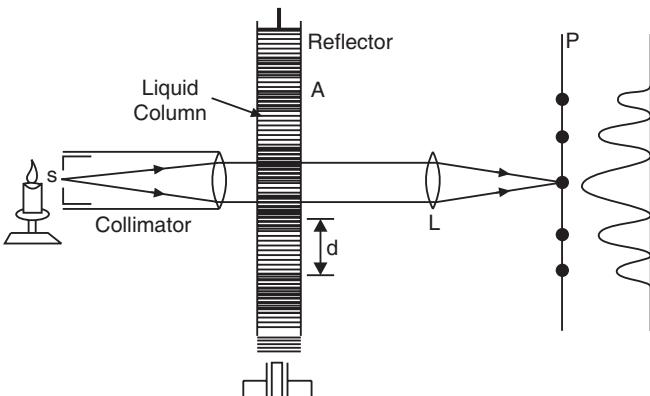


Fig. 12.6. An acoustic diffraction grating produced by a liquid column subjected to ultrasonic waves

Experimental setup

Fig. 12.6 shows the experimental arrangement. A glass tube is filled up with a liquid. The piezoelectric transducer (a quartz crystal, C) is positioned at the bottom of the glass tube and a reflector, R, is arranged at the top. The surfaces of the crystal, C and reflector, R are held perfectly parallel to each other. S is a slit through which monochromatic light passes and gets collimated and rendered parallel. The image of the slit is focused on the screen P by the lens L.

Working

When the quartz crystal, C is not excited and is at rest, the light forms a single image of the slit on the screen, P. When the electric circuit excites the crystal, it goes into a state of vibration and produces ultrasonic waves. The waves travel through the liquid column and get reflected at the reflector located at the top. The two waves combine to form stationary waves in the liquid column. The density and hence the refractive index of the liquid is maximum at nodal points and minimum at antinodal points. Therefore, the nodal points act as opaque regions while antinodal areas act as transparent regions for light. The liquid column thus resembles a ruled grating and causes diffraction of light. The image formed on the screen consists of a diffraction pattern having a central maxima flanked by first order, second order maxima and minima and so on.

Determination of wavelength

Grating equation $(a + b) \sin \theta = m\lambda$ is applicable to the acoustic grating also. The grating constant $(a + b)$ in this case equals $\lambda_u/2$ and is given by

$$\frac{\lambda_u}{2} \sin \theta = m\lambda \quad (12.7)$$

where λ_u is the wavelength of the ultrasonic waves,

λ is the wavelength of monochromatic light used to produce the diffraction pattern and m is the order of the maxima and takes integer values 1, 2, 3,...etc.

∴ The wavelength of the ultrasonic waves is

$$\lambda_u = \frac{2m\lambda}{\sin \theta} \quad (12.8)$$

Knowing the values of λ , m and measuring the angle θ , we can calculate the wavelength λ_u . This method of determining the wavelength of ultrasonic waves is known as **acoustic diffraction method**.

Determination of velocity

First, the wavelength is determined using the values of λ , m and θ in equ.(3.8). Then, the velocity of the ultrasonic wave in the liquid v_u is computed from the relation

$$\begin{aligned} v_u &= f\lambda_u \\ \text{or} \quad v_u &= \frac{2m\lambda f}{\sin \theta} \end{aligned} \quad (12.9)$$

where f is the frequency of the ultrasonic waves which is known from the frequency of the oscillator.

Example 12.4: The wavelength of light transmitted through a liquid is 6000 \AA . The first order angle of diffraction is 0.046° . Calculate the velocity of ultrasonic waves in the liquid. The frequency of the ultrasonic waves produced by the transducer is 2 MHz .

Solution: Velocity of ultrasonic waves in liquid,

$$v_u = \frac{2m\lambda f}{\sin \theta} = \frac{2 \times 1 \times 6000 \times 10^{-10} \text{ m} \times 2 \times 10^6 \text{ Hz}}{\sin 0.046^\circ} = 2989 \text{ m/s.}$$

12.9 MEASUREMENT OF ELASTIC CONSTANTS IN LIQUIDS

Once the velocity of the ultrasonic waves in a liquid is determined, we can calculate the bulk modulus and adiabatic compressibility of the liquid as follows.

Bulk Modulus of a liquid

The bulk Modulus of a liquid is given by the expression

$$v_u = \sqrt{\frac{K}{\rho}} \quad (12.10)$$

where v_u is the velocity of ultrasonic waves in the liquid,

K is the bulk modulus and

ρ is the density of the liquid

Adiabatic compressibility of liquids

Adiabatic compressibility of liquids, electrolyte solutions and binary systems is characterized by β . β is a significant parameter in Chemical Engineering and for Chemical industries.

It is given by the expression

$$\beta = \frac{1}{\rho v_u^2} \quad (12.11)$$

where v_u is ultrasonic velocity in the liquid and is ρ density of the liquid.

12.10 DETERMINATION OF VELOCITY OF ULTRASONIC WAVES IN SOLIDS

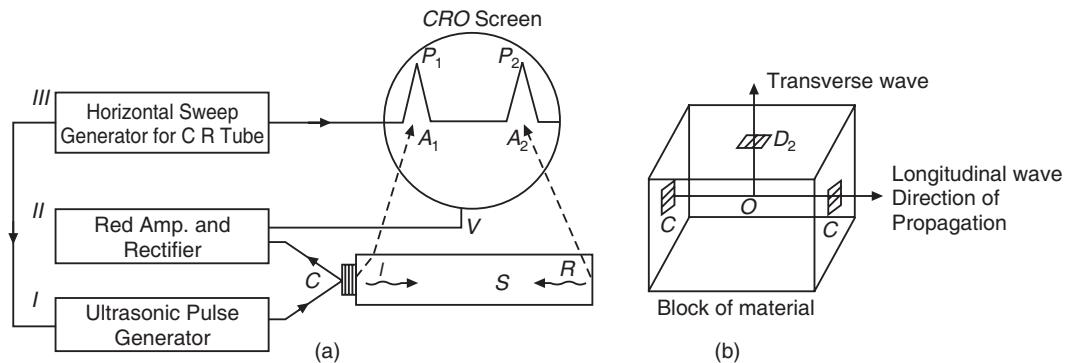


Fig. 12.7

The block diagram of the set up to determine the velocity of ultrasonic waves in longitudinal and transverse directions of a solid is shown in Fig. 12.7 (a). S is the specimen taken in the form of a block having plane faces. A crystal transducer C is kept in contact with one of the faces, which sends ultrasonic pulses of microsecond order in the solid. The pulses travel through the specimen and get reflected at its other end. The incident and reflected pulses, denoted as I and R, are fed to an amplifier-rectifier unit (block II). The output of this unit is applied to vertical plate system of the CRO. The initial of pulse generator is also fed to the III block which generates sweep for the CRO. The coupling of I to III block provides synchronization and sweep starts with the pulse fed into the specimen. Traces P_1 and P_2 obtained on the CRO screen correspond to the direct and reflected pulses.

The longitudinal and transverse directions in the specimen are shown in Fig. 12.7(b) and are determined in relation to the face of transducer C. D_1 and D_2 are two detectors of ultrasonic waves geometrically set to detect longitudinal and transverse waves respectively. The detectors D_1, D_2 are crystals identical to C.

Working:

The crystal transducer C is excited by a pulse generator and produces pulses of microsecond order. The pulses travel in the material block and are reflected back to produce stationary waves in the block. In longitudinal direction the pulse is received by detector D_1 . The pulse exciting C to produce ultrasonic waves is traced on CRO screen. The pulse of ultrasonic waves reaching at D_1 develops a piezo-electric voltage which is amplified and produces another trace on CRO screen. The separation $A_1 A_2$ gives the time taken by the ultrasonic wave in traversing the block in longitudinal direction. Time in traversing the block in transverse direction is determined in the same way.

The velocity of the longitudinal wave is computed from the formula

$$v_l = \frac{2L}{t} \quad (12.12)$$

where L is the length of the specimen in longitudinal direction and t is the time taken by the wave to traverse the specimen following a similar procedure, the transverse velocity of the ultrasonic waves is determined. The detector D_2 is used for detection of transverse waves.

12.11 MEASUREMENT OF ELASTIC CONSTANTS IN SOLIDS

It is to be noted that longitudinal as well as transverse (or shear) waves travel through solids. The velocity of propagation of longitudinal waves in solids is determined by Young's modulus and is given by

$$v_l = \sqrt{\frac{Y}{\rho}} \quad (12.13)$$

where Y is Young's modulus and ρ the density of the solid.

The velocity of shear waves in solids is determined by the rigidity modulus and is given by

$$v_t = \sqrt{\frac{\eta}{\rho}} \quad (12.14)$$

where η is the rigidity modulus of the solid.

It can be shown that the ultrasonic speeds of longitudinal and transverse waves are given by

$$v_l = \sqrt{\frac{Y(1-\sigma)}{\rho(1+\sigma)(1-2\sigma)}} \quad (12.15)$$

$$v_t = \sqrt{\frac{\eta}{\rho}} = \sqrt{\frac{Y}{2\rho(1+\sigma)}} \quad (12.16)$$

where σ is the Poisson's ratio of the solid.

Solving the equations (12.15) and (12.16), we get

$$\sigma = \frac{\left[1 - 2\left(\frac{v_t}{v_l}\right)^2\right]}{2\left[1 - \left(\frac{v_t}{v_l}\right)^2\right]} \quad (12.17)$$

$$Y = 2\rho(1+\sigma)\rho_t^2 \quad (12.18)$$

$$\text{and} \quad \eta = \rho v_t^2 \quad (12.19)$$

Knowing v_l and v_t of a solid, its elastic constants can be determined from the above equations.

12.12 INDUSTRIAL APPLICATIONS

Ultrasonic waves are extensively used in industry, medicine and marine applications.

1. Ultrasonic Drilling: Ultrasonic machining is a vibratory process that is now in common use for the mechanical treatment of hard and brittle solids such as ceramics, glasses, precious stones, semiconductors and hard alloys. The tool motion is produced by an acoustic concentrator to which the tool holder is threaded. The acoustic concentrator consists of a needle type magnetostriction vibrator, illustrated in Fig. 12.8. The vibrator is made of thin isolated ferromagnetic plates of high magnetostriction such as nickel. A coil is wound on

the needle, through which an alternating current of frequency passes. The resulting magnetic field magnetizes the core and thus changes its length. The core of the vibrator vibrates at a frequency $2f$. By choosing the frequency f to be equal to half the natural vibration frequency of the vibrator, the system is held at resonance and the vibrations of the needle will be of large amplitudes. A tapered waveguide of appropriate dimensions and rigidly attached to the vibrator concentrates the vibrational energy and communicates it to the tool. The tool oscillates linearly with an amplitude of 0.013 to 0.1 mm at ultrasonic frequency of 20 kHz to 30 kHz.

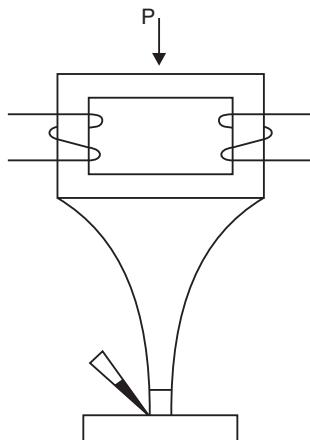


Fig. 12.8. Schematic of a typical ultrasonic drill

In operation, the needle vibrator is set in oscillation and the tool shank is pressed against the work piece. An aqueous suspension of a solid abrasive powder is then fed through a tube to the working zone. Abrasive particles bombard the work surface at high velocity and shear off small pieces of the material. This action rapidly chips away the work piece in a pattern controlled by the tool shape and contour.

2. Ultrasonic Welding: Practically, all metals and plastics can be welded using ultrasonic waves of suitable energy. The surfaces of the work pieces are cleaned and held together. They are subjected to ultrasonic oscillations at the spot where they are to be welded. The ultrasonic energy converts to heat at the contact area as a result of friction arising between the surfaces. As the temperature of surface layers exceeds the recrystallization point, the layers melt and bond together to form a strong joint. The merits of this process are that it does not cause stress at the spot of welding and that the structure of the materials remain unchanged.

3. Ultrasonic Soldering: Normally, surfaces are covered with contaminants, grease and oxide films. Such films prevent formation of a good joint. Therefore, prior to soldering, the surfaces are cleaned with active fluxes. The fluxes, when heated, dissolve the oxide films and uncover the clean metal surface which readily allow the molten solder to form a firm joint. This method however is not suitable for soldering aluminium. Active metals such as aluminium can be soldered without fluxes with the help of ultrasonic waves. In this case soldering is done by a special iron that vibrates at a frequency of tens of kilohertz.

4. Ultrasonic Cleaning: In the fabrication of electronic devices, it is highly essential to clean the surfaces of parts and components at different stages of production. Cleaning of the surfaces is commonly carried out in either organic solvents or weakly alkaline aqueous solutions containing surface-active agents. To scrub the surfaces more effectively, the phenomenon of cavitation is utilized. Ultrasonic cleaning baths are used for this purpose.

The hydraulic shock arising at the surface of a part due to cavitation destroys any layer of contaminants. Bubbles penetrate under the layer, tear it off and break it down into minute pieces. The surface-active agent pulls them away into the solution.

The chief advantage of this method is that it enables cleaning the surface of small products of intricate configuration. Jewelers make use of ultrasonic baths to clean jewellery.

5. Echo Sounder : Ultrasonic waves can be produced in the form of directed beams like beams of light. Further, ultrasonic waves can travel long distances in water. This property is utilized in measuring the depth of ocean. A ship equipped with an echo sounder sends out short pulses of ultrasonic waves towards the bed of the ocean (Fig. 12.9). These waves are reflected back from the bed and the receiver receives the reflected pulse. The time interval between the pulse sent and the pulse received is determined. Knowing the velocity v of the waves through the seawater, the depth of the ocean, l can be computed with the help of the following formula.

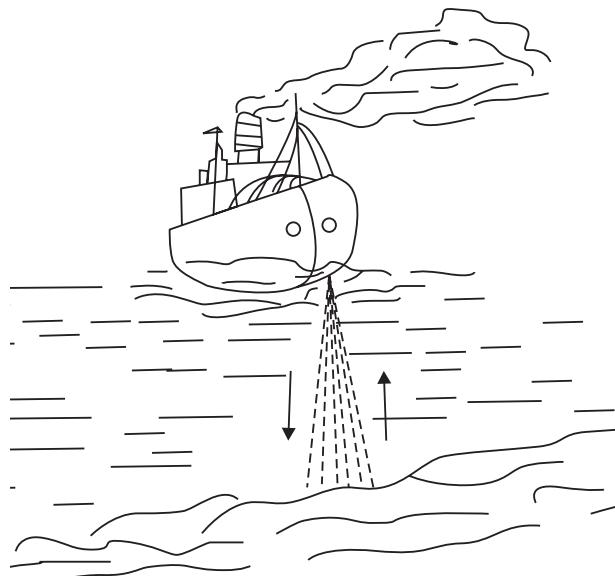


Fig. 12.9. Depth sounding

$$l = \frac{vt}{2} \quad (12.20)$$

where t is the time interval between the transmitted and reflected pulses.

6. SONAR: The word SONAR stands for Sound Navigation And Ranging. The ultrasonic waves, which are highly directional, can be used for locating objects submerged under seawater and determining their distance. The idea of ultrasonic sonar was put forward first by the French physicist Paul Langevin and was successfully used by him during the first world war for detecting enemy submarines. The sonar acts in a way much similar to an echo sounder. In sonar, an ultrasonic beam is directed in different directions into the sea. In the absence of an obstacle, the ultrasonic pulses do not return to the ship. In the presence of an obstacle, pulses are reflected from the obstacle and are picked up by the receiver. Knowing the speed of the ultrasonic waves in seawater and time elapsed between the transmitted and reflected pulses, the distance of the object is determined using the formula (12.20).

- Sonar is used to guide submarines in the seas.
- It is used to detect the presence of submerged icebergs.
- It is used for direction signaling in submarines.

7. Fish-finder: Ultra sound can be used to locate shoals of fish utilizing the fact that the swimming bladder of fish is filled with air that scatters ultrasonic waves. Ultrasonic sonar is used for this purpose. At present ultrasonic locators are mainly used for detecting icebergs, fish shoals and the like.

Some of the sea animals such as whales and dolphins use ultrasound to locate their prey, avoid collision with obstacles and even to converse with each other. In the depths of the sea,

visibility is highly restricted because of the strong absorption of light by water. It may be therefore that these animals use ultrasound that is relatively less absorbed.

8. Emulsification: Immiscible liquids like water and oil can mix thoroughly and form stable emulsions when their mixture is subjected to strong ultrasonic waves. The ultrasonic emulsification is used in industry to mix molten metals and form alloys of uniform composition.

Example 12.5: An ultrasonic source of 0.07 MHz sends down a pulse towards the seabed, which returns after 0.65 s. The velocity of sound in seawater is 1700 m/s. Calculate the depth of sea and wavelength of pulse.

$$\text{Solution: Depth of the sea, } l = \frac{vt}{2} = \frac{(1700\text{m/s})(0.65\text{s})}{2} = 552.5 \text{ m.}$$

$$\text{Wavelength of the pulse } \lambda = \frac{v}{f} = \frac{1700\text{m/s}}{0.07 \times 10^6 \text{s}^{-1}} = 24 \text{ mm.}$$

12.13 ULTRASONIC TESTING

Ultrasonic testing is a versatile and widely used non-destructive testing (NDT) method. It utilizes high frequency acoustic waves generated by piezoelectric transducers. In NDT, ultrasonic waves of frequencies from 100 kHz to about 25 MHz are generally used. The ultrasonic waves are generated with the help of piezoelectric devices. When bursts of alternating voltage are applied to the transducer, the transducer emits ultrasonic beam. The ultrasonic beam is then transmitted from the transducer into the specimen under testing. If even a slight discontinuity exists inside the specimen, the ultrasonic waves are reflected back to the transducer. The transducer converts these reflected waves again into electrical signal. This signal is displayed on a screen of CRT. The characteristics of the pulses produced by the transducer are used for interpretation of the nature of the defect in the specimen.

Several methods have been developed for the ultrasonic testing. Among them, pulse-echo methods are most popular and widely used. In some of such methods, normal beam probes are used in which a transducer crystal is fixed parallel to the bottom plate of the probe. The ultrasonic beam produced by the probe propagates into the object perpendicular to the surface of contact and travels in the material in the form of longitudinal waves. Hence, these methods are known as normal beam pulse testing methods.

A. Normal beam pulse-echo Testing

In this method, an ultrasonic pulse propagating perpendicular to the surface of the test object is reflected at the boundaries of the object and at the surfaces of defects. The reflected pulses are known as echoes. Hence, this method is known as a **normal beam pulse echo testing** method.

Fig. 12.10 shows the block diagram of a normal beam pulse-echo flaw detector. The essential units in the equipment are the pulse transmitter, clock or timer, receiver amplifier and

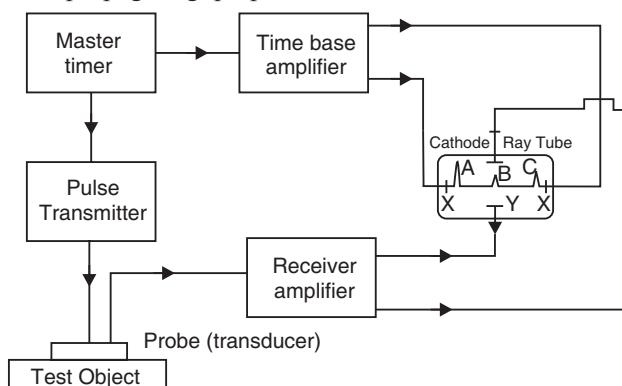


Fig. 12.10. Normal beam pulse-echo flaw detector

cathode ray oscilloscope. The time base generator used in the CRO here differs from that used in an ordinary laboratory oscilloscope. After the end of a sawtooth wave, the next cycle starts only after about 4 to 5 times the sweep time; thus it waits till the reverberations in the specimen diminish.

From the transmitter, the electric pulse is fed to the transducer probe. The piezoelectric transducer is excited by the electric pulse and vibrates at its resonant frequency. It produces a short ultrasonic pulse, which is propagated into the test object through the couplant layer. The electric pulse also triggers the time-base generator, so that the pulse of ultrasound starts to move through the object at the same time as the luminous spot moves across the CRO screen. The output of the transducer is applied to the Y-plates of CRO through an amplifier. It produces a transmission signal (A), which represents the initial ultrasonic pulse (see Fig. 12.10). The luminous spot continues to move across the screen, as the ultrasonic pulse travels through the object. If the ultrasonic pulse encounters a defect, part of the energy is reflected back from the defect surface. The reflected part of the ultrasound returns to the transducer. Under the action of this reflected energy, the transducer vibrates and produces a small voltage pulse. This induced voltage is fed to the Y-plates of CRO through the amplifier and produces signal (B), the echo pulse from the defect. The ultrasonic energy in the transmitted pulse travels further to the bottom surface of the object and gets reflected there back to the transducer. The transducer produces an electric voltage pulse (C), which is much smaller than the transmitted pulse (A). The signal (C) constitutes the bottom surface echo.

As the Y-axis of the display on CRO screen represents time, one can determine the location of the defect in the object. The time interval between the transmitted pulse and echo pulse from the defect equals the time, t , taken by the energy to travel from the transducer to the defect surface and back to the transducer. If v is the velocity of ultrasonic waves in the material, then the distance, d , of the defect from the top surface of the object is given by

$$d = \frac{vt}{2} \quad (12.21)$$

B. Normal beam pulse through-transmission Testing

In certain cases, the pulse-echo technique may not provide required information. It happens whenever a defect does not provide a suitable reflection surface or whenever its orientation is not favorable for detection. In such cases, the through-transmission method is adopted. The method uses two ultrasonic transducers on each side of the specimen being inspected. In this method, an ultrasonic pulse, from the transducer held at the front surface of the test object, propagates perpendicular to the surface and is transmitted through the boundaries of the object and the surfaces of defects. The transmitted pulses are detected by the second transducer held at the opposite end of the test object. Hence, this method is known as **normal beam pulse-through-transmission testing method**.

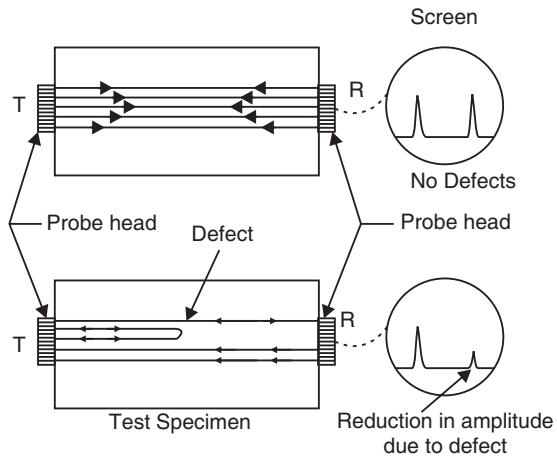


Fig. 12.11. Normal beam pulse through-transmission Testing

Fig. 12.11 shows the arrangement of a pulse-through transmission system. The transmitter and receiver probes are coupled to the test object through couplant oil. The pulse generator excites the transmitting transducer, and an ultrasonic pulse is propagated in the object. The pulse travels through the specimen to the other side. The receiver transducer on the opposite side receives the vibrations and converts them into an electrical pulse. It is amplified and displayed on an oscilloscope. If the ultrasonic pulse travels through specimen without encountering any defect, the signal received will be relatively large. If there is a defect in the path of the ultrasonic beam, part of the energy is reflected and hence the signal received at the opposite end will be reduced (see Fig. 12.11).

Example 12.6: A mild steel plate has a thickness of 18×10^{-3} m. An ultrasonic pulse travels in it with a velocity of 5.9×10^3 m/s. Calculate the echo time of the pulse.

Solution:

$$t = \frac{2l}{v} = \frac{2 \times 18 \times 10^{-3} \text{ m}}{5.9 \times 10^3 \text{ m/s}} = 6 \mu\text{s}$$

12.14 MODES OF DISPLAY

The CRO can be made to display the pulse information in various ways. They are known as modes of display.

1. A-scan display:

A-scan display is the most used mode of display in ultrasonic testing. In this mode of display, the X-axis represents time taken by the pulse to the reflecting surface and return back to the transducer. Y-axis represents the amplitude of the echoes. The location of the defect is estimated by the position of the echo given by it on the horizontal axis and size of the defect from the relative amplitude of the echo. The information that is available in A-scan is one-dimensional. A typical A-scan echo pattern is shown in Fig. 12.12.

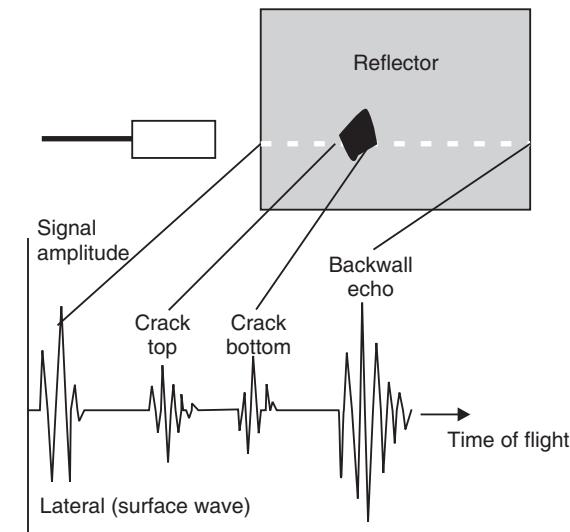


Fig. 12.12. A-scan presentation

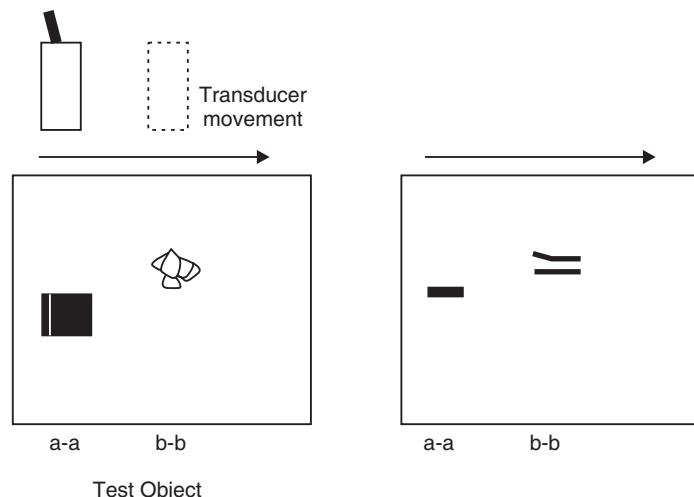


Fig. 12.13. B-scan presentation

2. B-scan display: B-scan display gives a cross-sectional view of the test object and shows the position, orientation and depth of defects in the specimen. In this mode of display, Y-axis represents elapsed time while X-axis represents the position of the transducer along a line on the surface of the test object relative to the starting position of the transducer. Thus, the probe movement is displayed in x-direction while the distance of the defect is displayed in y-direction. Echo amplitude is indicated by the relative brightness of echo indications. If a storage oscilloscope is used, the whole picture will be displayed, which reveals the depth of the defect beneath the surface and its size in the lateral direction (see Fig. 12.13).

3. C-scan display: The depth of defects is not relevant in some testing problems, but information about their distribution parallel to the test surface is required. In the C-scan mode, the transducer is moved over the surface of the test piece and the echo intensity is recorded as a variation in line shading. The image shows the plan of the object as viewed from the top and is a true-to-scale reproduction of the defect in the object (Fig. 12.14).

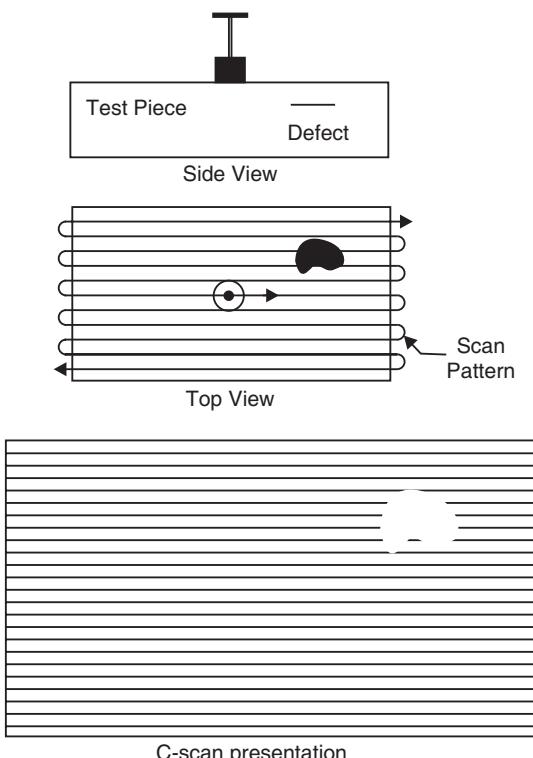


Fig. 12.14. C-scan presentation

12.15 MEDICAL APPLICATIONS—SONOGRAPHY

Ultrasound is widely used in imaging of internal organs or structures of the human body. Ultrasound imaging, also called ultrasound scanning or **sonography**, involves exposing part of the body to high-frequency sound waves to produce pictures of the inside of the body. Ultrasound imaging provides valuable information regarding the size, location, and displacement of a given structure. Tumors and other regions of organ that differ in density from surrounding tissues can be detected. Ultrasound imaging is a noninvasive medical test that helps physicians diagnose and treat medical conditions.

Principle:

Sonography uses a probe containing one or more acoustic transducers to send pulses of sound into a body. Whenever a sound pulse encounters a boundary between two tissue structures, it is partly reflected from, and partially transmitted. The sound pulse reflected back to the probe is detected as an echo (Fig. 12.15). The reflection depends on the difference in acoustic impedance of the two tissues. The acoustic impedance of a medium is the speed of sound in the material \times the density:

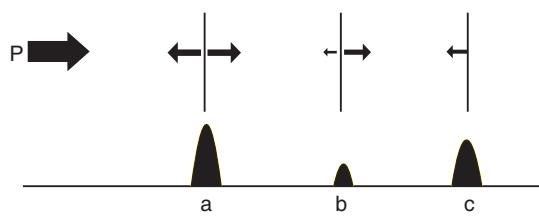


Fig. 12.15

$$Z = c \times \rho \quad (12.22)$$

The time it takes for the echo to travel back to the probe is measured and used to calculate the depth of the tissue interface causing the echo. The greater the difference between acoustic impedances, the larger the echo is. If the pulse hits gases or solids, the density difference is so great that most of the acoustic energy is reflected and it becomes impossible to see deeper.

The time lag, τ , between emitting and receiving a pulse is the time it takes for sound to travel the distance to the tissue boundary and back. Thus,

$$\tau = \frac{2l}{v} \quad (12.23)$$

12.16 ULTRASOUND SCANNER

Fig. 12.16 shows a block diagram of a simple ultrasound scanner. The **transmitter** supplies energy to the piezoelectric **transducer** which produces sharp pulses of ultrasound. The sound pulses travel through the body and get reflected from the organ or structure under investigation. The pulses with reduced strength are returned to the transducer probe which acts as a receiver as well. The returned echoes are amplified with the help of **swept-gain generator** and applied to the vertical deflection plates of CRO. Each time an echo reaches the probe, a vertical line appears on CRO screen. Corresponding to each reflecting surface in the body, one vertical line is produced on CRO screen. The **time base generator** supplies voltage to the horizontal deflection plates of CRO. The **rate generator** synchronizes the actions of the transmitter, time base generator and the swept-gain generator. The CRO displays the original pulse transmitted by the transducer into the body and the echoes received from different interfaces, as vertical lines on its screen. The CRO may be calibrated to give directly distances between the reflecting surfaces in the body. The display obtained using the above technique is known as a A-scan display and it is seen on CRO screen in Fig. 12.16.

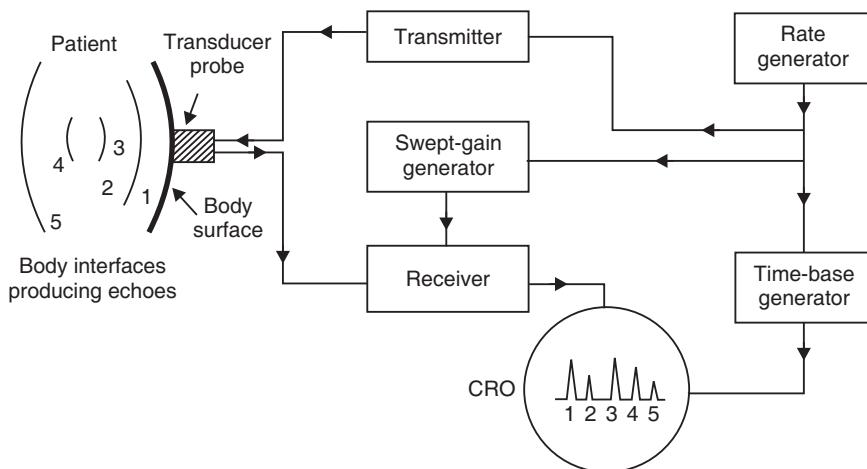


Fig. 12.16. Block diagram of a ultrasound scanner

The frequencies used for medical imaging are generally in the range of 1 to 18 MHz. Higher frequencies have a correspondingly smaller wavelength, and can be used to make sonograms with smaller details. However, the attenuation of the sound wave increases at higher frequencies, so in order to have better penetration of deeper tissues, a lower frequency (3-5 MHz) is used.

12.16.1 Display Modes

The echoes are displayed as a function of time which is proportional to the distance from the source to interface. The echo information is displayed in one of several different display modes.

A-mode (Amplitude mode):

A-mode is the simplest type of scan mode. It is a graphic depiction of amplitude of echo versus distance into the tissue. A high energy pulse from a pulser excites the transducer.

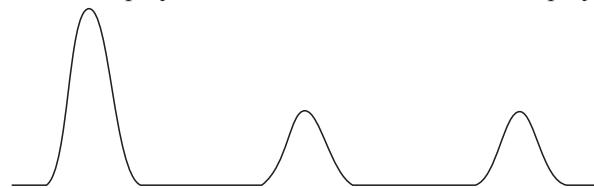


Fig. 12.17. Typical A-scan

Echoes returned from the tissue are detected by the same transducer, amplified and processed for display. The returning echoes are displayed as vertical deflections on the trace (see Fig. 12.17) which represent the amplitude of the reflected energy. In most cases, the transducer is kept stationary. Hence, the echoes are static and one dimensional. Pulses are typically a few milliseconds long and are emitted at 400 to 1000 pulses/s. It is used in ophthalmology and encephalography.

B-mode (Brightness mode):

The amplitude can also be displayed as the brightness of the certain point representing the structure, in a B-plot. In this scan, the echo signals are not applied to the horizontal deflecting plates of CRO. Instead they are used to control the brightness of the spot on the screen. Hence the reflecting surfaces appear as spots (Fig. 12.18). The brightness of the spot is proportional to the strength of returning echo. A linear array of transducers may be made to simultaneously scan a plane through the body so that we can obtain a two-dimensional image of a stationary organ or body structure on screen. It is used in diagnostic studies of liver, breast, heart, fetus etc.

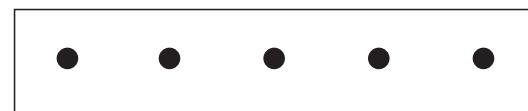


Fig. 12.18. Typical B-scan

M-mode (Motion mode): If some of the structures are moving, the motion curve can be traced by letting the B-mode image sweep across a screen. This is called the M (Motion) -mode. It enables measure range of motion, as the organ boundaries that produce reflections move relative to the probe. In this, the probe is fixed in position so that the movement of the dots along the sweep represents movement of targets. In this scan also, the echo pulses brighten the trace as is the case in B-scan. A stationary target will trace a straight line where as a moving target will trace the pattern of its movement with respect to time (see Fig. 12.19).

A typical ultrasound image is shown in Fig. 12.20.

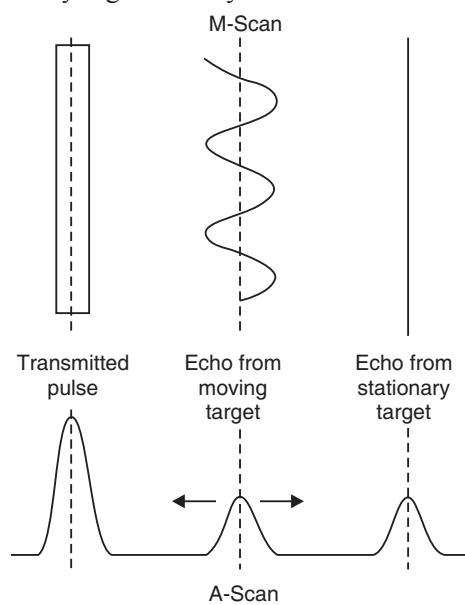


Fig. 12.19. Typical M-scan

12.16.2 Doppler Ultrasound Imaging

Doppler ultrasound imaging is based upon the Doppler effect. When the object reflecting the ultrasound waves is moving, it changes the frequency of the echoes, creating a higher

frequency if it is moving toward the probe and a lower frequency if it is moving away from the probe. The change in frequency depends upon how fast the object is moving. Doppler mode enables to measure the change in frequency of the echoes to calculate how fast an object is moving. Doppler ultrasound has been used mostly to measure the rate of blood flow through the heart and major arteries.

Advantages

- Most ultrasound scanning is noninvasive (no needles or injections) and is usually painless.
- Ultrasound is widely available, easy-to-use and less expensive than other imaging methods.
- Ultrasound imaging uses no ionizing radiation.
- Ultrasound scanning gives a clear picture of soft tissues that do not show up well on x-ray images.
- Ultrasound causes no health problems and may be repeated as often as is necessary.
- Ultrasound is the preferred imaging modality for the diagnosis and monitoring of pregnant women and their unborn babies.

Limitations

- Ultrasound waves are disrupted by air or gas; therefore ultrasound is not an ideal imaging technique for the bowel or organs obscured by the bowel.
- Ultrasound waves do not pass through air; therefore an evaluation of the stomach, small intestine and large intestine may be limited. Intestinal gas may also prevent visualization of deeper structures such as the pancreas and aorta.
- Ultrasound has difficulty penetrating bone and therefore can only see the outer surface of bony structures and not what lies within.

12.17 ULTRASONIC BLOOD FLOW METER

The ultrasonic blood flow meter utilizes the phenomenon of Doppler shift to measure the velocity of blood in veins and arteries. Ultrasound of frequency of about 5 to 10 MHz is directed at an angle to the blood stream. The particles of blood reflect the beam. The sound waves undergo a frequency change from f to f' . The frequency shift $\Delta f = f' - f$ is known as Doppler shift. It is measured by an appropriate electronic circuitry. In practice, the ultrasonic transmitter-receiver system is strapped to the limb with the help of a gel. The ultrasonic waves strike the blood stream at an angle θ . If v is the velocity of the blood with respect to the blood vessel, the component of velocity along the direction of the ultrasound beam is $v \cos \theta$. The beam reflected by the blood in the direction of the beam suffers a shift in the frequency. The Doppler shift is given by

$$\Delta f = \frac{2f v \cos \theta}{v_u}$$

where v_u is the velocity of ultrasonic waves in blood.

$$\therefore v = \frac{\Delta f v_u}{2f \cos \theta} \quad (12.24)$$



Fig. 12.20. A primary use of ultrasound is to monitor the progress of a pregnancy. The above sonogram shows the 27 weeks stage.

12.18 OTHER MEDICAL APPLICATIONS

- Ultrasonic therapy is used in treatment of rheumatic pains. Ultrasonic waves produce massaging action and relieves pain.
- The waves are useful for dental cutting and they make the cutting painless.
- Ultrasonic waves destroy bacteria and therefore they are used in sterilization of water and milk.

Example 12.7: In an ultrasonic Doppler flow meter the direction of blood flow is along the direction of ultrasonic beam. The frequency of ultrasonic waves is 2 MHz and the velocity of waves in blood is 1500 m/s. If the Doppler shift in frequency is 267 Hz, calculate the velocity of blood flow.

Solution: Velocity of blood flow, $v = \frac{\Delta f v_u}{2f \cos \theta} = \frac{267\text{Hz} \times 1500\text{m/s}}{2 \times 2 \times 10^6\text{Hz} \times 1} = 0.1 \text{ m/s.}$

QUESTIONS

- Define ultrasonics. (C.S.V.T.U.,2007)
- What is piezoelectric effect? (C.S.V.T.U.,2006)
- Why ultrasonic waves are used in detection of objects submerged in sea?
- Which crystal is used more generally in production of ultrasonic waves?
- What is meant by the natural frequency of a crystal?
- What is meant by non-destructive testing?
- What is meant by flaws in a material?
- What is a sonar?
- Why ultrasonic cleaners are used to clean mechanical components and semiconductor chips etc?
- Explain the direct and inverse piezoelectric effects. Describe the importance of each and one application of each.
- Write short notes on piezoelectric effect. (Calicut Univ.,2007)
- Define piezoelectric effect and magnetostriction effect. (Calicut Univ., 2005)
- What is piezoelectric effect? With necessary circuit diagram, explain the production of ultrasonics using piezoelectric crystal. (M.G.Univ., 2006), (Bombay Univ.), (Anna Univ., 2006)
- Define ultrasonic waves. Describe the piezoelectric method for their production. (C.S.V.T.U.,2008), (Anna Univ., 2006)
- Explain the phenomenon of magnetostriction. How will you produce high frequency sound waves with its help? (C.S.V.T.U.,2007), (Anna Univ., 2005)
- Explain the terms magnetostriction and piezoelectric effect. Discuss any one method of production of ultrasonic waves. (G.T.U., 2009)
- What is inverse piezoelectric effect? (Anna Univ., 2006)
- Explain what is Magnetostriction effect? Draw a neat labeled diagram for the production of ultrasonics by magnetostriction oscillator. (Univ. of Pune, 2008), (Bombay Univ.)
- What is magnetostriction effect? Draw circuit diagram of magnetostriction oscillator and explain its working. What are the advantages of magnetostriction method?(C.S.V.T.U., 2007)
- (a) Explain magnetostriction effect. Explain its use in production of ultrasonic waves.
(b) Explain the use of ultrasonic waves in sound ranging and also explain cavitation. (Bombay Univ.)

21. What is piezoelectric effect? Explain how ultrasonic waves are produced by a piezoelectric transducer. **(Calicut Univ., 2005)**
22. Explain the cavitation effect when an ultrasonic wave is passed through a liquid. **(Bombay Univ.)**
23. What are the different methods for the production of ultrasonic waves? Describe one of them in detail. How will you determine the wavelength of these waves? **(C.S.V.T.U., 2005)**
24. What are ultrasonic waves? Describe a method of measuring the velocity of ultrasonic waves in solids. **(V.T.U., 2007)**
25. What is an acoustic grating? Explain how an acoustic grating is used to determine the velocity of ultrasonic waves in liquids. **(V.T.U., 2007)**
26. Explain the ultrasonic diffractometer with neat diagram. How will you determine the velocity of ultrasonics in a liquid? **(Calicut Univ., 2007)**
27. Describe with theory a method of measuring velocity of ultrasonic waves in a liquid and mention how the bulk modulus of the liquid can be evaluated. **(V.T.U., 2008)**
28. (a) Determine the velocity of sound in a liquid with a neat sketch.
 (b) What are the applications of ultrasonics? **(Calicut Univ., 2006)**
29. Explain echo-sounding technique with one example. **(Univ. of Pune, 2007)**
30. Explain echo sounding technique and cavitation with one example each. **(Univ. of Pune, 2008)**
31. What are the properties of ultrasonic waves? **(C.S.V.T.U., 2006, 2007)**
32. Describe ultrasonic flaw detector. How is it used in detection of flaws in metals?
33. Explain the principle of sonar.
34. Explain the use of ultrasonic waves for non-destructive testing and in SONAR. **(Univ. of Pune, 2007)**
35. Draw a block diagram of ultrasonic flaw detector. Explain the three different scan modes used for presentation of data.
36. Discuss the use of ultrasonics for flaw detection. **(Univ. of Pune, 2008)**
37. What is the principle of ultrasonic testing?
38. What are the advantages and limitations of ultrasonic testing?
39. Explain application of ultrasound in medical field.
40. What is a sonogram and how is it obtained?
41. Explain the different scan modes used for obtaining a sonogram.
42. Explain the use of ultrasonic waves in non-destructive testing. **(M.G.Univ., 2005)**
43. What are the applications of ultrasonics? **(Calicut Univ., 2005)**
44. What are ultrasonics? Explain three applications of ultrasonics. **(M.G.Univ., 2005)**

PROBLEMS

1. Calculate the velocity of ultrasonic waves in a certain liquid using the following data obtained in an acoustic grating experiment.
 Frequency of ultrasonic waves = 100 MHz
 Wavelength of light used = 600 nm
 Angle of first order diffracted beam = 5° **[Ans: 1375m/s]**
2. An ultrasonic beam of 1 cm wavelength is sent by a ship, returns from the seabed after 2 sec. If velocity of the ultrasonic beam in seawater is 1510 m/s at 0°C , its salinity at 30°C is 29 gm/litre, calculate the depth of the seabed at 30°C and frequency of the ultrasonic beam. (Assume that the velocity of sound in sea water is given by the equation

$v = v_0 + 1.14s + 4.21t - 0.037t^2$ where v is the velocity of sound at $T^\circ C$ in seawater, v_0 the velocity of sound at $0^\circ C$, and s is the salinity [gm/litre]. [Ans: 1.6 km, 160 kHz]

3. A quartz crystal of length 0.05 cm is used in a piezoelectric oscillator. Calculate the fundamental frequency of oscillation if the velocity of the longitudinal waves in the crystal is 5.5 km/s. [Ans: 55 kHz]
4. An ultrasonic source of 70 kHz sends down a pulse towards the seabed which returns after 0.65s. The velocity of sound in sea water is 1700m/s. Calculate the depth of sea and the wavelength of pulse. [Ans: 552m, 2.4 cm]
5. Two ships are anchored at some distance away in the deep sea. An ultrasonic signal of 50 kHz is sent from one ship to another by two routes: one through water with velocity of 1372 m/s and the other through air with velocity of 342 m/s. Calculate the distance between the two ships. [Ans: 1372m]
6. A steel bar is tested using an ultrasonic flaw detector. The pulse arrival times were found to be 30 μ s and 80 μ s. If the bar is of 40 cm thick, find out the location of the defect. [Ans: 15 cm below the surface]
7. Calculate the depth of the sea if the time interval between the emitted signal and the echo received is 2 sec in sonar studies. Assume the velocity of sound in sea water as 1490 m/s. [Ans: 1490m/s]
8. Calculate the natural frequency of iron rod of 0.03 m length. The density of iron is 7.23×10^3 kg/m³ and Young modulus is 11.6×10^{10} N/m².
9. A quartz crystal of thickness 0.001 m is vibrating at resonance. Calculate its fundamental frequency if Young modulus of quartz = 7.9×10^{10} N/m² and the density is 2650 kg/m³. [Ans: 27.30 kHz]

CHAPTER

13

Electron Emission

13.1 INTRODUCTION

Modern technology is strongly electronics-oriented. A majority of the instruments, appliances and gadgets are based on the movement of electrons in their circuits for their operation. In many of the cases, the working of a device or a circuit cannot be understood without a proper knowledge of how electrons are controlled and manipulated at various stages. The understanding of the exciting developments would be possible only when the basic physical principles regarding the generation of electrons, their response to external stimuli are amply learnt.

The methods of generation of electrons are outlined in this chapter laying stress on thermionic emission which is widely used in many of the vacuum tube devices.

13.2 WORK FUNCTION

Any metal contains a large number of free electrons. Typically, the electron concentration in a metal is $10^{29}/\text{m}^3$. Periodic potential craters are located at the sites of positive ions. Each electron moves in the common field of all positive ions. Fig. 13.1 shows the potential energy of an electron as a function of distance in the metal. It is only at distances close to an ion that the variation in potential is perceptible. At greater distances, the metal may be viewed as an equipotential region. In that case, the details in Fig. 13.1 are ignored and the potential energy curve is represented by a simple potential well (Fig. 13.2). The electrons are distributed among various energy levels in the potential well. At normal temperatures, most electrons occupy the levels up to the energy level E_F and a small number with higher energy are at levels above E_F . The level E_F is known as **Fermi level**. A conduction electron may move freely within the interior of the metal but it cannot escape from the surface. The electron has to surmount the potential barrier on the metal surface if it is to escape from the metal. The existence of the surface barrier may be understood this way. If an electron attempts to escape from a metal, it will induce a positive charge on the surface, because the metal was originally neutral. Unless

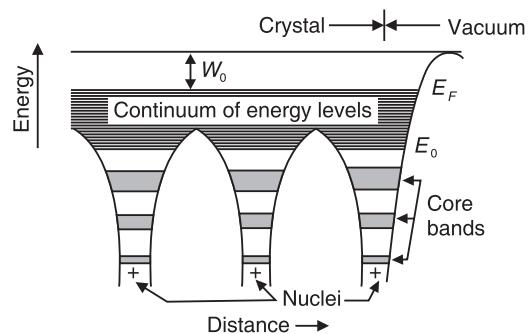


Fig. 13.1. Schematic representation of the energy barrier at the surface of a metal.

the electron possesses enough energy to move away from the region of influence of the induced positive charge, it will be compelled to fall back into the potential well.

Energy is to be supplied to the electron in order to remove it from the metal. An electron at the lowest level needs an energy E_0 , whereas an electron at the Fermi level needs an energy $(E_0 - E_F)$. The minimum energy required for an electron to be just emitted from a metal surface is called the **work function** of the particular metal. It is designated by W_0 and is expressed in electron – volts. Thus,

$$W_0 = E_0 - E_F \quad (13.1)$$

The magnitude of the work function depends on the energy level E_F and the height and shape of the energy barrier. Thus, the work function is a characteristic of the metal. The larger the work function, the harder it is to cause electron emission.

The work function for some metals are shown in Table 1.

Table 1: Work function of some metals

Metal	Work function, eV
Caesium	1.9
Lithium	2.5
Sodium	2.3
Potassium	2.2
Silver	4.7
Platinum	5.6

The work function for a metal is approximately half the ionization energy of a free atom of that metal. For example the ionization energy of caesium atom is 3.9 eV and the work function of caesium metal is 1.9 eV. Therefore, electron emission cannot occur at low temperature.

13.3 ELECTRON EMISSION

The liberation of electrons from a metal surface is known as **electron emission**. In order to liberate an electron from a metal, energy is to be supplied to the electron. The energy may be supplied in different ways. According to the way in which an electron receives the kinetic energy, different processes of electron emission are identified.

13.3.1 Types of Electron Emission

A metal can be made to emit electrons through the following four processes.

1. **Thermionic Emission:** Electrons are emitted from a metal when it is heated to a high temperature.
2. **Photoelectric Emission:** Electrons are emitted from a metal when it is illuminated with a high frequency light.
3. **Field Emission:** Electrons are emitted when a metal surface is subjected to a strong electric field of the order of millions of volts/m.
4. **Secondary Emission:** Electrons are emitted from a metal when it is bombarded by high speed particles.

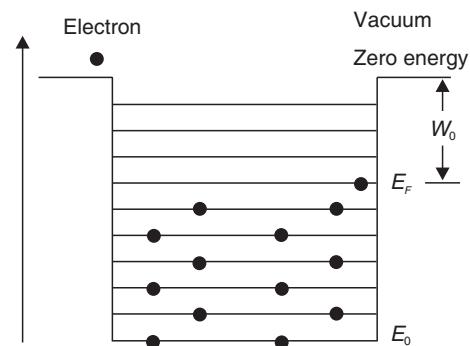


Fig. 13.2

Among the four methods, the thermionic emission is by far the most common method employed to generate free electrons in vacuum devices.

13.4 THERMIONIC EMISSION

The emission of electrons by a heated metal is known as *thermionic emission*. This effect was discovered by Thomas A. Edison in 1883 during the development of light bulbs. It is also called the *Edison effect*. In this process the metal is heated to about 2500°C . Even at lower temperatures there are a few electrons which can surmount the energy barrier. However, the number of such electrons increases sharply with increasing temperature. Metals with lower work function emit electrons at lower temperatures. The most commonly used materials are tungsten, thoriated tungsten etc.

13.4.1 Richardson-Dushman Equation

The emission current at a particular temperature can be calculated knowing the distribution of electrons in various energy levels. Such a calculation was performed and the result is known as **Richardson – Dushman equation**. It is given as,

$$J_S = AT^2 e^{-b/T} \text{ amp/m}^2 \quad (13.2)$$

where J_S = emission current density, i.e., current per square metre of the emitting surface,

T = Absolute temperature of emitter in K

A = Constant, depending on the type of emitter and is measured in $\text{amp/m}^2\text{K}^2$

$$= \frac{4\pi em k^2}{3}$$

The value of b is constant for a given metal and is given by $b = \frac{e\phi}{k} = 11,600 \frac{\phi}{k} \text{ K}$

where $\phi = W_0$, the work function of emitter.

Using the value of b into the equation (13.2), we get

$$J_S = AT^2 e^{\frac{-11600\phi}{T}} \quad (13.3)$$

The theoretical value of A is $A = 120 \text{ A/cm}^2\text{K}^2$.

The experiment values are considerably lower and differ from metal to metal.

The values of the emitter constants A and b may be obtained from the equation (13.2).

The equation (13.2) indicates that the saturation current is strongly dependent on the emitter temperature and work function. The higher the temperature the larger is the saturation current. Further, the smaller the work function the higher will be electron emission and hence the current density.

13.4.2 Thermionic Emission Materials

Thermionic emitters are operated at high temperature. They are expected to have a reasonable service life. Therefore, they have to fulfill a number of requirements.

- (i) **Low work function:** The material should have as low a work function as possible. Then the electron emission takes place at reasonably low temperatures.
- (ii) **High melting point:** The electron emission occurs at high temperatures ($>1500^{\circ}\text{C}$). Therefore, the material should have a high melting point.
- (iii) **Mechanical strength:** The emitter is held normally at zero potential or at a negative potential with respect to an anode. Hence, it is referred to as a **cathode**. The material used as cathode must be durable and the emission must be stable. Further, it should have high mechanical strength to withstand ion bombardment. Even in a high vacuum, there will be gas molecules present. The gas molecules get ionized

due to the collisions with energetic electrons. The resulting positive ions will get accelerated towards the cathode. As a result, the cathode will be subjected to ion bombardment. Unless the cathode is mechanically stronger it will get damaged.

13.4.3 Thermionic Cathodes

The requirement of operation at high temperatures restricts the number of suitable emitters to be used as cathodes. The most commonly used cathodes are tungsten, thoriated tungsten and oxide coated metals.

- (i) **Tungsten:** Tungsten is a high melting point metal (m.p. = 3300° C). It has a high work function (4.52 eV). It is operated at 2300° C. The advantages of tungsten are that it has high mechanical strength and highly consistent emission. Its use in ordinary devices such as vacuum diodes is discontinued. However, it is used in power transmitting tubes and X-ray tubes, where the anode voltages exceed 15 KV.
- (ii) **Thoriated Tungsten:** Thoriated tungsten is an improved material. A thoriated tungsten cathode consists of a tungsten filament coated with a layer of thorium. It has a lower work function of 2.63 eV and therefore it provides emission at a lower temperature of 1700° C.
- (iii) **Oxide Coated Cathode:** This is the most commonly used material. It consists of a nickel ribbon or cylinder coated with a thin layer of barium and strontium oxides. The oxide coated cathode has a low work function of 1.1 eV. It is operated at 700-900° C and has high emission efficiency.

The features of the various types of cathodes are summarized in Table 2.

Table 2: Thermionic cathode materials

Cathode	Work function eV	Working temperature °K	Js A/cm ²	A amp/cm ² deg ²	b °K	Life hr.	Emission efficiency mA/watt
Tungsten	4.52	2500	0.25	60	52,400	3000	4-20
Thoriated Tungsten	2.63	1900	1.5	3	30,500	10,000	50-100
Oxide	1.1	1000-1100	1.00	0.01	12,000	20,000	1000-10,000

13.4.4 Directly and Indirectly Heated Cathodes

There are mainly two types of thermionic cathodes, namely *directly heated cathodes* and *indirectly heated cathodes*. Some of the types are shown in Fig. 13.3.

(i) **Directly heated cathode:** A directly heated cathode is a wire or ribbon filament. The filament is usually bent in zig-zag fashion. The heating current passes directly through the filament and causes thermionic emission from the filament. The advantage is that the efficiency of conversion of electric power into the thermionic emission is more. But the disadvantage is that the variations in heater voltage affect the electron emission and thus produce hum in the circuit.

(ii) **Indirectly heated cathode:** An indirectly heated cathode consists of a nickel tube coated with barium and strontium oxides. A filament heater is enclosed within the tube and insulated from it by an aluminium oxide coating. The heating current is passed through the filament heater and the cathode is indirectly heated through the heat from the filament. The major advantage is that the cathode is completely isolated from the heating circuit. Therefore, it can be directly connected to any potential. Secondly, because the cathode has a larger mass,

temperature variations are practically non-existent. Finally, ac voltage can be used to heat the filament.

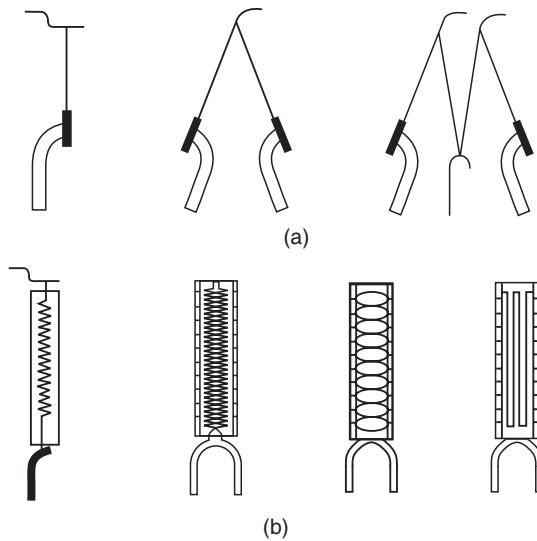


Fig. 13.3: Types of cathodes (a) directly heated (b) indirectly heated.

13.5 PHOTOELECTRIC EMISSION

Additional energy required by the electrons in a metal to surmount the surface barrier can be supplied by the application of light. When a metal is irradiated by a light of suitable frequency, electrons absorb the incident photon energy and leave the metal, as shown in Fig. 13.4. Such an emission of electrons caused by light is known as *photoelectric emission*.

In order to liberate an electron, the photon should have an energy greater than the work function of the metal. Thus, if the energy of the photon is $h\nu$, $h\nu$ should be greater than the work function, W_0 for electron emission. Thus,

$$\begin{aligned} h\nu &> W_0 \\ \text{or} \quad v &> \frac{W_0}{h} \\ \text{or} \quad \lambda &< \frac{ch}{W_0} = \frac{12400}{W_0} \text{ Å} \end{aligned} \tag{13.4}$$

The electron emission in this process depends upon the frequency of the incident radiation and on its intensity.

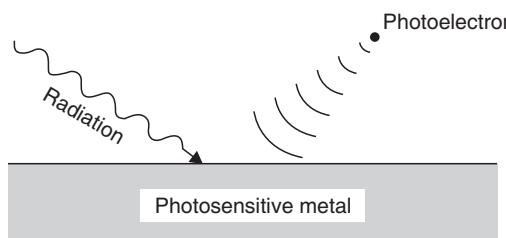


Fig. 13.4: Photoelectric emission

13.6 FIELD EMISSION

When an electric field is applied to a cathode surface, the surface barrier is reduced and the work function is also reduced. When an intense electric field of the order of 10^7 to 10^8 V/m is applied, the surface barrier is reduced to such an extent that electrons can be pulled from the surface. The emission of electrons due to an intense external electric field is known as **field emission**. It is also known as **cold emission** because the cathode is not heated in this process. This phenomenon is in fact due to a quantum mechanical process known as *tunneling*. The emission current varies approximately as $\exp(-\phi/E)$, where E is the intensity of the electric field.

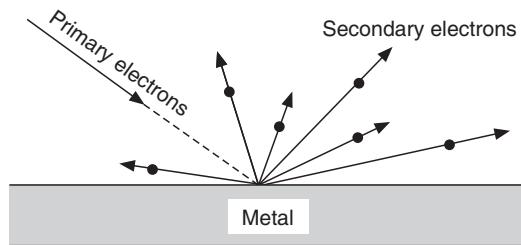


Fig. 13.5: Illustration of secondary emission

13.7 SECONDARY EMISSION

When high energy electrons strike a metallic surface they cause emission of more electrons as shown in Fig. 13.5. Such an electron emission from a metallic surface by energetic electrons is known as **secondary emission**. The electrons that strike the metal surface are called **primary electrons**. The electrons that are emitted from the metal surface are known as **secondary electrons**. The primary electrons penetrate the surface layer of the metal and give up their energy to electrons in that layer. Some of the electrons in the layer acquire sufficient energy and escape from the surface. Secondary electron emission takes place when the incident electrons have energies of the order of 10-15 eV and higher. This process has no direct relationship to the work function. Secondary emission occurs in different directions and with different energies. Often an incident primary electron causes emission of several secondary electrons. Secondary electron emission may also be caused due to bombardment by any energetic particles.

Secondary emission ratio, δ is defined by,

$$\delta = \frac{\text{Number of secondary electrons emitted}}{\text{Number of primary electrons incident}} \quad (13.5)$$

δ varies between 1.5 and 2 for most metals but can be as high as 10-15 for certain surfaces.

QUESTIONS

1. Explain the concept of work function.
2. Why are electrons not able to escape from a metal under normal conditions?
3. What are the different processes that cause electron emission from a metal?
4. What is meant by thermionic emission?
5. Write down Richardson-Dushman equation and explain the various terms.
6. What are the requirements that a metal has to fulfill in order to act as a thermionic emitter?
7. Which are the materials used commonly as thermionic emitters?
8. What are directly heated and indirectly heated cathodes? Explain their functioning.
9. Explain photoelectric emission.
10. Describe field emission.
11. Describe secondary emission.

CHAPTER

14

Electron Ballistics

14.1 INTRODUCTION

Electric and magnetic fields affect the motion of charged particles. This was found in the experiments on discharge through gas at low pressures. We can see that if one holds a strong permanent magnet near the face of a television, the picture gets distorted. Charged particles travel in straight line path in a vacuum and behave just as any classical point mass particle, when they are in a free state. They readily respond to the commands of electric and magnetic forces. In practice, macroscopic electric and magnetic fields are employed to subject charged particles to desired forces and to obtain predictable trajectories covering measurable distances. These forces are large enough to dominate the motion so that gravitational force and mutual repulsion between the particles become negligible. Further, the particle trajectories are confined to evacuated enclosures so that the particle paths are not distorted by collisions with atmospheric gas molecules and uninterrupted travel of the particles over measurable distances is ensured. Under these conditions particles follow geometrically simpler paths which can be predicted and analyzed by the application of the laws of classical Newtonian mechanics. Various combinations and configurations of electric and magnetic fields have been used to control and manipulate the particle trajectories. A thorough appreciation of electron dynamics led to the invention of an array of vacuum tubes whose applications constituted the discipline of electronics in its early stages. The word electronics is coined by fusing the word ‘mechanics’ to the word ‘electron’. Thus electron + mechanics = electronics.

We study in this chapter how the particle motion is affected by the electric and magnetic fields under variety of conditions, taking electron as representative of the group.

14.2 ELECTRIC FIELD

Pictorial Representation

Faraday proposed that the electric field can be conveniently visualized and represented in terms of **lines of force**. They are also called **electric field lines**. The electric field lines are supposed to emanate from positive charges and terminate on negative charges. Thus, the field lines indicate the **direction** of the electric field. The field lines are related to the electric field in the following manner. In a given region, the number of lines per unit area through a surface perpendicular to the lines is proportional to the **magnitude** of the electric field in that region. When the field lines are close, the electric field strength, E is large; and when the lines are apart, E is smaller, as illustrated in Fig. 14.1. The field lines are therefore used to indicate the magnitude of the field as well as its direction.

In a nonuniform electric field, the field lines are curved and spaced nonuniformly (Fig. 14.1). In a uniform electric field the field lines are straight, parallel, and uniformly spaced as shown in Fig. 14.2.

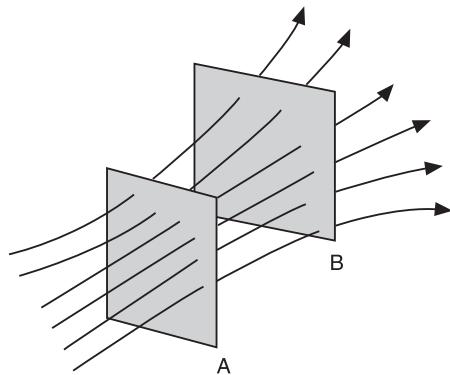


Fig.14.1: Field lines in a nonuniform field.
E is larger at A than at B.

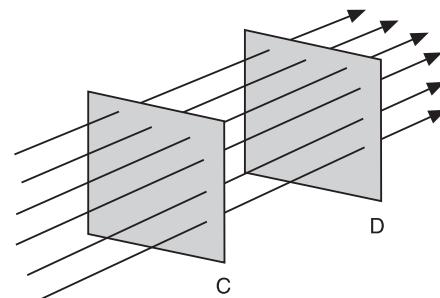


Fig.14.2: Field lines in a uniform field.
E is same at locations C and D.

Uniform Electric Field

An electric field in a region is said to be **uniform**, when the field lines are straight and parallel pointing in the same direction; and the strength of the field E is constant in that region. It is also known as a **homogeneous electric field**. For a uniform electric field,

$$E = \text{constant.}$$

Uniform electric fields are set up by a pair of metal plates held strictly parallel to each other and charged to opposite polarity (Fig. 14.3).

The electric field intensity E in the region between the plates A and B (Fig. 14.3) is given by

$$E = \frac{V}{d} \quad (14.1)$$

where $V = (V_A - V_B)$ is the potential difference between the plates A and B; and d is the distance of separation of the plates.

14.3 MOTION OF AN ELECTRON IN A UNIFORM ELECTRIC FIELD

When an electron enters a uniform electric field, it experiences a force and gets accelerated. The path of electron depends upon the angle between the applied field and the initial direction of the electron velocity. We assume that electric field alone acts on the particle and the gravitational force is negligible. The region between the plates is evacuated such that the electron is not deflected away from its path due to collisions with air molecules.

14.3.1 Electric Field Parallel to Initial Velocity

Let an electron of mass m and charge $-e$ enter the region of uniform electric field. In Fig. 14.4, the uniform electric field E is directed from positive plate A to the negative plate B.

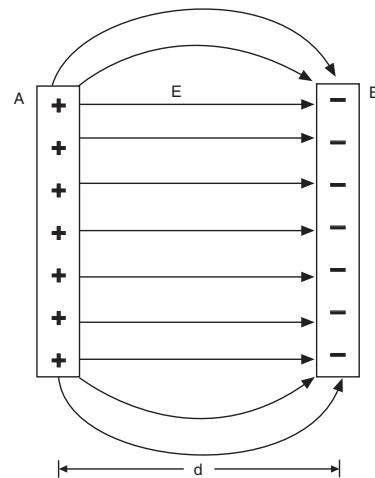


Fig.14.3 : Uniform electric field produced by a parallel plate capacitor. Field is nonuniform at the edges.

The electron experiences a force \mathbf{F} due to the electric field.

$$\mathbf{F} = -e\mathbf{E} \quad (14.2)$$

The negative sign indicates that the force acts in a direction opposite to that of electron motion. As a result, the electron moves with acceleration a in the direction opposite to that of the electric field. According to Newton's second law, the equation of motion of the electron is given by

$$ma = -eE \quad (14.3)$$

$$\text{Therefore, the acceleration } a \text{ is } a = \frac{-eE}{m}$$

The magnitude of the acceleration is given by

$$a = \frac{eE}{m} \quad (14.4)$$

As the parameters e and m in the above equation are constants and the electric field \mathbf{E} is uniform, the acceleration a is uniform. Hence, the electron is *uniformly accelerated* in a direction opposite to that of the electric field.

- As the electric field acts along a straight line, the path of the electron is a straight line. The electron motion resembles the motion of a body freely falling in a gravitational field.
- Equation (14.4) indicates that the acceleration of a charged particle in an electric field depends on the ratio e/m . Since the ratio e/m is different for different charged particles, their accelerations in a given electric field are different. In contrast, the acceleration due to gravity is the same for all bodies.

Equations of kinematics for rectilinear motion can now be applied to the motion of electron in an electric field. The distance traveled by the electron in the field during time t in x-direction (see Fig. 14.4) is

$$x = v_0 t + (-eE/2m)t^2 \quad (14.5)$$

where v_0 is the initial velocity of the electron. For convenience, we assume that the positive x-axis is towards left in Fig. 14.4. Note that v_0 is also in x-direction.

The velocity of the electron after time t is given by

$$v = v_0 + (-eE/m)t \quad (14.6)$$

and

$$v^2 = v_0^2 + \left(\frac{-2eE}{m} \right) x \quad (14.7)$$

The kinetic energy of the electron after moving through a distance 'x' in the field is

$$\text{K.E.} = \frac{1}{2}mv^2 = \frac{1}{2}mv_0^2 + \frac{1}{2}m\left(\frac{-2eEx}{m}\right) = (\text{K.E.})_0 + (-e)Ex \quad (14.8)$$

where $(\text{K.E.})_0$ is the initial kinetic energy of the electron.

Eq.(14.8) means that

- An electron acquires energy ' eEx ' when it travels in a uniform electric field parallel to the field lines.

If the electron is at rest initially, $v_0 = 0$ and the above equations reduce to

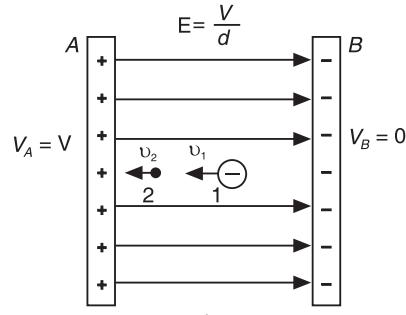


Fig.14.4: Electron motion in a uniform electric field

$$\left. \begin{array}{l} (i) \ x = (-eE / 2m)t^2 \\ (ii) \ v = (-eE / m)t \\ (iii) \ v^2 = \left(\frac{-2eE}{m} \right)x \\ (iv) \ K.E. = -eEx \end{array} \right\} \quad (14.9)$$

$$\text{As } -Ex = V, \text{ K.E.} = eV \text{ or } \frac{1}{2}mv^2 = eV \quad (14.10)$$

The equation (14.10) shows that the potential energy is converted into kinetic energy, when the electron moves from one potential difference to another. Equ.(14.10) helps us in finding the electron velocity v .

$$v = \sqrt{\frac{2eV}{m}} \quad (14.11)$$

That is,

$$v \propto \sqrt{V}$$

Thus, the velocity acquired by the electron in a uniform electric field is proportional to the square root of the potential difference through which it is accelerated.

Substituting the values of e and m of an electron into equ. (14.11), we obtain

$$v = 5.93 \sqrt{V} \times 10^5 \text{ m/s}$$

Example 14.1. An electron starts from rest and moves freely in an electric field of intensity 1500 V/m. Determine the force on the electron and acceleration attained by the electron.

Solution:

$$(i) F = eE = (1.602 \times 10^{-19} \text{ C})(1500 \text{ V/m}) = 2.4 \times 10^{-16} \text{ N}$$

$$(ii) a = \frac{eE}{m} = \frac{(1.602 \times 10^{-19} \text{ C})(1500 \text{ V/m})}{9.11 \times 10^{-31} \text{ kg}} = 2.6 \times 10^{14} \text{ m/s}^2.$$

Example 14.2: A proton is projected into a region of a uniform electric field of $3 \times 10^5 \text{ N/C}$. The proton travels 4 cm against the field direction before it comes to rest. Determine (i) the deceleration of the proton, (ii) its initial speed and (iii) the time taken by proton to come to rest.

Solution:

$$(i) \text{ The deceleration of the proton is given by } a = -\frac{eE}{m}$$

$$\therefore a = -\frac{(1.602 \times 10^{-19} \text{ C})(3 \times 10^5 \text{ N/C})}{1.67 \times 10^{-27} \text{ kg}} = -\frac{4.8}{1.67} \times 10^{13} \text{ N/kg}$$

$$\therefore a = -2.9 \times 10^{13} \text{ m/s}^2.$$

$$(ii) \text{ The initial speed is given by } v_0 = \sqrt{-2as}$$

$$\therefore v_0 = \sqrt{(-2)(-2.9 \times 10^{13} \text{ m/s}^2)(0.04 \text{ m})}$$

$$\therefore v_0 = 1.52 \times 10^6 \text{ m/s.}$$

$$(iii) \text{ Time taken by proton to come to rest } t = -\frac{v_0}{a} = \frac{1.52 \times 10^6 \text{ m/s}}{2.88 \times 10^{13} \text{ m/s}^2}$$

$$\therefore t = 5.3 \times 10^{-8} \text{ s} = 53 \text{ ns.}$$

14.3.2 Electron-Volt

An electron (or a charged particle) gains energy when it is accelerated in an electric field; the kinetic energy so gained is very small compared to a joule. Hence, in atomic physics, particle energies are expressed in terms of a smaller unit called electron-volt (eV). *An electron-volt is defined as the energy acquired by an electron when it gets accelerated through a potential difference of one volt.* We see that the electron volt is related to the joule through the relation

$$1 \text{ eV} = \text{Charge on the electron} \times 1 \text{ V} = (1.602 \times 10^{-19} \text{ C}) (1 \text{ V}) = 1.602 \times 10^{-19} \text{ J}$$

14.3.3 Electric Field Perpendicular to Initial Velocity

We now consider the case of a uniform electric field applied perpendicular to the initial direction of motion of the electron.

Let A and B be two plane parallel metal plates of length l oriented horizontally and separated by a small distance d . If a potential difference V is applied across the plates, a uniform electric field $E = (V/d)$ is produced in the region between the plates. The electric field is directed vertically downward from plate A to plate B, as shown in Fig. 14.5.

Let an electron be moving in x-direction with velocity v_0 . The electron velocity in y-direction is initially zero. At point K, the electron enters the uniform electric field. As the electric field acts in y-direction, the electron experiences an upward force and gets deflected upward in y-direction. The acceleration acquired by the electron in y-direction is given by

$$\therefore a_y = \frac{eE}{m} \quad (14.12)$$

The velocity attained by the electron after traveling for a time t in the electric field is

$$v_y = \frac{eE}{m} t \quad (14.13)$$

If y is the **displacement of electron** in the field direction during time t , then

$$y = \frac{eE}{2m} t^2 \quad (14.14)$$

Since the initial velocity v_0 is in a direction perpendicular to the electric field E , it remains unchanged. Therefore, the horizontal displacement of electron in x-direction in time t is

$$x = v_0 t \quad (14.15)$$

Therefore, the co-ordinates of the electron after time t are $\left\{ x = v_0 t, y = \frac{eE}{2m} t^2 \right\}$.

t is known as the **transit time** and is given by

$$t = \frac{x}{v_0} \quad (14.16)$$

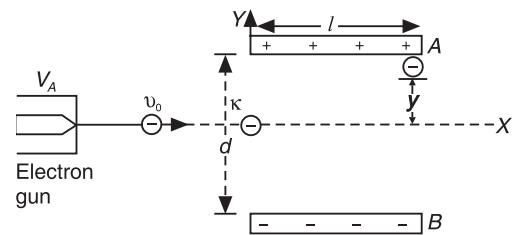


Fig. 14.5: Electron motion in a transverse uniform electric field. The electron describes a parabolic path in a transverse electric field.

Eliminating time t from equ. (14.14) and equ. (14.15), we obtain the equation of the path as

$$y = \left(\frac{eE}{2m} \right) \left(\frac{x}{v_0} \right)^2$$

$$y = \left(\frac{eE}{2mv_0^2} \right) x^2 \quad (14.17)$$

or

$$y = k x^2$$

where k is a constant.

This is an equation of a parabola. It means that

- An electron moving with uniform velocity follows a parabolic path when it passes through a transverse uniform electric field.

14.3.4 Electrostatic Deflection

The deflection of electron caused by an electrostatic field is known as **electrostatic deflection**. An electron follows a parabolic path in a transverse uniform electric field and finally, it emerges out of the electric field. Since, electron is an invisible particle, we cannot determine the electron displacement caused by the electric field, say at point M in Fig. 14.6. Therefore, we keep a fluorescent screen at some distance in the path of electron and locate its position. When the moving electron strikes fluorescent screen, it causes a luminous glow on the screen and makes its position known. Referring to Fig. 14.6, the electric field terminates at MN and the electron does not experience force to the right of the plates beyond the line MN. The electron then onwards follows a rectilinear path. It travels along the line MP with a velocity v and strikes the fluorescent screen at point P. If the electric field is switched off, the electron moves without deviation and strikes the screen at point Q. Therefore, QP is the linear deflection caused by the electric field. When the line MP is extended backward it cuts the axis KQ at O and the line PO represents the tangent to the parabola KM at M. The angle θ ($=\angle POQ$) represents the angular displacement of the electron path.

In Fig. 14.6, let $QP = D_E$, $OQ = L$, $AB = d$ and $AA' = l$.

From $\Delta^{le} OQP$, $QP = OQ \tan \theta$

\therefore Electrostatic deflection $D_E = L \tan \theta$

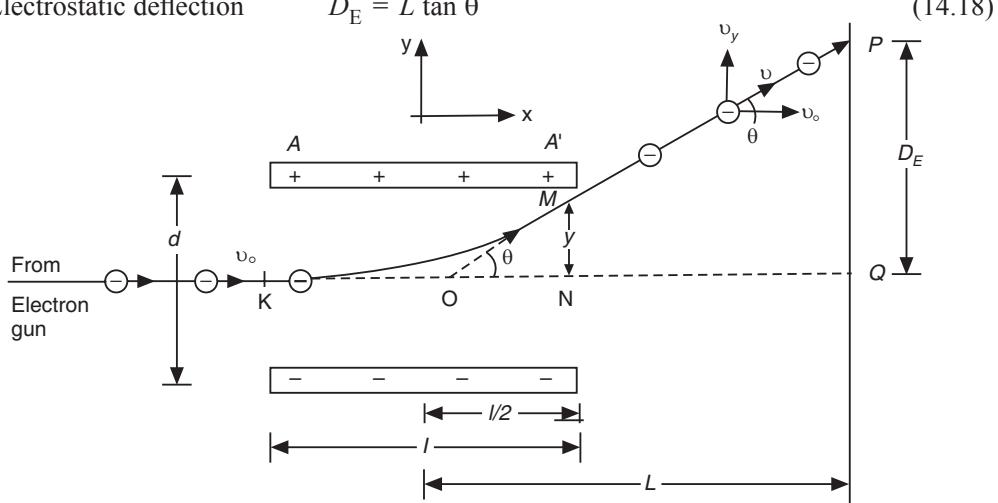


Fig. 14.6: Electrostatic deflection of an electron beam due to uniform electric field acting perpendicular to its path.

But $\tan \theta = \left[\frac{dy}{dx} \right]_{x=l} = \frac{d}{dx} \left[\frac{eEx^2}{2mv_0^2} \right]_{x=l} = \frac{eEl}{mv_0^2}$ (14.19)

Using equ.(14.19) into equ.(14.18), we get

$$D_E = \frac{LeEl}{mv_0^2}$$
 (14.20)

If we have a stream of electrons, they have the same e/m value, but may have different initial velocities in the x -direction. Therefore, the electrons are deflected according to their velocities. All electrons with a given value of v_0 , will reach a point P on the screen. If the initial velocity v_0 has been obtained by passing the electron through voltage V_A , we have

$$v_0 = \sqrt{\frac{2eV_A}{m}}$$
 (14.21)

Using the value for v_0 (equ.14.21) and $E = V/d$ into the equation (14.20), we obtain

$$\begin{aligned} D_E &= \frac{LeVl}{md} \cdot \frac{m}{2eV_A} \\ \therefore D_E &= \frac{LVl}{2dV_A} \end{aligned}$$
 (14.22)

Thus, D_E is proportional to the deflecting voltage V and inversely proportional to the accelerating voltage V_A .

- + When the accelerating voltage is smaller, the electron travels through the electric field sufficiently slowly and undergoes appreciable deflection. It is required that V_A be large so that the particles will have sufficient kinetic energy for the production of a luminous spot on the fluorescent screen. But a high accelerating voltage requires a high deflection potential V for causing a sufficient amount of deflection, D_E . Therefore, the value chosen for V_A has to be a compromise to meet these conflicting requirements.
- + The angle of deflection of the electron beam in electrostatic deflection is to be restricted to smaller values. Otherwise, at angles greater than a certain value, the electrons hit the deflection plates instead of reaching the screen. Because of this limitation, the area that can be covered on the screen by the electron beam on the screen is smaller.
- + The **deflection sensitivity**, S , of the deflection plates is given by the deflection caused by one volt of potential difference applied to deflection plates.

$$S = \frac{D_E}{V} = \frac{Ll}{2dV_A}$$
 (14.23)

The reciprocal of the deflection sensitivity S is called the **deflection factor**. Thus,

$$G = \frac{1}{S} = \frac{2dV_A}{IL}$$
 volts/m. (14.24)

Example 14.3: A proton has an initial velocity of 2.3×10^5 m/s in the x -direction. It enters a uniform electric field of 1.5×10^4 N/C in a direction perpendicular to the field lines.

- Find the time it takes for the proton to travel 0.05 m in the x -direction, and
- Find the vertical displacement of the proton after it has travelled 0.05 m in the x -direction.

Solution:

$$(i) \text{ Time taken by the proton } t = \frac{l}{v_o} = \frac{0.05 \text{ m}}{2.3 \times 10^5 \text{ m/s}} = 2.2 \times 10^{-7} \text{ s}$$

(ii) Displacement in the vertical direction due to electric field is

$$y = \frac{eEt^2}{2m} = \frac{(1.602 \times 10^{-19} \text{ C})(1.5 \times 10^4 \text{ N/C})(2.2 \times 10^{-7} \text{ s})^2}{2 \times 1.67 \times 10^{-27} \text{ kg}}$$

$$= 3.48 \times 10^{-2} \frac{(\text{kg.m/s}^2)\text{s}^2}{\text{kg}} = 3.5 \text{ cm.}$$

14.3.5 Electron Projected at an Angle into a Uniform Electric Field

In general, the initial velocity of the electron may not be parallel to x- or y-axes. Suppose an electron is projected into a uniform electric field at an angle θ , and with an initial velocity v_o as shown in Fig. 14.7.

As the electric field direction (Fig. 14.7) is along the positive y-direction, the electron gets accelerated in the negative y-direction. The acceleration is constant and is given by

$$a = \frac{eE}{m}$$

The acceleration is constant and the electron motion in the electric field closely resembles the motion of a projectile in the gravitational field. The component of electron velocity along x-axis is $v_o \cos \theta_o$ and the component along the y-direction is $v_o \sin \theta_o$. The horizontal component $v_o \cos \theta_o$ remains constant during motion, while $v_o \sin \theta_o$ decreases initially and again increases when the electron reverses its path.

The velocity components are:

$$(i) v_x = v_{x0} = v_o \cos \theta_o = \text{constant and}$$

$$(ii) v_y = v_{y0} + at = v_o \sin \theta_o + at$$

Therefore, the co-ordinates of the electron after time t are

$$x = v_{x0}t = (v_o \cos \theta_o)t \text{ and}$$

$$y = v_{y0}t + \frac{1}{2}at^2 = (v_o \sin \theta_o)t + \frac{1}{2}at^2.$$

Eliminating t from the above equations, we get the equation for the electron motion in the electric field. Thus,

$$y = (\tan \theta_o)x + \left(\frac{a}{2v_o^2 \cos^2 \theta_o} \right)x^2 \quad (14.25)$$

The above equation is of the form $y = ax + bx^2$ which represents a parabola. Therefore, The trajectory of an electron projected into a uniform electric field is a parabola.

As is the case of a projectile in a gravitational field, the parameters of motion of electron in the uniform electric field may be calculated.

The **maximum height** that the electron attains in the uniform electric field is

$$H = \frac{v_o^2 \sin^2 \theta_o}{2a} \quad (14.26)$$

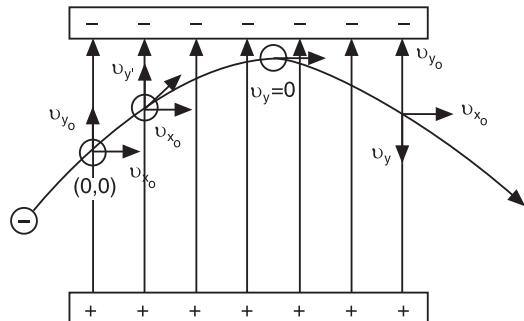


Fig. 14.7: Path of electron projected at an angle into a uniform electric field

The **time taken** by the electron to reach the maximum height, H is

$$t = \frac{v_0 \sin \theta_0}{a} \quad (14.27)$$

The **time of flight**, T , i.e., the time taken by electron to return to its initial position along x-direction is

$$T = \frac{2 v_0 \sin \theta_0}{a} \quad (14.28)$$

The **range**, R , i.e., the horizontal distance traveled by electron from the starting position to the point at which it returns to the initial position along x-direction is

$$R = \frac{v_0^2 \sin 2\theta_0}{a} \quad (14.29)$$

Example 14.4: An electron projected at an angle of 37° to the horizontal at an initial speed of 4.5×10^5 ms/ in a region of uniform electric field 200 N/C.

- (i) Find the time it takes the electron to return to its initial height.
- (ii) Find the maximum height reached by the electron.
- (iii) Find its horizontal displacement when it reaches its maximum height.

Solution:

$$(i) \text{ The time of flight, } T = \frac{2mv_0 \sin \theta}{eE} = \frac{2(9.11 \times 10^{-31} \text{ kg})(4.5 \times 10^5 \text{ m/s})(\sin 37^\circ)}{(1.602 \times 10^{-19} \text{ C})(200 \text{ N/C})}$$

$$= \frac{49.34}{320.4} \times 10^{-7} \frac{\text{kg.m/s}}{\text{N}} = 1.54 \times 10^{-6} \frac{\text{kg.m/s}}{\text{kg.m/s}^2} = 1.54 \mu \text{s}$$

$$(ii) \text{ Maximum Height, } H = \frac{m(v_0 \sin \theta)^2}{2eE} = \frac{(9.11 \times 10^{-31} \text{ kg})(4.5 \times 10^5 \text{ m/s})^2 (\sin 37^\circ)}{2(1.602 \times 10^{-19} \text{ C})(200 \text{ N/C})}$$

$$= 1.04 \times 10^{-3} \frac{\text{kg.m}^2/\text{s}^2}{\text{kg.m/sec}} = 104 \times 10^{-3} = 1.04 \text{ mm}$$

- (iii) Horizontal displacement at the position of maximum height,

$$s = \frac{(9.11 \times 10^{-31} \text{ kg})(4.5 \times 10^5 \text{ m/s})^2 \sin 74^\circ}{2(1.602 \times 10^{-19} \text{ C})(200 \text{ N/C})} = 2.77 \times 10^{-3} \frac{\text{kg.m}^2/\text{s}^2}{\text{kg.m/s}} = 2.77 \text{ mm}$$

14.4 UNIFORM MAGNETIC FIELD

A *uniform magnetic field* is a region of space where the lines of magnetic induction are parallel to each other and the magnetic induction is identical in magnitude and direction at all points in that region. It is also called a **homogeneous magnetic field**.

A uniform magnetic field is produced by a solenoid or an electromagnet.

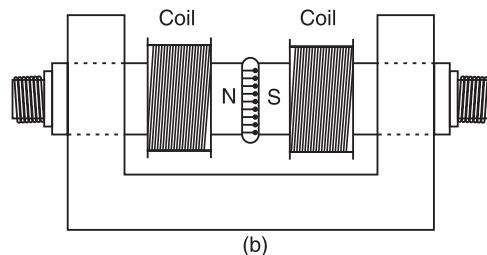
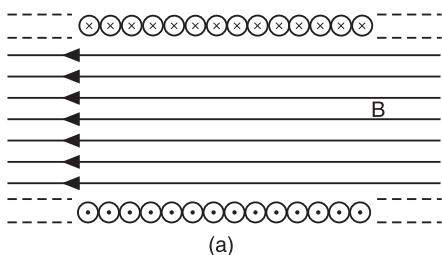


Fig. 14.8: Production of uniform magnetic field – (a) by a solenoid and (b) by an electromagnet.

The SI unit of the magnetic field is Weber per square metre (Wb/m^2), also called tesla (T). The cgs unit of magnetic field is gauss (G). It is related to tesla as follows:

$$1 \text{ T} = 10^4 \text{ G}$$

Thus the tesla is relatively a large unit. For comparison, the earth's magnetic field is about 0.5 G or 5×10^{-5} T, the field of a small permanent magnet is about 100 G or 10^{-2} T and that of large magnets is upto 20,000 G or 2T. The magnets of large particle accelerators produce 6T or 60,000 G while super conducting magnets generate fields as high as 25 T or 250,000 G.

14.5 MOTION OF AN ELECTRON IN A UNIFORM MAGNETIC FIELD

Let an electron enter the region where a uniform magnetic field acts. If θ is the angle between the velocity of the electron and the direction of the uniform magnetic field, then the magnetic force on the electron is given by

$$\mathbf{F} = e(\mathbf{v} \times \mathbf{B})$$

or

$$\mathbf{F} = ev \mathbf{B} \sin \theta \quad (14.30)$$

where \mathbf{B} is the magnetic induction, and v is the velocity of the electron. If v and B are at right angles to each other (as in Fig. 14.9), $\theta = 90^\circ$ and hence

$$\mathbf{F} = evB \quad (14.30a)$$

The direction of the magnetic force is determined as follows.

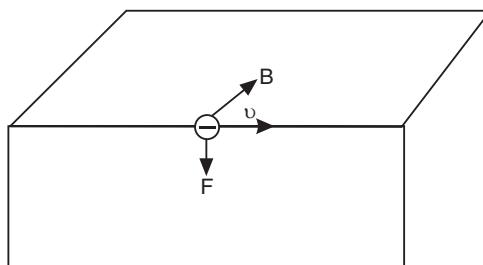


Fig. 14.9: Direction of force acting on a moving electron due to a transverse uniform magnetic field

The vectors \mathbf{v} and \mathbf{B} are drawn or visualized from a common point, as shown in Fig. 14.9. Next, a plane is visualized in which the two vectors lie. The force vector lies along a line perpendicular to this plane. To determine the direction of \mathbf{F} , the vector \mathbf{v} is turned into vector \mathbf{B} through the small angle between them. The direction of \mathbf{F} is the direction in which a left-hand screw will advance as \mathbf{v} rotates into \mathbf{B} .

Example 14.5: An electron having velocity 10^6 m/s experiences a maximum force of $1.6 \times 10^{-14} \text{ N}$ when it enters a uniform magnetic field. What is the magnitude of the magnetic field?

Solution: The magnitude of the magnetic field $B = \frac{F}{ev} = \frac{1.6 \times 10^{-14} \text{ N}}{(1.602 \times 10^{-19} \text{ C})(10^6 \text{ m/s})} = 0.1 \text{ T}$

14.5.1 No Energy is Gained by an Electron in a Uniform Magnetic Field

The magnetic force, \mathbf{F} , is *always* perpendicular to the velocity v and as well as the magnetic field, B . Therefore, the work done by this force on the electron is zero. In other words, the magnetic field does not produce any change in either the speed or the kinetic energy of the electron.

$$\mathbf{F} \cdot \mathbf{v} = 0 \quad (14.31)$$

i.e.

$$m \mathbf{a} \cdot \mathbf{v} = 0$$

$$m \left(\frac{d\mathbf{v}}{dt} \right) \mathbf{v} = 0 \quad \text{or} \quad \frac{d}{dt} \left(\frac{mv^2}{2} \right) = 0$$

$$\frac{1}{2} mv^2 = \text{constant} \quad (14.32)$$

It means that *an electron moves in the magnetic field without acquiring or losing energy*. It follows that the magnetic field cannot change the speed v of the electron. The *only* other parameter that can be affected due to the magnetic field is the direction of the velocity of the electron. Therefore, the effect of the magnetic field on the electron is to change the direction of motion of the electron. The path followed by the electron in a magnetic field depends on the angle at which it enters the magnetic field.

14.5.2 Magnetic Field Parallel to Initial Velocity

- (a) A static magnetic field does not act on an electron which is at rest. For an electron at rest, $\mathbf{v} = 0$ and hence $F = e\mathbf{v} \cdot \mathbf{B} \sin \theta = 0$.
- (b) When an electron enters a uniform magnetic field parallel to the magnetic field lines, the magnetic force acting on the electron is zero.

As $\theta = 0^\circ$, $F = e\mathbf{v} \cdot \mathbf{B} \sin \theta = 0$

Similarly, when an electron moves opposite to the field lines, i.e., $\theta = 180^\circ$, then again, $F = e\mathbf{v} \cdot \mathbf{B} \sin \theta = 0$

As the force is zero in the above two cases, the acceleration is zero. Therefore, the electron continues to move along the initial direction of motion without suffering any change in its speed or direction of motion.

14.5.3 Magnetic Field Perpendicular to Initial Velocity

We consider now the case of an electron entering a uniform magnetic field with its initial velocity perpendicular to the field. Then the angle θ between \mathbf{v} and \mathbf{B} is 90° and $\sin \theta = 1$.

$$\therefore F = e\mathbf{v} \cdot \mathbf{B} \quad (14.33)$$

We assume that the magnetic field \mathbf{B} is acting into the page (Fig. 14.11). This is indicated by the crosses, which represent the tail of \mathbf{B} vector. The force \mathbf{F} due to magnetic field acts perpendicular to the velocity \mathbf{v} and \mathbf{B} . This force does not perform work on the electron and hence cannot change the magnitude of electron velocity. However, the action of the force is to deflect the electron from its path and hence change the direction of velocity. After an infinitesimal change of direction, the electron has the same speed and is still moving perpendicular to

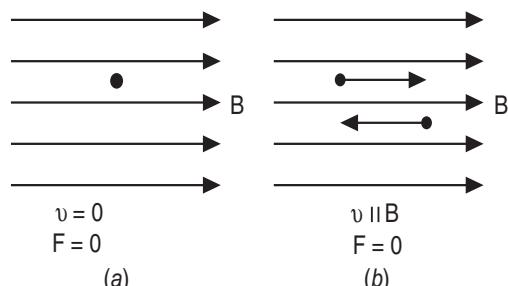


Fig. 14.10: Electron in a uniform magnetic field. (a) When an electron is at rest, it is not acted upon by the magnetic field. (b) A longitudinal field B does not act on a charged particle.

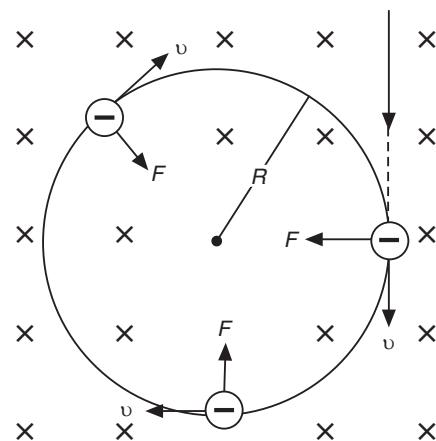


Fig. 14.11: An electron describes a circular path in a transverse uniform magnetic field

\mathbf{B} ; but the direction of \mathbf{F} changes. As the force \mathbf{F} deflects the electron, the directions of \mathbf{v} and \mathbf{F} change continuously and the electron follows a curved path. The force that produces the curvature is \mathbf{F} . Since the magnetic field is uniform everywhere, and the speed \mathbf{v} is constant and is unchanged by the field, the magnitude of \mathbf{F} is constant. As the direction of the force \mathbf{F} needs to be always perpendicular to the velocity of the particle, it is obviously a *centripetal force* and causes the electron to move in a circular path in a plane perpendicular to the magnetic field. The sense of rotation in Fig. 14.11 is clockwise for an electron. The centripetal acceleration caused by the force due to magnetic field is v^2/R , where R is the radius of the curvature. The centripetal force \mathbf{F} is given by

$$F = \frac{mv^2}{R} \quad (14.34)$$

Comparing equations (14.33) with (14.34), we get

$$evB = \frac{mv^2}{R} \quad (14.35)$$

$$\therefore R = \frac{mv}{eB} \quad (14.36)$$

It is seen that the radius of the circle depends on the momentum ' mv ' of the electron. Thus,

$$R \propto mv$$

The radius of the orbit of an electron moving at right angles to the magnetic field is proportional to its momentum.

The larger the momentum, the ~~larger~~ is the radius of the ~~of~~ electron path and the smaller is the curvature. It is because it becomes more difficult for the magnetic field to change the direction of motion of the electron with greater momentum.

The time taken by the electron to complete one revolution is called the **time period, T**. It is given by

$$\begin{aligned} T &= \frac{\text{Distance travelled by the electron in one revolution}}{\text{Speed of the electron}} \\ &= \frac{\text{Circumference of the path}}{\text{Speed of the electron}} \end{aligned}$$

$$\begin{aligned} \therefore T &= \frac{2\pi R}{v} = \frac{2\pi}{v} \cdot \frac{mv}{eB} \\ \therefore T &= \frac{2\pi m}{eB} \end{aligned} \quad (14.37)$$

The frequency of revolution in the orbit is given by

$$v = \frac{1}{T} = \frac{eB}{2\pi m} \quad (14.38)$$

Equ.(14.37) and equ.(14.38) indicate that

- The time period and frequency of revolution are independent of the electron velocity as well as the radius of circular path.

This is because v and R adjust themselves in such a way that for a given magnetic induction value \mathbf{B} , T and v remain constant. It implies that

- Slower electrons move in smaller circles while faster electrons move in larger circles but all of them take the same time for completion of one revolution.

Example 14.6: An electron is accelerated through a potential difference of 5 kV and enters a uniform magnetic field of 0.02 wb/m^2 acting normal to the direction of electron motion. Determine the radius of the path.

Solution:

The radius of the circular path described by the electron in the magnetic field is given by

$$R = \frac{mv}{eB}$$

The velocity of the electron accelerated through a potential V is given by

$$v = \sqrt{\frac{2eV}{m}}$$

$$\therefore R = \frac{1}{B} \sqrt{\frac{2mV}{e}}$$

$$= \frac{1}{0.02 \text{ wb/m}^2} \left[\frac{2(9.11 \times 10^{-31} \text{ kg})(5 \times 10^3 \text{ V})}{1.602 \times 10^{-19} \text{ C}} \right]^{1/2}$$

$$= 119.3 \times 10^{-4} \frac{\text{C.m N.s}}{\text{N.s C}} = 12 \text{ mm.}$$

14.5.4 Relation between K.E. of the Electron and Radius of the Circular Path

Suppose an electron is moving in a circular path under the action of transverse uniform magnetic field B. If v is the speed of the electron, the kinetic energy of the electron is given by

$$E_k = \frac{1}{2} mv^2 \quad (14.39)$$

$$\text{But, } v = \frac{eBR}{m} \quad (14.40)$$

$$\therefore E_k = \frac{e^2 B^2 R^2}{2m} \quad (14.41)$$

Note that this is *not* the kinetic energy gained by the electron due to its motion in the magnetic field. The kinetic energy of the electron given by the equation (14.39) is the energy that the electron has prior to its entry into the magnetic field, and is expressed in terms of the radius of the circular path in the magnetic field in equ. (14.41).

14.5.5 Magnetic Field Acting at an Angle to Initial Velocity

Let us consider a general case where an electron enters the uniform magnetic field B with its velocity at an angle θ , as shown in Fig. 14.12.

Let an electron enter a uniform magnetic field B with a velocity v making an angle θ with the magnetic field direction. The velocity v may be resolved into parallel and perpendicular components with respect to the magnetic field direction. The components are

$$v_{||} = v \cos \theta$$

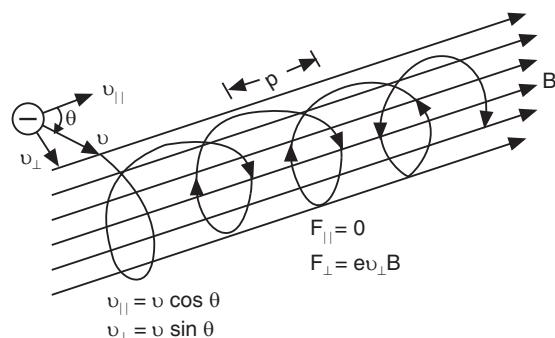


Fig. 14.12: An electron entering a uniform magnetic field at an angle describes a helical path

and

$$v_{\perp} = v \sin \theta$$

Axial velocity component causes uniform rectilinear motion: The force on the electron in the direction parallel to the magnetic field is zero.

$$F_1 = ev_{||} B = 0$$

Consequently, the velocity component, $v_{||}$ which is parallel to the magnetic field, does not undergo a change due to the magnetic field. The electron continues to move along the field direction with the velocity $v_{||}$. This is rectilinear motion.

Perpendicular velocity component causes circular motion: The force acting on the electron due to the magnetic field in a direction perpendicular to the field is

$$F_{\perp} = ev_{\perp} B$$

Due to this normal force component, the electron travels along a circular path around the field direction with constant speed $v_{\perp} = v \sin \theta$. This is uniform circular motion.

Resultant motion is helical motion around the magnetic field direction: The actual motion of the electron is the **resultant** of the above two motions—a uniform rectilinear motion parallel to the field and a uniform circular motion perpendicular to the field. The superposition of circular motion on the rectilinear motion results in motion along a helical or spiral path. Thus, the electron describes a *helical path* in the magnetic field, the axis of the helix being parallel to the magnetic field, B.

The projection of the path onto the xy plane is a circle, whose **radius** is given by

$$R = \frac{mv_{\perp}}{eB} = \frac{mv \sin \theta}{eB} \quad (14.42)$$

The projections of the path onto the xz and yz planes are sinusoids.

The **time period** of the revolution is given by

$$T = \frac{2\pi R}{v_{\perp}} = \frac{2\pi mv \sin \theta}{eBv \sin \theta} = \frac{2\pi m}{eB} \quad (14.43)$$

The distance covered in one revolution is the **pitch of the helix**. It is given by

$$p = v_{||} T = T v \cos \theta$$

$$\therefore p = \frac{2\pi mv \cos \theta}{eB} \quad (14.44)$$

Example 14.7: A 2 keV positron is projected into a uniform magnetic field of induction 0.1 T with its velocity vector making an angle of 89° with the field. Find the period, pitch and radius of the helical path of the positron.

Solution:

Time period is given by

$$T = \frac{2\pi m}{eB} = \frac{2(3.143)(9.11 \times 10^{-31} \text{ kg})}{(1.602 \times 10^{-19} \text{ C})(0.1 \text{ T})}$$

$$= 3.6 \times 10^{-10} \frac{\text{kg} \cdot \text{m}}{\text{s}} \frac{\text{s}^2}{\text{kg} \cdot \text{m}} = 0.36 \text{ ns}$$

Pitch of the helical path is given by $p = T v \cos \theta = T(593 \times 10^3 \sqrt{V}) \cos \theta$

$$= 3.6 \times 10^{-10} \text{ s} \times 593 \times 10^3 \times \sqrt{2000} \text{ m/s} \times \cos 89^{\circ} = 0.17 \text{ mm}$$

Radius of the helical path is given by R

$$= \frac{mv}{eB} \sin \theta$$

$$\begin{aligned}
 &= \frac{(9.11 \times 10^{-31} \text{ kg})(593 \times 10^3 \sqrt{2000} \text{ m/s}) \sin 89^\circ}{(1.602 \times 10^{-19} \text{ C})(0.1T)} \\
 &= 1.5 \text{ mm}
 \end{aligned}$$

Example 14.8: An electron shot into a uniform magnetic field at an angle 60° moves in a spiral of diameter of 10 cm and with a period of 6×10^{-5} s. Determine the magnetic induction and electron velocity.

Solution:

$$(i) \text{ The period of revolution } T = \frac{2\pi m}{eB}. \quad \therefore B = \frac{2\pi m}{eT}$$

$$\text{or } B = \frac{2 \times 3.143 \times 9.11 \times 10^{-31} \text{ kg}}{6 \times 10^{-5} \text{ s} \times 1.602 \times 10^{-19} \text{ C}} = 5.95 \times 10^{-7} \text{ wb/m}^2.$$

$$(ii) \text{ The radius of the spiral } R = \frac{mv \sin \theta}{eB} \quad \therefore v = \frac{eBR}{m \sin \theta}$$

$$\text{or } v = \frac{(1.602 \times 10^{-19} \text{ C})(5.95 \times 10^{-7} \text{ wb/m}^2)(10 \times 10^{-2} \text{ m})}{9.11 \times 10^{-31} \text{ kg} (\sin 60^\circ)} = 1.2 \times 10^4 \text{ m/s.}$$

14.6 MAGNETOSTATIC DEFLECTION

The deflection produced in the path of electron by a magnetic field is called *magnetostatic deflection*. Fig. 14.13 shows a uniform magnetic field B acting in a region of space over a length l .

Let us consider an electron beam traveling in the horizontal direction with velocity v which enters a region where a uniform magnetic field acts in a transverse direction. As the beam travels through the magnetic field, it bends through an arc of radius ' R '. Before it completes a revolution, it emerges from the field, as the magnetic field is of limited extension. After emerging from the magnetic field, the beam travels along a straight line and strikes a fluorescent screen, say at point Q . In the absence of magnetic field, the electron beam would have struck the screen at the point P . PQ represents the amount of deflection experienced by the electron beam due to the action of magnetic field. Let the amount of deflection be denoted by D_H .

$$\therefore PQ = D_H$$

The circular arc AC subtends an angle θ at F . The radii, FA and FC , of the circle have their centre at F . AO and OCQ are tangents to the arc AC at points A and C respectively.

$$\therefore \angle POQ = \theta$$

From the Fig. 14.13 we can write

$$D_H = L \tan \theta \quad (14.45)$$

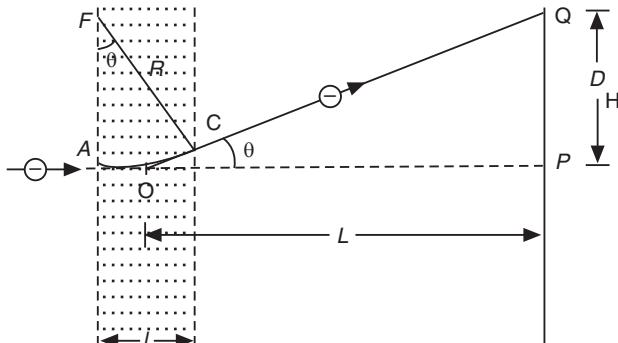


Fig. 14.13: Magnetostatic deflection. The dots represent uniform magnetic field emerging out of the plane of the page

The circular arc AC subtends an angle θ at F . The radii, FA and FC , of the circle have their centre at F . AO and OCQ are tangents to the arc AC at points A and C respectively.

$$\therefore \angle POQ = \theta$$

where L is the distance of the screen from the centre of the magnetic field. When θ is small, we can write

$$D_H \approx L \left(\frac{AC}{AF} \right) \approx \frac{IL}{R} \quad (14.46)$$

where $l (=AC)$ is the extent of the magnetic field. We know that

$$R = \frac{mv}{eB}$$

and

$$v = \sqrt{\frac{2eV_A}{m}}$$

Using the above relations into equ.(14.45), we obtain

$$D_H = \frac{LleB}{mv} = \left[\frac{LleB}{m} \right] \left[\sqrt{\frac{m}{2eV_A}} \right] \quad (14.47a)$$

where V_A is the voltage that caused electron to attain velocity v .

$$\therefore D_H = LIB \sqrt{\frac{e}{2mV_A}} \quad (14.47)$$

It is seen that *the magnetic deflection D_H is inversely proportional to the square root of the accelerating voltage V_A* .

Following the example of electrostatic deflection, we can define the *deflection sensitivity* in this case also.

$$S_H = \frac{D_H}{B} = Ll \sqrt{\frac{e}{2mV_A}} \quad (14.48)$$

Similarly, the *deflection factor* is expressed as

$$G_H = \frac{1}{S_H} = \frac{1}{Ll} \sqrt{\frac{2mV_A}{e}} \quad (14.49)$$

- The magnetostatic deflection is inversely proportional to the square root of accelerating potential V_A . Comparing with the electrostatic deflection (14.22), it is readily seen that
- For a given value of accelerating voltage the amount of deflection caused by a magnetic field will be more. Therefore, larger area of the fluorescent screen can be covered employing magnetic deflection than with electrostatic deflection. In view of this, magnetic deflection is employed in picture tubes of TV, radar etc.

Comparison between electrostatic and magnetic deflections

1. The angle of deflection of the electron beam in electrostatic deflection is to be restricted to smaller values. Otherwise, at angles greater than a certain value, the electrons hit the deflection plates instead of reaching the screen. Because of this limitation, the area that can be covered on the screen by the electron beam on the screen is smaller. In case of magnetic deflection it is possible to cover larger area on the screen.
2. In electrostatic deflection, the deflection sensitivity decreases more rapidly with increasing accelerating voltage than in magnetic deflection.
3. The electrostatic deflection requires little power for deflection. A large power is consumed in the coils for producing magnetic field in case of magnetic deflection.

Example 14.9: What transverse magnetic field acting over the entire length of a cathode ray tube must be applied to cause a deflection of 3 cm on the screen that is 15 cm away from the anode and if the accelerating voltage is 2000V?

Solution: Equ. (14.47) may be rewritten in terms of B as

$$\begin{aligned} B &= \frac{D_H \sqrt{2 m V_A}}{L l \sqrt{e}} = \frac{(0.03\text{m}) \left[2(9.11 \times 10^{-31} \text{kg})(2000\text{V}) \right]^{1/2}}{(0.075\text{m})(0.15\text{m})(1.602 \times 10^{-19} \text{C})^{1/2}} = 402.2 \times 10^{-6} \frac{(\text{kg.V})^{1/2}}{\text{C}^{1/2} \cdot \text{m}} \\ &= 4.02 \times 10^{-4} \frac{(\text{kg.J/C})}{\text{C}^{1/2} \cdot \text{m}} = 4 \times 10^{-4} \frac{\text{kg.m}}{\text{C.s.m}} = 4 \times 10^{-4} \text{Wb/m}^2 \left(\text{as } 1\text{kg.C.s} = 1\text{Wb/m}^2 \right) \end{aligned}$$

14.7 LORENTZ EQUATION

When an electron travels in a region where both electric and magnetic fields act, it will experience both an electric force $e\mathbf{E}$ and a magnetic force $e(\mathbf{v} \times \mathbf{B})$. So, the total force is the vector sum of the electric force and the magnetic force. Thus,

$$\mathbf{F}_L = e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \quad (14.50)$$

This expression is called the **Lorentz equation** and the force **Lorentz force** because it was identified in this form by H. Lorentz.

14.8 CROSSSED ELECTRIC AND MAGNETIC FIELD CONFIGURATION

When uniform electric and magnetic fields are perpendicular to each other and act over the same region of space, they are said to be in *crossed configuration*.

Let us consider uniform electric field acting vertically downward, which is set up by a pair of charged parallel plates, as shown in the Fig. 14.14. A uniform magnetic field is produced by, say, a set of Helmholtz coils placed on either side of the electric field plates and by passing current through the coils. The magnetic field is applied perpendicular to the page and acting into the page. When an electron beam passes through the region along a direction normal to both the field directions (Fig. 14.14), it is simultaneously subjected to the action of the two fields. The electric force acts upward whereas the magnetic force acts downward.

The electron beam is deflected in a direction along which the resultant force acts. The force on the electron due to electric field is

$$\mathbf{F}_E = e\mathbf{E}$$

The force on the electron due to magnetic field is

$$\mathbf{F}_B = ev\mathbf{B}$$

The net force acting on the electron is

$$\mathbf{F}_R = \mathbf{F}_E - \mathbf{F}_B$$

If $F_E > F_B$, the electron is deflected upward. On the other hand, if $F_B > F_E$, it is deflected downward. If the magnitudes of electric field \mathbf{E} and that of the magnetic field \mathbf{B} are adjusted such that $F_E = F_B$, then the net force on the electron beam is zero.

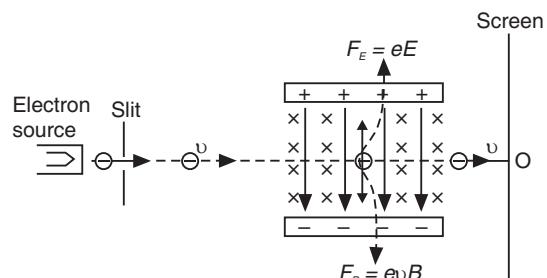


Fig. 14.14: Crossed fields acting normal to electron path. When the forces F_E and F_B balance each other the electron moves along the original straight line path

$$F_R = F_E - F_B = 0$$

or $eE - evB = 0 \quad (14.51)$

It means that the forces due to electric field and magnetic field balance each other and the electron beam passes through E and B fields without any deviation from its initial path. The forces balance each other when

$$eE = evB \quad (14.52)$$

or $E = vB$

The velocity of the electron may then be expressed as

$$v = \frac{E}{B} \quad (14.53)$$

This method of compensating fields was first used by J.J.Thomson in 1897 for measuring the *velocity* of an electron beam. He adjusted the field ratio E/B until the path of the investigated beam became rectilinear in the crossed fields.

Example 14.10: An electron beam passes through a magnetic field of 2×10^{-3} wb/m² and an electric field of 4×10^4 V/m acting simultaneously. If the path of the electron remains undeviated, determine the speed of the electron.

Solution: The speed of the undeviated electron is given by $v = \frac{E}{B}$

$$\therefore v = \frac{4 \times 10^4 \text{ V/m}}{2 \times 10^{-3} \text{ wb/m}^2} = 2 \times 10^7 \frac{\text{V/m}}{\text{kg/C.s}} = 2 \times 10^7 \frac{\text{N.s}}{\text{kg}} = 2 \times 10^7 \text{ m/s}$$

14.9 VELOCITY SELECTOR

A **velocity selector** is an electro-optic device which utilizes uniform electric and magnetic fields in crossed configuration for selecting a stream of single-velocity charged particles from a beam of particles having a range of velocities. This device is also known as a **velocity filter**.

A narrow beam of electrons traveling in vacuum enters a region where a uniform electric field **E** and a uniform magnetic field **B** are acting at right angles to each other. Let the electrons in the beam have velocities in the range of $v_0 \pm \Delta v$ spread around a central value v_0 . Let the beam enter the crossed field region in a direction normal to both the fields, as shown in Fig. 14.15.

The uniform electric field **E** is produced in the vertical direction by a set of charged parallel plates. The uniform magnetic field **B** is applied perpendicular to **E** over the same region of space. Both **E** and **B** act normal to the direction of velocity v_0 of electrons. The electric field produces an upward force on the electrons in the beam, whereas the magnetic field produces a downward force. Now, if the field values are adjusted so that the electric force balances the magnetic force, then the electrons will not be subjected to any net force. Thus, the resultant force is

$$F_R = F_E - F_B = 0$$

or

$$eE = ev_0B$$

$$\therefore v_0 = \frac{E}{B}$$

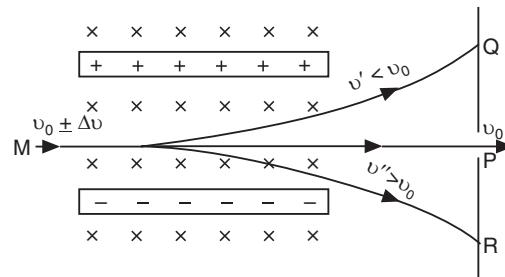


Fig. 14.15 Schematic of a velocity selector

Thus, there exists a unique velocity v_0 for which the electric and magnetic forces exactly cancel. Hence, the electrons traveling with the velocity $v_0 = E/B$ pass through the region of fields without suffering any deflection. Subsequently, they pass through the slit QR. Electrons traveling with velocities less than v_0 spend more time in the electric field region and are subject to electric force for a longer time. As a result, they get deflected upward and hit the slit walls at points such as Q. The force due to magnetic field is proportional to the velocity of electron. Therefore, electrons moving with velocities greater than v_0 are deflected more by the magnetic field. Hence, electrons having larger velocities are deflected downward and strike the slit at points such as R. The deflected electrons are absorbed by the slit walls. Electrons that have velocity v_0 are not deflected and pass through the aperture at P. The electron beam emerging from the aperture has a strictly single velocity.

14.10 PARALLEL ELECTRIC AND MAGNETIC FIELD CONFIGURATION

Case (A)

- (i) If the initial velocity of the particle is zero, it is accelerated by the electric field but does not experience force due to magnetic field, because the particle has velocity parallel to magnetic field.
- (ii) If the initial velocity of the particle is directed along the fields the magnetic field does not exert force on the particle. The resultant motion solely depends on the electric field.

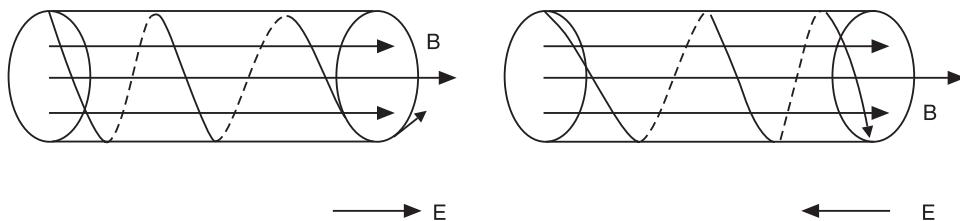


Fig. 14.16: (a) The path of a particle in the parallel uniform electric and magnetic fields.
 (b) The path of a particle in antiparallel uniform electric and magnetic fields.

In both the cases, the charged particle travels in a direction parallel to the fields with a constant acceleration, $a = eE/m$.

Case (B)

If the charged particle has an initial velocity component perpendicular to the magnetic field, then this component gives rise to a circular motion. The radius of the path is given by

$$r = \frac{mv \sin \theta}{qB}$$

which is independent of \mathbf{E} . However, the velocity component parallel to the field direction goes on changing continuously because of the acceleration due to the electric field. It may be noted that the resulting motion occurs along a helical path having a pitch $p = v_{\parallel} T$, which changes with time.

$$v_{\parallel} = \frac{qE}{m} t \quad (14.54)$$

14.11 e/m OF ELECTRON

Thomson's apparatus consisted of a highly evacuated glass tube, similar to the one shown in Fig. 14.17. Electrode C is the cathode from which electrons emerged. Electrode A is the

anode, which was maintained at a high positive potential. The high potential difference applied between the cathode C and anode A caused emission of electrons from the surface of the cathode. These electrons are accelerated and hit the electrode A but some of them passed through the aperture in A. These rays were further collimated by another electrode A', in which there was another aperture. The narrow electron beam passed through a region between plates P and P' and hit the fluorescent coating on the opposite side of the tube. A luminous spot is produced at point O. A uniform electric field \mathbf{E} was set up in the region between the plates P and P' when a potential difference is applied across them. A pair of Helmholtz coils was placed on either side of the tube. When a current was passed through the coils, a uniform magnetic field \mathbf{B} was set up perpendicular to \mathbf{E} in the same region where electric field acted. Now let us say the electric field alone is switched on and the electrons are deflected upward along OO₁. Then the magnetic field is switched on and by properly adjusting the magnetic field B, we can make the magnetic force equal to electric force. Then the two forces balance each other and the luminous spot returns from O to O₁. For the condition of balance

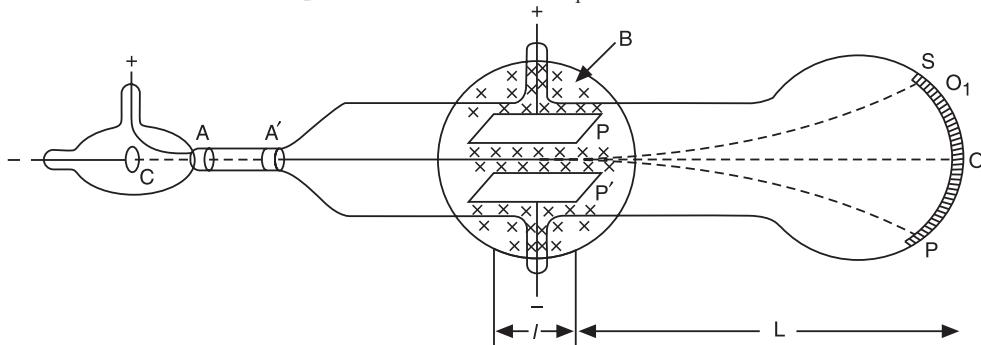


Fig.14.17: Schematic diagram of Thomson's apparatus for the determination of e/m of electron

$$F_R = F_E - F_B = 0$$

$$\therefore eE = evB$$

$$\therefore v = \frac{E}{B}$$

The electric field is now switched off. The electron beam experiences now a force $F = evB$ due to the magnetic field alone. It gets deflected along a circular arc of radius R and strikes the screen at P. The centre of curvature of the trajectory is at C and radius, R, of the curvature of the path is given by

$$R = \frac{mv}{eB}$$

$$\therefore \frac{e}{m} = \frac{v}{BR} \quad (14.55)$$

Using equ. (14.54) into equ. (14.55), we get

$$\frac{e}{m} = \frac{E}{B^2 R} \quad (14.56)$$

14.11.1 Determination of R

From Fig. 14.18 it is seen that CE and CF are radii of the arc of the electron path in the magnetic field. EO is tangent to the arc at E while GP is the tangent to the same arc at F. As

the radius CE moved to a position CF through angle θ , the tangent EO moves through the same angle θ and goes to the position GP.

$$\begin{aligned}\therefore \angle ECG' &= \angle OGP = \theta \\ \angle G'EC &= \angle GOP = 90^\circ \\ \therefore \Delta^{le} ECG' &\equiv \Delta^{le} OGP \\ \frac{EG'}{EC} &= \frac{OP}{OG} \\ \frac{l}{R} &= \frac{D}{L} \\ \text{or } R &= \frac{IL}{D} \quad (14.57)\end{aligned}$$

Now, using equ.(14.57) into equ. (14.56) we get

$$\frac{e}{m} = \frac{E}{B^2} \cdot \frac{D}{IL} \quad (14.58)$$

As $E = V/d$, we may write equ. (14.58) as

$$\frac{e}{m} = \frac{VD}{LdB^2} \quad (14.59)$$

The quantities l , L , and d are geometric quantities. The voltage V , magnetic field B , and deflection D are measurable quantities. Knowing the values of these quantities, the value of e/m can be computed. This experiment provided one of the first reliable experiments for measuring e/m . Thomson measured e/m for electrons and found a unique value for this quantity which was independent of the cathode material and the residual gas in the tube. The modern value of e/m of electrons is found to be $1.759 \times 10^{11} \text{ C/kg}$.

Example 14.11: A beam of electrons passes undeflected through two mutually perpendicular electric and magnetic fields. If the electric field is cut off and the same magnetic field is maintained, the electrons move in a circular path of radius 1.14 cm. Determine the ratio e/m , if $E = 8 \text{ kV/m}$ and $B = 2 \times 10^{-3} \text{ T}$.

Solution:

The velocity of the undeflected electrons is given by $v = E/B$.

The radius of the circular path is given by $R = \frac{mv}{eB} = \frac{mE}{eB^2}$.

$$\therefore \frac{e}{m} = \frac{E}{RB^2} = \frac{8 \times 10^3 V/m}{(0.014m)(2 \times 10^{-3} T)^2} = 1.754 \times 10^{11} \text{ C/kg.}$$

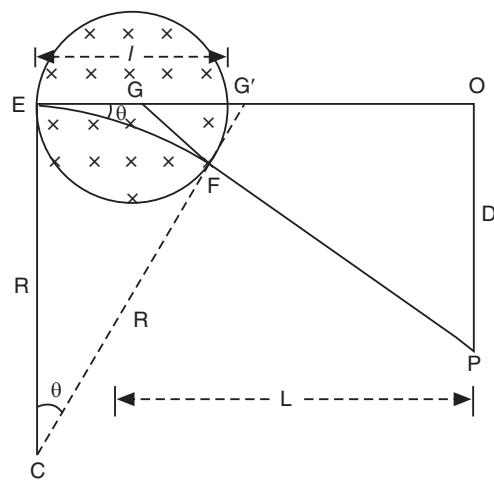


Fig.14.18: Determination of radius of the circular path from the geometry of magnetostatic deflection

14.12 CHARGE OF THE ELECTRON

Thomson's experiment enabled the determination of the value of e/m but from it either the charge e or mass m of the electron could not be determined independently. In 1917, Robert A. Millikan (1868-1953), the American physicist determined the magnitude of the elementary charge with high accuracy, using charged oil droplets supported by electric field between two parallel horizontal plates. The method is now popularly known as **Millikan oil drop method**.

Fig. 14.19 shows a schematic of the set up used by Millikan. It mainly consists of an observation chamber which contains two optically plane metal discs A and B of about 20 cm in diameter. The discs are held strictly parallel to each other at a distance of about 1.5 cm with the help of insulating rods of glass or ebonite. The upper disc A is provided with a tiny hole. The observation chamber is provided with three windows W_1 , W_2 and W_3 . One of the windows facilitates illumination of the chamber by an intense beam of light. The second window permits observation of the region between the discs. The air in this region can be ionized by X-rays which are allowed through the third window. An atomizer is used to spray minute droplets of a heavy nonvolatile oil which find their way into the region between the discs through the hole in A . The droplets of oil get charged due to the frictional effects at the nozzle of the atomizer. A high voltage power supply establishes a suitable potential difference between the discs A and B . The chamber is enclosed in a thermostat in order to avoid disturbances by thermal convection currents.

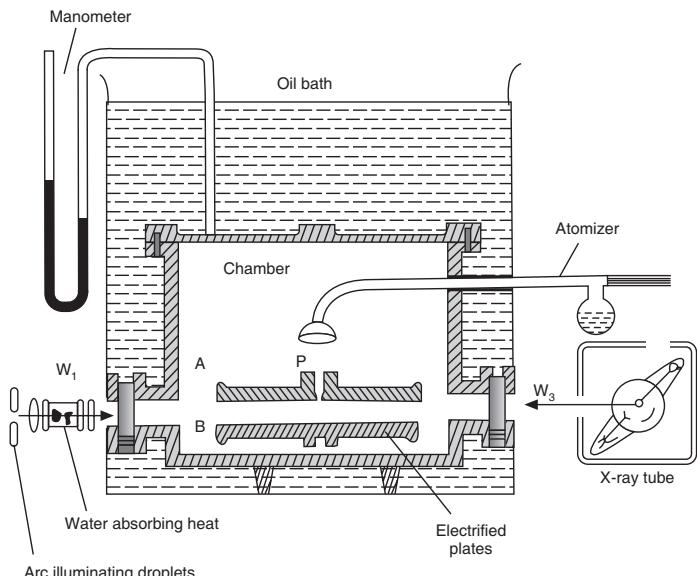


Fig. 14.19

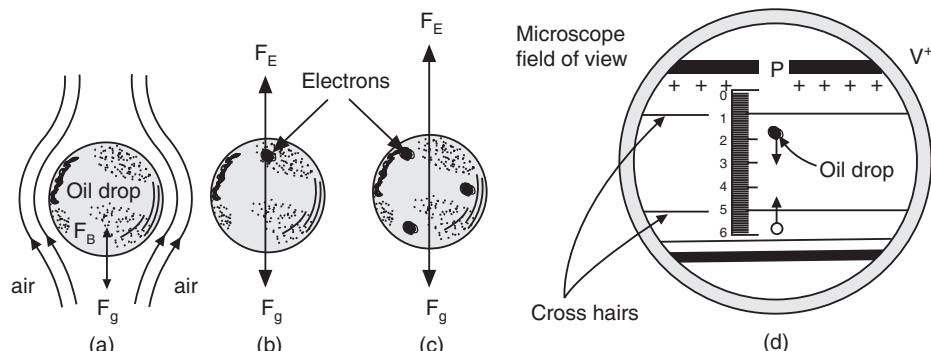


Fig. 14.20

Now, let the atomizer spray minute oil droplets into the closed space between the horizontal discs A and B . The oil drops are illuminated by a light and one of them is selected for observation through a microscope provided with a millimeter scale in the eyepiece. As an oil droplet falls in air, three forces act on it as shown in Fig. 14.20. They are the force of gravity W , the buoyant force F_B and the force of resistance F_R . The net force

$$F = W - F_B - F_R \quad (14.60)$$

- (i) If ρ is the density of the oil, r the radius of the oil drop, then the force of gravity is given by

$$W = mg = (\text{volume of the drop} \times \text{density of oil}) g$$

or
$$W = \frac{4}{3}\pi r^3 \rho g \quad (14.61)$$

- (ii) The buoyant force is the upward thrust experienced by the oil drop. According to Archimedes principle, it equals to the weight of the air displaced by the oil drop. Thus,

$$F_B = \frac{4}{3}\pi r^3 \sigma g \quad (14.62)$$

where σ is the density of the air.

- (iii) The gravitational and buoyant forces are constants. The force of resistance F_R depends on the velocity of the oil drop. F_R increases rapidly with velocity. As the oil drop continues to fall, its velocity increases leading to an increase in F_R . When the retarding force F_R equals $(W - F_B)$, the net force F on the drop becomes zero. At this instant, acceleration of the oil drop becomes zero and the drop falls with constant velocity v_o which is known as the *terminal velocity*.

According to Stoke's law, the resistive force on a spherical body due to the viscosity of the medium is given by

$$F_R = 6\pi r \eta v_o \quad (14.63)$$

where v_o is the terminal velocity of the body, and η is the coefficient of viscosity of air.

Under equilibrium, the net force on the oil drop is zero.

$$\begin{aligned} W - F_B - F_R &= 0 \\ \frac{4}{3}\pi r^3 \rho g - \frac{4}{3}\pi r^3 \sigma g - 6\pi r \eta v_o &= 0 \end{aligned} \quad (14.64)$$

$$\frac{4}{3}\pi r^3 (\rho - \sigma) g = 6\pi r \eta v_o \quad (14.65)$$

$$\therefore v_o = \frac{2 r^2 (\rho - \sigma) g}{9\eta} \quad (14.66)$$

or
$$r = \left[\frac{9\eta v_o}{2(\rho - \sigma) g} \right]^{\frac{1}{2}} \quad (14.67)$$

Experimentally, the terminal velocity of the oil drop is measured by determining the time during which it descended through the distance between two cross-wires in the field of view of the microscope (Fig. 14.20d). As all other quantities in equ.(14.67) are known, the radius r of the oil drop can be determined.

Now, let the electric field be switched on. If the drop carries a charge q , it experiences an additional upward force $F_E = -qE$ along with F_B and F_R . Let the terminal velocity be reduced to v_E .

The equation of motion of the oil drop in electric field is given by

$$\begin{aligned} \frac{4}{3}\pi r^3 (\rho - \sigma) g - 6\pi r \eta v_E - qE &= 0 \\ \therefore \frac{4}{3}\pi r^3 (\rho - \sigma) g &= 6\pi r \eta v_E + qE \end{aligned} \quad (14.68)$$

Using equ.(14.65) into equ.(14.68), we get

$$\begin{aligned} 6\pi r \eta v_o &= 6\pi r \eta v_E + qE \\ \therefore q &= \frac{6\pi r \eta}{E} (v_o - v_E) \end{aligned} \quad (14.69)$$

Using equ.(14.67) for r into equ.(14.69) one obtains

$$q = \frac{6\pi\eta}{E} \left[\frac{9\eta v_o}{2(\rho-\sigma)g} \right]^{\frac{1}{2}} (v_o - v_E) \quad (14.70)$$

or

$$q = \frac{9\pi}{E} \left[\frac{2\eta^3 v_o}{(\rho-\sigma)g} \right]^{\frac{1}{2}} (v_o - v_E)^* \quad (14.71)$$

Thus, by measuring the speed of free fall of an oil droplet v_o and the speed v_E in a known electric field of strength E , one can find the charge on the droplet. In measuring the speed v_E at a certain value of the charge q , Millikan ionized the air in the region with X-rays. New ions adhere to the droplet under observation and changed its charge, as a result of which the speed v_E is changed. After measuring the new value of v_E , the space between the discs was again irradiated and so on. Repeating the measurements a great number of times, Millikan established that the charge q on an oil droplet and the changes in charge Δq were some integral multiples of the same elementary charge. The value of the elementary charge obtained by Millikan was $e = 1.592 \times 10^{-19}$ C, quite close to the best modern result

$$e = 1.602 \times 10^{-19} \text{ C} \quad (14.72)$$

Millikan was awarded the Nobel prize in Physics in 1923.

14.13 MASS OF THE ELECTRON

From the values of (e/m) and e , one can compute the mass of the electron.

$$m = \frac{e}{e/m} = \frac{1.602 \times 10^{-19} \text{ C}}{1.759 \times 10^{11} \text{ C/kg}}$$

$$\therefore m = 9.109 \times 10^{-31} \text{ kg} \quad (14.73)$$

The mass as given by equ.(14.73) is known as the **rest mass** of the electron. It is $\frac{1}{1836}$ th of the mass of the hydrogen atom.

14.14 RADIUS OF THE ELECTRON

If the electron is assumed to be a charged particle having a spherical shape, its electrostatic potential energy is given by

$$E = \frac{e^2}{4\pi\epsilon_0 r} \quad \text{where } r \text{ is the radius of the electron.}$$

According to Einstein's mass-energy equivalence relation, the electron rest mass energy must be equal to its electrostatic potential energy. Therefore,

$$E = m_o c^2 = \frac{e^2}{4\pi\epsilon_0 r}$$

$$\text{or } r = \frac{e^2}{4\pi\epsilon_0 m_o c^2} \quad (14.74)$$

*When the drop begins to move upward instead of downwards, the equation of motion is

$$qE - \frac{4}{3}\pi r^3 (\rho - \sigma)g = 6\pi r\eta v_E$$

which yields

$$q = \frac{9\pi}{E} \left[\frac{2\eta^3 v_o}{(\rho - \sigma)g} \right]^{\frac{1}{2}} (v_o + v_E)$$

$$= \frac{(1.602 \times 10^{-19} \text{ C})^2 (9 \times 10^9 \text{ N.m}^2/\text{C}^2)}{(9.11 \times 10^{-31} \text{ kg})(3 \times 10^8 \text{ m/s})^2}$$

$$\therefore r = 2.82 \times 10^{-15} \text{ m} \quad (14.75)$$

Estimates from other methods give the same order of magnitude.

14.15 POSITIVE RAYS

E.Goldstein (1850-1930) the German physicist designed a special discharge tube while investigating the properties of cathode rays and discovered positive rays in 1886. A perforated cathode is kept between an anode and a fluorescent screen in the discharge tube, as illustrated in Fig. 14.21. When a discharge is initiated in the tube, scintillations are observed on the screen. Goldstein discovered that the scintillations were due to rays passing through the canal in the cathode and hence called them **canal rays**.

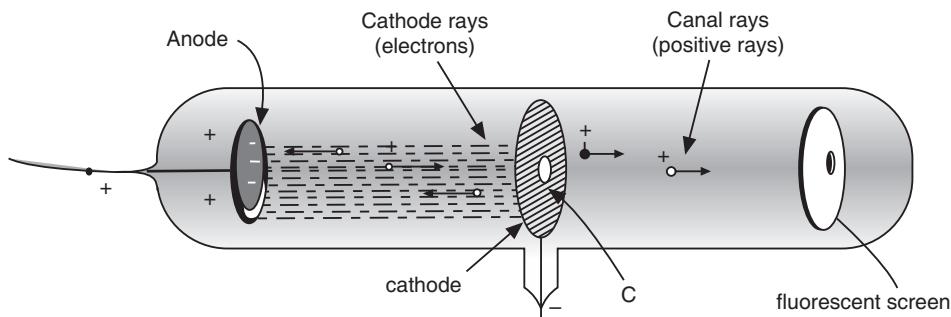


Fig. 14.21

In 1896, W.Wien observed that the canal rays could be deflected by a magnetic field and from the direction of deflection he established that the canal rays consisted of positively charged particles. Therefore, the canal rays are commonly known as **positive rays**. The origin of the positively charged particles is as follows. In the discharge tube, the electrons emitted by the cathode are accelerated toward the anode and on their way collide with the atoms of the gas in the tube. As a result of collisions, electrons are knocked off the neutral atoms and the atoms become positively charged. Any process by which an electron is removed from an atom is called **ionization** and the resulting positively charged particle is called a **positive ion**. The positive ions generated during electron and atom collisions are accelerated toward the cathode. Some of the ions pass through the hole in the cathode and strike the fluorescent screen giving out tiny flashes of light. It is thus seen that the positive rays are positively charged atoms of the gas contained in the discharge tube.

Since the mass of the electron is very small, the mass of the positive ion will be almost equal to the mass of the atom. Therefore, determination of the mass of positive ion yields the mass of the corresponding atom. The mass M and electric charge q of the positive ions can be determined by deflecting them in electric and magnetic fields.

The method of determining the charge to mass ratio using crossed field configuration described in Art. 14.11 is suitable only when all particles in a beam posses the same speed. In case of electrons, all electrons are emitted at cathode and all of them are accelerated by the same potential difference applied between the cathode and anode. Consequently, all the

electrons in a beam possess almost the same speed. Positive ions, on the other hand, are formed as a result of ionization of atoms of a gas taking place at different regions in the discharge tube. Therefore, some of the positive ions would be accelerated through only a small part of the field, some through the entire field and others in between. The resulting positive ions consist of a stream of ions with a wide range of speeds. In view of this, the method described in Art. 14.11 cannot be applied to determine q/M of positive ions.

14.16 THOMSON'S PARABOLA METHOD

In 1911, J.J.Thomson developed a method of measuring the relative masses of different atoms by employing a parallel configuration of electric and magnetic fields.

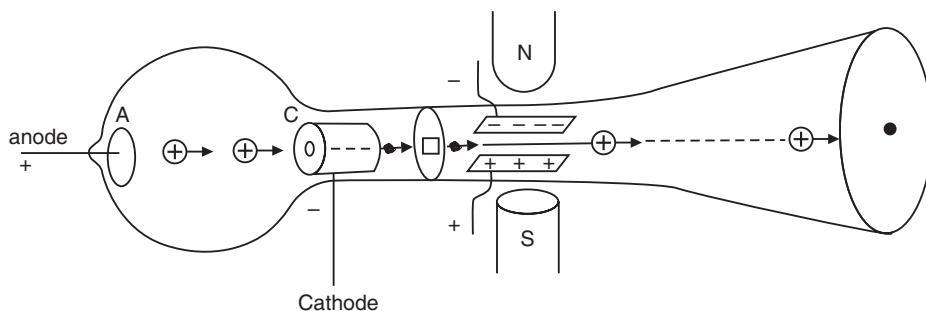


Fig. 14.22

Description

The apparatus used by Thomson is illustrated in Fig. 14.22. The discharge tube, in which positive rays are produced, is a large flask. A perforated cathode C is placed in the neck of the flask. An anode A is held opposite to the cathode at the head of the flask. The discharge tube is evacuated with the help of a vacuum pump. A small quantity of the gas, the mass of whose atoms is to be determined, is admitted into the tube. A pair of horizontal metal plates is arranged in the narrow tube beyond the cathode. When a potential difference is applied across these plates, a vertically acting electric field will be set up in the region. The narrow portion of the tube is held between the pole-pieces of an electromagnet in such a way that the magnetic field acts in the vertical direction over the same region where the electric field would be acting. A photographic plate is arranged at the far end of the chamber.

Working

When a large potential difference is applied between the cathode and anode, electrons emitted from the cathode ionize the gas atoms in the region between the anode and the cathode. Getting accelerated toward the cathode, many of the positive ions pass through the narrow hole in the cathode. The narrow pencil of positive rays then passes through electric and magnetic fields and strikes the photographic plate. When the electric and magnetic fields are not switched on, the positive ions will strike the photographic plate at a point in line with the initial direction of the beam.

When electric field alone is switched on the ions will be deflected upwards and hit the photographic plate at a distance y from the undeflected position. The deflection y is given by

$$y = \frac{EqL}{Mv^2} \quad (14.76)$$

a relation similar to eq. (14.20).

In the above equation

E is the strength of the electric field,

L , the distance from the centre of the field to the photographic plate,

l , the length of the electric field region/plates,

v , the velocity of the positive ion,

q , the charge of the ion and,

M , the mass of the ion.

When magnetic field alone is switched on the ions will be deflected into ‘the pages’ and hit the photographic plate at a distance x from the undeflected spot. x is given by

$$x = \frac{BqLl}{Mv} \quad (14.77)$$

where B is the magnetic field induction. This equation is similar to (14.47a).

When both the electric and magnetic fields are switched on and act simultaneously the positive ion beam experiences the forces due to electric and magnetic fields. The resultant force causes the ion beam to strike the photographic plate at a point (x,y) .

Eliminating v from equations (14.76) and (14.77), we get

$$y = \frac{E}{B^2 L l} \cdot \frac{M}{q} \cdot x^2 \quad (14.78)$$

This is the equation of a parabola. Therefore, ions having identical values of q/M and different values of speed produce a trace in the form of a parabola on the photographic plate. If there are ions of different masses but with the same charge, different parabolas are traced. Fig. 14.23(b) shows the first parabolas obtained by Thomson.

We can rewrite equ. (14.78) as

$$\frac{q}{M} = \frac{x^2}{y} \cdot \frac{E}{B^2 L l} \quad (14.79)$$

The value of q/M can be found by measuring the coordinates of the parabolic trace and using them into equ.(14.79) along with the known values of E , B , L and l . In practice, one parabola associated with ions of known mass is identified and all the other parabolas are compared with it and their masses are deduced. Usually hydrogen is present in the discharge tube and the hydrogen parabola is taken as the standard. Then the mass of the ions of the gas under investigation is found in terms of the hydrogen ion mass.

In Fig. 14.24, let the parabola corresponding to hydrogen ions be AB , while the ion of unknown mass but of the same charge as the hydrogen ion produce the parabola CD . Since hydrogen is the lightest element, its displacement in magnetic field is more and the parabola AB is located at the top. The initial line OX cannot be identified on the plate. When the direction of magnetic field is reversed, the relevant coordinate reverses its sign and parabolas GH and EF symmetrical to the initial ones are obtained. A vertical line such as GA is drawn to cut the parabolas. Then,

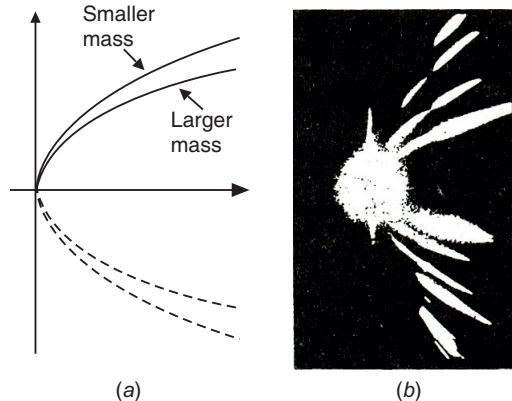


Fig. 14.23: (a) The locus of ions on the face of the positive ray tube is one of half of a parabola. Ions with different values of q/m fall on different parabolas. (b) Photograph of parabolas obtained by J.J.Thomson with ions of xenon, krypton, argon, and neon.

$$\frac{q/M_H}{q/M_u} = \frac{y_H^2/x}{y_u^2/x} = \frac{y_H^2}{y_u^2}$$

$$\therefore M_u = M_H \left[\frac{GA}{EC} \right]^2 \quad (14.80)$$

By measuring GA and EC on the photographic plate, and using the values into eqn. (14.80), M_u can be readily determined.

In 1912, Thomson attempted to determine the atomic weight of neon using parabola method. He discovered two parabolas, a strong one corresponding to a mass 20 and a weaker one corresponding to mass 22. Thus, for the first time, Thomson found two kinds of neon atoms, identical in chemical nature and having the same optical spectra but different in mass. In fact, it was shown later by Thomson's coworker, F.W.Aston that there are three kinds of neon atoms: 90.5% have a mass value of 20, 9.2% have a mass 22 and 0.3% have a mass 21.

Fredrick Soddy (1877-1956), the British Chemist, gave the name **isotopes** to the atoms of different weight belonging to the same element. Thus, neon has three isotopes. Thomson's method provided the first evidence to the existence of stable isotopes.

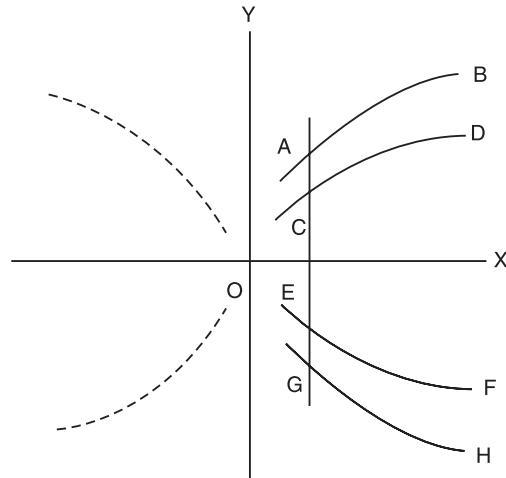


Fig. 14.24

QUESTIONS

- What is Lorentz force? Discuss the path traveled by an electron in a uniform transverse electric field. **(R.T.M.N.U., 2006)**
- Show that the velocity acquired by an electron in a uniform electrostatic field varies as the square root of potential difference through which it is accelerated.
- When and why a charged particle entering magnetic field follows a helical path?
- Define an electron volt. Express it in joules.
- Describe the motion of an electron subjected to the following forces:
 - A uniform electric field acting normal to the electron velocity;
 - A uniform magnetic field acting perpendicular to the direction of the electron motion,
 - When a uniform electric and magnetic field which are perpendicular to each other act normal to the direction of the electron motion.
- Show that the energy of a charged particle remains constant when it moves in magnetic field.
- Describe the trajectories of electron of mass ' m ' and charge ' e ' moving with velocity v in :
 - Uniform magnetic field
 - Uniform electric field
 In both the cases field is applied perpendicular to the direction of motion of particle.
- Explain how a charged particle fired into uniform electric field describes a parabolic motion, and when fired into a uniform magnetic field describes circular and helical motion. Find the expression for radius of the circle and pitch of the helix.
- Prove that the electron moving with uniform velocity in transverse electric field traces parabolic path. **(Amaravati Univ., 2003,2005)**

10. Show that an electron moving with uniform velocity follows a parabolic path in a transverse uniform electric field. **(Amaravati Univ., 2004)**
11. Explain the motion of electron in transverse electric field and obtain expression for deflection. **(Amaravati Univ., 2006)**
12. Derive an expression for vertical displacement when electron moves perpendicular to the uniform electric field. **(Amaravati Univ., 2008)**
13. Discuss the motion of an electron projected into uniform electric field at an acute angle with the field direction. **(R.T.M.N.U., 2007)**
14. Show that an electron moves along a parabolic path, when it enters in a uniform electric field applied perpendicular to its motion. **(R.T.M.N.U., 2006)**
15. Show that charged particle does not gain additional kinetic energy when it travels in uniform transverse magnetic field.
16. Prove that kinetic energy of a charge remains constant during its motion in magnetic field. **(Amaravati Univ., 2003, 2005)**
17. Show that the radius of orbit of a charged particle moving at right angles to magnetic field is proportional to its momentum. **(Amaravati Univ., 2004), (R.T.M.N.U., 2006)**
18. Derive an expression for the radius and time period for an electron in a transverse magnetic field of uniform intensity. **(R.T.M.N.U., 2006)**
19. State the shapes of trajectories of a charged particle of velocity v moving in an electric field E when
 - (a) $v \perp E$
 - (b) v makes an angle with E
 What are these shapes if E is replaced by a magnetic field B ?
20. Discuss the trajectory of a charged particle moving with uniform velocity traversing uniform magnetic induction B if:
 - (a) v_0 is parallel to B
 - (b) v_0 is perpendicular to B , and
 - (c) v_0 makes an acute angle with B . **(C.S.V.T.U., 2008)**
21. State conditions under which a charged article moves in a straight line in
 - (a) An electric field E
 - (b) A magnetic field B and
 - (c) In a region having both E and B .
22. Consider a particle of mass m and charge q moving with velocity v . The particle enters a region where a perpendicular uniform magnetic field B acts. Show that in the region the kinetic energy of the particle is proportional to the square of the radius of its orbit.
23. A charged particle of charge q coulombs and mass m kg enters the field with velocity v m/s under the following conditions:
 - (a) velocity perpendicular to the magnetic field
 - (b) angle between the directions of velocity and magnetic field is at an acute angle.
 Indicate the trajectory of motion of charged particle and derive the concerned mathematical relationship in each case. **(R.T.M.N.U., 2005)**
24. A particle of mass “ m ” and charge “ q ” enters a region of uniform magnetic induction field “ B ” with a velocity ‘ v ’ making an angle θ with the direction of field. Discuss its motion and derive expressions for: (i) radius of path (ii) pitch of helical path and (iii) frequency of revolution. **(C.S.V.T.U., 2005)**
25. How is it possible for a charged particle to pass through a combination of electric and magnetic field with out any deviation? What will be its velocity? What will happen if the particle is moving slowly or faster than its velocity? **(R.T.M.N.U., 2007)**
26. Obtain expressions for radius and pitch of helical path described by a charged particle when it enters a uniform magnetic field making an acute angle θ with the direction of the magnetic field.
27. Derive expression for deflection of electron beam in transverse magnetic field. **(Amaravati Univ., 2003, 2005)**
28. Describe Thomson’s method for determination of e/m of the electron. Derive necessary formula. **(Amaravati Univ., 2005), (R.T.M.N.U., 2006)**

29. Discuss the measurements of e/m for electron by Thomson's method. (Amaravati Univ., 2003, 2008)
30. Explain the method by which the specific charge of an electron can be determined. Is it essential in this method that the electrons have a constant speed?
31. Give the properties of positive rays. (Amaravati Univ., 2002)
32. Explain the method for analysis of the parabolic traces for positive rays. (Amaravati Univ., 2002)
33. Explain why:
- (a) No parabola passes through origin
 - (b) Weak parabolic traces are observed on the left side of the plate
 - (c) Thomson's parabola method for q/M . (Amaravati Univ., 2002)
34. What is mass spectrograph? How will you determine the isotopes of Neon using parabola method? (Amaravati Univ., 2008)

PROBLEMS

1. An electron is liberated at rest from one of two large parallel plates separated by 2 cm. The other plate has a relative potential of 2.4 kV. How long does the electron take to reach it? [Ans: 6.9×10^{-10} s]
2. A proton, a deuteron and an alpha particle with the same kinetic energies enter a region of uniform magnetic field moving at right angles to B. Compare the radii of their circular paths. [Ans: $R_p : R_d : R_\alpha = 1 : \sqrt{2} : 1$]
3. An electron describes a helix of radius 6 cm and pitch of 2 cm in a magnetic field of intensity 30 Gauss. Obtain the component of its velocity along and perpendicular to the magnetic field. [Ans: $v_{\perp} = 3.164 \times 10^7$ m/s, $v_{\parallel} = 1.68 \times 10^6$ m/s]
4. A proton traveling at 23° with respect to magnetic field of strength 2.63 mT experiences a magnetic force of 6.48×10^{-17} N. Calculate (i) the speed and (ii) the kinetic energy in eV of the proton. [Ans: $v = 3.93 \times 10^5$ m/s, K.E. = 0.805 keV]
5. A positive ion beam moving along X-axis enters a region of uniform electric field of 3kV/m along Y-axis and magnetic field of 1 Kilogauss along Z-axis acting simultaneously. Calculate speed of those ions, which shall pass undeviated. What will happen if ions are moving slowly or faster than this velocity? [Ans: 3×10^4 m/s]
6. In an experiment for determination of e/m for electrons by Thompson's method, electric field between plates of 6 cm length and 2 cm separation a potential difference of 480 V produces a deflection of 10.9 cm on the screen 40.0 cm away. A magnetic field of 8.0×10^{-4} T restores the beam to its original position. Calculate e/m . [Ans: $e/m = 1.75 \times 10^{11}$ C/kg]
7. The electrons circulating in a uniform magnetic field of 4.55×10^{-4} wb/m² have a kinetic energy of 22.5 eV. What are the radius and period of revolution of the electron path? [Ans: $r = 3.5$ cm; $T = 7.9 \times 10^{-8}$ s]
8. The mass of an alpha particle is 7344 times that of electron and its charge is twice that of an electron. It is accelerated from rest through a potential difference of 5×10^6 volts. Calculate the flux density of magnetic field required to bend its path into a circle of radius 1 m if it enters the field
 - (a) Normally and
 - (b) At an angle of 45° with the field. [Ans: 0.45 wb/m²; 0.31 wb/m²]
9. A proton accelerates from rest in a uniform electric field of 400V/m. After 't' seconds, its speed is 3×10^6 m/s.
 - (a) Find the acceleration of a proton
 - (b) Find time 't'. (c) How much is the distance traveled in the time 't'? [Ans: 3.8×10^{10} m/s²; 7.8×10^{-5} s; 117m]
10. A proton and an α -particle both move in circular paths in a uniform magnetic field with same tangential speeds. Compare the number of revolutions they make per second.

CHAPTER

15

Electron Optics

15.1 INTRODUCTION

Electrons travel along straight-line path in uniform electric fields. However, in many of the practical cases the electric fields happen to be nonuniform. The analysis of electron motion in nonuniform fields is not easy to carry out and poses serious mathematical difficulties. Fortunately, there is a close resemblance between the motion of electrons in an electrostatic field and the propagation of light in an optical medium. Light rays travel along straight line path in homogeneous media and along curved paths in inhomogeneous media. Electrons travel along straight line path in an equipotential region and follow a curved path in passing through regions of varying potential. Light undergoes refraction at an optical boundary. In a similar fashion, electron path bends when electrons travel from a region of one potential into the region of another potential. The resemblance suggests that the concepts of geometrical optics can be applied to the motion of the electrons in homogeneous as well as inhomogeneous electric fields. Such an extension of the concepts of geometrical optics to electron motion is known as **electron optics**. This approach led to the discovery of TV tube, particle accelerators, electron microscope and perfecting of photomultiplier tube, microwave tube etc. Electron optics had its beginning in the works of H. Busch in 1926 and C.J. Davisson and C.J. Calbick in 1931.

The principles of electron lens and magnetic lens are introduced and the working of CRO and its applications are explained in this chapter.

15.2 BETHE'S LAW

Let us consider a narrow stream of electrons (or any charged particles) traveling in space. These electrons bend and follow curved path when they travel through regions of different electric field strengths. This behaviour is similar to the behaviour of light rays that change their direction of motion while passing through an optical medium of inhomogeneous refractive index. By analogy with optical refraction, the bending of electrons by electric fields is called **electron refraction**. Bethe's law deals with electron refraction while Snell's law describes refraction of light.

On account of nonuniform electric fields, the equipotential surfaces shall be curved. Let us consider two equipotential surfaces 1 and 2 separated by an infinitesimal region (see Fig.15.1). Let the potential on surface 1 be V_1 and the potential on surface 2 be V_2 . Let $V_2 > V_1$.

An electron starting from rest and traveling in the region above the surface 1 acquires a velocity v_1 when it reaches a point K on the surface V_1 . This velocity is related to the potential V_1 according to the relation

$$v_1 = \sqrt{\frac{2eV_1}{m}} \quad (15.1)$$

where 'e' is the electronic charge and 'm' the mass of the electron.

Let the angle between the path of the electron at K and the normal to the equipotential surface 1 be θ_1 . Then the velocity component of the electron perpendicular to the equipotential surface is $v_1 \cos \theta_1$ and this is also the electric field direction at K. The tangential velocity component is equal to $v_1 \sin \theta_1$. Electron velocity increases from v_1 to v_2 , as the electron travels from the surface 1 to surface 2. When the electron reaches the point L on the second equipotential surface, its speed is v_2 .

$$v_2 = \sqrt{\frac{2eV_2}{m}} \quad (15.2)$$

Note that the electron is moving in nonuniform electric field since V_2 is different from V_1 . Consequently, the direction of velocity shall continually change as the electron goes from K to L. If $V_2 > V_1$, the path of the electron shall be as shown in the Fig. 15.1. If the angle between the path of the electron and the normal to the equipotential surface 2 at the point L is θ_2 , then the velocity component of the electron perpendicular to the equipotential surface is $v_2 \cos \theta_2$ and is along the electric field direction at L. The tangential velocity component is equal to $v_2 \sin \theta_2$.

As the electric field is perpendicular to the equipotential surface, the electron does not experience any force tangential to the surface. However, the electron is acted upon by a force due to the electric field that is acting perpendicular to the equipotential surface. Since $V_2 > V_1$, $v_2 \cos \theta_2$ is greater than $v_1 \cos \theta_1$. Since the surfaces V_1 and V_2 are very close to each other they may be taken as plane, parallel surfaces. Hence, the components $v_1 \sin \theta_1$ and $v_2 \sin \theta_2$ are parallel to each other and equal to each other. This is due to the fact that as the electron goes from K to L, there is no force in the horizontal direction. Hence, we have

$$\begin{aligned} v_1 \sin \theta_1 &= v_2 \sin \theta_2 \\ \therefore \frac{\sin \theta_1}{\sin \theta_2} &= \frac{v_2}{v_1} \\ \therefore \frac{\sin \theta_1}{\sin \theta_2} &= \sqrt{\frac{2eV_2 / m}{2eV_1 / m}} \\ \text{or } \frac{\sin \theta_1}{\sin \theta_2} &= \left[\frac{V_2}{V_1} \right]^{1/2} \end{aligned} \quad (15.3)$$

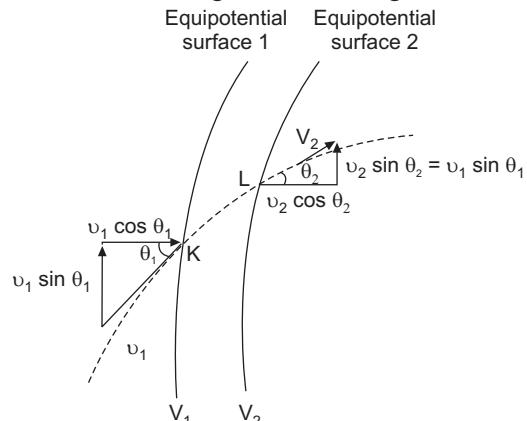


Fig. 15.1. Electron refraction through equipotential surfaces

The above equation is known as **Bethe's law**.

We may regard that the electron is effectively refracted at the equipotential surface lying midway between surfaces 1 and 2. Then this refraction can be seen to be similar to the optical refraction occurring at the boundary of two refracting media. In the present case, the electron refraction is similar to the optical refraction from rarer to the denser medium, where light bends towards the normal. The light refraction is governed by the Snell's law.

$$\frac{\sin i}{\sin r} = \frac{c_1}{c_2} = \frac{\mu_2}{\mu_1} \quad (15.4)$$

where c_1 and c_2 are the velocities of light in the optical media 1 and 2 respectively.

- In case of electron refraction, the rarer medium corresponds to the lower potential region and the denser medium corresponds to the higher potential region.
- When light enters a denser medium, it slows down and bends toward the normal to the surface. On the other hand, when an electron ray enters higher potential region, its velocity increases and bends toward the normal.
- The order of velocities v_1 and v_2 in Bethe's law is reversed from that in Snell's law.

Example 15.1: An electron beam enters from a region of potential 75 volts into a region of potential 100 volts, making an angle of 45° with the direction of electric field. Find the angle through which the beam refracts.

Solution:

$$\begin{aligned} \frac{\sin \theta_1}{\sin \theta_2} &= \left[\frac{V_2}{V_1} \right]^{1/2} \\ \therefore \frac{\sin 45^\circ}{\sin \theta_2} &= \left[\frac{100 V}{75 V} \right]^{1/2} \\ \therefore \sin \theta_2 &= \frac{0.7071}{1.1547} = 0.612 \\ \therefore \theta_2 &= 37.76^\circ. \end{aligned}$$

Example 15.2: Electrons accelerated under a potential of 250 V enter an electric field at an angle of incidence of 50° and gets refracted through an angle of 30° . Find the potential difference between the two equipotential surfaces.

Solution:

$$\begin{aligned} \frac{\sin \theta_1}{\sin \theta_2} &= \left[\frac{V_2}{V_1} \right]^{1/2} \\ \therefore \frac{\sin 50^\circ}{\sin 30^\circ} &= \left[\frac{250 V}{V_1} \right]^{1/2} \\ \text{or } V_1 &= \left[\frac{0.5}{0.766} \right]^2 \times 250 V = 106.5 \text{ V.} \end{aligned}$$

15.3 ELECTRON LENS

An **electron lens** is an electrical component that focuses an electron beam to a point.

Principle: A stream of electrons experiences a change in direction of motion when it passes through a non-uniform electric field. Its path is bent at each equipotential surface in the same way as a light ray is bent at an optical boundary. Hence, non-uniform electric fields can be used to focus a bundle of electron rays. In 1931 C.J.Davisson and C.J. Calbick achieved

such focusing action. Just as a convex lens focuses light rays, a non-uniform electric field produced by two coaxial metal tubes maintained at different potentials can focus electron rays and therefore, such a system is called an *electron lens*.

Construction: An electron lens is made of two coaxial short cylindrical metal tubes T_1 and T_2 separated by some distance, as shown in Fig.15.2. The tubes are held at different potentials V_1 and V_2 respectively. V_2 is greater than V_1 . A non-uniform electric field is produced in the gap between the two tubes, as shown in Fig.11.2.

Working: Let us consider a thin bundle of electron rays entering the lens system through the tube T_1 as shown in Fig.15.2. The electrons are moving from a lower potential region to a higher potential region. Fig.15.2 shows the equipotential surfaces in the region between the tubes.

Electrons, labeled 1, are moving along the axis of the system. When the electrons travel through the tube T_1 , they do not experience any force. As the electrons approach the gap between the cylindrical tubes, they come across first the convex shaped equipotential surfaces. The electric field is perpendicular at each point on the equipotential surface and is directed from tube T_2 to tube T_1 . At point A, the electric field is along the axis and hence the electrons labeled 1 get accelerated forward along the axis. They travel forward without any deviation from their initial path. Beyond the mid-plane MM' , the equipotential surfaces are concave shaped but the electric field still acts along the axis in a direction which accelerates the electrons labeled 1. Therefore, these electrons fly forward with increased speed along the axial direction. Electrons labeled 2, reaching at B on the convex shaped equipotential surface, experience an electric force that acts at an angle to the direction of their motion. The force can be resolved into its rectangular components F_{11} and F_{\perp} . Due to the force F_{\perp} , electrons labeled 2 are deflected down toward the axis. Due to the component F_{11} , they are also accelerated toward tube T_2 . Hence these electrons are continually bent downwards till the midplane MM' is reached. In the same way, the electrons, labeled 3 on reaching at C on the equipotential surface are continually deflected up toward the axis and are also accelerated forward. Because of cylindrical symmetry, all off-axis electron paths around the lens axis tend to converge toward a point on the axis, as shown in Fig.15.2.

On crossing the mid-plane MM' of the gap, the electron rays encounter equipotential surfaces of concave shape. In the second half of the gap, the normal component of electric force, F_{\perp} is directed away from the axis for all off – axis electron rays. As a result, the electron rays tend to diverge. The parallel component F_{11} is directed forward and hence the electrons are accelerated further. Before the electron rays can converge, they are diverged to some extent. However, because the potential is everywhere higher in the second half of the gap, the electrons travel faster in the second half than in first half of the gap. The electrons spend a greater time in the first half of the gap and the impulse ' $F_{11}t$ ' is greater for the convergence interval. In the second-half the electrons move faster because of higher potential and the impulse " $F_{\parallel} t$ " is smaller for the divergence interval. Consequently, the electron rays are less diverged by the second half than converged by the first half of the gap. Therefore,

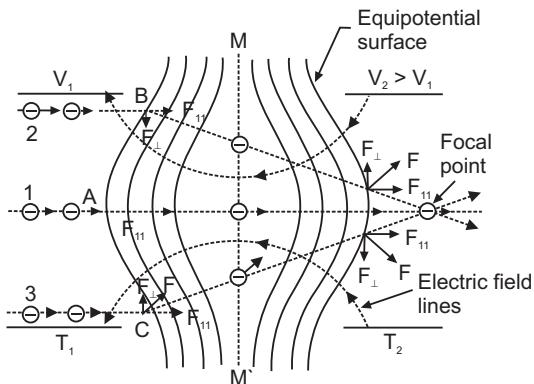


Fig. 15.2. Focusing of electron beam by an electron lens

the converging action of the first half of the gap will be stronger than diverging action of the second half of the gap and the electron rays emerge from the tube T_2 still sufficiently convergent, as shown in Fig. 15.3. By adjusting the potentials on the tubes appropriately, the electron beam can be focused to a suitable required point on the axis.

Comparison with a glass lens

- Light rays are bent only at the two surfaces of a glass lens but electron rays are bent continuously through successive equipotential surfaces.
- Secondly, the focal length of a glass lens is fixed whereas the focal length of an electron lens may be varied by varying the potentials V_1 and V_2 on the cylinders.

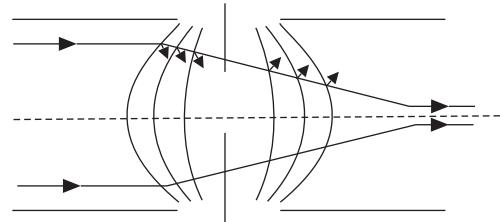


Fig. 15.3. Electron rays converging toward the lens axis

Application

Electron lens forms the most important component of an electron gun used for producing a narrow intense electron beam. Electron lens action is utilized in particle accelerators to focus charged particles into a narrow beam.

15.4 FOCUSING BY UNIFORM MAGNETIC FIELDS

An electron beam can be focused with the help of a magnetic field also. Depending upon the nature of the magnetic field, i.e., uniform or non-uniform, and the direction of its application, different types of focusing are possible.

15.4.1 Longitudinal Uniform Magnetic Field Focusing

This method employs uniform magnetic field acting along the direction of motion of the electron beam. The path of electron in a uniform magnetic field would be a helix if electrons enter the field at an angle. The pitch of the helical path is given by

$$p = \frac{2\pi m v}{eB} \cos \phi \quad (15.5)$$

Let us consider a bundle of electron rays of a given velocity 'v' originating at a point O with the rays making different angles with the uniform magnetic field as shown in Fig. 15.4. If the divergence in the beam is assumed to be small ($\phi \leq 10^\circ$), $\cos \phi$ can be taken as unity. Any electron ray leaving at point O again intersects the field lines after the time interval T . Therefore, the electron rays leaving at point O at small angles will again cross the same field line at a common point P after the time period T . The distance of P from O is l which is given by

$$l = \frac{2\pi m v}{eB} \quad (15.6)$$

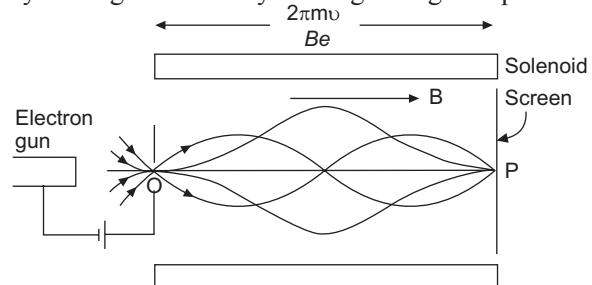


Fig. 15.4. Focusing of electron beam in a longitudinal uniform magnetic field

Thus a longitudinal homogeneous magnetic field focuses electrons at a distance l or at any distance which is an integral multiple of l (i.e. $n l$).

Application

This method is used for determining the e/m ratio of electron.

15.4.2 Transverse Uniform Magnetic Field Focusing

A transverse uniform magnetic field would also focus a beam of charged particles. The focusing would be affected after rotating the beam through 180° . The method of *semicircular focusing* was first employed by Rutherford in 1914. The principle of the method is shown in Fig.15.5. If a beam of positively charged particles with the same velocity enter uniform magnetic field acting in a transverse direction to its motion, and if the angular divergence of the beam is very small, it gets focused after completing a semicircle. This type of focusing effect is widely used in mass spectroscopy. To separate and identify the isotopes of an element, its atoms are ionized and allowed through a slit into the region of a uniform magnetic field acting transversely as shown in Fig.15.5. The ions belonging to different isotopes would have different masses. Therefore, even though all of them have the same velocity, they describe semicircles of different radii and come to focus at different points. The radii of the circles depend on the momenta of the different isotopes. As the separation of the isotopes is based on momentum selection by the magnetic field, such an arrangement is also known as a **momentum selector**.

Application

A momentum selector is used in Bainbridge mass spectrograph for separating positive ions of different masses.

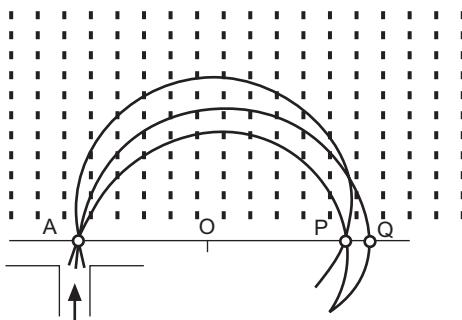


Fig. 15.5. Focusing of positively charged particles in a transverse uniform magnetic field

15.5 FOCUSING BY AXIALLY SYMMETRIC MAGNETIC FIELD

Magnetic fields that are axially symmetrical have a focusing effect on an electron beam passing through them.

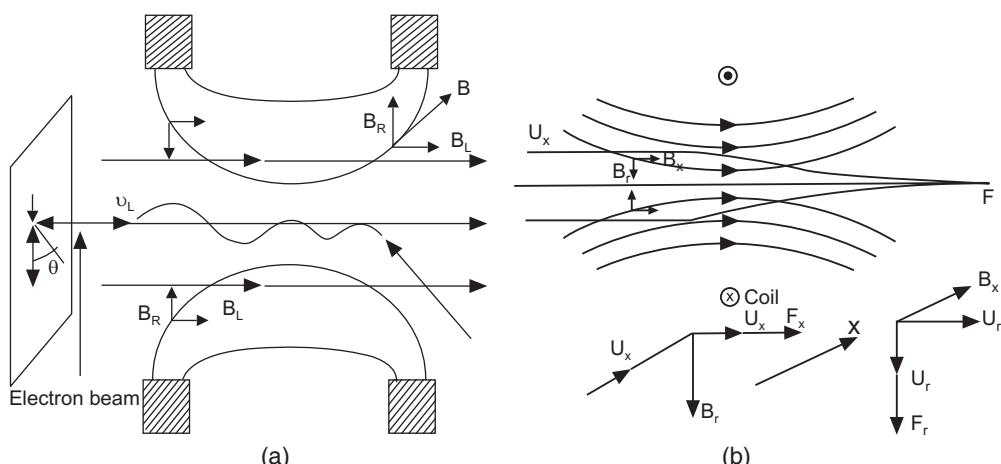


Fig. 15.6. Converging action of an axially symmetrical magnetic field

Let us consider a short coil carrying a current as shown in Fig.15.6. The magnetic field due to the current can be resolved into a radial component B_R , and an axial component B_L . B_R is directed towards the axis on the left-hand side; it becomes zero in the plane of the coil; and it is directed away from the axis on the right-hand side. An electron moving parallel to the axis with the velocity v_L will be unaffected by the magnetic field B_L , and experiences equal radial forces due to the components B_R along its path. As a result, the electron travels along the axial direction without any deviation. On the other hand, electrons traveling along the non-axial lines, experience unequal forces and are deflected towards the axis of the lens. Therefore, the beam converges towards the axis intersecting it at F. The effect of the coil is thus analogous to the effect of a convex lens on a beam of light. With the adjustment of current through the coil and the initial velocity of the electrons, the focal distance of the magnetic lens can be altered.

15.5.1 Magnetic Lens

Magnetic fields, which are axially symmetric, have a focusing effect on an electron beam passing through them. The axially symmetrical magnetic fields are produced by short solenoids. By encasing the coils in hollow iron shields the magnetic fields are concentrated and improved focusing action is obtained. Such solenoids are called *thin magnetic lenses*. Fig.15.7 shows the schematic of a thin magnetic lens.

AB is the section of the coil of an electromagnet and the hatched portion represents a soft iron shield. P and Q represent the gaps at diametrically opposite positions. A strong non-uniform magnetic field is produced across the gap and it will be symmetrical about the axis OI of the coil. O is the source of electrons, which are focused at the point I due to the magnetic field. The focal length of magnetic lens is very small of the order of a few centimeters.

A magnetic lens can only converge an electron beam. Diverging action is impossible in magnetic lenses. With the adjustment of current through the solenoid and the initial accelerating voltage of the electron, the focal distance of the magnetic lens can be adjusted.

Application

Magnetic lenses are widely employed in electron microscope and in electron beam machining equipments.

15.6 CATHODE RAY TUBE

A *cathode ray tube* (CRT) is a specially designed vacuum tube (Fig.15.8). In the CRT, an electron beam controlled by an electric field, generates a visual display of input signal on a fluorescent screen. It consists of the following three principal parts:

- (i) electron gun,
- (ii) deflection system and
- (iii) a fluorescent screen.

(i) Electron gun: The electron gun consists of several electrodes (see Fig.15.9) and is mounted at one end of the cathode ray tube. An indirectly heated cathode K emits a stream of electrons from its front face. The electrons pass through the control grid G held

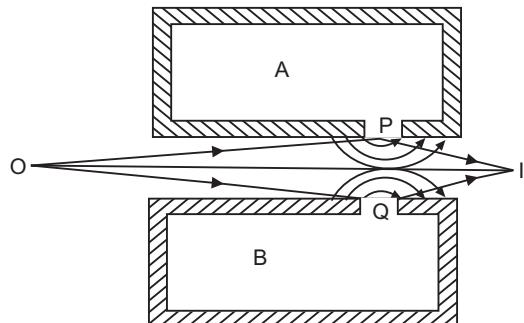


Fig. 15.7. Schematic diagram of a magnetic lens

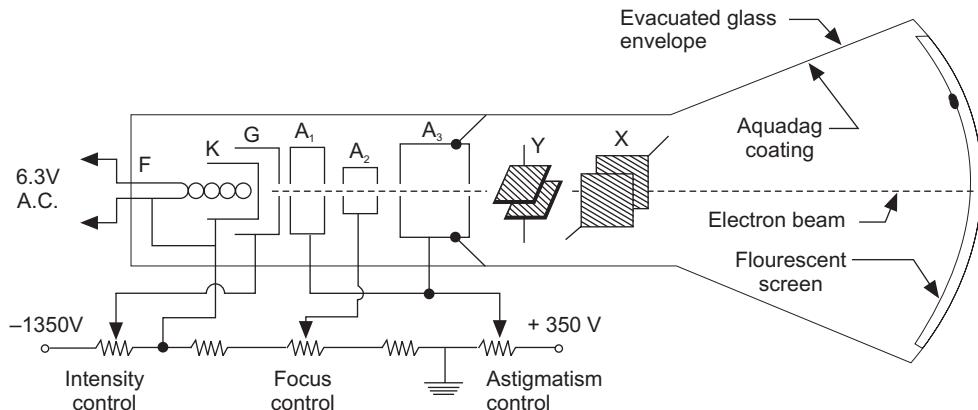


Fig. 15.8. Schematic of Cathode ray tube (electrostatic deflection type)

at negative potential. The effective size of the aperture in the grid varies depending upon the potential difference between the grid and cathode. The number of electrons striking the screen determines the **intensity** of the glow produced at the screen. Therefore, by varying the negative d.c. voltage on the grid, the intensity of the luminous spot on the screen may be controlled. Two anodes A_1 and A_2 are positioned beyond the control grid, which are coaxial with the grid. The electron lens system formed by the combination of the grid and anodes focus the electron rays into a sharp narrow pencil of beam. The electron beam travels along the axis of the tube and strikes at the centre of a fluorescent coating located at the opposite end of the CRT. The electron beam may be brought to a sharp focus on the coated face by varying the relative magnitudes of voltages applied to the anodes.

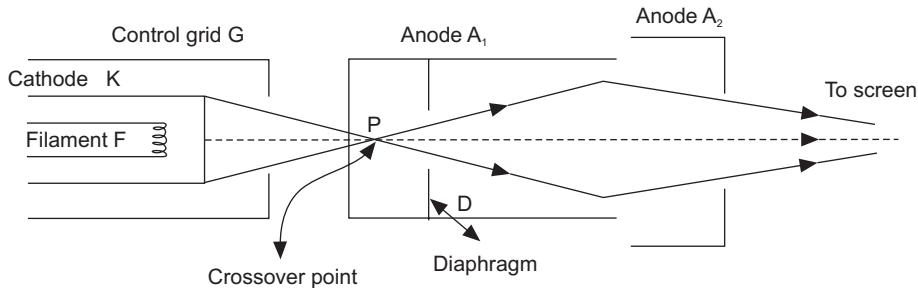


Fig. 15.9. Schematic of an electron gun

(ii) Deflection System: In an electrostatic deflection type CRT, the deflection system consists of two pairs of parallel metal plates Y and X mounted in the neck of the tube. The metal plates are positioned symmetrically around the electron beam path. The electron beam emitted by the electron gun travels towards the fluorescent coating, and on the way pass through the deflection system. The two plates in each pair are strictly held parallel to each other. One pair of plates is arranged horizontally and the other pair is arranged vertically. When a potential difference is applied to the plates, uniform electric field is produced in the region between the plates. The set of *horizontal plates* produces a uniform electric field in the *vertical direction*. When the electron beam passes through the field, it gets deflected upward or downward in the vertical direction, as shown in Fig.15.10 (a). Therefore, this set of plates is called **vertical deflection plates** or **Y-plates**. The second set of plates is oriented *vertically* and, when a potential difference is applied between the plates, it produces uniform electric field in *horizontal direction*. The electron beam passing through the field gets deflected horizontally

from left to right, as shown in Fig.15.10 (b). Therefore, this set of plates is called horizontal deflection plates or **X-plates**. In either case, the amount of deflection of the electron beam is directly proportional to the magnitude of the voltage applied to the plates.

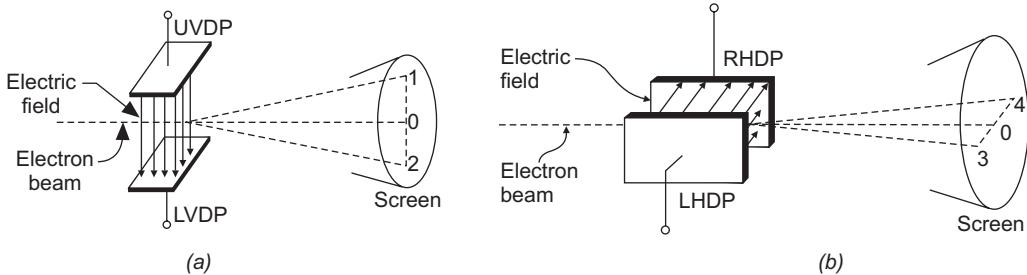
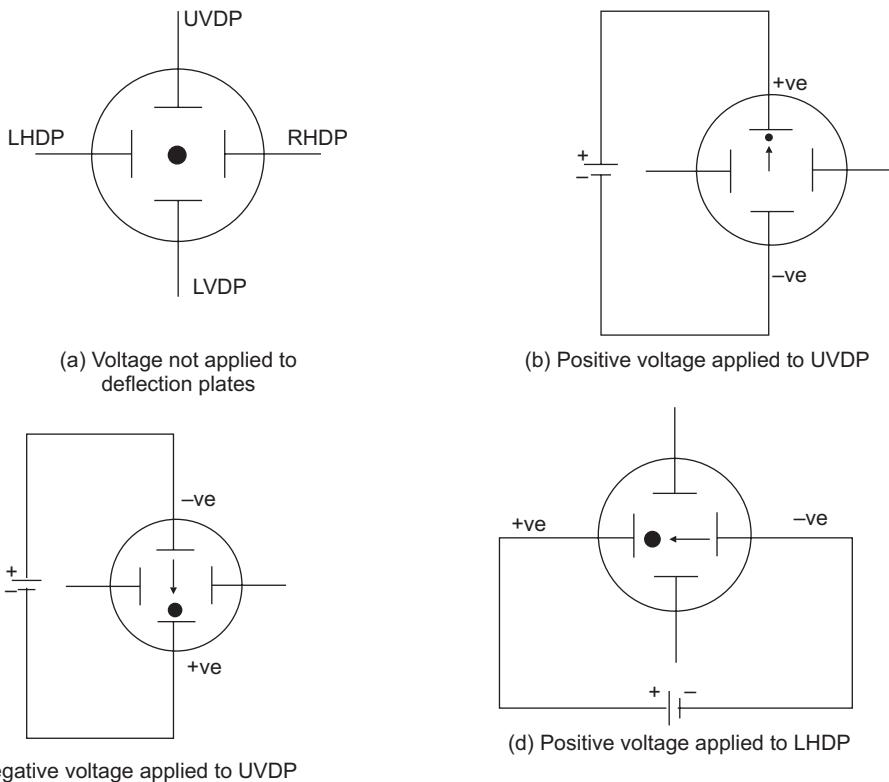
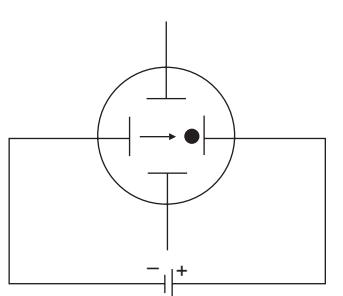


Fig. 15.10. Electron beam deflection (a) Vertical deflection (b) Horizontal deflection

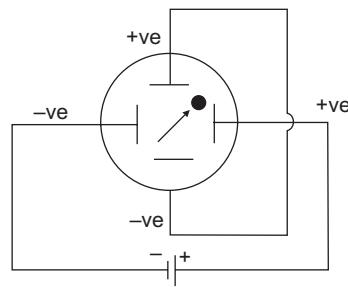
The direction of the deflection depends on the polarity of the applied voltage, as shown in Fig.15.11. When dc voltages are applied to both the X- and Y-plates, the electron beam will be subjected to two orthogonal forces and gets deflected along the direction of their resultant, as shown in Fig.15.11. By varying the voltages on the vertical and horizontal deflection plates, the luminous spot may be moved to any position on the screen.

(iii) Fluorescent Screen: Electrons are invisible particles. However the electron beam position can be located with the help of a fluorescent screen. The interior surface of circular front face of the CRT is coated with a thin translucent layer of phosphor. The front face acts as a fluorescent screen. Phosphors have the property of emitting light when high-energy particles





(e) Negative voltage applied to LHDP



(f) Positive voltage simultaneously applied to UVDP and RHDP

Fig. 15.11. Electron beam deflection under the action of dc voltages given to Y and X-plates

or high-energy radiation hits them. Therefore, when the high-energy narrow electron beam strikes the fluorescent screen at a point, it emits a glow at that point. The glow can be seen through the thin glass face and makes the position of the electron beam known.

15.6.1 Acquadag Coating

In a CRT, electrons striking the fluorescent screen emit secondary electrons, which tend to charge the screen negatively. Consequently, the screen repels the electrons that arrive afterwards. This leads to a decrease in the brightness of the trace. On the other hand, the cathode loses electrons through emission and attains positive polarity. Therefore, it drags the emitted electrons back, and this also leads to a decrease in the brightness of the trace. This problem can be eliminated if the electrons accumulating on the screen are conducted away and the shortage of electrons at the cathode is made up. This is accomplished by the acquadag coating.

The inner surface of the flare of the glass envelope of CRT is coated with conductive graphite coating called *acquadag*. It is connected internally to the positive anode. The acquadag coating, which is at a positive potential, attracts these secondary electrons and returns them to the cathode via the ground. Thus, the acquadag coating completes the electrical circuit from screen to cathode and keeps them in electrically neutral condition. The acquadag coating also serves as an electrostatic shield and shields the electron beam from the influence of external electric fields.

Application: CRTs are widely used in many display applications such as CRO, radar, and video terminals.

15.6.2 Limitation of Electrostatic Deflection

The angle of deflection of the electron beam in the electrostatic deflection type CRT is given by

$$\theta = \tan^{-1} \left[\frac{IV}{2dV_A} \right]$$

Theoretically, a large angular deflection of the beam can be attained by selecting longer deflection plates (larger l) and by keeping them closer (smaller d). However, in practice, the angle θ is to be restricted to smaller values. Otherwise, at angles greater than a certain value, the electrons strike the deflection plates instead of hitting the screen. A slightly larger beam deflection is achieved by flaring the deflection plates at their exit ends. Because of the restricted angular deflection, the area that can be covered by the electron beam on the screen becomes smaller. It is thus not possible to obtain bigger displays using electrostatic deflection.

For a given deflection angle θ larger area of the screen can be covered if the screen is kept at a farther distance (large L) from deflection plates. However, it is impracticable to increase L beyond a certain limit. Optimum result is achieved for a fixed L by positioning the Y -plates before X -plates. The information to be displayed by CRT is applied conventionally to Y -plates and therefore by keeping Y -plates next to the electron gun and before X -plates, the Y -plates to screen distance is increased. Such an arrangement allows a larger area of the screen to be covered by the electron beam.

The electrostatic deflection type CRT is chiefly used in Cathode Ray Oscilloscopes (CROs) where larger display is not the primary requirement. On the other hand, the CRT screen is required to be as big as possible in TVs, in order to obtain larger pictures such that the viewers get viewing satisfaction.

15.7 ELECTROMAGNETIC DEFLECTION TYPE CRT

The CRT with electromagnetic deflection is very much similar in construction to a CRT with electrostatic deflection except for the deflection system. The magnetically deflected CRT does not contain deflection plates. The electron beam deflection is caused by an external magnetic field produced by electromagnets. Two sets of coils are placed at right angles to each other and mounted on the tube neck where the electron beam leaves the electron gun. There are four coils in all, with opposite ones comprising one set. The coils in each set are connected in series such that magnetic north and south poles appear when a dc current flows through them.

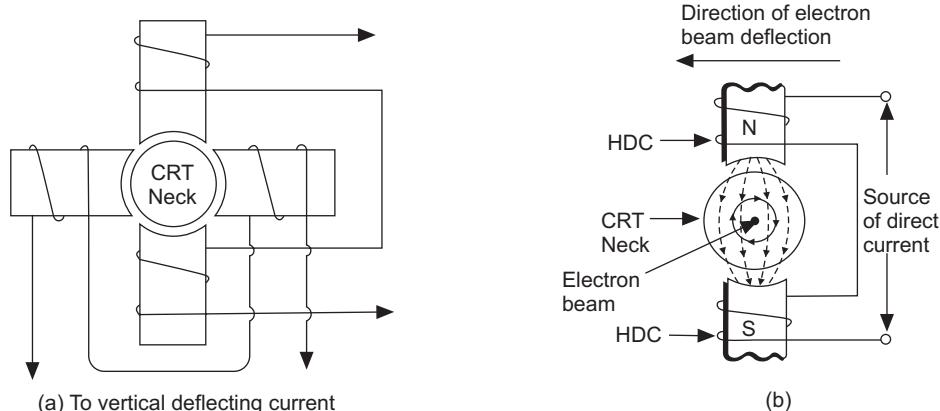


Fig. 15.12. Deflection of luminous spot under the action of magnetic field:

- Arrangement of deflection coils around CRT neck
- Deflection of electron beam in horizontal direction due to X-coils.

Fig.15.12 shows the physical placement of the coils. The horizontally mounted coils produce horizontal magnetic fields when direct current flows through them. The magnetic field acts at right angles to the electron beam. The electrons experience a Lorentz force as they pass through the magnetic field and are deflected vertically, in a direction normal to both the field direction and direction of motion. If the direction of the dc current through the coils is reversed, the magnetic field direction is reversed and the electron beam is deflected in the opposite direction. Thus, the horizontally placed coils, cause vertical deflection and are called **vertical deflection coils** or **Y-coils**. The coils mounted in vertical direction produce a vertically acting magnetic field which deflects the electron beam horizontally. These coils are therefore called **horizontal deflection coils** or **X-coils**. Both the vertical deflection coils and horizontal deflection coils are enclosed in a common former, called the **yoke**, which fits over the CRT neck. The amount of beam deflection is proportional to the strength of the

magnetic field which in turn is proportional to the current in the deflection coils. By properly controlling the amount of current flowing through each set of coils, the electron beam can be made to strike any desired point on the screen.

Advantages

1. As the deflection system is mounted outside, the fabrication of CRT becomes relatively simpler.
2. The deflection system does not obstruct the deflected electron beam and as such larger angles of beam deflection can be realized leading to bigger display on the screen.

Electromagnetic deflection type CRTs are used in TVs, Radar receivers, VDU applications etc.

15.8 CATHODE RAY OSCILLOSCOPE

A cathode ray oscilloscope (CRO) is a very important electronic measuring instrument. It is used to measure parameters such as voltages and time intervals. The most advantageous feature of CRO is that it gives a visual display of waveform under test.

15.8.1 Block Diagram of CRO

The block diagram of a cathode ray oscilloscope is shown in Fig.15.13.

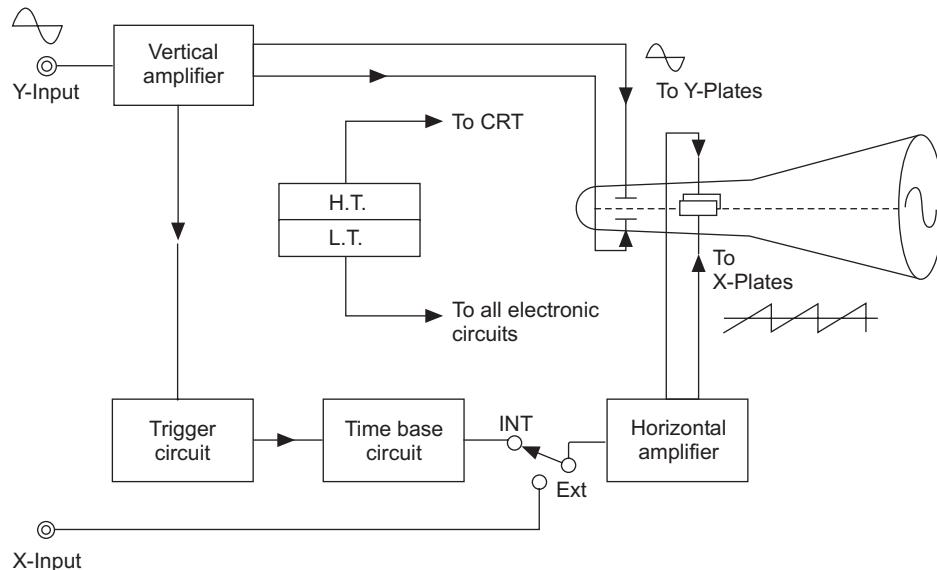


Fig. 15.13. Block diagram of a cathode ray oscilloscope

Any CRO contains the following seven basic sections:

- (i) Cathode ray tube (CRT)
- (ii) Time base generator
- (iii) Trigger circuits
- (iv) Vertical circuits
- (v) Horizontal circuits
- (vi) Low voltage power supply and
- (vii) High voltage power supply

15.8.2 Cathode Ray Tube

The CRT constitutes the central part of CRO. With the help of the other sections in the CRO, the CRT produces a trace on its fluorescent screen and maintains it. An electron gun, located at one end of the CRT, produces a sharp electron beam. A small luminous spot caused on the fluorescent coating indicates the position of the electron beam. The brightness and sharpness of the luminous spot on the screen are adjusted with the help of intensity and focus controls. The vertical and horizontal controls are used to adjust the position of the luminous spot at any desired position on the screen. When voltages are not applied to Y-plates and X-plates, the luminous spot stays at the centre of the screen. When a voltage is applied to the Y-plates, an electric field is produced in the vertical direction and the electron beam passing through the field is deflected upward or downward. When a voltage is given to the X-plates, an electric field is produced horizontally and electron beam passing through the electric field is deflected sideways either to the left or right of the screen. The Y- and X-plates are electrically independent and different voltages can be applied to each set. A transparent graph, called a **graticule**, marked in centimetre lines both vertically and horizontally, is attached to the face of the CRT for making measurements.

Signal:

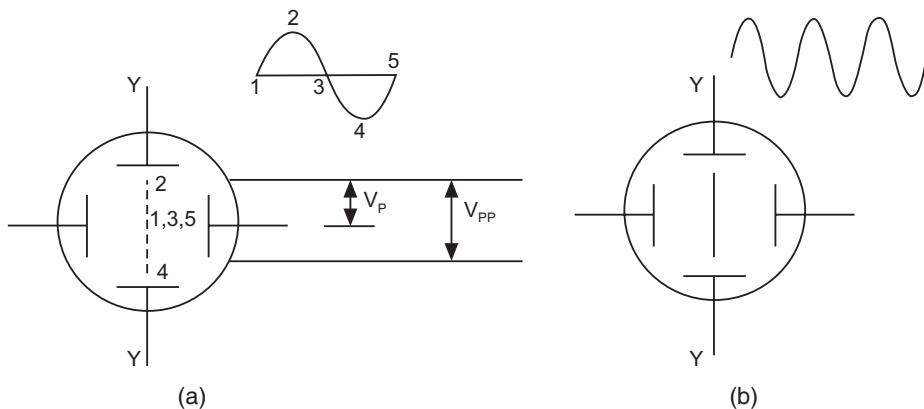


Fig. 15.14. A vertical trace is observed when ac voltage is applied to Y-plates

When a signal is to be displayed, it is applied to the Y-plates of a CRO by connecting it to the Y-input of CRO. For example, let the signal be a simple harmonic wave. Because of the application of the signal to the Y-plates, the luminous spot moves up and down on the screen at the same frequency as that of the applied voltage provided X-plates are not given any voltage. The successive positions of the luminous spot can be seen (Fig. 15.14a) when the frequency of the signal is less than about 20 Hz. At higher frequencies the path of the beam is seen as a vertical luminous line (Fig. 15.14b). This is due to persistence of vision and as well as the phosphorescence of the coating. The luminous line is called a **trace**. The length of the vertical trace corresponds to the peak-to-peak value of the applied voltage, as shown in Fig. 15.14 (a). In the same way a horizontal trace is produced on the screen, when an AC voltage is applied to X-plates.

15.8.3 Time Base

In a signal the voltage varies as a function of time in a specific way. It is obvious that the shape of the signal cannot be observed only through the vertical motion of the beam. It can be seen if the beam is made to move simultaneously in a horizontal direction from left to right. The signal variations can be faithfully displayed on the CRT screen when the electron beam

is deflected horizontally through equal distances in equal intervals of time. Secondly, at the end of the horizontal motion, the electron beam should return to the starting point in order to repeat the motion, just as after completion of writing a line on a paper the pen is brought back to the left side to commence the next line. It requires that the voltage applied to horizontal plates increases uniformly with time. A **ramp voltage** given to the X-plates causes uniform motion of the electron beam across the screen and also causes the electron beam to return to the starting point. The shape of the ramp voltage is shown in Fig. 15.15.

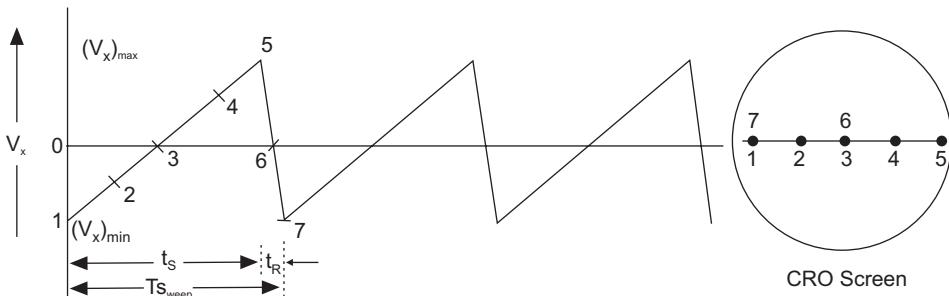


Fig. 15.15. A saw-tooth voltage applied to X-plates produces a horizontal trace on the screen.

The ramp voltage is also known as **sweep voltage** or **sawtooth voltage** or **time base**. It is generated by an oscillator, known as time base generator. *The time base generator is a variable frequency oscillator, which produces an output voltage of sawtooth shape.* Ideally, the voltage increases uniformly with time from a minimum to a maximum value and from there it abruptly falls to minimum. The process is repeated again and again. When the ramp voltage is applied to X-plates of CRT, the electron beam is deflected horizontally from the left edge to the right edge of the screen through equal distances in equal intervals of time. Therefore, the luminous spot will traverse on the face of the screen at a uniform velocity. As the voltage attains a maximum and drops instantaneously to the minimum value, the spot is whipped back across the screen to the starting point at the left edge. The cycle is repeated. The sweep time may be typically from 10 ns to 5s/division.

Let us assume that at $t = 0$, the ramp voltage is at its negative maximum $(V_x)_{\min}$ on the RHD (Fig. 15.15). It sets the beam at the left extreme on the screen. With time the voltage rises uniformly and the beam moves towards the right side on the screen. When ramp voltage reaches zero value, the beam will come to the centre and as it rises further through positive values the beam continues the motion towards right end of the screen. At $(V_x)_{\max}$ the beam will be at the right extreme. As $(V_x)_{\max} = (V_x)_{\min}$, the amounts of beam deflection to the left and right from the centre will be equal. V_x abruptly falls from $(V_x)_{\max}$ to $(V_x)_{\min}$ and the electron beam flies back to left extreme. Again the above cycle repeats. Thus, the luminous spot sweeps from left to right along a straight line, in step with the cycles of ramp voltage. Because of this reason, the ramp voltage is often called the **sweep voltage**. It is also called the **sawtooth voltage** for obvious reason of its shape.

It is easy to see that the length of the trace on the screen is a measure of the period of the oscillator frequency in seconds and each point along the path will represent a proportionate time interval measured from the beginning of the trace. Hence, the x-axis of the screen not only denotes the amount of horizontal deflection but also the time elapsed. Due to this reason, the ramp voltage is also known as the **time base**. In fact, the horizontal axis of the screen is calibrated in milliseconds and microseconds.

Different sweep times are obtained by varying the frequency of the time base generator.

Referring to Fig. 15.15, the time taken by the sweep voltage to rise from its maximum negative voltage to its maximum positive voltage is known as **sweep time** or **trace time** (t_s). The time taken by the sweep voltage to dip from its positive maximum to negative maximum value is called **fly back time** or **retrace time** (t_r). The sum of sweep time and retrace time constitutes the **sweep period**, T_{sweep} . Thus,

$$T_{\text{sweep}} = t_s + t_r \equiv t_s \quad (15.7)$$

Display of Signal Shape

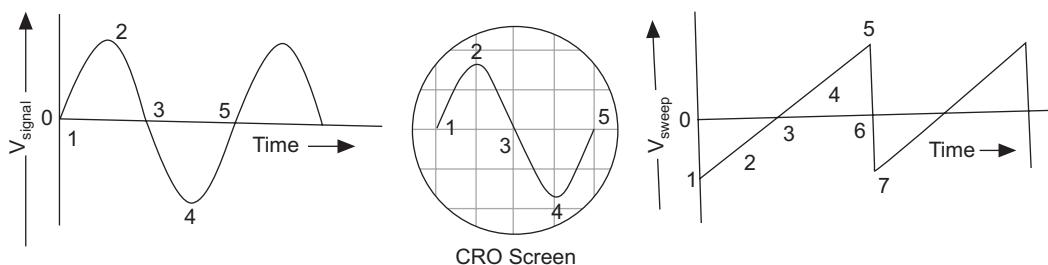


Fig. 15.16. Display of Signal Shape

For displaying the signal on the CRO screen it is necessary that the signal is fed to Y-input and the time base voltage is applied to X-plates. The electron beam is then simultaneously subjected to two forces: one acting in the vertical direction and the second in the horizontal direction. The deflection of the beam at any instant is determined by the resultant of these two forces. Referring to Fig. 15.16, it is seen that at the instant 1 the input signal is zero and the sweep voltage is $(V_x)_{\min}$, the resultant of the forces due to them acts along the left direction and the beam is deflected to the left extreme. At the instant 2, the signal amplitude is positive and the sweep voltage is at a lesser negative value. The beam is deflected left upward in the second quadrant of the screen. At the instant 3, both the input and sweep voltages are zero. The resultant force is zero and the beam stays at the center of the screen. At the instant 4, the signal amplitude is negative and the sweep voltage is positive. The beam is deflected to right down into the fourth quadrant of the screen. At the instant 5, the signal voltage is zero and the sweep voltage is $(V_x)_{\max}$. The electron beam is deflected toward the right extreme along the horizontal direction. Then, the beam returns to position 1 and the process repeats. By joining the resultant positions of the spot, it is seen that the waveform of the input voltage is faithfully displayed.

Blanking

If the backward trace of the electron beam is displayed on the screen along with the forward trace, it gives a bad visual effect. In principle, by making the retrace time (time for the backward trace) equal to zero, the retrace path can be eliminated. But in practice, the retrace time cannot be reduced to zero; it can be only curtailed to a minimum value. To prevent the electron beam trace during retrace period, a high negative voltage pulse is applied to the control grid in the electron gun so that the electron beam is switched off momentarily. This process is known as *blanking* of the trace.

15.8.4 Trigger Circuit

Synchronization

To display a stationary wave pattern on the CRO screen, the horizontal deflection should start at the same point of the input signal in each sweep cycle. When this happens we say that the horizontal sweep voltage is *synchronized* with the input signal. If the sweep and signal

voltages are not synchronized a stand still pattern is not displayed on the screen; the wave pattern moves continuously to the right or left of the screen. **Synchronization** is the method of locking the frequency of the time base generator to the frequency of the input signal so that a stationary display of wave pattern is seen on the CRO screen.

The signal is synchronized when its frequency equals the sweep frequency or an integral multiple of the sweep frequency. That is,

$$f_{\text{signal}} = n f_{\text{sweep}} \quad (15.7)$$

or $T_{\text{sweep}} = n T_{\text{signal}}$

As an example, if the sweep frequency is 50 Hz and the signal frequency is 50 Hz, one wave is displayed on the screen (Fig. 15.17a). On the other hand, if the sweep frequency is 50 Hz but the signal frequency is 100 Hz, the time period of sweep voltage is 20 ms and the time period of signal is 10 ms. In the sweep time which actually is horizontal trace length, the signal goes through two complete cycles. As a result two cycles of the signal voltage are displayed on the screen as seen in Fig. 15.17(b).

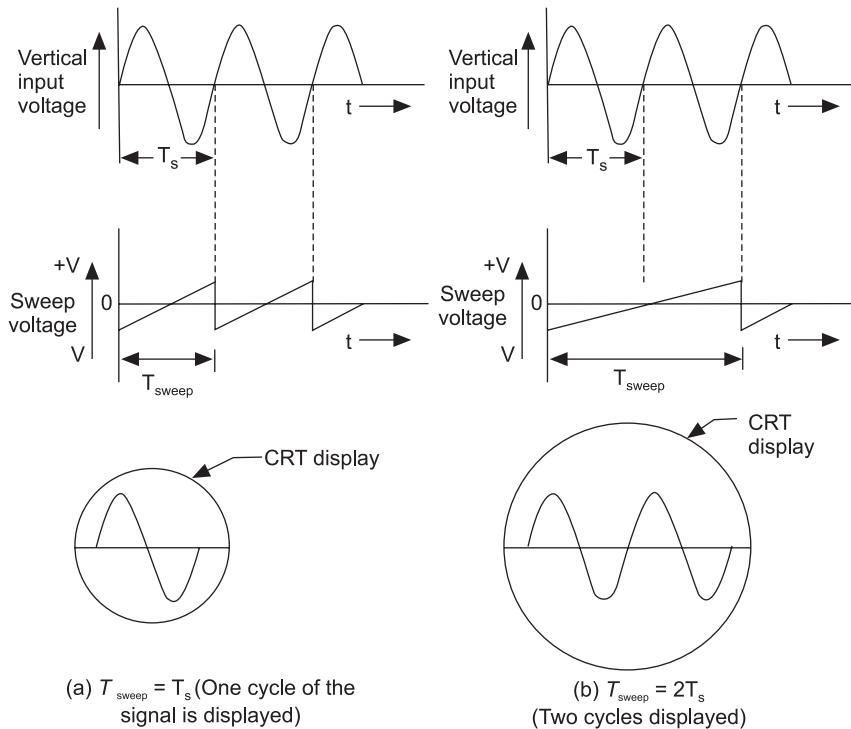


Fig. 15.17. Synchronization of signal and sweep voltages (a) $T_{\text{sweep}} = T_{\text{signal}}$, (b) $T_{\text{sweep}} = 2T_{\text{signal}}$

Achieving Synchronization

One of the methods of achieving synchronization is using a trigger circuit. In this method, a part of the output obtained from the vertical amplifier is fed to a trigger generator. Trigger generator is sensitive to the level of the voltage applied at its input. The circuit monitors the input signal and detects the point when it reaches a selected level while moving toward a selected polarity. When the predetermined level is reached, the circuit produces a trigger pulse. This trigger pulse is fed to the time base generator, and it acts as a command signal to the time base generator and starts one sweep cycle of the time base. The sweep voltage is not

developed in the trigger mode, if the input signal is not given. A portion of the trigger pulse is fed to a second circuit, which produces an un-blanking bias voltage to bring the grid of CRT to a potential, which allows electron beam to appear. Thus, a stationary display of the wave is seen only above a predetermined level of the input voltage. It happens in each cycle. Because the signal voltage is initiating the sweep cycle, both voltages will be synchronized. By proper adjustment of controls, the trigger pulse may be made to originate when the input signal is going positive or negative or at any particular voltage level.

However, in “auto” trigger mode, the trigger circuit will automatically provide a trigger pulse to the sweep generator even when the input signal is not applied to it and the horizontal trace is seen even without signal at Y-input.

15.8.5 Vertical Circuits

The vertical circuits mainly consist of an attenuator, and a voltage amplifier. The signal is applied at the Y-input. It goes to the input of the attenuator. The signal amplitude is increased or decreased by changing the amount of attenuation and then fed to the input of the voltage amplifier so that adequate deflection is obtained on the screen.

15.8.6 Horizontal Circuits

The sweep generator output cannot directly drive the horizontal plates. Therefore, it must be initially amplified. The horizontal circuits mainly consist of a voltage amplifier. When the sweep selector switch is in ‘INT’ position, the sweep voltage is applied to the horizontal amplifier. The output of the amplifier is fed to X-plates, and a linear trace is produced on the CRO screen. When the sweep selector switch is held in ‘EXT’ position, the horizontal amplifier input is disconnected from the internal sweep generator and is instead connected to the horizontal input jack. In this position, the electron beam remains stationary and produces a luminous spot at the centre of the CRO screen.

15.8.7 Low Voltage Power Supply

The low voltage power supply powers the electronic circuits such as the amplifiers, trigger generator, time base generator. It gives an output of the order of a few tens to a few hundreds of volts.

15.8.8 High Voltage Power Supply

The high voltage Power Supply Provides voltages to the electrodes in the electron gun assembly. It supplies voltages of the order of 1600 to 2200 volts.

15.8.9 Advantages

1. CRO is the only instrument that can give a visual display of the actual shape of the signal.
2. Electron beam, due to its negligible mass, can respond instantaneously to the variations of a signal, how rapidly they may occur. Therefore, CRO can be used to display and test signals of frequencies up to a few hundred mega hertz.

15.9 APPLICATIONS

A CRO is a universal measuring instrument capable of measuring the characteristics of a wide variety of electrical signals. The signals may be repetitive or occur only once and last a fraction of a microsecond. With the oscilloscope, we can determine the signal amplitude, time period (or duration), and also find out the phase relationship between two harmonic signals. We study here the application of CRO for measuring the frequency of a given wave and the phase difference between two harmonic waves.

15.9.1 Determination of Frequency

(i) **Time period measurement method:** The simpler method of determining the signal frequency is done by first measuring its period T and calculating the frequency using the relation

$$f = \frac{1}{T}$$

The time period of signal is measured as follows. The signal is fed to the vertical input of CRO. The scope vertical sensitivity, sweep speed and triggering controls are adjusted to obtain a stable waveform of one or two complete cycles covering a large area on the CRO screen. The number of horizontal divisions for one complete cycle is determined and the sweep speed is noted. The time period of the signal, T , is found by the formula

$$T = (\text{Number of horizontal divisions for 1 cycle}) \times (\text{sweep speed}) \quad (15.8)$$

Then the signal frequency is calculated.

$$f = \frac{1}{T}$$

This method of measuring frequency is accurate to approximately $\pm 3\%$, which is generally the sweep generator accuracy.

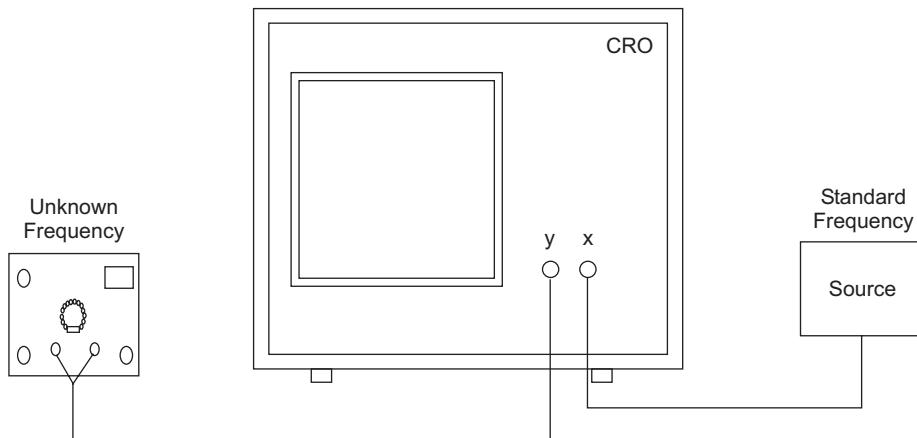


Fig. 15.18. Experimental set up for frequency measurement

(ii) **Lissajous Pattern method:** A more accurate method is to use the CRO as a frequency comparator (see Fig. 15.18). The signal of unknown frequency is applied to the vertical input and a voltage of known frequency is given to the horizontal input. By varying the frequency of the known source, a stable loop pattern known as Lissajous pattern is obtained on the screen. The number of points at which the loops touch the horizontal and vertical tangents is noted. If L_H and L_V are the number of points touching the horizontal and vertical tangents respectively, then the unknown frequency is calculated from

$$f_y = f_x \left[\frac{L_H}{L_V} \right] \quad (15.9)$$

where f_x is the known frequency. Example of measurement is shown in Fig. 15.19.

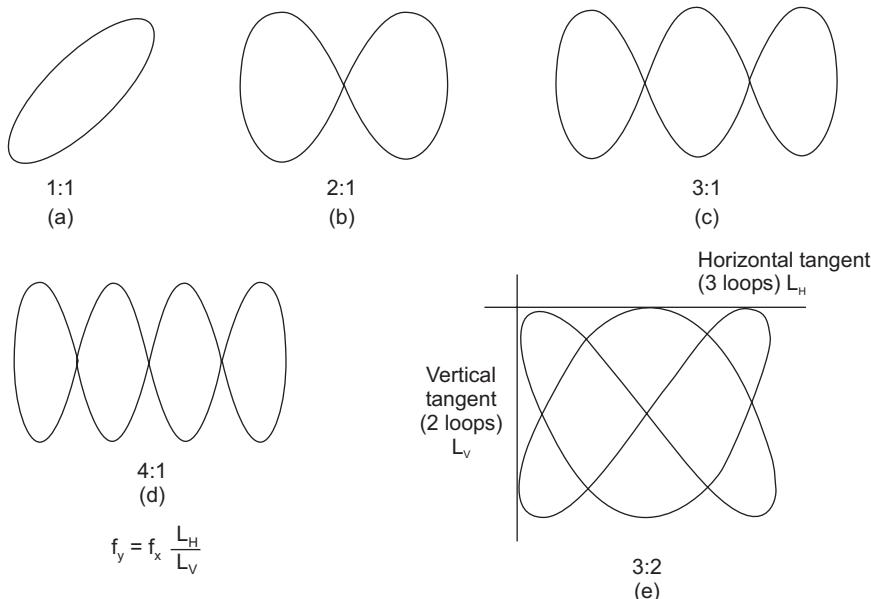


Fig. 15.19. Lissajous patterns and frequency measurement

15.9.2 Phase Measurements

(i) Dual sweep method: It requires a dual trace CRO. The phase relationship between two sinusoidal signals of same frequency may be directly measured by displaying both waveforms on the CRO screen and determining the delay time between the two waveforms. The sensitivity and trigger controls of each channel are adjusted for two stationary sinusoidal signals. The sweep speed is initially adjusted such that the time period T of the sine wave is measured. Then the sweep speed is increased and the delay time T_d between the two sine waves is accurately determined.

The phase difference is calculated using the relation

$$\phi = \frac{T_d}{T} \times 360^\circ \quad (15.10)$$

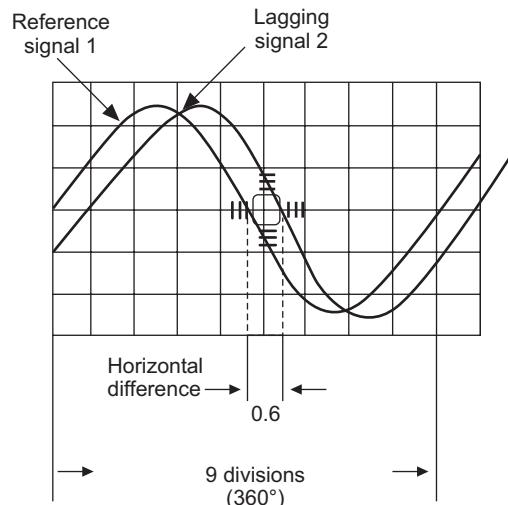
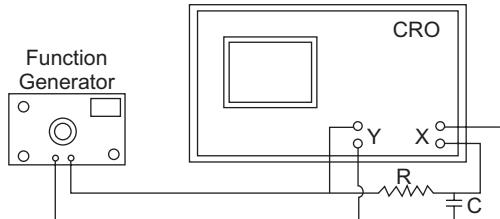
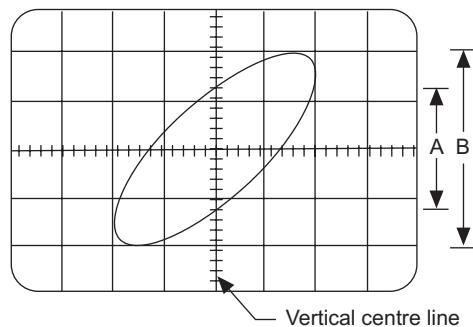


Fig. 15.20. Phase measurement using dual sweep method

(ii) Lissajous Pattern method: A second method for determining the phase difference of two sine waves of same frequency is to feed one sine wave to vertical input and the other sine wave to horizontal input (see Fig. 15.21). The sweep selector switch is kept in 'EXT' position.

A Lissajous pattern, namely ellipse (Fig. 15.22), is obtained on the screen. By measuring the lengths A and B of the elliptical pattern the phase shift ϕ is calculated.

$$\phi = \sin^{-1} \left(\frac{A}{B} \right) \quad (15.11)$$

**Fig. 15.21.** Experimental set up for phase measurement**Fig. 15.22.** Phase measurement

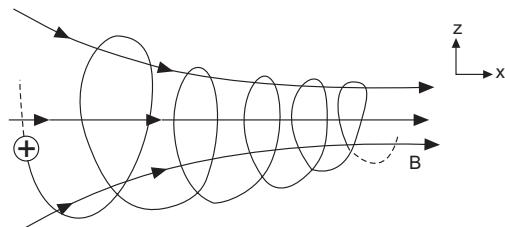
15.10 OTHER APPLICATIONS OF AN ELECTRON BEAM

A focused electron beam finds important use in machining processes such as cutting, welding, etc. In the electron beam machining (EBM) a high energy electron beam is produced with the help of accelerating voltages of the order of 50 to 200 KV on the anode. The beam is focused to a sharp point of a few microns diameter using magnetic focusing. When such a high energetic beam hits the work piece, it produces very high temperatures of the order of 6000° C. Therefore, the beam can cut the material quickly and easily. In electron beam welding (EBW) the heat generated can weld materials. Materials which are difficult to weld by other means can be successfully welded with the help of EBW. It is chiefly used to weld metals such as zirconium, beryllium, tungsten, etc. In these processes, the electron gun system as well as work piece are to be kept in a vacuum.

The electron beam is also used in preparing high purity materials used in atomic, rocket and electronics engineering.

15.11 MOTION OF CHARGED PARTICLES IN A NONUNIFORM MAGNETIC FIELD

It is learnt earlier that when a charged particle enters a uniform magnetic field at an angle with respect to the field lines it spirals around the field direction. If the uniform field were of infinite extent, the charge would spiral indefinitely and the cross-section of the spiral would be uniformly the same. On the other hand, if the field is nonuniform the motion will be very complex. Let us assume that the magnetic field increases along the X-direction as shown in Fig. 15.23. It is seen that the field is converging in the direction of the axis of the spiral. As the field is increasing in the X-direction the particle spirals in the X-direction with increasing cycling frequency. Also, the radius of the turns gradually decrease in the X-direction. Consequently, the particle spirals faster and faster in ever tightening loops. We have seen that a magnetic field cannot change the energy of the particle and hence it also cannot change the velocity of the particle. Therefore, with the increasing spiraling velocity v_z which is normal to the field B , the velocity parallel to the field direction v_x should decrease such that the sum remains constant. The particle velocity gradually decreases till it becomes zero. Then onwards it gets reversed. As such the particle now starts spiraling backwards. It can be viewed as being reflected by the strong converging

**Fig. 15.23.** Reflection of a charged particle in a nonuniform magnetic field.

magnetic field. The kinetic energy remains constant during the process of reflection. Therefore, the backward velocity of the particle will be equal to the velocity in forward direction for the same points in space. As v_z is increasing in magnitude, v_x decreases and the pitch of the spiral increases again. It amounts to state that the stronger field regions act as **mirrors** for charged particles. This effect is used to trap charged particles in a finite volume.

15.12 THE MAGNETIC BOTTLE

A **magnetic bottle** or a magnetic trap uses the above principle. It is a magnetic system with a weak field at its central region and increasingly stronger regions on both sides as illustrated in Fig. 15.24.

When a charged particle, say an electron enters the field at an angle with the field, it starts spiraling. In the region between A and B the field is diverging and the electron gets accelerated towards the right. When the particle passes into the region BC, the field lines are converging. Therefore, the electron experiences deceleration and gradually slows down and reverses its path to left. Thus the electron spirals back and forth between points M_1 and M_2 , which act as mirror points. Consequently, the electron gets confined within the region AC of the nonuniform magnetic field which thus acts as a **magnetic bottle**. A cylindrical tube fitted with magnetic fields coils at its two ends will serve as a **magnetic bottle**. The increased fields at the ends of the cylinder serve as magnetic mirrors.

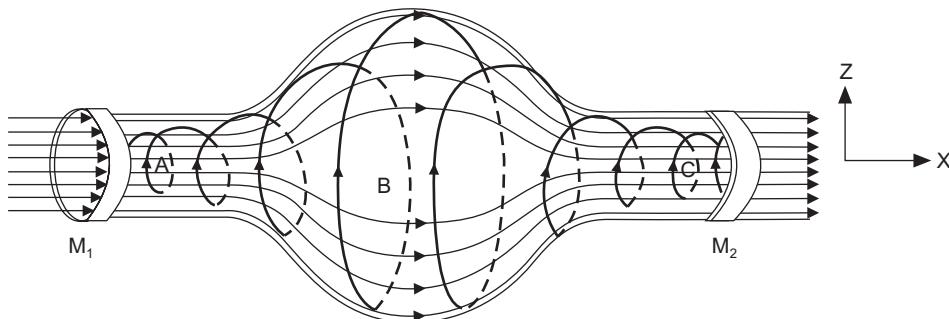


Fig. 15.24. The magnetic bottle

It was discovered in 1958 by James A. van Allen that the earth's magnetic field forms a huge magnetic bottle in space. As charged particles of cosmic origin rush toward the earth, they get captured by the nonuniform magnetic field and spiral back and forth between the two magnetic mirrors formed by the converging field near the north and south poles. The trapped particles form radiation belts around the earth and they are called **van Allen belts**.

The magnetic traps play a very important role in confining extremely hot plasmas. Plasmas are so hot ($T = 10^6$ K) that their contact with the walls of an ordinary container would melt the material of the container. Therefore, to confine plasma, magnetic bottles are used. Such a plasma confinement scheme plays a crucial role in achieving a controlled nuclear fusion process, which is almost an endless source of energy.

QUESTIONS

1. Discuss refraction of electron beam across an equipotential surface. Hence explain the working of electrostatic lens. **(R.T.M.N.U., 2007)**
2. What do you mean by a non-uniform electric field? Discuss Bethe's law. What is the refractive index of electrostatic field? **(C.S.V.T.U., 2006)**
3. Explain Bethe's law of electron refraction. Compare it with Snell's law in optics. **(C.S.V.T.U., 2005, 2008)**

4. State Bethe's law for electron refraction. (RGPV, 2007)
5. State Bethe's law. How is it analogous to Snell's law? (R.T.M.N.U., 2006)
6. What is Bethe's law? Show how this concept helps in understanding the focusing of electron beam by a symmetrical electron lens. (R.T.M.N.U., 2005)
7. Explain how an electron beam can be made to bend either towards or away from the normal to an equipotential surface. (Univ. of Pune, 2008)
8. Explain briefly the electrostatic focusing.
9. Discuss refraction of electron beam across an equipotential surface. Hence explain the working of the symmetrical and unsymmetrical electron lenses.
10. Explain working of an electrostatic electron lens.
11. Explain the working of a magnetic lens.
12. Derive an expression for electron deflection of electron beam in transverse magnetic field. (Amaravati Univ., 2005)
13. Draw a schematic of a cathode ray tube and show its principal parts. Briefly describe their functions.
14. What is the need of the aquadag coating inside the glass bulb in a CRT?
15. Draw a schematic of an electrostatic CRT. Describe the role of
 - (i) Electron gun, (ii) Deflection system
 - (iii) Fluorescent screen, and (iv) Aquadag coating
16. Derive an expression for electrostatic deflection sensitivity of a C.R.T. in terms of d , V_A , l and L where the terms have their usual meaning. (R.T.M.N.U., 2007)
17. Draw the block diagram of a CRO and write in brief the working of each block. (R.T.M.N.U., 2005)
18. Draw the block diagram of a cathode ray oscilloscope and explain the various parts. (Amaravati Univ., 2006)
19. Draw block diagram of CRO. How can the frequency of ac mains be determined using CRO? Explain.
20. Draw a neat block diagram of CRO. Explain how intensity and sharpness of the trace on the screen is controlled.
21. Draw the block diagram of CRO. (Amaravati Univ., 2003, 2004, 2006, 2007), (R.T.M.N.U., 2006)
22. Draw a neat block diagram of a CRO. Explain how the actual waveform of the signal is traced on the CRO screen with the help of time base generator. (Amaravati Univ., 2006)
23. Draw the block diagram of CRO. Explain the function of trigger circuit. (Amaravati Univ., 2004)
24. Draw block diagram and explain construction, working and applications of cathode ray oscilloscope (CRO). (RGPV, 2007)
25. What are the important parts of CRO ? Explain the function of sweep generator. (Amaravati Univ., 2002)
26. Explain the working of time base circuit and trigger circuit in CRO. (R.T.M.N.U., 2007)
27. Explain how the true shape of a voltage waveform is displayed on a CRO screen.
28. In the normal operating condition of CRO, what are the voltages present on X-plates and Y-plates of its CRT?
29. How can the frequency of a given signal be determined using the CRO?
30. Discuss two applications of a CRO.
31. Draw a block diagram of CRO and explain the function of time base circuit. (R.T.M.N.U., 2006)
32. What is CRO? Give its applications. (Amaravati Univ., 2002, 2003, 2005)
33. Explain the effect of a nonuniform magnetic field on the motion of a charged particle.

CHAPTER

16

Elements of Thermodynamics

16.1 INTRODUCTION

The various efforts of converting heat to obtain mechanical work led to the development of steam engine in the eighteenth century. Initially, the steam engine was developed to pump water out of deep coal mines which were flooded with water. These steam engines were developed by practical inventors who followed trial and error methods rather than any scientific principles. In 1765, James Watt invented a superior steam engine with a separate condenser. It stimulated the development of engines that could do many kinds of jobs and marked the beginning of industrial revolution. In subsequent years, many ambitious inventors claimed to have invented perpetual motion machines which could do work without any input. The analysis of these false claims and the genuine efforts to improve the efficiency of heat engines resulted in the birth of thermodynamics as a separate branch of physics. The laws of thermodynamics comprise the essence of 200 years of experimentation and theoretical interpretation.

The works of Julius Robert Mayer (1814-1878), James Prescott Joule (1818-1889) and Herman von Helmholtz (1821-1894) established the law of equivalence of heat and work. Rudolf Clausius (1822-1888) formulated the second law of thermodynamics mathematically. Ludwig Boltzmann (1844-1906) gave the statistical interpretation of the second law of thermodynamics in 1877. Josiah Willard Gibbs (1839-1903) generalized the methods of Boltzmann.

Thermodynamics is the science that deals with the rules according to which bodies exchange energy. In the early stages, thermodynamics was concerned with the relationship between mechanical and heat energy. Further developments turned thermodynamics into a science that is concerned with the relationships between heat and all other kinds of energy such as chemical, electrical etc. The entire structure of thermodynamics rests on two laws known as first and second laws of thermodynamics. The *first law of thermodynamics* embodies the law of conservation of energy which encompasses all the fields of science and engineering. The *second law of thermodynamics* expounds the idea that it is impossible to convert a given amount of heat fully into work. The second law is also known as the *law of increasing entropy*, on the basis of its interpretation at molecular level.

16.2 CONCEPT OF TEMPERATURE

The concept of temperature originated from our subjective sense of hot and cold. We use these words to denote the degree of heating to which a body is subjected. The quantity that

characterizes the degree of heating to which a body is subjected is termed the **temperature of that body**.

Experience shows that when a hot body is placed in contact with a cold body, their temperatures change. The body with the higher temperature always gets cooler and the body with the lower temperature always gets hotter. It indicates that energy exchange takes place between the two bodies. Such an exchange of energy is called **heat exchange**. Eventually, the temperature of the bodies become equal and the change in temperature stops. In technical terms, the bodies are said to be in **thermal equilibrium**. Thus, if the temperatures of two bodies are equal, no heat exchange takes place between them and the energy of each body remains constant. We may therefore define temperature as a quantity indicating the direction of heat exchange.

The concept of thermal equilibrium leads us to the **zeroth law of thermodynamics**. It may be stated formally as follows:

"Two bodies, A and B, each in thermal equilibrium with a third body C are in thermal equilibrium with each other".

This law implies that equality of temperature is necessary and sufficient to ensure thermal equilibrium. This law is really the basis of temperature measurements, for numbers can be marked on the thermometer and every time a body has equality of temperature with the thermometer, we can say that the body has the temperature we read on the thermometer.

According to kinetic theory of gases, temperature is a quality characterizing the average kinetic energy of translational motion of the molecules of an ideal gas. It implies that the temperature of every body is closely related to the energy of motion of its molecules. The greater the average kinetic energy per molecule of a body, the higher its temperature. Therefore, to heat a body, energy must be supplied to it, and to cool the body energy must be taken away from it.

16.3 HEAT

In the eighteenth century, heat was considered to be a massless and invisible substance called **caloric**. According to the caloric theory, every body contains an amount of caloric that depends on the temperature of the body. Thus, a hot body contains more caloric than a cold body. This theory could explain a number of observations concerning heat and temperature effects. Count Rumford (Benjamin Thompson, American – British Physicist, 1753-1814) was observing in 1798 the drilling of artillery gun barrels in Munich Armory. Every object was supposed to contain a definite amount of caloric, but Rumford noticed that heat continued to be generated as long as the boring tool scraped away at the barrel. Further, blunt drills generated more heat than sharpened ones. He observed that drilling was somehow capable of generating an almost unlimited amount of heat. He concluded that the generation of heat was due to the mechanical work done in the process of boring and continued as long as this work was being done.

Julius Robert Mayer (1814-1878), a German doctor serving as a physician on a ship, conceived on 1842 the idea of the equivalence of heat and work. He hypothesized that *heat is a form of energy*. However, the relationship between work and heat was established in 1850 experimentally by James Joule (1818-1889), the British Physicist. He conducted a long series of experiments that conclusively demonstrated the equivalence of mechanical energy and heat energy.

According to the modern view, heat is the energy that is transferred from a hot body to a cold body. For example, let us consider the case of a beaker of hot water (hot body) left on the table in an environment at room temperature (cold body). The hot water gets colder as its temperature tends to approach the room temperature. It suggests that there is some sort of

exchange of energy between water (the system) and the surroundings. We now define heat in a more general way.

Heat is energy that flows from one body or system to another solely as the result of a temperature difference between them.

Thus, heat is not an intrinsic property of a body. A statement that a body contains a certain amount of heat is meaningless. On the other hand, we can say that a body can transfer a certain amount of energy as heat under certain specified conditions.

16.4 THERMODYNAMICS

The conversion of mechanical energy into heat and the reverse process of obtaining mechanical work at the expense of heat are of the greatest interest in engineering. For example, in a thermal power station, thermal energy in the form of steam at high temperature and pressure drives a turbine (Fig.16.1) and is converted into mechanical energy. The turbine in its turn drives the rotor of a generator which produces electricity.

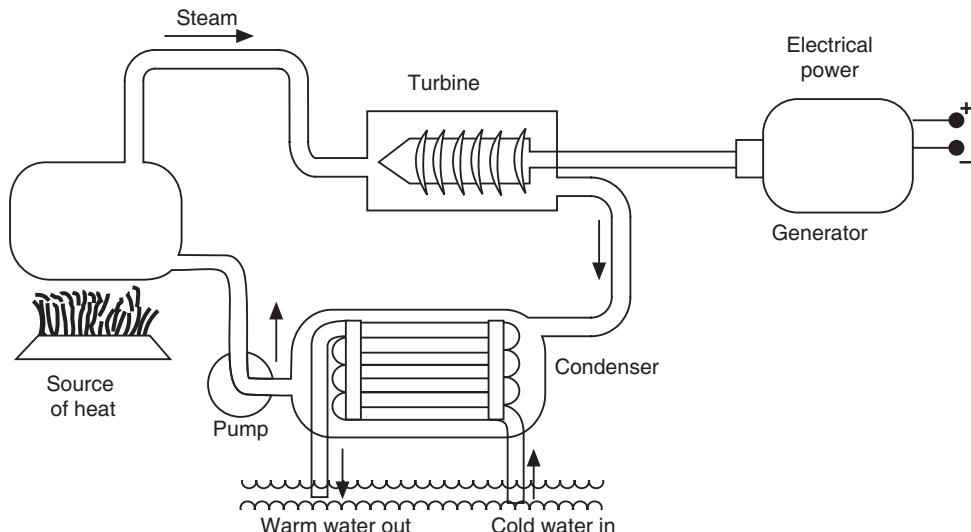


Fig. 16.1: Schematic diagram of an electrical power plant

Thermodynamics is the science that deals with work and heat, and those properties of substances related to heat and work. Many of the properties of a substance and many phenomena can be studied either from a microscopic or macroscopic point of view. For instance, let us consider the familiar example of a monoatomic gas confined in a small container. Quantities, characteristic of the atoms, such as the energy of an atom, its velocity, its mass etc can be used to describe the properties of the gas. These quantities can be computed from theoretical analysis but cannot be measured. All such quantities are said to be **microscopic**. We may also describe the properties of the gas using quantities whose values can be directly measured. Such quantities are called **macroscopic** quantities. Temperature, pressure and volume are examples of macroscopic quantities. These are the quantities the gas as a whole has, but have no meaning when applied to individual atoms. Thus, we cannot speak of the pressure or temperature of one molecule. It is possible to develop the principles of thermodynamics from a microscopic point of view. Historically, the central concepts of thermodynamics were developed from a macroscopic view point without reference to microscopic models and details of the structure of matter. We study here the principles of thermodynamics from the macroscopic point of view.

16.5 TERMINOLOGY

There are certain terms in thermodynamics which are used with specific connotation. It is necessary for us to acquaint with these terms before we embark on the study of thermodynamics.

16.5.1 System

The principles of thermodynamics are usually stated with reference to a well defined system.

A system is defined in thermodynamics as a quantity of matter of fixed mass and identity.

A familiar example of a thermodynamic system is a quantity of gas enclosed in a cylinder fitted with a movable piston (Fig.16.2). The system and its environment are distinctly identified by drawing a definite boundary between them. The system can interact with its environment, mainly in two ways – by way of transfer of heat or by way of doing work.

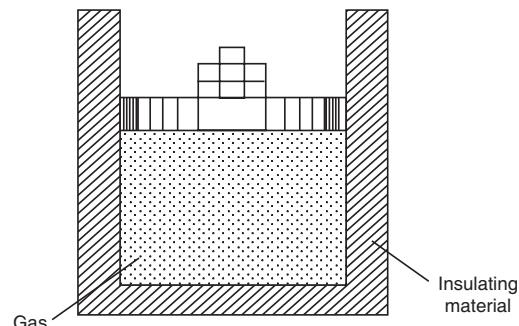


Fig. 16.2. A gas in cylinder fitted with a movable piston is an example of a thermodynamic system

16.5.2 State

The thermodynamic state of a system is described with the macroscopic quantities such as pressure and volume. The physical quantities which unambiguously determine the state of a body are called **thermodynamic variables**. The variables that are needed to define completely the state of the system, illustrated in Fig.16.2, are the temperature T , the pressure P , the volume V , and the mass, m . Two states of a body are considered to be different if the value of even one of the thermodynamic variables is different for them.

16.5.3 Thermodynamic Equilibrium

In mechanics, equilibrium means a state of rest. In thermodynamics the concept is somewhat broader. *A system is said to be in thermodynamic equilibrium if none of the thermodynamic variables determining its state changes with time.* Thermodynamic equilibrium is easily understood in the case of a monoatomic gas confined in a cylinder, as illustrated in Fig.16.2. If the temperature of the gas in the cylinder is the same at all points in the cylinder and the temperature of the walls of the cylinder is also the same, then the gas is in thermal equilibrium with the cylinder. The heat of the gas does not flow from one part of the cylinder to another. Further, when neither the pressure nor chemical composition of the gas changes, it is said to be in thermodynamic equilibrium. Thus, a system will be in a state of thermodynamic equilibrium if it satisfies the conditions for mechanical, thermal and chemical equilibrium.

A Cartesian coordinate system is used to plot sets of values of P , V , and T to indicate equilibrium states of a system, as in Fig.16.3. Different points on the graph correspond to different equilibrium states of the gas.

An isolated system always reaches a state of thermodynamic equilibrium in course of time but can never depart from it spontaneously.

16.5.4 Process

Any change in the state of a system is called a thermodynamic process. In a thermodynamic process one or more of the thermodynamic variables change. For instance, an increase in the pressure exerted on a gas causes a decrease in its volume. The path of the succession of states

through which the system passes is called a **process**.

Different thermodynamic processes are studied and compared by depicting them graphically. Such graphs are called process charts and ***thermodynamic diagrams***. Two varying parameters of state, for example the pressure P and volume V are laid off along the axes of a two – dimensional Cartesian coordinate system. It is called a ***PV-diagram***. The dependence of P on V shows the given thermodynamic process.

16.5.5 Quasistatic Process

If a thermodynamic process can be represented as a continuous succession of equilibrium states of the system, it is termed to be an equilibrium process or a **quasistatic process**. Therefore, a quasistatic process is a thermodynamic process in which a system passes through a series of equilibrium states. Geometrically, a quasistatic process can be represented by a curve that joins the initial and final equilibrium states by a succession of intermediate equilibrium states, as in Fig. 16.4.

Only equilibrium states and quasistatic processes can be depicted graphically.

If a system, which is originally in an equilibrium state, is momentarily disturbed and then left alone, it passes on to equilibrium state spontaneously. The process of transition to equilibrium is called **relaxation**. The time needed for the transition is known as the **relaxation time**. If a process is to be quasistatic, it must take place during a time that is longer than the relaxation time. During such a slow process the system can be considered to pass through an infinitely large number of equilibrium states.

A quasistatic process is illustrated in Fig. 16.5. A cylinder filled with a gas has piston at one end. There is no friction between the piston and the cylinder walls. Initially, the gas is in the state of equilibrium. Now, if we drop sand, grain by grain on the piston, it moves down slowly through infinitesimally small distances. The gas is compressed slightly near the piston each time a grain is added and its pressure, density and temperature are increased. In other

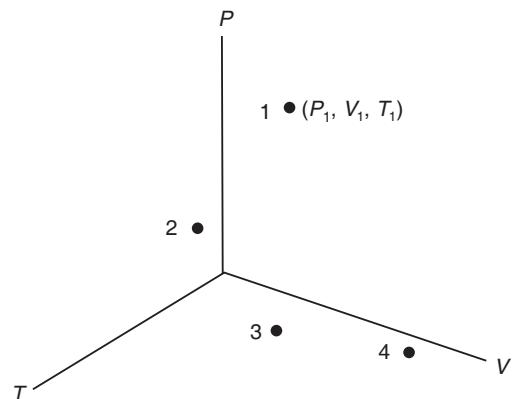


Fig. 16.3: Different equilibrium states of a gas can be represented by points specified by values of pressure, volume and temperature

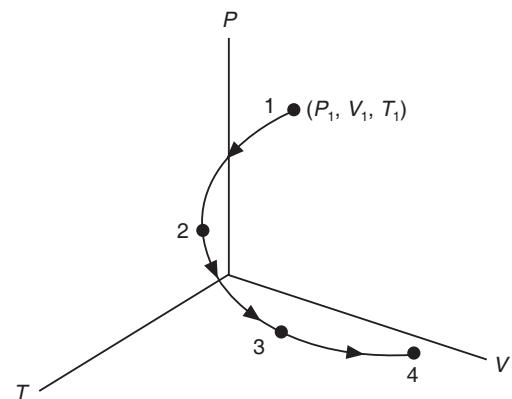


Fig. 16.4: The curve connecting the equilibrium states of a system represents the path of the process.

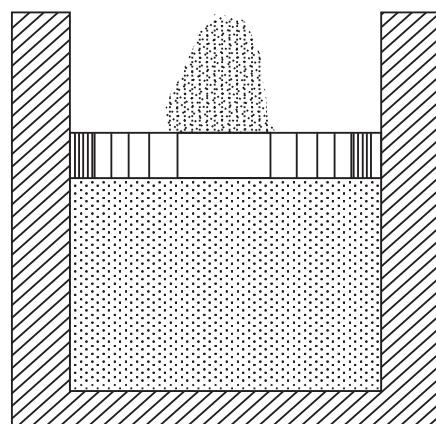


Fig. 16.5: A gas in a cylinder is compressed slowly by adding grains of sand onto the piston

parts of the gas, the variables have not had sufficient time to change. After a time, equal to relaxation time, equilibrium is restored. It means that the thermodynamic variables take new values, but again the same throughout the gas. If the grains of sand are added slowly such that the time of each micro-compression is greater than the relaxation time of the gas, then the time will be sufficient for restoration of the equilibrium. Each added grain of sand takes the system into a new equilibrium state. Then, the whole process of compression can be regarded as the sum of a large number of micro-compressions and the thermodynamic process as the totality of transitions through a large number of equilibrium states.

16.5.6 Reversible and Irreversible Processes

Suppose a perfectly elastic ball drops in a vacuum onto a perfectly elastic surface. As the ball travels down, its potential energy (mgh) transforms into kinetic energy $\left(\frac{1}{2}mv^2\right)$. At the instant when the ball touches the surface, its kinetic energy exactly equals to the initial potential energy $\left(\frac{1}{2}mv^2 = mgh\right)$. Upon the impact elastic forces appear due to deformation of the ball and the surface. These forces will make the ball begin reverse motion upward. The energy of deformation will transfer into the kinetic energy of the motion of the ball and it will rise to the same height from which it began to drop. This process repeats many times. The upward motion of the ball is a process that is reverse of downward motion. The ball when rising passes through the same intermediate states determined by its coordinates and velocities as in falling, but in the reverse sequence. Consequently, this mechanical process is a **reversible process**.

Basing on the above example, we may define a reversible process as a process that satisfies the following conditions:

- i. The process can be carried out with equal facility in two opposite directions.
- ii. In each direction the system passes through the same intermediate states; and
- iii. After the completion of the process, the system and its environment return to their initial state.

Any process which does not satisfy even one of these conditions is **irreversible**.

In thermodynamics a reversible process is a quasistatic process that can be reversed by an infinitesimal change in the surroundings. Thus, a thermodynamic process is **reversible** if the system passes from the initial state to final state through a succession of equilibrium states. A reversible thermodynamics process may be understood with the help of the example given in Fig. 16.5.

In the example cited there, adding or removing a grain of sand constitutes an infinitesimal change in the surroundings. If the sand is added grain by grain, the gas is compressed quasistatically and the gas passes through a series of equilibrium states to reach the final equilibrium state. Now, if the sand is removed by grain by grain, the gas passes through the same sequence of equilibrium states during the expansion, and reaches the initial state.

Since a reversible process is defined by a succession of equilibrium states, it can be represented by a line on a *PV* diagram, as in Fig. 16.6. The line represents the *path* of the process. Each point on the curve represents one of the intermediate equilibrium states. On the other hand, an irreversible process passes from the initial state to the final state through a series of nonequilibrium states. In this case, only the initial and final equilibrium states can be represented on the *PV* diagram. The intermediate nonequilibrium states are not characterized by uniquely defined sets of values for thermodynamic parameters. Hence, an irreversible process cannot be represented by a line on a *PV* diagram.

There are many factors that make processes irreversible. There are no strictly quasistatic processes in nature because all thermal processes occur at a finite rate and not infinitely slowly. Consequently, all real processes in nature are irreversible. Let us consider three examples that demonstrate the irreversibility of real thermal processes.

(i) Free expansion of a gas:

Consider the example of a gas separated from a vacuum by a membrane, as in Fig. 16.7 (a). If the membrane breaks, the gas fills the entire vessel (Fig. 16.7 b). The system cannot be restored to its initial state unless the gas is compressed and the resulting heat is taken away. This, however, causes a change in the surroundings. Therefore, at the end of the process, the surroundings are not restored to their initial state. For this reason, the free expansion of a gas is an irreversible process.

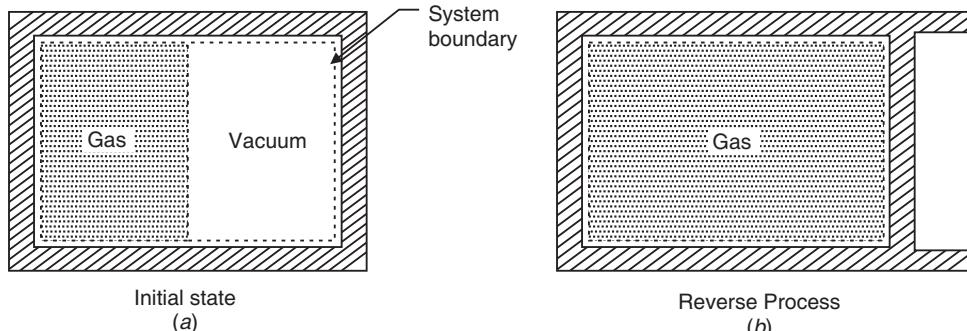


Fig. 16.7: Free expansion of a gas. When the partition separating the gas from the evacuated space is ruptured, the gas expands freely and irreversibly.

(ii) Diffusion: Let us consider the case of two different gases, oxygen and hydrogen, separated by a membrane, as in Fig. 16.8 (a). If the membrane breaks, diffusion of gases takes place spontaneously and a homogeneous mixture of oxygen and hydrogen fill the entire volume, as in Fig. 16.8 (b). But the process will never reverse itself. The mixture of gases will never divide by itself into the initial components.

In practice, a mixture of gases can be divided into its initial components. But first, it involves application of energy. Secondly, the system will not pass again through the same intermediate states it passed through in the diffusion process. Thirdly, the system cannot be returned to its initial state without substantially altering the properties of the surroundings.

(iii) Heat Exchange: Experience shows that heat exchange, like diffusion, is a one-way process. In heat exchange energy is always transmitted from a body at a higher temperature to another body at a lower temperature. Consequently, heat exchange is always accompanied by an equalization of temperatures. The reverse process of transferring energy in the form of heat from cold bodies to hot bodies never occurs by itself. For example, a hot cup of coffee cools by virtue of heat transfer to the surroundings, but heat will not flow from the cooler surroundings to the hot cup of coffee.

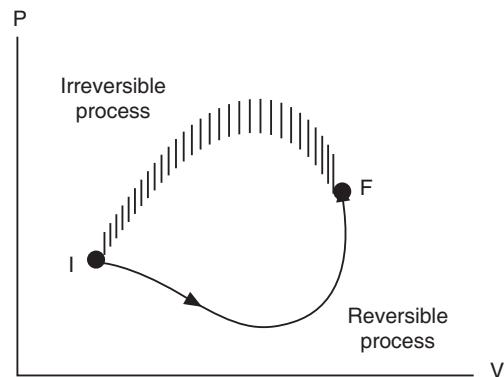


Fig. 16.6: A reversible process between the initial state and final state can be represented by a line on PV diagram. An irreversible process passes through a series of nonequilibrium state and cannot be represented by a line on PV diagram.

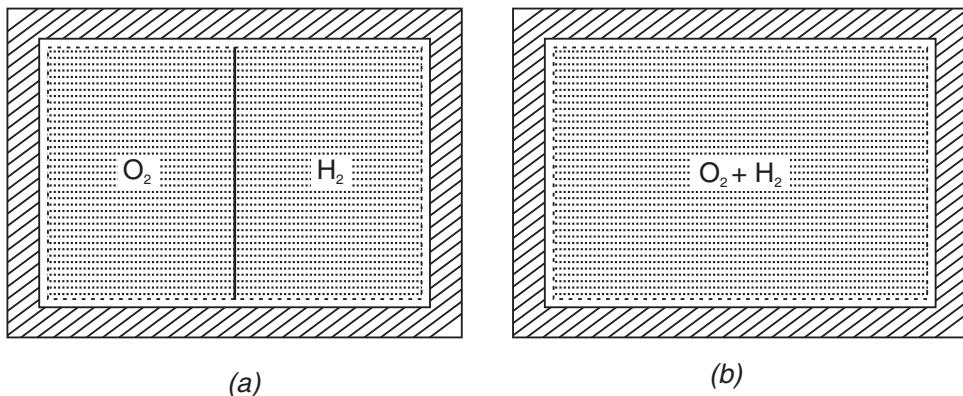


Fig. 16.8: (a) Oxygen and hydrogen gases separated by a partition.
(b) Removal of partition leads to the irreversible mixing of the two gases

16.5.7 Cycle

When a system in a given initial state passes through a number of different changes of state or processes and eventually returns to its original equilibrium state, then the system is said to have undergone a **thermodynamic cycle**. The steps that constitute the cycle may be reversible or irreversible. If the system consists of a single homogeneous substance and if all the steps are reversible, the cycle can be represented by a closed curve in a PV diagram, as in Fig.16.9.

16.5.8 Heat Reservoir

A **heat reservoir** is a body from (or to) which heat can be transferred without causing a change in the temperature of the reservoir.

Such a heat reservoir remains at constant temperature always. For example, the atmosphere is a good heat reservoir. A heat reservoir from which heat is extracted by a body is called a **high-temperature reservoir**.

A heat reservoir to which heat is ejected from a body is called a **low-temperature reservoir**.

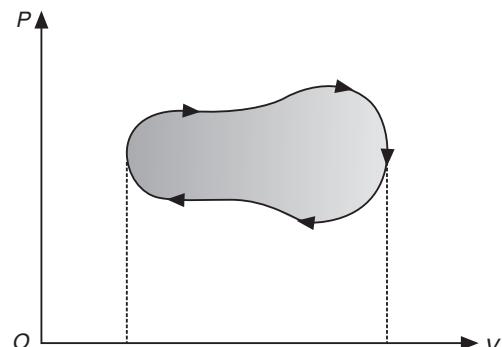


Fig.16.9: The PV diagram for an arbitrary cyclic process. The net work done in the process equals the area enclosed by the curve.

16.6 WORK

The mechanical work performed by a force is the most familiar form of work. In mechanics, work is said to have been performed when a force acting on a body displaces it through a distance, the displacement being in the direction of force. That is,

$$W = F \cdot x \quad (16.1)$$

When mechanical work is done on a body, its mechanical energy changes. Therefore, it serves as the measure of transfer of mechanical energy from one body to another. In general, we say work is a form of energy. Work is not stored in the body. **Work is a form of transferring energy and is a measure of the transferred energy.**

Now let us examine the work done in a thermodynamic process. Consider a thermodynamic system such as a gas contained in a cylinder fitted with a movable piston, as in

Fig.16.10 (a). In equilibrium the gas occupies a volume V and exerts a uniform pressure P on the piston and the walls of the cylinder. If the cross-sectional area of the piston is A , the force exerted by the gas on the piston is

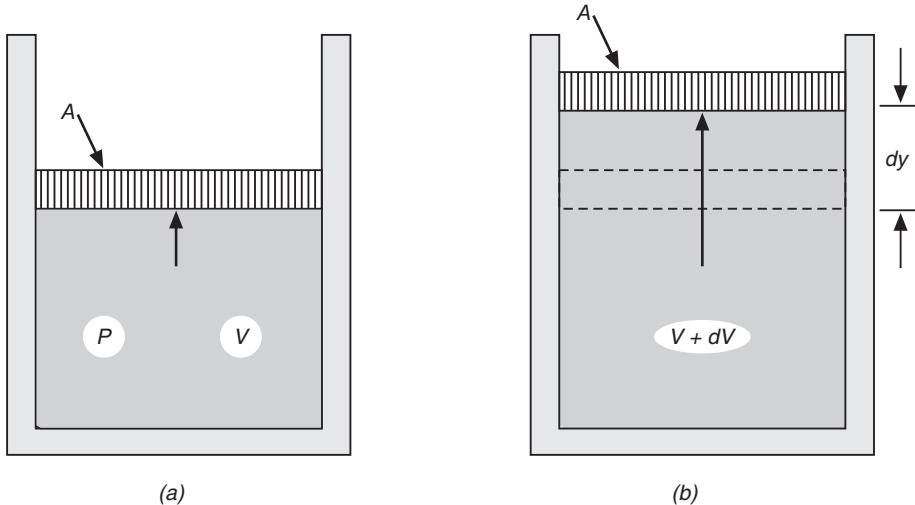


Fig.16.10: (a) Gas contained in a cylinder at a pressure P does work on the piston. (b) As the piston moves the system expands from a volume V to a volume $V + dV$.

$$F = PA$$

Now, let us assume that the gas expands quasistatically. As the piston moves up a distance (Fig.16.10 b), the work done by the gas on the piston is

$$\delta W = F dy = PA dy$$

But $A dy = dV$, the increase in the volume of the gas. Therefore,

$$\delta W = P dV \quad (16.2)$$

The relation (16.2) expresses the work solely in terms of the thermodynamic variables of the system. The nature of the external force and other characteristics of the surroundings do not appear in this relation. The work $\delta W = P dV$ is often called *thermodynamic work*. In practice, we refer to it simply as work.

The convention is that the work done by a system is regarded as positive and the work done on a system is negative. Thus, the work done by the gas expanding against a piston is positive and the work done by a piston compressing a gas is negative.

The total work done by the gas as its volume changes from V_1 to V_2 can be found by integrating equ.(16.2). Thus,

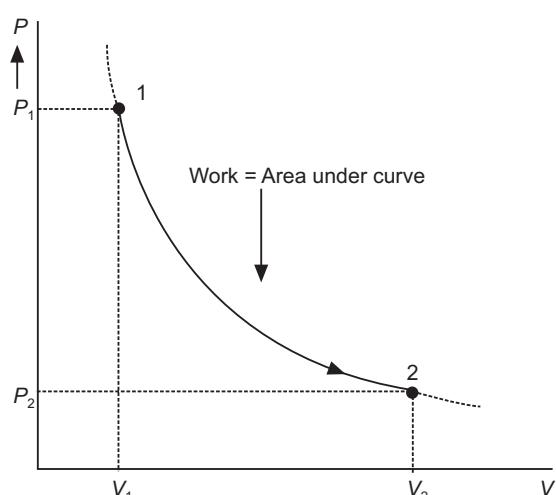


Fig. 16.11: A gas expands from an initial state 1 to final state 2. The work done by the gas equals the area under the PV curve.

$$W_2 = \int_1^2 \delta W = \int_{V_1}^{V_2} P dV \quad (16.3)$$

The above integration can be performed only if we know the relationship between P and V during the process. In general, the pressure is not constant but depends on the volume and temperature. If the pressure and volume are known at each step of the process, the work done can be obtained graphically from the PV diagram (Fig.16.11). The work done during the process is given by the area $V_1-1-2-V_2-V_1$ under the curve 1-2 on Fig.16.11. Therefore, the total work done during the expansion of the gas from the initial state to final state is the area under the curve in a PV diagram.

It is possible to go from the initial state 1 to the final state 2 along many different paths such as A, B or C, as depicted in Fig.16.12. Since, the area underneath each curve represents the work for each process, it is evident that the amount of work performed in each case is a function of the end states of the process and also it is dependent on the path that is followed in going from one state to another. To illustrate this important point, we consider three different paths (Fig.16.13) connecting initial and final states.

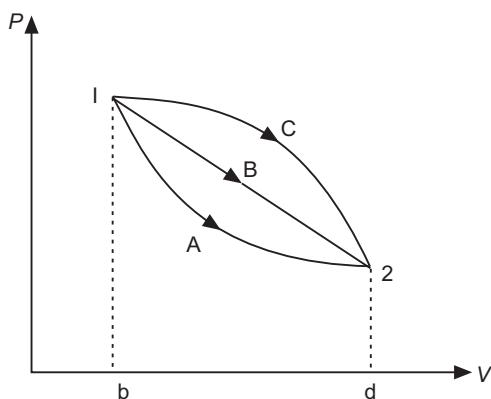


Fig. 16.12: Three different paths connecting the same initial and final states. The work done by the gas is a maximum for the path marked C which encloses the largest area.

In the process represented in Fig.16.13 (a) the pressure of the gas is first reduced from P_1 to P_2 by cooling it at a constant volume V_1 . Next, the gas is allowed to expand from V_1 to V_2 at constant pressure P_2 . The work done along this path is $W_A = P_2(V_2 - V_1)$

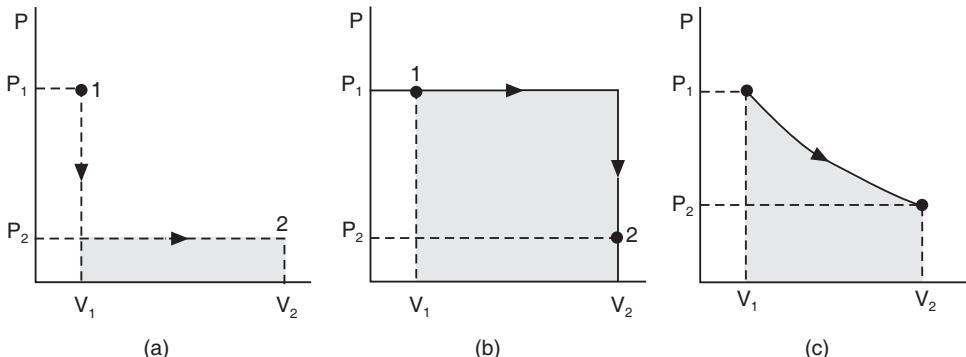


Fig. 16.13: The work done by an ideal gas as it is taken from an initial state 1 to final state 2 depends on the path between these states 1 & 2.

In the process represented in Fig.16.13 (b), the gas is first allowed to expand from V_1 to V_2 at constant pressure P_1 and then the pressure is reduced to P_2 at constant volume V_2 . The work done along this path is $W_B = P_1(V_2 - V_1)$.

Finally in the process described in Fig.16.13 (c), both P and V change continuously. To compute the work in this case, the shape of the PV diagram must be known unambiguously.

It is clear from the PV diagrams in Fig.16.13 that W_A is smaller than W_B and W_C has a value intermediate between W_A and W_B . This example amply demonstrates that the work done by a system depends on how the system goes from the initial to the final state. For this reason, thermodynamic work is not a point function but is a path function. dW cannot be therefore treated as exact differential in the mathematical sense.

16.7 HEAT IN THERMODYNAMICS

If a hot body is immersed in cold water, we know from experience that the hot body cools down and the water warms up until the body and water attain the same temperature. It suggests a transfer of energy from the hot body to the water. Such a process of transfer of energy without work being done is called **heat exchange**. We may define heat as a form of energy that is transferred from a system at a given temperature to another system at a lower temperature solely by virtue of the temperature difference between the two systems. Thus, heat is energy in transit. Heat transferred to a system is considered positive and heat transferred from a system is negative.

Like work, heat also is a path function. The heat transferred to or from a system depends on the process. We illustrate this point with the help of the following examples. Let us consider the case of a gas of volume V_1 which is in thermal contact with a heat reservoir (Fig.16.14a). If the pressure of the gas is infinitesimally greater than atmospheric pressure, the gas will expand and pushes up the piston. Ultimately the gas expands to a final volume V_2 . The heat required to maintain the gas at a constant temperature T , during the process of expansion, is supplied from the reservoir to the gas. Work is done by the gas in this case.

In the second case, illustrated in Fig.16.14 (b), a gas of the same volume V_1 is thermally insulated. It can neither receive nor give away heat. When the membrane is broken, the gas freely and rapidly expands into the vacuum until it occupies a volume V_2 . No work is done in this case. Further, no heat is transferred through the walls. The initial and final states of both processes are identical but the paths followed are different. We conclude that heat, like work, depends on how the system goes from its initial state to final state. Therefore, it is path function. Like δW , δQ is an inexact differential.

16.8 COMPARISON OF HEAT AND WORK

In thermodynamics, **work includes all forms of energy except heat**. When a battery is charged, electrical work was performed and stored in chemical form. When a piece of steel is magnetized, magnetic work is performed.

Any kind of energy in the course of transformations may pass through many forms but invariably ends in the form of heat energy. In the process of mechanical motion, the kinetic energy of a body decreases owing to the action of friction forces and gets transformed into heat. Similarly, the energy of an electric current, and of chemical reactions transform into heat.

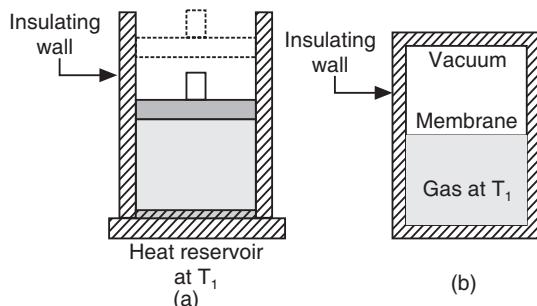


Fig. 16.14: (a) A gas at temperature T_1 expands slowly by absorbing heat from a reservoir at the same temperature. (b) A gas expands rapidly into the evacuated region due to the rupture of the membrane.

Both work and heat are *transient* phenomena. Systems never contain work or heat. Heat and work exist only in a process of energy transfer, and their numerical value depends on the kind of this process. Either work or heat or both are transferred when a system undergoes a change of state. In real conditions, both ways of transferring energy to a system accompany each other. For instance, when a metal rod is heated, heat exchange takes place and at the same time thermal expansion of rod occurs. The latter implies that the work of expansion is done.

Heat and work are not equivalent forms of energy transfer from a qualitative viewpoint. The energy of ordered motion is transferred in the form of work. The energy of chaotic motion of particles constituting the body is increased when energy is transferred to a body in the form of heat.

Work can be completely converted into heat. On the other hand, heat cannot be fully converted into work. Hence work is said to be *high grade energy* and heat as *low grade energy*.

Both heat and work are path functions and inexact differentials. Since they depend on the path, neither quantity is independently conserved during a thermodynamic process.

16.9 INTERNAL ENERGY

The total energy E_T of a system consists of (i) the kinetic energy of its macroscopic motion as a whole (ii) the potential energy due to the presence of external fields and (iii) internal energy, U . Thus,

$$E_T = K.E. + P.E. + U \quad (16.4)$$

The internal energy U of the system depends on the nature of the motion and interaction of the particles in the system. It consists of (a) the kinetic energy of random thermal motion of molecules (b) the potential energy of molecules due to intermolecular interactions, (c) the kinetic and potential energies of atoms and electrons, and (d) the nuclear energy. In thermodynamics, we do not concern ourselves with the form of internal energy.

Experiments have shown that the internal energy is determined by the thermodynamic state of the system and does not depend on how the system acquired the given state. Consequently, the internal energy is not related to the process of a change in the state of the system. For example, the internal energy of a given amount of gas depends on its pressure and temperature. It makes no difference how the gas achieved a particular pressure and temperature. All processes leading to a particular pressure and temperature leave the gas with the same internal energy. Keeping this in view, we may define **internal energy** as the energy stored in a system.

The internal energy is not of practical interest. The change in internal energy ΔU , when a system changes from one state to another, is of actual interest. It is generally assumed that the internal energy of a system is zero at 0K.

16.10 LAW OF CONSERVATION OF ENERGY

In a closed mechanical system, the sum of kinetic energy and potential energy is constant. *The total energy can neither be created nor destroyed.* It means that energy is conserved. This is the law of conservation of mechanical energy. This law is applicable only in situation where there is no transformation of mechanical energy into heat energy. In real systems the mechanical motion is in general accompanied with heating. When the engine of a running car is switched off, it gradually slows down and ultimately stops. Apparently, its kinetic energy has disappeared. In fact, its kinetic energy is transformed into heat energy by the friction

forces. Both the tire of the car and the ground are heated up. As a result, the random motion of particles constituting the interacting tires and road acquired more velocity. To sum up, the mechanical energy (K.E. of car) is transformed into the internal energy of the interacting bodies. If we take into account the internal energy, the law of conservation of energy can be extended to include thermodynamic systems also. The first law of thermodynamics generalizes the law of consideration of mechanical energy.

16.11 FIRST LAW OF THERMODYNAMICS

Thermodynamics is concerned with three general forms of energy, namely heat, work and internal energy. Heat is the energy transferred by virtue of temperature difference, thermodynamic work is energy (exclusive of heat) transferred between a system and surroundings, and internal energy is energy stored within a system.

If the state of a system such as the one described in Fig.16.2 changes as a result of supplying a quantity of heat Q to it, and as a consequence the system does the work W , then the law of conservation of energy states that the quantity of heat supplied to the system will be equal to the sum of the work performed by the system and the change in the internal energy of the system. That is,

$$\text{Net heat transfer} = \text{Work} + \text{Change in internal energy}$$

$$\text{Or mathematically, } Q = W + \Delta U \quad (16.5)$$

This is known as the **first law of thermodynamics**. It is in effect a statement of the conservation of energy. The net change in the internal energy of the system is always equal to the net transfer of energy as heat and work across the boundary of the system.

While the quantities Q and W are path-dependent, the internal energy does not depend on the path of the process.

Suppose a thermodynamic system undergoes a change from an initial state 1 to a final state 2 in which Q units of heat are absorbed (or removed) and W is the work done by (or on) the system. Expressing both Q and W in the same units (either thermal or mechanical) the difference $(Q - W)$ can be calculated. If now we carry out this calculation for different paths between the same states 1 and 2, the quantity $(Q - W)$ will be the same for all paths connecting the states 1 and 2. It follows that the internal energy change of a system is independent of the path. If U_1 is the internal energy in state 1 and U_2 is the internal energy in state 2, then the relation (16.5) can be rewritten as

$$U_2 - U_1 = \Delta U = Q - W \quad (16.6)$$

When a thermodynamic process proceeds smoothly, it can be treated as a continuous sequence of small changes. Mathematically, we write equ.(16.5) in the differential form as

$$dQ = dU + dW \quad (16.7)$$

It is important to note that expressing heat and work as dQ and dW does not imply the existence of properties Q and W that measure the heat and work content of a system. dQ and dW denote small amounts of heat and work but they are not true differentials.

There is a serious limitation on the first law of thermodynamics. The law tells us whether energy considerations permit a particular process to take a system from one equilibrium state to another. But it does not tell us whether this process will actually occur or not. A certain process might be entirely consistent with the principle of energy conservation but still it will not take place.

16.12 APPLICATIONS OF THE FIRST LAW

Let us now study some special processes and the application of the first law to them.

1. Isolated system: An *isolated system* does not interact with its surroundings. Therefore, there is no heat flow and the work done is zero. That is, $Q = 0$ and $W = 0$. It follows from equ. (16.6) that $\Delta U = 0$.

$$\therefore \quad U_2 - U_1 = 0 \\ \text{or} \quad U_2 = U_1 \quad \text{Isolated system} \quad (16.8)$$

Equ.(16.8) means that the **internal energy of an isolated system remains constant.**

2. Cyclic process: In a *cyclic process*, the initial and final states of the system are the same. Thus,

$$\begin{aligned} U_2 &= U_1 \\ \therefore \quad \Delta U &= 0 \end{aligned}$$

It allows from equ.(16.5) that

$$\begin{aligned} Q - W &= 0 \\ \therefore \quad Q &= W \quad \text{Cyclic process} \quad (16.9) \end{aligned}$$

Equ.(16.9) means that the **net work done by the system over a cycle equals the net heat absorbed over the cycle.**

3. Adiabatic process: A process in which no heat is absorbed or ejected by the system is called an *adiabatic process*. Thus, $Q = 0$.

$$\therefore \quad \Delta U = -W \quad \text{Adiabatic process} \quad (16.10)$$

Thus, the change in the internal energy of a system is equal in magnitude to the work done by the system. Heat flow into the system from the surroundings may be prevented in two ways – (i) by surrounding the system with a thick layer of insulating material or (ii) by performing the process quickly. The flow of heat requires finite time. So any process performed quickly enough will be adiabatic.

4. Isochoric process: When a substance undergoes a process in which the volume remains unchanged, the process is called *isochoric*. If the volume of a system remains constant, it can do no work. Thus, $W = 0$ and the first law gives

$$\Delta U = Q \quad \text{Isochoric process} \quad (16.11)$$

In this case, the **heat that entered the system is stored as internal energy.**

5. Isothermal process: A process taking place at **constant temperature** is said to be *isothermal*. In an isothermal process, the quantities, Q , W and ΔU are in general nonzero. The first law does not assume any special form for an isothermal process.

6. Isobaric process: A process taking place at **constant pressure** is called an *isobaric* process. As in the case of an isothermal process, Q , W and ΔU are all nonzero. The work done by a system that expands or contracts isobarically has a simple form. As the pressure is constant,

$$W = \int_1^2 P dV = P(V_2 - V_1) \quad \text{Isobaric Process} \quad (16.12)$$

7. Isothermal expansion of an ideal gas: Let an ideal gas be allowed to expand quasistatically at constant temperature by placing the gas in good thermal contact with a heat reservoir at the same temperature. Since the gas is ideal, and the process is quasistatic, we can apply the ideal gas equation, namely $PV = nRT$ for each point on the path.

$$\therefore \quad W = \int_{V_1}^{V_2} P dV = \int_{V_1}^{V_2} \frac{nRT}{V} dV$$

At T is constant in the process, we can write

$$W = nRT \int_{V_1}^{V_2} \frac{dV}{V}$$

or

$$W = nRT \ln\left(\frac{V_2}{V_1}\right) \quad (16.13)$$

8. Adiabatic expansion of an ideal gas: Let us find the relation between P and V for an *adiabatic process* carried out on an ideal gas. According to equ.(16.10), $\Delta U = -W$ in an adiabatic process.

For an ideal gas $\Delta U = nC_v dT$ (at constant volume)

The work done during the process is given by $W = P dV$.

$$\therefore PdV = -nC_v dT \quad (16.14)$$

We can write the equation of the state of the gas in differential form as

$$d(PV) = d(nRT)$$

$$\therefore PdV + VdP = nR dT \quad (16.15)$$

Using equ.(16.14) into equ.(16.15), we get

$$V dP = nC_v dT + nRdT = n dT(C_v + R)$$

But

$$C_v + R = C_p$$

\therefore

$$V dP = nC_p dT \quad (16.16)$$

Taking the ratio between equ.(16.16) and equ.(16.14), we get

$$\frac{V dP}{P dV} = -\frac{C_p}{C_v} = -\gamma$$

$$\therefore \frac{dP}{P} = -\gamma \frac{dV}{V}$$

Integrating on both sides of the above equation, we get

$$\int_{P_1}^{P_2} \frac{dP}{P} = -\gamma \int_{V_1}^{V_2} \frac{dV}{V}$$

or

$$\ln \frac{P_2}{P_1} = -\gamma \ln \frac{V_2}{V_1}$$

\therefore

$$P_1 V_1^\gamma = P_2 V_2^\gamma \quad (16.17)$$

or

$$PV^\gamma = \text{Constant} \quad (16.18)$$

16.13 HEAT ENGINE

A *heat engine* is a device that converts heat energy to work. The automobile engines and steam turbines are examples of heat engines. Basically, they all operate on the same principle. For theoretical purposes, any heat engine can be conveniently represented by a diagram, as shown in Fig.16.15, regardless of its particular design and features.

Any heat engine consists of three main parts: the working medium, the high-temperature reservoir and the low-temperature reservoir. The working medium is some gas (or steam) which receives a certain quantity of heat Q_H from the high-temperature reservoir, some of which is used to do useful work W and the remainder Q_L is transferred to the low-temperature reservoir.

For example, in an actual engine the energy produced in the combustion of fuel is transmitted by heat exchange to some gas. As a result the gas expands and the expanding gas does work on a piston that is coupled to a crankshaft to produce useful work output. The unused heat is rejected to the surroundings through an exhaust system. However, the gas cannot continue to expand without any limit because the engine is of finite size. Secondly, in practical applications we want an engine to deliver work continuously. For these reasons, the gas must be compressed, after expansion, so that the gas and all the parts of the engine return to their initial state. Again the expansion and compression can be repeated. This implies that a heat engine must operate in cycles so that we have work output during each cycle. As mentioned above, in the real heat engine, the gas is discharged at the exhaust. A new portion of the gas is admitted and compressed instead. It is called an **open cycle** operation. However, we assume, for the sake of simplicity, a **closed cycle** in which the same portion of gas expands and is compressed as it does not influence the thermodynamics of the process.

We understand the necessity of the low-temperature reservoir as follows. For an engine to do useful work during a cycle, it is necessary that the work done in expansion by the gas must be greater than the work required to compress the gas. Then only we obtain more mechanical energy. From *PV* diagram, Fig.16.16, it is seen that the work in expansion will be larger than that in compression only if the pressure in the compression process is lower than that in expansion for all intermediate states. This is possible only when the temperature of the gas is lower in compression than in expansion at all intermediate points. Therefore, an engine will do useful work if the temperature of the gas is lower in compression than in expansion.

16.13.1 Thermal Efficiency

Efficiency is a measure of how economical an engine is. We express cost efficiency in terms of fuel economy. In the case of a mechanical device,

$$\text{Mechanical efficiency} = \frac{\text{Work output}}{\text{Work input}} \quad (16.19)$$

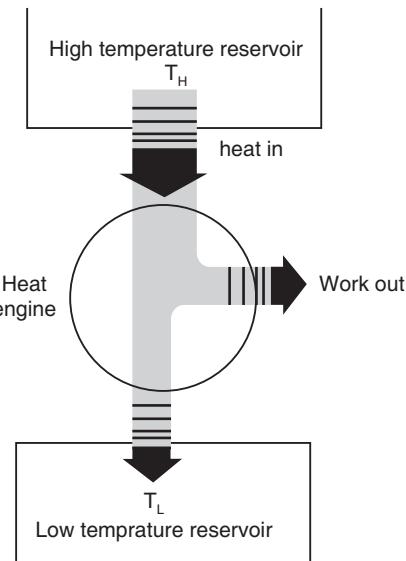


Fig.16.15. Schematic representation of a heat engine. The engine absorbs heat Q_H from the high temperature reservoir, expels heat Q_L to low temperature reservoir and does work W .

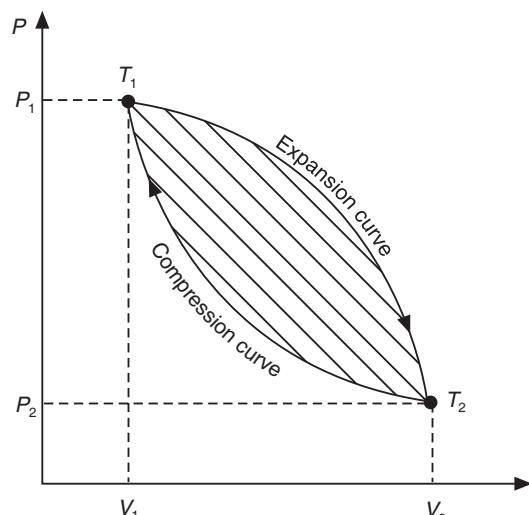


Fig. 16.16

It is the fraction or percent of useful work done by a machine.

In heat engines, the work input is the heat input. The work output equals the heat in minus the heat out.

Let us consider a heat engine operating in a cycle over and over again. Let Q_H and Q_L , be the heat received and heat rejected by the working substance per cycle. The net heat absorbed is

$$Q = Q_H - Q_L \quad (16.20)$$

The useful output of the engine is the net work W done by the working substance. Since the working substance returns to its initial state after the cycle is completed, its internal energy remains unchanged. That is, $\Delta U = 0$. It follows from the first law that

$$W = Q = Q_H - Q_L \quad (16.21)$$

Basing on the definition (16.19), we can define

$$\text{Thermal efficiency, } \eta = \frac{W}{Q_H} = \frac{Q_H - Q_L}{Q_H} \quad (16.22)$$

$$\therefore \eta = 1 - \frac{Q_L}{Q_H} \quad \text{Ideal engine} \quad (16.22 \text{ a})$$

In practice, the useful work delivered by an engine is less than the work W owing to friction losses. Therefore, the overall efficiency is less than the thermal efficiency. Thus,

$$\eta \leq 1 - \frac{Q_L}{Q_H} \quad \text{Real engine} \quad (16.23)$$

where the less-than sign refers to real engines and the equal sign to an ideal engine in which there are no losses.

Equation (16.23) points out that even an ideal engine has efficiency less than 100%. It could be 100% only if heat is not delivered ($Q_L = 0$) during the compression. But this is impossible, because the gas must be cooled in compression. Therefore, a certain amount of heat ($Q_L \neq 0$) must be delivered to the low temperature reservoir. It means that $\eta < 100\%$. A good automobile engine works at an efficiency of about 20% and diesel engines have efficiencies ranging from 35% to 40%.

16.14 THE CARNOT CYCLE

If the efficiency of all heat engines is less than 100%, what is the most efficient cycle we can have? Let us consider the case of a heat engine that receives heat from a high-temperature reservoir and rejects heat to a lower temperature reservoir. In 1824 a French engineer, Sadi Carnot (1796-1832) showed that a heat engine operating in a reversible cycle between the two heat reservoirs would be the most efficient engine possible. Such an ideal heat engine operates in a cycle in which every process is reversible. If every process is reversible, the cycle is also reversible and if the cycle is reversed, the heat engine becomes a refrigerator. Such a cycle is now called a **Carnot cycle**. The ideal engine is called the **Carnot Engine** which establishes an upper limit on the efficiencies of all engines. That is, the net work done by working substance taken through the Carnot cycle is the largest possible for given amount of heat supplied to the substance.

The Carnot cycle is an idealization of the cycle of a real heat engine. It is assumed that there are no losses of energy by heat exchange with the environment, that there is no friction in the machine and that processes of gas expansion and compression are quasistatic and therefore reversible.

To describe the Carnot cycle, we shall assume that the working substance is an ideal gas contained in a cylinder with a movable piston at one end. Fig.16.17 (a) is a schematic diagram of a Carnot engine and Fig.16.17 (b) shows the plot for a Carnot cycle.

1. The cycle begins at the equilibrium state A [Fig.16.17(b)] with the cylinder in contact with high-temperature reservoir at a temperature T_H (Fig.16.17. a). We assume that the base of the cylinder is **diathermic** (*i.e.* perfectly conducting) and that the walls and the piston are ideally nonconducting. The working substance (the ideal gas) undergoes a slow quasistatic isothermal expansion to state B . During this part of the cycle $A \rightarrow B$, the temperature and hence the internal energy of the gas are constant. Let the total amount of heat absorbed by the gas from the reservoir during the process be Q_H . This heat energy input is converted directly to the work W_{AB} done in moving the piston.

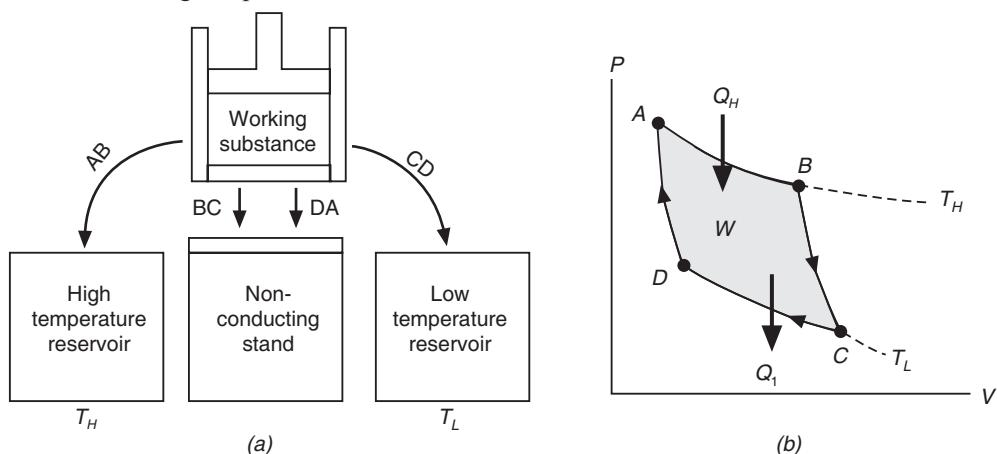


Fig. 16.17: The Carnot cycle (a) In process $A \rightarrow B$ the gas expands isothermally while in contact with a heat reservoir T_H . In process $B \rightarrow C$, the gas expands adiabatically. In process $C \rightarrow D$, the gas is compressed isothermally while in contact with a reservoir at T_L . In process $D \rightarrow A$, the gas is compressed adiabatically. (b) The PV diagram for the Carnot Cycle. The net work W done equals the net heat received in one cycle.

2. With the system at B , the cylinder is removed from contact with the high temperature reservoir and placed on the nonconducting stand. The gas expands adiabatically along the path $B \rightarrow C$ to the state C . During this part of the cycle $B \rightarrow C$ no heat enters or leaves the system. However, work W_{BC} is done at the expense of reducing the internal energy. Consequently, the temperature falls from T_H to T_L .
3. Next the cylinder is placed in contact with the low temperature reservoir at a temperature T_L . The gas is now slowly compressed (by an external agency) isothermally to a predetermined volume V_D at point D . During this part of cycle $C \rightarrow D$, the temperature and hence the internal energy of the working substance are constant. During this process $C \rightarrow D$, the gas expels heat to the reservoir and the work done on the gas by the external agent is W_{CD} .
4. In the final part $D \rightarrow A$, the cylinder is removed from contact with the low-temperature reservoir and again placed on the non-conducting stand. The gas is compressed adiabatically and brought to the initial state A . The adiabatic compression is also the result of work W_{DA} done on the gas by an external energy. Consequent to the adiabatic compression, the system temperature increases from T_L to T_H .

This step completes the Carnot cycle and returns the system to its initial condition.

The net work done in this reversible cyclic process is equal to the area enclosed by the path ABCDA of the PV-diagram (Fig.16.17b). In a reversible cycle, the change in internal energy is zero. It follows from first law that the net work done in one cycle equals the net heat transferred into the system. For a system that undergoes a Carnot cycle, no heat is supplied to or rejected by the system during the adiabatic paths, *BC* and *DA* (Fig.16.17b). An amount of heat Q_H is supplied to the system during the isothermal expansion *AB*, and an amount Q_L is rejected during the isothermal compression *CD*. Thus, the first law can be written as

$$W = Q_H - Q_L \quad \text{All heat engines} \quad (16.24)$$

Equ.(16.24) is applicable for a complete cycle of any heat engine. The thermal efficiency is given by equ.(16. 22).

$$\eta = \frac{W}{Q_H} = \frac{Q_H - Q_L}{Q_H} = 1 - \frac{Q_L}{Q_H}$$

16.14.1 Efficiency

For a Carnot engine, the heat exchanges take place during the isothermal processes. During the isothermal expansion $A \rightarrow B$, the heat absorbed from the high temperature reservoir is given by

$$Q_H = W_{AB} = nRT_L \ln \frac{V_B}{V_A} \quad (16.25)$$

In a similar manner, the heat rejected to the low temperature reservoir during the isothermal compression $C \rightarrow D$ is given by

$$Q_L = W_{CD} = nRT_L \ln \frac{V_C}{V_D} \quad (16.26)$$

Dividing these expressions, we get

$$\frac{Q_L}{Q_H} = \frac{T_L \ln(V_C / V_D)}{T_H \ln(V_B / V_A)} \quad (16.27)$$

This can be further simplified by use of the temperature-volume relation for an adiabatic process. For an adiabatic process, the pressure and volume are related by

$$PV^\gamma = \text{Constant}$$

During any reversible process, the ideal gas must also obey the equation of state,

$$PV = nRT \quad (16.28)$$

Eliminating pressure between the above equations, we get

$$T V^{\gamma-1} = \text{Constant} \quad (16.29)$$

Applying this result to the adiabatic processes $B \rightarrow C$ and $D \rightarrow A$, we get

$$T_H V_B^{\gamma-1} = T_L V_C^{\gamma-1} \quad (16.30)$$

$$\text{and} \quad T_H V_A^{\gamma-1} = T_L V_D^{\gamma-1} \quad (16.31)$$

Dividing equ.(16.30) by equ.(16.31) we get

$$\begin{aligned} \left(\frac{V_B}{V_A} \right)^{\gamma-1} &= \left(\frac{V_C}{V_D} \right)^{\gamma-1} \\ \therefore \left(\frac{V_B}{V_A} \right) &= \left(\frac{V_C}{V_D} \right) \end{aligned} \quad (16.32)$$

Substituting equ. (16.32) into equ.(16.27) , we obtain

$$\frac{Q_L}{Q_H} = \frac{T_L}{T_H} \quad \text{Carnot Engine} \quad (16.33)$$

Using the result equ.(16.33) into equ.(16.23) we obtain the thermal efficiency of the Carnot engine as

$$\eta = 1 - \frac{T_L}{T_H} = \frac{T_H - T_L}{T_H} \quad \text{Carnot Engine} \quad (16.34)$$

This surprisingly simple result says that the efficiency of a Carnot engine depends only on the temperatures of the two reservoirs. When ($T_H \gg T_L$) the difference of the two temperatures is large, the efficiency is nearly unity; when the difference is small the efficiency is less than unity. Thus, the greater the temperature difference $T_H - T_L$, the greater will be the efficiency of a Carnot engine.

We see from equ. (16.34) that all Carnot engines operating between the same two temperatures in a reversible manner have the same efficiency.

It brings us to the important difference in principle between heat engines and mechanical or electrical machines. In improving the design of mechanical or electrical machines we try to make their efficiency as near as possible to the theoretical limiting value of 100%. Though this limit is unattainable under real conditions, we can approach it by reducing losses. When we improve heat engines, we do not try to bring their efficiency to 100%, but to that of a Carnot engine operating in the same temperature interval. Naturally, the reduction of all kinds of losses also raises the efficiency of a heat engine but the most effective measure is to increase the temperature difference between the high temperature reservoir and low temperature reservoir.

16.14.2 Carnot's Theorem

The most important aspect of the Carnot engine is that it is the most efficient heat engine operating between the two heat reservoirs. This is summarized in the **Carnot's theorem** as follows:

No heat engine operating between two heat reservoirs can be more efficient than a Carnot engine operating between the same two reservoirs.

The Carnot's theorem reveals a fundamental limitation on the conversion of heat into work.

16.15 HEAT PUMP

A **heat pump** is a device that transfers heat from a low-temperature reservoir to a high-temperature reservoir. This is the reverse function of a heat engine. A heat pump is represented by a diagram as shown in Fig. 16.18. It requires work input, because the heat transfer from low temperature reservoir to high-temperature reservoir will not take place spontaneously.

The household refrigerator and air conditioner are examples of the most common heat pumps. In an ordinary household refrigerator, the working substance is a liquid that circulates within the system. It takes heat from the low temperature reservoir, which is the cold chamber in which food articles are stored, and transfers it to the outside air in the room where the unit is kept. The work is supplied by a compressor that uses electrical energy. The working substance is a fluid that readily undergoes a liquid to gas phase change at the operating temperature. The most common refrigerants are Freon – 12 (CCl_2F_2) and ammonia.

A schematic diagram of a refrigerator is shown in Fig. 16.19. Starting with the gaseous refrigerant in the compressor, the gas is compressed and emerges at a high temperature and pressure. It then passes into the condenser, where it is cooled and liquefies, and the heat from the refrigerant is rejected to the surroundings. The liquid refrigerant then goes through a throttle valve. Work is done by the liquid in getting through the valve at the expense of its internal energy. It leads to the drop in the temperature of the liquid. The cooled liquid flows through the evaporation coils where it absorbs heat and boils. The gaseous refrigerant goes into the compressors and the cycle repeats.

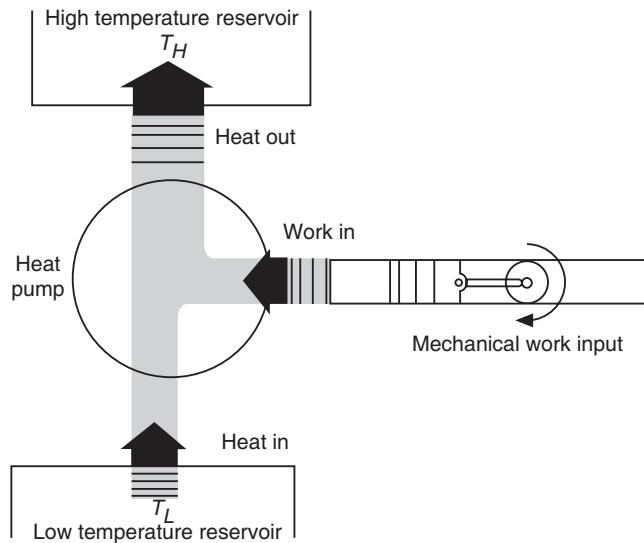


Fig. 16.18. Schematic representation of a refrigerator. It absorbs heat Q_L from the low temperature reservoir and expels and heat Q_H to the high temperature reservoir. Work W is done on the engine.

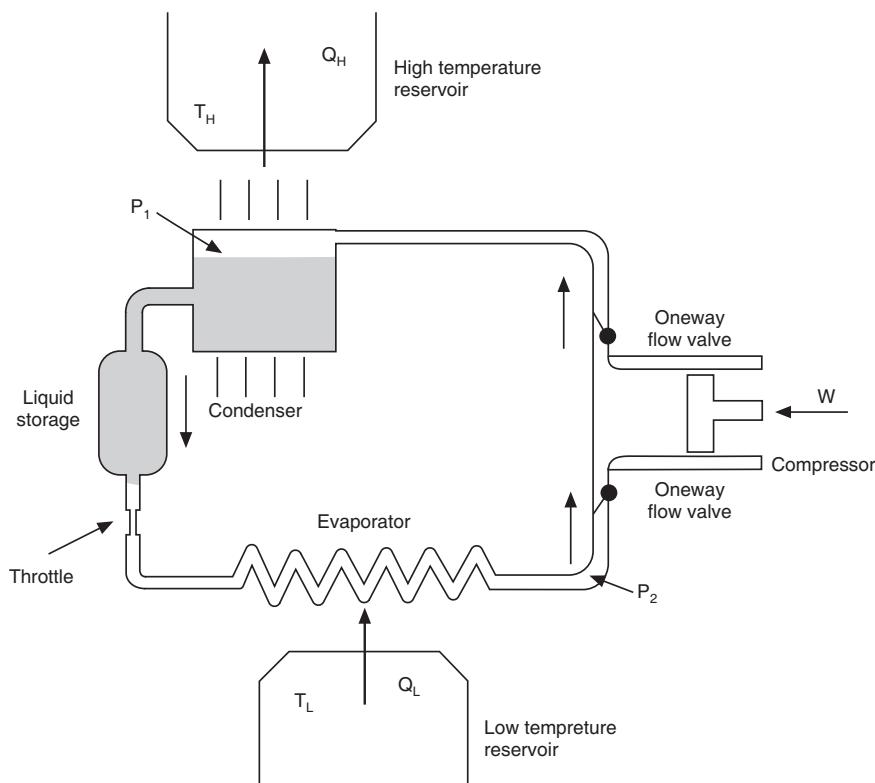


Fig. 16.19. Essential features of a common refrigerator.

The performance of a refrigerator is specified by a figure of merit K_r called the **coefficient of performance**.

$$K_r = \frac{Q_L}{W} = \frac{Q_L}{Q_H - Q_L} \quad (16.35)$$

The quantity Q_L is the amount of heat absorbed from the cold compartment at the temperature T_L . The quantity Q_H is the heat rejected into the surroundings. Q_H is equal to the heat absorbed Q_L plus the work done on the refrigerant by the compressor.

Thus, $W = Q_H - Q_L$. Assuming that the cycle is a reverse Carnot cycle, we can write.

$$K_r = \frac{T_L}{T_H - T_L} \quad \text{Carnot cycle} \quad (16.36)$$

For a house-hold refrigerator, typical temperatures are $T_H = 305\text{ K}$ (32° C) and $T_L = 260\text{ K}$ ($= -13^\circ\text{ C}$) which give a typical value of 5.8 for K_r .

16.16 SECOND LAW OF THERMODYNAMICS

There are two classical statements of the second law of thermodynamics. They are known as the Kelvin-Planck statement and the Clausius statement.

An analysis of operation of heat engines led William Thomson (later Lord Kelvin) and Planck to arrive at the following conclusion:

It is impossible to construct a heat engine that will operate in a cycle and which will receive a given amount of heat from a high-temperature reservoir and does an equal amount of work. The only alternative is that some heat must be transferred from the working fluid to a low temperature reservoir. Therefore, work can be done by the transfer of heat only if there are two temperature levels involved.

The analysis of operation of heat pumps led Rudolf Clausius to conclude as follows:

It is impossible to construct a device that operates in a cycle and produces no effect other than the transfer of heat from a cold body to a hot body. In effect, it is impossible to construct a refrigerator that operates without an input of work.

The first law of thermodynamics is concerned with conservation of energy. As long as the energy is conserved in a process, the first law is satisfied. A cycle in which a given amount of heat is transferred from the system and an equal amount of work is done on the system satisfies the first law but it does not ensure that the cycle will actually occur. The second law emphasizes the fact that processes proceed in a certain direction but not in the opposite direction. For instance, a hot cup of coffee cools by virtue of heat transfer to the surroundings but heat will not flow from the cooler surroundings to the hot cup of coffee. Thus, real processes proceed only in one direction. It is important to take note of the fact that a cycle will occur *only* if both the first and the second laws of thermodynamics are satisfied.

The first law of thermodynamics is a general statement of the conservation of energy. It makes no distinction between the different forms of energy. The second law of thermodynamics asserts that thermal energy is different from all other forms of energy. Various forms of energy can be converted into thermal energy spontaneously and completely, whereas the reverse transformation is never complete. The impossibility of converting heat completely into mechanical energy forms the basis of Kelvin-Planck statement of second law. The fact that work may be dissipated completely into heat whereas heat may not be converted entirely into work expresses the essential one sidedness of nature.

The basis of second law lies in the difference between the nature of mechanical energy and the nature of internal energy. Mechanical energy is the energy of ordered motion of a body. Internal energy is the energy of random motion of molecules within the body. When the body moves, the motion is due to the ordered motion of molecules within it as a whole in the direction of velocity of the body. The energy associated with this ordered motion of molecules is the kinetic energy. The kinetic and potential energies associated with the random motion is the internal energy. When a moving body comes to rest due to friction, the ordered portion of the kinetic energy becomes converted into energy of random molecular motion. It is impossible to reconvert the energy of random motion completely to the energy of ordered motion, since we cannot control the motions of individual molecules. We can convert only a portion of it. That is what a heat engine does.

16.17 ENTROPY

The differential form of the first law of thermodynamics is written as

$$dQ = dU + dW$$

The work dW done depends on the path of process and therefore it is not a function of the state. The same is the case with dQ , the quantity of heat supplied or taken away. The work dW can be expressed in terms of thermodynamic variables and their changes. For instance, we have expressed dW in equ.(16.2) as

$$dW = P dV$$

It is found that dQ can also be expressed in a similar fashion, in case of reversible processes. We write

$$dQ = T dS \quad (16.37)$$

where dS is called the **change in entropy** and T is the temperature. Now, we define the change in entropy as

$$dS = \frac{dQ}{T} \quad \text{Reversible process} \quad (16.38)$$

The change in entropy dS in the course of an infinitesimal change is equal to the quantity of heat dQ divided by the absolute temperature T , where dQ is the heat absorbed (or rejected) when the change is carried out in a reversible manner.

The total entropy change in a reversible process may be obtained by integrating equ.(16.38). Thus,

$$\Delta S = S_2 - S_1 = \int_1^2 \frac{dQ}{T} \quad (16.39)$$

where S_1 and S_2 are the entropies of the initial and final states of the system.

The importance of the entropy S is that it is a function of state like the internal energy U . Both these parameters depend only on the initial and final states of the system and not on the path of the process that takes the system from the initial state to final state. Equation (16.39) assumes a simpler form when the process is an isothermal process. As T is constant in isothermal process, equ.(16.39) may be written as

$$\Delta S = \int_1^2 \frac{dQ}{T} = \frac{1}{T} \int_1^2 dQ = \frac{Q}{T}$$

Thus,

$$\Delta S = \frac{Q}{T} \quad \text{Isothermal Process} \quad (16.40)$$

The units of entropy and entropy change are J/K .

In practice, the value of entropy S is not of much interest. We have to know the change in entropy when the system changes from one state to another.

16.17.1 Entropy, Disorder and Second Law

When processes occur, in general, they are irreversible and the degree of disorder increases as a result of these processes. As an example, let us take the case of isothermal expansion of an ideal gas (Fig. 16.10). As the gas absorbs heat, it slowly expands. At the end of the process the gas occupies a greater volume than at the beginning. The gas molecules are more disordered now. The gas will not, by its own accord, give up its thermal energy and segregate itself to confine to the initial volume. We, thus, observe that the flow of heat takes place in the direction that increases the amount of disorder. The same type of order to disorder change occurs when free expansion of gas occurs, when one gas diffuses into another, and in similar other spontaneous processes.

Rudolf Clausius (1822-1888), the German physicist, introduced the quantity entropy which is regarded as a measure of disorder in a system. An increase in disorder is equivalent to an increase in entropy. Irreversible processes are processes for which entropy increases.

These considerations led Clausius to reformulate the second law of thermodynamics in terms of entropy. According to it, the entropy of an isolate system always tends to increase. Mathematically, it is expressed as

$$\Delta S_{\text{isolated system}} \geq 0 \quad (16.41)$$

16.17.2 Some Interesting Points

1. The net change in entropy in any reversible cycle is zero.

Let us take the case of Carnot cycle as an example. There is no change in entropy of the working substance during the two adiabatic paths. Either during adiabatic expansion or compression, $Q = 0$. Therefore, $S = 0$. However, there is an increase in entropy during isothermal expansion, as heat Q_H is added at a constant temperature T_H . The consequent increase in entropy $\Delta S_1 = \frac{Q_H}{T_H}$. There is a decrease in entropy during the isothermal compression in which heat Q_L is rejected at a temperature T_L . Thus, $\Delta S_2 = -\frac{Q_L}{T_L}$.

\therefore The net change in entropy is given by

$$\Delta S = \Delta S_1 + \Delta S_2 = \frac{Q_H}{T_H} - \frac{Q_L}{T_L}$$

But

$$\frac{Q_H}{T_H} = \frac{Q_L}{T_L}$$

\therefore

$$\Delta S = 0$$

Reversible Cycle (16.42)

2. Entropy increases in all irreversible processes.

It is proved that there is a net increase in the entropy during irreversible processes. Since all real processes taking place in the universe are irreversible, there is a continuous increase

in its entropy. For this reason, **entropy is not conserved**. In this respect entropy differs from energy.

We can illustrate this by taking a simple example. Suppose a small quantity of heat dQ is radiated away from a hot body A, at a temperature T_H , to a cold body B at a temperature T_C . Let dQ be so small that T_H and T_C are not altered appreciably, due to the exchange of heat. However, the entropy of A decreases by $-dQ/T_H$ whereas that of B increases by dQ/T_C in this process.

$$\therefore \Delta S = \frac{dQ}{T_C} - \frac{dQ}{T_H} \quad \text{As } T_H > T_C \quad \Delta S > 0 \quad (16.43)$$

3. Entropy indicates the direction in which processes proceed in nature.

All natural processes are irreversible. They proceed in the direction of increasing entropy.

4. Entropy represents the unavailability of energy.

In the thermodynamic sense, entropy is a measure of the capability to do work or transfer heat. A system at a higher temperature will tend to do work on and/or transfer heat to its lower temperature surroundings. In the process, the entropy of the system increases and the greater the entropy the **less available** is the energy.

Let us consider the example of Carnot engine. The efficiency of Carnot engine is given by

$$\eta = 1 - \frac{T_L}{T_H}$$

As Q_H is the heat input, heat converted into work = $Q_H \left(1 - \frac{T_L}{T_H}\right)$

\therefore Heat unavailable for work = $Q_H = T_L$

But Q_H / T_H represents the increase in entropy ΔS during isothermal expansion.

$$\therefore \text{Energy wasted} = T_l \Delta S \quad (16.44)$$

If T_L is constant, the amount of energy wasted is proportional to the increase in entropy.

16.18 THIRD LAW OF THERMODYNAMICS

With a decrease in temperature, a greater degree of order prevails in any system. If we could cool a system to 0K, the maximum conceivable order would be established in the system and the minimum entropy would correspond to this state. Now, suppose we apply a pressure on the system at 0K. What does happen to the entropy of the system? On the basis of experiments conducted at low temperatures, W.Nernst concluded that “at 0K, any change in the state of a system takes place without a change in the entropy”. This is called **Nernst’s theorem**. It is also called the **third law of thermodynamics**. Third law of thermodynamics is sometimes known as the principle of unattainability of absolute zero. It is stated as follows:

It is impossible to attain a temperature of 0K.

QUESTIONS

1. Show that work is a path function, and not a property.
2. State the first law for a closed system undergoing a change of state.
3. What is thermodynamics? State the first, second and third laws of thermodynamics and discuss their significance. **(Andhra Univ.)**
4. What is a cyclic heat engine?
5. Define the thermal efficiency of a heat engine cycle. Can this be 100%?
6. Give the Kelvin-Planck statement of the second law.
7. Give the Clausius statement of the second law.
8. What is a reversible process? A reversible process should not leave any evidence to show that the process had ever occurred. Explain.
9. All spontaneous processes are irreversible. Explain.
10. Distinguish between a reversible and an irreversible process. Illustrate your answer by some examples. **(Andhra Univ.)**
11. What is a Carnot Cycle? What are the four processes which constitute the cycle?
12. Describe the different operations involved in a Carnot's cycle. Derive the efficiency of a Carnot engine in terms of source and sink temperatures. **(Andhra Univ.)**
13. What is a reversed heat engine?
14. Show that the efficiency of a reversible engine operating between two given constant temperatures is the maximum.
15. How does the efficiency of a reversible engine vary as the source and sink temperatures are varied? When does the efficiency become 100%?
16. What do you understand by the entropy principle?
17. Show that the transfer of heat through a finite temperature difference is irreversible.
18. Show that the adiabatic mixing of two fluids is irreversible.
19. What is the maximum work obtainable from two finite bodies at temperatures T_1 and T_2 ?
20. Why is the second law called a directional law of nature?
21. Give the Nernst statement of the third law of thermodynamics.

CHAPTER

17

Thermoelectricity

17.1 INTRODUCTION

Thermoelectricity refers to a class of phenomena in which a temperature difference creates an electric potential or an electric potential creates a temperature difference. The term refers collectively to the Seebeck effect, Peltier effect and the Thomson effect. The principle of thermoelectric effects has been known for over 100 years. It is only recently that practical applications have become more viable. Thermoelectricity is a unique part of solid state technology.

The thermoelectric effect is the direct conversion of temperature differences to electric voltage and vice versa. Joule heating, the heat that is generated whenever a voltage difference is applied across a resistive material, is somewhat related, though it is not generally termed a thermoelectric effect. The Peltier–Seebeck and Thomson effects are reversible, whereas Joule heating is not.

This effect can be used to generate electricity, to measure temperature, to cool objects, or to heat them. Because the direction of heating and cooling is determined by the sign of the applied voltage, thermoelectric devices make very convenient temperature controllers.

W. Thomson provided the first important theoretical basis for thermoelectricity. He related, mathematically, the Peltier and Seebeck effects and predicted the existence of the Thomson effect.

17.2 SEEBECK EFFECT

In 1821 Thomas Johann Seebeck discovered that an electromotive force appears in a circuit composed of two dissimilar metals, if the junctions between the metals are held at different temperatures (Fig.17.1). The e.m.f. is known as *thermoelectric e.m.f.* The thermoelectric e.m.f. causes a continuous current in the conductors if they form a complete loop and the current is known as *thermoelectric current*. This phenomenon is called the Seebeck effect. The voltage (thermoelectric e.m.f.) created is of the order of several microvolts per Kelvin difference.

The thermoelectric e.m.f. will exist and the current will flow in the circuit as long as the two junctions, known as the “hot” junction and “cold” junction, are at different temperatures. Thus, the Seebeck effect is the conversion of temperature differences

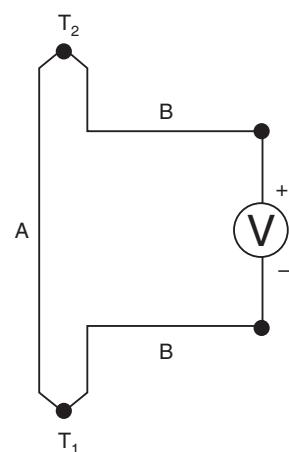


Fig. 17.1: Seebeck Effect

directly into electricity. The magnitude and direction of thermoelectric current is affected by the types of metals used, and the temperature difference between the hot and cold ends; and does not depend on the temperature distribution along the conductors.

The voltage developed in the circuit, Fig.17.1, is proportional to the temperature difference between the two junctions.

$$V = \alpha(T_2 - T_1) \quad (17.1)$$

where $\alpha = \alpha_B - \alpha_A$. α_A and α_B are known as the **Seebeck coefficients** of the metals *A* and *B* and T_1 and T_2 are the temperatures of the two junctions.

Seebeck effect is observed not only in metals but as well in semiconductors also. It is not necessarily a junction phenomenon, but arises in a single conductor also. If a temperature gradient (difference) is caused in a conductor, electrons diffuse from the hot side to the cold side. Electrons migrating to the cold side leave behind their oppositely charged and immobile nuclei on the hot side and thus give rise to a thermoelectric voltage.

17.2.1 Origin of the Thermoelectric e.m.f.

If a temperature gradient (difference) is caused in a conductor, electrons diffuse from the hot side to the cold side. Electrons migrating to the cold side leave behind their oppositely charged and immobile nuclei at the hot side and thus give rise to a thermoelectric voltage. Since a separation of charges also creates an electric potential, the buildup of charged carriers onto the cold side eventually ceases at some maximum value since there exists an equal amount of charged carriers drifting back to the hot side as a result of the electric field at equilibrium. Only an increase in the temperature difference can resume a buildup of more charge carriers on the cold side and thus lead to an increase in the thermoelectric voltage.

17.2.2 Seebeck Coefficient

Seebeck coefficient or *thermopower* of a material measures the magnitude of an induced thermoelectric voltage in response to a temperature difference across that material. It is defined as the open circuit voltage produced between two points on a conductor, where a uniform temperature difference of 1K exists between those points.

If the temperature difference ΔT between the two ends of a material is small, then the *thermopower* or *Seebeck coefficient* of a material may be written as

$$\alpha = \frac{\Delta V}{\Delta T} \quad (17.2)$$

This can also be expressed in terms of the electric field E and the temperature gradient ∇T , as

$$\alpha = \frac{E}{\nabla T} \quad (17.3)$$

The thermopower has units of (V / K), though in practice it is more common to use microvolts per Kelvin.

The thermopower is an important material parameter that determines the efficiency of a thermoelectric material. A larger induced thermoelectric voltage for a given temperature gradient will lead to a larger efficiency. Ideally one would want very large thermopower values since only a small amount of heat is then necessary to create a large voltage. This voltage can then be used to provide power.

17.3 THERMOCOUPLE

The Seebeck effect is commonly used in a device called a *thermocouple*, a widely used method of temperature measurement. A pair of junctions formed by two dissimilar metals constitutes a thermocouple. It is used to measure a temperature difference directly or to measure an

absolute temperature by setting one end to a known temperature, say 0°C (Fig. 17.2).

17.4 THERMOELECTRIC SERIES

Seebeck arranged the metals into the following series, known as *Seebeck series*.

Bi, Ni, Co, Pd, U, Cu, Mn, Tl, Hg, [Pb], Sn, Cr, Mo, Rh, Ir, Au, Zn, W, Cd, Fe, As, Sb, Te.

The metals to the left side of Pb are called *thermoelectrically positive metals*, while those to its right side are called *thermoelectrically negative metals*.

If a thermocouple is constructed with any two metals in this series, the thermo current flows across the hot junction from the metal appearing earlier in the series to the one appearing later. The magnitude of the thermo e.m.f. is of the order of a few microvolts per Kelvin temperature difference between the two junctions. The greater the separation of the two metals in the series, the higher the e.m.f. for a given temperature difference between the junctions. For example, the thermo e.m.f. for Ni-Fe couple is greater than for Cu-Fe couple. The direction of the current will be from a metal occurring earlier in the series to a metal occurring later in the series through the cold junction. For example, in Cu-Fe thermocouple, the current flows from Cu to Fe through the cold junction and from Fe to Cu through the hot junction.

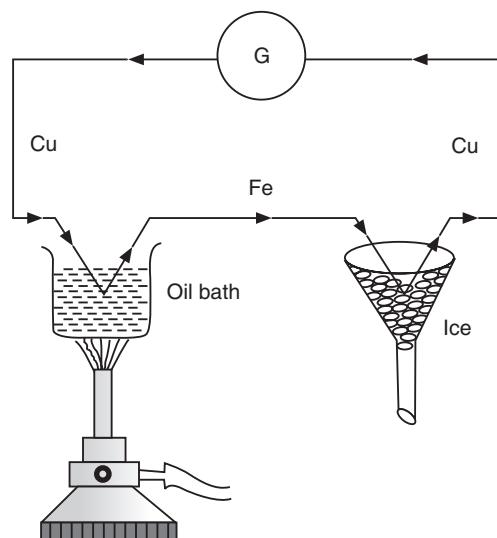


Fig. 17.2: Cu – Fe Thermocouple used to measure the temperature of oil bath.

17.5 VARIATION OF THERMOELECTRIC E.M.F. WITH TEMPERATURE

If the temperature of the cold junction of a thermocouple is kept at 0°C and the thermoelectric e.m.f. ‘e’ is plotted against the temperature T of the hot junction, we obtain a parabolic curve, as shown in Fig. 17.3.

It is seen that the thermo e.m.f. increases with the temperature of the hot junction and becomes a maximum at a particular temperature, T_n . T_n is known as the **neutral temperature** which is a constant for the given pair of metals forming the thermocouple. *The temperature of the hot junction at which maximum thermo e.m.f. flows is a constant for a given couple and is known as neutral temperature T_n for that couple.* If the temperature of the hot junction is increased beyond the neutral temperature, the e.m.f. decreases and becomes zero at a temperature T_i , known as the **inversion temperature**. *The temperature at which the thermo e.m.f. is zero, is known as inversion temperature.* Beyond the temperature of inversion, the e.m.f. again increases but in the reverse direction.

The thermo e.m.f. varies with temperature according to the following relation.

$$e = at + \frac{1}{2}bt^2 \quad (17.4)$$

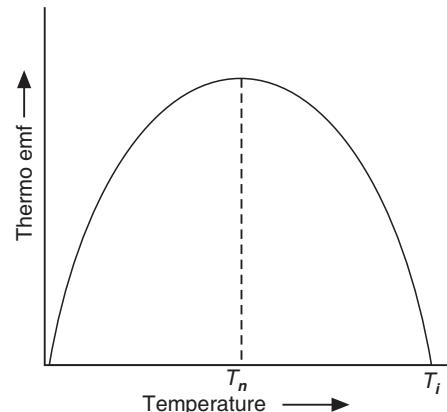


Fig. 17.3

where a and b are *Seebeck constants* for the thermocouple. Equ.(17.4) is known as *Seebeck equation*, and $t = T_2 - T_1$.

Differentiation of equ.(17.4) gives

$$\frac{de}{dT} = a + bt \quad (17.5)$$

At $T = T_n$, $\frac{de}{dT} = 0$ and e is a maximum. Therefore,

$$\begin{aligned} 0 &= a + bT_n \\ \therefore T_n &= -\frac{a}{b} \end{aligned} \quad (17.6)$$

At $T = T_i$, $e = 0$. Therefore, it follows from equ.(17.4) that

$$\begin{aligned} 0 &= aT_i + \frac{1}{2}bT_i^2 \\ \text{or } T_i(a + \frac{1}{2}bT_i) &= 0 \\ \therefore T_i &= -\frac{2a}{b} \end{aligned} \quad (17.7)$$

It follows from equ.(17.6) and (17.7) that

$$T_i = 2T_n \quad (17.8)$$

Example 17.1. Calculate the inversion temperature for an iron-cadmium thermocouple. The value of a and b are 17.5 and -0.1 for iron and for cadmium these values are 3 and $+0.08$ respectively. (Bombay Univ., 96)

Solution:

$$\begin{aligned} a_{\text{iron} - \text{cadmium}} &= a_{\text{iron}} - a_{\text{cadmium}} = 17.5 - 3 = 14.5 \\ b_{\text{iron} - \text{cadmium}} &= b_{\text{iron}} - b_{\text{cadmium}} = -0.1 - 0.08 = -0.18 \end{aligned}$$

$$\text{Inversion temperature } T_i = -\frac{a_{Fe-cd}}{b_{Fe-cd}} = -\frac{14.5}{-0.18} = 80.55^\circ\text{C}$$

Example 17.2. EMF of a thermocouple is $1200 \mu\text{V}$, when working between 0°C and 100°C . Its neutral temperature is 300°C . Find the values of a and b for it. (Bombay Univ., 95)

Data: Emf of a thermocouple $e = 1200 \mu\text{V} = 1200 \times 10^{-6} \text{ V}$

Solution:

$$T = 100^\circ\text{C}$$

Neutral temperature

$$T_n = 300^\circ\text{C}$$

$$e = aT + bT^2$$

$$(1200 \times 10^{-6}) \text{ V} = a(100) + b(100)^2$$

$$\text{i.e., } a + 100b = 12 \times 10^{-6} \quad (i)$$

$$T_n = -\frac{a}{2b}$$

$$300^\circ\text{C} = -\frac{a}{2b}$$

$$\therefore a = -600b \quad (ii)$$

$$\text{From relation (i), } -600b + 10b = 12 \times 10^{-6}$$

$$\text{i.e., } b = \frac{12}{5} \times 10^{-8} = 2.4 \times 10^{-8}$$

$$\text{From relation (ii)} a = (-600)(-2.4 \times 10^{-8}) = 14.4 \times 10^{-6}$$

Example 17.3. The thermo e.mf of a Cu-Fe thermocouple is $2160 \mu\text{V}$ when the cold junction is at 0°C and the hot junction at 250°C . Calculate the constants a and b if the neutral temperature is 330°C .
(Bombay Univ.,94)

Solution:

$$e = aT + bT^2$$

$$2160 \times 10^{-6} \text{ volts} = a(250) + b(250)^2$$

$$\therefore a + 250b = 8.64 \times 10^{-6} \quad (i)$$

Neutral temperature

$$T_n = -\frac{a}{2b}$$

$$\therefore a = T_n(-2b) = -660b \quad (ii)$$

Using (ii) in (i), we get

$$-660b + 250b = 8.64 \times 10^{-6}$$

$$b = -0.02107 \times 10^{-6}$$

From (ii)

$$a = 660(-0.02107 \times 10^{-6})$$

$$a = 13.91 \times 10^{-6}$$

Example 17.4. For Fe-Cu thermocouple it is observed that the thermo e.m.f is zero when one of the junction is at 20°C and the other one is at some higher temperature. If the neutral temperature is 285°C , calculate the higher temperature. Hence find out the temperature of inversion, if the cold junction temperature is at -20°C .
(Bombay Univ., 95)

Solution: For Fe – Cu thermocouple,

$$\text{e.mf } e = 0 \text{ at } T_1 = 20^\circ\text{C} \text{ and } T_2 = T \text{ (say)}$$

$$e = 0 = a(T_2 - T_1) + 2b(T_2 - T_1)$$

$$\therefore a(T_2 - T_1) + 2b(T_2 - T_1) = 0$$

$$\text{i.e., } (T_2 - T_1) = \frac{-a}{2b}$$

We know that

$$-a/2b = T_n \quad (\text{neutral temperature})$$

$$\therefore T_2 - T_1 = T_n$$

$$\therefore T_2 = T_n + T_1 = 285^\circ\text{C} + 20^\circ\text{C}$$

$T_2 = 305^\circ\text{C}$ is the higher temperature. If cold junction temperature $T_1 = -20^\circ\text{C}$, then the neutral temperature $T_n = T_2 - T_1 = 305^\circ\text{C} - (-20^\circ\text{C})$

$$\text{or } T_n = 325^\circ\text{C}$$

$$\therefore \text{Inversion temperature } T_i = 2T_n = 650^\circ\text{C}$$

17.6 THE PELTIER EFFECT

In 1834 Peltier discovered that when electric current is passed in a circuit consisting of two dissimilar metals, heat is evolved at one junction and is absorbed at the other junction. This is known as **Peltier effect**. It is the inverse of the Seebeck effect. The Peltier effect is a junction phenomenon.

There is heat absorption or generation at the junctions depending on the direction of current flow. Heat generated by current flowing in one direction was absorbed if the current was reversed.

As an example, consider the circuit Fig.17.4. Under these conditions it is observed, as indicated in the diagram, that the right-hand junction is heated, showing that electrical energy is being transformed into heat energy. Meanwhile, heat energy is transformed into electrical

energy at the left junction, thereby causing it to be cooled. When the current is reversed, heat is absorbed at the right junction and produced at the left one.

The Peltier effect is found to be proportional to the first power of the current.

17.6.1 The Peltier Coefficient

The *Peltier coefficient* is defined as the amount of heat energy absorbed or evolved at the junction of two dissimilar metals when one ampere of current flows through it for one second. It is denoted by π and expressed in volts. It is a property that depends on both materials of the junction.

The heat absorbed per second at a junction carrying a current I amperes is given by

$$\text{Heat absorbed/sec} = \pi_{ab} I \quad (17.9)$$

$$\therefore \text{Heat absorbed in } t \text{ seconds}, H = \pi_{ab} I t = \pi_{ab} q.$$

where the current is from metal a to metal b . The junction emf, π_{ab} , is known as the *Peltier coefficient*.

$$\pi_{ab} = \frac{H}{I t} \quad (17.10)$$

π_{ab} is positive if metal a is positive with respect to metal b (thus π_{Cu-Fe} is positive). The magnitude of π_{ab} is a function of the temperature of the junction. For identical temperatures $\pi_{ab} = -\pi_{ba}$. Thus, if the direction of the current in equation (17.9) is reversed, the heat absorbed per second is

$$\text{Heat absorbed/sec} = \pi_{ba} I \quad (17.11)$$

which is opposite in sign to equation (17.9).

If V is the potential difference applied, then

$$\text{Heat absorbed} = V_q = VIt$$

Equating the above with equ. (17.9), we get

$$\pi = V \quad (17.12)$$

Thus, the Peltier coefficient is numerically equal to the applied potential difference expressed in volts.

17.7 THE THOMSON EFFECT

When a current flows through an unequally heated metal, there is an absorption or evolution of heat in the body of the metal. In 1851, W. Thomson (later Lord Kelvin) predicted this effect. Therefore, it is known as **Thomson effect**. In Thomson effect we deal with only metallic rod and not with thermocouple as in **Peltier effect** and **Seebeck effect**. Some of the metals show positive Thomson effect while others exhibit negative Thomson effect.

(i) **Positive Thomson effect:** Positive Thomson effect is shown by Cu, Sn, Ag, Cd, Zn... etc metals. In metals such as zinc and copper, hotter end will be at a higher potential and cooler end at a lower potential. When current moves from the hotter end to the colder end, it is moving from a high to a low potential, so there is evolution of heat. Heat is absorbed when

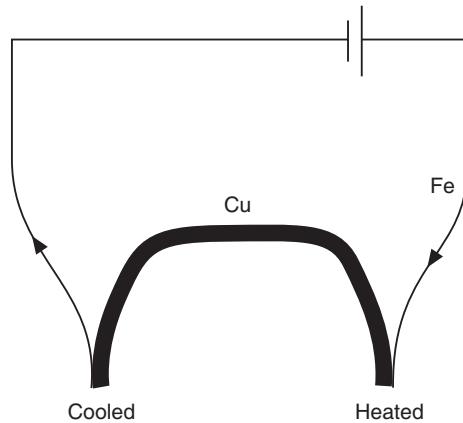


Fig. 17.4: Peltier Effect

current is passed from colder end to hotter end, as shown in Fig. 17.5(a). This is called the positive Thomson effect.

(ii) Negative Thomson effect:

In metals such as cobalt, nickel, and iron, the cooler end will be at a higher potential and hotter end at a lower potential. When current moves from the hotter end to the colder end, it is moving from a low to a high potential. So there is absorption of heat. Heat is evolved when current is passed from colder end to the hotter end (Fig. 17.5 b). This is called the negative Thomson effect. The metals which show negative Thomson effect are Fe, Co, Bi, Pt, Hg... etc.

Now let us consider a circuit consisting of two metals A and B as shown in Fig. 17.6. The hot junction is at a temperature T_2 °K and the cold junction is at a temperature T_1 °K. Due to the Seebeck effect, i.e., due to temperature difference between the junctions, thermoelectric current flows through the circuit. As the current flows through the hot and cold junctions, heat is absorbed at the hot junction and evolved at the cold junction due to Peltier effect.

Let π_1 and π_2 be the Peltier coefficients at T_1 and T_2 . During the passage of current an amount of heat energy equal to $\pi_2 q$ is absorbed at hot junction and heat energy $\pi_1 q$ is evolved at the cold junction. Then, the energy $(\pi_2 - \pi_1)q$ is used in driving the current through the circuit. As π_2 and π_1 are equal to the potential differences at hot and cold junctions respectively, the thermo e.m.f. developed is given by

$$e = (\pi_2 - \pi_1)q \quad (17.13)$$

The current in the circuit is small, and the joules heating effect is negligible. As the Peltier effect is reversible, a thermocouple may be regarded as a reversible heat engine taking heat from the source at the hot junction at temperature T_2 , does work in driving the current through the circuit, and rejecting heat to the sink, the cold junction at temperature T_1 . By the Carnot's engine we have

$$\frac{Q_2}{T_2} = \frac{Q_1}{T_1} \quad (17.14)$$

Now during the flow of current in the thermocouple, heat absorbed at the hot junction is $Q_2 = \pi_2 q$ joules while the energy given out to the sink is $Q_1 = \pi_1 q$ joules.

$$\frac{\pi_2 q}{T_2} = \frac{\pi_1 q}{T_1} \quad \therefore \frac{\pi_2}{\pi_1} = \frac{T_2}{T_1} \quad \therefore \frac{\pi_2}{\pi_1} - 1 = \frac{T_2}{T_1} - 1$$

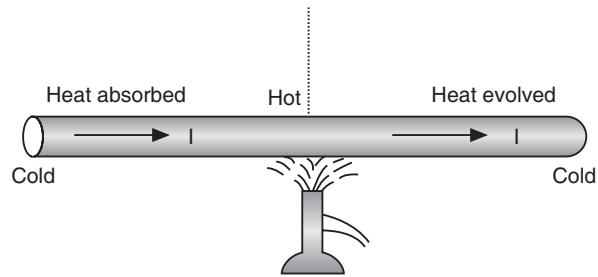


Fig. 17.5(a)

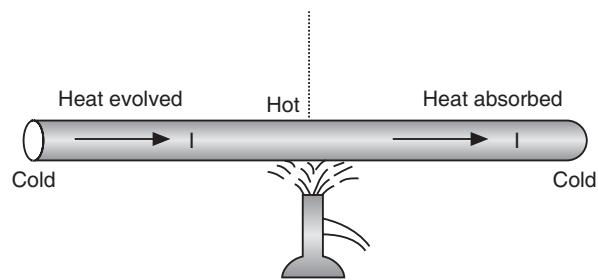


Fig. 17.5(b)

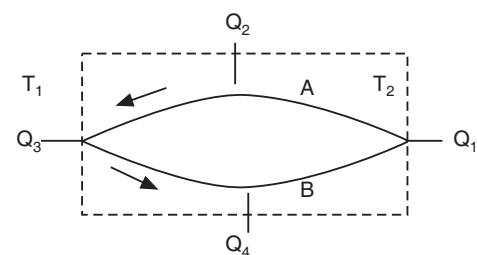


Fig. 17.6

or

$$\frac{\pi_2 - \pi_1}{\pi_1} = \frac{T_2 - T_1}{T_1} \quad (17.15)$$

or

$$\pi_2 - \pi_1 = \pi_1 \left(\frac{T_2 - T_1}{T_1} \right)$$

$$\text{But } \pi_2 - \pi_1 = e. \quad \therefore e = \frac{\pi_1}{T_1} (T_2 - T_1) \quad (17.16)$$

If the cold junction temperature is held constant, the Peltier coefficient π_1 will also be constant. Then,

$$e \propto (T_2 - T_1) \quad (17.17)$$

The above relation shows that the thermoelectric e.m.f. is directly proportional to the difference in temperatures between hot and cold junctions. If a graph is plotted between e and $(T_2 - T_1)$, it would be a straight line. Experiments show that the graph is a parabola. It means that Peltier effect alone cannot account for thermo e.m.f. in a thermocouple. Thomson predicted that there is some additional source of e.m.f. besides the Peltier effect. He concluded that this additional e.m.f. is due to the temperature gradient in the individual metals.

Thus, Thomson effect is the absorption or evolution of energy in a single current carrying conductor subjected to a temperature gradient.

It can be interpreted as follows. Consider a material with a current flowing through it and a temperature gradient applied to it. In this situation, thermal energy is generated (or absorbed) *all along* the sample. The reason for this is that the Seebeck coefficient is different at different places along the sample. So the sample can be thought of as a series of many small Peltier junctions, each of which is generating (or absorbing) heat.

The Thomson coefficient is unique among the three main thermoelectric coefficients because it is the only thermoelectric coefficient directly measurable for individual materials. The Peltier and Seebeck coefficients can only be determined for pairs of materials. Thus, there is no direct experimental method to determine an absolute Seebeck coefficient (i.e. thermopower) or absolute Peltier coefficient for an individual material.

17.7.1 The Thomson Coefficient

According to Thomson effect, there is a variation of potential along a metal when its different parts are kept at different temperatures. Consider two points in a conductor which are very close together and at temperatures T and $T + dT$. The difference of potential between the two points may be expressed as σdT , where σ is the Thomson coefficient. Hence, *Thomson coefficient of a metal is numerically equal to the difference in potential between two parts whose temperature difference is 1°C*.

Consider a material with a current flowing through it and a temperature gradient applied to it. In this situation, thermal energy is generated (or absorbed) all along the sample. The energy absorbed or evolved in a time ' t ' seconds is $(\sigma dT)I t$.

Thomson coefficient may then be expressed as the ratio of the power evolved per unit volume in the sample to the applied current and temperature gradient.

$$\sigma = \frac{P}{IdT} \quad (17.18)$$

The amount of heat energy absorbed or evolved per second between two points of a conductor having a unit temperature difference, when a unit current is passed is known as **Thomson coefficient** for the material of a conductor.

17.8 E.M.F. IN A THERMOCOUPLE

Let us consider a thermocouple made of two dissimilar metals A and B as shown in Fig. 17.7. Let the hot junction be at absolute temperature T_2 and the cold be at T_1 . Let π_1 and π_2 be the Peltier coefficients at the cold and hot junctions respectively directed from metal B to metal A through the junctions. Then, the e.m.f. due to Peltier effect is in the anticlockwise direction and is equal to $(\pi_2 - \pi_1)$.

Let σ_A and σ_B be the Thomson coefficients of the metals A and B respectively, which are assumed to be positive. Then, the Thomson e.m.f. in the metal A in the anticlockwise direction is

$$E_A = - \int_{T_1}^{T_2} \sigma_A dT \quad (17.19)$$

Similarly, Thomson e.m.f. in metal B in anticlockwise direction is

$$E_B = + \int_{T_1}^{T_2} \sigma_B dT \quad (17.20)$$

The resultant e.m.f. in the circuit is

$$\begin{aligned} E_{AB} &= (\pi_2 - \pi_1) - \int_{T_1}^{T_2} \sigma_A dT + \int_{T_1}^{T_2} \sigma_B dT \\ \therefore E_{AB} &= (\pi_2 - \pi_1) - \int_{T_1}^{T_2} (\sigma_A - \sigma_B) dT \end{aligned} \quad (17.21)$$

17.9 THE THERMOELECTRIC POWER

The rate of change of e.m.f. with temperature is called *thermoelectric power* and is denoted by P . Thus,

$$P = \frac{de}{dT} \quad (17.22)$$

Let us consider a thermocouple made of two dissimilar metals A and B with their junctions at temperatures $(T + dT)$ and T . The Peltier coefficient at temperature $(T + dT)$ is $(\pi + d\pi)$ and at temperature T , it is π . Let σ_A and σ_B be the Thomson coefficients of the metals A and B respectively, which are assumed to be positive.

Current flows in the thermocouple circuit due to Seebeck effect. Heat is absorbed at the hot junction and is evolved at the cold junction due to the Peltier effect. And as temperature gradient exists along A and B, heat is evolved in A and is absorbed in B, due to the Thomson effect.

If a unit charge passes through the circuit, the energy absorbed due to Peltier effect at the hot junction $= (\pi + d\pi)$.

Energy evolved due to Peltier effect at the cold junction $= -\pi$

Energy absorbed in metal B due to Thomson effect $= \sigma_B dT$

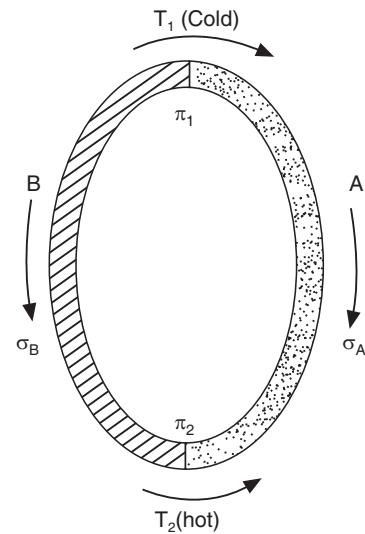


Fig. 17.7

$$\begin{aligned}
 \text{Energy evolved in metal } A \text{ due to Thomson effect} &= -\sigma_A dT \\
 \text{The total gain of energy is} & de = (\pi + d\pi) - \pi + \sigma_B dT - \sigma_A dT \\
 \therefore & de = d\pi - (\sigma_A - \sigma_B) dT \\
 \therefore \text{Thermoelectric power, } P &= \frac{de}{dT} = \frac{d\pi}{dT} + (\sigma_B - \sigma_A)
 \end{aligned} \tag{17.23}$$

17.10 RELATION BETWEEN PELTIER COEFFICIENT AND THERMOELECTRIC POWER

The Peltier and Thomson effects are reversible and the joules heating can be neglected in a thermocouple. Therefore, the thermocouple may be considered as identical to a reversible engine and the second law of thermodynamics can be applied to it.

Let us consider a thermocouple made of Cu-Fe junctions (Fig. 17.8). Let the temperature of the hot junction be $T + dT$ and that of cold junction be T . Further, let the Peltier coefficient of the hot junction be $\pi + d\pi$ and that of cold junction be π . If unit current passes for unit time through the thermocouple, then the heat absorbed at hot junction will be $\frac{\pi + d\pi}{J}$ and heat given out at the cold junction will be $\frac{\pi}{J}$,

where J is the mechanical equivalent of heat.

$$\text{The heat energy absorbed due to the Thomson effect} = \frac{(\sigma_{Cu} - \sigma_{Fe})}{J} dT \tag{17.24}$$

Using the second law of thermodynamics, $\frac{Q_1}{T_1} = \frac{Q_2}{T_2}$, we obtain

$$\begin{aligned}
 \frac{\pi + d\pi}{T + dT} - \frac{\pi}{T} + \frac{(\sigma_{Cu} - \sigma_{Fe})}{T} dT &= 0 \\
 \frac{T d\pi}{(T + dT)T} - \frac{\pi dT}{(T + dT)T} + \frac{(\sigma_{Cu} - \sigma_{Fe})}{T^2} T dT &= 0
 \end{aligned}$$

As dT is a very small quantity, we can approximate $(T + dT) T \approx T^2$.

$$\therefore \frac{T d\pi}{T^2} - \frac{\pi dT}{T^2} + \frac{(\sigma_{Cu} - \sigma_{Fe})}{T} dT = 0$$

$$\text{or } d\pi - \frac{\pi dT}{T} + (\sigma_{Cu} - \sigma_{Fe}) dT = 0$$

$$\text{or } d\pi + (\sigma_{Cu} - \sigma_{Fe}) dT = \frac{\pi dT}{T} \tag{17.25}$$

But

$$d\pi + (\sigma_{Cu} - \sigma_{Fe}) dT = de$$

$$\therefore de = \frac{\pi dT}{T}$$

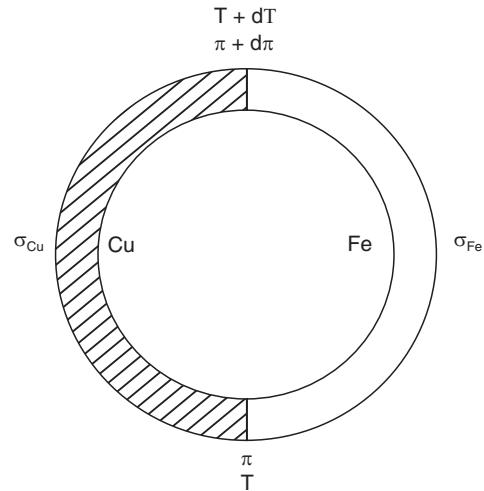


Fig. 17.8

or $\pi = T \cdot \frac{de}{dT}$ (17.26)

or $\pi = TP$ (17.27)

The above equation shows that the Peltier coefficient of a thermocouple is equal to the product of the absolute temperature of the junction and the thermoelectric power at that temperature.

Example 17.5: The e.m.f. in micro volts (e) of a thermocouple, one junction of which is at 0°C e is given by $e = 1600 T - 4 T^2$ where $T^\circ\text{C}$ is the temperature of hot junction. Find (i) neutral temperature (ii) Peltier coefficient.

Solution: (i) $e = aT + bT^2 = 1600 T - 4 T^2$

$$\therefore a = 1600 \text{ and } b = -4$$

$$\therefore \text{Neutral temperature, } T_n = \frac{-a}{2b} = \frac{-1600}{2(-4)}$$

$$\therefore T_n = 200^\circ\text{C}$$

(ii) Peltier coefficient $\pi = T \frac{de}{dT} = T(a + 2bT)$

$$\therefore \pi = (T)(1600 - 8T) = 1600 T - 8 T^2$$

17.11 RELATION BETWEEN THOMSON COEFFICIENT AND THERMOELECTRIC POWER

From equ.(17.25), we have

$$d\pi + (\sigma_{\text{Cu}} - \sigma_{\text{Fe}}) dT = \frac{\pi dT}{T}$$

$$(\sigma_{\text{Cu}} - \sigma_{\text{Fe}}) = \frac{\pi}{T} - \frac{d\pi}{dT} \quad (17.28)$$

Using equ. (17.26), we may write equ. (17.28) as

$$(\sigma_{\text{Cu}} - \sigma_{\text{Fe}}) = \frac{de}{T} - \frac{d\pi}{dT}$$

Differentiating equ. (17.26), we obtain

$$\frac{d\pi}{dT} = \frac{de}{dT} + T \frac{d^2e}{dT^2}$$

$$\therefore \frac{de}{dT} - \frac{d\pi}{dT} = -T \frac{d^2e}{dT^2} \quad (17.29)$$

$$\therefore (\sigma_{\text{Cu}} - \sigma_{\text{Fe}}) = -T \frac{d^2e}{dT^2} \quad (17.30)$$

or $(\sigma_{\text{Fe}} - \sigma_{\text{Cu}}) = T \frac{d^2e}{dT^2}$ (17.31)

If we consider a thermocouple of copper and lead, the relation (17.31) becomes

$$(\sigma_{\text{Cu}} - \sigma_{\text{Pb}}) = T \frac{d^2e}{dT^2} \quad (17.32)$$

But for lead, $\sigma_{Pb} = 0$.

$$\therefore \sigma_{Cu} = T \frac{d^2 e}{dT^2} \quad (17.33)$$

Similarly, for a thermocouple of iron and lead

$$\sigma_{Fe} = T \frac{d^2 e}{dT^2} \quad (17.34)$$

or

$$\sigma_{Fe} = T \frac{d}{dT} \left(\frac{de}{dT} \right) = T \frac{dP}{dT} \quad (17.35)$$

Thus, Thomson coefficient is the product of absolute temperature and the rate of change of thermoelectric power.

Example 17.6: The thermoelectric power for steel is $18 \mu V / ^\circ C$ at $0^\circ C$ and zero at $400^\circ C$. That for copper is $6 \mu V / ^\circ C$ at $500^\circ C$ and zero at $-50^\circ C$. Find the e.m.f. for steel-copper thermocouple with one junction at its neutral temperature and other at $0^\circ C$.

(Bombay Univ., 96)

Solution: Thermoelectric power for steel

$$\left(\frac{de}{dT} \right)_{T=0^\circ C} = 18 \mu V / ^\circ C = a + 2b (0^\circ C)$$

i.e.

$$a = 18 \times 10^{-6} V / ^\circ C$$

$$\left(\frac{de}{dT} \right)_{T=400^\circ C} = 0 = a + 2b (400^\circ C)$$

\therefore

$$b = \frac{-18 \times 10^{-6} V / ^\circ C}{800^\circ C} = -2.25 \times 10^{-8} \frac{V / ^\circ C}{^\circ C}$$

Thermoelectric power for Cu

$$\left(\frac{de}{dT} \right)_{T=500^\circ C} = 6 \mu V / ^\circ C = a + 2b (500^\circ C)$$

$$\left(\frac{de}{dT} \right)_{T=-50^\circ C} = 0 = a + 2b (-50^\circ C)$$

i.e

$$a = 100 b^\circ C$$

\therefore

$$100 b^\circ C + 100 b^\circ C = 6 \times 10^{-6} V / ^\circ C$$

$$b = 0.545 \times 10^{-8} V / ^\circ C / ^\circ C$$

and

$$a = 0.545 \times 10^{-6} V / ^\circ C$$

For the Steel-Copper thermocouple

$$\begin{aligned} a &= 18 \times 10^{-6} V / ^\circ C - 0.545 \times 10^{-6} V / ^\circ C \\ &= 17.455 \times 10^{-6} V / ^\circ C \end{aligned}$$

$$\begin{aligned} b &= -2.25 \times 10^{-8} V / ^\circ C / ^\circ C - 0.545 \times 10^{-8} V / ^\circ C / ^\circ C \\ &= -2.755 \times 10^{-8} V / ^\circ C / ^\circ C \end{aligned}$$

Neutral temperature $T_n = -\frac{a}{2b} = \frac{-17.455 \times 10^{-6} V/^\circ C}{2 \left(-2.795 \times 10^{-8} \frac{V/^\circ C}{^\circ C} \right)} = 3.122 \times 10^2 {}^\circ C$
 $= 312.2 {}^\circ C$

The e.m.f. for the Steel – Copper thermocouple

$$e = aT_n + bT_n^2 \text{ (according to data)}$$

$$\therefore e = (17.455 \times 10^{-6} V/^\circ C)(312.2 {}^\circ C) + \left(-2.795 \times 10^{-8} \frac{V/^\circ C}{^\circ C} \right) 312.2^2 ({}^\circ C)^2$$

$$e = 272868 \times 10^{-8} V = 2.7 \text{ mV}$$

Example 17.7. The e.m.f. in a thermocouple, one junction of which is kept at $0 {}^\circ C$ is given by $E = a\theta + b\theta^2$. Find the neutral temperature and Peltier and Thomson coefficients.

Solution: Let $\theta {}^\circ$ be the temperature of hot junction

$$0 {}^\circ C = (T - 273) {}^\circ C, \text{ Where } T \text{ is in absolute degrees}$$

$$\therefore E = a\theta + b\theta^2 = a(T - 273) + b(T - 273)^2$$

$$\therefore \frac{dE}{dT} = a + 2b(T - 273)$$

and

$$\frac{d^2E}{dT^2} = 2b$$

$$\text{At neutral temperature} \quad T = T_n$$

$$\text{and} \quad \frac{dE}{dT} = 0$$

$$\therefore 0 = a + 2b(T_n - 273)$$

$$\text{or} \quad T_n - 273 = -\frac{a}{2b}$$

$$\text{or} \quad T_n = \left[\left(273 - \frac{a}{2b} \right) \right] \text{ Absolute}$$

$$\text{Peltier coefficient, } \pi = T \cdot \frac{dE}{dT} = T [a + 2b(T - 273)]$$

$$\text{and Thomson coefficient, } \sigma = T \cdot \frac{d^2E}{dT^2} = 2b \cdot T$$

Example 17.8. The e.m.f. in lead-iron thermocouple, one junction of which is at $0 {}^\circ C$, is given by $E = 1784 t - 2.4 t^2$ (in μ volts) where t is the temperature in ${}^\circ C$. Find the neutral temperature, π and σ .

Solution:

$$E = at + \frac{1}{2}bt^2$$

$$a = 1784, b = -4.8$$

$$\text{Now} \quad t_n = -\frac{a}{b} = -\frac{1784}{-4.8} = 371.67 {}^\circ C$$

$$\text{Peltier coefficient, } \pi = T \cdot \frac{dE}{dT} = T [a + b(T - 273)] = 3094.4T - 4.8T^2$$

and Thomson coefficient,

$$\sigma = T \cdot \frac{d^2 E}{dT^2} = T \cdot \frac{d}{dT} (a + bt) = 3094.4T - 4.8T^2$$

$$\sigma = T \cdot b = -4.8T$$

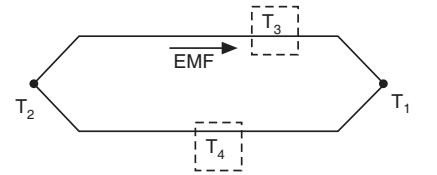
17.12 THE THERMOELECTRIC LAWS

After much investigation of thermoelectric circuits several basic concepts have been conceived. These concepts can be summed up according to three fundamental “laws”.

1. The Law of Homogeneous Circuits

An electric current cannot be sustained in a circuit of a single homogeneous metal, however varying in section, by the application of heat alone.

What this means for thermocouples, is that if there is a temperature distribution along the wires between the hot and cold junctions, the total thermal EMF will be unaffected. Only the temperature at the junction between the two dissimilar metals will have an effect on the EMF produced.



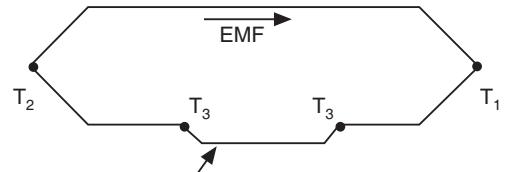
T₃ and T₄ have no effect on the total EMF produced

Fig. 17.9

2. The Law of Intermediate Metals

If two dissimilar metals A and B with their junctions at T₁ and T₂ are joined to a third metal C at one leg, if C is kept at a uniform temperature along its entire length, the total EMF in the circuit will be unaffected.

This comes into play with thermocouples because there is usually a need to introduce extra metals into the circuit. This generally occurs when instrumentation (lead wires) is connected to measure the EMF, or when the junction is welded together on the hot end (weld rod). One would assume that the introduction of these extra undesired “junctions” would destroy the calibration and throw off the EMF measurement. However, this law states that the addition of these extra metals will not have an effect on the total EMF as long as they are kept at the same temperature at the point where they are connected.



This extra connection has no effect on emf as long as it has uniform temperature along its length

Fig. 17.10

3. The Law of Intermediate Temperatures

Let one thermocouple be with its junction at T₁ and some reference temperature, T₂ and another thermocouple be at the same reference temperature, T₂ and the temperature, T₃ to be measured. This is equivalent to a single thermocouple with its junction at T₁ and T₃.

17.13 APPLICATIONS OF THERMOCOUPLE

The Seebeck effect is the basis for the thermocouple, which is used in measurement of temperature and detection of heat radiation.

1. Measurement of Temperature

An unknown temperature can be measured with the help of thermocouple. First the thermocouple is calibrated. For the purpose of calibration, one of the thermocouple junctions is placed in a liquid bath while the other junction is kept in a bath of constant temperature, usually held at 0°C (see Fig. 17.2). As the liquid in the hot bath is heated, the thermo e.m.f. varies. The e.m.f. developed at different temperatures is recorded and a graph is plotted between the

temperature and the corresponding thermo e.m.f. This graph is known as the *calibration graph*. Next, the hot junction is placed in a bath of unknown temperature and the e.m.f. developed in the thermocouple is noted. The temperature corresponding to this e.m.f. is read from the calibration graph. It is the temperature of the bath whose temperature is to be measured.

2. Thermopile

In the detection and measurement of radiation, a thermopile is usually used.

A series-connected array of thermocouples is known as a **thermopile**. It is constructed in order to increase the output voltage since the voltage induced over each individual thermocouple is small. The thermopile was developed by Leopoldo Nobili (1784-1835) and Macedonio Melloni (1798-1854). It was initially used for measurements of temperature and infra-red radiation, but was also rapidly put to use as a stable supply of electricity for other physics experimentation.

The thermopile consists of thermocouples of metals antimony and bismuth placed end to end to enhance the thermoelectric e.m.f. The metals are widely separated in the thermo-electric series and comparatively large e.m.f. is produced for a small difference temperature between the junctions. The bismuth and antimony metals are taken in the form of thick strips which reduces the resistance of the circuit. A galvanometer of resistance equal to the resistance of the thermocouple is connected to the end of the thermocouple (Fig. 17.11). The whole assembly is enclosed in a box and the front surface of the metals is blackened. A conical piece is attached to the front side so that the incoming radiation falls on the blackened surface and increases the temperature of the hot junction. With the help of a thermopile temperature differences up to 0.001°C can be detected.

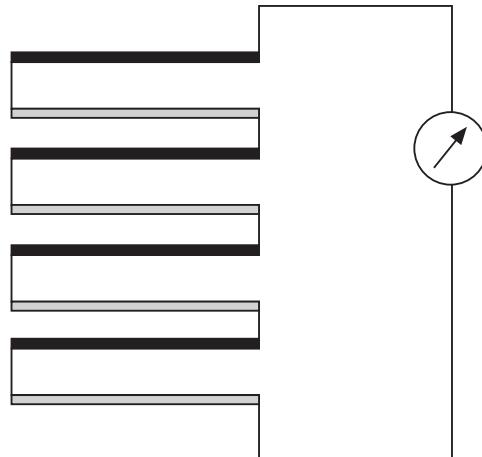


Fig. 17.11: Thermopile

17.14 FIGURE-OF-MERIT, Z

The efficiency of conversion of thermal energy to electrical energy is denoted by the parameter called the figure-of-merit of a thermoelectric material. It is defined as:

$$Z = \frac{\alpha^2 \sigma}{K} \quad (17.36)$$

where α is the Seebeck coefficient of the material (measured in microvolts/K), σ is the electrical conductivity of the material and K is the total thermal conductivity of the material.

Typically metals have small thermopowers. In 1909 Altenkirch showed that good thermoelectric materials should possess large Seebeck coefficients, high electrical conductivity and low thermal conductivity. A high electrical conductivity is necessary to minimize Joule heating, whilst a low thermal conductivity helps to retain heat at the junctions and maintain a large temperature gradient. These three properties were embodied in *figure-of-merit*, Z . Since Z varies with temperature, a useful dimensionless figure-of-merit can be defined as ZT .

Semiconductors

In contrast, semiconductors have large positive or negative values of the thermopower depending on the charge of the excess carriers. The sign of the thermopower can determine which charged carriers dominate the electric transport in both metals and semiconductors.

During the 1920's synthetic semiconductors were developed with Seebeck coefficients in excess of 100 microvolts/K.

This is more commonly expressed as the *dimensionless figure of merit ZT* by multiplying it with the average temperature $((T_2 + T_1) / 2)$. Greater values of ZT indicate greater thermodynamic efficiency. ZT is a very convenient figure for comparing the potential efficiency of devices using different materials. Values of $ZT = 1$ are considered good, and values of at least the 3–4 range are considered to be essential for thermo electrics to compete with mechanical generation and refrigeration in efficiency. To date, the best reported ZT values have been in the 2–3 range. Much research in thermoelectric materials has focused on increasing the Seebeck coefficient and reducing the thermal conductivity, especially by manipulating the nanostructure of the materials.

Heat produced at a junction is a combination of Joule heat and “borrowed” heat contributed by connections elsewhere in the circuit, whether purposeful or not. Similarly, heat absorbed by a junction will be “pumped” into the other junctions distributed around the circuit.

Semiconductor materials formulated especially for use in Peltier devices show “heat pumping” efficiency several times that of the most active metal pairings. These devices are used in applications ranging from lunch box coolers/warmers to laboratory instruments and communications systems. Quite large temperature differences can be generated by cascading or pyramiding Peltier devices such that each level in the pyramid acts as a heat sink for the level above.

17.15 THERMOELECTRIC POWER GENERATION

The **Seebeck** effect forms the basis for power generation. Thermo-electric generators convert heat directly into electricity, using the voltage generated at the junction of two different metals.

Thermoelectric power generators convert heat energy to electricity. When a temperature gradient is created across the **thermoelectric device**, a DC voltage develops across the terminals. When a load is properly connected, electrical current flows. Typical applications for this technology include providing power for remote telecommunication, navigation, and petroleum installations.

As early as 1929, A.F. Ioffe (1880-1960) showed that a thermoelectric generator utilizing semiconductors could achieve a conversion efficiency of 4%, with further possible improvement in its performance. The simplest thermoelectric generator consists of a thermocouple, comprising a p-type and n-type thermo-element connected electrically in series and thermally in parallel (Fig. 17.12). Heat is pumped into one side of the couple and rejected from the opposite side. An electrical current is produced, proportional to the temperature gradient between the hot and cold junctions.

Of the great number of materials studied, semiconductors based on bismuth telluride, lead telluride and silicon-germanium alloys are found to be the best.

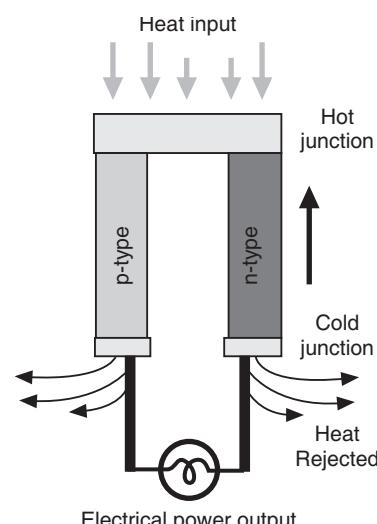


Fig. 17.12. Illustration of thermoelectric generation (Seebeck effect)

17.16 THERMOELECTRIC COOLING

The basic concept behind thermoelectric cooling is the **Peltier effect**. Semiconductors, usually Bismuth Telluride, are the material of choice. In its simplest form, a **Peltier device** is a single semiconductor ‘pellet’ soldered to electrically-conductive material on each end to plated copper. The copper connection paths serve as the second dissimilar material required for the formation of the junctions, i.e., thermocouple. While we can make a simple thermoelectric device with a single semiconductor pellet, we cannot pump an appreciable amount of heat through it. In order to give a thermoelectric device greater heat-pumping capacity, multiple pellets are used together.

By arranging n- and p-type pellets in a ‘couple’ and forming a junction between them with a plated copper tab, it is possible to configure a simple series circuit (Fig. 17.13).

The bottom end of the p-type pellet is connected to the positive voltage potential and the bottom end of the n-type pellet is similarly connected to the negative side of the voltage. The holes in the p-material are repelled by the positive voltage potential and attracted by the negative pole; the negative charge carriers (electrons) in the n-material are likewise repelled by the negative potential and attracted by the positive pole of the voltage supply. The electrons flow continuously from the negative pole of the voltage supply, through the n-pellet, through the copper tab junction, through the p-pellet, and back to the positive pole of the supply—yet because we are using the two different types of semiconductor material, the current is flowing in the same direction through the pellets.

The heat will be moved (or ‘pumped’) in the direction of charge carrier movement throughout the circuit as it is actually the charge carriers that transfer the heat. Thus, heat is absorbed at the bottom junction and actively transferred to the top junction. Therefore, the bottom junction is the cold junction while the top junction forms the hot junction. The cold junction will rapidly drop below ambient temperature provided heat is removed from the hot side. The temperature gradient will vary according to the magnitude of current applied.

17.17 THE THERMOELECTRIC COOLERS

Thermoelectric coolers are solid state heat pumps used in applications where temperature stabilization, temperature cycling, or cooling below ambient are required.

A thermoelectric cooling arrangement is shown in Fig. 17.14. It consists of a thermoelectric module, a heat sink and the object to be cooled. A typical thermoelectric module consists of an array of bismuth telluride semiconductor pellets that have been “doped” so that one type of charge carrier—either positive or negative carries the majority of current. The pairs of P/N pellets are configured so that they are connected electrically in

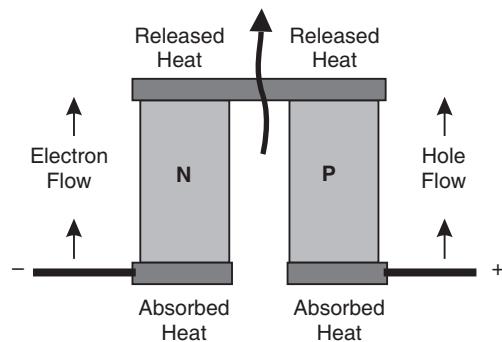


Fig. 17.13. Illustration of thermoelectric cooling (Peltier effect)

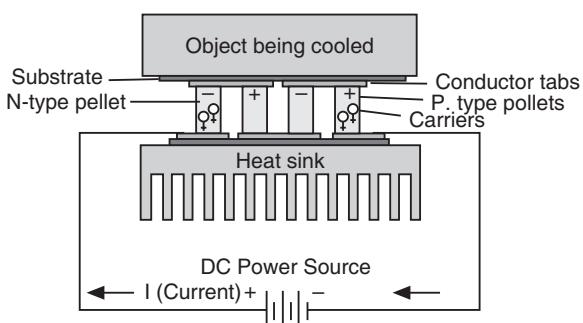


Fig. 17.14

series, but thermally in parallel. Metalized ceramic substrates provide the platform for the pellets and the small conductive tabs that connect them. The ceramic material on both sides of the thermoelectric adds rigidity and the necessary electrical insulation. The pellets, tabs and substrates thus form a layered configuration. Module size varies from less than 0.25" by 0.25" to approximately 2.0" by 2.0". Thermoelectric modules can function singularly or in groups with either series, parallel, or series/parallel electrical connections. Some applications use stacked multi-stage modules.

When DC voltage is applied to the module, the positive and negative charge carriers in the pellet array absorb heat energy from one substrate surface and release it to the substrate at the opposite side. The surface where heat energy is absorbed becomes cold; the opposite surface where heat energy is released, becomes hot.

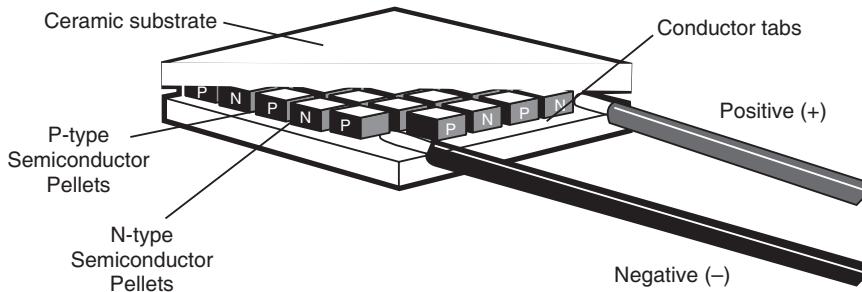


Fig. 17.15

These devices cannot only pump appreciable amounts of heat, but with their series electrical connection, are suitable to be used as DC power supplies. Thus, the most common **thermoelectric devices** now in use—connecting 254 alternating P and N-type pellets—can run from a 12 to 16 VDC supply and draw only 4 to 5 amps. A means to mechanically hold everything together is to mount the conductive tabs to thin ceramic substrates (Fig. 17.15); the outer faces of the ceramics are then used as the thermal interface between the **Peltier device** and the ‘outside world’. Ceramic materials represent the best compromise between mechanical strength, electrical resistivity, and thermal conductivity.

Advantages

- Thermoelectric coolers offer several distinct advantages over other technologies:
- Thermoelectric coolers have no moving parts and, therefore, need substantially less maintenance.
- Thermoelectric devices have life times of the order of 100,000 hrs of steady state operation.
- Thermoelectric coolers contain no chlorofluorocarbons or other materials which may require periodic replenishment.
- Temperature control to within fractions of a degree can be maintained using thermoelectric devices and the appropriate support circuitry.
- Thermoelectric coolers function in environments that are too severe, too sensitive, or too small for conventional refrigeration.
- Thermoelectric coolers are not position-dependent.
- The direction of heat pumping in a **thermoelectric system** is fully reversible. Changing the polarity of the DC power supply causes heat to be pumped in the opposite direction—a cooler can then become a heater!

Thermoelectrics can be used to heat and to cool, depending on the direction of the current. In an application requiring both heating and cooling, the design should focus on the cooling mode. Using a thermoelectric in the heating mode is very efficient because all the internal heating (Joulian heat) and the load from the cold side is pumped to the hot side. This reduces the power needed to achieve the desired heating.

Applications

Thermoelectric technology is applied to many widely-varied applications.

There are many products using thermoelectric coolers, including CCD cameras (charge coupled device), laser diodes, microprocessors, blood analyzers and portable picnic coolers, portable refrigerators, scientific thermal conditioning, liquid coolers, and beyond. There are an increasing number and variety of products which use thermoelectric technology such as highly-specialized instrumentation and testing equipment. The compatibility of many thermoelectrics with automotive voltages, makes them especially suitable for small cooling jobs in that industry. With each new year, the imaginations of design engineers widen with the immense possibilities of thermoelectric heating and cooling.

QUESTIONS

1. Explain the terms Seebeck effect, Peltier effect and Thomson effect.
2. Give the theory of thermo e.m.f. Show that $\pi = T.de/dT$.
3. Explain Seebeck effect and Peltier effect. Prove that Peltier coefficient of a pair of metals is product of absolute temperature and thermoelectric power.
4. Explain characteristics of a thermocouple. Explain the terms (i) neutral temperature and (ii) inversion temperature. **(Bombay Univ.)**
5. Define thermocouple, neutral temperature, thermoelectric power and temperature of inversion.
6. What is thermoelectric emf ? Explain (i) the law of intermediate metals and (ii) the law of intermediate temperatures. **(Bombay Univ.)**
7. (a) Explain the Seebeck effect. Describe how the e.m.f. of a thermocouple is measured experimentally.
 (b) State the expression for the total e.m.f. of a thermocouple and hence define neutral temperature and inversion temperature in terms of the thermoelectric constants. **(Bombay Univ.)**
8. Explain: Neutral temperature, inversion temperature and thermoelectric power. Describe the method for using thermocouple as a thermometer. **(Bombay Univ.)**
9. Prove the thermoelectric relation

$$\frac{d}{dT} \left(\frac{\pi}{T} \right) + \left(\frac{\sigma_A - \sigma_B}{T} \right) = 0$$
10. Prove that in a thermocouple

$$\sigma = T \cdot \frac{d^2 e}{dT^2}$$
11. (a) Describe how a thermocouple is used as thermometer and discuss its advantages.
 (b) Obtain the law of intermediate metals for thermocouples. What is its significance? **(Bombay Univ.)**
12. What is thermoelectric effect? From the formula , find out neutral temperature and thermoelectric power. **(Bombay Univ.)**
13. (a) What is a thermocouple? Describe the Seebeck effect and the Peltier effect.
 (b) What are the advantages of using thermocouple as a thermometer? **(Bombay Univ.)**
14. Write a short note on thermopile.

15. Explain the laws of intermediate metals and intermediate temperatures.
 16. Show that in a thermocouple composed of two metals, the total e.m.f. is given by

$$e = \pi_2 - \pi_1 + \int_{T_1}^{T_2} (\sigma_A - \sigma_B) dT$$

17. Deduce from thermodynamic consideration the expressions for Peltier and Thomson coefficients.
 18. Explain thermoelectric power generation.
 19. Explain the principle of thermoelectric cooling.
 20. Describe the construction and working of a thermoelectric cooler.
 21. Explain the advantages of thermoelectric cooling and some of the areas where thermoelectric cooling is applied profitably.

PROBLEMS

- Calculate the neutral temperature for an iron - silver thermo couple. The values of a and b are 16. 65 and -0.096 for iron and 2.86 and 0.017 for silver respectively. **(Bombay Univ., 92)**
[Ans: 61°C]
- The neutral temperature of a thermocouple is 300 °C. When its junctions are kept at temperatures 0°C and 100 °C, the e.m.f. generated is 1300 µV. Calculate the values of the coefficients a and b . **(Bombay Univ., 97)** **[Ans: 15.6 × 10⁻⁶]**
- For Fe-Cu thermocouple it is observed that the thermo e.m.f is zero when one of the junction is at 20°C and the other one is at some higher temperature. If the neutral temperature is 285°C, calculate the higher temperature. Hence find out the temperature of inversion, if the cold junction temperature is at -20°C. **(Bombay Univ., 95)** **[Ans: 650°C]**
- The e.m.f of Fe-Pb thermocouple when one junction is at 0°C and the other at 100 °C is 1185 µV. When the second junction is at 300 °C the e.mf. is 675 µV. Similar readings with Ag-Pb thermocouple are 371 µV and 1623 µV respectively. Calculate the neutral temperature for F e-Ag thermocouple. **[Ans: 122°C]**
- A thermocouple is made of iron and constantan. Find the e.m.f. developed per °C difference of temperature between the junctions, given that thermo e.m.f.s of iron and constantan against platinum are +1600 and -3400 microvolts per 100°C difference of temperature. **[Ans: - 16 µV/°C (- 34)µV/°C 50 µV/°C]**
- The thermoelectric power for steel is 18 µ V / °C at 0 °C and zero at 400 °C. That for copper is 6 µV / °C at 500 °C and zero at -50°C. Find the e.m.f. for steel-copper thermocouple with one junction at its neutral temperature and other at 0°C. **(Bombay Univ., 96)** **[Ans : 2.7 mV]**
- The thermoelectric power of Fe-Pb thermocouple is 17.5 µV/ °C at 0°C and zero at 360°C and that for Cu-Pb is 5 µV / °C at 450° C and zero at -50° C. Find the e.m.f for Fe-Cu thermocouple with its one junction at 0° C and other at its neutral temperature. **(Bombay Univ., 95)** **[Ans:4.97 mV]**
- The thermoelectric power of Fe - Pb thermocouple is 17.5 µV/ °C at 0°C and 5µV/°C at 125°C. The thermoelectric power of Cd-Pb thermocouple is 3 µV/ °C at 0°C and 15µV/°C at 150°C. Calculate the neutral temperature of the Fe – Cd junction. **(Bombay Univ., 95)** **[Ans:80.6°C]**
- The thermoelectric power of iron is 1734 – 4.87 t and that of copper is 136–0.95 t , where t is the temperature in °C. Show that the e.m.f. of thermocouple of iron-copper, the junctions of which are at 0°C and 100°C is 0.14 V.
- The e.m.f. of iron-lead thermocouple, where one junction is at 0°C and the other at 100°C is 1185 µV. When the second junction is at 300°C, the e.m.f. is 675 µV. Similar readings with silver-lead couple are 371 µV and 1623 µV respectively. Calculate the neutral temperature of an iron-silver thermocouple. **[Ans: 122°C]**

CHAPTER

18

Special Theory of Relativity

18.1 INTRODUCTION

In 1905, at the age of only 26, Albert Einstein (1879-1955) published four scientific papers that revolutionized physics. One of the papers dealt with the photoelectric effect, which was responsible for the birth of quantum physics. Two of the papers dealt with special theory of relativity. Relativity represents the greatest intellectual achievement of twentieth century physics. Just as quantum theory showed that the classical concepts are to be revised in case of microscopic world, the relativity theory established that the classical notions are not applicable to bodies moving with velocities approximately nearer to that of light. Classical mechanics regarded space and time to be absolute and separate entities. It assumed the flow of time to be uniform in all situations. As such the moments of time and time intervals are supposed to be identical in all frames of reference. Similarly, lengths are assumed to be identical in all frames. An analysis of these concepts at high velocities revealed that they are not correct. The relativity theory leads to many unusual conclusions. It shows that the length of moving bodies contract in the direction of motion and the clocks in motion slow down. It is difficult to comprehend these conclusions because of our habit founded on routine experiences. At high velocities, space and time are no more separate but merge into space-time continuum. The results of experiments on high-energy particles provide the proof for the predictions of relativity.

18.2 SPACE, TIME AND MOTION

The concepts of space, length, time and mechanical motion appear to us as self-evident and obvious.

Classical mechanics presumes the space to be homogeneous in all its parts and also isotropic. It means that the properties of space are identical at all points and in all directions at each point. Classical mechanics further supposes the existence of an absolute space, which is absolutely motionless and irrelevant to the existence of any bodies. According to Newton “absolute space, in its own nature, without regard to anything external, remains always similar and immovable”.

In classical mechanics, time is understood as a measure of absolute duration, which exists irrespective of physical bodies. Newton regarded that the true course of time is not liable to change; in his opinion the course of time belongs to absolute category.

If we are asked, “What is mechanical motion?”, we sometimes simply say that the displacement of a body in space with time is mechanical motion. In fact, the displacement of

a body can be recognized only when there is a fixed body relative to which the given body changes its position. Therefore, we mean by mechanical motion the displacements of bodies relative to some other bodies. Thus, for the description of motion it is essential that there are certain real physical bodies, which can be assumed to be fixed. When we say that a train is moving, we are referring to the change of position of the train with respect to the platform, which is stationary. It is not possible to observe the change of position of a single isolated body and, therefore, it is senseless to speak of motion in such a case.

18.3 FRAME OF REFERENCE

Space cannot be thought of without physical bodies. To locate a body in space, the Cartesian coordinate system is used normally. The body may be a fixed body used for reference or it may be a body which is in motion. A Cartesian coordinate system attached to the reference fixed body is called a “frame of reference” or a “coordinate system”. It is impossible to attach a set of coordinate axes to empty space.

The choice of a frame of reference is determined by our own convenience. For describing the motion of bodies on the earth, we choose a frame of reference rigidly connected to the earth, which is regarded as a fixed body. However, in reality the Earth is in circular motion. In investigation of the Earth’s motion, we attach the coordinate system to the Sun. In studying the Sun’s motion, we choose a reference frame connected to the stars. Sometimes, we choose the floor or walls of a room as the reference frame which may be at rest with respect to the Earth or which may be in motion if it is on a train or in a spacecraft.

The choice of a frame of reference is arbitrary. Therefore, a passenger in a moving train may claim to be at rest and declare that the electric poles and trees are moving backward while a person standing on the ground claims that the train is moving forward. Both are equally right. On the basis of our daily experience, we intuitively choose a frame of reference attached to the Earth and describe the motion of various bodies. Thus, in the reference system attached to the Earth, it is the train that moves forward.

18.4 INERTIAL FRAMES OF REFERENCE

Basically, there are infinite number of reference frames available and any one of them can be used. But the laws of mechanics may take different forms in different reference frames. For instance, let us consider acceleration of a body relative to an arbitrary frame of reference. The acceleration of the body may be due to its interaction with other bodies or it may be due to the properties of the reference frame itself. If the acceleration of the body arises solely due to its interaction with other bodies, then the frame of reference is said to be an *inertial reference frame*. In an inertial frame of reference, a *free body* moves rectilinearly and uniformly, without exhibiting acceleration.

Newton’s first law states that every body continues in its state of rest or of uniform rectilinear motion, unless it is compelled to change that state by forces impressed on it. This is known as the *law of inertia*. This law, in effect, defines an inertial frame of reference. An **inertial frame of reference** is one in which a body, not subjected to a force, moves with constant velocity. A reference frame that moves with constant velocity relative to the distant

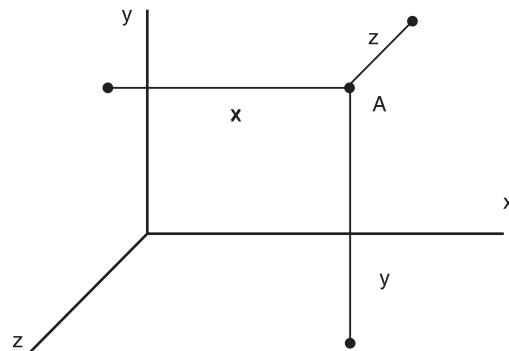


Fig. 18.1: A frame of reference

stars is the best approximation of an inertial frame. In reality, the Earth is not an inertial frame because of its orbital motion about the sun and rotational motion about its own axis. However, the Earth may be assumed to be an inertial frame in many situations.

Any other reference frame moving rectilinearly and uniformly relative to an inertial frame is also inertial. Thus, there is a vast number of inertial reference frames moving relative to one another uniformly and rectilinearly.

18.5 NON-INERTIAL REFERENCE FRAME

A frame of reference, which is in an accelerated motion with respect to an inertial frame of reference, is known as a **non-inertial frame** of reference. The law of inertia is valid in an inertial frame, whereas it is not valid in an accelerated reference frame. A ball placed on the floor of a train will move to the rear if the train accelerates forward even though no forces act on it; likewise, a coin placed on a rotating turntable will slide to the periphery though no visible force pushes it away from the centre.

18.6 GALILEO'S PRINCIPLE OF RELATIVITY

A reference frame for which Newton's laws of dynamics hold is an inertial frame. If a body is in uniform motion ($v = \text{constant}$) an observer in an inertial frame, which is at rest with respect to the body, will find that the acceleration and the resultant force on the body are zero (i.e., $a = 0$ and $F = 0$). An observer in another frame of reference moving with uniform velocity in a straight line with respect to the first frame will also find that $a = 0$ and $F = 0$ for the body. Consequently, the second reference frame is inertial to the same degree as the first. It follows that any reference frame moving with uniform velocity with respect to an inertial frame is also an inertial frame.

Thus, reference frames fixed in a railway bogie or a ship that travels with a uniform velocity in a straight line, are inertial frames. Experience shows that in a railway car or on a ship traveling with uniform velocity, it is equally easy to move in any direction as it is on the earth. A body released at a height falls vertically downward. Thus, the results of an experiment performed in a uniformly moving vehicle will be the same as those from the same experiment performed in the stationary laboratory. Therefore, it is not possible for us to tell by any experiment whether we are at rest or moving with uniform velocity.

It follows that the laws of mechanics have the same form in all inertial frames and none of the reference frames have any advantage over the others. It implies that there is no "exclusive" or "preferred" frame of reference and every inertial frame is as good as the other. Absolute rest or absolute motion of bodies has no sense. We can speak of their relative motion in some inertial frame reference.

This basic law of nature was recognized by Galileo and is summed up in the form of principle of relativity. **Galileo's principle of relativity states that the laws of mechanics are the same in all inertial frames of reference.**

18.7 GALILEAN TRANSFORMATIONS

The transformation from one inertial frame of reference to another is called **Galilean transformation**. Knowing the laws of motion of a body in a reference system S , one can derive the laws of motion of the same body in another inertial system S' .

(a) Coordinate Transformation

When a point has coordinates x , y , and z at the instant " t " we regard it as an event. It means that the location and time of occurrence of the event can be specified by the coordinates (x , y , z , and t), as shown in Fig. 18.2.

An event may be in reality some physical phenomenon such as the explosion of a flashbulb. The Galilean transformation establishes the relationship between the coordinates x, y, z and t of an event in system S and the coordinates x', y', z' and t' of the same event in system S' .

Let an event occur in an inertial frame of reference S and let the event be located by the coordinates (x, y, z , and t) in the system S . Let us consider another inertial frame of reference S' , which moves in a straight line with respect to the frame S at a constant speed v . For the sake of simplicity, let us assume that the reference frame S' moves so that x and x' are in one straight line and have the same positive directions. Also let the origins O and O' coincide at the moment $t = 0$.

Let the point P be at rest with respect to S . In such a case, it moves with respect to the second frame S' and its coordinates change. The coordinate x' depends on time, since the system S' moves at a speed v . During the time t , the moving system covers a distance equal to vt . Hence $x' = x - vt$. Obviously, $y' = y$ and $z' = z$. Also $t' = t$. Thus,

$$\begin{aligned}x' &= x - vt \\y' &= y \\z' &= z' \quad \text{Galilean Coordinate Transformation (18.1)} \\t' &= t\end{aligned}$$

Galilean transformation expresses the space-time relation of an event in different inertial frames. It is seen from equ. (18.1) that the coordinates of an event are relative and have different values in different reference frames. In the above transformation we tacitly assumed that the passage of time is the same in both the inertial systems. That is, the time of an event for an observer in S is the same as time for the same event in S' . This intuitive assumption is taken for granted in classical mechanics. It is considered natural and does not require any proof. All the phenomena in mechanics where $v < c$ do not contradict this assumption.

(b) Velocity transformation

Let us now consider the passage of a moving body. For the sake of simplicity, let us assume that the body moves parallel to the $O'x'$ -axis in the positive direction at a speed v' with respect to the system S' . Let the body be at P_1 at time t_1 and at P_2 at time t_2 , as shown in Fig. 18.3.

If the coordinates of the body in S are x_1, y_1, z_1 , when it is at P_1 and x_2, y_2, z_2 , at P_2 , then

$$x'_1 = x_1 - vt_1; \quad y'_1 = y_1; \quad z'_1 = z_1 \quad (18.2)$$

$$x'_2 = x_2 - vt_2; \quad y'_2 = y_2; \quad z'_2 = z_2 \quad (18.3)$$

Taking the difference, we get

$$x'_2 - x'_1 = (x_2 - x_1) - v(t_2 - t_1)$$

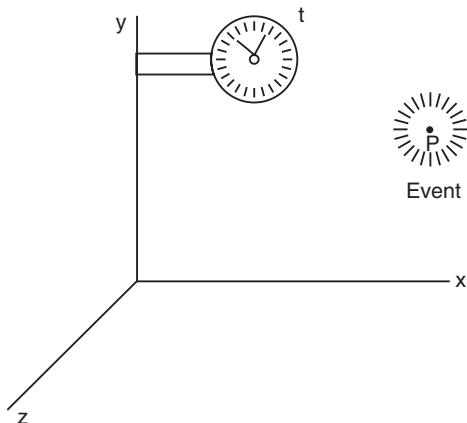


Fig. 18.2: An event is specified by the coordinates (x, y, z) at time t

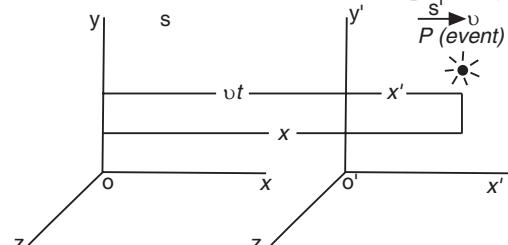


Fig. 18.3: An event occurring at point P is observed by an observer in two different inertial frames S and S' , where S' moves with a velocity v relative to S .

$$\therefore \frac{x'_2 - x'_1}{t_2 - t_1} = \frac{x_2 - x_1}{t_2 - t_1}$$

$$\text{But } \frac{x'_2 - x'_1}{t_2 - t_1} = u'_2 \quad \text{and } \frac{x_2 - x_1}{t_2 - t_1} = u_x$$

$$\therefore u'_x = u_x - v \quad (18.4)$$

It is obvious that $u'_y = u_y$ and $u'_z = u_z$. Thus,

$$u'_x = u_x - v$$

$$u'_y = u_y$$

$$u'_z = u_z$$

Galilean Velocity Transformation (18.5)

Equation (18.5) is known as the Galilean velocity transformation. It states that the velocity of a body located at a point P measured by an observer in the moving frame equals the velocity as measured in the stationary frame minus the velocity of the frame S'.

Equation (18.4) may be rewritten as

$$u_x = u'_x + v \quad (18.6)$$

Equation (18.6) implies the well-known law of addition of velocities in the classical mechanics.

The Galilean transformation equations agree with our daily experience. For example, a person in a moving railcar (Fig. 18.4) throws a ball with a speed u in the direction of motion of the railcar, where u is measured by the person in the railcar. If v is the velocity of the moving railcar, the speed of the ball relative to a stationary observer on the ground will be $u + v$.

(c) Acceleration Transformation

Let us now consider the case where the body moves along x -axis of reference frame S with acceleration. At some moment of time, say

$$u = u' + v \quad (18.7)$$

After a time interval Δt , the velocities in both systems increase and the new velocities are given by

$$(u + \Delta u) = (u' + \Delta u') + v \quad (18.8)$$

Subtracting eqn. (18.7) from eqn. (18.8), we get

$$\Delta u = \Delta u'$$

Dividing both sides of the above equation with Δt , we obtain

$$\frac{\Delta u}{\Delta t} = \frac{\Delta u'}{\Delta t'}$$

or

$$a = a' \quad \text{Galilean Acceleration Transformation} \quad (18.9)$$

Multiplying eqn. (18.9) by the mass of the body, we get

$$ma = ma'$$

or

$$F = F' \quad (18.10)$$

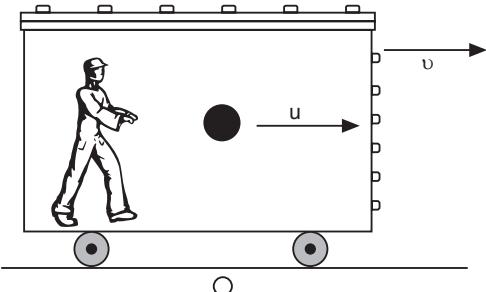


Fig.18.4: A person in a moving railcar throws a ball with a speed u . The speed of the ball relative to a stationary observer O is $u + v$

It follows from equs. (18.9) and (18.10) that accelerations and forces do not change from one inertial frame to another inertial frame. Such quantities, which remain invariable under the transformation, are termed **invariants**. It is seen from the above that Newton's laws are invariant with respect to Galilean transformations.

18.8 THE ETHER

Newton's postulate of absolute space implied the existence of an absolutely stationary reference frame. He assumed that out of the infinite inertial frames moving uniformly with respect to another, there is a preferred frame fixed in absolute space, which is really stationary. However, the mechanical experiments fail to establish the existence of an absolute reference frame. According to the Galilean principle of relativity, all inertial reference frames are equivalent as far as laws of mechanics are concerned. Mechanics was the only branch of physics advanced sufficiently at that time. With the developments in electrodynamics and optics, scientists naturally were concerned with the questions whether the principle of relativity is applicable to the phenomena in these fields and whether one can find an absolute frame of reference. As a result of the importance attached to the absolute frame, it became of considerable interest to prove its existence by experiment.

The interference and diffraction effects of light clearly established wave nature of light. As all known waves are transmitted by some medium, it was natural to assume that light waves are also transmitted by a medium, which obviously fills the space around us and the interplanetary and the interstellar space. It was supposed that this all pervasive medium would have no effect on the motion of planets and other celestial bodies. Failing to identify such an unusual medium, the scientists of 19th century concocted the medium and named it **luminiferous ether** or in short the **ether**. Whatever may be its nature, the ether cannot be at rest in all inertial frames of reference simultaneously. Therefore, one can distinguish an inertial frame that is stationary with respect to the ether. It was supposed that, in that and only in that reference frame, light propagates with the same velocity c in all directions. That frame of reference would be the absolute frame.

If a certain inertial frame moves with the velocity v relative to the ether, the velocity of light c' in that reference frame must obey the conventional law of velocity composition. Thus, $c' = c - v$. For instance, the velocity of light on the Earth must depend on the velocity of the motion of the Earth relative to the reference system of the ether. If the velocity of the Earth in this system is v and the speed of light relative to the medium is c , then the speed of light must be equal to $c - v$ in the direction of the Earth's motion and $c + v$ in the opposite direction.

The speed of the Earth in its orbit is about 30 km/s, which is only about 1/10,000 times the speed of light. Therefore, the observation and the measurement of the influence of the Earth's motion on the speed of light face many experimental difficulties. High precision apparatus is required to verify the addition rule for the velocities of the light and of the Earth. First experiments of this kind were carried out in 1881 by the American Physicists Albert A. Michelson (1852–1931) and again in 1887 by A.A. Michelson and E.W. Morley (1838–1923).

18.9 MICHELSON-MORLEY EXPERIMENT

A schematic diagram of the Michelson-Morley experiment is shown in Fig.18.5. A collimated beam of light from the source S is split into two beams (A and B) by the half-silvered mirror M. Beam A travels to mirror M_1 and is reflected back to M, where it is reflected down. Beam B proceeds to mirror M_2 and is reflected back to M and is transmitted through M. The component of beam B, which is transmitted through M, combines with the reflected part

of beam A. Since beams A and B are derived from the same beam falling on M, they are coherent and produce interference.

Let us assume that the Earth travels through stationary ether with a speed v and that light has a speed c in the ether. We further assume that the instrument is arranged such that one of its two equal arms is parallel to the Earth's velocity v (Fig. 18.6). The speed of beam A traveling from M to M_1 is $c - v$ relative to the instrument and the time required for light to travel from M to M_1 is $L/(c - v)$. Similarly, the speed of light beam is $(c + v)$ in the trip from M_1 to M and the time required for the return trip to M is $L/(c + v)$. The time for the round trip MM₁M is thus

$$T_A = \frac{L}{c-v} + \frac{L}{c+v} = \frac{2Lc}{c^2 - v^2} = \frac{2L/c}{1 - \frac{v^2}{c^2}} \quad (18.11)$$

The beam B that travelled toward mirror M_2 will get reflected at M_2' but only after M has moved to a new position M' (Fig. 18.6). The component of velocity of light in the direction perpendicular to the motion of the instrument is $\sqrt{c^2 - v^2}$. Therefore, the time for the round trip MM₂M is

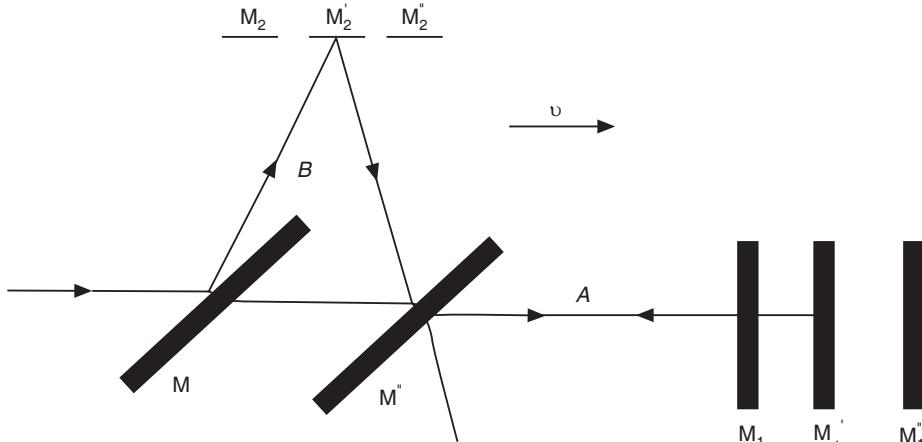


Fig. 18.6: The paths of beam A and B in reference frame at rest relative to the other. The interferometer moving with a velocity v relative to the other.

$$T_B = \frac{2L}{\sqrt{c^2 - v^2}} = \frac{2L/c}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (18.12)$$

The time difference between the two paths due to the motion of the instrument relative to the ether is given by

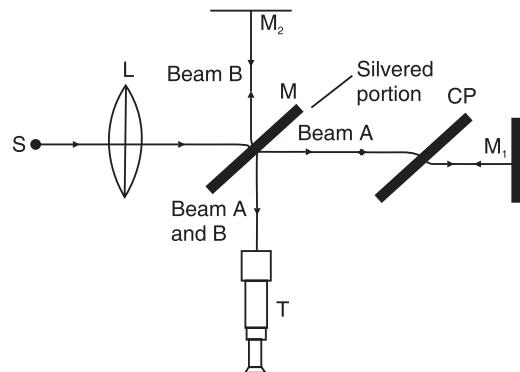


Fig. 18.5: A schematic diagram of the Michelson-Morley interferometer experiment

$$\Delta T = T_A - T_B = \frac{2L}{c} \left[\frac{1}{1 - \frac{v^2}{c^2}} - \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \right] \quad (18.13)$$

Since v is very small compared to c , we may use the binomial theorem to simplify equ. (18.13). We can approximate

$$\frac{1}{1 - \frac{v^2}{c^2}} = 1 + \frac{v^2}{c^2} \text{ and } \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = 1 + \frac{v^2}{2c^2}$$

Using the above approximations into equ. (18.13), we get

$$\Delta T = \frac{2L}{c} \left[1 + \frac{v^2}{c^2} - 1 - \frac{v^2}{2c^2} \right] = \frac{Lv^2}{c^3} \quad (18.14)$$

If the instrument is turned through 90° , the role of paths A and B are interchanged. In this case, the times T'_A and T'_B required to travel the paths MM'_1M' is given by

$$T'_A = \frac{2L/c}{\sqrt{1 - \frac{v^2}{c^2}}} \text{ and } T'_B = \frac{2L/c}{1 - \frac{v^2}{c^2}}$$

$$\Delta T' = T'_A - T'_B = \frac{2L}{c} \left[\frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} - \frac{1}{1 - \frac{v^2}{c^2}} \right]$$

or $\Delta T' = -\frac{Lv^2}{c^3}$ (18.15)

$$\Delta T_{\text{total}} = \Delta T - \Delta T' = \frac{Lv^2}{c^3} - \left(-\frac{Lv^2}{c^3} \right)$$

$$\therefore \Delta T_{\text{total}} = \frac{2Lv^2}{c^3} \quad (18.16)$$

The path difference introduced between the beam components A and B will be, therefore,

$$\Delta = c(\Delta T_{\text{total}}) = \frac{2Lv^2}{c^2} \quad (18.17)$$

The number of fringes passing through a reference mark will be

$$N = \frac{\text{Path difference}}{\lambda} = \frac{2Lv^2}{\lambda c^2} \quad (18.18)$$

Michelson and Morley mounted the entire apparatus on a large stone slab, which floated on a mercury pool (Fig. 18.7). By using multiple reflections (Fig. 18.8), the effective path L is made 11 m which improves the accuracy of the measurement. Using the light of wavelength $\lambda = 5500 \text{ \AA}$, and taking $v = 30 \text{ km/s}$, we obtain

$$N = \frac{2 \times 11 \text{ m} (3 \times 10^4)^2 \text{ m}^2/\text{s}^2}{5500 \times 10^{-10} \text{ m} (3 \times 10^8)^2 \text{ m}^2/\text{s}^2} = 0.4$$

A fringe shift of this amount is readily detected with the apparatus. It should then be possible to measure the fringe shift and from it determine the velocity of the Earth relative to the ether. Michelson and Morley failed to observe any fringe shift. Measurements were made during day and night, at various locations in Europe and U.S.A., and in various seasons of the year.

The results of the experiment always proved to be negative. The negative result of Michelson-Morley experiment contradicts Galileo's relativity principle from which the ordinary addition rule for the velocities follows. It also showed that motion relative to the ether cannot be detected and the velocity of light is independent of the motion of a light source, since its motion with respect to the ether is different at different seasons of the year.

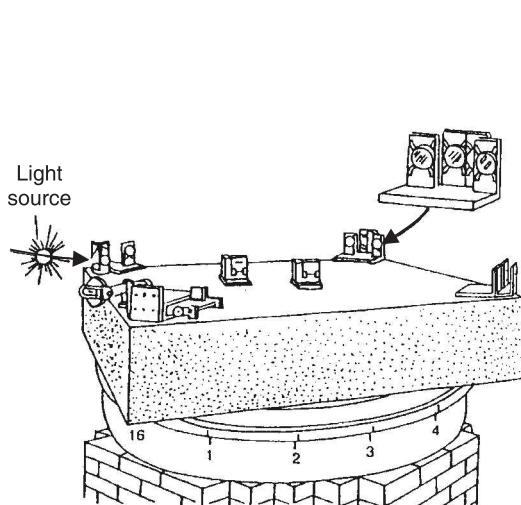


Fig. 18.7: Drawing of the Michelson-Morley Interferometer mounted on a stone slab which floats on a pool of mercury.

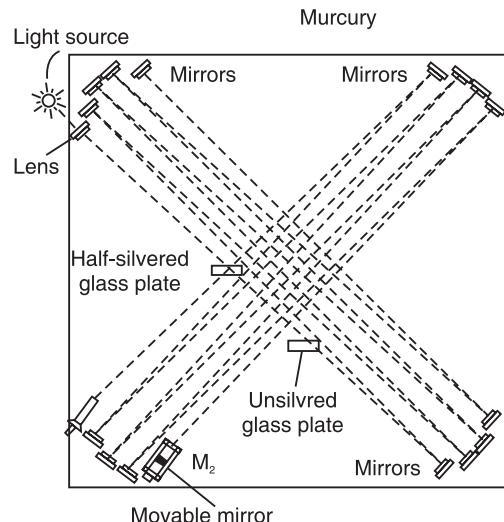


Fig. 18.8: A top view of the interferometer using additional mirrors to increase the optical path length.

Some astronomical observations such as double stars also indicate the fact that the velocity of light does not depend on the velocity of the source. A number of special experiments carried out later gave the same evidence.

Following the publication of the Michelson-Morley experiment in 1887, several explanations were put forward to explain the null result of the experiment. One of the ingenious explanations was offered in 1889 by G.F. Fitzgerald, an Irish Physicist. He suggested that all objects contract in the direction of their motion through the ether. It was a clever suggestion according to which the arm of the instrument moving parallel to the ether was assumed to be shortened, but similar contractions were not observed in other situations.

18.10 FAILURE OF GALILEAN TRANSFORMATIONS

We have seen that the laws of mechanics are invariant under a Galilean transformation. However, the laws of electromagnetism are found to change their form under a Galilean transformation. This fact can be demonstrated with the following simple example.

Consider two equal and like charges q which are stationary in one reference frame S. They repel each other according to Coulomb's law. Now, suppose these charges are examined by an observer in a reference frame S' moving with velocity v along a line perpendicular to the line joining the two charges. The observer will find the two charges moving with velocity $-v$. The two moving charges constitute two currents in the same direction and two parallel currents attract each other. Therefore, the observer in the moving frame finds that the charges exert a mutual repulsive force as well as an attractive force.

It means that either the Galilean transformation is wrong or Maxwell's equations are wrong. But Maxwell's equations are in total agreement with all known experiments. Albert Einstein (1879–1955) was convinced that Maxwell's equations are true and must yield the same result in all inertial frames of reference.

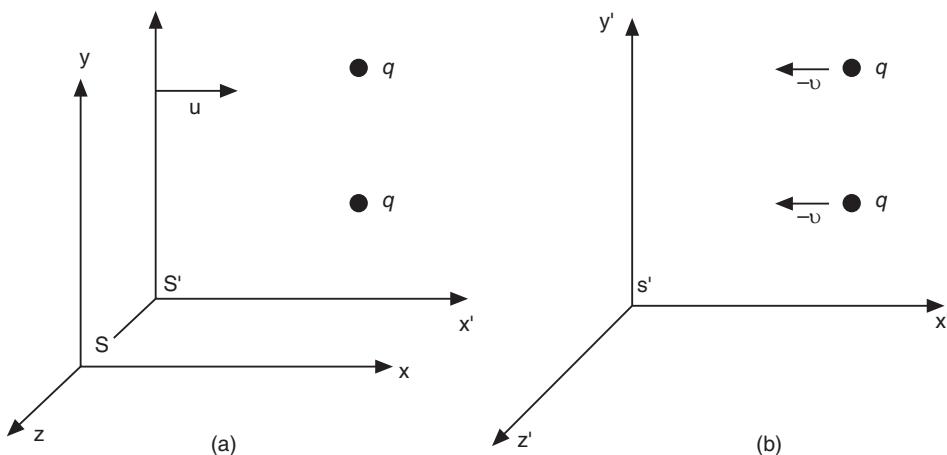


Fig. 18.9: (a) Two electrical charges q at rest in reference frame S. (b) The same charges as seen by an observer in reference frame S' which moves with a velocity u relative to S.

18.11 EINSTEIN'S PRINCIPLE OF RELATIVITY

Basing on thorough analysis of all experimental and theoretical data accumulated by the beginning of the twentieth century, Einstein concluded that the classical concept of space and time are to be revised. In 1905, at the age of 26, Einstein published a paper entitled "**On the Electrodynamics of Moving Bodies**" in the Seventeenth volume of *Annalen der Physik*, a German scientific journal. In this paper, he made the revolutionary proposal that the Newtonian concepts of space and time are to be revised; and propounded the special theory of relativity. He suggested that the experimental failure to detect uniform motion relative to an absolute reference system simply means that there is no such system at all and that all the inertial frames are completely equivalent.

The special theory of relativity is based on the following two postulates:

- 1. The Principle of relativity:** All the laws of physics are the same in all inertial frames of reference.
- 2. The Principle of independence of the velocity of light:** The speed of light in a vacuum is independent of the motion of the light source or receiver.

The term "special" implies that this theory considers phenomena only in inertial reference frames.

The first postulate is, in effect, a generalization of Galilean principle of relativity to cover all physical processes. All physical phenomena proceed identically in all reference frames.

All physical laws are absolutely identical in all inertial systems. Basically, no experiment can distinguish one of the frames as preferable. Thus, Einstein's principle of relativity establishes the complete equality of all inertial frames and rejects the Newton's ideas of absolute space and absolute motion.

The second postulate states that the velocity of light in a vacuum has the same value for all observers and is independent of their motion or of the motion of the light source. In contrast to all other velocities, which change on transition from one reference frame to another, the velocity of light in a vacuum is invariant. This invariance of velocity of light requires that we modify some of our intuitive, everyday notions of space and time.

Let us take once again the example of an observer in a railcar moving with a velocity v (Fig. 18.10). Suppose a flash of light is sent by him. The flash travels with a velocity c relative to the observer in the railcar. What will be the speed of the flash relative to the stationary observer on the ground? We expect the velocities to add as in classical case and the speed of the flash should be $c + v$. According to the Einstein's second postulate, both observers would measure the speed of light to be c . This conclusion appears to violate our common sense.

The second postulate explains the null result of Michelson-Morley experiment clearly. It was presumed in the experiment that when light travels against the ether wind, its speed would be $c - v$ and after reflection at mirror M_1 , it returns to M with a speed $c + v$. Because of the differences in speed, the beams A and B would acquire a phase shift and hence a fringe shift was predicted. According to Einstein's second postulate the value of c is the same in all directions irrespective of the direction of the motion of the Earth. Consequently, the path difference between beams A and B would be zero and as such shift in the fringe pattern does not take place.

18.12 THE LORENTZ TRANSFORMATIONS

In Newtonian mechanics the Galilean transformation equation (18.1), relate the space and time coordinates in one inertial frame to those in the other frame. The equations are not valid for cases where v approaches the value of c . The transformation equations that apply for all speeds up to c and incorporate the invariance of speed of light were developed in 1890 by the German Physicist H.A. Lorentz (1853–1928). They are known as the **Lorentz transformations**. Their real physical significance was later established by Einstein.

(a) Coordinate transformations

Let us consider two inertial reference frames S and S' in which the standards for measuring distances and time are the same. Let the reference system S be stationary while the system S' , moves with constant velocity v relative to system S along the directions of the axes x and x' . The axes x and x' are in one line and the axes y and y' , and z and z' are parallel. Let us assume that at the inertial instant, $t = t' = 0$ and the origin O and O' of the reference systems coincide [Fig. 18.11a]. We further assume that at the inertial instant $t = t' = 0$, a light source placed at the origin emits a light pulse (flash). The light pulse (flash) travels in the form of a spherical wave at the speed c in both the reference frames. Then during the time interval t' the light pulse travels a distance of ct' in the reference frame S . Similarly, in the reference frame S' , the light pulse travels a distance of ct' during the corresponding time interval t' . In both the

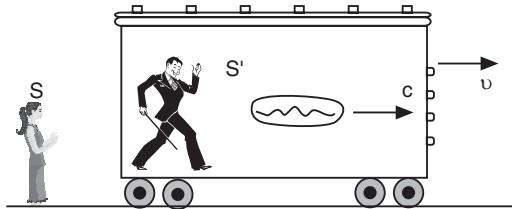


Fig. 18.10

frames the points reached by the light pulse at t and t' respectively lie on spherical surfaces of radii ct and ct' .

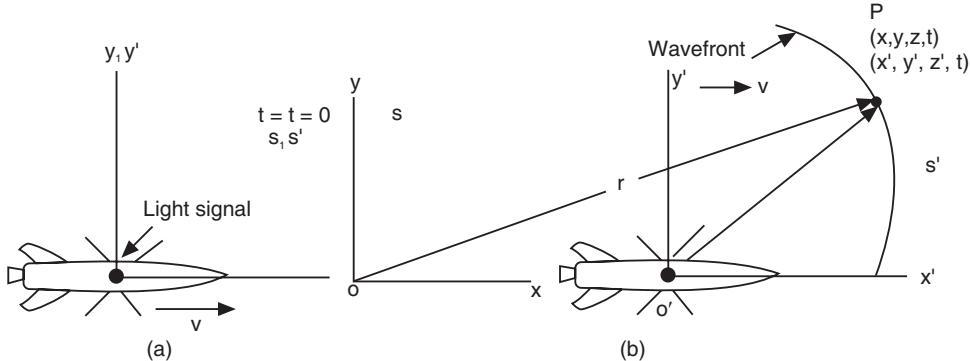


Fig. 18.11

$$r = ct \quad (18.19)$$

Thus,

$$r' = ct' \quad (18.20)$$

The wave surface in frame S is described by the equation

$$x^2 + y^2 + z^2 = r^2 = c^2 t^2 \quad (18.21)$$

and in frame S' by the equation

$$(x')^2 + (y')^2 + (z')^2 = c^2 (t')^2 \quad (18.22)$$

As the space and the time are regarded as homogeneous, the relations between the coordinates and the time in different frames must be linear. Thus,

$$x' = ax + bt \quad (18.23)$$

$$t' = fx + gt \quad (18.24)$$

Equation (18.23) can be rewritten as

$$x' = a[x + (b/a)t] \quad (18.25)$$

At time t , the origin S' is at $x = vt$. For the position $x' = 0$, it follows from equ. (18.25), that

$$0 = a[vt + (b/a)t]$$

or

$$(b/a) = -v$$

Using the above relation into equ. (18.25), we get,

$$x' = a(x - vt) \quad (18.26)$$

Similarly,

$$x = a(x' + vt') \quad (18.27)$$

Since the systems only move in the direction of x and x' axes, the coordinates y , y' and z , z' do not change.

$$y' = y \text{ and } z' = z \quad (18.28)$$

From equations (18.22), (18.26) and (18.21), we may write

$$x' = ct' = a(ct - vt) = act(1 - v/c) \quad (18.29)$$

and

$$x = ct = a(ct' - vt') = act'(1 - v/c) \quad (18.30)$$

From equation (18.29), we get

$$t' = at(1 - v/c)$$

Using the relation into equ. (18.30), we obtain

$$ct = ac [at(1 - v/c)] (1 + v/c)$$

$$= a^2 ct(1 - v^2/c^2)$$

$$\therefore a = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

More often a is denoted by γ .

$$\therefore \gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (18.31)$$

Using equ. (18.31) into equ. (18.26), we get

$$x' = \frac{x - vt}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (18.32)$$

$$t' = \frac{x'}{c} = \frac{\frac{x}{c} - \frac{vt}{c}}{\sqrt{1 - \frac{v^2}{c^2}}}$$

But, $x = ct$,

$$t' = \frac{t - \frac{vx}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (18.33)$$

Thus, the equation for transformation of coordinates from system S to system S' are

$$\begin{aligned} x' &= \gamma(x + vt) \\ y' &= y \\ z' &= z \quad \textbf{Lorentz Transformation} \\ t' &= \gamma\left(t - \frac{vx}{c^2}\right) \end{aligned} \quad (18.34)$$

where

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

These equations were derived by Lorentz in connection with electromagnetic phenomena. Einstein pointed out that they have a universal character because they involved only spatial coordinates and time.

The inverse transformation from reference frame S' to S is given by

$$\begin{aligned} x &= \gamma(x' + vt) \\ y &= y' \\ z &= z' \quad \textbf{Lorentz Inverse Transformation} \\ t &= \gamma\left(t' - \frac{vx'}{c^2}\right) \end{aligned} \quad (18.35)$$

The equations (18.34) and (18.35) demonstrate that the space and time are intimately related to each other. What is a pure space interval or a pure time interval in one frame becomes a mixture of space and time intervals in another frame of reference. One important consequence of this mixing of space and time intervals is that the simultaneity of two events depends on the reference frame.

For the sake of comparison, the Galilean and Lorentz transformations are listed below:

Galilean Transformation	Lorentz Transformation
$x' = x - vt$	$x' = \frac{x - vt}{\sqrt{1 - v^2/c^2}}$
$y' = y$	$y' = y$
$z' = z$	$z' = z$
$t' = t$	$t' = \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}}$

It is easy to see from the table that the Lorentz transformation reduces to the Galilean transformation when speeds of motion of one inertial frame with respect to another frame are small compared to c . i.e., $v \ll c$.

(b) Velocity Transformation

Let us again consider two inertial frames S and S' moving with a relative velocity v along XX' axes. Let us suppose that a body moving with a constant velocity u is observed in the reference frame S . Its positions are at x'_1 at time t'_1 and x'_2 at time t'_2 as measured by an observer in S' . Therefore, its speed u'_x measured in the system S' is given by,

$$u'_x = \frac{x'_2 - x'_1}{t'_2 - t'_1} = \frac{dx'}{dt'} \quad (18.36)$$

Differentiating Eq. (18.32), we get

$$dx' = \frac{dx - vdt}{\sqrt{1 - v^2/c^2}}$$

Differentiating Eq. (18.33), we get

$$dt' = \frac{dt - (v/c^2)dx}{\sqrt{1 - v^2/c^2}}$$

Using the expressions for dx' and dt' into Eq. (18.36), we obtain

$$u'_x = \frac{dx'}{dt'} = \frac{dx - vdt}{dt - (v/c^2)dx} = \frac{(dx/dt) - v}{1 - (v/c^2)(dx/dt)}$$

But $\frac{dx}{dt} = u_x$ is the velocity of the body measured in S frame,

$$\therefore u'_x = \frac{u_x - v}{1 - u_x v/c^2} \quad (18.37)$$

It can also be shown that

$$u_x = \frac{u'_x + v}{1 + u'_x v/c^2} \quad (18.38)$$

If the body has velocity components along y and z directions of S frame, we can find the corresponding components in S' as follows:

$$\begin{aligned} y' &= y \\ dy' &= dy \\ \frac{dy'}{dt'} &= \frac{dy\sqrt{1-v^2/c^2}}{dt - (v/c^2)dx} = \frac{(dy/dt)\sqrt{1-v^2/c^2}}{1-(v/c^2)(dx/dt)} \\ \text{or } u'_y &= \frac{u_y\sqrt{1-v^2/c^2}}{1-u_xv/c^2} \end{aligned} \quad (18.39)$$

Similarly, we can show that,

$$u'_z = \frac{u_z\sqrt{1-v^2/c^2}}{1-u_xv/c^2} \quad (18.40)$$

The following important conclusions may be drawn from the Lorentz transformation:

- (i) The Lorentz coordinate transformation differs drastically from the Galilean transformation. However, in the limiting case $v \ll c$, both the transformation laws coincide. It means that the theory of relativity does not reject the Galilean transformation but includes it as a special case.
- (ii) It may be seen that at $v > c$ the radicands of the Lorentz transformation become negative and the formulae lose physical meaning. It means that bodies cannot move with a velocity exceeding that of light in a vacuum. Thus no mechanical or electromagnetic agent can transport energy from one point to another with a speed exceeding c .
- (iii) If $v = c$, the radicands become equal to zero and in this case also, the formulae do not have physical meaning. It means that there is no reference frame in which a photon is at rest.
- (iv) The transformation law for the velocities differs radically from the ordinary addition rule known in classical mechanics. In the relativistic case $v \approx c$, the sum of the components of the velocity along the x -axis is multiplied by the quantity $\left(1+u'_x v/c^2\right)^{-1}$ which is dependent on u'_x , v and c . The components of the velocity along y -and z -axes also change, because the time intervals change and $dt \neq dt'$.
- (v) The Newtonian addition rule for the velocity does not apply to the speed of light. If we have $u'_x = c$ for the speed of light in system S' , then in system S we also have $u_x = c$.

$$u_x = \frac{u'_x + v}{1 + u'_x v/c^2} = \frac{c + v}{1 + cv/c^2} = c$$

- (vi) If a light pulse propagates along the y' -axis in S' frame ($u'_y = c$), then the components of velocity of light in system S are

$$u_y = \frac{u'_y \sqrt{1-v^2/c^2}}{1 + u'_x v/c^2} = c \sqrt{1-v^2/c^2}$$

and

$$u_x = v$$

$$u_z = 0$$

Consequently, a ray of light, which is normal to the x' -axis in system S' , has a different direction in system S. Of course, the magnitude of the speed of light remains c .

$$\sqrt{u_x^2 + u_y^2} = \left[v^2 + c^2 \left(1 - v^2/c^2 \right) \right]^{1/2} = 0$$

This change of the direction of the rays of light accounts for the aberration of starlight, which is an observed seasonal change in the position of stars. It appears due to the Earth's orbital motion.

18.12.1 Relativistic law of addition of velocities

According to equ.(18.37), the x' -component of velocity of the body in the reference frame S'

is given by $u'_x = \frac{u_x - v}{1 - u_x v/c^2}$ and that of the body in the reference frame S is $u_x = \frac{u'_x + v}{1 + u'_x v/c^2}$.

Now, if u' is along the x' -axis, $u'_x = u'$, $u'_y = 0$ and $u'_z = 0$. Therefore, the above equations may be rewritten as

$$u' = \frac{u - v}{1 - (uv/c^2)} \text{ and } u = \frac{u' + v}{1 + (u'v/c^2)} \quad (18.41)$$

Equ.(18.41) is known as the **relativistic law of addition of velocities**. In classical mechanics, the law of addition of velocities is simply $u = u' + v$ and $u' = u - v$.

Example 18.1. Prove that $x^2 + y^2 + z^2 - c^2 t^2$ is invariant under Lorentz transformation.

Solution: We have $x = \frac{x' + vt'}{\sqrt{1 - (v^2/c^2)}}$, $y = y'$, $z = z'$ and $t = \frac{t' + (x'v/c^2)}{\sqrt{1 - (v^2/c^2)}}$. Substituting

these values in the equation given, we obtain

$$\begin{aligned} x^2 + y^2 + z^2 - c^2 t^2 &= \frac{(x' + vt')^2}{1 - (v^2/c^2)} + (y')^2 + (z')^2 - \frac{c^2 [t' + (x'v/c^2)]^2}{1 - (v^2/c^2)} \\ \text{or} \quad &= (y')^2 + (z')^2 - \frac{1}{1 - v^2/c^2} \left[c^2 (t')^2 + \frac{v^2 (x')^2}{c^2} - (x')^2 - v^2 (t')^2 \right] \\ &= (y')^2 + (z')^2 - \frac{1}{1 - (v^2/c^2)} \left[(c^2 t'^2 - x'^2) \left(1 - \left(\frac{v^2}{c^2} \right) \right) \right] \\ &= (y')^2 + (z')^2 - (c^2 t'^2 - x'^2) \\ &= x'^2 + y'^2 + z'^2 - c^2 t'^2 \end{aligned}$$

$$\therefore x^2 + y^2 + z^2 - c^2 t^2 = x'^2 + y'^2 + z'^2 - c^2 t'^2$$

That is, $x^2 + y^2 + z^2 - c^2 t^2$ is invariant under Lorentz transformation.

Example 18.2. Two particles come towards each other with speed $0.7c$ with respect to laboratory. What is their relative speed?

Solution: The relative speed is given by $u = \frac{u' + v}{1 + (u'v/c^2)}$. Here $u' = v = 0.7c$.

$$\therefore u = \frac{2 \times 0.7c}{1 + \left(0.7c \times 0.7c / c^2\right)} = \frac{1.4c}{1.49} = 0.94c.$$

18.13 CONSEQUENCES OF SPECIAL RELATIVITY

The predictions of the special theory of relativity are very strange and startling. The strange effects, which we call *relativistic effects*, appear to conflict with our commonsense. Actually, relativistic effects require speeds that are close to the velocity of light. The relative velocities in our everyday experience are far smaller than the velocity of light. Therefore, we do not commonly observe the strange effects. The special relativity theory predicts that an observer will measure different times and length in different inertial frames. An observer finds that a clock in the moving frame appears to run more slowly. When comparing two events, he finds the events, which are simultaneous in his frame, occur at different times in a moving frame. The observer further finds moving bodies contract along the line of motion. For instance, a meter-stick is measured to be shorter in the direction of motion in a moving frame. It means that we must give up the intuitively obvious notion of absolute time and absolute length on which Newtonian mechanics is based. The new mechanics based on relativity is called **relativistic mechanics**. In relativistic mechanics, there is no such thing as absolute time and absolute length. Events that occur simultaneously at different locations in one frame are not simultaneous in another frame.

It should not be construed from the above discussion that Newtonian mechanics is incorrect. Newtonian mechanics is very much correct and valid for macroscopic bodies moving with speeds much less than c . However, in case of the motion of high-speed particles, the relativistic mechanics is to be employed. In fact, the laws of relativistic mechanics become the same as Newton's laws for cases of $v < c$. It means that Newtonian mechanics is the limiting case of relativistic mechanics.

18.14 SIMULTANEITY OF EVENTS

Before proceeding to analyze the consequence of relativity, let us first understand the concept of simultaneity of two events occurring at different places.

The location of the point at which an event occurs is expressed in Cartesian coordinates. The corresponding moment of time can be determined by means of a clock placed at that point of the reference frame where the given event occurs. This method is not satisfactory when we have to compare events occurring at different points.

When two events occurs approximately at the same place, their simultaneity can be detected by means of a simple observation. When they occur at points A and B removed from each other, one has to compare the time indicated on clocks kept at A and B. There is no sense in comparing the time indicated on the two clocks unless they were synchronized before hand. For this purpose, one can bring the clocks together to some point, make them synchronous and take them apart to points A and B.

A more elegant method to synchronize the clocks positioned at different points of a reference frame is by using light or radio signals. It can be done as follows: an observer located at the origin O of a given reference frame sends a signal at the moment t_o as per his clock. At the moment when this signal reaches the clock located at A at a known distance r from the point O, the clock is set such that it registers $t = t_o + r/c$ which takes into account the signal delay. The same procedure is followed at different points of the reference frame. The repetition of signal after specific time intervals permits all observers to synchronize the rate of

their clocks with that at the origin O. Having done this, we can claim that all the clocks of the reference frame show the same time at each moment.

In classical mechanics, temporal relationship between events is assumed to be independent of the reference frame. It implies that if two events, which occur simultaneously in one reference frame, will also be simultaneous in all other frames moving with constant velocity relative to the first frame.

According to Einstein, time and time interval measurements depend on the reference frame in which they are made.

Let us assume that two events A and B occur in the stationary reference frame S at $x_A = -X$ and $x_B = +X$ at time $t_A = t_B = t$. According to Eq. (18.33) event A occurs at

$$t'_A = \frac{t - (\nu/c^2)(-X)}{\sqrt{1-\nu^2/c^2}} = \frac{t + \nu X/c^2}{\sqrt{1-\nu^2/c^2}} \quad (18.42a)$$

$$\text{And event B occurs at } t'_B = \frac{t - (\nu/c^2)(X)}{\sqrt{1-\nu^2/c^2}} = \frac{t - \nu X/c^2}{\sqrt{1-\nu^2/c^2}} \quad (18.42b)$$

It is seen that $t'_A \neq t'_B$. Event B occurs before event A in the moving frame and consequently, an observer in frame S' find that the two events did not occur simultaneously.

Two events that are simultaneous in one inertial frame of reference are in general not simultaneous in another reference frame moving with respect to the first.

Therefore, simultaneity is not an absolute concept, but is a relative concept.

Einstein devised the following thought experiment to explain this concept of relativity of events. Let us consider a railcar moving with uniform velocity. Let two lightning bolts strike at its two ends as shown in Fig. 18.12. They leave marks on the railcar as well as on the ground. Let A and B be the marks left on the ground and A' and B' be the marks on the railcar. An observer is on the ground at O midway between A and B. An observer at O' midway between A' and B' is moving with the railcar. If the two light signals reach the observer at O at the same time he concludes that the events at A and B occurred simultaneously.

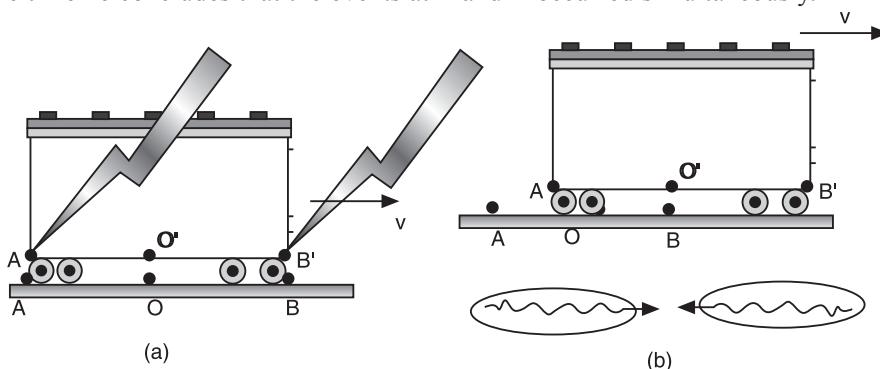


Fig. 18.12: A moving rail car is hit by two lightning bolts (a) The events appear to be simultaneous to the observer O on the ground. (b) The events appear to have occurred at different times to the observer O' in the rail car.

However, the observer O' observes that the light signal from B' reaches him before the light signal coming from A' , because he is moving forward with the railcar. Therefore, he concludes that the front of the railcar was struck before the rear of the railcar was. Thus events, which appear simultaneous in one reference frame, appear not to be simultaneous in another inertial frame.

18.15 LENGTH CONTRACTION

Let us consider a rigid rod at rest in a moving frame S' , say a spaceship or railcar moving along x -axis with a speed v . The length of the rod is equal to the difference between the coordinates of its two end points. Thus, an observer in the spaceship measures the length of the rod to be

$$L_0 = x'_2 - x'_1 \quad (18.43)$$

L_0 is called the **proper length** of the rod. The *proper length* of a body is defined as the length of the body measured in the reference frame in which the body is at rest.

An observer on the Earth frame S will have to determine the length of the rod by marking the forward end position and the rearward end position of the rod at the same instant of time in his stationary frame S . Thus, he will mark and measure the coordinates of rod as x_1 and x_2 in his frame. According to Lorentz coordinate transformation Equ. (18.32)

$$\begin{aligned} x'_1 &= \frac{x_1 - vt}{\sqrt{1-v^2/c^2}} \\ x'_2 &= \frac{x_2 - vt}{\sqrt{1-v^2/c^2}} \\ x'_2 - x'_1 &= \frac{x_2 - x_1}{\sqrt{1-v^2/c^2}} \\ \text{or } L_0 &= \frac{L}{\sqrt{1-v^2/c^2}} \end{aligned} \quad (18.44)$$

where $L = x_2 - x_1$ is the length of the rod as measured by an observer in stationary reference frame S . It follows that

$$L = L_0 \sqrt{1-v^2/c^2} \quad (18.45)$$

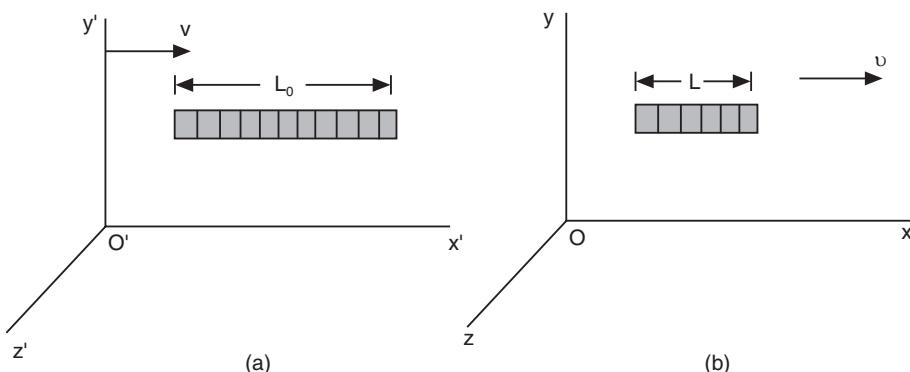


Fig. 18.13

Since $\sqrt{1-v^2/c^2}$ is always less than 1, L is less than L_0 . It means that, if an observer at rest with respect to a body measures its length to be L , an observer moving with a relative speed v with respect to the body will find it shorter than its proper length by a factor $\sqrt{1-v^2/c^2}$ as shown in Fig. 18.13. Hence as measured from the Earth, the rod in the spaceship is contracted. This effect is symmetric; a rod at rest on Earth will undergo length contraction when measured from spaceship. This effect is known as the **length contraction**. The length contraction occurs only along the direction of motion. Thus in the case discussed above, the

rod contracts only along its length, i.e., in x direction, whereas it does not suffer any change in its width and thickness along y and z directions. As compared to the shape of a body in the reference frame where it is at rest, its shape in the moving reference frame would be an oblate in the direction of motion, as shown in Fig. 18.14.

It follows from Eq.(18.45) that the degree of contraction depends on the velocity v . Therefore, in different inertial reference frames the length of the same rod turns out to be different. In other words, length is a relative notion.

Fitzgerald and Lorentz proposed independently the idea of contraction in length long before Einstein's special theory of relativity was proposed, to explain the negative result of Michelson-Morley experiment. Hence, the length contraction is also known as Lorentz – Fitzgerald contraction.

The length contraction has not been tested directly by experiments, as there is no practical method for the high precision measurement of the length of a fast moving body.

Example 18.3. The length of a rocket ship is 100 m long on the ground. During its flight, the apparent length is found to be 99 m when measured from the ground. What is its speed?

Solution:

$$L = L_0 \sqrt{1 - v^2 / c^2}$$

$$\therefore v = c \sqrt{1 - (L/L_0)^2}$$

$$\therefore v = 3 \times 10^8 \text{ m/s} \left[1 - \left(\frac{99}{100} \right)^2 \right]^{1/2} = 4.2 \times 10^7 \text{ m/s}$$

Example 18.4. Calculate the percentage contraction in the length of a rod moving with a speed of $0.8c$ in a direction at an angle 60° with its own length.

Solution: Let the rod of length L_0 be at rest in the reference frame S. Let S' be a reference frame which moves with a speed of $0.8c$ in a direction making an angle 60° with x-axis. The component of L_0 along the x'-axis of S' frame will be $L_0 \cos 60^\circ$. Since S' is moving with a speed of $0.8c$, the component L'_x is given by

$$\begin{aligned} L'_x &= L_0 \cos 60^\circ \sqrt{1 - v^2 / c^2} \\ &= \frac{1}{2} L_0 \sqrt{1 - 0.64} = 0.3 L_0 \end{aligned}$$

The component of L_0 along the y'-axis of S' frame will be $L_0 \sin 60^\circ$.

Therefore,

$$L'_y = L_0 \sin 60^\circ = \sqrt{3} L_0 / 2$$

Now,

$$L' = \sqrt{L'^2_x + L'^2_y} = \sqrt{\left(0.3 L_0\right)^2 + \left(\sqrt{3} L_0 / 2\right)^2} = 0.87 L_0$$

$$\% \text{ contraction} = \frac{L_0 - L'}{L_0} \times 100 = \frac{L_0 - 0.87 L_0}{L_0} \times 100 = 13\%.$$

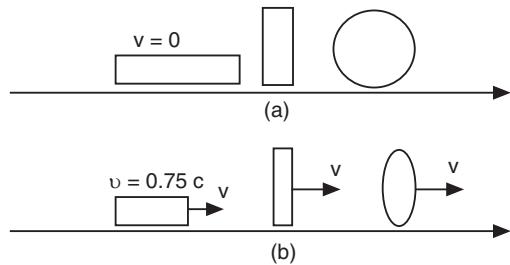


Fig. 18.14

18.16 THE TIME DILATION

Let us now compare the rate of time flow in different inertial reference frames, by making use of a so-called light clock.

Let us consider a railcar S moving to the right with a speed v , as shown in Fig.18.15. Let the car have a mirror fixed to its ceiling as shown in Fig.18.15. An observer at rest in the car sends off a light pulse at some instant toward the mirror. The light pulse travels (upward) at the speed c gets reflected at the mirror at the top and comes back to the observer. Let us say the observer measures the time interval $\Delta t'$ for the round trip of the pulse with the help of a clock. As $\Delta t'$ is the time measured by a stationary observer in the S' frame, it is the **proper time** T_0 . $\Delta t'$ is given by

$$\Delta t' = \frac{2d}{c} \quad (18.46)$$

where d is the distance between the source of light pulse and the mirror at the ceiling.

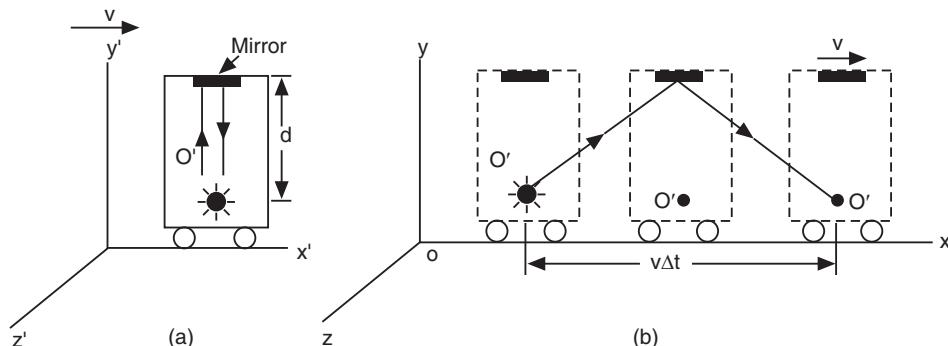


Fig. 18.15

From the point of view of an observer on the ground (stationary frame S), the mirror and the light source are moving to the right with a speed v . By the time the light pulse hits the mirror, the mirror will have moved a distance equal to $v(\Delta t/2)$ where Δt is the time taken by the light pulse for its round trip as measured in S frame. From the geometry of the situation, we find that

$$\left[\frac{c(\Delta t)}{2} \right]^2 = \left(\frac{v\Delta t}{2} \right)^2 + d^2$$

$$(c^2 - v^2)(\Delta t)^2 = (2d)^2$$

$$\therefore \Delta t = \frac{2d}{\sqrt{c^2 - v^2}} = \frac{2d/c}{\sqrt{1 - v^2/c^2}}$$

But $2d/c = \Delta t'$

$$\therefore \Delta t = \frac{\Delta t'}{\sqrt{1 - v^2/c^2}}$$

Designating $\Delta t = T$, we rewrite the above equation as

$$T = \frac{T_0}{\sqrt{1 - v^2/c^2}} \quad (18.47)$$

It follows that a stationary clock measures a longer time interval between events occurring in a moving frame of reference than does a clock in the moving frame. Therefore, from the point of view of the observer in the stationary frame, events will be found to be happening at a slower rate in the moving frame. In other, he concludes that a moving clock runs slower compared to an identical stationary clock by a factor $\sqrt{1 - v^2/c^2}$. This relative showing of time is known as **time dilation**.

If an observer on the platform watches the events taking place in the moving railcar, he gets the impression that the clock in the car runs slower than his own. His clock measures a longer time between two events taking place in the car than the clock in the car. Conversely, if an observer is in a moving railcar and watches the events on the platform he gets the impression that the clock on the platform runs slower than clock in the railcar. From the point of view of each observer, the moving clocks slow down as compared to his clock. We can generalize by saying that all physical processes including chemical and biological reactions slow down relative to a stationary clock when they are in a moving frame.

Thus, the rate of time flow actually depends on the state of motion. There is no such thing as universal time and the concept of time is relative.

18.16.1 Experimental Evidence

Time dilation has been verified by various experiments. We consider here two of them.

(a) Decay of Muons

One of the direct experimental evidences of time dilation was provided by the tests conducted by B. Rossi and D.B. Hall in 1941. Rossi and Hall measured the flux of one of the elementary particles known as muons at an altitude of about 2000 m above sea level and again at sea level. Muons are created in upper layers of earth's atmosphere as a result of cosmic ray collisions with the air molecules. From there they approach the earth with speeds of the order of velocity of light ($\sim 0.998 c$). Muons are unstable and quickly decay into other particles.

The average lifetime of a muon at rest in the laboratory is about $2 \mu s$. During this time, even traveling at the speed of light, a muon would travel only a distance of $d = ct = (3 \times 10^8 \text{ m/s}) (2 \times 10^{-6} \text{ s}) = 600 \text{ m}$. Therefore, muons created at high altitudes of several kilometers are expected to decay away before reaching the earth. However, measurements showed that an appreciable number of muons do reach the earth's surface (the sea level). The finding can be explained using the concept of time dilation.

The muon decays by its own clock. Relative to an observer on the Earth, muons have a lifetime equal to $\tau / \sqrt{1 - v^2/c^2}$. At a speed of the order of $0.998 c$, $\frac{1}{\sqrt{1 - v^2/c^2}} = 15$. Therefore, time on earth clock is $t = \gamma \tau = 15 (2 \mu s) = 30 \mu s$.

According to an observer on Earth, the average distance travelled by the muon is

$$d = vt = (0.998c)(30 \mu s) = 8980 \text{ m.}$$

Therefore, an appreciable number of muons reach the earth before decaying.

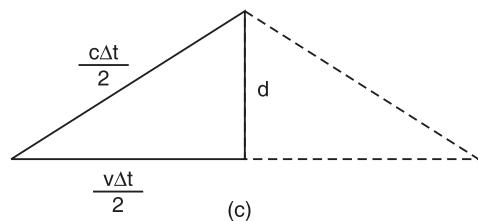


Fig. 18.15(c): The triangle for calculating the relationship between Δt and $\Delta t'$

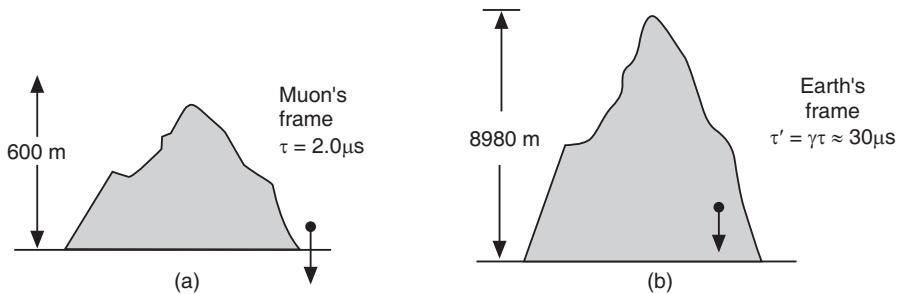


Fig.18.16

The above result can be interpreted from the viewpoint of muons as follows:

For muons moving at a speed close to c , the terrestrial measures of length appear greatly contracted in the direction of their motion. Because of length contractions, the muon measures the distance to be

$$L = L_0 \sqrt{1 - v^2/c^2} = 599 \text{ m}$$

This distance can be covered by the muon in its lifetime, i.e., in $2 \mu\text{s}$. Therefore, the existence of appreciable number of muons at the sea level is the direct consequence of time dilation.

(b) Experiment with Clocks

Further experimental evidence for the phenomenon of time dilation was provided by the tests conducted by NASA. In 1971, J.C. Hafele, as astronomer and R.E. Keating, a physicist circled the earth twice in a jet plane, once from east to west for two days and then from west to east for two days carrying two cesium-beam atomic clocks capable of measuring time to a nanosecond. After the trip the clocks were compared with identical clocks. The clocks on the plane lost $59 \pm 10 \text{ ns}$ during their eastward trip and gained $273 \pm 7 \text{ ns}$ during the westward trip. These results are consistent with Einstein's theory of time dilation.

Example 18.5. The proper life of π^- mesons is $2.5 \times 10^{-8} \text{ s}$. What is the velocity of these mesons if the observed mean life is $2.5 \times 10^{-7} \text{ s}$?

Solution.

$$\begin{aligned} \Delta t &= \frac{\Delta t}{\sqrt{1 - v^2/c^2}} \\ \therefore 2.5 \times 10^{-7} \text{ s} &= \frac{2.5 \times 10^{-8} \text{ s}}{\sqrt{1 - v^2/c^2}} \\ \text{or } \sqrt{1 - v^2/c^2} &= \frac{1}{10} \quad \therefore v = 0.995 c. \end{aligned}$$

Example 18.6. The average lifetime of a free neutron at rest is 15 minutes. It disintegrates spontaneously into an electron, proton and neutrino. What is the average velocity with which a neutron must leave the Sun in order to reach the earth before breaking up? Given the distance of Earth from Sun is $11 \times 10^{10} \text{ m}$.

Solution. Let $t_o = 15 \times 60 \text{ s}$ be the proper lifetime of the neutron. The average lifetime t of the moving neutron, as measured by an observer on the earth is $t = \frac{t_o}{\sqrt{1 - v^2/c^2}}$. Also, t is the time for the neutron to reach the earth after leaving the sun.

Hence,

$$t = \frac{d}{v} = \frac{11 \times 10^{10} \text{ m}}{v}$$

∴

$$\frac{15 \times 60 \text{ s}}{\sqrt{1 - v^2/c^2}} = \frac{11 \times 10^{10} \text{ m}}{v}$$

or

$$\frac{v}{\sqrt{1 - v^2/c^2}} = \frac{11 \times 10^{10} \text{ m}}{900 \text{ s}}$$

or

$$v^2 = \left[\frac{11 \times 10^{10} \text{ m}}{900 \text{ s}} \right]^2 \left(1 - \frac{v^2}{c^2} \right)$$

$$= 1.49 \times 10^{16} \text{ m}^2/\text{s}^2 \left(1 - \frac{v^2}{(3 \times 10^8)^2} \right)$$

$$= 0.16 (9 \times 10^{16} - v^2) \text{ m}^2/\text{s}^2$$

$$= \frac{0.16 \times 9 \times 10^{16}}{1.16} \text{ m}^2/\text{s}^2 = 1.24 \times 10^{16} \text{ m}^2/\text{s}^2$$

∴

$$v = 1.11 \times 10^8 \text{ m/s.}$$

18.17 THE TWIN PARADOX

The time dilation effect leads to the famous “**Twin Paradox**” of special relativity. Let us consider a hypothetical experiment involving twin sisters Seetha and Geetha. After celebrating their 20th Birthday, the adventurous twin Geetha sets out on a space voyage leaving behind her sister Seetha on Earth. Her spaceship travels at a speed close to the speed of light ($v = 0.998 c$). After a year she returns back to the Earth to celebrate her 21st birthday. On return she is shocked to find her sister Seetha to be a 70 years old feeble lady while she herself is only 21 years old. It means while Geetha has aged only one year, her twin sister has aged about 50 years. This is a paradox, which challenges our common sense.

As reckoned in the reference frame of Earth, the clocks and all the physical and biological processes on the spaceship slow down due to time dilation. Hence, Geetha ages at a slower rate and at the end of her trip, when she joins her twin on Earth, she will be younger than Seetha. But this explanation appears to contradict the postulate of special theory of relativity, which asserts that all inertial frames are equivalent. The aging effect seems to provide a way of distinguishing among the frames. Further, from the point of view of the spaceship, the Earth is in motion and hence the clocks and biological processes on the Earth are slowed down due to time dilation. This leads us to the conflicting conclusion that Seetha ages at a slower rate than Geetha and when the twins are united, Seetha will be younger than Geetha.

All these contradictions disappear when we take into account of the fact that Geetha the space traveler, was not in an inertial frame most of the time. She experienced a series of accelerations when leaving the Earth and decelerations when coming back to the Earth. Thus, she was in an accelerated frame for a greater part of her trip. Hence, the predictions based on special theory of relativity are not valid in her reference frame. On the other hand, Seetha is in an inertial frame all the time and, therefore, her predictions are reliable. Therefore, Geetha will indeed be younger than Seetha on returning to Earth.

18.18 THE RELATIVISTIC MASS

In Newtonian mechanics, the momentum of a body is defined as $p = mv$ and the mass of the body is supposed to be independent of its velocity. Further, it is supposed that the total momentum of an isolated system is conserved. Because of the immense significance of the conservation laws, the law of conservation of momentum is regarded as fundamental in the theory of relativity.

In order that the total momentum of an isolated system remains constant, it is found that the mass of a body must depend on the velocity of the body. The following relation governs the dependence of mass m on the velocity of the body:

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}} \quad (18.48)$$

where m is called the **relativistic mass** and m_0 is known as the **rest mass** of the body.

It is seen that the relativistic mass is greater than the rest mass. It depends on the velocity of the body. It means that the mass of the same particle is different in different inertial reference frames. The rest mass m_0 , on the other hand, is the same in all reference frames.

At velocities very small in comparison with c , the mass may be regarded as independent of the speed of the body. With increasing speed, the mass of the body steadily increases and requires a steadily increasing force to impart a constant speed to the body. At $v \approx c$, the mass becomes infinite and hence it is impossible to make a body move at the speed of light.

Relativistic mass increases are significant only at speed approaching that of light. At a speed $v = 0.1c$, the mass increases only by 0.5% but at speeds of $0.9c$, the increase is more than 100%. The relativistic effects are of no significance in space flight. Atomic particles such as electrons, protons etc have such high speeds which cause relativistic effects.

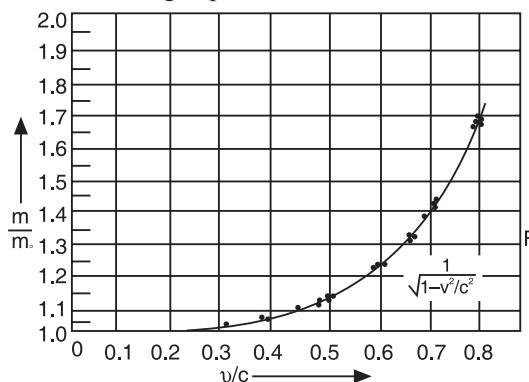


Fig. 18.17

There are experimental observations of the increase in mass of high speeds (see Fig. 18.17). It is found that a greater magnetic field is required to deflect a high-speed charged particle than that would be required if its mass is constant.

Example 18.7: At what speed the mass of an object will be three times of its value at rest?

Solution: Relativistic mass $m = \frac{m_0}{\sqrt{1 - v^2/c^2}}$. Here $m = 3m_0$.

$$\therefore 3m_0 = \frac{m_0}{\sqrt{1 - v^2/c^2}} \text{ or } \sqrt{1 - v^2/c^2} = \frac{1}{3}$$

$$\text{or } 1 - (\frac{v^2}{c^2}) = \frac{1}{9} \quad \frac{v^2}{c^2} = \frac{8}{9} = 0.89$$

$$\therefore v = \sqrt{0.89 \times 3 \times 10^8 \text{ m/s}} = 2.67 \times 10^8 \text{ m/s}$$

18.19 THE RELATIVISTIC MOMENTUM

The momentum of a relativistic particle is given by

$$p = m_0 v = \frac{m_0 v}{\sqrt{1 - v^2/c^2}} \quad (18.49)$$

This is known as the **relativistic momentum** of the particle. The momentum thus defined obeys the law of conservation regardless of the inertial reference frame chosen.

For velocities $v \ll c$, (equ.18.49) yields the classical momentum $p = m_0 v$. The velocity dependence of the classical and relativistic momentum of a particle is depicted in Fig.18.18. The difference between the momenta grows substantially as the velocity of a particle approaches that of light.

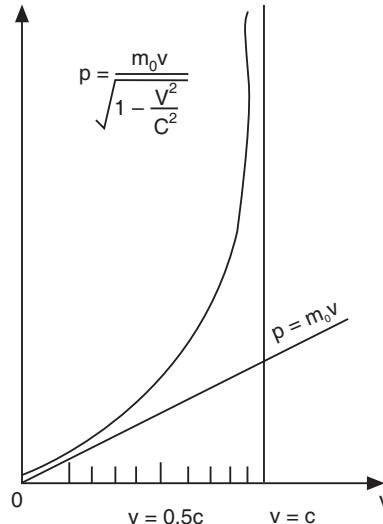


Fig.18.18

18.20 KINETIC ENERGY

In Newtonian mechanics force is defined as the time rate of change of momentum. This definition of force is valid in relativistic mechanics also, but the mass is a variable quantity in this case.

$$F = \frac{d}{dt}(mv) \quad (18.50)$$

The kinetic energy E_k of a moving body may be defined as the work done in accelerating it from rest to its final speed v . That is,

$$E_k = \int_0^s F ds \quad (18.51)$$

where F is the force acting in the direction of displacement.

$$\therefore E_k = \int_0^s \frac{d}{dt}(mv) ds = \int_0^m v d(mv) = \int_0^v v d \left[\frac{m_0 v}{\sqrt{1 - v^2/c^2}} \right]$$

Integrating the above equation by parts, we get

$$E_k = \frac{m_0 v^2}{\sqrt{1-v^2/c^2}} - m_0 \int_0^v \frac{v \, dv}{\sqrt{1-v^2/c^2}} = \left[\frac{m_0 v^2}{1-v^2/c^2} + m_0 c^2 \sqrt{1-v^2/c^2} \right]_0^v$$

or

$$E_k = \frac{m_0 c^2}{\sqrt{1-v^2/c^2}} - m_0 c^2 \quad (18.52)$$

∴

$$E_k = mc^2 - m_0 c^2 \quad (18.53)$$

or

$$E_k = (m - m_0) c^2 \quad (18.54)$$

It follows that the increase in kinetic energy of the body is due to the increase in mass. Equation (18.52) may be rewritten as

$$E_k = m_0 c^2 \left[\frac{1}{\sqrt{1-v^2/c^2}} - 1 \right]$$

This is the expression for the relativistic kinetic energy of a particle. At low velocities ($v \ll c$). We can write using binomial theorem that

$$\frac{1}{\sqrt{1-v^2/c^2}} = 1 + \frac{v^2}{2c^2} + \frac{3v^4}{8c^4} + \dots$$

At $v \ll c$, $v^2/c^2 \ll 1$ and we can ignore the higher terms in the series. Thus,

$$\frac{1}{\sqrt{1-v^2/c^2}} \approx 1 + \frac{v^2}{2c^2}$$

$$\therefore E_k = m_0 c^2 \left[1 + \frac{v^2}{2c^2} - 1 \right]$$

$$\text{or } E_k = \frac{1}{2} m_0 v^2$$

Thus, at velocities $v \ll c$, the relativistic formula reduces to the classical formula. The relativistic kinetic energy E and classical K.E. are plotted as a function of v/c in Fig. 18.19. Their difference becomes very large at $v \approx c$.

18.21 MASS-ENERGY EQUIVALENCE

We rewrite equation (18.53) as $mc^2 = E_k + m_0 c^2$.

If we interpret mc^2 as the total energy E of the

body, equ.(18.55) indicates that the total energy of a freely moving body consists of its rest energy plus its energy due to motion E_k . When the body is at rest, it possesses the energy $m_0 c^2$. Therefore, $m_0 c^2$ is called the **rest energy** E_o .

$$E = E_o + E_k$$

where

$$E = mc^2 \quad (18.56)$$

When $E_k = 0$, the body is motionless. However, it possesses the energy

$$E_o = m_0 c^2 \quad (18.57)$$

The above equations are known as **mass-energy equivalence** relations. They imply that energy manifests as mass. This is one of the most remarkable results of Einstein's theory. Till the special theory of relativity was propounded, energy and mass were believed to be independent. The theory of relativity showed that mass is a form of energy.

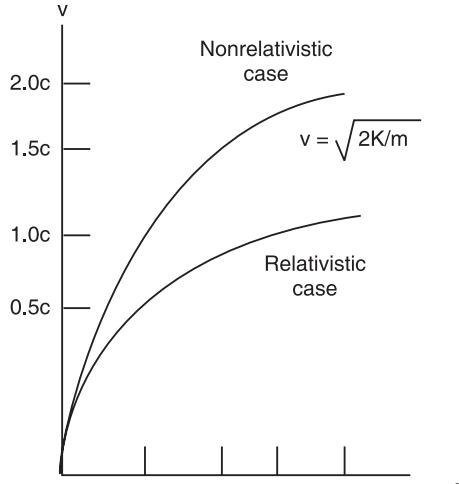


Fig.18.19

The validity of the equation (18.56) is evident in nuclear processes such as fusion and fission.

Example 18.8: *The sun radiates away energy at the rate of $4 \times 10^{26} \text{ J/s}$. Calculate the rate at which its mass is decreasing.*

Solution: The change in mass $\Delta m = \frac{\Delta E}{c^2} = \frac{4 \times 10^{26} \text{ J/s}}{9 \times 10^{16} \text{ m}^2/\text{s}^2} = 4.4 \times 10^9 \text{ kg}$.

18.22 RELATION BETWEEN MOMENTUM AND ENERGY

It is clear that both the energy E and the momentum of a particle have different values in different reference frames. However, the quantity $(E^2 - p^2 c^2)$ is invariant and has the same value in different reference frames.

$$\begin{aligned} E &= mc^2 \\ \text{or} \quad E &= \frac{m_0 c^2}{\sqrt{1-v^2/c^2}} \end{aligned} \quad (18.58)$$

Squaring the equation (18.58), we get

$$\begin{aligned} E^2 &= \frac{m_0^2 c^4}{1-v^2/c^2} \\ \text{or} \quad E^2 (1-v^2/c^2) &= m_0^2 c^4 \\ E^2 - \frac{E^2 v^2}{c^2} &= m_0^2 c^4 \\ E^2 &= m_0^2 c^4 + \frac{E^2 v^2}{c^2} = m_0^2 c^4 + m^2 c^4 (v^2/c^2) \\ E^2 &= m_0^2 c^4 + c^2 (mv)^2 \end{aligned}$$

Denoting mv by p , we write

$$E^2 = p^2 c^2 + m_0^2 c^4 \quad (18.59)$$

$$\therefore E = c \sqrt{p^2 + m_0^2 c^2} \quad (18.60)$$

We can also write Eq.(18.59) as

$$\begin{aligned} E^2 &= E_0^2 + p^2 c^2 \\ \text{or} \quad E^2 - p^2 c^2 &= E_0^2 \end{aligned}$$

The above equation does not contain v , which implies that the quantity $(E^2 - p^2 c^2)$ is independent of the velocity of the particle. It has the same value E_0^2 in all inertial reference frames.

Alternately, we can express momentum p in terms of energy E as

$$p = \frac{\sqrt{E^2 - E_0^2}}{c} \quad (18.61)$$

18.23 DOPPLER EFFECT IN LIGHT

The frequency of light radiation changes when the light source or the observer moves with respect to one another. This is known as Doppler effect. Johann Christian Doppler (1803-1853) discovered the effect in case of sound waves. The response of sound waves to moving bodies is illustrated in the example of the sounding of the whistle of a moving train. When a train, at

rest in the station, blows the whistle, **stationary** listeners, who are either ahead of the engine or behind it, will hear the same pitch made by the whistle. But as the train **moves**, those who are **ahead** will hear the sound of the whistle at a **higher** pitch. Listeners **behind** the train, as it pulls further away from them, hear a **lower** pitch. The faster the train moves the greater will be the effect of the rising and falling of the pitch.

Also, if the train remains at rest but the listeners either move toward the sounding train whistle or away from it, the effect will be the same. Those who move toward the train will hear a higher pitch, while those who travel away from the train will hear a lower pitch.

When a sound source is moving towards a stationary observer, the apparent frequency is given by

$$v' = v \left[\frac{c - v}{c} \right] \quad (18.62)$$

On the other hand, when the observer is moving towards a stationary sound source, the apparent frequency is given by

$$v' = v \left[\frac{c + v}{c} \right]$$

Thus, the Doppler effect is asymmetric in case of sound waves. Sound waves require a material medium for their propagation whereas light waves do not require a medium. Therefore, the Doppler effect is symmetric in case of light and the apparent frequency is the same when either the light source moves towards the stationary observer or the observer moves towards a stationary source.

- (i) Let us suppose that the observer is stationary and the light source is moving towards the observer with a velocity v . As there is no medium in between, the observer will receive more waves due to the motion of the source. But the wavelength of the waves is not changed. The apparent frequency is given by

$$v' = v + \frac{v}{\lambda}$$

As, $c = v\lambda$, we write the above relation as

$$v' = v + \frac{vv}{c}$$

or $v' = v \left[1 + \frac{v}{c} \right] \quad (18.63)$

- (ii) Let the source be stationary and the observer move towards the source with a velocity v . Then, also

$$v' = v + \frac{v}{\lambda}$$

As, $c = v\lambda$, we write the above relation as

$$v' = v + \frac{vv}{c}$$

or $v' = v \left[1 + \frac{v}{c} \right] \quad (18.64)$

The equations (18.63) and (18.64) are similar.

- (iii) Now, let the light source move away from the stationary observer or let the observer move away from the stationary source with a velocity v . Then,

$$\begin{aligned}
 v' &= v - \frac{v}{\lambda} \\
 \text{or} \quad v' &= v - \frac{v\lambda}{c} \\
 \therefore \quad v' &= v \left[1 - \frac{v}{c} \right] \tag{18.65}
 \end{aligned}$$

- (iv) If the source and the observer move towards each other and each moves with a velocity v , the apparent frequency is given by

$$v' = v \left[1 + \frac{v}{c} \right]^2 \tag{18.66}$$

- (v) If the source and the observer move away from each other and each moves with a velocity v , the apparent frequency is given by

$$v' = v \left[1 - \frac{v}{c} \right]^2 \tag{18.67}$$

18.23.1 Applications

1. The Doppler effect in light waves can be observed by the spectral analysis of light emitted by luminous objects such as stars. Using a prism or a diffraction grating, we can spread this light out into a spectrum. If we look at the spectrum of the Sun or any other star, we see not only the rainbow of colors from red, orange and yellow through to violet, but also a distinctive pattern of dark lines. The wavelengths of lines are based on atomic physics and can be measured extremely accurately in a laboratory on Earth.

If a star is moving towards us, the whole pattern of the spectrum gets shifted to shorter wavelengths, i.e. towards the blue end of the spectrum. This is a **BLUESHIFT**, and we can measure it very accurately by comparing the apparent wavelengths of the spectral lines with the known laboratory wavelengths. If the star is receding, the pattern moves to longer, redder wavelengths, and this is a **REDSHIFT**.

Such an analysis of light from distant galaxies shows that the light experiences a red shift. These galaxies are moving away from the Earth. Extrapolation of this evidence, along with other findings, lends support to the “big bang” theory of the origin of the universe.

2. Police use Doppler effect to track speed. Radio waves are transmitted out from radars, collide with a vehicle, and bounce back. The speed of the vehicle (which acts as the source of the reflected wave) determines the change in frequency, which can be detected with the radar.
3. We can measure the masses of binary stars, by detecting the velocity changes as they orbit around their common center of gravity, and combining these velocities with the period of the orbit.

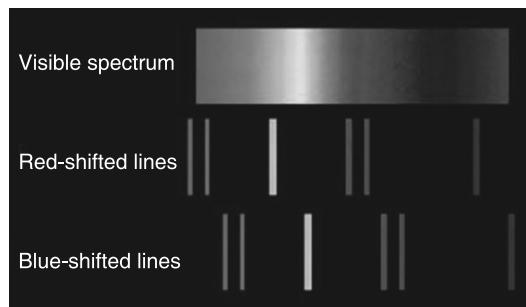


Fig.18.20: If the star is moving towards the observer the light is blueshifted. If the star is moving away from the observer the light is redshifted.

4. In the same way, we can look for evidence of massive planets in orbit around other stars, by searching for the small velocity changes in the central stars caused by their unseen planetary companions.
5. Certain types of stars known as CEPHEID VARIABLES have atmospheres that swell and shrink periodically. Using the Doppler technique, we can tell that the surface of such a star may be moving at speeds up to 50 km/sec. As the period of the pulsation is related to the average brightness of the star, we can tell roughly how far away the star is by combining its apparent brightness with its pulsation period. So the Doppler Effect can help us determine distances as well as speeds.
6. We can detect the rotation of the Sun easily enough by following the positions of SUNSPOTS, but the spectrum of the light from various regions of the Sun also shows this rotation. The equator at the Sun's east limb is approaching us at 2 km/sec, and the west limb is retreating from us at the same speed. The same thing happens in other stars. They are further away, so we cannot study each edge of the star in the same way. However, we can see that the absorption lines are BROADENED, or smeared out, by the rotation. Some stars have equatorial speeds of 400 km/sec.
7. When we look at the spectra of other galaxies and map their velocities, we find that on average, all galaxies are speeding away from each other, with very high redshifts. The Doppler Shift tells us that the universe as a whole is expanding.

Example 18.9. A spectral line of wavelength 4000\AA in the spectrum of light from a star is found to be displaced from its normal position towards the red end of the spectrum by 1\AA . What is the velocity of the star in the line of sight?

Solution. The star is moving away from the stationary observer. Using equ.(18.65), we obtain

$$v' = v \left[1 - \frac{v}{c} \right]$$

$$\frac{c}{\lambda'} = \frac{c}{\lambda} \left[1 - \frac{v}{c} \right] \text{ or } \frac{1}{\lambda'} = \frac{1}{\lambda} \left[1 - \frac{v}{c} \right] \text{ or } \frac{1}{4001\text{\AA}} = \frac{1}{4000\text{\AA}} \left[1 - \frac{v}{3 \times 10^8 \text{ m/s}} \right]$$

or

$$v = 75 \text{ km/s.}$$

Example 18.10. The wavelength 5000\AA of a spectral line emanating from a distant star is observed in the laboratory to be 5200\AA . Is the star approaching the earth or receding from it? Calculate the velocity of approach or recession, as the case may be.

Solution.

$$v' = v \left[1 - \frac{v}{c} \right]$$

$$\therefore \frac{c}{\lambda'} = \frac{c}{\lambda} \left[1 - \frac{v}{c} \right] \text{ or } \frac{1}{\lambda'} = \frac{1}{\lambda} \left[1 - \frac{v}{c} \right] \text{ or } \frac{\lambda}{\lambda'} = 1 - \left(\frac{v}{c} \right)$$

$$\therefore \left(\frac{v}{c} \right) = 1 - \frac{\lambda}{\lambda'} = 1 - \frac{5000\text{\AA}}{5200\text{\AA}} = \frac{1}{26}$$

$$\therefore v = \frac{3 \times 10^8 \text{ m/s}}{26} = 1154 \text{ km/s.}$$

The star is moving away from the stationary observer with a velocity of 1154 km/s.

Example 18.11. A space ship is moving with a velocity of 10^{10} cm/s towards a star that emits radiations of wavelength 6×10^{-5} cm. Calculate the wavelength of the radiations as received by the crew in the ship.

Solution.

$$\text{or} \quad v' = v \left[1 + \frac{v}{c} \right]$$

$$\lambda' = \frac{\lambda c}{c + v} = \frac{(6 \times 10^{-5} \text{ cm})(3 \times 10^{10} \text{ cm/s})}{3 \times 10^{10} \text{ cm/s} + 10^{10} \text{ cm/s}} = 4500 \text{ \AA.}$$

QUESTIONS

1. Define an inertial frame of reference. Distinguish between special theory of relativity and general relativity. **(C.S.V.T.U., 2006)**
2. Describe Michelson-Morley experiment. Show how negative results obtained from Michelson orley experiment were interpreted. **(C.S.V.T.U., 2006, 2007, 2008)**
3. What are non-inertial frames of reference?
4. Give Galileo's principle of relativity. **(C.S.V.T.U., 2006)**
5. State the fundamental postulates of special theory of relativity? **(Univ. of Pune, 2007)**
6. What is proper time?
7. What is proper length?
8. Does $E = mc^2$ apply to particles that travel with the speed of light?
9. Is mass a conserved quantity in special theory of relativity?
10. Show that the laws of mechanics are identical in any inertial reference frame.
11. Describe Michelson-Morley experiment and explain the significance of the null result of the experiment.
12. State fundamental postulates of special theory of relativity. Deduce the Lorentz transformation of space. **(Shivaji Univ.)**
13. State the fundamental postulates of the special theory of relativity and deduce the Lorentz transformation equation of space and time. **(C.S.V.T.U., 2005)**
14. Prove that $x_2^2 + y_2^2 + z_2^2 = c^2 t_2^2$ is invariant under Lorentz transformation. **(C.S.V.T.U., 2006)**
15. What is Galilean transformation? Derive Galilean transformation equations. **(C.S.V.T.U., 2008)**
16. Discuss Galilean transformation for position, velocity and acceleration. **(UPTU, Lucknow)**
17. State the fundamental postulates of the special theory of relativity and deduce the Lorentz transformations for space and time coordinates.
18. State and explain the relativistic law of addition of velocities.
19. Discuss simultaneity of events in two inertial frames of reference. Hence explain the concept of length contraction and time dilation. **(C.S.V.T.U., 2005)**
20. Show that the apparent length of a rigid body is decreased by the factor $\sqrt{1 - v^2/c^2}$ in the direction of its motion.
21. Explain the concept of time dilation. Describe experimental verification of time dilation.
22. Explain the concept of time-dilation and length contraction in theory of relativity. **(Univ. of Pune, 2007)**
23. What is time dilation in special relativity? Obtain an expression for time dilation in regard to the time interval between two events measured from two different inertial frames. **(Sivaji Univ.)**
24. What is time dilation? Deduce an expression for the time dilation as regards to interval between two events measured from two different inertial frames. **(Univ. of Pune, 2007)**
25. What is time dilation in special relativity? Obtain an expression for time dilation in regard to the time interval between two events measured from two different inertial frames. **(C.S.V.T.U., 2006)**

26. What is the length contraction in special theory of relativity? Deduce an expression for length contraction. **(Shivaji Univ.)**
27. Show how the relativistic invariance of the law of conservation of momentum leads to the concept of variation of mass with velocity.
28. Show that the addition of a velocity to the velocity of light gives the velocity of light.
29. Obtain the relativistic formula for the addition of velocities.
30. Derive the formula for the variation of mass with velocity according to special theory of relativity. **(C.S.V.T.U., 2005, 2007)**
31. Deduce Einstein's mass-velocity relation $E = mc^2$, considering the variation of mass with velocity.
32. Deduce Einstein's expression for mass-energy equivalence. **(Univ. of Pune, 2007)**
33. Deduce Einstein's mass-energy relation considering the variation of mass with velocity. **(C.S.V.T.U., 2005, 2008)**
34. Deduce an expression for the variation of mass with velocity. **(UPTU, Lucknow)**
35. State and explain the Doppler's effect of light. **(Shivaji Univ.)**
36. What is Doppler's effect? **(Univ. of Pune, 2007)**

PROBLEMS

- A person on a rocket traveling at a speed of $0.5c$ with respect to the earth observes a meteor coming from behind and passing him at a speed of $0.5c$. How fast is the meteor moving with respect to the earth? **[Ans: $0.8c$]**
- A rocket leaves the earth at a speed of $0.6c$. A second rocket leaves the first at a speed of $0.9c$ with respect to the first. Calculate the speed of the second rocket with respect to the earth if (i) it is fired in the same direction as the first one and (ii) if it is fixed in a direction opposite to the first. **[Ans: $0.974c$, $-0.652c$]**
- A particle with a mean lifetime of $10^{-6}s$ moves through the laboratory at a speed of $0.8c$. What will be its lifetime as measured by an observer in the laboratory? **[Ans: $1.67 \times 10^{-6}s$]**
- A meson having a mass of $2.4 \times 10^{-28} \text{ kg}$ travels at a speed of $0.8c$. What is its kinetic energy? **[Ans: $1.44 \times 10^{-11} \text{ J}$]**
- The mean lifetime of a muon at rest is $2.4 \times 10^{-6}s$. What will be its mean lifetime as measured in the laboratory, if its velocity is $0.6c$ with respect to the laboratory? **[Ans: $3 \times 10^{-6}s$]**
- The π -mesons coming out of an accelerator have a velocity of $0.99c$. If they have a mean lifetime of $2.6 \times 10^{-8}s$ in the rest frame, how far can they travel before decay? **[Ans: 54.6 m]**
- A man leaves the earth in a rocket ship that makes a round trip to the nearest star, which is 4 light years away. If the speed of the rocket ship is $0.8c$, how much younger will he be on his return than his twin brother who preferred to stay behind? **[Ans: 3.4 yrs]**
- Two particles are moving in opposite directions with speeds $0.9c$ as observed in laboratory frame. What is the velocity of one particle relative to the other? **[Ans: $0.994c$]**
- In the laboratory the lifetime of a particle moving with speed $2.8 \times 10^8 \text{ m/s}$ is found to be $2.5 \times 10^{-7}s$. Calculate the proper lifetime of the particle. **[Ans: $8.9 \times 10^{-8}s$]**
- Find the momentum and velocity of an electron having a kinetic energy of 10 MeV. The rest energy of electron is 0.51 MeV. **[Ans: $0.998c$]**

CHAPTER

19

Atomic Physics

19.1 INTRODUCTION

The aim of physics is *to understand the natural phenomena* around us. 19th century witnessed a rapid growth in physics. Newtonian mechanics was nearly perfected and Maxwell synthesized the fields of electricity and magnetism into a single theory which permitted inclusion of optics into the frame work of electromagnetic phenomenon. Towards the end of the 19th century it was generally believed that all that to be discovered in nature was discovered, and all the laws of nature were formulated. The Newtonian mechanics, Maxwell's electromagnetic theory and thermodynamics came to be known later as classical physics. Classical physics was developed assuming that particles are localized and we can observe them without appreciably disturbing them. The three laws of conservation, namely conservation of linear momentum, angular momentum and energy formed the basis for classical mechanics. However, discoveries that followed dispelled such illusions very soon. At about the turn of 20th century a number of fundamental discoveries which could not be explained within the frame work of the existing theories in physics were reported.

Classical Physics

A material particle is a point object having mass and a definite position at any instant of time. The *trajectory* or *path* followed by the particle is then pictured by a sharply defined curve or line. The instantaneous positions of the particle on the trajectory, its velocity, acceleration all have definite numerical values for their components. The position, velocity, momentum, energy etc are called **dynamical variables**. The manner in which they vary with time is governed by the Newton's second law. In classical mechanics, the state of a particle is uniquely defined by specifying its coordinates and the three components of its momentum. At each instant both the coordinates and momentum components are thought to have strictly definite values and can be measured. If the state of a particle is known at any instant of time t_o , the state of the particle at any other time is determined by the Newton's second law. The set of successive positions of a particle moving in space determines its path. An electromagnetic wave is characterized by the instantaneous values of electric and magnetic fields. All properties like energy density, momentum density etc of electromagnetic field are functions of the instantaneous values of E and B. Therefore, the state of the electromagnetic field at that instant is completely specified by E and B for all x at t_o . If the state of the field at some instant t_o is known, the state at any other time can be determined. The classical concepts of particle and wave are mutually *exclusive*.

Quantum Physics

At about the turn of 20th century a number of fundamental discoveries were reported which could not be explained within the framework of the above classical theories of physics. The inadequacy of classical theories was noticed first when they were applied to explain the black body radiation emitted by a body hotter than its surroundings. To explain the blackbody radiation, Max Planck put forward a revolutionary hypothesis that the molecules in a source emit energy not continuously but in small discrete packets called *quanta*. This was a radical departure from the classical theory and contrary to day-to-day experience. Other experimental results also showed that the classical concepts were entirely inadequate to explain the behaviour of atoms and subatomic particles. A new body of ideas based on Planck's work was developed, and the new theory came to be known as quantum theory or **quantum physics**. *Quantum physics explains the behaviour of matter and radiation at the microscopic (atomic) level.*

19.2 WAVE-PICTURE OF RADIATION—ENERGY FLOW IS CONTINUOUS

Radio waves, microwaves, heat waves, light waves, UV-rays, x-rays and γ -rays belong to the family of electromagnetic waves. All of them are known as **radiation**. Electromagnetic waves consist of varying electric and magnetic fields traveling at the velocity of ' c '. The propagation of electromagnetic waves and their interaction with matter can be explained with the help of Maxwell's electromagnetic theory.

Maxwell's theory treated the emission of radiation by a source as a *continuous* process. A heated body may be assumed to be capable of giving out energy that travels in the form of waves of *all possible wavelengths*. In the same way, the radiation incident on a body was thought to be absorbed at all possible wavelengths. The intensity of radiation is given by

$$I = |E|^2 \quad (19.1)$$

where E is the amplitude of the electromagnetic wave.

The phenomena of interference, diffraction and polarization of electromagnetic radiation proved the wave nature of radiation. Therefore, it is expected that it would explain the experimental observations made on *thermal (heat) radiation* emitted by a blackbody.

19.3 BLACKBODY RADIATION

It is well known that when a body is heated, it emits electromagnetic radiation. The radiation emitted by the source by virtue of its temperature is called **thermal radiation**. Thermal radiation is electromagnetic in nature and its energy is smoothly distributed over all wavelengths. Therefore, a thermal source produces **continuous spectrum**. The intensity and the predominant wavelength of radiation vary with the temperature of the body. At low temperatures the radiation mainly lies in the infra red region. As the temperature of the body is increased, the component of maximum intensity shifts to a higher and higher frequency. For example, the filament of an incandescent bulb appears dark at room temperature and as current is passed, it gets heated up. As the current increases through the filament, it appears initially red, orange gradually and yellow and finally it emits white light. At temperatures above 1000°C, a heated body is capable of giving out energy in the form of waves of all possible wavelengths. Normally the amount of radiation emitted by a hot body depends on factors such as the properties of its surface. Apart from emitting electromagnetic radiation, a body also absorbs electromagnetic radiation incident on it. A body that absorbs the entire radiation incident on it is called a **perfect blackbody**. When a perfect blackbody is heated, it emits radiation at all frequencies and thus, it is a good radiator as well as a good absorber. The radiation emitted by a perfect blackbody is called **blackbody radiation**.

In practice, there are no perfect blackbodies but an ideal black body can be made by taking a hollow sphere (cavity) and drilling a small hole in it (see Fig. 19.1). Its inner surface is coated with lampblack. Any radiation entering the cavity through O is incident on the inner surface of the cavity and is partly absorbed and partly reflected. The reflected component is incident at another point on the inner surface where again it is partly absorbed and partly reflected. Thus, light entering the cavity undergoes multiple reflections at the walls and gets trapped inside the cavity. The probability of the radiation of any wavelength escaping out of the hole is negligible. Hence, the hole acts a perfect absorber and appears perfectly dark. Conversely, when the cavity is heated, the radiation produced in the cavity comes out through the aperture and contains all the wavelengths. Therefore, the hole acts as a perfect emitter and has the characteristics of blackbody radiation. Its spectrum can be analysed by an infra red spectrometer using a bolometer as a detector. Thus, the emissive power of the blackbody at different wavelengths can be determined. If the distribution of radiant energy as a function of wavelength at different temperatures is plotted, we obtain a set of curves as shown in Fig.19.2.

The experimental results (Fig.19.2) show that at a given temperature the radiation energy density initially increases with increasing wavelength, then peaks at around a particular wavelength λ_m and after that decreases finally to zero at very high wavelengths. The spectral distribution of that radiation is a function of temperature alone and the material as such plays no role.

19.3.1 Laws of Blackbody Radiation

- (i) **Stefan-Boltzmann law:** The Stefan-Boltzmann law is an empirical relationship obtained by Stefan and later derived theoretically by Boltzmann. It connects the intensity of radiation to the temperature. It states that *the total radiation emitted from a blackbody at temperature T is proportional to the fourth power of the absolute temperature of the body T^4 .*

$$E = \sigma T^4 \quad (19.2)$$

where σ is called **Stefan's constant** having a numerical value of $5.67 \times 10^{-8} \text{ W/m}^2\text{-K}^4$.

- (ii) **Wien's law:** It is seen from Fig.19.2 that the value of λ_m depends only on the temperature T of the blackbody and decreases with increasing temperature. λ_m is independent of the

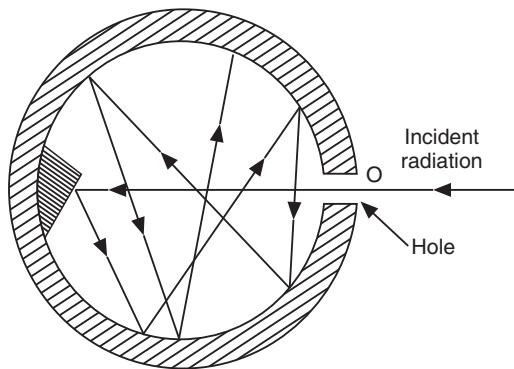


Fig. 19.1

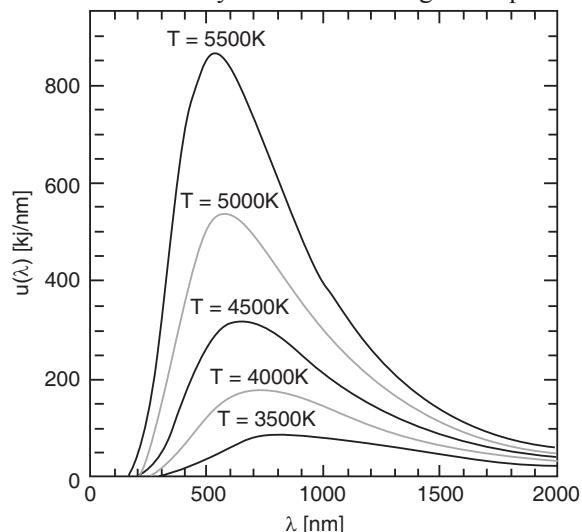


Fig. 19.2

nature of the emitting body. The shift in the peak of the intensity distribution curves obeys **Wien's displacement law**. The Wien's law states that *the peak wavelength, λ_{max} at which the maximum emission occurs for any given temperature is inversely proportional to the absolute temperature of the body*. Thus,

$$\lambda_{max} = \frac{2.8978 \times 10^{-3}}{T} \text{ mK} \quad (19.3)$$

- (iii) Rayleigh-Jean's Law and the ultraviolet catastrophe:** Various efforts were made to calculate theoretically the frequency distribution of thermal radiation. At that period of time electron was discovered but the inner structure of the atom was unknown. The generally accepted model was that a blackbody was made up of a huge number of atoms and the atoms are regarded as small harmonic oscillators. The random thermal motion of atoms within the walls generates electromagnetic waves, which is the thermal radiation emitted from the walls of the cavity. The radiation emitted by the atoms is reflected back and forth by the cavity walls to form a system of standing waves for each frequency present. There would be many modes of vibration present in the cavity space. Finally, when thermal equilibrium is attained, the average rate of emission of radiant energy by atomic oscillators in the walls equals the rate of absorption of radiant energy by the interior walls. According to Boltzmann's principle of equipartition of energy, each simple harmonic oscillator has an average thermal energy of ' kT ' at thermal equilibrium. Rayleigh assumed that each of the standing waves ought to have energy ' kT ' and derived an expression for the energy density of radiation distribution within the cavity.

$$E(v) = \frac{8\pi v^2 kT}{c^3} \quad (19.4)$$

This result is known as **Rayleigh-Jeans law**. The energy density calculated with the above formula agrees well with the experimental results at long wavelength end of the spectrum but goes toward infinity at the short wavelength end. It predicted that the intensity of thermal radiation should increase with square of frequency. It implied that the radiation emitted by a hot body should have a large portion of UV rays. This is contrary to our experience and violates the law of conservation of energy. This contradiction came to be known as **ultraviolet catastrophe**. The failure of Rayleigh-Jeans formula presented a crisis and gave the first indication of inadequacy of classical physics.

- (iv) The Planck's Radiation Law:** The failure of classical mechanics to account for black body radiation led Max Planck to approach the problem from another direction. That is to find a mathematical expression that could account for the black body radiation spectrum, and then seek to derive this spectrum from a new equation, hitherto unknown. This led to the postulation of Planck's Radiation Law for the spectral energy density of black body radiation in 1900.

After long years of struggle, Planck succeeded in 1900 to obtain the correct mathematical law for distribution of energy in the blackbody radiation. He recognized that Rayleigh's assumption about the equipartition of energy among all available modes in the cavity was leading to the absurd result that the total energy in the cavity would be infinite. Therefore, Planck put forward a revolutionary postulate that an oscillating atom can absorb or reemit energy only in quantities that are integer multiples of ' hv '. All other values of energy are

forbidden. The indivisible discrete unit of energy $h\nu$ is called an **energy quantum**. The assumption which allowed Planck's Radiation Law to be derived, was that the energies of the molecular oscillators must be quantized, not continuous, as it had been considered to be the case until then.

In general, the possible values of the energy of an oscillator of frequency ν are

$$E = n h\nu \quad (n = 1, 2, 3, \dots) \quad (19.5)$$

where n is called the **quantum number** of the oscillator, h is a constant now known as **Planck's constant**.

Planck further assumed that the change in energy of the oscillator due to emission or absorption of radiation could take place by a discrete amount $h\nu$. Since radiation is emitted from the oscillators, the energy carried by the emitted radiation will be $h\nu$ which is equal to the loss of energy of the oscillator. Obviously, this is the energy gain of the oscillator when it absorbs the radiation. Thus, the atomic oscillators can exist in a set of discrete energy states, $0, h\nu, 2h\nu, 3h\nu$, and so on. The number of oscillators in an energy state $E_n = n h\nu$ is determined by the Maxwell-Boltzmann distribution function.

$$N_n = N_0 \exp(-E_n/kT) = N_0 \exp(n h\nu/kT) \quad (19.6)$$

where N_0 is the number of oscillators in the ground state ($E_n = 0$).

Let there be N oscillators in the system in equilibrium at absolute temperature T . According to Maxwell-Boltzmann distribution law, the number of oscillators N_n with energy E_n is given by

$$N_n = Ne^{-E_n/kT}$$

Since the energies of the oscillators assume discrete values, we have to sum over all possible oscillator states to determine $\langle E \rangle$. The average energy $\langle E \rangle$ of an oscillator is then given by

$$\begin{aligned} \langle E \rangle &= \frac{\sum_{n=0}^{\infty} E_n N_n}{\sum_{n=0}^{\infty} N_n} = \frac{\sum_{n=0}^{\infty} E_n e^{-E_n/kT}}{\sum_{n=0}^{\infty} e^{-E_n/kT}} \\ \langle E \rangle &= \frac{\sum_{n=0}^{\infty} n h\nu e^{-n h\nu/kT}}{\sum_{n=0}^{\infty} e^{-n h\nu/kT}} \\ &= \frac{h\nu e^{-h\nu/kT} + 2h\nu e^{-2h\nu/kT} + \dots}{1 + e^{-h\nu/kT} + e^{-2h\nu/kT} + \dots} \end{aligned} \quad (19.7)$$

Substituting $e^{-h\nu/kT} = x$, we get

$$\begin{aligned} \langle E \rangle &= \frac{h\nu x(1+2x+3x^2+\dots)}{1+x+x^2+\dots} \\ &= \frac{h\nu x(1-x)^{-2}}{(1-x)^{-1}} = \frac{h\nu x}{(1-x)} \\ \therefore \langle E \rangle &= \frac{h\nu e^{-h\nu/kT}}{1-e^{-h\nu/kT}} = \frac{h\nu}{e^{h\nu/kT}-1} \end{aligned} \quad (19.8)$$

According to classical physics, the average energy of an oscillator is kT . The concept of quantization of energy shows that the theorem of equipartition of energy is not valid in the microscopic world. The number of oscillators in the frequency range v and $v + dv$ are estimated to be

$$\frac{8\pi v^2 dv}{c^3}$$

Multiplying this number by the average energy of an oscillator, Planck obtained the energy density in blackbody radiation as

$$E(v) = \frac{8\pi h v^3}{c^3} \left[\frac{1}{e^{hv/kT} - 1} \right] \quad (19.9)$$

or $E(\lambda) = \frac{8\pi h c}{\lambda^5} \left[\frac{1}{e^{hc/\lambda kT} - 1} \right]$ (19.10)

This is **Planck's radiation law**. It is seen from the formula that the exponential term in the denominator prevents $E(v)$ from rising to infinity at small wavelength values. Planck's formula gives complete fit with the experimental observations at all wavelengths and for all temperatures.

All the laws of blackbody radiation can be derived from Planck's law as follows:

(i) Wien's law

When v is large, $(hv/kT) \gg 1$, and $(e^{hv/kT} - 1) \approx e^{hv/kT}$. Hence equ.(19.9) reduces to

$$E(v) = \frac{8\pi h v^3}{c^3} \cdot \frac{1}{e^{hv/kT}} = \frac{8\pi h v^3}{c^3} e^{-hv/kT}$$

This represents Wien's law.

(ii) Rayleigh-Jean's law

When v is small, $(hv/kT) \ll 1$, and $(e^{hv/kT} - 1) \approx \frac{hv}{kT}$. Hence equ.(19.9) reduces to

$$E(v) = \frac{8\pi h v^3}{c^3} \cdot \frac{kT}{hv} = \frac{8\pi v^2 kT}{c^3}$$

This represents the Rayleigh-Jean's law.

(iii) Wien's displacement law

Planck's formula (19.10) shows a maximum at $\lambda = \lambda_m$ when the denominator becomes a minimum. The denominator can be written as

$$z = \lambda^5 (e^{hc/\lambda kT} - 1)$$

Then $\frac{dz}{d\lambda} = 5\lambda^4 (e^{hc/\lambda kT} - 1) - \lambda^5 \frac{hc}{\lambda^2 kT} e^{hc/\lambda kT} = 0$ for $\lambda = \lambda_m$.

$$1 - e^{(-hc/\lambda_m kT)} = \frac{hc}{5\lambda_m kT} \quad (19.11)$$

or $1 - e^{-x} = \frac{x}{5}$

where

$$x = hc/\lambda_m kT. \quad (19.12)$$

The above equation is a transcendental equation which cannot be solved analytically. It has to be solved graphically. If we put $y = 1 - e^{-x}$ and $y = x/5$, then the point of

intersection of the two graphs given by these two relations gives the solution. It comes out to be $x = 4.9651$. Using this value into equ.(19.12), we get

$$\lambda_m T = \frac{hc}{kx} = \frac{hc}{4.9651k} = 0.0029 \text{ m} \cdot K = \text{constant.}$$

This is Wien's displacement law.

19.4 PLANCK'S QUANTUM HYPOTHESIS – Energy is quantized

Max Planck empirical formula explained the experimental observations. In the process of formulation of the formula, he assumed that the atoms of the walls of the blackbody behave like small harmonic oscillators, each having a characteristic frequency of vibration. He further made two radical assumptions about the atomic oscillators.

- (i) *An oscillating atom can absorb or reemit energy in discrete units.* The indivisible discrete unit of energy $h\nu$ is the smallest amount of energy which can be absorbed or emitted by the atom and is called an **energy quantum**. A quantum of energy has the magnitude given by

$$E = h\nu \quad (19.13)$$

where ν is the frequency of radiation and ' h ' is a constant now known as the **Planck's constant**.

- (ii) *The energy of the oscillator is quantized. It can have only certain discrete amounts of energy E_n .*

$$E_n = n h\nu; \quad n = 1, 2, 3 \dots \quad (19.14)$$

The hypothesis that *radiant energy is emitted or absorbed basically in a discontinuous manner and in the form of quanta* is known as the **Planck's quantum hypothesis**. Planck's hypothesis states that radiant **energy is quantized** and implies that an atom exists in certain discrete energy states. Such states are called **quantum states** and n is called the **quantum number**. The atom emits or absorbs energy by jumping from one quantum state to another quantum state.

The assumption of discrete energy states for an atomic oscillator (Fig.19.3) was a departure from the classical physics and our everyday experience. If we take a mass–spring harmonic oscillator, it can receive any amount of energy from zero to some maximum value (Fig.19.4). Thus, in the realm of classical physics energy always appears to occur with continuous values and energy exchange between bodies involves any arbitrary amounts of energy.

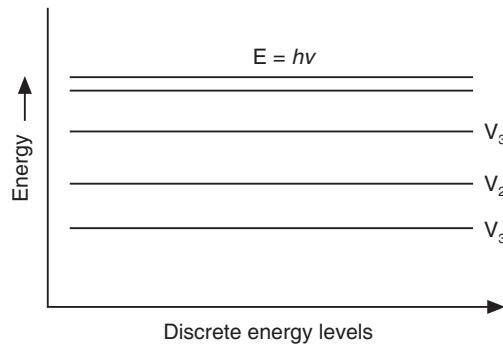


Fig. 19.3

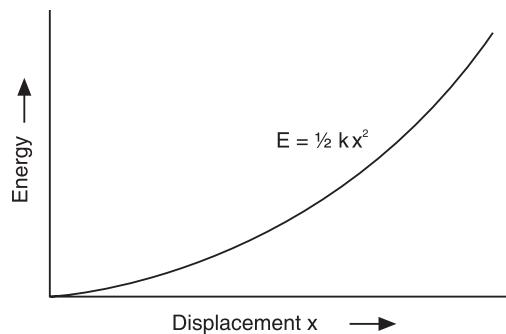


Fig. 19.4

The quantum concept can be understood with the help of the following example:

Consider a person walking up a hill and another person climbing a stair case, as shown in Fig. 19.5. Both of them gain in potential energy. The potential energy of the person going up the hill increases *continuously* through *arbitrary* amounts. In contrast, the potential energy of the person climbing the stair case increases through fixed doses. If all the steps are of the same height ' h ', the person acquires a quantum of energy ' mgh ' each time he climbs one step. Consequently, his potential energy increases by $1 mgh, 2 mgh, 3 mgh, \dots N mgh$. It is not possible to acquire energy which is a fraction of mgh or of any intermediate value.

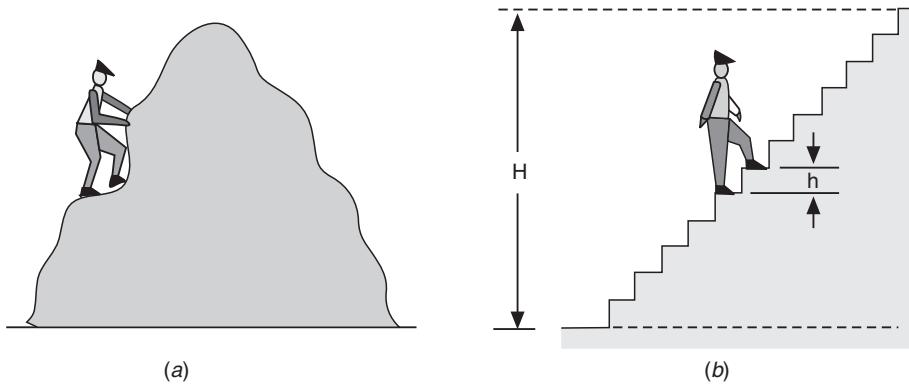


Fig. 19.5

19.5 PARTICLE PICTURE OF RADIATION – Radiation is a stream of photons

Max Planck introduced the concept of discontinuous emission and absorption of radiation by bodies but he treated the propagation through space as occurring in the form of continuous waves as demanded by electromagnetic theory. Einstein refined the Planck's hypothesis and invested the quantum with a clear and distinct identity. He successfully explained the experimental results of the photoelectric effect in 1905 and the temperature dependence of specific heats of solids in 1907 basing on Planck's hypothesis. The photoelectric effect conclusively established that *light behaves as a stream of particles*. Einstein extended Planck's hypothesis as follows.

1. Einstein considered that the quantization of energy, which is evident in the emission and absorption processes, is retained as the energy propagates through space. He assumed that the light energy is not distributed evenly over the whole expanding wave front but rather remains concentrated in discrete quanta. He named the energy quanta as *photons*. Accordingly, a light beam is regarded as a stream of photons travelling with a velocity ' c '.
2. An electromagnetic wave having a frequency ν contains identical photons, each having an energy $h\nu$. The higher the frequency of the electromagnetic wave, the higher is the energy content of each photon. Thus, x-ray and γ -ray photons are far more energetic compared to optical photons while the photons of r.f. frequencies are the feeblest.
3. An electromagnetic wave would have energy $h\nu$ if it contains only one photon, $2h\nu$ if it contains 2 photons and so on. Therefore, the intensity of a monochromatic light beam I , is related to the concentration of photons, N , present in the beam. Thus,

$$I = N h\nu \quad (19.15)$$

Note that according to electromagnetic theory, the intensity of a light beam is given by

$$I = |E|^2$$

4. When photons encounter matter, they impart all their energy to the particles of matter and vanish. That is why absorption of radiation is discontinuous. The number of photons emitted by even a weak light source is enormously large and the human eye cannot register the photons separately and therefore light appears as a continuous stream. Thus, the discreteness of light is not readily apparent.

19.5.1 The Photon

As the radiant energy is viewed as made up of spatially localized photons, we may attribute particle properties to photons.

1. **Energy:** The energy of a photon is determined by its frequency ν and is given by $E = h\nu$. Using the relation $\omega = 2\pi\nu$ and writing $h/2\pi = \hbar$, we may express

$$E = \hbar\omega \quad (19.16)$$

2. **Velocity:** Photons always travel with the velocity of light ' c '.

3. **Rest Mass:** The rest mass of photon is zero since a photon can never be at rest. Thus,

$$m_0 = 0$$

4. **Relativistic mass:** As photon travels with the velocity of light, it has relativistic mass, given by $m = \frac{E}{c^2} = \frac{h\nu}{c^2}$.

$$(19.17)$$

5. **Linear Momentum:** The linear momentum associated with a photon may be expressed as $p = \frac{E}{c} = \frac{h\nu}{c} = \frac{h}{\lambda}$

$$\text{As the wave vector } k = \frac{2\pi}{\lambda}, \quad p = \frac{h}{2\pi}k = \hbar k. \quad (19.18)$$

6. **Angular Momentum:** Angular momentum is also known as spin which is the intrinsic property of all microparticles. Photon has a spin of one unit. Thus, $s = 1\hbar$.

7. **Electrical Charge:** Photons are electrically neutral and cannot be influenced by electric or magnetic fields. They cannot ionize matter.

Example 19.1. The wavelength of yellow light is 5890 \AA . What is the energy of the photons in the beam? Express in electron volts.

Solution:

$$\begin{aligned} E &= h\nu = \frac{hc}{\lambda} = \frac{(6.63 \times 10^{-34} \text{ J.s})(3 \times 10^8 \text{ m/s})}{5.89 \times 10^{-7} \text{ m}} \\ &= \frac{19.89}{5.89} \times 10^{-19} \text{ J} = (3.38 \times 10^{-19})(6.24 \times 10^{18} \text{ eV}) \\ &= \mathbf{2.11 \text{ eV}} \end{aligned}$$

Example 19.2. The light sensitive compound on most photographic films is silver bromide, AgBr . A film is exposed when the light energy absorbed dissociates this molecule into its atoms. The energy of dissociation of AgBr is 23.9 k.cal/mol . Find the energy in electron volts, the wavelength and the frequency of the photon that is just able to dissociate a molecule of silver bromide.

Solution: The number of molecules contained in one mole of any substance is equal to the Avogadro's number, N_A . The energy of dissociation of one Ag Br molecule is therefore given by $E = \frac{E_D}{N_A}$, This is equal to the energy of the photon.

$$(i) E = \frac{E_D}{N_A} = \frac{(23.9 \text{ k.cal/mol})(4185 \text{ J/kcal})}{6.023 \times 10^{23} / \text{mol}} \\ = \frac{1.0}{6.023} \times 10^{-18} \text{ J} = (0.166 \times 10^{-18})(6.24 \times 10^{18} \text{ eV}) = 1.04 \text{ eV}$$

$$(ii) v = \frac{E}{h} = \frac{0.166 \times 10^{-18} \text{ J}}{6.63 \times 10^{-34} \text{ J.s}} = 0.025 \times 10^{16} \frac{1}{\text{s}} = 2.5 \times 10^{14} \text{ Hz}$$

$$(iii) \lambda = \frac{hc}{E} = \frac{(6.63 \times 10^{-34} \text{ J.s})(3 \times 10^8 \text{ m/s})}{0.166 \times 10^{-18} \text{ J}} = \frac{19.89}{0.166} \times 10^{-8} \text{ m} = (119.82 \times 10^{-8})(10^{10} \text{ \AA}) \\ = 11982 \text{ \AA}$$

Example 19.3. What is the photon energy in joules corresponding to a 50 Hz wave emitted by the power line? How does it compare with the energy of yellow light?

Solution. Photon energy, $E_{PL} = hv_{PL} = (6.63 \times 10^{-34} \text{ J.s})(50 \text{ Hz}) = 3.32 \times 10^{-32} \text{ J} \\ = (3.32 \times 10^{-32})(6.24 \times 10^{18} \text{ eV}) = 2.07 \times 10^{-13} \text{ eV}$

Yellow light photon energy, $E_y = \frac{hc}{\lambda} = \frac{12400}{5890} \text{ eV} = 2.11 \text{ eV}$

$$\therefore E_y : E_{PL} = \frac{2.11 \text{ eV}}{2.07 \times 10^{-13} \text{ eV}} = 1.02 \times 10^{13}$$

Thus the energy of light photon is about 10^{13} times greater than the energy of the photon associated with a 50 Hz wave.

Example 19.4. A blue lamp emits light of mean wavelength of 4500 \AA . The lamp is rated 150 W and 8% of the energy appears as light. How many photons are emitted per second by the lamp?

Solution. Total energy supplied by the lamp

$$E_T = P \times 1 \text{ sec} = P(J) = 6.24 \times 10^{18} \text{ PeV}$$

Energy converted into light

$$E_L = \eta E_T = 6.24 \times 10^{18} \text{ PeV}$$

$$\text{Number of photons emitter per second } N = \frac{E_L}{E}$$

$$\text{Where } E \text{ is the energy of a photon, } E = \frac{12400}{\lambda(\text{\AA})}$$

$$\therefore N = \frac{(6.24 \times 10^{18})(P\eta\lambda(\text{\AA}))}{12,400} = \frac{(6.24 \times 10^{18} \text{ eV})(150)(0.08)(4500)}{12,400 \text{ eV}} = 27.17 \times 10^{18}$$

$$\therefore N = 2.7 \times 10^{19} \text{ photon/sec.}$$

19.6 PHOTOELECTRIC EFFECT

In 1887 Hertz discovered that electrons are ejected when ultraviolet radiation strikes a metal surface. Photoelectric effect is the phenomenon in which electrons are knocked off a metal surface when light is incident on it (see Fig.19.6). The metal is said to be a *photosensitive material* and the liberated electrons are called *photoelectrons*.

Electronic emission increases with the intensity of the radiation falling on the metal surface, since more energy is available to release electrons. But characteristic frequency dependence is also observed. For each substance there is a minimum frequency v_0 of light and for light of frequency less than v_0 , photoelectrons are not produced, however intense the incident light may be. The minimum frequency v_0 is called **threshold frequency**. The photoelectrons that are emitted in the process are found to have a range of energies. The maximum kinetic energy of the photoelectrons varies with the frequency of the incident light (see Fig. 19.7) and is independent of the intensity of the light.

Attempts to explain the photoelectric effect on the basis of wave theory of light were unsuccessful. In 1905 Einstein successfully explained the photoelectric effect by treating the incident light as a stream of photons. In a beam of monochromatic light of frequency, v , all photons have the same energy hv . When the photons encounter electrons in a metal, they give up their energy to the electrons. However, the electrons are held within the metal by a **potential barrier** at the surface. In order to escape from the metal, an electron should have enough energy to overcome the potential barrier at the surface. It requires an input energy that is equal to the *work function*, W_0 . This work function is specific for the metal. When an electron absorbs a photon, the energy gained by it is equal to hv . If $hv \geq W_0$, the electron will escape from the metal. Thus, out of the total energy hv , a small portion, W_0 , is spent on surmounting the potential barrier and the balance energy ($hv - W_0$) is given to the electron as kinetic energy. Applying the law of conservation of energy to this phenomenon, we find that

$$hv = W_0 + \frac{1}{2}mv_{\max}^2 \quad (19.20)$$

The above equation is known as **Einstein's photoelectric equation**. It explains all the features of the photoelectric effect. From this equation it is seen that a photoelectron is emitted from the metal only if the energy of the incident photon $hv \geq W_0$. If $hv < W_0$, electrons cannot surmount the barrier at the metal surface and emerge from the metal, however intense the incident light may be. The minimum energy that a photon should have to dislodge an electron from the metal, without imparting it any kinetic energy ($K.E. = 0$), is given by $hv_0 = W_0$. Hence, photoelectric effect can occur only when the frequency of the incident light is equal to or greater than the minimum value v_0 , which is given by

$$v_0 = \frac{W_0}{h} \quad (19.21)$$

Thus, the existence of the threshold frequency is explained.

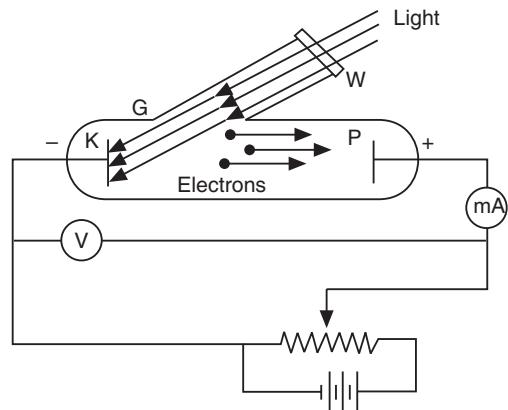


Fig. 19.6

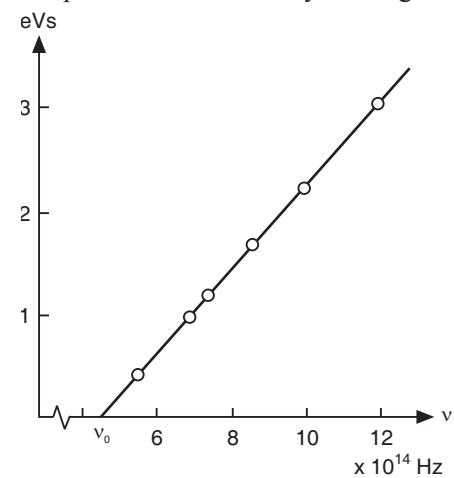


Fig. 19.7

The kinetic energy of the photoelectrons can be measured by subjecting them to a repelling electrostatic force, by giving negative potential to the anode. The electrons ejected from the cathode are decelerated as they travel toward the anode. When the repelling force is made large enough, even the most energetic electrons are stopped from reaching the anode and hence the current will drop to zero. The potential difference that stops the electrons reaching the anode and reduces the current in the circuit to zero, is called **stopping potential**, V_s . The kinetic energy of the most energetic electrons is therefore given by

$$K.E_{\max} = eV_s \quad (19.22)$$

Using the relations (19.20) and (19.21) into (19.22), we get

$$V_s = \frac{h}{e}[v - v_0] \quad (19.23)$$

The above relation shows that the stopping potential and hence the kinetic energy of photoelectron varies linearly with the frequency of the incident light.

Finally, as soon as a photon of energy $hv > W_0$ is incident on the metal, one of the electrons in it absorbs the photon and is ejected instantaneously. Thus, there is no delay between the incidence of light and emission of electrons.

In 1916 Millikan verified Einstein's theory. If a plot is drawn between the potential V_s and the frequency of the incident light, the resulting graph will be a straight line (see Fig. 19.8). The slope of this graph is equal to (h/e) and is the same for all materials. Millikan showed that the value of h determined from experiments on the photoelectric effect agrees well with the value obtained from other experiments.

It means that the photoelectric effect confirmed the existence of energy in the form of discrete quanta.

Example 19.5. Light of a wavelength 2000 Å falls on an aluminium surface with work function 4.2 eV. Calculate (i) threshold wavelength and (ii) stopping potential. (N.U., W-94)

Solution:

$$(i) \text{ Threshold wavelength, } \lambda_0 = \frac{hc}{W_0} = \frac{12400}{W_0(eV)} \text{ Å} = \frac{12400}{4.2} \text{ Å} = 2952 \text{ Å}$$

(ii) If V_s is the stopping potential, then

$$eV_s = \frac{hc}{\lambda} - W_0 = \left[\frac{12400}{\lambda(\text{Å})} - W_0 \right] eV = \left[\frac{12400}{2000} - 4.2 \right] eV = 2 \text{ eV}$$

$$\therefore V_s = \frac{2 \text{ eV}}{e} = 2 \text{ volts}$$

Example 19.6. Work function of sodium is 2.3 eV. Obtain the maximum wavelength which will cause emission of photoelectrons from the material. What will be the maximum kinetic energy of the photoelectrons emitted by the surface exposed to the radiation of 2000 Å?

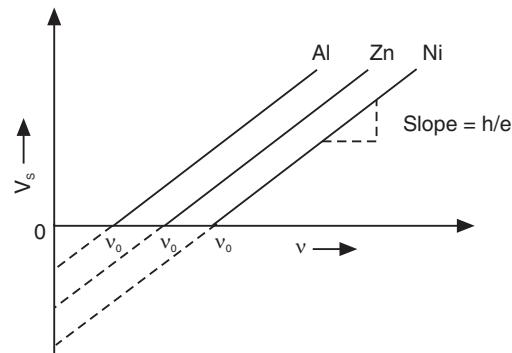


Fig. 19.8

Solution: Maximum wavelength, $\lambda = \frac{hc}{W_0} = \frac{12400}{(W_0)eV} \text{ Å} = \frac{12400}{2.3eV} \text{ Å} = 5391 \text{ Å}$

$$\begin{aligned}\text{Max. kinetic energy, } K.E_{\max} &= \frac{hc}{\lambda} - W_0 = \left[\frac{12400}{\lambda} - (W_0) \right] eV \\ &= \left[\frac{12400}{2000} - (2.3) \right] eV = 3.9 \text{ eV}\end{aligned}$$

Example 19.7. The work function for cadmium is 4.08 eV. What must be the wavelength of radiation incident on cadmium so that the maximum velocity of photoelectrons will be $7.2 \times 10^5 \text{ m/s}$.

Solution.

$$\begin{aligned}E &= W_0 + K.E_{\max} = W_0 + \frac{1}{2}mv_{\max}^2 = 4.08eV + \frac{1}{2}(9.11 \times 10^{-31} \text{ kg})(7.2 \times 10^5 \text{ m/s})^2 \\ &= 4.08eV + 2.36 \times 10^{-19} \text{ kg.m}^2/\text{s}^2 \\ &= 4.08eV + (2.36 \times 10^{-19})(6.24 \times 10^{18} \text{ eV}) \\ &= 4.08eV + 1.47eV = 5.55eV \\ E &= \frac{hc}{\lambda} = \frac{12400}{\lambda(\text{\AA})} = W_0 + \frac{1}{2}mv_{\max}^2 \\ \therefore \lambda &= 12400 \left[\frac{1}{W_0 + \frac{1}{2}mv_{\max}^2} \right] \text{\AA} \\ &= \left[\frac{12400}{W_0 + \frac{1}{2}mv_{\max}^2} \right] \text{\AA} = \frac{12400}{5.55} \text{\AA} = 2234 \text{\AA}\end{aligned}$$

19.7 X-RAYS

Wilhelm Rontgen discovered X-rays in 1895 during the course of some experiments with a discharge tube. He noticed that a screen coated with barium platinocyanide present at a distance from the discharge tube fluoresced brilliantly. Rontgen called these invisible radiations X-rays. He concluded that X-rays are produced due to the bombardment of cathode rays on the walls of the discharge tube. Now it is well known that X-rays are produced when an obstacle stops fast moving electrons, and that metals of high atomic weight are most effective for this purpose. X-rays are highly penetrating and can pass through many solids. It is subsequently showed that X-rays are electromagnetic waves with very short wavelengths. They occur beyond the UV region in the electromagnetic spectrum. Their wavelengths range from about 0.01 Å to about 10 Å.

19.8 GENERATION OF X-RAYS

X-rays are produced by an X-ray tube. The schematic of the modern type of X-ray tube designed by Coolidge is shown in Fig.19.9. It is an evacuated glass bulb enclosing two

electrodes, a cathode and an anode. The cathode consists of a tungsten filament which when heated emits electrons. The electrons are focused into a narrow beam with the help of a metal cup C. The anode consists of a target material, made of tungsten or molybdenum, which is embedded in a copper bar. Water, circulating through a jacket surrounding the anode, cools the anode. Further, large cooling fins conduct the heat away to the atmosphere. The face

of the target is kept at an angle relative to the oncoming electron beam. A very high potential difference of the order of 50 kV is applied across the electrodes. The electrons emitted by the cathode are accelerated by the anode and acquire high energies of the order of 10^5 eV. When the target suddenly stops these electrons, X-rays are emitted. The magnetic field associated with the electron beam undergoes a change when the electrons are stopped and electromagnetic waves in the form of X-rays are generated. The greater the speed of the electron beam, the shorter will be the wavelength of the radiated X-rays. Only about 0.2% of the electron beam energy is converted into X-rays and the rest of the energy transforms into heat. It is for this reason that the anode is intensively cooled during the operation of the X-ray tube. The intensity of the electron beam depends on the number of electrons leaving the cathode. This can be adjusted by passing a suitable current through the filament. The hardness of the X-rays emitted depends on the energy of the electron beam striking the target. It can be adjusted by varying the potential difference applied between the cathode and the anode. Therefore, the larger the potential difference, the more penetrating or **harder** are X-rays.

19.9 X-RAY SPECTRUM

When the X-rays produced by a metallic target are examined with the help of a spectrometer and their intensity is plotted against wavelength, a curve such as the one shown in Fig. 19.10 is obtained. This curve is called an *X-ray spectrum*. An analysis of the spectrum of radiation emitted by an X-ray tube shows the presence of two distinct contributions. One portion of the spectrum consists of X-rays of all possible wavelengths with a lower limit. This is known as the **continuous spectrum**. The continuous distribution of wavelengths emitted in the process is analogous to the continuous distribution of wavelengths in white light. It is sometimes called *white radiation*. The continuous spectrum is also called **bremsstrahlung** meaning breaking radiation. It is found that there is a minimum wavelength below which X-rays are not emitted. This minimum wavelength is called the *cut-off wavelength*. The other portion of the spectrum consists of several sharp lines of discrete wavelengths superim-

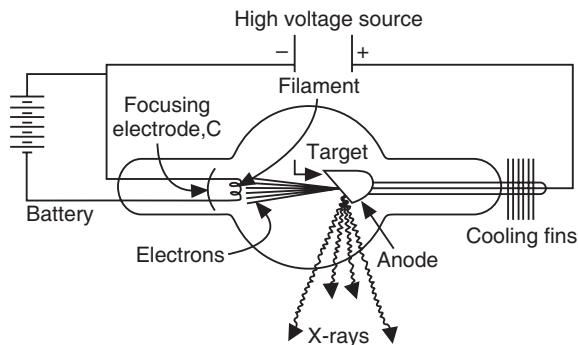


Fig. 19.9

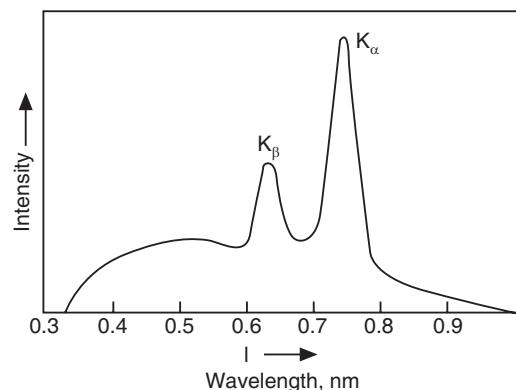


Fig. 19.10: The characteristic K_α and K_β lines emitted by a molybdenum target superposed on the continuous spectrum.

Hence, the continuous X-ray radiation is sometimes called *white radiation*. The continuous spectrum is also called **bremsstrahlung** meaning breaking radiation. It is found that there is a minimum wavelength below which X-rays are not emitted. This minimum wavelength is called the *cut-off wavelength*. The other portion of the spectrum consists of several sharp lines of discrete wavelengths superim-

posed over the continuous spectrum. This is known as the **characteristic spectrum**. The wavelengths of the line spectrum are characteristic of the target element.

19.10 ORIGIN OF CONTINUOUS X-RAY SPECTRUM

Continuous X-rays are produced by the phenomenon of **bremsstrahlung**, which is a German word meaning ‘breaking’ or ‘slowing down’ radiation. Electrons emitted from the cathode in the X-ray tube are accelerated toward the target. They strike the target and penetrate deep into the interior of atoms and are attracted by the nuclei due to strong Coulomb attraction. The electrons deviate from their original path (see Fig.19.11). The deviation of an electron from its original path is equivalent to its collision with the nucleus. The electron loses some of its kinetic energy due to such collision. The energy lost appears as an X-ray photon. If E_i is the initial kinetic energy and E_f the final kinetic energy of the electron, the energy of X-ray photon will be equal to $(E_i - E_f)$. If v is the frequency of the X-ray photon, then the energy of the photon is

$$hv = E_i - E_f$$

The electron may make several collisions before coming to rest. Therefore, the frequency of the emitted photon is not unique. During each collision of the electron, a photon may be given out. The energies of these photons will range from very small value up to the maximum energy E_f . The maximum energy corresponds to an electron that loses all its energy in a single encounter.

The spectral distribution of continuous X-rays and its dependence on the applied potential difference is shown in Fig.19.12. The important features of the curves are as follows:

1. The wavelengths at which the maximum intensity occurs shifts to lower values for higher accelerating potentials.
2. With increase in accelerating potential, the total intensity of X-rays increases and X-rays of shorter wavelengths are emitted.
3. The radiation emitted has a continuous spectral distribution.
4. For any given accelerating potential V , there is a sharply defined minimum wavelength λ_{\min} called **short-wavelength limit** or **cut-off wavelength**.
5. The short wavelength limit shifts to lower wavelength values with increasing accelerating potential.

The production of X-rays can be considered to be an *inverse photoelectric emission*. In the case of photoelectric effect, photons incident on a surface liberate electrons. In the present case, electrons cause emission of X-ray photons. Hence X-ray phenomenon is considered as **inverse photoelectric effect**.

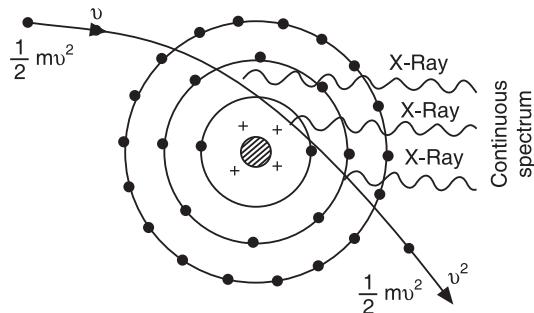


Fig. 19.11

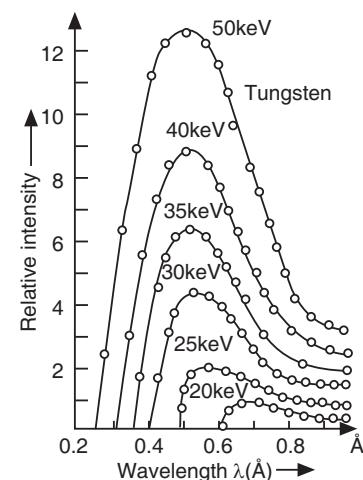


Fig.19.12

19.11 DUANE-HUNT FORMULA

It was found experimentally by Duane and Hunt that the short wavelength limit λ_{\min} is inversely proportional to the accelerating potential V and is given by

$$\lambda_{\min} = \frac{12400}{V} \text{ Å} \quad (19.24)$$

This equation is known as **Duane-Hunt formula**.

The short-wavelength limit and its dependence on applied voltage cannot be explained on the basis of classical theory. According to classical theory, the continuous spectrum should consist of all frequencies from zero to a certain maximum. The short wavelength limit can be explained only on the basis of quantum theory. In an X-ray tube, the electrons emitted at the cathode are all accelerated through the same potential difference V and they reach the target with nearly the same energy eV. When the electrons are stopped, they lose the energy eV suddenly. This energy eV goes into X-rays. According to Planck's hypothesis, the emission of X-rays occurs in the form of photons, each photon carrying energy $h\nu$. Therefore, the maximum energy that an X-ray photon can have is

$$\begin{aligned} h\nu_{\max} &= eV \\ \therefore \nu_{\max} &= \frac{eV}{h} \\ \text{But } \nu_{\max} &= \frac{c}{\lambda_{\min}} \\ \therefore \lambda_{\min} &= \frac{hc}{eV} \end{aligned} \quad (19.25)$$

Using the values of h , c and e into the above equation, we get

$$\lambda_{\min} = \frac{(6.626 \times 10^{-34} \text{ J.s})(3 \times 10^8 \text{ m/s})}{(1.602 \times 10^{-19} \text{ C})V} = \frac{1.24}{V} \times 10^{-6} \text{ m}$$

or $\lambda_{\min} = \frac{12,400}{V} \text{ Å.}$

This is the *Duane-Hunt formula*. Thus, the quantum theory explains the short-wavelength limit λ_{\min} and its dependence on the accelerating voltage.

Example 19.8:

- (i) At what potential difference must an X-ray tube operate to produce X-rays with minimum wavelength of 0.01 Å ?
(ii) What is the maximum frequency of the X-rays produced in a tube operating at 50 kV ?

(Bombay Univ., 94)

Solution:

$$(i) \lambda_{\min} = \frac{hc}{eV} = \frac{12400}{V} \text{ Å} \quad \therefore V = \frac{12400}{\lambda_{\min} (\text{Å})} = \frac{12400}{0.01} V = 1.24 \text{ MV}$$

$$(ii) h\nu_{\max} = eV \quad \therefore \nu_{\max} = \frac{eV}{h} = \frac{1.602 \times 10^{-19} \text{ C} \times 50 \times 10^3 \text{ V}}{6.63 \times 10^{-34} \text{ J.s}} = 1.2 \times 10^{19} \frac{CV}{J.s} = 1.2 \times 10^{19} \text{ Hz.}$$

Example 19.9: An X-ray tube operated at 40 kV emits a continuous X-ray spectrum with a short wavelength limit $\lambda_{\min} = 0.31 \text{ \AA}$. Calculate Planck's Constant.

Solution:

$$\lambda_{\min} = \frac{hc}{eV} = \frac{12400}{V} \text{ \AA}$$

$$\therefore h = \frac{\lambda_{\min} eV}{c} = \frac{(0.31 \times 10^{-10} \text{ m})(1.602 \times 10^{-19} \text{ C})(40 \times 10^3 \text{ V})}{(3 \times 10^8 \text{ m/s})} = 6.61 \times 10^{-34} \text{ J.s}$$

Units: $1 \frac{\text{CV m}}{\text{m/s}} = 1 \text{ CVs} = 1 \text{ J.s}$

19.12 COMPTON SCATTERING

Photoelectric effect provided the evidence for the existence of photons and their energy. According to the theory of relativity mass and energy are identical. If photons have energy, then they must possess mass and exhibit the property of momentum. The fact that photons have momentum was established by Compton scattering studies.

Arthur H. Compton carried out investigations on the scattering of x-rays by materials of low atomic number. Compton had observed that when a beam of x-rays was scattered by a material, the scattered radiation consisted of two components of x-rays. One of the components had the same wavelength as that of the incident beam and the other having a longer wavelength than that of the incident x-rays. **Compton effect** is the name given to this phenomenon.

Scattering of X-rays by electrons of atoms in a material without any change in their wavelength is known as **classical** or **coherent scattering**. Scattering of X-rays by electrons of atoms in a material with a change in wavelength is called **incoherent scattering**. It is now known as **Compton scattering**. In a Compton scattering event, an electron is also ejected and the electron is known as **recoil electron**.

19.12.1 Importance of Compton effect

- Historically, Compton effect provided the direct confirmation of the particle nature of electromagnetic radiation. It had conclusively showed that photons carry energy and momentum like any material particle. Further, it proved that the momentum as well as the energy of electromagnetic radiation is quantized.
- Compton effect showed that the photon description applies not only to visible light but also to x-rays.

19.12.2 Experimental Arrangement

A schematic diagram of the apparatus for studying the Compton scattering is shown in Fig. 19.13(a). A monochromatic beam of K_α X-rays of wavelength λ_i from the molybdenum target were collimated by passing it through the slits L_1 and L_2 . The collimated beam was then made incident on a graphite block, G. The graphite block scattered X-rays in different directions. The scattered X-rays were detected by a Bragg's X-ray spectrometer. Bragg's law was used in determining the wavelength of scattered X-rays. X-rays diffracted by the crystal in the spectrometer were passed through an ionisation chamber to measure their intensity. The scattering angle θ was varied and corresponding X-ray wavelengths and their intensities were determined.

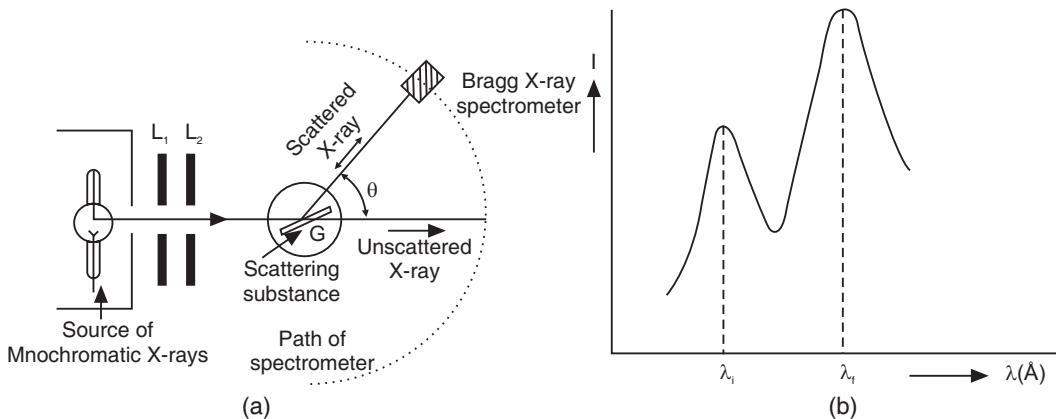


Fig. 19.13: (a) Compton's apparatus for the study of scattered X-rays
 (b) Unmodified and modified components

19.12.3 Experimental Results

- Graphs were plotted for the variation of intensity versus wavelength of the scattered X-rays. The curves were found to have only two peaks, as illustrated in Fig. 19.13(b). Thus, the measurements showed that though the incident X-rays are of only a single wavelength, the scattered X-rays contained two wavelengths, the original wavelength λ_i and another wavelength λ_f which is longer than λ_i . The primary (original) wavelength is called **unmodified component** and the second wavelength the **modified component**. The peak at λ_i is caused by X-rays scattered from electrons that are tightly bound to the target atoms, and the other peak of longer wavelength, λ_f is caused by X-rays scattered from free electrons in the target.
- The difference in the wavelength between the modified and unmodified components is called the **Compton shift**. It is denoted by

$$\Delta\lambda = \lambda_f - \lambda_i.$$

The Compton shift was found to depend solely on the scattering angle, θ and is completely independent of the primary wavelength, and the target material. The shift is also independent of the intensity of the incident radiation. It is found to vary with θ according to the following relation.

$$\Delta\lambda = \frac{h}{m_0 c} (1 - \cos \theta) \quad (19.26)$$

- The Compton shift $\Delta\lambda = 0$ when $\theta = 0^\circ$ and increases to a maximum value of $2h/m_0 c$ along $\theta = 180^\circ$.
- Only the intensity of the scattered radiation depends on the target material.

19.12.4 Wave Theory Fails to Explain Compton Effect

According to the wave theory, x-rays force electrons in the atoms of the target material to execute forced oscillations. The oscillating electrons emit radiation with the same frequency as that of the incident radiation. This radiation is called *Rayleigh-scattered radiation*. Further, the electrons radiate waves uniformly in all directions. Thus, as per wave theory

- the scattered radiation should have the same wavelength as that of incident radiation.
- the wavelength of the scattered radiation should not show dependence on the scattering angle, θ .

The above conclusions are contrary to the experimental observations. It means that the wave theory fails to explain the Compton effect.

19.12.5 Compton's Explanation

In 1923, Compton explained the effect on the basis of quantum theory of radiation. By Einstein's photon theory, the x-ray radiation consists of photons each having energy $h\nu$. The energy of a photon is directly proportional to the frequency of the radiation. So a change in wavelength implies a change in photon energy. An increase in wavelength of the scattered photons indicates that they had less energy than the incident ones. When a photon collides with an electron initially at rest, the photon transfers some energy and momentum to the electron. Hence, the energy of the scattered photon should decrease and its wavelength increase. Due to transfer momentum the electron gains kinetic energy and recoils with velocity v .

Assumptions

In the above explanation Compton assumed that

- (i) X-ray radiation is quantized and may be regarded as a stream of photons. A single photon carries an energy $h\nu$. The photons behave like true particles and have momentum like any mechanical particle.
- (ii) The electron in the atom of the target material is loosely bound and may be considered as at rest initially.
- (iii) The scattering event between a photon and an electron may be treated as a simple *elastic collision* between two particles. As a result of collision, the incident photon loses a part of its energy and is scattered with reduced energy at an angle θ with the direction of incident radiation. The electron gains kinetic energy equal to that of the energy lost by the photon, and recoils in a direction making an angle ϕ with the direction of incident radiation.

The geometry of Compton scattering is shown in Fig. 19.14.

19.12.6 Derivation of Expression for Compton Shift

Let an X-ray photon of energy $h\nu_i$ and momentum $h\nu_i/c$ collide with an electron at rest. As the collision is an elastic collision, the energy and momentum of the system are conserved.

The momentum and the kinetic energy of the electron before the collision is zero. However, according to relativity theory, the electron rest mass energy is m_0c^2 . After the collision event, the electron acquires a momentum ' p ' and recoils with kinetic energy. Since the electron may recoil at speeds comparable to the speed of light, we must use the relativistic expression for K.E., which is given by

$$K.E. = h\nu_i - h\nu_f = mc^2 - m_0c^2 \quad (19.27)$$

The *relativistic momentum* of the recoiling electron is $p = mv$. The relativistic momentum is related to the total energy through the following equation.

$$E = c\sqrt{p^2 + m_0^2c^2} \quad (19.28)$$

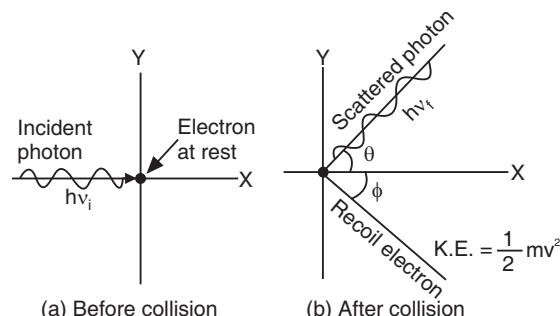


Fig. 19.14 Geometry of Compton scattering. (a) Before collision (b) After collision-The scattered photon has longer wavelength (less energy) than the incident photon.

Let the energy of the scattered photon be $h\nu_f$. We now apply the laws of conservation of energy and momentum to the *elastic collision* of photon and electron.

A. Conservation of Energy

According to the principle of conservation of energy,

$$\left. \begin{array}{l} \text{Total Energy of the} \\ \text{system before collision} \end{array} \right\} = \left. \begin{array}{l} \text{Total Energy of the} \\ \text{system after collision} \end{array} \right\}$$

That is

$$\left. \begin{array}{l} \text{Initial Energy of the Photon, } h\nu_i \\ + \\ \text{Initial Energy of the electron, } m_0 c^2 \end{array} \right\} = \left. \begin{array}{l} \text{Final Energy of the photon, } h\nu_f \\ + \\ \text{Final Energy of the electron, } c\sqrt{p^2 + m_0^2 c^2} \end{array} \right\}$$

$$\therefore h\nu_i + m_0 c^2 = h\nu_f + c\sqrt{p^2 + m_0^2 c^2}$$

The above equation can be rewritten as

$$h(\nu_i - \nu_f) + m_0 c^2 = c\sqrt{p^2 + m_0^2 c^2}$$

$$\frac{h}{c}(\nu_i - \nu_f) + m_0 c = \sqrt{p^2 + m_0^2 c^2}$$

Squaring the above equation on both sides and on simplification, we get

$$P^2 = \frac{h^2}{c^2}(\nu_i - \nu_f)^2 + 2m_0 h(\nu_i - \nu_f) \quad (19.29)$$

B. Conservation of Linear Momentum

Linear momentum is a vector quantity and the x - and y -components of linear momentum are conserved individually.

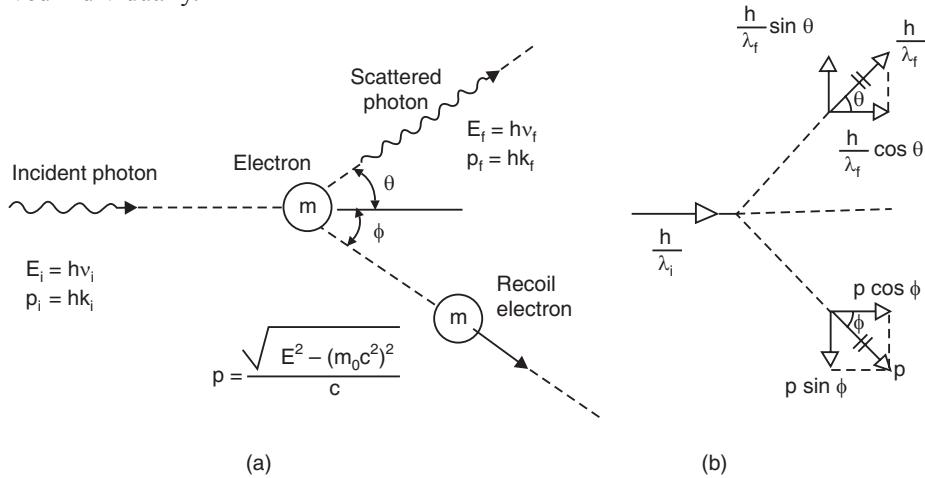


Fig. 19.15: Components of momentum before and after collision.

For the momentum in the x -direction before and after collision (Fig. 19.15), we have

$$\frac{h\nu_i}{c} + 0 = \frac{h\nu_f}{c} \cos \theta + p \cos \phi$$

$$\text{or } \frac{h\nu_i}{c} - \frac{h\nu_f}{c} \cos \theta = p \cos \phi \quad (19.30)$$

For the momentum in the y -direction (Fig. 19.15),

$$\begin{aligned} 0 &= \frac{h\nu_f}{c} \sin \theta - p \sin \varphi \\ \therefore \frac{h\nu_f}{c} \sin \theta &= p \sin \phi \end{aligned} \quad (19.31)$$

Squaring the equ.(19.30) and equ.(19.31), we respectively get

$$p^2 \cos^2 \phi = \frac{h^2}{c^2} (v_i - v_f \cos \theta)^2 \quad (19.32)$$

$$\text{and } p^2 \sin^2 \phi = \frac{h^2}{c^2} v_f^2 \sin^2 \theta \quad (19.33)$$

Adding equ.(19.32) to equ.(19.33), we get

$$p^2 = \frac{h^2}{c^2} (v_i^2 + v_f^2 - 2v_i v_f \cos \theta) \quad (19.34)$$

Equating equ.(19.29) to equ.(19.34), we obtain

$$\begin{aligned} \frac{h^2}{c^2} (v_i - v_f)^2 + 2m_o h (v_i - v_f) &= \frac{h^2}{c^2} (v_i^2 + v_f^2 - 2v_i v_f \cos \theta) \\ \text{or } (v_i - v_f)^2 + \frac{2m_o c^2}{h} (v_i - v_f) &= (v_i^2 + v_f^2 - 2v_i v_f \cos \theta) \\ \text{or } \frac{2m_o c^2}{h} (v_i - v_f) &= 2v_i v_f (1 - \cos \theta) \\ \text{or } (v_i - v_f) &= \frac{h}{m_o c^2} v_i v_f (1 - \cos \theta) \\ \text{or } \left[\frac{c}{\lambda_i} - \frac{c}{\lambda_f} \right] &= \frac{h}{m_o c^2} \frac{c^2}{\lambda_i \lambda_f} (1 - \cos \theta) \\ \text{or } (\lambda_f - \lambda_i) &= \frac{h}{m_o c} (1 - \cos \theta) \end{aligned} \quad (19.35)$$

The above equation is known as **Compton equation**. It does not contain the wavelength of the incident radiation and any parameter characteristic of the target material. Therefore, the change in wavelength is independent of both of them. It contains only the scattering angle θ . Therefore, *the wavelength shift depends solely on the scattering angle θ* .

Case 1: When $\theta = 0^\circ$, $\cos \theta = 1$. Therefore, $\Delta\lambda = 0$ and therefore $\lambda_f = \lambda_i$. That is, the change in wavelength is zero in the direction of the incident photon.

Case 2: In a direction normal to the direction of incident photon, i.e., when $\theta = 90^\circ$, $\cos \theta = 0$. Therefore, $\Delta\lambda = \lambda_f - \lambda_i = \frac{h}{m_o c} = 0.02426 \text{ \AA}$.

As the unit of the ratio $h/m_o c$ is angstroms, the ratio is designated as **Compton wavelength**, λ_C .

$$\therefore \lambda_C = \frac{h}{m_o c} = 0.02426 \text{ \AA} \quad (19.36)$$

Case 3: When $\theta = 180^\circ$, $\cos \theta = -1$. Therefore, $\Delta\lambda = \frac{2h}{m_o c} = 0.04852 \text{ \AA}$.

The change in wavelength is maximum at $\theta = 180^\circ$.

Thus, as θ varies from 0° to 180° , the wavelength of the scattered photon varies from λ_i to $\lambda_i + \frac{2h}{m_o c}$ and the Compton shift varies from 0 to a value $\frac{2h}{m_o c}$.

Compton effect is purely a quantum phenomenon. According to classical model, it should not occur. By letting h equal to zero, the quantum prediction should agree with the laws of classical physics. It is readily seen that as $h \rightarrow 0$, $\Delta\lambda \rightarrow 0$, which is in agreement with classical physics.

19.12.7 Experimental Verification

Compton measured the change in wavelength of X-rays scattered by graphite. The schematic diagram of his apparatus is shown in Fig.19.13(a). The monochromatic X-rays of wavelength $\lambda_i = 0.071 \text{ nm}$ were made to be incident on a graphite block. The X-rays scattered by the graphite block were detected by a Bragg's X-ray spectrometer. Bragg's law was used in determining the wavelength of scattered X-rays. X-rays diffracted by the crystal in the spectrometer were passed through an ionisation chamber to measure their intensity. The scattering angle is θ is varied and corresponding X-ray wavelengths are determined. Thus, Compton measured how X-ray intensity depends on wavelength at various scattering angles. The wavelength of primary X-rays was measured by making them fall directly on the crystal of the spectrometer. The results obtained for four values of θ are shown in Fig.19.16. The solid vertical line corresponds to the *primary (unmodified) wavelength*, λ_i , and the broken vertical lines correspond to the *modified wavelength* λ_f .

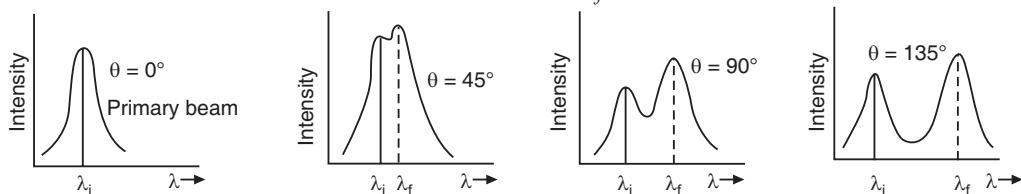


Fig. 19.16

Compton found that the shift in wavelength depended solely on the scattering angle, θ and it is completely independent of the primary wavelength and the target material. Compton's experimental results were in excellent agreement with the predictions of equation (19.35).

19.12.8 Modified and Unmodified Components

In Compton scattering, the scattered radiation consists of two component wavelengths, namely, the modified and unmodified wavelengths.

The *modified wavelength* arises due to the collision of X-ray photons with either *free or loosely bound* electrons. For high-energy photons, most of the atomic electrons appear *free* and hence a large fraction of incident X-rays undergoes a wavelength shift. Therefore, in the case of low atomic number targets, the modified component will be more intense (see Fig. 19.17).

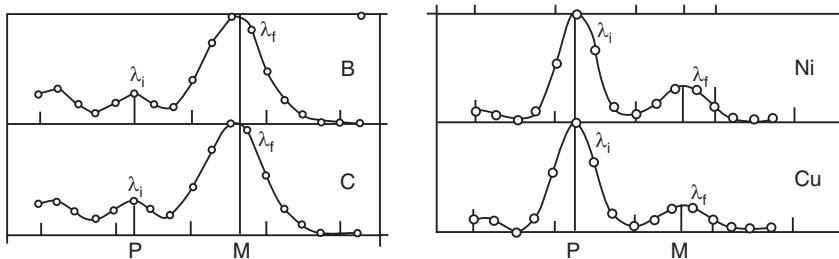


Fig. 19.17: Modified M (λ_f) and unmodified P (λ_i) wavelengths in case of different scattering materials. Note that the modified component is more intense in lower atomic number materials whereas the unmodified component is stronger in higher atomic number materials.

The *unmodified component* arises when photons are scattered by *tightly bound* electrons. In this case, the whole atom is involved in the collision and hence the value of m_o in equ. (19.35) is to be replaced by the mass of the whole atom. Since the mass of the atom is many times larger than that of an electron, the Compton shift is negligible and correspondingly there is no change in wavelength. Hence, the intensity of unmodified component is large.

In the case of scatterers of large atomic number, there are many electrons, which require high-energy X-ray photons for being scattered. Therefore, in the case of large atomic number targets, the unmodified component will be more intense (see the curves for Ni and Cu in Fig. 19.17).

For visible photons, all the atomic electrons appear bound and hence Compton shift is not observed.

Secondly, it is seen from Fig. 19.17 that the modified component is broader than the unmodified component. The broadening occurs because electrons in the target are not at rest and hence the change in the photon energies (Compton shift) is spread out.

Example 19.10. *X-rays of 0.5 \AA are scattered by free electrons in a block of carbon through 90° . Find the velocity of recoil electrons.*

Solution:

$$\text{K.E. of the recoil electrons} = h\nu_i - h\nu_f = 12400 \left(\frac{1}{\lambda_i} - \frac{1}{\lambda_f} \right) eV (1.602 \times 10^{-19} \text{ J/eV})$$

$$\therefore \text{K.E.} = 12400 \left(\frac{1}{0.5} - \frac{1}{0.5243} \right) (1.602 \times 10^{-19}) \text{ J} = 1.84 \times 10^{-16} \text{ J}$$

$$\begin{aligned} \therefore v &= \left[\frac{2(\text{K.E.})}{m} \right]^{1/2} = \left[\frac{2(1.84 \times 10^{-16} \text{ J})}{9.11 \times 10^{-31} \text{ Kg.}} \right]^{1/2} \\ &= \left\{ 4.04 \times 10^{14} \frac{\text{kg.m}^2/\text{s}^2}{\text{kg}} \right\}^{1/2} = 2 \times 10^7 \text{ m/s.} \end{aligned}$$

Example 19.11. *X-ray photon of wavelength 0.3 \AA is scattered through an angle 45° by a loosely bound electron. Find the wavelength of scattered photon.*

Solution: The wavelength of the scattered photon is given by $\lambda_f = \frac{h}{m_o c} (1 - \cos \theta) + \lambda_i$

$$\begin{aligned} \therefore \lambda_f &= \frac{6.63 \times 10^{-34} \text{ Js}}{9.11 \times 10^{-31} \text{ kg} \times 3 \times 10^8 \text{ m/s}} (1 - \cos 45^\circ) + 0.3 \times 10^{-10} \text{ m} \\ &= 0.307 \text{ \AA}. \end{aligned}$$

Example 19.12. Photon of initial energy 90 keV undergoes Compton scattering at an angle 60° . Find:

- (i) the energy of the scattered photon and
- (ii) the recoil energy of the electron.

Solution. (i)

$$E_i = \frac{hc}{\lambda_i} = 90 \text{ keV} = 90 \times 10^3 \left(1.6 \times 10^{-19} \text{ J} \right) = 1.44 \times 10^{-14} \text{ J}$$

∴

$$\lambda_i = \frac{hc}{E_i} = \frac{6.63 \times 10^{-34} \text{ Js} \times 3 \times 10^8 \text{ m/s}}{1.44 \times 10^{-14} \text{ J}}$$

$$= \frac{1.989 \times 10^{-25} \text{ J} \cdot \text{m}}{1.44 \times 10^{-14} \text{ J}} = 0.138 \times 10^{-10} \text{ m}$$

$$\lambda_f = \frac{h}{m_o c} (1 - \cos \theta) + \lambda_i$$

$$= 0.02426 \times 10^{-10} \text{ m} (1 - \cos 60^\circ) + 0.138 \times 10^{-10} \text{ m}$$

$$= 0.15 \times 10^{-10} \text{ m}$$

∴

$$E_f = \frac{hc}{\lambda_f} = \frac{6.63 \times 10^{-34} \text{ Js} \times 3 \times 10^8 \text{ m/s}}{0.15 \times 10^{-10} \text{ m}}$$

$$= \frac{1.989 \times 10^{-25} \text{ J} \cdot \text{m}}{0.15 \times 10^{-10} \text{ m}} = 1.32 \times 10^{-14} \text{ J} = 82.73 \text{ keV}.$$

(ii) Recoil energy of electron = $E_i - E_f = (90 - 82.73) \text{ keV} = 7.27 \text{ keV}$.

Example 19.13. X-rays with initial wavelength $0.5 \times 10^{-10} \text{ m}$ undergo Compton scattering. For what scattering angle is the wavelength of the scattered X-rays greater than that of the incident X-rays by one percent?

Solution: $\lambda_i = 0.5 \times 10^{-10} \text{ m}$. It is given that $\frac{\lambda_f - \lambda_i}{\lambda_i} = 1\% = 0.01$

∴

$$\Delta\lambda = \lambda_f - \lambda_i = 0.5 \times 10^{-10} \text{ m} \times 0.01 = 0.5 \times 10^{-12} \text{ m}$$

$$\Delta\lambda = \frac{h}{m_o c} (1 - \cos \theta) = 0.02426 \times 10^{-10} \text{ m} \times (1 - \cos \theta)$$

∴

$$0.5 \times 10^{-12} \text{ m} = 0.02426 \times 10^{-10} \text{ m} \times (1 - \cos \theta)$$

or

$$(1 - \cos \theta) = 0.2061$$

∴

$$\theta = 37.4^\circ.$$

Therefore, the required angle is 37.4° .

19.13 PAIR PRODUCTION

Although the photoelectric effect and Compton effect provided the earliest experimental evidence in support of the quantisation of electromagnetic radiation, there are numerous other experiments that can be interpreted only when we assume the existence of photons as discrete quanta of electromagnetic radiation. Pair production is one of such processes. When a photon encounters atoms, the photon loses all its energy and two particles are created: an electron and a positron. The photon energy is converted into the relativistic total energies E_+ and E_- of the positron and electron.

$$h\nu = E_+ + E_- = (m_e c^2 + K_+) + (m_e c^2 + K_-) \quad (19.37)$$

Since K_+ and K_- are always positive, the photon must have energy at least $2m_e c^2 = 1.02$ MeV in order for this process to occur; such high energy photons are in the region of nuclear γ -rays. This process will not occur unless there is an atom nearby to supply the necessary recoil momentum.

19.14 WAVE-PARTICLE DUALITY

The ability to produce interference effects is an essential characteristic of waves. We are familiar with the Young's famous double-slit experiment. In this experiment, light passes through two closely spaced slits and produces a pattern of bright and dark fringes on a screen. The fringe pattern is a direct indication that interference is occurring between the light waves coming from each slit. The phenomena of interference, diffraction and polarization give exclusive evidence for the wave behaviour of light. These phenomena require two waves to be present at the same position at the same time. It is impossible for two particles to occupy the same position at the same time to produce the observed effects.

However, blackbody radiation, photoelectric effect, Compton effect etc give exclusive evidence for the particle behaviour of light. Wave theory miserably failed to explain these effects. Thus, on one hand light resembles a continuous wave of frequency, v and on the other hand it resembles a collection of particles having energy E and momentum p . The particle nature and wave nature are in fact contradictory.

- (i) A particle occupies a definite position in space and hence it must be very small.
- (ii) A wave spreads out and occupies a relatively large region of space and cannot be attributed to a particular location in space.

As we need both descriptions to complete our understanding, we have to reconcile that light behaves as an advancing wave in some phenomena and it behaves as a flux of particles in some other phenomena. In short, light exhibits **wave-particle duality**.

Let us consider the interference pattern obtained on a screen in the double-slit experiment. We may interpret the pattern as follows: at positions of bright fringes, many photons are hitting the screen and at the positions of dark fringes no photons are incident. Even if we reduce the intensity of light to the extent that only a few photons pass through the slits, we observe the usual interference pattern. Such an experiment was performed in 1909 by G.I.Taylor. We know from experiments that sending light through slits gives interference patterns and that the patterns are built by individual photons. To reconcile the wave and particle aspects, we have to regard the pattern as a *statistical distribution* that tells us how many photons, on an average go to different places. In other words, the pattern indicates the *probability* for an individual photon to arrive at a particular place on the screen. However, we cannot predict exactly where an individual photon will go.

The reason for the manifestation of wave-particle dualism can be understood if one considers the entire electromagnetic spectrum. At the lower frequency end of the spectrum (see Fig.19.18) are radio waves whose wavelengths are of the order of a few hundred metres. The energy of these waves therefore spreads over such a large volume of space that the energy available at any point is insignificantly small and therefore, the particle nature is not observable. On the higher frequency side of the spectrum are UV rays, X-rays and γ -rays. The wavelengths of these waves are so short ($<10^{-10}$ m) that the wave energy is literally concentrated in a point of negligibly small dimension and the wave properties are less noticeable compared to the particle properties. Hence, the wave nature stands out at

lower frequencies and at higher frequencies the particle nature dominates. The visible region represents the *transition region* where both aspects can be observed.

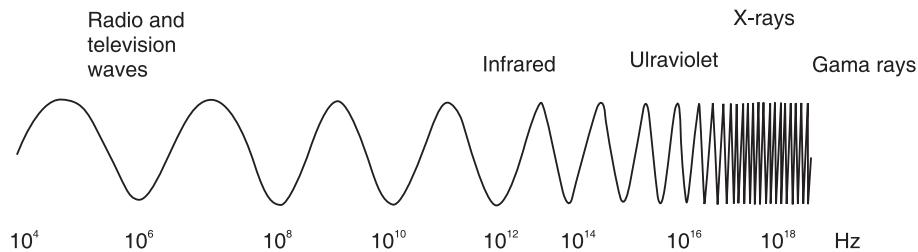


Fig. 19.18

We know that the intensity of light is given by $I = |E|^2$

According to equ.(19.15) $I = Nhv$.

A comparison of these equations indicates that the square of the amplitude of a light wave at a point in space is proportional to the number of photons arriving at that point. That is

$$N \propto |E|^2 \quad (19.38)$$

Therefore, we conclude that

$$\text{Probability to observe photons} \propto |\text{Electric field amplitude}|^2 \quad (19.39)$$

Equ.(19.39) provides the ultimate connection between the wave behaviour and the particle behaviour of light. The amplitude of a light wave determines the **probability** that a photon can be found at a particular point in space.

An acceptable explanation of wave-particle duality has been given by the '**quantised electromagnetic field theory**', i.e. '**quantum electrodynamics**' (**QED**). This theory extended the concept of energy levels of an atomic system and the Heisenberg uncertainty principle to electromagnetic fields. It combines particle and wave properties into a single theory which predicts as well as explains both types of behaviour.

19.15 SPECTRAL LINES

It had been known for a long time that light is emitted by gases when they are heated to high temperature. The light emitted from a hot gas may be examined with a prism or grating. In practice, the source of light is placed behind a narrow linear slit and the narrow beam of light emerging from the slit is passed through a grating or a prism as shown in Fig.19.19 (a). The light breaks up into several beams according to wavelength. These beams produce a set of narrow lines with dark gaps between them on a photographic plate kept in their path. The lines recorded on the photographic plate are called **spectral lines** and the whole set of lines is known as a **spectrum**. The term spectral line denotes light with a discrete wavelength and corresponds to the line image of the linear slit. The spectrum obtained from a hot body is called **emission spectrum**. Each chemical element has unique spectrum lines that are characteristic of the element. Some elements, such as hydrogen and neon, have relatively few lines; others, such as iron have many thousand lines. The line spectra of elements are distinguished by their wavelength, the relative position, intensity, and number of lines. Lines characteristic of each individual element occur not only in the visible region but also in IR and UV regions. Kirchhoff and Bunsen were the first to carry out systematic investigations on line spectra during the years 1854–1859.

In addition to emitting light at specific wavelengths, an element can also absorb light at those specific wavelengths. The spectrum in the latter case is known as the **absorption**

spectrum. The absorption spectrum of an element is obtained by passing white light through the element in its gaseous state. It is found that each line in the absorption spectrum of a given element coincides with a line in the emission spectrum of that element. The conclusion is that an element can absorb *only* light of those wavelengths, which when excited, it can emit.

Line spectra are produced by individual atoms of elements. The emission of light is the outcome of certain processes taking place inside the atom. The study of emission and absorption spectra of elements helped us understand the structure of the atom. In the spectrum, the wavelengths of lines form an infinite sequence following each other in an *orderly* fashion and converging to a lower limit in wavelength. A group of such lines exhibiting systematic regularities are called **spectral series**.

Of all spectra, the emission spectrum of hydrogen is the simplest [see Fig. 19.19 (b)]. In 1885 Johann Balmer found a regularity in the position of the lines of the hydrogen spectrum. He obtained the following simple empirical relation to compute the wavelengths of the lines in the visible region of the spectrum.

$$\lambda = 3646 \left[\frac{n^2}{n^2 - 4} \right] \text{ Å} \quad n = 3, 4, 5, \dots \quad (19.40)$$

Balmer's formula was rearranged by J.R. Rydberg in the following form.

$$\frac{1}{\lambda} = R \left[\frac{1}{2^2} - \frac{1}{n^2} \right] \quad (19.41)$$

where R is known as **Rydberg constant** and has the value

$$R = 1.0973732 \times 10^7 \text{ m}^{-1}$$

The reciprocal of wavelength $1/\lambda$ is called the **wave number**.

All the lines of the hydrogen spectrum satisfying the relation (19.41) are termed **Balmer series**. Subsequently, additional series of hydrogen spectrum were discovered. They are called Lyman, Paschen, Brackett, and Pfund series named after their discoverers. Such regularities were also observed for other atoms. An over-all formula

$$\frac{1}{\lambda} = R \left[\frac{1}{n_f^2} - \frac{1}{n_i^2} \right] \quad (19.42)$$

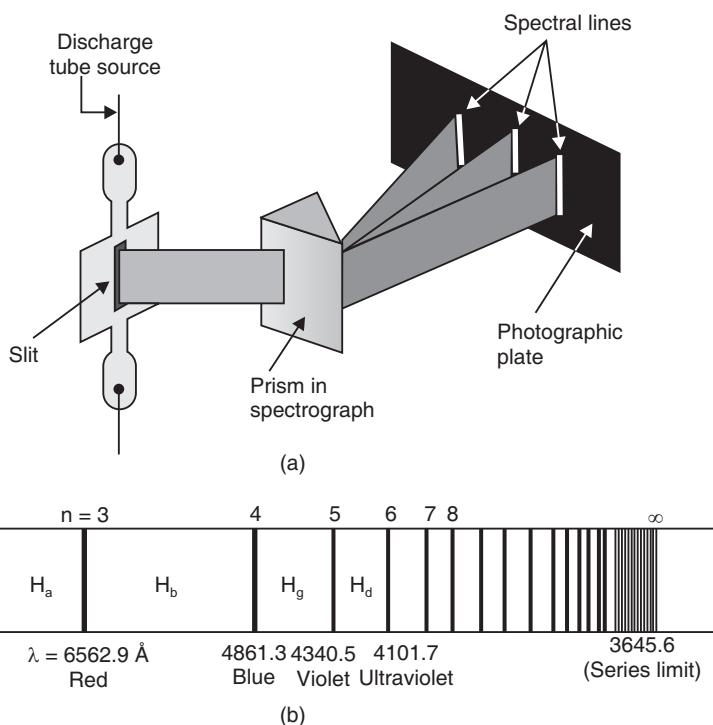


Fig. 19.19

describes all the series. The above formula which modified Balmer's formula is merely descriptive and does not provide any insight into the mechanism responsible for the emission of spectral lines. Attempts to explain the origin of discrete line spectra on the basis of classical physics met with severe failure. The Rutherford's planetary model of atom also was not successful in explaining the atomic spectra.

19.16 ATOMIC STRUCTURE

J.J.Thomson discovered electron in 1897. He put forward the first ever model of the atom. Thomson represented it as a positively charged sphere with electrons immersed in it here and there. When unperturbed, both electrons and the positive charge were believed to be at rest. When in motion the electrons lose their energy in the form of radiation.

Basing on the results of α -scattering experiments conducted by him, Rutherford proposed 'planetary model' of the atom. The atom consists of a relatively small positive nucleus at the centre and electrons orbiting well away from it. But an electron moving in an orbit around the nucleus is in a state of accelerated motion and according to Maxwell, any accelerated charge generates electromagnetic waves. Consequently, the orbiting electron loses continually its energy until it approaches the nucleus and falls into it. This leads to the prediction that an atom exists for only about 10^{-8} s; but in fact atoms have remarkable stability. Secondly, if an electron is rotating around the nucleus, it radiates light whose frequency is the same as that of the rotation. As the electron spirals and falls onto the nucleus the light frequency continually increases and therefore the radiation given out by Rutherford atom must be continuous. However, the atomic spectra were in fact a set of discrete lines.

19.17 BOHR'S MODEL OF ATOM

In 1913, Niels Bohr removed the deficiencies of Rutherford model by incorporating some adhoc quantum hypotheses into it. Bohr explained the origin of line spectra in general terms on the basis of two central ideas. One is the concept of photon and the other is the concept of energy levels of atoms. Bohr combined these two concepts nicely to explain how light is emitted and why the spectral lines are arranged in an orderly fashion.

Bohr fused the quantum idea with the purely mechanical model of Rutherford and introduced the following three revolutionary adhoc postulates.

- Electrons revolve about a nucleus only in certain special orbits called **stationary orbits**, though an infinite number of orbits are mechanically allowed. While moving in the permitted (stationary) orbits, electrons do not emit or absorb electromagnetic radiation though they are in accelerated motion. Hence the atom is stable.
- The allowed electron orbits are those for which the angular momentum is an integral multiple of \hbar , where \hbar is Planck's constant. The angular momentum of electron is $I\omega = (mr^2)(v/r) = mv r$. Accordingly

$$mv r = n\hbar \quad (n = 1, 2, 3, \dots) \quad (19.43)$$

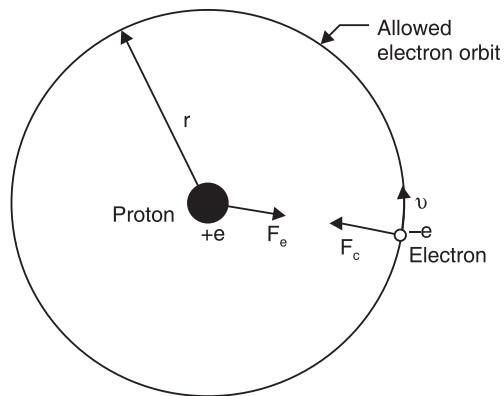


Fig. 19.20: Bohr's model of the hydrogen atom.

- (iii) The atom radiates energy *only* when the electron jumps from one of the upper allowed orbits, with energy say E_2 , to a lower allowed orbit, with energy say E_1 . The change in energy during the transition is given by

$$E_2 - E_1 = h\nu \quad (19.44)$$

Note that the frequency of the emitted light is not determined by the motion of the electron but is governed by changes in its motion.

Bohr obtained values of the energies of various states of hydrogen atom by assuming that an electron with velocity 'v' revolves in a circular orbit of radius 'r' around the nucleus. For a dynamically stable orbit, the centripetal force experienced by the electron equals the force of electrical attraction between the nucleus and the electron. Thus,

$$\frac{mv^2}{r} = \frac{e^2}{4\pi\epsilon_0 r^2} \quad (19.45)$$

Using (19.43) and (19.45) to eliminate v, we obtain for the radius of the orbits as

$$r = \frac{\epsilon_0 h^2}{\pi me^2} n^2 = r_o n^2 \quad (19.46)$$

where r_o is the radius of the first Bohr orbit ($n = 1$) which is given by

$$r_o = \frac{\epsilon_0 h^2}{\pi me^2} = 0.53 \text{ \AA}. \quad (19.47)$$

The above value for the size of the Bohr radius agrees with the value 0.5 \AA obtained from the estimates based on kinetic theory of gases.

The total energy for a specific electron orbit is obtained by adding the kinetic energy and the potential energy of the electron with respect to the nucleus. Thus,

$$E = \frac{1}{2}mv^2 - \frac{e^2}{4\pi\epsilon_0 r}$$

where the first term represents the kinetic energy of electron due to its angular motion and the second term is the potential energy due to electrical attraction. Substituting the value of r from (19.46) and eliminating v using (19.43), we get

$$E = -\frac{e^2}{8\pi\epsilon_0 r^2} \quad (19.48)$$

Using the value of r into the above equation (19.48), we find that electron can have only certain discrete negative values of energy E_n given by

$$E_n = -\frac{me^4}{8\epsilon_0^2 h^2} \cdot \frac{1}{n^2} = -\frac{E_o}{n^2} \quad (19.49)$$

where

$$E_o = \frac{me^4}{8\epsilon_0^2 h^2} = 13.6 \text{ eV} \quad (19.50)$$

Equ. (19.49) shows that the energy of an orbit varies as $-1/n^2$. The energy is larger for larger values of n and hence for bigger orbits. For each value of n there is a corresponding definite energy E_n . The values of E_n are called **allowed energy values** and the corresponding energy levels are called **allowed energy states**. The other values of energy or intermediate energy states are forbidden. The allowed energy states correspond to stationary orbits of electron in the atom. The energy of the system is negative. The negative energy implies that

the electron is in a **bound state** and is not free to move away. It also indicates that energy is to be supplied from outside in order to separate the electron and the nucleus.

19.17.1 Rydberg Constant

According to Bohr's third postulate, an atom emits a photon of frequency ν when it jumps from a higher energy orbit to a lower energy orbit. Thus,

$$\nu = \frac{E_f - E_i}{h} \quad (19.51)$$

Now from equ. (19.49) we can write $E_f = -\frac{me^4}{8\epsilon_o^2 h^2} \cdot \frac{1}{n_f^2}$ and $E_i = -\frac{me^4}{8\epsilon_o^2 h^2} \cdot \frac{1}{n_i^2}$

$$\therefore h\nu = \frac{me^4}{8\epsilon_o^2 h^2} \left[\frac{1}{n_f^2} - \frac{1}{n_i^2} \right]$$

$$\text{or } \nu = \frac{me^4}{8\epsilon_o^2 h^3} \left[\frac{1}{n_f^2} - \frac{1}{n_i^2} \right]$$

As $\nu = \frac{c}{\lambda}$, we write the above equation as

$$\frac{c}{\lambda} = \frac{me^4}{8\epsilon_o^2 h^3} \left[\frac{1}{n_f^2} - \frac{1}{n_i^2} \right]$$

$$\text{or } \frac{1}{\lambda} = \frac{me^4}{8\epsilon_o^2 c h^3} \left[\frac{1}{n_f^2} - \frac{1}{n_i^2} \right]$$

$$\text{or } \frac{1}{\lambda} = R \left[\frac{1}{n_f^2} - \frac{1}{n_i^2} \right]$$

$$\text{where } R = \frac{me^4}{8\epsilon_o^2 c h^3} \quad (19.52)$$

Substituting the known values into equ. (19.52), we obtain $R = 1.097 \times 10^7 \text{ m}^{-1}$ which is in excellent agreement with the experimental value for Rydberg constant. It is a triumph for Bohr's theory.

19.17.2 Spectral Series of Hydrogen Atom

The origin of the spectral series can easily be understood from Bohr's theory. When hydrogen atom is excited, it returns to its ground state by emitting the energy it had absorbed earlier. The energy is emitted by the excited electron in the form of radiation of different wavelengths as it jumps down from a higher to lower orbit.

If the electron in the hydrogen atom is initially in one of the stationary orbits with $n_i = 2, 3, 4, \dots$ and makes transition to the final stationary orbit with $n_f = 1$, then the wavelengths emitted due to these transitions given by equ.(19.42) are given by $\frac{1}{\lambda} = R \left[\frac{1}{1^2} - \frac{1}{n_i^2} \right]$. These lines belong to **Lyman series**.

Again, if transitions take place from initial stationary orbits with $n_i = 3, 4, 5, \dots$ to the final stationary orbit with $n_f = 2$, then the wavelengths emitted due to these transitions are given by $\frac{1}{\lambda} = R \left[\frac{1}{2^2} - \frac{1}{n_i^2} \right]$. These lines belong to Balmer series. Similarly, transitions taken place from initial stationary orbits with $n_i = 4, 5, 6, \dots$ to the final stationary orbit with $n_f = 3$, produce lines with wavelengths given by $\frac{1}{\lambda} = R \left[\frac{1}{3^2} - \frac{1}{n_i^2} \right]$. These lines belong to **Paschen series**.

The remaining series were not observed by the time Bohr proposed his theory of atom. He predicted in 1913 in his theory the other series in UV and far infra red regions. They were subsequently discovered establishing the validity of Bohr's theory. If transitions take place from initial stationary orbits with $n_i = 5, 6, 7, \dots$ to the final stationary orbit with $n_f = 4$, then the wavelengths emitted due to these transitions are given by $\frac{1}{\lambda} = R \left[\frac{1}{4^2} - \frac{1}{n_i^2} \right]$. These lines belong to **Brackett series**, and so on.

19.17.3 Limitations of Bohr's Theory

Bohr's theory employed a semi-classical model where quantum postulates are introduced into a mechanical model. But it explained with remarkable success the origin of line spectra in case of hydrogen atom, which is a one-electron atom. However the theory could not give correct predictions for two electron atoms. The Bohr hypothesis established the relation of wavelengths to energy levels, but it did not provide any general principles for predicting the energy levels of a particular atom. It did not explain why angular momentum is quantized and why atoms make transitions. However, Bohr's theory contributed immensely to the understanding of atomic structure. A more general understanding of atomic structure and energy levels is provided by the quantum mechanics.

19.18 FRANK-HERTZ EXPERIMENT

The atomic energy levels are quantized according to Bohr model. The existence of discrete energy levels of an atom was confirmed by the experiments carried out in 1914 by the German physicists James Franck and Gustav Hertz.

A schematic view of the apparatus used by Franck and Hertz is shown in Fig. 19.21. It consists of a gas tube filled with mercury vapour at low pressure of about 1 mm Hg. The glass tube contained three electrodes, namely cathode C, grid G and anode P. The cathode emits electrons through thermionic emission when it is heated by the filament F. The grid and the anode are kept close and away from the cathode. The electrons emitted by the cathode are accelerated by the grid which is maintained at a positive potential V with respect to the cathode. The potential difference between the cathode and grid can be smoothly varied with the help of potentiometer. The anode is kept at a small negative potential with respect to the grid such that the potential difference between the anode and the grid is of the order of 0.5 V. The electrons reaching the grid acquire energy eV and after passing the grid they experience a retarding electric field.

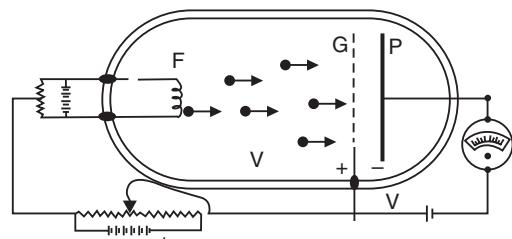


Fig. 19.21

Electrons that have enough kinetic energy, overcome the retarding field and reach the anode. Such electrons cause the anode current which is measured by a suitable milli-ammeter.

The variation of anode current I with the potential difference V between the cathode and grid was studied. The results obtained in the experiment are shown in Fig.19.22. The current was found to increase first and after reaching a maximum at about 4.9 V, it decreased sharply with a further increase in voltage. It reached a minimum and then again began to increase. Maxima of current occurred at 9.8V, 14.7V,.... and so on.

The above results can be interpreted as follows. For low values of V , the collisions between an electron and a mercury atom are of elastic nature. The electron does not lose its energy in the process but only changes its direction. And also the atom is not excited. The electron on reaching G has sufficient kinetic energy to reach the anode P. As V is increased, more and more electrons are accelerated and anode current increases.

According to Bohr's third postulate, the amount of energy that a mercury atom can receive from a colliding electron would be equal to the energy difference between its ground state and one of the excited states. The excited state nearest to the ground state for a mercury atom is at about 4.9 eV above the ground state. It may therefore be expected that as long as the energy of an electron is less than 4.9 eV, it experiences only elastic collisions. When the electron accumulates energy of 4.9 eV or more, the collisions cease to be elastic. The electron has enough energy now to undergo inelastic collision with a mercury atom and transfer to the atom an energy of 4.9 eV. After that the electron continues its motion with a lower velocity. Consequently, it cannot overcome the retarding field beyond the grid and returns to the grid. Therefore, it cannot contribute to the anode current and the anode current decreases.

The mercury atom that received an energy of 4.9 eV gets excited to its first excited state. Such events occur again and again, whenever the potential difference attains the values $2 \times 4.9 \text{ V} = 9.8 \text{ V}$, $3 \times 4.9 \text{ V} = 14.7 \text{ V}$,....etc. at these values of potential difference electrons experience two, three,... etc inelastic collisions. Therefore, a sharp drop in the anode current is observed each time when the grid potential reaches the values given by

$$V = nV_1 \quad (n = 1, 2, 3, \dots)$$

where V_1 corresponds to the energy of the first excited state, namely 4.9 eV.

The mercury vapour in the tube was observed to emit ultraviolet light at a wavelength of 2537 Å. The UV light must have been emitted by the excited mercury atoms returning to the ground state. According to Bohr's postulate, the wavelength of light emitted by mercury atoms should be

$$\begin{aligned} \lambda &= \frac{hc}{E_f - E_i} = \frac{hc}{\Delta E} \\ &= \frac{(6.62 \times 10^{-34} \text{ Js})(3 \times 10^8 \text{ m/s})}{(4.9V)(1.602 \times 10^{-19} \text{ C})} = 2533 \text{ Å}. \end{aligned}$$

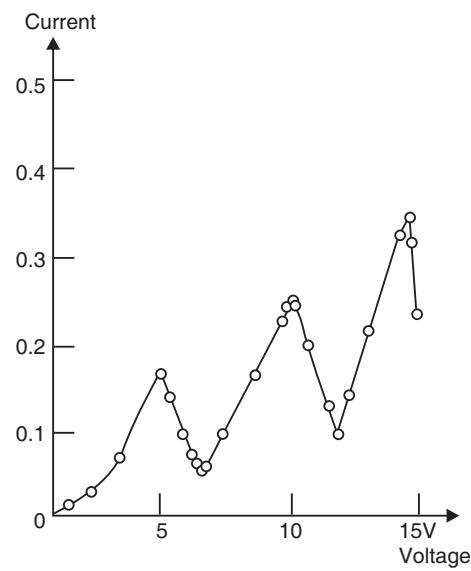


Fig. 19.22

The good agreement between the calculated and experimental values of the wavelength lends support to the Bohr model. Thus, the Franck-Hertz experiment is a direct proof to the existence of discrete energy levels in atoms and hence to the Bohr model of the atom. Franck and Hertz were awarded the Nobel Prize in physics in 1925.

The potentials required to accelerate electrons for imparting them kinetic energy sufficient to raise the atoms to different excited levels are called **critical potentials**. The critical potential corresponding to the ionization of the bombarded atom is known as **ionization potential**.

19.19 ENERGY LEVEL DIAGRAM

The electron in a hydrogen atom can have any one of a series of negative energies, E_1, E_2, E_3, \dots which are referred to as **energy states**. It is easier to represent the *energy states* of an atom in the form of **energy levels**, drawn in the form of horizontal lines rather than in the form of circular *orbits*. The words energy state and energy level are used interchangeably. (To be more specific, the hydrogen atom has two states in its ground level corresponding to -13.60 eV , eight states in its -3.40 eV level that is the first excited level and so on). The sequential representation of energy levels, as shown in Fig. 19.23, is known as an **energy level diagram**.

The energy-level diagram of hydrogen is shown in Fig. 19.23 and is the simplest. The diagrams are characteristic of atom and are complicated for complex atoms. In an energy level diagram, the vertical axis represents electron energy and the horizontal lines are the allowed energy states each identified by a specific value of quantum number n .

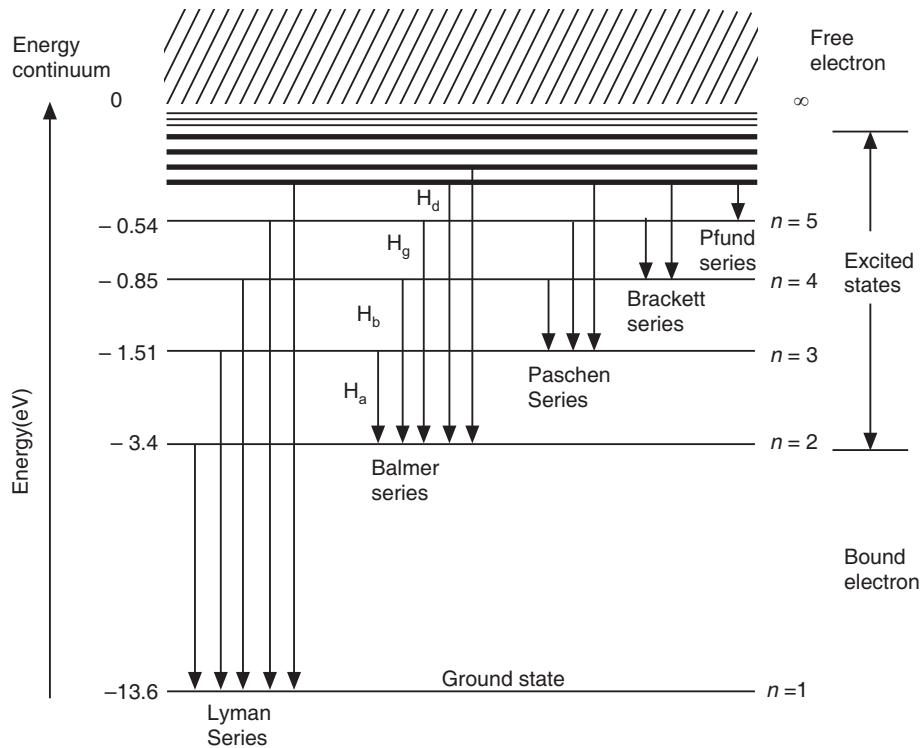


Fig. 19.23

The lowest energy level is the level in which the electron revolves in the innermost orbit ($n = 1$); that level is called the **ground state**. It is the *normal unexcited state* of the atom. The atom can reside in its ground state for an indefinite time. For hydrogen, the ground state is at

– 13.6 eV. An atom can get raised to a higher level only by absorbing energy from an external source. The energy may be supplied through inelastic collisions with other atoms, when the hydrogen gas is heated or subjected to an electric discharge. When the atoms are raised to higher levels, they are said to be **excited**. For example, the first excited level, $n = 2$, is at –3.4 eV; and so on. The energy difference between the energy of an excited state and the energy of the ground state is called **excitation energy**. An atom cannot stay in an excited state for more than 10^{-8} s and spontaneously returns to the ground state. It is due to Coulomb attraction exerted by the positive nucleus on the electron moved to a higher orbit. The time for which an atom stays in an excited state is known as its **lifetime** in that state.

The energy levels in the energy level diagram are not evenly spaced. They get closer and closer at the upper end of the diagram. It is because the energies of the allowed states are proportional to $1/n^2$ and as n increases the energy difference between successive levels rapidly decreases. The upper end of the diagram is bounded by zero energy level for which $n = \infty$. It is called the **ionization level** of the atom. When the electron reaches the ionization level it gets separated completely from the nucleus and the atom is said to be **ionized**. The energy needed to remove an electron from an atom in its ground state is called its **ionization energy**. In case of hydrogen it is equal to 13.6 eV.

The above picture can be extended to atoms of all elements. Each atom has a set of possible energy levels. An atom can have an amount of internal energy equal to any one of these levels, but it cannot have energy intermediate between two levels. All isolated atoms of a given element have the same set of energy levels, but atoms of different elements have different sets.

Every atom has a **ground level** that represents the minimum internal energy that the atom can have. An atom shut off from outside influences will always lie in the ground level. All the levels higher to it are **excited levels**. If the atom is disturbed by exposing it to some radiation, the total energy of electron is increased and the atom is raised to an excited level.

19.19.1 Atomic Transitions

According to Bohr's third postulate, the *transition* of atom from one energy state to another is accomplished by transfer of energy. If energy is supplied to the system consisting of atoms, the atom is raised from a lower energy state E_1 to a higher excited state, E_2 . Such a transition, $E_1 \rightarrow E_2$ is called **absorption**. It is customary to indicate an atomic transition between two energy states by a vertical arrow in the energy level diagram. An upward arrow represents an upward transition of the electron consequent to absorption of energy (Fig.19.24).

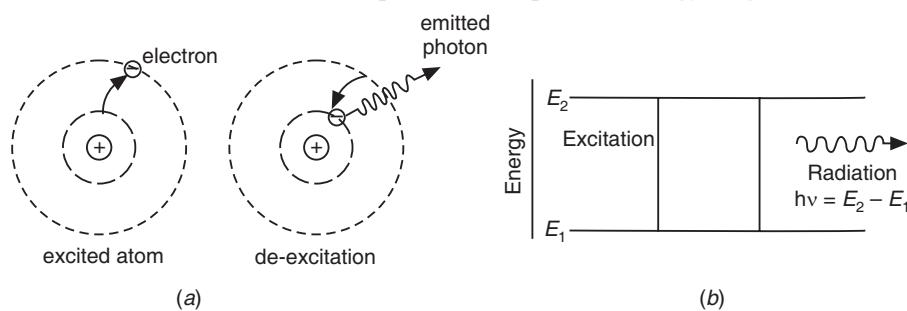


Fig. 19.24

The transition from a lower state to an excited state can occur only if the difference in energy is exactly equal to the photon energy $h\nu$. The atom does not stay in an excited state indefinitely. It usually returns on its own to the lower state after about 10 ns by emitting a photon.

The downward transition $E_2 \rightarrow E_1$ is called **emission**. A downward arrow represents a downward transition. During the emission process, as the atom returns from a higher energy state E_2 to a lower energy state E_1 , it emits a quantum of energy, $h\nu$. If the photon energy is $h\nu = \frac{hc}{\lambda}$, then conservation of energy gives

$$h\nu = \frac{hc}{\lambda} = E_2 - E_1 \quad (19.53)$$

Vertical lines in Fig.19.23 and Fig.19.24 show the jumps that electrons can make from one level to another. The location of these transitions, or **emission lines**, can be shown on a wavelength graph. The ensuing pattern corresponds exactly to that recorded by the spectrographic plates. Thus, **position** is one of the important properties of spectral lines.

The average time spent by the atom in an excited level is called the **lifetime** of the level. The lifetime is usually of the order of $10^{-8}s$. Besides the short-lived excited states, there are states with average lifetimes greater than 10 milliseconds or even as long as several seconds. They are called **metastable states**.

Example 19.14. The energy of a particular state of an atom is 5.36 eV and the energy of another state is 3.45 eV. Find the wavelength of the light emitted when the atom makes a transition from one state to the other.

Solution.

$$(E_1 - E_2) = h\nu = \frac{hc}{\lambda}$$

$$\therefore \lambda = \frac{hc}{(E_1 - E_2)}$$

$$= \frac{(6.626 \times 10^{-34} \text{ J.s})(3 \times 10^8 \text{ m/s})}{(5.36 - 3.45)(1.602 \times 10^{-19} \text{ J})} = 6496 \text{ Å}$$

19.20 ELECTRON SHELLS

The simple model of hydrogen was extended to other elements by Bohr and Stoner. Each atom is composed of a positively charged nucleus and a number of electrons revolving in allowed orbits around it. The number of electrons in an atom are determined by the atomic number, Z. The orbits to which electrons are confined are the orbits characterized by different values of the quantum number n . The orbits are called **electron shells**. These are named as K, L, M, N, O, etc.

When an atom is built, electrons are added one after the other filling one shell and then another starting from the shell nearest to the nucleus. A shell is filled when $2n^2$ electrons occupy it. Thus, K shell is filled when it has 2 electrons, the L shell is filled when 8 electrons occupy it and so on.

Shell	Quantum number, n	Capacity
K	1	2
L	2	8
M	3	18
N	4	32
O	5	50

There are certain departures from the above order in case of heavier elements. The reasons are now well understood and are indicative of the chemical behaviour of the heavy elements. The theoretical justification for the electron shells was provided by Wolfgang Pauli in 1925. The existence of characteristic line spectrum in the x-ray radiation emitted by elements lent a strong support to Bohr's scheme of organization of electrons into shells in an atom.

19.21 CHARACTERISTIC X-RAY SPECTRUM

X-ray radiation emitted by a target consists of a continuous spectrum superimposed with high intensity peaks, as shown in Fig. 19.10. The peaks occur at wavelengths characteristic of the target material. They occur only at specific wavelengths and hence the spectrum is a line spectrum. The following features are observed regarding the line spectrum.

1. The X-ray line spectra of all elements are strikingly similar.
2. The line spectrum given off by an element remains the same irrespective of the compounds that the element forms.
3. As one proceeds from the elements of low atomic weight to elements of higher atomic weight, the line spectrum shifts progressively towards the shorter wavelengths.

The origin of X-ray line spectra could not find explanation on the basis of classical theory and is accounted only by the quantum theory.

19.21.1 Origin of Characteristic X-Rays

The origin of X-rays is explained on the basis of Bohr's theory. According to Bohr model, electrons in an atom are organized into K, L, M, N,...shells. An atom radiates energy when an electron jumps from one allowed orbit to another allowed orbit. Normally, electron transitions cannot take place from the outermost orbits to the innermost orbits because the inner orbits are occupied. However, when a high energy electron knocks off an inner core electron belonging to K, L, M, N,...shells, an electron from an outer orbit jumps to fill up the vacated inner orbit. Such a transition causes emission of high-energy photon, namely X-ray photon. For example, the binding energy of an electron in K-shell of sodium atom is 1041 eV. The binding energy of a 2s electron in L shell is 63 eV. When the 2s electron jumps into 1s level, an X-ray photon of energy 978 eV is emitted.

The characteristic spectrum consists of a series of discrete lines. The group of lines having the shortest wavelength is called **K-series**. These lines occur when an electron from the K-shell is knocked away and the resulting vacancy is filled by an electron from the next

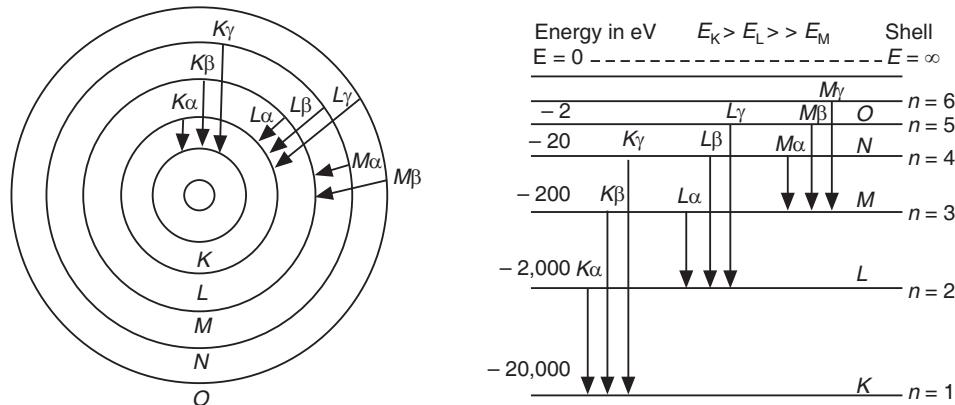


Fig.19.25

higher shells L, M, N, ...etc (Fig. 19.25). If an electron from L-shell jumps into K-shell, the energy difference is emitted in the form of K_{α} radiation; the transition of an electron from M-shell to K-shell produces K_{β} radiation. In a similar way the **L-series** consisting of $L_{\alpha}, L_{\beta}, \dots$ lines is produced when vacancies in L-shell are filled by electrons jumping from next higher shells. K-series form the most energetic and hard X-rays whereas L- and M-series form soft X-rays. We can summarize some more features of characteristic spectrum (in addition to those described in §19.21) as follows:

- (i) Line spectrum is produced when electrons are dislodged from the innermost orbits of the atoms of the target material followed by electron jumps from the outer orbits.
- (ii) It consists of discrete spectral lines which constitute K-series, L-series and M-series. K series consists of those lines for which electron jumps end at the K-level. Similarly, L-series and M-series are produced when electron jumps end at L- and M-levels respectively.
- (iii) K-series lines are the most energetic and constitute the **hard X-rays** whereas L- and M-series form the **soft X-rays**.
- (iv) There is a regular shift towards shorter wavelength in the K-spectrum as the atomic number of the target is increased. The exact relationship as found by Moseley is

$$\frac{\lambda_2}{\lambda_1} = \frac{v_1}{v_2} = \frac{(Z_1 - 1)^2}{(Z_2 - 1)^2}$$

where v_1 is the frequency of the K_{α} line for a target material having an atomic number of Z_1 and v_2 is the frequency of the K_{α} line for the target material having an atomic number of Z_2 .

19.22 MOSELEY'S LAW

During the years 1913-1914, Moseley made a systematic study of characteristic X-ray spectra of various elements. He found that different elements produce similar spectra consisting of K, L, and M series. He measured the wavelengths of the characteristic X-rays of a large number of elements using Bragg's X-ray spectrometer. He found a striking similarity in the spectral lines of each series. The spectrum emitted by each element was found to be exactly similar to that of the other, the only difference being that the lines had different wavelengths. The frequency of any particular line in a series varied regularly from one element to the next one in the periodic table. On plotting a graph between the square root of the frequency of any particular line in a series and the atomic number of the element, Moseley found that it gives a straight line, as seen in Fig.19.26. Such a graph is known as a **Moseley's plot**. Each series is governed by the following mathematical relationship:

$$v = a^2 (Z - \sigma)^2 \quad (19.54)$$

where Z is the atomic number of the target material, and a and σ are constants. The constant σ is known as **screening constant**. The above equation is known as **Moseley's law** and is usually written in the form

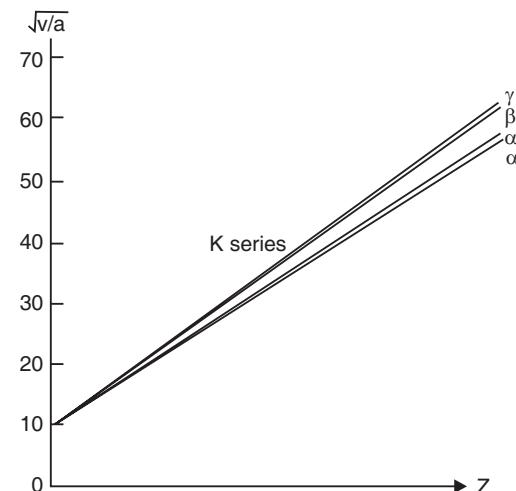


Fig.19.26

$$\sqrt{v} = a(Z - \sigma) \quad (19.55)$$

Moseley's law states that *the frequency of a spectral line in the characteristic X-ray spectrum varies directly as the square of the atomic number of the element emitting it.*

19.22.1 Derivation of Moseley's Law

Moseley's law can be derived from the Bohr's theory. Let us consider an atom of fairly large atomic number Z . For an electron in an inner orbit nearer to the nucleus, the Coulomb attraction due to positive nucleus will be dominant. The energy of the electron is given by

$$E_n = \frac{-m(Ze)^2(e)^2}{32\pi^2\varepsilon_0^2\hbar^2} \left[\frac{1}{n^2} \right] \quad (19.56)$$

where (Ze) is the nuclear charge. Accordingly, the energies of electron in an orbit of principal quantum number n_1 and that of n_2 can be written respectively as

$$E_{n_1} = \frac{-m(Ze)^2(e)^2}{32\pi^2\varepsilon_0^2\hbar^2} \left[\frac{1}{n_1^2} \right] \quad (19.57)$$

and

$$E_{n_2} = \frac{-m(Ze)^2(e)^2}{32\pi^2\varepsilon_0^2\hbar^2} \left[\frac{1}{n_2^2} \right] \quad (19.58)$$

Taking into account of the screening effect due to electrons, the positive charge of the nucleus seen by the electron under consideration reduces from Z to $(Z - \sigma)$.

$$\therefore E_{n_1} = \frac{-m(Z - \sigma_1)^2(e)^4}{32\pi^2\varepsilon_0^2\hbar^2} \left[\frac{1}{n_1^2} \right]$$

and

$$E_{n_2} = \frac{-m(Z - \sigma_2)^2(e)^4}{32\pi^2\varepsilon_0^2\hbar^2} \left[\frac{1}{n_2^2} \right]$$

The difference in energy is given by $\Delta E = E_{n_2} - E_{n_1}$.

$$\Delta E = \frac{me^4}{32\pi^2\varepsilon_0^2\hbar^2} \left[\frac{(Z - \sigma_1)^2}{n_1^2} - \frac{(Z - \sigma_2)^2}{n_2^2} \right] \quad (19.59)$$

Now let us assume that $\sigma_1 \approx \sigma_2 \approx \sigma$. Then

$$\Delta E = \frac{me^4}{32\pi^2\varepsilon_0^2\hbar^2} (Z - \sigma)^2 \left[\frac{1}{n_1^2} - \frac{1}{n_2^2} \right]$$

The frequency of the X-ray photon emitted in the transition is given by

$$v = \frac{\Delta E}{h} = \frac{me^4}{8\varepsilon_0^2 h^3} (Z - \sigma)^2 \left[\frac{1}{n_1^2} - \frac{1}{n_2^2} \right] \quad (19.60)$$

The K-series is due to transitions from higher states to the $n = 1$ state. Therefore, $n_1 = 1$. Then,

$$v = \frac{me^4}{8\varepsilon_0^2 h^3} (Z - \sigma)^2 \left[1 - \frac{1}{n_2^2} \right] \quad (19.61)$$

or

$$v = a^2 (Z - \sigma)^2 \quad (19.62)$$

where $a^2 = \frac{me^4}{8\varepsilon_0^2 h^3} \left[1 - \frac{1}{n_2^2} \right]$ is a constant. Thus, Moseley's law is derived from Bohr's theory.

19.22.2 Importance of Moseley's Law

Moseley's work gave a precise method for determination of the atomic numbers of the elements. Basing on his work, Moseley concluded that the determining factor in the arrangement of elements in the periodic table should be the atomic number and not the atomic weight. Earlier, Mendeleev arranged the elements in increasing atomic weight to form the periodic table. Moseley proposed that *the elements must be arranged in the periodic table according to their atomic numbers and not according to their atomic weights*. Moseley found that if he arranged the wavelengths of K_{α} lines in the order of atomic weights, these wavelengths formed a remarkable regular progression. But, the sequence was not perfect, since some gaps were found and some of the wavelengths were out of order. Moseley attributed the gaps in the series to undiscovered elements. The atomic number is the number of positive units of electric charge on the nucleus of the atom and it is the fundamental quantity, which increases by regular steps as we pass from one element to the next. Following his conclusions, Moseley corrected the positions of the elements argon, cobalt and tellurium in the periodic table. According to the arrangement with ascending order of atomic weights, the element cobalt ($Z = 27$) with atomic weight 58.94 should be placed after the element nickel ($Z = 28$) having atomic weight 58.69. Similarly, the element tellurium ($Z = 52$) with atomic weight 127.61 should be placed after the element iodine ($Z = 53$) having atomic weight 126.92. So is the case with the elements argon and potassium. Argon ($Z = 18$) with atomic weight 40 should come after element potassium ($Z = 19$) having atomic weight 39. However, Mendeleev observed that the above elements are to be placed in reverse order to preserve the periodicity of chemical and physical properties. The discrepancy is removed if the elements are arranged as per the ascending order of atomic numbers. Thus, Moseley had shown that the atomic number is much more important than the atomic weight in determining the properties of the elements. Moseley pointed out gaps in the group of rare earth elements at values $Z = 43, 61, 72$, and 75 . These elements technetium (43), promethium (61), hafnium (72) and rhenium (75) were subsequently discovered.

Example 19.15. Cobalt ($Z = 27$) gives strong K_{α} line of wavelength $\lambda = 1.785 \text{ \AA}$ and weak lines of 2.285 \AA and 1.537 \AA due to impurities. Assume screening constant $\sigma = 1$ for K -series and find the atomic number Z of each of the two impurities.

Solution:

$$\lambda_{K_{\alpha}} = \frac{1}{R} \cdot \frac{1}{(Z-\sigma)^2} \left[\frac{1}{\left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right)} \right] = \frac{1}{R} \cdot \frac{1}{(Z-1)^2} \left[\frac{1}{\left(\frac{1}{1^2} - \frac{1}{2^2} \right)} \right]$$

$$\therefore (Z-1)^2 = \frac{1}{R(1.785)} \left(\frac{4}{3} \right) = \frac{4}{3R(1.785)} \quad (1)$$

Similarly for impurity elements, we write

$$(Z_1-1)^2 = \frac{1}{R(2.285)} \left(\frac{4}{3} \right) = \frac{4}{3R(2.285)} \quad (2)$$

$$\text{and } (Z_2-1)^2 = \frac{1}{R(1.537)} \left(\frac{4}{3} \right) = \frac{4}{3R(1.537)} \quad (3)$$

Dividing (2) by (1), we obtain

$$\frac{(Z_1 - 1)^2}{(Z - 1)^2} = \frac{1.785}{2.285}$$

or $\frac{(Z_1 - 1)^2}{(27 - 1)^2} = \frac{1.785}{2.285}$

or $(Z_1 - 1)^2 = 528.08$
 $\therefore Z_1 = 23.9 \approx 24$

Similarly dividing (3) by (1), we obtain

$$\frac{(Z_2 - 1)^2}{(Z - 1)^2} = \frac{1.785}{1.537}$$

or $\frac{(Z_2 - 1)^2}{(27 - 1)^2} = \frac{1.785}{1.537}$

or $(Z_2 - 1)^2 = 785.08$
 $\therefore Z_2 = 28.02 \approx 28$

19.23 THE SOMMERFELD RELATIVISTIC ATOM MODEL

The spectral lines emitted by hydrogen atom appeared to be single lines with spectrometers of low resolving power. However, spectrometers of high resolving power showed that these lines consist of groups of separate and distinct lines located very close to each other. It is known as **fine structure** of spectral lines. The origin of more than one spectral line can be explained only when several orbits of slightly different energies exist corresponding to a given quantum number n . But according to Bohr's theory, only one electron orbit is possible for each value of n . Obviously, Bohr's theory could not explain the fine structure of spectral lines.

In 1915, A. Sommerfeld, the German physicist extended Bohr's theory by incorporating the idea of elliptical electronic orbits and taking into consideration the relativistic variation of electron mass. The atom proposed by Sommerfeld is therefore called the Sommerfeld relativistic atom model.

19.23.1 Elliptical Orbits

In general, periodic motion under the influence of a central force leads to elliptical orbits. Therefore, it is postulated that in atom an electron moves in an elliptical orbit with the nucleus located at a focus. In the elliptical orbit the position of the electron at any instant is defined by two coordinates, namely the radius vector r and the azimuthal angle ϕ (Fig. 19.27). The electron momentum may be resolved into two components p_r along the radius vector and p_ϕ at right angles to it. The momentum component p_ϕ at right angles to the radius vector represents the angular momentum. Sommerfeld assumed that the generalized condition, postulated by Bohr in case of circular orbits, holds good for elliptical orbits also.

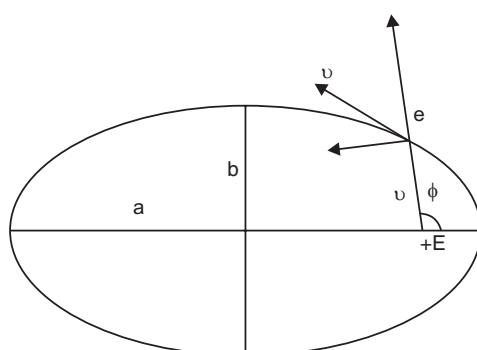


Fig. 19.27

Thus,

$$p_r dr = n_r h \quad (19.63)$$

and

$$p_\phi d\phi = n_\phi h \quad (19.64)$$

where n_r is called the **radial quantum number** and n_ϕ is the **azimuthal quantum number**. Further,

$$n_r + n_\phi = n$$

where n is the **principal quantum number**.

19.23.2 Allowed Elliptical Orbits

It is known from classical dynamics that the angular momentum p_ϕ remains constant during the motion. Therefore, from equ.(19.64), we have

$$\begin{aligned} \int_0^{2\pi} p_\phi d\phi &= p_\phi 2\pi = n_\phi h \\ \therefore p_\phi &= \frac{n_\phi h}{2\pi} \end{aligned} \quad (19.65)$$

In an elliptical orbit, the radius vector is a variable quantity. The momentum may be expressed as

$$\begin{aligned} p_r &= m \left(\frac{dr}{dt} \right) = m \left(\frac{dr}{d\phi} \right) \left(\frac{d\phi}{dt} \right) \\ &= m \left(\frac{dr}{d\phi} \right) \left(\frac{p_\phi}{m r^2} \right) = \left(\frac{p_\phi}{r^2} \right) \left(\frac{dr}{d\phi} \right) \\ \therefore \oint p_r dr &= \oint \left(\frac{p_\phi}{r^2} \right) \left(\frac{dr}{d\phi} \right) dr = n_r h \\ \text{or } p_\phi \oint \left(\frac{1}{r} \frac{dr}{d\phi} \right)^2 d\phi &= n_r h \end{aligned} \quad (19.66)$$

In an ellipse

$$\frac{1}{r} = \frac{1+e \cos\theta}{a(1-e)^2} \quad (19.67)$$

where e is the eccentricity of the ellipse and a is the semi-major axis. On differentiation of equ. (19.67), we get

$$\frac{1}{r^2} \frac{dr}{d\phi} = \frac{e \sin\phi}{a(1-e)^2} \quad (19.68)$$

Dividing equ.(19.68) by equ.(19.67), we obtain

$$\frac{1}{r} \frac{dr}{d\phi} = \frac{e \sin\phi}{1+e \cos\phi} \quad (19.69)$$

Using equ.(19.69) into equ.(19.66), one obtains

$$p_\phi \oint \left[\frac{e \sin\phi}{1+e \cos\phi} \right]^2 d\phi = n_r h$$

or

$$p_\phi \int_0^{2\pi} \frac{e^2 \sin^2\phi}{(1+e \cos\phi)^2} d\phi = n_r h \quad (19.70)$$

The integration of the above expression yields

$$\frac{2\pi p_\phi}{\sqrt{1-e^2}} - 2\pi p_\phi = n_r h$$

Putting $p_\phi = \frac{n_\phi h}{2\pi}$ into the above equation and simplifying it, we get

$$\begin{aligned} \sqrt{(1-e^2)} &= \frac{n_\phi}{n_r + n_\phi} = \frac{n_\phi}{n} \\ \therefore (1-e^2) &= \frac{n_\phi^2}{n^2} \end{aligned} \quad (19.71)$$

For an ellipse we have the relation

$$(1-e^2) = \frac{b^2}{a^2} \quad (19.72)$$

where a and b are the semi-major and semi-minor axes of the ellipse respectively.

Comparing the equations (19.71) and (19.72), we get

$$\frac{n_\phi}{n} = \frac{b}{a} \text{ Condition for allowed elliptical orbits} \quad (19.73)$$

Thus, the allowed elliptical electron orbits are those for which the ratio of minor to major axes is in the ratio of n_ϕ/n . A series of elliptical orbits of different eccentricities are thus allowed for values of $n_\phi = (n-1), (n-2), (n-3), \dots, 1$. The orbit is circular for $n_\phi = n$. However, n_ϕ cannot assume the value of zero since it would imply that the electron oscillates back and forth in a straight line penetrating through the centre of nucleus, which is most unlikely. Hence, n_ϕ can have in total n possible values from n to 1.

The net result of Sommerfeld's extension is that one elliptical orbit is added to Bohr's second orbit; two elliptical orbits are added to Bohr's third orbit; three elliptical orbits are added to Bohr's fourth orbit and so on (See Fig. 19.28).

19.23.3 Energy

It is shown that the total energy of the electron in an elliptical orbit is given by

$$E_n = -\frac{m e^4 Z^2}{8\epsilon_o^2 h^2} \left(\frac{1}{n^2} \right) \quad (19.74)$$

This energy is exactly identical to that of an electron in a Bohr' circular orbit. The introduction of elliptical orbits has increased the number of orbits but their energies being

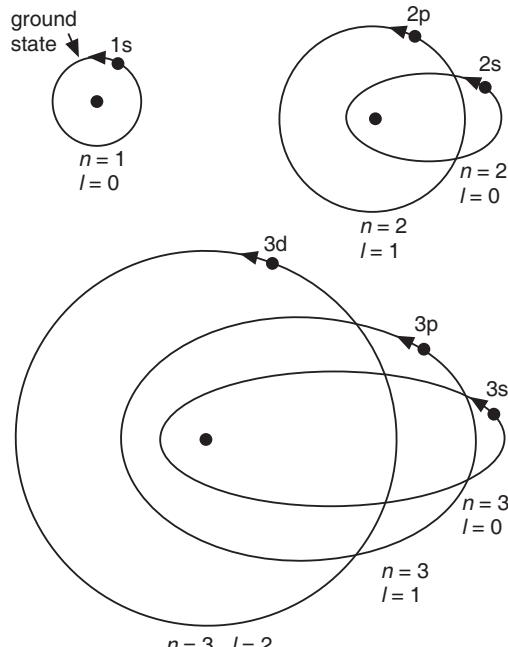


Fig. 19.28

identical, the number of atomic energy levels has not increased. Therefore, additional electron transitions and spectral lines are not possible. It means that the concept of elliptical orbits could not explain the fine structure of spectral lines.

19.23.4 Relativistic Correction

Owing to the motion of an electron in an elliptical orbit, its velocity would not be uniform. When the electron comes close to the nucleus, its velocity must be much higher than its velocity in a circular orbit of the same energy. Similarly, electron velocity would be lower when it is farthest from the nucleus. Such variations in velocity cause variation of the electron mass according to the theory of relativity. Sommerfeld showed that this assumption of the relativistic variation of electron mass transforms the simple elliptical orbit into a complicated curve known as a **rosette**, as depicted in Fig. 19.29. As the electron describes the elliptical path, the major axis of the ellipse revolves slowly in the plane of the ellipse around the nucleus located at the focus of the ellipse.

After relativistic correction, the energy of the electron in a particular state designated by quantum numbers n and n_k is obtained as

$$E_n = -\frac{mZ^2e^4}{8\varepsilon_0^2n^2h^2} \left[1 + \frac{\alpha^2Z^2}{n} \left(\frac{1}{n_\phi} - \frac{3}{4n} \right) \right] \quad (19.75)$$

where α is called the **fine structure constant**.

The electron energy in an orbit depends not only on the principal quantum number n but also on the azimuthal quantum number n_ϕ . Consequently, transitions among different levels corresponding to the various possible n_ϕ values can occur. As the energy differences of these levels are very small, the wavelengths of the resulting spectral lines would be very close. Therefore, a group of closely associated lines is produced instead of a single line.

At first sight it would seem that the number of fine structure lines should be very large in view of the many possible values of n_ϕ . To restrict the possible number of allowed transitions, Sommerfeld proposed the following *selection rule*:

$$\Delta n_k = \pm 1.$$

Because of this **selection rule** the number of allowed electronic transitions and hence the actual spectral lines are limited in number.

19.23.5 Limitations of Sommerfeld Theory

- (i) Sommerfeld theory gave the total number of possible azimuthal quantum numbers correctly, but the actual values are incorrect. Spectroscopic observations showed that the azimuthal quantum number can take the value of zero so that the total values are 0, 1, 2, 3, ..., $(n - 1)$. Therefore, a new azimuthal quantum number l is introduced where $l = (n_\phi - 1)$.
- (ii) Sommerfeld theory provided a theoretical justification to the existence of fine structure of spectral lines but it failed in predicting the correct number of fine structure lines.
- (iii) The theory does not give information about the relative intensities of the lines.
- (iv) Sommerfeld theory does not explain the distribution and arrangement of electrons in multielectron atoms.

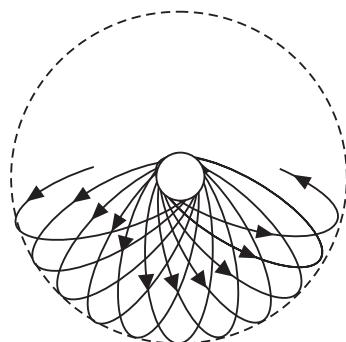


Fig. 19.29

19.24 THE VECTOR ATOM MODEL

The vector atom model is an extension of Bohr-Sommerfeld atom model. It was intended to overcome the limitations of the Bohr-Sommerfeld model while it also attempted to explain the new body of experimental observations such as anomalous Zeeman effect, Paschen-Back effect, Stark effect etc. This model incorporated new concepts partly by analogy and partly by empirical methods. The two central features of the vector atom model are:

- (i) space quantization of the electron orbits, and
- (ii) the electron spin.

1. Space Quantization

In Bohr's model of atom, the electron is assumed to move in circular orbits. The orbits are quantized as regards their magnitude, that is, their size. Thus, the electron in Bohr model has only one degree of freedom. One quantum number, namely the principal quantum number, n is enough to describe the electron motion. According to Sommerfeld theory, the electron moves in elliptical orbits, which are quantized as regards their magnitude and shape. Thus, the electron possesses two degrees of freedom. Hence, two quantum numbers, namely the principal quantum number, n and the azimuthal quantum number, n_k are required to explain the motion of the electron in an atom. Description of the electron motion in terms of these two quantum numbers implies that the electron motion is confined to a single plane.

As an atom is a three-dimensional entity, the orbital plane of an electron may take different orientations in the atom. Classically speaking, the electron orbit may orient in all probable directions in space (Fig. 19.30). Sommerfeld extended the concept of quantization to the orientation of the electron orbits in space. According to Sommerfeld, out of the possible infinite orientations, only certain discrete orientations are allowed for the electron orbits.

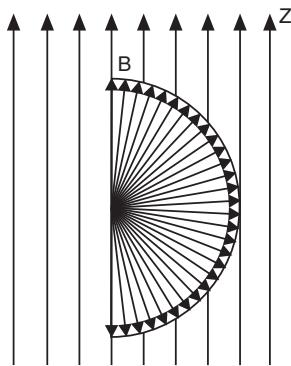


Fig. 19.30

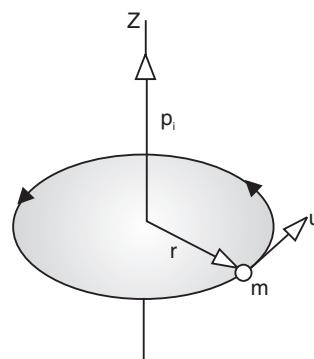


Fig. 19.31

Orientations of the orbit

The orientation of an orbit can be identified with the help of **orbital angular momentum vector**, p_l , which is directed along the axis of rotation of the electron and is perpendicular to the plane of the orbit (see Fig. 19.31). We require a reference direction, with respect to which the orientation of the vector p_l could be defined. Such a reference direction may be obtained by *imagining* the atom to have been placed in an external magnetic field, \mathbf{B} , whose strength tends to zero so that it may not disturb the electron orbits. Then, the magnetic field direction serves as the reference direction. Generally, magnetic field acting along z-direction is taken

as the **reference direction**. According to Sommerfeld, the vector p_l can orient itself only in certain discrete directions relative to the external magnetic field direction. This is known as **quantization of direction or space quantization**. Thus, in the vector atom model, the orbits are quantized in magnitude, shape and direction. The space quantization of an electron orbit is specified by the projection of its orbital angular momentum on to the reference direction; such projections being themselves quantized (see Fig.19.32). In vector atom model, the orbital angular momentum of the electron is given by

$$p_l = \frac{l\hbar}{2\pi} = l\hbar \quad (19.76)$$

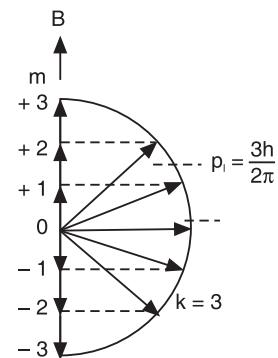


Fig.19.32

Allowed Orientations

According to the concept of space quantization, p_l can take only such orientations in space for which its component in the field direction \mathbf{B} will take integral values of \hbar (see Fig.19.32). Thus, $p_l \cos \theta$ is quantized and is given by

$$p_l \cos \theta = m_l \hbar$$

where m_l is called the **orbital magnetic quantum number** and θ is the angle between p_l and the field direction in space.

The largest component of p_l along the field direction is thus $l\hbar$ when $m_l = 1$. This value is less than the magnitude of p_l which means that the vector p_l cannot align in the direction of magnetic field. The angle between p_l and the z-axis is given by

$$\cos \theta = \frac{m_l \hbar}{l \hbar} = \frac{m_l}{l} \quad (19.77)$$

Since m_l has to be an integer and $\cos \theta$ can never exceed unity, the permitted values of m_l range from $+l$ to $-l$ at unit intervals. Thus, m_l takes the following values

$$l, (l-1), (l-2), \dots, 1, 0, -1, -2, \dots, -(l-2), -(l-1), -l.$$

It means that corresponding to each value of l , there will be $(2l+1)$ possible values of m_l . Hence, the angular momentum vector p_l can take $(2l+1)$ discrete orientations with respect to the reference field.

According to Bohr-Sommerfeld theory, in the absence of the external field, the energy of the electron in its orbit is the same for all orientations of p_l in space. However, when the reference field is applied, the energy of the electron will depend upon the relative orientation of p_l with respect to the field direction. Now the electron motion is visualized as three-dimensional motion. Space quantization, therefore, requires three different quantum numbers n , l and m_l . The energy of the electron now depends on all the three quantum numbers in an applied field.

Orbital magnetic moment

In fact, the discrete orientations of the electron orbit with respect to the reference magnetic field direction occur due to the interaction of the orbital magnetic moment of the electron and the reference magnetic field. Since an electron has a negative charge, its orbital motion is equivalent to a tiny current loop, which produces a magnetic field perpendicular to the plane of the orbit. The direction of the magnetic field is given by the left-hand rule and is pointed opposite to the direction of vector p_l . The magnetic field set up by the orbital electron is quite

similar to the field around a bar magnet and is characterized by a **magnetic dipole moment**. The magnetic moment of a current loop is given by

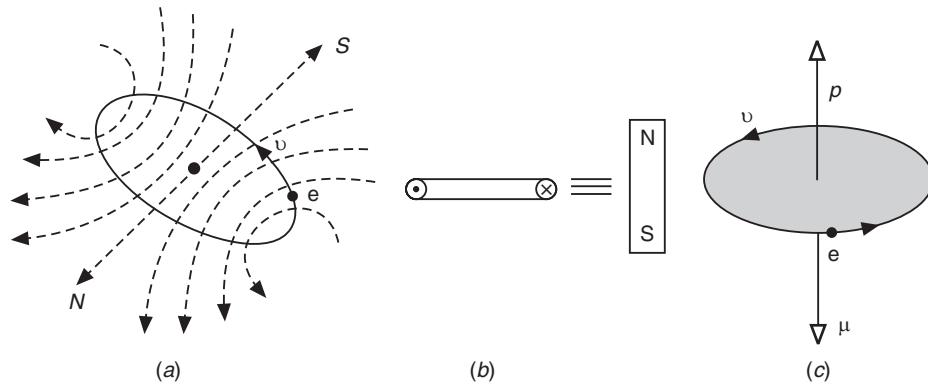


Fig. 19.33

$$\mu = IA$$

The magnetic moment of the electron orbit can be expressed as

$$\mu_l = IA = \frac{e}{T} \pi r^2$$

where r is the radius of the electron orbit and T is the time period for electron revolution.

But

$$\frac{2\pi r}{T} = v$$

Therefore,

$$\mu_l = e\pi r^2 \left(\frac{v}{2\pi r} \right) = \frac{evr}{2}$$

Using $p_l = mvr$ into the above equation, we get

$$\mu_l = \left(\frac{e}{2m} \right) p_l = \left(\frac{e}{2m} \right) l\hbar \quad (19.78)$$

All the symbols on the right side of the above equation are atomic constants. Hence, we write

$$\mu_l = \mu_B l$$

Bohr Magneton

$$\mu_B = \left(\frac{e\hbar}{2m} \right) \quad (19.79)$$

is called the Bohr magneton. **Bohr magneton** is the basic unit of atomic magnetic moments. The numerical value of Bohr magneton is given by

$$\mu_B = 9.273 \times 10^{-24} Am^2 \quad (19.80)$$

Gyromagnetic Ratio

The ratio

$$G = \frac{\mu_l}{p_l} = \frac{e}{2m} \quad (19.81)$$

is called the *gyromagnetic ratio*. It is written as

$$G = g \cdot \frac{e}{2m} \quad (19.82)$$

where g is known as the **Lande splitting factor**. $g = 1$ for orbital motion of electron.

2. Electron Spin

In 1925, Uhlenbeck and Goudsmith put forward the hypothesis of the spinning electron, in order to explain some of the spectral phenomena such as fine structure, Zeeman effect etc. According to it, the electron revolves about its own axis while orbiting round the nucleus. The concept is very much akin to the spinning of the planets in the solar system. The spinning motion, like the orbital motion, is quantized not only in magnitude but also in direction according to the rules of space quantization.

The intrinsic **spin angular momentum** is given by

$$p_s = s\hbar \quad (19.83)$$

where $s = \frac{1}{2}$ is known as the *spin quantum number*. The spin angular momentum is a vector.

As both the orbital angular momentum \mathbf{p}_l and the intrinsic spin angular momentum \mathbf{p}_s that determine the state of the electron are quantized vectors and are added as per the rules of vectors, the atom model is called **vector atom model**.

Allowed Orientations

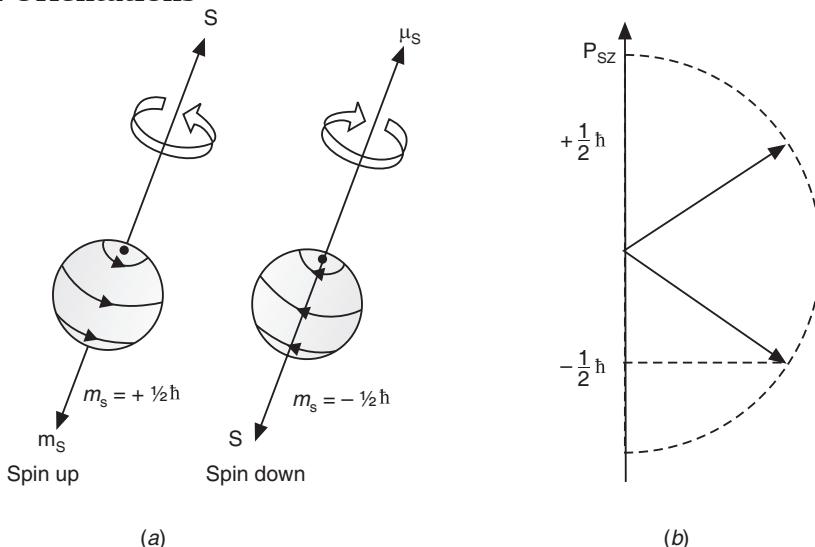


Fig. 19.34

Classically speaking, a body can spin clockwise or anti-clockwise. Correspondingly, the spin angular momentum can take only two orientations with respect to the reference magnetic field direction, as illustrated in Fig. 19.34. In the presence of magnetic field, the projection of p_s on to the reference direction is given by

$$p_{sz} = m_s \hbar \quad (19.84)$$

By analogy with the orbital angular momentum vector, the spin angular momentum vector can take $(2s + 1)$ orientations in space relative to the reference magnetic field direction. As $s = \frac{1}{2}$, p_s can take only two orientations. It implies that the spin angular momentum p_s can align either parallel or antiparallel to the applied field direction. The spin magnetic quantum number m_s can have only two values, namely $+\frac{1}{2}$ or $-\frac{1}{2}$. They correspond to "spin-up" (parallel) and "spin-down" (antiparallel) orientations.

Intrinsic magnetic moment

The spinning of an electron around its own axis generates a magnetic field. This field is similar to the magnetic field around a bar magnet and is characterized by a magnetic moment. The magnetic moment μ_s of a spinning electron is given by

$$\mu_s = 2 \frac{e}{2m} p_s$$

The Lande splitting factor $g = 2$ in case of spin motion.

$$\mu_s = 2 \frac{e}{2m} \cdot \frac{\hbar}{2} = \frac{e\hbar}{2m} \quad (19.85)$$

or

$$\mu_s = 1\mu_B$$

Resultant angular momentum

Since an electron in an atom has two different angular momenta p_l and p_s . The resultant angular momentum is obtained by the vector addition of p_l and p_s . Thus,

$$p_j = p_l + p_s \quad (19.86)$$

where $p_l = j\hbar$. j is called the **total angular momentum quantum number**. Thus,

$$p_j = \frac{j\hbar}{2\pi} = j\hbar \quad (19.87)$$

For each value of l , there are two possibilities.

$$j\hbar = l\hbar + s\hbar \quad (19.88a)$$

and

$$j\hbar = l\hbar - s\hbar \quad (19.88b)$$

As all the vectors have the common factor \hbar , the quantum numbers themselves are used as vectors. Thus, equ.(19.88) may be written in simple terms as

$$j = l + s \text{ and } j = l - s \quad (19.89)$$

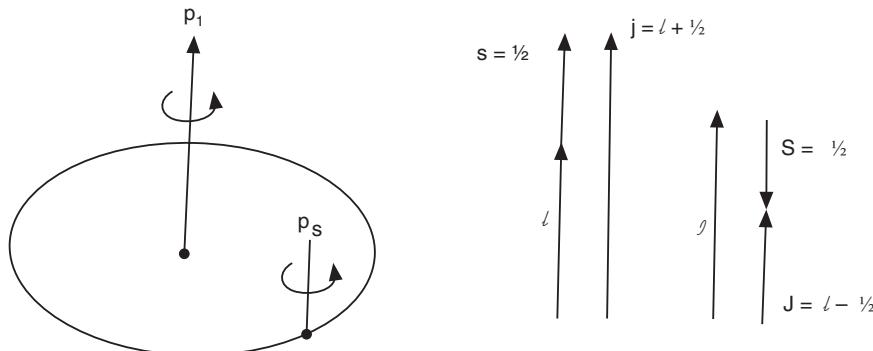


Fig. 19.35

For a single electron, $s = \frac{1}{2}$ and hence j can take one of the two values, namely $j = l + \frac{1}{2}$ or $j = l - \frac{1}{2}$. For example, if $l = 0$, j can have only one value, i.e., $j = \frac{1}{2}$. If $l = 1$, then j can be either $3/2$ or $1/2$.

19.24.1 Quantum Numbers Associated With the Vector Atom Model

Let us now sum up the various quantum numbers used in the vector atom model at one place. In this model, a quantum number is assigned to each of the component parts. The numerical value of the quantum number may be thought of as the length of the vector. The additions of the components are carried out as per the rules of vector analysis. A straight line, whose direction is parallel to the axis of rotation and whose length is proportional to the magnitude of the momentum, represents angular momentum. The quantum numbers associated with an electron in a given atom (in the absence of reference field) are as follows:

- 1. Principal Quantum Number, n :** An electron is characterized by a principal quantum number, n which describes the quantization of the electron energy. It can take only integral values, 1,2,3,..... n .
- 2. Orbital Quantum Number, l :** The shape of the orbit is characterized by l which may take any integral value between 0 and $(n - l)$. Thus, if $n = 4$, l can take the values 1,2,3. In vector atom model, the orbital angular momentum of the electron is given by

$$p_l = \frac{lh}{2\pi} = l\hbar.$$

According to quantum mechanical model, it is given by

$$p_l = \sqrt{l(l+1)} \hbar \quad (19.90)$$

It is a common practice to designate the states having different l values as follows.

Orbital quantum number	0	1	2	3	4	5
Designation of the state	s	p	d	f	g	h

- 3. Spin Quantum Number, s :** The spin quantum number has a magnitude $\frac{1}{2}$ always. In the vector atom model, the electron spin angular momentum is given by

$$p_s = \frac{sh}{2\pi} = s\hbar.$$

According to quantum mechanical model, it is given by

$$p_s = \sqrt{s(s+1)} \hbar \quad (19.91)$$

- 4. Total Angular Quantum Number, j :** This quantum number refers to the resultant angular momentum of the electron due to both of its orbital and spin motions. It is the numerical value of the vector sum of l and s . Thus, $j = l + s$ and the value of j is evidently half-integer, since $s = \frac{1}{2}$ always. It is usually expressed as $j = l \pm s$; the plus sign when s is parallel to l , minus sign when s is antiparallel to l .

In vector atom model, the total angular momentum of the electron p_j is given by

$$p_j = \frac{jh}{2\pi} = j\hbar.$$

According to quantum mechanical model, it is given by

$$p_j = \sqrt{j(j+1)} \hbar \quad (19.92)$$

When the atom is placed in a reference magnetic field, three more quantum numbers are associated with the electron due to space quantization.

- 5. Orbital magnetic quantum number, m_l :** m_l is the numerical value of projection of the orbital vector l in the reference field direction. It is an integer and takes a total of

$(2l + 1)$ values ranging from $-l$ to $+l$ including zero. According to space quantization, the projection of l in the field direction must itself be quantized. Hence l can orient only at such discrete angles that its projection m_l may also be an integer. For example, if the vector l is inclined to the field direction at an angle θ , the projection is $m_l = l \cos \theta$. Now since m_l has to be an integer and $\cos \theta$ can never exceed unity, the permitted values of m_l are from $-l$ to $+l$ including zero.

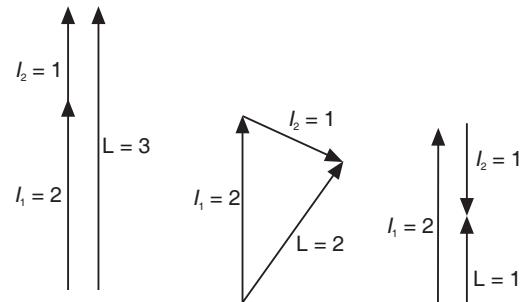
The total number of permitted orientations of the l vector relative to the field direction is also $(2l + 1)$. Thus, if $l = 1$, the permitted orientations of l are three for which m_l has the values $+1, 0, -1$.

6. **Spin magnetic quantum number, m_s :** m_s is the numerical value of the projection of the spin vector s on the reference field direction. By analogy with the orbital angular momentum vector l , the spin vector s can have only $(2l + 1)$ values from $-s$ to $+s$ at unit intervals. Therefore, the value zero is excluded and m_s can take only two values $+\frac{1}{2}$ and $-\frac{1}{2}$.
7. **Total magnetic quantum number, m_j :** m_j is the numerical value of the projection of the total angular momentum vector j on the reference field direction. In case of a single electron, j can have only half-integral value, i.e., $l \pm \frac{1}{2}$. Consequently, m_j has to take only half-integral values. The permitted orientations of j with respect to the field direction are $(2j + 1)$ and hence m_j can have only $(2j + 1)$ values from $-j$ to $+j$, excluding zero.

19.24.2 Multielectron Atom

When there is more than one electron in an atom, the l vectors for the different electrons must be added up vectorially to obtain the resultant orbital angular momentum of the atom.

$$\mathbf{L} = \sum_i \mathbf{l}_i = \mathbf{l}_1 + \mathbf{l}_2 + \mathbf{l}_3 + \dots \quad (19.93)$$



where $\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3, \dots$ are the orbital angular momenta of the individual electrons in the atom. From now onwards, we use the symbol \mathbf{L} in place of \mathbf{p}_L .

The value of the total orbital quantum number L of the atom must be an integer. For example, if $l_1 = 2$ and $l_2 = 1$, then L can have one of the values 3, 2, or 1 (see Fig. 19.36).

The resultant spin angular momentum is given by

$$\mathbf{S} = \sum_i \mathbf{s}_i = \mathbf{s}_1 + \mathbf{s}_2 + \mathbf{s}_3 + \dots \quad (19.94)$$

where $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \dots$ are the spins of the individual electrons in the atom. Each of these is equal to $\frac{1}{2}$.

The rule for the vector addition of the spins of the different electrons requires that these can be either parallel or antiparallel to one another. Therefore, the value of S can be either

zero or an integer, if the atom has an even number of electrons. It will be half-odd integer, if the atom has an odd number of electrons. These two cases are illustrated in Fig. 19.37.

19.24.3 Coupling Schemes

There are several ways in which the different vectors of the electrons may combine to give the vectors representing the atom as a whole. The method of combination depends on the interaction or “coupling” between the component vectors, since the orbital and spin motions of the electron produce magnetic fields and thereby result in mutual perturbation. It is usual to distinguish two types of coupling known as the *L-S coupling* and *j-j coupling*.

(i) L-S Coupling

It is also known as Russell-Saunders coupling. It is the most frequently occurring coupling and hence is known as the *normal coupling*. In this, all the spin vectors of the electrons combine to form a resultant vector \mathbf{S} and all the orbital angular momentum vectors of the electrons likewise combine to form a resultant vector \mathbf{L} . Then the resultant vectors \mathbf{S} and \mathbf{L} combine to produce the total angular momentum \mathbf{J} of the atom. The scheme may be represented as

$$\mathbf{L} = \sum_i l_i; \mathbf{S} = \sum_i s_i \text{ and } \mathbf{J} = \mathbf{L} + \mathbf{S} \quad (19.95)$$

This sort of coupling is the most natural one when the interaction between the individual spins on the one hand and the individual orbital momenta on the other hand are very strong.

The principles that govern this coupling are:

- (a) All the three vectors \mathbf{L} , \mathbf{S} , and \mathbf{J} are quantized.
- (b) \mathbf{L} is always an integer. It may take values $0, 1, 2, 3, \dots$ etc
- (c) \mathbf{S} is an integer or half-integer depending on the number of electrons involved and the direction of the spin vectors. Thus, for an atom containing a single electron \mathbf{S} can have only one value $\frac{1}{2}$; for a two electron system \mathbf{S} may be either 1 or 0 depending on whether the spin vectors are parallel or antiparallel and so on.

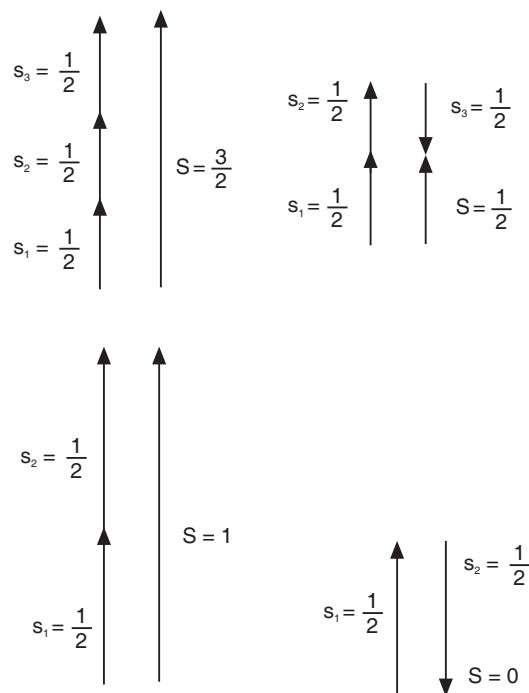


Fig. 19.37

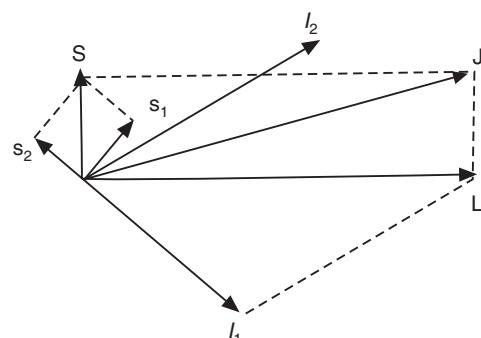


Fig. 19.38

(d) \mathbf{J} , the vector sum of \mathbf{L} and \mathbf{S} must be an integer (0, 1, 2, 3,..) for odd electron systems and it is an half-integer ($\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \dots$) for even electron systems. In general, the possible number of values which \mathbf{J} can assume are

$$(2S + 1) \text{ if } L \geq S$$

$$\text{and} \quad (2L + 1) \text{ if } L < S.$$

The quantity $(2S + 1)$ is known as the multiplicity of the \mathbf{L} state. It gives the permitted values of \mathbf{J} for a given value of \mathbf{L} . \mathbf{J} is always positive since it represents the total angular momentum of atom.

(ii) The j-j coupling

This coupling is found when the interaction between the spin and orbital angular momentum vectors of each electron is stronger than that between either the spin vectors or orbital angular momentum vectors of different electrons. In such a case, each electron is considered separately and its individual spin and orbital angular momenta are combined first using the relation, $j = l + s$. Then, the total angular momentum \mathbf{J} of the atom is obtained by taking the vector sum of all individual \mathbf{j} vectors of different electrons. This coupling scheme may be symbolically represented as

$$j_i = l_i + s_i \text{ and } \mathbf{J} = \sum_i \mathbf{j}_i \quad (19.96)$$

This type of coupling exists mainly in heavy atoms.

19.24.4 Pauli's Exclusion Principle

Pauli's exclusion principle is an empirical rule, which is used in conjunction with the vector atom model to explain the electronic structure in an atom. Among the several quantum numbers associated with the electron, four quantum numbers, namely n , l , m_l and m_s are needed in order to specify completely its state. According to exclusion principle, every completely defined quantum state in an atom can be occupied by only one electron. In other words, it is impossible for two electrons in an atom to be identical as regards all their quantum numbers. It is variously stated as "no two electrons in an atom can occupy the same quantum state" or "no two electrons in an atom can have the same four quantum numbers."

The exclusion principle explains certain experimental observations such as the occurrence or non-occurrence of spectra lines and the scheme of arrangement of electrons in atoms etc.

19.24.5 Selection Rules and Intensity Rules

It has been deduced from spectroscopic observations that electronic transitions from each energy state to all other states do not occur. In the vector model of atom, three selection rules have been devised and they are supplemented by the intensity rules in order to predict the intensity of the lines that occur.

Selection Rules

The rules that govern the electronic transitions are known as *selection rules*. They are devised one for L , another for J , and a third for S .

(a) *The selection rule for L:* Most of the observed spectral lines are due to transitions between states, in which a single electron jumps from one orbit to another. The rule that

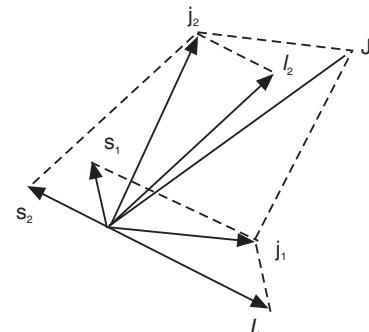


Fig.19.39

governs the transitions is that $\Delta L = \pm 1$. Only those transitions are observed for which the value of L changes by ± 1 .

- (b) *The selection rule for J:* Spectral lines are observed only when transitions take place between states for which $\Delta J = \pm 1$ or 0. the transition $0 \rightarrow 0$ is not permitted.
- (c) *The selection rule for S:* Transitions take place between states for which $\Delta S = 0$.

Intensity Rules

Whether a spectral line due to an allowed transition is weak or intense is determined by the intensity rules. They are as follows:

- (a) A spectral line will be *intense* for which electron transition takes place in such a way that L and J change in the same sense.
- (b) A spectral line will be *weak* for which electron transition L and J change in the opposite sense.
- (c) A spectral line will be *intense* for which electron transition L and J decrease, that is, $L \rightarrow (L-1)$ and $J \rightarrow (J-1)$.
- (d) A spectral line will be *less intense* for which electron transition L and J increase, that is, $L \rightarrow (L+1)$ and $J \rightarrow (J+1)$.
- (e) The oppositely directed transitions do not occur.

The above rules may be summarized as follows:

$\Delta L = -1$,	$\Delta J = -1$	Very intense line
$\Delta L = -1$,	$\Delta J = 0$	Less intense line
$\Delta L = +1$,	$\Delta J = +1$	Weak line
$\Delta L = +1$,	$\Delta J = 0$	Weak line
$\Delta L = -1$,	$\Delta J = +1$	No transition
$\Delta L = +1$,	$\Delta J = -1$	No transition

19.25 APPLICATIONS OF THE VECTOR ATOM MODEL

The vector atom model provided a rational explanation of several complex atomic and spectral phenomena. Out of them, the most important are

- (i) the electronic structure in atoms
- (ii) the Zeeman effect
- (iii) anomalous Zeeman effect
- (iv) the Stark effect

We now study these phenomena and understand how vector atom model succeeded in explaining them.

19.26 THE ELECTRONIC STRUCTURE OF ATOMS

An atom of a certain kind cannot be said to have a definite size but from a practical point of view a definite size is attributed to it. Thus, hydrogen atom is supposed to have a diameter of about 1\AA .

The electrons in an atom occupy energy or orbitals characterized by the quantum numbers n , l and m_l . A set of these numbers defines the **state** of the electron. The states (or orbitals) are organized into subshells and shells. The possible orbitals corresponding to a particular value of n are said to constitute a **shell**. All the electrons in an atom which have the same principal quantum number n belong to the electron shell. The electrons having $n = 1$ are said to form the K shell of the atom; electrons in the $n = 2$ state form the L shell and so on. The designation of the electron shells is as follows,

Principal quantum number, n	1	2	3	4	5	6
Designation of the electron shell	K	L	M	N	O	P

A shell characterized by the principal quantum number n accommodates a maximum of $2n^2$ electrons. The value of n increases as the distance of a shell from the nucleus increases. The number of electrons that a shell can accommodate increases in accordance with the value of n . Thus, a K shell with $n = 1$ accommodates 2 electrons, an L shell with $n = 2$ accommodates 8 electrons and so on. The maximum capacities of the shells are given below.

Electron shell	K $n = 1$	L $n = 2$	M $n = 3$	N $n = 4$	O $n = 5$	P $n = 6$
Maximum capacity, $2n^2$	2	8	18	32	50	72

Shells are built from subshells which accommodate electrons of the same value of the orbital quantum number l . Thus, electrons that share a certain value of l in a shell said to occupy the same subshell. The number of subshells in a shell is equal to the value of n . Thus a K shell has only one subshell corresponding to $l = 0$, an L shell has two subshells and so on. In atomic physics, states with a particular value of l have a particular name. A state with $l = 0$ is called an *s* state; a state with $l = 1$ is called a *p* state and so on. The designations are listed in the table below;

Orbital quantum number, l	0	1	2	3	4	5
Designation of the state or subshell	s	p	d	f	g	h

For a given set of values of n and l , there are $(2l + 1)$ possible values of m_l . It means that for a given value of n and l , there are $(2l + 1)$ orbitals or electron **states**. Thus, if $l = 0$, there is only one electron state; if $l = 1$, three electronic states etc. The electronic states correspond to the different orientations of the orbitals.

An energy state described by the three quantum numbers n , l and m_l can be occupied by only two electrons which have opposite spin directions. This is known as **Pauli's exclusion principle**. This principle indicates why electrons do not crowd into the lowest energy state. The electron spin is characterized by quantum number m_s . The state of the electron described by the four quantum numbers n , l , m_l and m_s is called the **quantum state**. Thus each energy state consists of two quantum states.

The terms energy level and energy state are used often in the same meaning. It may be noted that an energy level is not equivalent with an electronic state. An energy level is determined by the value of the principal quantum number n , and such a level corresponds to n^2 electronic states. Thus, for $n = 2$, there are $2^2 = 4$ different electronic states. The energy of an electron is mainly determined by the value of n , but to some extent by the quantum numbers l and m_l also.

The occupancy of various subshells and shells in a complex atom is governed by three basic rules;

1. **Minimum energy condition.** Electrons tend to occupy the lowest available energy state such that their total energy is a minimum. They go to higher energy states only when the lower energy states are not vacant.
2. **Pauli's exclusion principle.** An orbital in an atom described by the quantum numbers n , l and m_l can be occupied by only two electrons having opposite spin directions.
3. **Hund's rule.** The order of filling of the orbitals of subshells obeys Hund's rule. According to this rule the *total spin number of the electrons of a shell must be maximum*.

It means that the orbitals of a subshell are filled first with one electron each and then with the second electron respectively.

For instance in nitrogen atom, there are three electrons in $2p$ subshell. Three electrons occupy the three orbitals p_x , p_y and p_z instead of occupying only p_x and p_y orbitals.

The sequence of the energy states in order of increasing energy of the orbitals of a multi-electron atom is in the following order.

$$1s < 2s < 2p < 3s < 3p < 4s < 3d < 4p < 5s < 4d < 5p < 6s < 4f < 5d < 6p.$$

As each orbital characterized by the set of quantum numbers n , l , m_l can accommodate two electrons having m_s values $+\frac{1}{2}$ and $-\frac{1}{2}$, the electron capacity of a subshell is a maximum of $2(2l + 1)$ electrons.

Therefore, the maximum capacity N of a shell is given by,

$$\begin{aligned} N &= \sum_{l=0}^{n-1} 2(2l+1) = 2 [1 + 3 + 5 + \dots + 2(n-1) + 1] \\ &= 2[1 + 3 + 5 + \dots + 2(n-1) + 1] \\ &= 2 \times \frac{n}{2} [1 + 2(n-1)] \\ &= 2n^2 \end{aligned} \quad (19.97)$$

The following table shows at a glance the capacities of individual orbitals and the shells in total.

Table 1

n	shell	$l=0$	Electron capacity	$l=1$	Electron capacity	$l=2$	Electron capacity	$l=3$	Electron capacity	$l=4$	Electron capacity	$l=5$	Electron capacity	Maximum capacity of shell $2n^2$
1	K	1s	2	-	-	-	-	-	-	-	-	-	-	2
2	L	2s	2	2p	6	-	-	-	-	-	-	-	-	8
3	M	3s	2	3p	6	3d	10	-	-	-	-	-	-	18
4	N	4s	2	4p	6	4d	10	4f	14	-	-	-	-	32
5	O	5s	2	5p	6	5d	10	5f	14	5g	18	-	-	50
6	P	6s	2	6p	6	6d	10	6f	14	6g	18	6h	22	72

The electron configuration of an atom is described using nl^x notation, where x denotes the number of electrons occupying the subshell. For example, we write the electron configuration of the sodium atom. $1s^2, 2s^2, 2p^6, 3s^1$.

A shell or subshell that contains its full quota of electrons is said to be **closed** or **completely filled**. Thus, the K shell and L shell of sodium atom are closed while the M shell is partially filled. The 3s subshell of M shell contains only one electron and is half filled.

The completely filled shells and subshells of atoms are stable and are not readily disturbed. The electrons in the outer shell will be in a position to interact with similar electrons in adjacent atoms. The number of such electrons determines the **valency** of the atom and the chemical behaviour of the element.

19.27 ZEEMAN EFFECT

One of the basic concepts on which vector atom model rests is the concept of space quantization. Sommerfeld proposed the concept of space quantization and the Zeeman effect is considered as one of the experiments that confirmed the existence of space quantization.

The Zeeman effect is the name given to the splitting of the energy levels when a magnetic field acts on atoms. Splitting of energy levels leads to the splitting of the spectral lines into several components. In 1896, Zeeman found that if a source of light giving line spectrum is placed in a magnetic field, the spectral lines are split up into several components distributed symmetrically about the original lines. The splitting of spectral lines under the action of a magnetic field is known as **Zeeman effect**.

Zeeman observed that in the presence of a strong magnetic field of the order of 1 tesla (10^4 gauss), each spectral line is split into two components when viewed along the field direction and into three components when viewed in a direction perpendicular to the magnetic field direction. This is known as **normal Zeeman effect**.

Experimental Arrangement

The normal Zeeman effect can be readily observed. The experimental arrangement is shown in Fig.19.40. It consists of an electromagnet capable of producing strong magnetic field. A source of light such as a sodium flame or a mercury arc is placed between the pole pieces of the electromagnet. The pole pieces are conical in shape. The light coming from the source is examined by means of a spectroscope of high resolving power. In order to view the light in longitudinal direction (i.e., parallel to the field) a hole is drilled lengthwise in one of the pole pieces.

Initially, the magnetic field is not switched on and the spectroscope is focused on one of the spectral lines emitted by the source of light. When the magnetic field is switched on, the spectral line is split into three components with the original line. One of the lines is in the same position as the original line and the other two lines are located symmetrically about the original line. The splitting is illustrated in Fig.19.41. When the outer two lines are viewed through a Nicol prism, it was found that they are polarized at right angles to the undisplaced line. This effect is sometimes called *transverse Zeeman effect*.

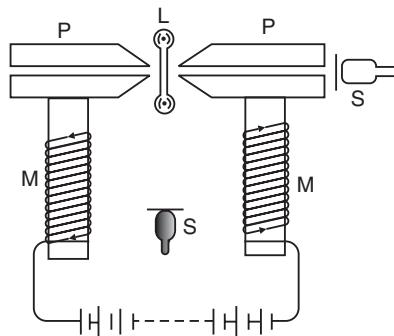


Fig.19.40

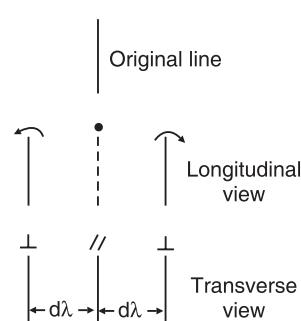


Fig.19.41

When the line is viewed through the hole in the pole piece, i.e., in a direction parallel to the magnetic field direction, the spectral line is found to split into two lines. There is no line in the position of the original line whereas the two outer lines are present. When the outer two lines are viewed through a Nicol prism, it was found that they are circularly polarized in opposite directions. This effect is known as *longitudinal Zeeman effect*.

19.27.1 Classical Explanation of Zeeman effect by Lorentz on the Basis of Electron Theory

A satisfactory explanation of the normal Zeeman effect was given by H.A. Lorentz on the basis of the electron theory.

Let us consider an electron moving in a circular orbit of radius r with a velocity v . The centripetal force acting on the electron is

$$F = \frac{mv^2}{r}. \quad (19.98)$$

When an external magnetic field is applied, an additional force acts on the electron. This force acts perpendicular to the direction of the magnetic field and along the radius of the orbit. If the force acts inwards along the radius, the velocity of the electron increases. On the other hand, if it acts outward, the velocity of the electron decreases. Let the force due to magnetic field be F_1 and let the velocity increase to v_1 due to this force. Then,

$$F_1 = Bev_1 \quad (19.99)$$

As this force is in a direction to add to the initial force, the total force on the electron now is

$$\begin{aligned} F_R &= F + F_1 \\ \text{or } \frac{mv_1^2}{r} &= \frac{mv^2}{r} + Bev_1 \end{aligned} \quad (19.100)$$

But

$$v = \omega r \text{ and } v_1 = \omega_1 r$$

where ω and ω_1 are the respective angular velocities.

From equ.(19.100), we get

$$\begin{aligned} \frac{mr^2\omega_1^2}{r} &= \frac{mr^2\omega^2}{r} + Ber\omega_1 \\ \text{or } \omega_1^2 - \omega^2 &= \frac{eB\omega_1}{m} \\ \text{or } (\omega_1 - \omega)(\omega_1 + \omega) &= \frac{eB\omega_1}{m} \\ \text{or } (\omega_1 - \omega) &= \frac{eB\omega_1}{m(\omega_1 + \omega)} \end{aligned} \quad (19.101)$$

$(\omega_1 + \omega)$ is approximately equal to $2\omega_1$.

$$\therefore (\omega_1 - \omega) = \frac{eB}{2m}$$

$$\therefore \omega_1 = \omega + \frac{eB}{2m} \quad (19.102)$$

If v_1 and v are the corresponding frequencies, then $\omega_1 = 2\pi v_1$ and $\omega = 2\pi v$. Then,

$$2\pi v_1 = 2\pi v + \frac{eB}{2m}$$

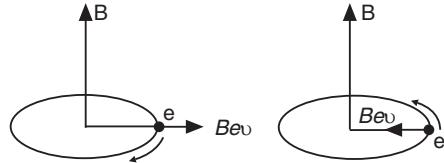


Fig. 19.42

or $v_1 = v + \frac{eB}{4\pi m}$ (19.103)

When the electron revolves in an opposite direction, the magnetic field produces a force in the opposite direction as a result of which its velocity decreases to v_2 . In that case, the total force on the electron is

$$\begin{aligned} F_R &= F - F_2 \\ \therefore \frac{mv_2^2}{r} &= \frac{mv^2}{r} - Bev_2 \\ \text{or } \frac{mr^2\omega_2^2}{r} &= \frac{mr^2\omega^2}{r} - Ber\omega_2 \\ \therefore \omega_2^2 - \omega^2 &= -\frac{eB\omega_2}{m} \\ \therefore (\omega_2 - \omega) &= -\frac{eB}{2m} \\ \therefore v_2 &= v - \frac{eB}{4\pi m} \end{aligned} \quad (19.104)$$

Subtracting equ.(19.104) from equ.(19.103), we obtain

$$\begin{aligned} v_1 - v_2 &= \frac{eB}{4\pi m} \\ \text{or } \Delta v &= \frac{eB}{4\pi m} \end{aligned} \quad (19.105)$$

The quantity $\frac{eB}{4\pi m}$ is known as normal Zeeman separation. Knowing the values of v and B , the ratio e/m can be calculated. The value of e/m calculated from these measurements is in satisfactory agreement with the value $1.75 \times 10^{11} \text{ C/kg}$ obtained from Thomson's experiment.

This effect established that electrons emit the spectral lines.

19.27.2 Explanation of Zeeman Effect by Debye on the Basis of Vector Atom Model

P. Debye explained the Zeeman effect on the basis of vector atom model. He, however, did not take into account the concept of electron spin.

Let us consider an electron revolving in an orbit in an atom. Its orbital angular momentum is given by $p_l = l\hbar$. The magnetic moment of the orbit μ_l is directed opposite to p_l and is related to it through the expression

$$\begin{aligned} \mu_l &= -\left(\frac{e}{2m}\right)p_l \\ \therefore \mu_l &= -\left(\frac{eh}{4\pi m}\right)l \end{aligned}$$

Let the energy of the electron in its orbit be E . When a magnetic field of induction B is applied, the orbital angular momentum vector l precesses around the magnetic field direction with the Larmor frequency. The change in the energy of the electron is now given by

$$dE = -\mu_l \cdot B = \left(\frac{e\hbar}{4\pi m} \right) lB \cos \theta \quad (19.106)$$

where θ is the angle between l and B . θ is given by

$$\cos \theta = \frac{m_l}{l}$$

where m_l is the projection of l on to B .

$$\therefore dE = \left(\frac{e\hbar}{4\pi m} \right) m_l B$$

or $dE = m_l \mu_B B$ (19.107)

The eqn. (19.107) suggests that, in the absence of the magnetic field, the states differing in the values of the quantum number m_l have identical energy determined by the quantum numbers n and l . Consequently the electron orbit does not prefer any particular orientation. In the presence of a magnetic field, the energy depends on the value of m_l as well as on that of l . A state characterized by orbital quantum number l breaks up into $(2l + 1)$ substates due to the magnetic field, since m_l can take $(2l + 1)$ values. The substates will have slightly more or slightly less energy than the energy of the state in the absence of the field. All this means that the electron orbits prefer only certain orientations with respect to magnetic field and the energy of the orbit varies from orientation to orientation by a discrete amount $\mu_B B$.

If E_1 represents the energy of a state of the atom before the application of the magnetic field and E_A its energy after applying the magnetic field, then

$$E_A = E_1 + (m_l)_1 \mu_B B \quad (19.108)$$

If E_2 represents the energy of a second state which changes to E_B after applying the magnetic field, then

$$E_B = E_2 + (m_l)_2 \mu_B B \quad (19.109)$$

Both E_A and E_B are multiple states, as $(m_l)_1$ and $(m_l)_2$ can take $(2l_1 + 1)$ and $(2l_2 + 1)$ values respectively. Several components are likely to arise due to transitions between these various states. The frequency v of any of these components may be computed from the eqns. (19.108) and (19.109). Thus,

$$v = \frac{(E_B - E_A)}{\hbar} \quad (19.110)$$

$$\begin{aligned} &= \frac{E_2 - E_1}{\hbar} + \frac{(m_l)_2 - (m_l)_1}{\hbar} \mu_B B \\ &= v_o + \Delta m_l \left(\frac{eB}{4\pi m} \right) \end{aligned} \quad (19.111)$$

where $\Delta m_l = (m_l)_2 - (m_l)_1$. $v_o = \frac{E_2 - E_1}{\hbar}$ represents the frequency of the line in the absence of the magnetic field. The allowed transitions are determined by the selection rule $\Delta m_l = 0, \pm 1$. Using this selection rule, we find that the allowed transitions are only the following three.

$$\left. \begin{array}{ll} v_1 = v_o - \frac{eB}{4\pi m} & \text{corresponding to } \Delta m_l = -1 \\ v_1 = v_o & \text{corresponding to } \Delta m_l = 0 \\ v_1 = v_o + \frac{eB}{4\pi m} & \text{corresponding to } \Delta m_l = +1 \end{array} \right\} \quad (19.112)$$

The frequency shift due to the magnetic field is

$$\Delta v = \frac{eB}{4\pi m} \quad (19.113)$$

Fig.19.43 shows the splitting of the levels and spectral lines for the transition between the states $l = 1$ and $l = 0$. In the absence of the magnetic field, one spectral line is observed whose frequency is denoted by v_o . When the magnetic field is switched on, two lines appear in addition to the line v_o . The two lines having frequencies $v_o + \Delta v$ and $v_o - \Delta v$ are symmetrically located relative to the initial line v_o . The satisfactory agreement between the measured values of frequency shift and the values calculated from the eqn. (19.113) provided supporting evidence to the concept of space quantization. Thus, it confirmed that the electrons in an atom take on certain specified orientations in space and that these orientations are determined by the angular momentum projections on to the field direction.

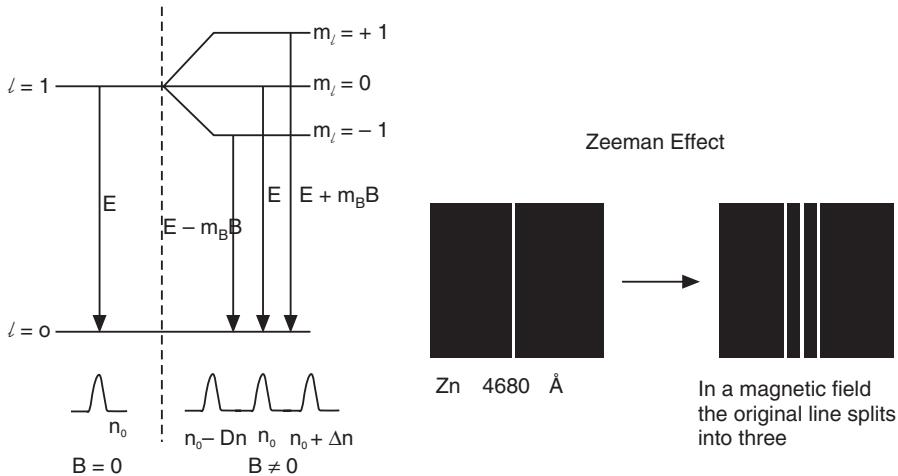


Fig. 19.43

Example 19.16. Determine the normal Zeeman splitting of the cadmium red line of 6438\AA when the atoms are placed in a magnetic field of 0.009 T .

Solution: We have $E = \frac{hc}{\lambda}$. $\therefore dE = -hc \frac{d\lambda}{\lambda^2}$ or $|dE| = hc \frac{|d\lambda|}{\lambda^2}$ $\therefore |d\lambda| = \frac{\lambda^2 |dE|}{hc}$

$$\therefore |dE| = \Delta E_{\text{Zeeman}} = \frac{e\hbar}{4\pi m} B = \left(5.79 \times 10^{-5} \frac{eV}{T} \right) (0.009T) = 5.21 \times 10^{-7} \text{ eV}$$

$$|d\lambda| = \frac{\lambda^2 |dE|}{hc} = \frac{(6438 \times 10^{-10} \text{ m})^2 (5.21 \times 10^{-7} \text{ eV})}{12.4 \times 10^3 \text{ eV\AA}} = 1.74 \times 10^{-3} \text{ \AA.}$$

19.28 THE STERN-GERLACH EXPERIMENT

In 1921 O. Stern and W. Gerlach directly measured the magnetic moments of atoms. These celebrated experiments provided direct proof for the space quantization of the angular momentum and for the spin of the electron.

Experimental Arrangement

Fig.19.44 shows a schematic of their experimental set up. A beam of silver atoms is produced by evapourating silver in a small electric oven. The beam is collimated by the slit S and passed through a very inhomogeneous magnetic field produced between two pole pieces and strikes a photographic plate. The pole pieces are 6 cm long and arranged 2 mm apart.

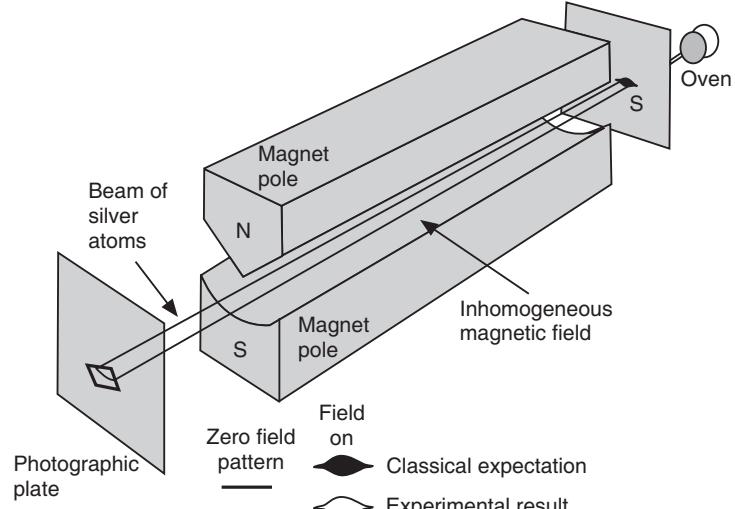


Fig. 19.44

One of the pole pieces is shaped in the form of a knife-edge and the other has a cylindrical groove, the lines of force converge very strongly towards the sharp edge. The magnetic field is thus highly non-uniform, having a gradient $\frac{dB}{dz}$ parallel to the flux density B acting in the z-direction.

Principle

The experiment is based on the behaviour of a magnetic dipole in a non-uniform magnetic field. In a uniform magnetic field, the magnetic dipole merely experiences a torque that tends to align the dipole in the field direction. If a beam of atomic magnets pass through a homogeneous magnetic field in a direction normal to the field, the beam will travel along a straight line path without any deviation. In an inhomogeneous magnetic field, the dipoles rotate into the field direction and in addition experience a translatory force in the direction of the field. Therefore, the atomic beam gets deflected along the field direction.

Let the magnetic field vary along the z-direction so that the field gradient is $\frac{dB}{dz}$ and is positive. Let MN be the atomic magnet having a pole strength m , length l and dipole moment $\mu = ml$. Let MN be inclined at an angle θ to the field direction. If the field strength at the pole M is B, then the field strength at the other pole N will be $B + \frac{dB}{dz}l \cos\theta$. Hence, the forces acting on the two poles of the magnet are now different and are given by mB and $m\left(B + \frac{dB}{dz}l \cos\theta\right)$. The translatory force on the magnet is

$$F_z = \frac{dB}{dz} ml \cos \theta$$

or $F_z = \mu_z \frac{dB}{dz}$ (19.114)

where $\mu_z = ml \cos \theta$ is the projection of the magnetic moment on to the field direction. This force deflects the atoms up or down depending on whether μ_z is positive or negative.

Expression for displacement in the field direction:

The acceleration, a , acquired by the atomic magnet along the field direction = $\frac{F_z}{m_A}$.

$$\text{The displacement of the atom along the field direction, } D = \frac{1}{2}at^2 = \frac{1}{2}\left(\frac{F_z}{m_A}\right)t^2$$

where m_A is the atomic mass. If L is the length of the magnetic field and v is the velocity of the atoms normal to the field direction, then

$$\begin{aligned} t &= \frac{L}{v} \\ \therefore D &= \frac{1}{2}\left(\frac{F_z}{m_A}\right)\left(\frac{L}{v}\right)^2 \\ \text{or } D &= \frac{\mu_z}{2m_A}\left(\frac{dB}{dz}\right)\left(\frac{L}{v}\right)^2 \end{aligned} \quad (19.115)$$

μ_z can be calculated knowing all the other quantities in the above equation.

(i) **Space quantization:** The distribution of directions of the magnetic moments in the incident beam is random. According to classical concepts, any orientation of the magnetic moment with respect to z-axis is equally probable. Therefore, each atom would then execute a Larmor precession about the field direction at some tilt angle θ . Since all values of θ occur among the atoms, the projection of the magnetic moment may have any value from $-\mu_z$ to $+\mu_z$, including zero value. Since there are several atoms in the beam, their deflections would be distributed in a continuous fashion and the trace formed by them on the photographic plate should be a *continuous band*.

But in the Stern-Gerlach experiment, two discrete traces are found for silver or hydrogen atomic beam. That is, the atomic beam is divided into two, each being deflected by the same amount but in opposite directions. It suggests that the z-projection of the magnetic moment can have only two values, equal in magnitude and opposite in direction. It proves that the classical theory, which expects equally probable orientation of the magnetic moment in random directions, is wrong. Discrete orientations are expected according to vector atom model. Owing to space quantization restriction, each atomic magnet can take up only certain specific orientations in the magnetic field. It appears in case of silver or hydrogen atoms, only two orientations are allowed. They correspond to μ_z parallel to B and μ_z antiparallel to B . This result establishes the existence of space quantization of angular momentum.

(ii) **Electron spin:** By the Stern-Gerlach experiment, the magnetic moment of the silver atom was found to be equal to one Bohr magneton.

$$\mu = \frac{e\hbar}{4\pi m} = \mu_B \quad (19.116)$$

The magnetic moment μ was initially attributed to the orbital magnetic moment of the electron in a fundamental Bohr orbit. At the time of Stern-Gerlach experiments, the electron configuration of silver atom was not clearly known. It came to be known later that monovalent atoms like silver have all their electrons, except one, arranged in shells, which are completed and therefore cannot contribute to the magnetic moment. The lone valence electron occupies a $S-$ subshell which has no orbital angular momentum. The absence of angular momentum in the atom leads to the conclusion that the quantized magnetic moment must be associated with some kind of non-orbital angular momentum. Obviously, it must be due to spinning of the electron. Thus, the Stern-Gerlach experiment provided direct experimental evidence of electron spin.

19.29 ANOMALOUS ZEEMAN EFFECT

The normal Zeeman effect is observed at high magnetic field strengths. In weak fields, a spectral line is found to split into more than three components. The splitting of spectral lines into even number of components in the presence of weak fields is known as *anomalous Zeeman effect*. The anomalous Zeeman effect can be explained only when the existence of electron spin is taken into account in addition to its orbital motion.

If \mathbf{L} and \mathbf{S} represent the orbital and spin angular momentum vectors, then their resultant \mathbf{J} will precess about the direction of the magnetic field. There are $(2J + 1)$ possible orientations of \mathbf{J} with respect to the field direction. Its components in the field direction are given by $m_j \hbar$ where $m_j = J, (J - 1), (J - 2), \dots, -J$. In a given magnetic field, an atomic energy level with a definite set of values for \mathbf{L} and \mathbf{J} splits up into $(2J + 1)$ closely lying component levels. The splitting between the adjacent levels depends on the value of J .

We know that the vectors \mathbf{L} and \mathbf{S} precess about the direction of \mathbf{J} . Since the magnetic moments μ_L and μ_S due to the orbital and spin motions are aligned antiparallel to the vectors \mathbf{L} and \mathbf{S} respectively, these two vectors also precess about the direction of \mathbf{J} . However, the magnitudes of the vectors μ_L and μ_S are different and hence, their resultant $\mu = (\mu_L + \mu_S)$ has a direction different from that of \mathbf{J} . As μ_L and μ_S precess about the direction of \mathbf{J} , their resultant also precesses about \mathbf{J} (see Fig. 19.45). The resultant magnetic moment μ can be resolved into its rectangular components μ_\perp and μ_J with respect to \mathbf{J} . The component μ_\perp averages out to zero whereas μ_J is equal to the sum of the components of μ_L and μ_S in the direction of \mathbf{J} . Thus,

$$\mu_J = \mu_L \cos(\mathbf{L}, \mathbf{J}) + \mu_S \cos(\mathbf{S}, \mathbf{J}) \quad (19.117)$$

where $\cos(\mathbf{L}, \mathbf{J})$ and $\cos(\mathbf{S}, \mathbf{J})$ are the cosines of the angles between \mathbf{J} and the vectors \mathbf{L} and \mathbf{S} respectively. Using cosines law, we get

$$\cos(\mathbf{L}, \mathbf{J}) = \frac{L(L+1) + J(J+1) - S(S+1)}{2\sqrt{L(L+1)} \cdot \sqrt{J(J+1)}} \quad (19.118)$$

$$\cos(\mathbf{S}, \mathbf{J}) = \frac{S(S+1) + J(J+1) - L(L+1)}{2\sqrt{S(S+1)} \cdot \sqrt{J(J+1)}} \quad (19.119)$$

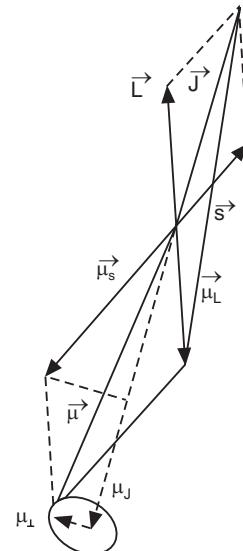


Fig. 19.45

Further, the quantum mechanical values of μ_L and μ_S are given by

$$\mu_L = \frac{e}{2m} p_L = \frac{e\hbar}{2m} \sqrt{L(L+1)} \quad (19.120)$$

and

$$\mu_S = \frac{e}{m} p_S = \frac{e\hbar}{2m} \cdot 2\sqrt{S(S+1)} \quad (19.121)$$

Using the above expressions into eqn. (19.110), we get

$$\mu_J = \frac{e\hbar}{2m} \left[\sqrt{L(L+1)} \frac{L(L+1) + J(J+1) - S(S+1)}{2\sqrt{L(L+1)}\sqrt{J(J+1)}} + 2\sqrt{S(S+1)} \frac{S(S+1) + J(J+1) - L(L+1)}{2\sqrt{S(S+1)}\sqrt{J(J+1)}} \right]$$

$$\text{or } \mu_J = \frac{e\hbar}{2m} \sqrt{J(J+1)} \left[1 + \frac{J(J+1) + S(S+1) - L(L+1)}{2J(J+1)} \right] \quad (19.122)$$

$$\text{or } \mu_J = g\mu_B \sqrt{J(J+1)} \quad (19.123)$$

where g is the Lande splitting factor given by

$$g = 1 + \frac{J(J+1) + S(S+1) - L(L+1)}{2J(J+1)} \quad (19.124)$$

The additional energy ΔE due to the action of the magnetic field on the atom is

$$\Delta E = \mu_J B \cos(\mathbf{J}, \mathbf{B}) = g\mu_B B \sqrt{J(J+1)} \cos(\mathbf{J}, \mathbf{B}) \quad (19.125)$$

The factor $\sqrt{J(J+1)} \cos(\mathbf{J}, \mathbf{B})$ is the component of the vector \mathbf{J} along the magnetic field \mathbf{B} and has the value m_J . Therefore,

$$\Delta E = \mu_B B g m_J \quad (19.126)$$

As the values of L , S , and J are different for different energy levels, g is different for different energy levels. Hence the magnetic splitting is different for different energy levels. The upper and lower levels split up by different amounts in a magnetic field. The transitions between the levels are governed by the selection rule

$$\Delta m_J = 0, \pm 1$$

For example, the transitions for the sodium lines are shown in Fig.19.46.

The ground state $^2S_{1/2}$ splits into two sublevels. This state is characterized by $L = 0$, and $S = \frac{1}{2}$. Therefore, $J = \frac{1}{2}$ and

$$g = 1 + \frac{\frac{1}{2}\left(\frac{1}{2}+1\right) + \frac{1}{2}\left(\frac{1}{2}+1\right) - 0}{2 \times \frac{1}{2}\left(\frac{1}{2}+1\right)} = 2.$$

Since m_J can have the values $\frac{1}{2}$ and $-\frac{1}{2}$, gm_J can have the values $+1$ and -1 . Thus, the

ground state splits into two levels, as illustrated in Fig.19.46. Similar calculations are carried out for gm_J corresponding to the other states. Applying the selection rule.

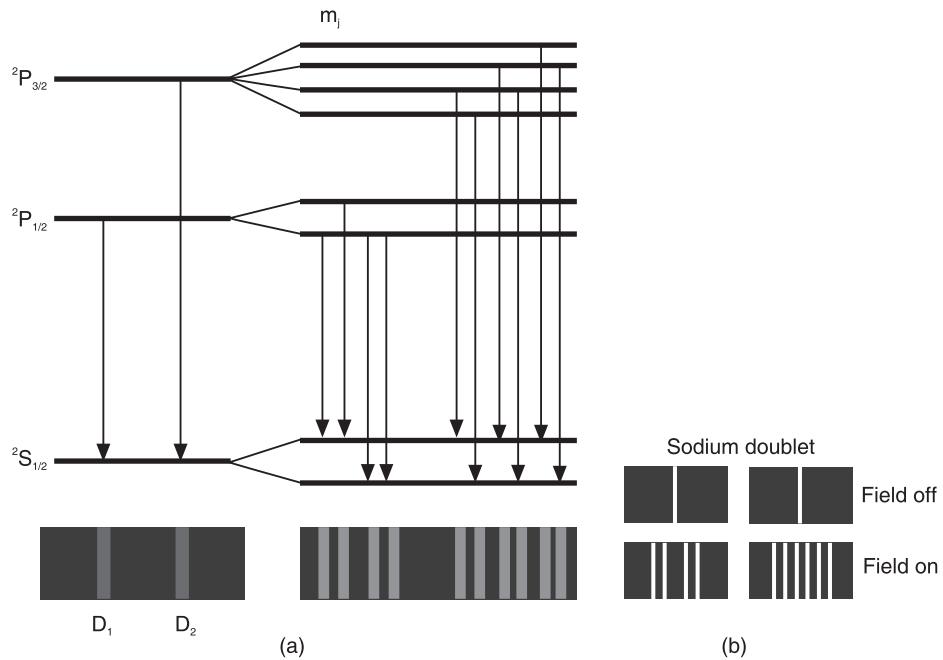


Fig. 19.46

$$\Delta m_J = 0, \pm 1$$

It is predicted that the spectral line D_1 corresponding to the transition $3^2P_{1/2} \rightarrow 3^2S_{1/2}$ splits into four levels and the spectral line D_2 corresponding to the transition $3^2P_{3/2} \rightarrow 3^2S_{1/2}$ splits into six levels. These conclusions agree with the experimentally observed splitting.

19.30 PASCHEN-BACK EFFECT

The anomalous Zeeman effect is observed when the applied magnetic field is relatively weak. In this effect, a spectral line splits into several lines. As the magnetic field is progressively increased, the splitting pattern becomes similar to normal Zeeman splitting where the spectral line splits into two or three components only. The reduction of the number of component lines takes place either through the coalescence of the lines or through the disappearance of certain lines. This transition is called **Paschen-Back effect**.

As the magnetic field becomes stronger, the coupling between the orbital and spin angular momentum vectors \mathbf{L} and \mathbf{S} breaks down. The vector \mathbf{J} loses its significance and the two vectors \mathbf{L} and \mathbf{S} now independently precess about the magnetic field direction, \mathbf{B} . The projection of \mathbf{L} in the field direction, i.e., m_l takes $(2L + 1)$ values. Similarly, the projection of \mathbf{S} in the field direction, i.e., m_s takes $(2S + 1)$ values. The magnitude of the splitting of an energy level in the magnetic field is given by

$$\begin{aligned}\Delta E &= \mu_B B m_L + 2\mu_B B m_S = \mu_B B(m_l + 2m_s) \\ \therefore \Delta v &= \mu_B B(\Delta m_L + 2\Delta m_S)\end{aligned}\quad (19.127)$$

The quantity $(m_l + 2m_s)$ is known as the *strong field quantum number* and is an integer. Since $\Delta m_L = 0, \pm 1$, and $\Delta m_S = 0$, $(\Delta m_L + 2\Delta m_S) = 0, \pm 1$. Hence, a given spectral line splits into three components only which is the pattern characteristic of normal Zeeman effect.

19.31 STARK EFFECT

In 1913, J. Stark discovered that a spectral line splits into several components by the action of the electric field. The splitting of the spectral lines in the presence of the electric field is known as **Stark effect**. This effect is analogous to the Zeeman effect. However, the splitting is very much smaller than in case of Zeeman effect and can be observed only with the help of high resolving power spectroscopes.

The experimental arrangement for the study of Stark effect is shown in Fig.19.47. It is an ordinary glass discharge tube provided with a perforated cathode, C. An auxiliary electrode F is placed close to the cathode. A very strong electric field of several thousand volts per metre is maintained between F and C. The discharge tube is filled with hydrogen gas. When the pressure in the tube is not very low, discharge takes place between the anode and cathode maintained at a suitable potential difference. The canal rays stream through the perforations in the cathode. The effect of the electric field can be studied both in transverse and longitudinal directions. With the help of a spectrometer, Stark noticed that the spectral lines emitted by the canal rays were split into numerous components under the action of the electric field. Observation perpendicular to the direction of the electric field showed that the components were polarized; some of them were polarized parallel to the direction of the field and others perpendicular to it.

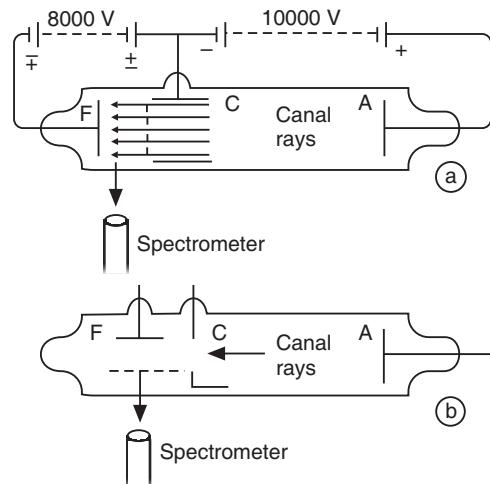


Fig.19.47

QUESTIONS

1. What is a black body? What are the salient features of black body radiation?
2. State and explain the laws relating to the radiation and temperature of a radiating body.
3. State and explain Stefan-Boltzmann law.
4. State and explain Wien's displacement law.
5. What is a blackbody? Give Planck's hypothesis. (Calicut Univ., 2006)
6. Discuss Planck's radiation law. (VTU, 2007)
7. Discuss in brief Planck's quantum hypothesis. (R.T.M.N.U., 2007)
8. State the basic postulate of Planck's theory. How did Einstein modify this theory in the light of experimental results?
9. (a) Explain briefly the distribution of energy in a blackbody spectrum. (b) What is Wien's displacement law? Explain the Wien's law of energy distribution. (Calicut Univ., 2006)
10. Explain the energy distribution in the spectrum of a black body. Give an account of the attempts made through various laws to explain the spectrum. (VTU, 2008)
11. Derive the Planck's law of radiation in terms of wavelength. (Calicut Univ., 2006)
12. Photoelectric effect is a frequency dependent phenomenon and not intensity dependent one. Explain.
13. The photon model succeeds whereas the wave model fails in explaining the photoelectric effect. Explain.

14. In a photoelectric effect, explain how a change in intensity and frequency affects the number of photoelectrons and the kinetic energies of these electrons, on the basis of Einstein's theory.
15. Describe Millikan's oil drop method of finding electronic charge. **(Shivaji Univ.)**
16. What are X-rays? How are they produced?
17. What factors influence (i) cut-off wavelength and (ii) the wavelengths of lines emitted by an X-ray tube?
18. Explain the production of X-rays by using Coolidge tube. **(Calicut Univ., 2007)**
19. What are continuous and characteristic X-rays? How are they produced? **(RGPV,2007)**
20. What is Duane-Hunt limit? **(RGPV,2007)**
21. Explain the origin of the continuous spectrum.
22. What is meant by minimum wavelength limit? Derive Duane-Hunt formula for a minimum wavelength limit of continuous X-rays?
23. What are the conclusions obtained from Duan-Hunt formula?
24. What are the types of X-ray spectra? How do we get the continuous and sharp line spectrum in X-rays?
25. What is the characteristic X-ray spectrum? How do we explain it?
26. What is the origin of the cut-off wavelength? Why is it an important clue for photonature of X-rays?
27. What are characteristic X-rays? Give its origin. Explain Moseley's law of characteristic X-rays. **(CSVTU,2006)**
28. State Duane and Hunt limit and Moseley's law and its significance. **(RGPV,2007)**
29. What are characteristic X-rays of an element? How can they be used to determine the atomic number of an element?
30. Discuss the origin and features of characteristic X-ray spectrum.
31. Discuss how Moseley explained the characteristic spectrum on the lines similar to line spectra of hydrogen-like atoms.
32. What is Compton effect? Why is it observed in light elements? **(RTMNU, 2007)**
33. Explain in brief Compton effect on the basis of quantum hypothesis.
34. Explain how wave theory of light failed to explain Compton effect? How the quantum theory explained this phenomenon?
35. Explain Compton effect and derive expression for Compton shift. **(CSVTU,2005, 2008)**
36. According to the special theory of relativity, the effective mass of the particle moving with large velocity is $m = \frac{m_0}{\sqrt{1 - v^2/c^2}}$. Show that energy of the particle is equal to $\sqrt{m^2 v^2 c^2 + m_0^2 c^4}$, where letters have their usual meanings.
37. How energy and momentum conservation occurs in Compton effect?
38. Write down an expression for Compton shift and explain in brief the existence of unmodified component in Compton scattering. **(RTMNU, 2007)**
39. How Planck's quantum theory can be used to explain qualitatively the existence of modified and unmodified component in Compton scattering.
40. Derive an expression for Compton shift in the wavelength of a photon after scattering from an electron. **(RGPV,2007)**
41. In compton effect what happens:
- (i) When photon collides with free electron in the scattering block?
 - (ii) When photon collides with bound electron in the scattering block? **(RTMNU-2010)**
42. What is Compton effect? Explain Compton scattering, by deriving the expression of Compton shift of wavelength with the help of a diagram. **(RGPV-2010)**

43. In Compton effect, considering elastic collision between a photon and a free electron, write down equations of energy and momentum conservation.

44. In Compton effect what happens: (i) when photon collides with free electron in the scattering block? Write down the equations of energy and momentum conservation; (ii) when photon collides with bound electron?

45. Explain the occurrence of an intensity peak corresponding to an unmodified component of wavelength in a scattered X-ray beam, when a beam of X-rays is arranged to fall on a graphite target. Why it is less pronounced for scatterers of low atomic number than those belonging to high Z group?

46. Does Compton scattering expression explain unmodified scattering? Explain.

47. Explain the existence of modified and unmodified radiations in Compton effect. (RTMNU, 2006)

48. Derive the expression for Compton shift. Why it is not observable in the visible region of electromagnetic radiation? (CSVTU,2007)

49. Write short notes on spectral series of atomic hydrogen.

50. Derive expression for radius of electron orbit and velocity of electron revolving in the orbit of hydrogen atom. (Amaravati Univ.,2003,2004)

51. Calculate radius of Bohr's orbit in ground state in a hydrogen atom. (Amaravati Univ.,2003,2005,2008)

52. Derive an expression for allowed energy of an electron in hydrogen atom. (Amaravati Univ.,2004)

53. Obtain an expression for energy of an electron in n^{th} orbit of Bohr's atom. (Amaravati Univ.,2004,2005,2006,2007,2008)

54. Give an account of Bohr's theory of hydrogen atom and deduce the expression for the wavelength of the first two lines of Balmer series. (Amaravati Univ.,2006)

55. Describe the various series of lines in the spectrum of atomic hydrogen and show how they have been explained in Bohr's theory.

56. Explain excitation and ionization of atom. (Amaravati Univ.,2003,2006)

57. On the basis of Bohr's theory deduce expression for (i) the ground state radius of the hydrogen atom and (ii) the ionization energy of the hydrogen atom.

58. Determine shortest wavelength in Brackett series of spectral lines. (Amaravati Univ.,2003)

59. Describe Frank-Hertz experiment and explain how the results of the experiment confirm Bohr's postulates.

60. Explain Frank-Hertz experiment and discuss its results. (Amaravati Univ.,2003,2004)

61. Describe Frank-Hertz experiment and prove that energy of radiation is absorbed by atom in quanta. (Amaravati Univ.,2003)

62. Describe Frank-Hertz experiment to prove the existence of discrete stationary states for electrons in atoms. (Amaravati Univ.,2005,2006)

63. Draw an energy level diagram for spectral series of hydrogen atom. (Amaravati Univ.,2003)

64. Draw energy level diagram for hydrogen atom and show different spectral series of hydrogen atom. (Amaravati Univ.,2005,2006)

65. What do you mean by the fine structure of spectral lines? Give an account of Bohr-Sommerfeld model of elliptical orbits of hydrogen atom. How does it account for the fine structure of hydrogen?

66. What are the different suggestions given by Sommerfeld in order to explain fine structure of H_{α} line of Balmer series and why? (Amaravati Univ.,2008)

67. Explain the concepts of vector atom model. (Calicut Univ.,2007)

68. Describe the vector model of the atom and explain the different quantum numbers associated with it. (Amaravati Univ.,2004)

69. Explain space quantization and spin of electron in vector atom model. (Amaravati Univ.,2004)

70. What do you mean space quantization and explain spinning of electron? (Amaravati Univ.,2008)

71. Show that magnetic moment due to spin motion of electron is always 1 Bohr magneton. (Amaravati Univ.,2008)

72. Derive an expression for magnetic moment of orbital electron. **(Calicut Univ.,2007)**
73. What is Pauli's exclusion principle? **(Amaravati Univ.,2004,2005,2006)**
74. State Pauli's exclusion principle. Explain how it helped in fixing up the electronic configuration of the elements in the periodic table.
75. Justify Pauli's exclusion principle for $n = 2$ when spin orbit interaction is considered. **(Amaravati Univ.,2008)**
76. Describe the Stern and Gerlach experiment and explain the importance of the results obtained.
77. Show that total magnetic moment offered by S electron is always 1 Bohr magneton. **(Amaravati Univ.,2008)**
78. What are the four quantum numbers required to specify the energy state of an electron in an atom having strong spin-orbit interaction? Explain their significance. **(Amaravati Univ.,2006)**
79. Prove that maximum number of electrons in a shell is $2n^2$ where n is the number of shell. **(Amaravati Univ.,2005)**
80. Write n, l, m_l and m_s for $3d^{10}$ electrons. **(Amaravati Univ.,2006)**
81. What is Zeeman effect? Explain normal Zeeman effect on the basis of classical ideas.
82. Outline the quantum theory of Zeeman effect and illustrate it with specific reference to sodium D-line.
83. Explain clearly the phenomenon of anomalous Zeeman effect. Describe the spectral patterns expected for the yellow lines of sodium.
84. Write short notes on: (i) Paschen-Back effect and (ii) Stark effect.

PROBLEMS

- A black body radiator at 0°C radiates energy of $3.2 \times 10^2 \text{ J.m}^{-2}.\text{s}^{-1}$. Deduce the value of Stefan's constant. **[Ans: $5.7 \times 10^{-8} \text{ J.m}^{-2}.\text{s}^{-1}$]**
- Using Wien's displacement law, estimate the temperature of sun. Given: $\lambda_m = 4900 \text{ \AA}$ and Wien's constant = 0.292 cm.K . **[Ans: 5963 K]**
- In an atomic explosion, the maximum temperature reached was of the order of 10^7 K . Calculate the wavelength of maximum energy. **[Ans: 2.93 \AA]**
- In order to break a chemical bond in the molecules of human skin, a photon energy of 3.5 eV is required. To what wavelength and region does this correspond? **[Ans : 3547 \AA]**
- A source is emitting 100 W of green light at a wavelength of 5000 \AA . How many photons are emerging from the source per second? **[Ans : $2.5 \times 10^{20} \text{ Photons/Second}$]**
- Calculate the wavelength of which the energy of a photon becomes equal to the average thermal energy of atoms in a solid at room temperature. **[Ans : 48 \mu\text{m}]**
- The work function of gold is 4.59 eV . What is the photoelectric threshold wavelength for gold? **[Ans : 2702 \AA]**
- The photoelectric threshold wavelength of tungsten is 2730 \AA . Determine the maximum kinetic energy of the electrons ejected from a tungsten surface by UV light of wavelength 1800 \AA . **[Ans:2.35 eV]**
- What is the potential difference that must be applied to stop the fastest photoelectrons emitted by a nickel surface under the action of UV light wavelength 2000 \AA ? The work function of nickel is 5 eV ? **[Ans:1.2V]**
- UV radiation of wavelength 750 \AA is incident on the surface of cesium. If the work function for cesium is 1.97 eV , determine the maximum speed with which the electrons are ejected from its surface. **[Ans : $2.3 \times 10^6 \text{ ms}$]**
- Calculate the energy gap in silicon if it is given that it is transparent to radiation of wavelength greater than 11000 \AA but it absorbs radiation of shorter wavelength. **(Shivaji University,1997)**
[Ans: 1.12 eV]
- Find the energy and momentum of an x-ray photon whose wavelength is 0.2 \AA .
[Ans: $9.95 \times 10^{15} \text{ J}$, $3.32 \times 10^{-23} \text{ kgm/s}$]

13. X-rays of wavelength 22×10^{-12} m (photon energy = 56keV) are scattered from a carbon target, the scattered radiation being viewed at 85° to the incident beam. (i) How much is Compton shift? (ii) What percentage of its initial energy does an incident X-ray photon lose?
[Ans: 0.02214 Å, 9.15%]
14. In a TV tube, the operating potential is 10 kV. Find the shortest wavelength emitted by screen when struck by electrons.
[Ans: 1.238 Å]
15. What is the maximum frequency of X-ray emitted if 12.4 kV potential is applied to a X-ray tube?
[Ans: 3×10^{10} GHz]
16. An X-ray tube operates at 40 kV emits continuous X-ray spectrum with a short wavelength limit of 0.31 Å. Calculate Planck' constant.
[Ans: 6.61×10^{-34} J.s]
17. A photon is Compton-scattered by a free electron at rest through an angle of 90° . What is the energy of the scattered photon if the energy of the incident photon is 10 MeV? **[Ans: 0.49 MeV]**
18. A photon of 4 Å strikes an electron at rest and is scattered at an angle of 150° to its original direction. Find the wavelength and speed of the photon after collision. **[Ans: 4.045 Å, 3×10^8 m/s]**
19. Calculate Compton wavelength for an electron.
[Ans: 0.0242 Å]
20. X-rays of wavelength 22×10^{-12} m (photon energy = 56keV) are scattered from a carbon target, the scattered radiation being viewed at 85° to the incident beam. (i) How much is Compton shift? (ii) What percentage of its initial energy does an incident X-ray photon lose?
[Ans: 0.02214 Å, 9.15%]
21. An X-ray photon of frequency 10^{19} Hz is scattered through an angle of 45° after colliding with a stationary electron. (a) What is its new frequency? (ii) What is the kinetic energy of the electron after collision?
[Ans: 9.8×10^{18} Hz, 938 eV]
22. Calculate the radii of the first, second and third permitted electron Bohr-orbits in a hydrogen atom.
[Ans: 0.0527 Å, 2.108 Å, 4.743 Å]
23. If the Rydberg constant is 1.097×10^7 m $^{-1}$, what are the wavelengths of the first three lines of the Paschen series?
[Ans: 18750 Å, 12810 Å, 10930 Å]
24. Calculate the frequency of electron in the first Bohr orbit in hydrogen atom. **[Ans: 6.5×10^{15} Hz]**
25. When hydrogen was bombarded in Frank-Hertz experiment by 10.21 eV and 12.10 eV electrons, emission of three spectral lines was observed. Calculate their wavelengths.
[Ans: 1026 Å, 1216 Å, 5672 Å]
26. If the average life time of an excited state of hydrogen is 10^{-8} s, estimate how many orbits an electron makes when it is in the state $n = 2$ and before it undergoes a transition to state $n = 1$. Given Bohr radius = 0.53 Å.
[Ans: 80,000]
27. In a normal Zeeman experiment, the calcium 4226 Å splits into 3 lines separated by 0.25 Å in a magnetic field of 3 T. Find e/m for the electron.
[Ans: 1.76×10^{11} C/kg]
28. Determine normal Zeeman splitting in mercury 4916 Å line when in a magnetic field of 0.3 T.
29. A 5000 Å line exhibits a normal Zeeman splitting of 1.1×10^{-3} Å. Find the magnetic field.
30. What magnetic flux density is required to observe the normal Zeeman effect if a spectrometer can resolve spectral lines separated by 0.5 Å at 500 Å.
[Ans: 4.28 T]
31. A sample of certain element is placed in magnetic field of density 0.3 T. How far apart are the Zeeman components of wavelength 4500 Å, if $e/m = 1.76 \times 10^{11}$ C/kg?
[Ans: 0.0283 Å]
32. The experimental value of Bohr magneton is 9.21×10^{24} SI units. Calculate the value of e/m for the electron.
33. Find the precessional frequency of an electron orbit when placed in a magnetic field of 5T.
[Ans: 7×10^{10} Hz]
34. A beam of silver atoms in a Stern-Gerlach experiment obtained from an oven heated to a temperature of 1500 K, passes through an inhomogeneous magnetic field having a field gradient of 2 Wb/m 2 /cm perpendicular to the beam. The pole pieces are 10 cm long. What is the separation between the two components of the beam on a photographic plate placed at a distance of 50 cm?
[Ans: 0.15 mm]

CHAPTER

20

Quantum Mechanics

20.1 INTRODUCTION

In 1925, de Broglie introduced the concept of matter waves and the idea of wave-particle duality. He suggested that the wave-particle duality observed in case of light should be extended to microparticles also. The combination of the idea of quantization with the idea of wave-particle duality proved to be very fruitful for the development of quantum mechanics. The whole apparatus of quantum mechanics was built in 1925-26. In 1925 Heisenberg suggested that any reference to conceptual pictures which are not amenable to direct experimental verification, should be discarded. He formulated matrix mechanics which is set in terms of the observable quantities alone. In 1926, Schrödinger developed wave mechanics. His theory is based on explicit use of a mental picture of matter wave which replaced the classical picture of point particle. He developed the well-known differential equation for a wave function. The problem of calculating the energy levels of a bound microparticle was reduced by Schrödinger to the problem of finding eigenvalues. Heisenberg's theory came to be known as **matrix mechanics** while that of Schrödinger as **wave mechanics**. In 1926, Max Born proposed the probability interpretation of the wave function. In 1927 the wave behaviour of microparticles was confirmed by experiments on electron diffraction conducted simultaneously in several different laboratories. In 1927, Heisenberg introduced the uncertainty principle. Through this principle Heisenberg showed how the concepts of coordinate, momentum, energy etc should be applied to microparticles. The uncertainty principle marked the final break of quantum mechanics from classical determinism and established quantum mechanics as a statistical theory. In quantum mechanics, the waves and particles are not classes of objects. They are distinct modes of behaviour shared by all atomic particles. Every microparticle can behave like a particle and like a wave too. In 1930, P.A.M. Dirac proposed a general formalism which is a unifying concept of the matrix mechanics and wave mechanics.

The new laws applicable for atoms and subatomic particles constitute **quantum mechanics**. The laws of conservation of momentum, angular momentum and energy are still valid but we are not in a position to obtain on their basis a detailed description of the motion of the subatomic particles. New ideas such as quantization of physical quantities and allowed values of physical quantities are required to be incorporated into the theory. Planck's hypothesis of energy quanta, Einstein's ideas on photons, de Broglie's visualization of the wave properties of micro-particles and the Heisenberg's uncertainty principle provided the basis for the development of quantum mechanics.

20.2 DE BROGLIE HYPOTHESIS

In 1924, Louis de Broglie extended the wave–particle dualism of light to the material particles. He reasoned out that nature exhibits a great amount of symmetry. Therefore, if a light wave can act as a wave sometimes and as a particle at other times, then particles such as electrons should also act as waves at times. This is known as *de Broglie hypothesis*.

According to de Broglie hypothesis any moving particle is associated with a wave. The waves associated with particles are known as **de Broglie waves** or **matter waves**. The wavelength λ of matter waves associated with a particle moving with velocity v is inversely proportional to the magnitude of the momentum of the particle. Thus,

$$\lambda = \frac{h}{mv} \quad (20.1)$$

De Broglie deduced the connection between the particle and wave properties as follows.

De Broglie wavelength of Matter waves

As a photon travels with the velocity c , we can express its momentum as

$$p = \frac{E}{c} = \frac{hv}{c} = \frac{h}{\lambda}$$

Thus, the wavelength and momentum p of a photon are related to each other through the expression

$$\lambda = \frac{h}{p} \quad (20.2)$$

De Broglie proposed that the relation (20.2) between the momentum and the wavelength of a photon is a universal one and must be applicable to photons and material particles as well.

The quantities v and λ are wave properties and the quantities E and p are particle properties. They are tied to each other through the relations

$$E = hv \text{ and } p = h/\lambda$$

which demonstrate that the wave and particle natures of a photon are intimately tied up to each other. These equations reflect the wave-particle dualism of light.

Now let us consider a moving particle. A particle of mass ‘ m ’ moving with a velocity v carries a momentum $p = mv$ and it must be associated with a wave of wavelength

$$\lambda = \frac{h}{p} = \frac{h}{mv} \quad (20.3)$$

The waves associated with moving particles are called **matter waves** or **de Broglie waves**.

The relation $\lambda = h/mv$ is known as **de Broglie equation** and the wavelength λ is called the **de Broglie wavelength**.

From equ.(20.3), we may draw the following conclusions.

1. $\lambda \rightarrow \infty$ when the velocity of the particle is zero. It means that matter waves are detectable only for moving particles.
2. Lighter the particle, smaller the value of mass m and hence the longer is the wavelength of the matter wave associated with it. Therefore, wave behaviour of micro-particles will be significant whereas waves associated with macro-bodies can never be detected.

3. The smaller the velocity of the micro-particle, the longer is the wavelength of the matter wave associated with it.

If a photon is considered to be a particle, then the corresponding electromagnetic wave is the de Broglie wave for the photon. Similarly, atomic particles also can be viewed as associated with matter waves, which do not have any similarity to any known waves. It is later understood that the waves associated with particles are not real three dimensional waves in the way sound waves are, but are **probability waves** related to the probabilities of finding the particles in various places and with various properties.

De-Broglie wavelength associated with an accelerated charged particle

If a charged particle, say an electron is accelerated by a potential difference of V volts, then its **kinetic energy** is given by $K.E. = eV$.

Or

$$\frac{1}{2}mv^2 = eV$$

∴

$$v = \sqrt{\frac{2eV}{m}}$$

Then the electron wavelength is given by

$$\lambda = \frac{h}{mv} = \frac{h}{m} \cdot \sqrt{\frac{m}{2eV}}.$$

$$\therefore \lambda = \frac{h}{\sqrt{2emV}}$$

(20.4)

De Broglie Wavelength expressed in terms of K.E.

If a particle has kinetic energy K.E., then $K.E. = \frac{1}{2}mv^2 = \frac{m^2v^2}{2m} = \frac{p^2}{2m}$

or

$$p = \sqrt{2m(K.E.)}$$

∴

$$\lambda = \frac{h}{\sqrt{2m(K.E.)}}$$

(20.5)

De Broglie wavelength associated with particles in thermal equilibrium

If particles are in thermal equilibrium at temperature T , then their kinetic energy is given by

$$K.E. = \frac{3}{2}kT$$

∴

$$\lambda = \frac{h}{\sqrt{2m(K.E.)}} = \frac{h}{\sqrt{3mkT}}$$

(20.6)

20.3 DE BROGLIE'S JUSTIFICATION OF BOHR'S POSTULATE

Bohr's theory of atomic structure was successful in explaining a large body of experimental observations concerning atomic behaviour but the three adhoc postulates underlying his theory remained without a valid theoretical justification for a long time. In support of his hypothesis of matter waves, de Broglie demonstrated that it could provide an explanation for the postulate regarding quantization of angular momentum of electron in Bohr's model of atom.

One of the postulates that Bohr used in formulating a model of atom is that the angular momentum L of the electron revolving in a stationary orbit is quantized. Thus,

$$L = n\hbar \quad (20.7)$$

The above postulate of Bohr follows directly from the concept of matter waves.

As the electron travels round in one of its circular orbits, the associated matter waves propagate along the circumference again and again. A wave must meet itself after going round one full circumference. If it did not meet, the wave would be out of phase with itself after going round one orbit. After a large number of orbits, all possible phases would be obtained and the wave would be annihilated by destructive interference. It implies that the wave should produce a standing wave profile in the orbit to preclude the electron energy from radiating away.

If a stretched string is fastened at both ends and is made to vibrate, standing waves are formed provided the length of the string is an integral number of half-wavelengths of the disturbance. If the string is formed into a circular loop, the condition for standing waves is that the circumference of the loop should be an integral number of whole wavelengths of the disturbance. Thus, if r is the radius of the circular loop,

$$2\pi r = n\lambda \quad n = (1, 2, 3, \dots) \quad (20.8)$$

We may regard the stationary electron orbits in an atom to be analogous to the circular loop of string. We conclude that stationary electron wave pattern can form in the orbit if only an integral number of electron wavelengths fit into the orbit, as shown in Fig. 20.1 (b).

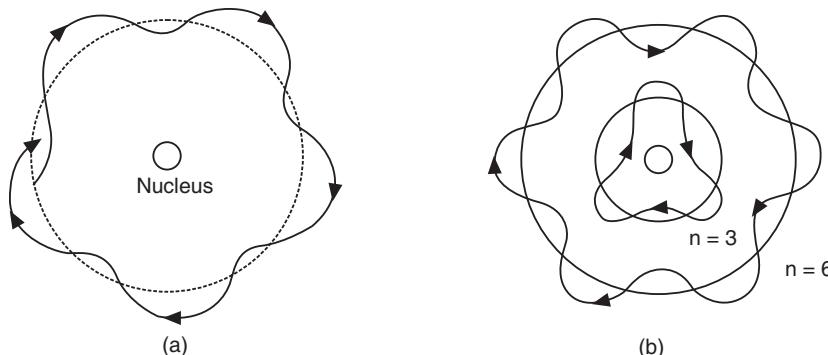


Fig. 20.1

The above equation (20.8) can be applied for electron waves, taking λ as de Broglie wavelength of electron waves.

The de Broglie wavelength of electron wave is given by

$$\lambda = \frac{h}{mv}$$

where v is the speed of the electron in the orbit. Using the de Broglie wavelength into equ.(20.8), we obtain

$$2\pi r = \frac{nh}{mv} \\ \therefore mv r = \frac{nh}{2\pi} \quad (20.9)$$

But the quantity ' mvr ' is the angular momentum, L , of electron in the orbit of radius r . Thus,

$$L = mvr$$

It, therefore, follows that

$$L = n\hbar$$

which is precisely Bohr's postulate. De Broglie thus demonstrated that the quantization of angular momentum is a direct consequence of wave nature of electron.

Let us calculate the wavelength of the electron in the first orbit of hydrogen atom. The electron speed v in the orbit is given by

$$\begin{aligned} v &= \frac{e}{\sqrt{4\pi\varepsilon_0 mr}} \\ \therefore \lambda &= \frac{h}{e} \sqrt{\frac{4\pi\varepsilon_0 r}{m}} \end{aligned} \quad (20.10)$$

Taking $r = 5.3 \times 10^{-11}$ m, we get $\lambda = 3.3$ Å. The circumference of the orbit is $2\pi r = 3.3$ Å.

Hence, the first orbit of the electron in a hydrogen atom corresponds to one complete electron wave joined on itself. The de Broglie hypothesis thus offered a new meaning to the quantum number ' n '. n is the number of de Broglie wavelengths that fit into the circumference of Bohr allowed orbits.

We may now picture the electron in the atom in two ways: either as a particle moving in an orbit with a certain quantized value of mvr , or as a standing de Broglie wave occupying a certain region around the nucleus.

20.4 DE BROGLIE WAVES ARE INSIGNIFICANT IN CASE OF MACRO-BODIES

According to de Broglie hypothesis a moving body is associated with matter waves and the wavelength of the waves is given by

$$\lambda = \frac{h}{mv}$$

where v is the velocity with which the body moves.

As the mass m of the body increases, the wavelength tends to be insignificant. Therefore, the wavelength associated with macroscopic bodies become insignificant in comparison to the size of the bodies themselves even at very low velocities. Because of the smaller magnitude of Planck's constant h , the wavelength λ will be significant only in case of micro-particles.

For example, if we consider a cricket ball of mass 500 gm flying with a velocity of 50 km/hr, its wavelength comes to

$$\lambda = \frac{6.62 \times 10^{-34} J.s}{0.5kg \times 13.9m/s} = 10^{-34} m = 10^{-24} \text{ Å.}$$

It is easy to see that this wavelength is insignificant in comparison to the size of the ball.

On the other hand, if we consider the case of an electron, having energy 100 eV, the de Broglie wavelength of the electron is given by

$$\begin{aligned} \lambda &= \frac{h}{\sqrt{2meV}} \\ &= \frac{6.62 \times 10^{-34} J.s}{\sqrt{2 \times 9.11 \times 10^{-31} kg \times 1.602 \times 10^{-19} C \times 100 V}} = 1.33 \text{ Å.} \end{aligned}$$

The size of an electron is about 10^{-5} Å, which is far smaller than the wavelength of 1.33 Å. It means that the electron behaves more as a wave than a particle under the circumstances.

20.5 PROPERTIES OF MATTER WAVES

1. Matter waves are produced by the motion of the particles and are independent of the charge. Therefore, they are neither electromagnetic nor acoustic waves but are new kind of waves.
2. They can travel through vacuum and do not require any material medium for their propagation.
3. The smaller the velocity of the particle, the longer is the wavelength of the matter wave associated with it.
4. The lighter the particle, the longer is the wavelength of the matter wave associated with it.
5. The velocity of matter waves depends on the velocity of the material particle and is not a constant quantity.
6. The velocity of matter waves is greater than the velocity of light.
7. They exhibit diffraction phenomenon as any other waves.

20.6 DAVISSON-GERMER EXPERIMENT

Waves exhibit diffraction. If the de Broglie hypothesis is valid, then the matter waves should exhibit diffraction effects. Diffraction is observed when the wavelength is comparable to the size of the object causing diffraction. The wavelength of 100 eV electrons is of the order of 1 Å and the interatomic spacing in a crystal is of the order of 2 to 3 Å. Therefore, we expect that the wave behaviour of a micro-particle such as an electron becomes noticeable when a beam of particles interact with crystals. In 1927, Davisson and Germer observed the diffraction of an electron beam incident on a nickel crystal. The experiment provided a convincing proof of the wave nature of matter.

Apparatus

The experimental arrangement of Davisson and Germer is shown in Fig. 20.2. The apparatus consisted of an electron gun, which produced collimated beam of electrons. An anode, A, connected to a variable voltage source accelerated the electrons. The energy of the electrons can be computed from the accelerating potential. These electrons were scattered by a nickel crystal located at C. The crystal can be rotated on the axis. The number of electrons scattered by the crystal in different directions was measured with the help of a detector D, which can be moved on a scale.

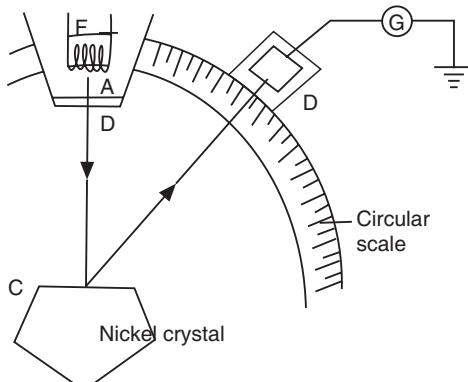


Fig. 20.2

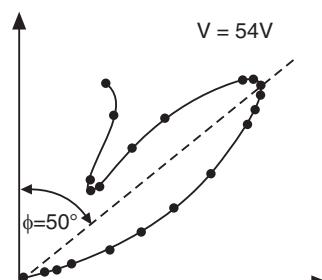


Fig. 20.3

The experimental arrangement of Davisson and Germer is shown in Fig. 20.2. An electron beam is generated from a hot tungsten filament F and an anode A connected to a variable voltage source accelerated the electrons. The energy of the electrons can be computed from the accelerating potential, V applied between the filament F and the anode A. The electrons emerge through an opening in the anode and fall normally on the surface of a nickel crystal, C. These electrons are scattered by a nickel crystal located at C. The crystal can be rotated on the axis. The detector D measured the number of electrons scattered by the crystal in different directions. The detector could be moved on a graduated semicircular scale. Thus, the intensity of the scattered electron beam was determined as a function of the scattering angle, ϕ .

Investigations

During their experiments Davisson and Germer moved the detector on the circular scale to various positions and the current was measured. The detector current is a measure of the intensity of the diffracted beam. A polar graph was then plotted between the detector current and the angle between the incident beam and the diffracted beam. Such polar curves were obtained for electrons accelerated through different voltages. It was found that a hump appears in the polar curve when 44 eV electrons were incident on the crystal. The hump grew in size as the accelerating voltage is increased and became most pronounced at 54 volts. The polar curve corresponding to 54 V is shown in Fig. 20.3. It is found that for the accelerating voltage of 54 volts, the electrons are scattered more pronouncedly at an angle of 50° with the direction of the incident beam. The maximum is an indication that electrons are being diffracted.

Analysis

It may be interpreted that the rows of atoms at the surface of the nickel crystal act like rulings of a natural diffraction grating and the de Broglie waves associated with the electrons underwent diffraction when they were incident on the crystal. The hump produced at 50° in Fig. 20.3 then corresponds to the first order diffraction maxima. Braggs' law, applicable for X-ray diffraction by crystals, would be valid for electron wave diffraction also. Fig. 20.4 shows atomic planes and the incident and scattered beams. The interplanar spacing is obtained from X-ray analysis to be $d = 0.91 \text{ \AA}$.

From the Fig. 20.4, it is seen that the glancing angle $\theta = 65^\circ$. Applying Braggs' equation,

$$\lambda = 2d \sin \theta = 2 \times 0.91 \text{ \AA} \times \sin 65^\circ = 1.65 \text{ \AA}$$

The wavelength of the electron wave is thus determined to be 1.65 \AA . The wavelength of electron wave can be computed from the accelerating potential V using de Broglie equation.

$$\begin{aligned} \lambda &= \frac{h}{\sqrt{2meV}} \\ &= \frac{6.63 \times 10^{-34} \text{ Js}}{\left[2 \times 9.1 \times 10^{-31} \text{ kg} \times 1.602 \times 10^{-19} \text{ C} \times 54 \text{ V} \right]} = 1.66 \text{ \AA} \end{aligned}$$

It is seen that the values obtained experimentally using Braggs' equation and de Broglie equation agreed well. Therefore, Davisson-Germer experiment gave conclusive evidence that electrons exhibit diffraction property.

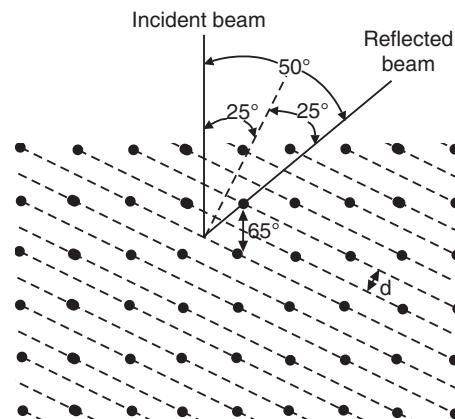


Fig. 20.4

20.7 G.P. THOMSON'S EXPERIMENT

The de Broglie hypothesis was further confirmed in 1927 by the experiments conducted independently by G.P.Thomson in England and by Kikuchi in Japan.

Thomson's experimental arrangement is shown in Fig. 20.5. Electrons are produced from a heated filament F and accelerated through a high positive potential given to the anode A. The whole apparatus is kept highly evacuated. The electron beam passes through a fine hole in a metal block B and falls on a gold foil of thickness $0.1 \mu\text{m}$. The electrons passing through the foil are received on a photographic plate P.

Metals are polycrystalline in which the grains are oriented completely at random. Therefore, some grains have always the right inclination θ towards the incident beam in order to produce a Bragg reflection. Owing to the random orientation, the reflections from a given set of lattice planes at the glancing angle occur in every azimuth about the incident beam. Consequently, the reflected beams form a cone of semi-vertical angles 2θ . Each set of lattice planes with its particular spacing d in the grain produces its own cone of diffracted rays. A concentric rings pattern is produced on the photographic plate when it intercepts the coaxial cones of diffracted rays.

The diffraction pattern produced by the electron beam was strikingly similar to the x-ray diffractions obtained from powder samples. Thus, the experiments of G.P. Thomson and Kikuchi provided irrefutable proof to the existence of de Broglie waves.

In 1937, C.J.Daavisson and G.P.Thomson were jointly awarded the Noble Prize in physics for their experimental discovery of electron diffraction.

In 1929, soon after the discovery of wave properties of electrons, the German physicist Otto Stern and his coworkers detected diffraction phenomena with neutral atomic and molecular beams.

20.8 VELOCITY OF DE BROGLIE WAVES

Any harmonic wave is characterized by a precise wavelength λ and constant amplitude. It is non-localized and has no beginning and end. It means that such a wave extends over a very large volume of space.

20.8.1 Phase Velocity

If we consider a harmonic wave, the wave has a single wavelength and a single frequency. The velocity of propagation of the wave is given by

$$v_p = v\lambda$$

Using, $v = \omega/2\pi$ and $\lambda = 2\pi/k$ into the above equation, we get

$$v_p = \frac{\omega}{2\pi} \cdot \frac{2\pi}{k} = \frac{\omega}{k} \quad (20.11)$$

v_p is called the *phase velocity*. The velocity with which the plane of equal phase travels through a medium is known as the phase velocity. It thus represents the velocity of propagation of the wave front.

As

$$E = hv \text{ and } p = h/\lambda, \text{ we get}$$

$$v_p = \frac{E}{h} \cdot \frac{h}{p} = \frac{E}{p} \quad (20.12)$$

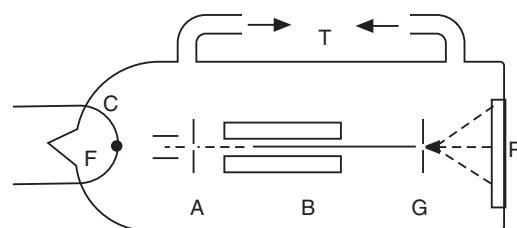


Fig. 20.5

- (i) When the atomic particle velocity is *non-relativistic*, the total energy $E = mc^2$ and momentum $p = mv$.

Therefore, the phase velocity of the de Broglie wave associated with the particle is

$$v_p = \frac{E}{p} = \frac{mc^2}{mv} = \frac{c^2}{v} \quad (20.13)$$

As $v < c$, the phase velocity of the de Broglie wave associated with the atomic particle is always greater than c .

- (ii) When the atomic particle velocity is *relativistic*, the total energy $E = \sqrt{m_o^2 c^4 + p^2 c^2}$, where m_o is the rest mass of the particle.

Therefore, the phase velocity of the de Broglie wave associated with the particle is

$$\begin{aligned} v_p &= \frac{E}{p} = \left[\frac{m_o^2 c^4 + p^2 c^2}{p^2} \right]^{1/2} \\ &= c \left[\frac{m_o^2 c^2}{p^2} + 1 \right]^{1/2} = c \left[\frac{m_o^2 c^2 \lambda^2}{h^2} + 1 \right]^{1/2} \end{aligned} \quad (20.14)$$

As the term $\frac{m_o^2 c^2 \lambda^2}{h^2}$ is always a positive quantity, the phase velocity of the de Broglie wave

associated with the atomic particle is always greater than c .

According to the theory of relativity, it is not possible that the velocity of the particle wave be greater than or equal to the velocity of light. Hence, a harmonic wave of wavelength λ cannot represent a moving atomic particle. Thus, ***de Broglie waves cannot be harmonic waves.***

20.9 WAVE PACKET – REPRESENTS A MICROPARTICLE

We have so far assumed that a particle may be represented by a monochromatic de Broglie wave. However, a wave spreads over a large region of space and cannot represent a highly localized particle. Schrödinger postulated that a **wave packet** rather than a single harmonic wave represents a particle. A wave packet consists of a group of harmonic waves. Each wave has slightly different wavelength. The superposition of a very large number of harmonic waves differing infinitesimally in frequency will produce a single wave packet (see Fig. 20.6 c). The waves interfere constructively over only a small region of space and cancel each other everywhere except in that small region. The position of the particle would then be approximately determined by the position of the wave packet.

The velocity with which the wave packet propagates is called the **group velocity** v_g .

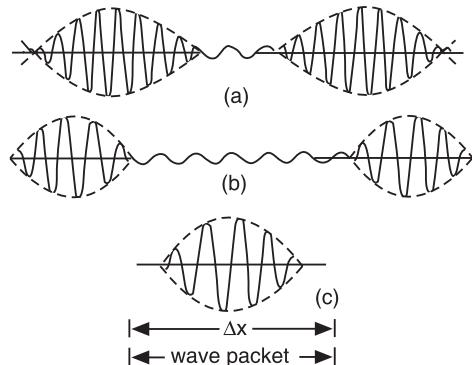


Fig. 20.6: Formation of a Wave packet. (a) two waves of slightly different frequencies produce constructive interference. (b) three waves produce interference maxima of larger size separated by larger distance. (c) A large number of waves having slightly different frequencies produces only one maximum and it is called a wave packet.

The individual waves forming the wave packet propagate at a velocity known as the **phase velocity** v_p .

20.9.1 Group Velocity

When a number of plane waves of slightly different wavelengths travel in the same direction, they form wave groups or wave packets. The velocity with which the wave group advances in the medium is known as the *group velocity* v_g . Each component wave has its own phase velocity, $v_p = v\lambda$. The wave packet has amplitude that is large in a small region and very small outside it. The amplitude of the wave packet varies with x and t . Such a variation of amplitude is called the *modulation* of the wave. The velocity of propagation of the modulation is known as the *group velocity*, v_g .

Here, we should note that *wave packets are only theoretical artifices to aid our visualization of various phenomena in the micro-world.*

Expression for the Group Velocity

We derive now an expression for group velocity considering a group of waves consisting of two components of equal amplitude and slightly differing angular velocities ω_1 and ω_2 .

Let the waves in Fig. 20.7 (a) be represented by the equations

$$\begin{aligned}y_1 &= A \sin(\omega_1 t - k_1 x) \\y_2 &= A \sin(\omega_2 t - k_2 x)\end{aligned}$$

The superposition of these two waves is given by

$$y_1 + y_2 = A \sin(\omega_1 t - k_1 x) + A \sin(\omega_2 t - k_2 x)$$

Using the trigonometric relation $\sin \alpha + \sin \beta = 2 \sin\left(\frac{\alpha+\beta}{2}\right) \sin\left(\frac{\alpha-\beta}{2}\right)$, we write the above equation as

$$\begin{aligned}y_1 + y_2 &= 2A \sin\left[\frac{(\omega_1 + \omega_2)}{2}t - \frac{(k_1 + k_2)}{2}x\right] \cos\left[\frac{(\omega_1 - \omega_2)}{2}t - \frac{(k_1 - k_2)}{2}x\right] \\&= 2A \sin(\omega t - kx) \cos\left(\frac{\Delta\omega t}{2} - \frac{\Delta k x}{2}\right) \quad (20.15)\end{aligned}$$

where $\omega = (\omega_1 + \omega_2)/2$, $k = (k_1 + k_2)/2$, $\Delta\omega = \omega_1 - \omega_2$ and $\Delta k = k_1 - k_2$. Eq.(20.15) represents the resultant wave which is seen to have the following two parts.

(i) A wave of angular frequency ω and propagation constant k , moving with a velocity

$$v_p = \frac{\omega}{k} = v\lambda \text{ and}$$

(ii) A second wave of angular frequency $\Delta\omega/2$ and propagation constant $\Delta k/2$, moving with a velocity $v_g = \frac{\Delta\omega}{\Delta k}$.

When $\Delta\omega$ and Δk are very small, we can write the above equation as

$$v_g = \frac{d\omega}{dk} \quad (20.16)$$

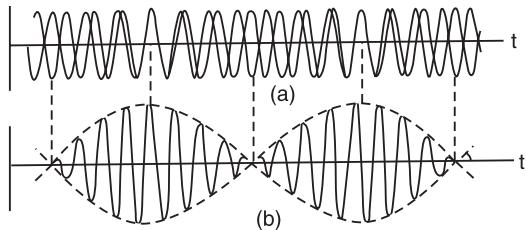


Fig. 20.7 Beats are formed when two waves of slightly different frequencies combine
(a) the individual waves (b) the resultant wave.

or

$$v_g = \frac{2\pi d\nu}{2\pi d(1/\lambda)} = -\lambda^2 \frac{d\nu}{d\lambda}$$

20.9.2 Relation between Phase Velocity and Group Velocity

The velocity of the individual component wave of the wave packet is given by

$$v_p = v\lambda$$

Using, $v = \omega / 2\pi$ and $\lambda = 2\pi / k$ into the above equation, we get

$$v_p = \frac{\omega}{2\pi} \cdot \frac{2\pi}{k} = \frac{\omega}{k} \quad (20.17)$$

∴

$$\omega = kv_p$$

The group velocity is given by the relation (20.16) as

$$v_g = \frac{d\omega}{dk} = \frac{d}{dk}(kv_p) = v_p + k \frac{dv_p}{dk}$$

But

$$k = \frac{2\pi}{\lambda}.$$

Therefore,

$$dk = -\frac{2\pi}{\lambda^2} d\lambda$$

and

$$\frac{k}{dk} = -\frac{\lambda}{d\lambda}.$$

∴

$$v_g = v_p - \lambda \frac{dv_p}{d\lambda} \quad (20.18)$$

Group velocity will be the same as phase velocity if the entire constituent waves travel with the same velocity. It means that in a nondispersive medium, $v_g = v_p$. However, the waves of different wavelengths travel in a medium with different velocities. Therefore, the group velocity is in general less than the phase velocity.

20.9.3 The Velocity of a Particle Equals the Group Velocity of the Associated Matter Waves

A particle moving with a velocity v is supposed to consist of a group of de Broglie waves. The group velocity of a wave packet is given by

$$v_g = \frac{d\omega}{dk}$$

which we can write as

$$v_g = \left(\frac{d\omega}{dE} \right) \left(\frac{dE}{dp} \right) \left(\frac{dp}{dk} \right) \quad (20.19)$$

As

$$E = h\nu = h \cdot \frac{\omega}{2\pi} = \frac{h}{2\pi} \cdot \omega = \hbar\omega, \quad \frac{d\omega}{dE} = \frac{1}{\hbar}$$

and

$$p = \frac{h}{\lambda} = h \cdot \frac{1}{\lambda} = h \cdot \frac{k}{2\pi} = \frac{h}{2\pi} \cdot k = \hbar k, \quad \frac{dp}{dk} = \hbar$$

∴

$$v_g = \left(\frac{d\omega}{dE} \right) \left(\frac{dE}{dp} \right) \left(\frac{dp}{dk} \right) = \frac{1}{\hbar} \left(\frac{dE}{dp} \right) \hbar = \left(\frac{dE}{dp} \right) \quad (20.20)$$

For a particle, $E = \frac{1}{2}mv^2 = \frac{1}{2} \frac{(mv)^2}{m} = \frac{p^2}{2m}$.

$$\therefore v_g = \frac{dE}{dp} = \frac{p}{m} = v. \quad (20.21)$$

Thus, the de Broglie wave group associated with an atomic particle travels with the same velocity as that of the particle itself.

20.9.4 Relation Between the Group Velocity and Particle Velocity (in a Non-dispersive Medium)

A particle moving with a velocity v is supposed to consist of a group of de Broglie waves. For an atomic particle of rest mass m_o moving with a velocity v , the total energy and momentum are given by

$$E = mc^2 = \frac{m_o c^2}{\sqrt{1-v^2/c^2}} \text{ and } p = mv = \frac{m_o v}{\sqrt{1-v^2/c^2}} \text{ respectively.}$$

The frequency of the associated de Broglie wave is

$$\nu = \frac{E}{h} = \frac{m_o c^2}{h \sqrt{1-v^2/c^2}} \text{ and } \omega = 2\pi\nu = \frac{2\pi m_o c^2}{h \sqrt{1-v^2/c^2}}.$$

Therefore,

$$d\omega = \frac{2\pi m_o}{h(1-v^2/c^2)^{3/2}} v \cdot dv. \quad (20.22)$$

The wavelength of the de Broglie wave is

$$\lambda = \frac{h}{p} = \frac{h(1-v^2/c^2)^{1/2}}{m_o v} \text{ and } k = \frac{2\pi}{\lambda} = \frac{2\pi m_o v}{h(1-v^2/c^2)^{1/2}}$$

$$\therefore dk = \frac{2\pi m_o}{h} \left[(1-v^2/c^2)^{-1/2} dv + v \cdot \frac{v}{c^2} (1-v^2/c^2)^{-3/2} dv \right]$$

or

$$dk = \frac{2\pi m_o dv}{h(1-v^2/c^2)^{3/2}} \quad (20.23)$$

Dividing eq. (20.22) by (20.23), we get

$$v_g = \frac{d\omega}{dk} = v \quad (20.24)$$

Thus, the de Broglie wave group associated with an atomic particle travels with the same velocity as that of the particle itself in a non-dispersive medium.

20.10 APPLICATIONS OF DE BROGLIE WAVES

We discuss here some of the applications of de Broglie waves.

1. Possible energy states of a microparticle trapped in a box

Let us consider a microparticle trapped in a one-dimensional box of length L . According to de Broglie hypothesis, the particle is associated with a wave having a wavelength λ . The particle cannot move beyond the walls of the box. Hence, the amplitude of the de Broglie wave drops to zero at the walls. It implies that the de Broglie wave of the particle forms a standing wave pattern with nodes at the walls. The formation of standing wave pattern requires that the distance L must be an integral multiple of half-wavelength. Thus,

$$L = n \frac{\lambda}{2} \text{ or } \lambda = \frac{2L}{n} \quad (20.25)$$

where $n = 1, 2, 3, \dots$

The possible values of linear momentum are given by

$$p = \frac{h}{\lambda} = n \frac{h}{2L} \quad (20.26)$$

The possible values of the kinetic energy of the microparticle are given by

$$K.E. = \frac{p^2}{2m} = \frac{n^2 h^2}{8mL^2} \quad (20.27)$$

The above expression indicates that a microparticle trapped in a box can take only certain discrete energy states. Secondly, the particle cannot have zero energy and the minimum kinetic energy that it can take is $\frac{h^2}{8mL^2}$ when $n = 1$.

2. Neutron diffraction

Experiments have showed that neutrons exhibit diffraction. The diffraction of neutrons is used to study atomic structures of solids containing hydrogen atoms. Normally, x-ray and electron diffraction analysis are used to determine the structures of solids. In solids containing hydrogen atoms, the scattering of x-rays and electrons by hydrogen is not sufficient. Therefore, it is difficult to locate the position of hydrogen atoms using these techniques. In contrast, neutrons interact largely with atomic nuclei, notably hydrogen nuclei. Therefore, the scattering of neutrons by hydrogen-containing atoms in solids reveals the presence and location of hydrogen atoms in the lattice.

Matter waves are less penetrating than x-rays and so are useful in studying surface features of materials.

3. Electron microscope

A microscope is an optical instrument used to magnify small objects in order to study their structural details. It consists of two high power lenses called the *objective* and the *eyepiece*. The objective forms a real image of the object kept in front of it and the image is viewed by the eye through the eyepiece. The magnification of the object by the microscope is given by the product of the magnifications of the objective and eyepiece. It may appear at the first instance that one may get an image magnified to any desired extent by increasing the magnifications of the objective and eyepiece. It does not happen so in practice. The light reflected from each point of the object has to pass through the objective, which is a circular aperture. Consequently, circular diffraction is produced by the objective corresponding to each point. The higher the objective power, the smaller the aperture and the larger is the diffraction pattern. When the image is viewed further through the eyepiece of higher power, the points appear as blurred patches overlapping on each other and the details are not discernible. Thus, the diffraction effects restrict the ultimate useful magnification achievable by a microscope. The maximum useful magnification of an optical microscope is about $1000\times$. Any further magnification does not show more details though a larger image is obtained. The diffraction effects depend on the wavelength of light. The useful magnification can be increased by making use of shorter wavelength radiation. Thus, UV radiation can give higher magnification of around $2000\times$.

Very large magnification is obtained by exploiting the wave character of electrons. The de Broglie wavelengths of electrons are extremely small, of the order of 0.001 to 1 Å as against the light wavelength of 5000 Å. As a result, magnifications of the order of $10^6 \times$ can be easily attained using electron waves. The Transmission Electron Microscope (TEM) was the first type of Electron Microscope to be developed and is built exactly on the model of Light Transmission Microscope except that a focused beam of electrons is used instead of light to "see through" the specimen. It was developed by Max Knoll and Ernst Ruska in Germany in 1931.

Construction: A schematic diagram of the electron microscope is shown in Fig. 20.8. It is essentially a very large modified cathode ray tube. It consists of an electron gun at one end of the tube, a number of magnetic lenses in the path of the electron beam and a fluorescent screen at the other end of the tube. Each magnetic lens is a solenoid encased in soft iron and has a soft iron pole piece to concentrate the magnetic field lines. The focal lengths are of the order of a few millimeters and can be varied by varying current through the solenoid. The tube is mounted vertically as illustrated in Fig. 20.8.

Working: Electrons are emitted by a hot cathode and are accelerated to high velocities with the help of an anode held at about 50 to 100 kV. The electrons pass through a magnetic lens that acts as a condenser lens and is formed into a parallel beam. The electron beam then passes through the specimen to be viewed. The specimen is prepared in the form of a very thin slice of thickness of about 100 to 1000 Å so that electrons are not scattered and blur the image. Different number of electrons passes through different portions of the specimen depending on its structure. After passing through the specimen, the electron beam goes through a second lens, which acts as the objective lens and forms an intermediate image of the object. After that the beam passes through a third lens that acts as the eyepiece and forms the final image of the object on a fluorescent screen. The fluorescent screen converts the image into an optical image. The magnified image of the object is viewed through a side window. The resolving power of an electron microscope is of the order of 10 to 100 Å and the magnification of the order of 10^6 to 10^7 is easily attainable.

In view of high magnification power the electron microscope proved to be a very valuable tool for studying the microstructures of a variety of materials and also for studying micro-organisms. The structural details of virus, proteins etc could be understood only with the help of an electron microscope.

4. Scanning electron microscope (SEM)

An important variation is the scanning electron microscope. A schematic diagram of the scanning electron microscope is shown in Fig. 20.9. In this, a beam of electrons is generated in the electron gun, located at the top of the column. This beam is attracted through the anode, condensed by a condenser lens, and focused as a very fine point on the sample by the objective

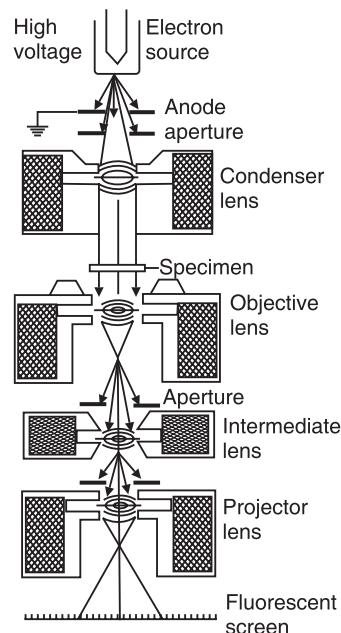


Fig. 20.8: Transmission Electron Microscope (TEM)

lens. The scan coils are energized (by varying the voltage produced by the scan generator) and create a magnetic field which deflects the beam back and forth in a controlled pattern. The varying voltage is also applied to the coils around the neck of the Cathode-ray tube (CRT) which produces a pattern of light deflected back and forth on the surface of the CRT. The pattern of deflection of the electron beam is the same as the pattern of deflection of the spot of light on the CRT. The electron beam hits the sample, producing secondary electrons from the sample. These electrons are collected by a collecting anode that is held at a positive potential with respect to the specimen. The current in the electron collector is converted to a voltage, and amplified. The amplified voltage is applied to the grid of the CRT and causes the intensity of the spot of light to change. The image consists of thousands of spots of varying intensity on the face of a CRT that correspond to the topography of the sample.

Advantages: Generally, the TEM resolution is about an order of magnitude greater than the SEM resolution. However, the advantages of SEM are as follows. (i) As the process of forming image involves surface processes, bulk samples can be used. (ii) It provides a much greater depth of view, and so can produce images that are a good representation of the 3D structure of the sample.

Example 20.1: An electron beam is accelerated from rest through a potential difference of 200 V.

- Calculate the associated wavelength.
- This beam is passed through a diffraction grating of spacing 3 \AA . At what angle of deviation from the incident direction will be the first maximum observed?

Solution: The wavelength of the waves associated with the electron beam is given by

$$\begin{aligned} \lambda &= \frac{h}{\sqrt{2meV}} \\ &= \frac{6.63 \times 10^{-34} \text{ Js}}{\sqrt{2 \times 9.11 \times 10^{-31} \text{ kg} \times 1.602 \times 10^{-19} \text{ C} \times 200 \text{ V}}} \\ &= 0.86 \text{ \AA}. \end{aligned}$$

The diffraction is governed by the equation $2d \sin \theta = m\lambda$

For first order $m = 1$ and $2d \sin \theta = \lambda$.

$$\therefore \theta = \sin^{-1} \left(\frac{\lambda}{2d} \right) = \sin^{-1} \left(\frac{0.86 \times 10^{-10} \text{ m}}{2 \times 3 \times 10^{-10} \text{ m}} \right) = 8.31^\circ.$$

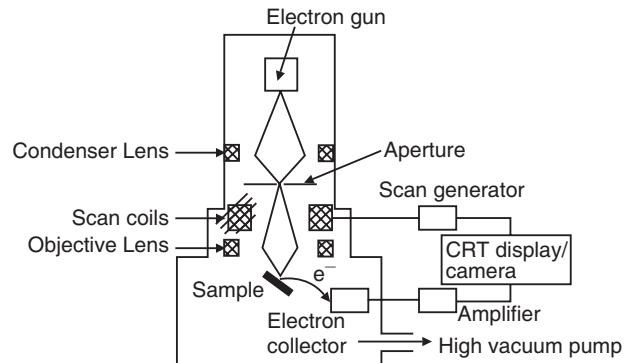


Fig. 20.9: Scanning electron microscope (SEM)

Example 20.2. An enclosure filled with helium is heated to 400K. A beam of He-atoms emerges out of the enclosure. Calculate the de Broglie wavelength corresponding to He atoms. Mass of He atom is $6.7 \times 10^{-27} \text{ kg}$.

$$\begin{aligned}\text{Solution. De Broglie wavelength } \lambda &= \frac{h}{\sqrt{2mkT}} \\ &= \frac{6.63 \times 10^{-34} \text{ Js}}{\sqrt{2 \times 6.7 \times 10^{-27} \text{ kg} \times 1.376 \times 10^{-21} \text{ J / deg} \times 400}} \\ &= 0.769 \text{ \AA}\end{aligned}$$

Example 20.3: Find the de Broglie wavelength of

- (i) an electron accelerated through a potential difference of 182 volts, and
- (ii) a 1 kg object moving with a speed 1 m/s. Comparing the results explain why the wave nature of matter is not more apparent in daily observations.

Solution:

$$\begin{aligned}(i) \lambda_e &= \frac{h}{\sqrt{2emV}} = \frac{6.626 \times 10^{-34} \text{ Js}}{\sqrt{2(1.602 \times 10^{-19} \text{ C})(9.11 \times 10^{-31} \text{ kg})182 \text{ V}}} \\ &= \frac{6.626 \times 10^{-34} \text{ Js}}{7.29 \times 10^{-24} \text{ kg.m/s}} = 0.91 \times 10^{-10} \frac{\text{kg.m}^2/\text{s}}{\text{kg.m/s}} = 9.1 \times 10^{-11} \text{ m} = 0.91 \text{ \AA} \\ (ii) \lambda_m &= \frac{h}{Mv} = \frac{6.626 \times 10^{-34} \text{ Js}}{1 \text{ kg} \times 1 \text{ m/s}} = 6.6 \times 10^{-34} \frac{\text{kg.m}^2/\text{s}}{\text{kg.m/s}} = 6.6 \times 10^{-34} \text{ m}.\end{aligned}$$

It is seen from the above that the wavelength of the accelerated electron is about 10^5 times larger than its own size ($\approx 10^{-15} \text{ m}$) and is therefore significant. On the other hand, the wavelength associated with the macroscopic object is negligibly small and is thus not apparent in its interactions with other objects.

20.11 HEISENBERG UNCERTAINTY PRINCIPLE

The wave nature of atomic particles leads to some inevitable consequences. Classically, the state of a particle can be defined by specifying its position and momentum at any given time t . If a body is moving along x -direction with a velocity v , its position is given by $x = vt$ and its momentum by $p = mv$. From this,

$$x = \frac{p}{m}t \quad (20.28)$$

At each instant, the position and momentum can be measured to a very high accuracy. When an atomic particle is conceptualized as a de Broglie wave packet such a precision becomes restricted.

Schrödinger postulated that a moving microparticle is equivalent to a wave packet. A wave packet spreads over a region of space. Therefore, it is difficult to locate the exact position of the microparticle. Although the particle is somewhere within the wave packet, it is impossible to know where exactly the particle is at a given instant. If the linear spread of the wave packet is Δx , the particle would be located somewhere within the region Δx . The probability of finding the particle is a maximum at the centre of the wave packet and falls off to zero at its ends. Therefore, there is an **uncertainty** Δx in the position of the particle. As a

result, the momentum of the particle at that instant cannot be determined precisely. It means that the location and momentum of a microparticle cannot be **simultaneously** determined with certainty. Any attempt to determine these variables will lead to uncertainties in each of the variables.

In 1927 Heisenberg showed that the product of uncertainty Δx in the x -coordinate of a quantum particle and the uncertainty Δp_x in the x -component of the momentum would always be of the order of Planck's constant \hbar . Thus,

$$\Delta x \cdot \Delta p_x \approx \hbar$$

or more precisely

$$\Delta x \cdot \Delta p_x \geq \frac{\hbar}{2} \quad (20.29)$$

This is known as Heisenberg's uncertainty principle for position and momentum, which may be stated as follows:

"It is not possible to know simultaneously and with exactness both the position and the momentum of a microparticle".

The Uncertainty Principle implies a built-in, unavoidable *limit to the accuracy with which we can make measurements*. Classically, it is thought that the precision of any measurement was limited only by the accuracy of the instruments the experimenter used. Heisenberg showed that whatever may be the accuracy of the instruments used, quantum mechanics limits the precision when two properties are measured at the same time. These are *not just any two properties* but pairs of measurable quantities whose product has dimensions of energy \times time. Such quantities are called *conjugate quantities* in quantum mechanics, and have a special relation to each other. Position-linear momentum, energy-time, time-frequency and angular momentum-angular displacement are conjugate pairs of variables.

The uncertainty principle asserts that it is physically impossible to know simultaneously the exact position ($\Delta x = 0$) and exact momentum ($\Delta p_x = 0$) of a micro-particle. According to it, the more precisely we know the position of the particle, the less precise is our information about its momentum. To localize a wave packet, we have to add more wavelengths to form the wave packet. More wavelengths mean larger $\Delta\lambda$ and more uncertainty in momentum (note that $\Delta p \propto \Delta\lambda$). Conversely, in order to have more precise value of momentum, the wave packet should contain less number of waves. Less number of waves produces a longer wave packet. Thus, the momentum of a particle cannot be precisely specified without our loss of knowledge of the position of the particle at that time. Similarly, a particle cannot be precisely localized in a particular direction without our loss of knowledge of momentum in that particular direction. We can at best specify that certain momentum of the particle is more probable than the other or that the particle is more likely to be here than there. We cannot use classical notions like coordinates and momentum to describe the motion of quantum particles. Thus, the uncertainty principle implies that we can never define the path of an atomic particle with the absolute precision indicated in classical mechanics. Therefore, concepts such as velocity, position, and acceleration are of limited use in quantum world.

Relations similar to (20.29) hold good for other components of position and linear momentum. Thus,

$$\Delta y \cdot \Delta p_y \geq \frac{\hbar}{2} \quad (20.29a)$$

$$\Delta z \cdot \Delta p_z \geq \frac{\hbar}{2} \quad (20.29b)$$

20.11.1 Energy - Momentum Uncertainty

The uncertainty relation for the simultaneous measurement of energy E and time t is expressed as

$$\Delta E \cdot \Delta t \geq \frac{\hbar}{2} \quad (20.30)$$

The physical significance of the energy-time uncertainty relation is different from that of the position-momentum uncertainty. If ΔE is the maximum uncertainty in the determination of the energy of a particle, then the minimum time interval for which the particle remains in that state is given by

$$\Delta t = \frac{\hbar / 2}{\Delta E}$$

And, if a particle remains in a particular energy state for a maximum time Δt , then the minimum uncertainty in the particle energy is given by

$$\Delta E = \frac{\hbar / 2}{\Delta t}$$

Derivation: We can obtain the result (20.30) as follows. Let us consider a microparticle of mass m moving with a velocity v . Its kinetic energy will be

$$E = \frac{1}{2}mv^2$$

If the uncertainty in the energy is ΔE , then $\Delta E = \Delta \left[\frac{1}{2}mv^2 \right] = mv \Delta v = v \Delta p$

As the velocity $v = \frac{\Delta x}{\Delta t}$, the uncertainty in energy may be written as $\Delta E = \frac{\Delta x}{\Delta t} \Delta p$

$$\begin{aligned} \text{Thus,} \quad & \Delta E \cdot \Delta t = \Delta x \cdot \Delta p \\ \text{But} \quad & \Delta x \cdot \Delta p \geq \frac{\hbar}{2} \\ \text{Therefore,} \quad & \Delta E \cdot \Delta t \geq \frac{\hbar}{2} \end{aligned}$$

The above relations are to be supplemented by the following uncertainty relation

$$\Delta M_x \cdot \Delta \varphi_x \geq \frac{\hbar}{2} \quad (20.31)$$

where ΔM_x is the uncertainty in the projection of the angular momentum on the x-axis and $\Delta \varphi_x$ is the uncertainty in the angular coordinates of the microparticle.

By analogy with (20.28a) and (20.28b), we may write down relations for other projections of momentum and angular momentum as follows:

$$\Delta M_y \cdot \Delta \varphi_y \geq \hbar/2 \quad \text{and} \quad \Delta M_z \cdot \Delta \varphi_z \geq \hbar/2 \quad (20.31a)$$

In general if q and p denote two canonically conjugate variables, the uncertainty relation is given by

$$\Delta q \cdot \Delta p \geq \hbar/2 \quad (20.32)$$

The above relations do not mean that the uncertainty principle creates certain obstacles to the understanding of the atomic phenomena; it only reflects certain peculiarities of the objective properties of a quantum particle.

20.12 ELEMENTARY PROOF OF UNCERTAINTY PRINCIPLE USING DE BROGLIE WAVE CONCEPT

A wave packet produced by a superposition of large number of harmonic waves is shown in Fig. 20.10. Since a wave packet is not an infinite harmonic wave, it has a range of wave numbers Δk instead of one definite wave number.

Δk is thus the uncertainty in wave number. Further, the position of the particle cannot be given with certainty. It will lie somewhere between the two consecutive nodes (see Fig. 20.7b and Fig. 20.10). Thus, the uncertainty in the position of the particle is equal to the distance between two consecutive nodes. Referring to equ. (20.15), the condition for formation of a node is

$$\cos\left(\frac{\Delta\omega t}{2} - \frac{\Delta k x}{2}\right) = 0$$

$$\text{i.e., } \frac{\Delta\omega}{2}t - \frac{\Delta k}{2}x = \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \dots, (2n+1)\frac{\pi}{2}$$

Thus, if x_1 and x_2 are the positions of two consecutive nodes, then

$$\frac{\Delta\omega}{2}t - \frac{\Delta k}{2}x_1 = (2n+1)\frac{\pi}{2}$$

$$\text{and } \frac{\Delta\omega}{2}t - \frac{\Delta k}{2}x_2 = (2n+3)\frac{\pi}{2}$$

Subtracting the above upper equation from the lower one, we find that

$$\frac{\Delta k}{2}(x_2 - x_1) = \pi \quad \text{or} \quad \frac{\Delta k}{2}\Delta x = \pi$$

where $\Delta x = (x_2 - x_1)$ is the uncertainty in the position of the particle. Thus,

$$\Delta x = \frac{2\pi}{\Delta k}.$$

Since

$$k = \frac{2\pi}{\lambda} = \frac{2\pi p}{h},$$

$$\Delta k = \frac{2\pi}{h} \Delta p$$

Therefore,

$$\Delta x = \frac{2\pi}{\Delta k} = \frac{2\pi h}{2\pi \Delta p} = \frac{h}{\Delta p}$$

or

$$\Delta x \cdot \Delta p = h$$

When we consider a group consisting of very large number of harmonic waves of continuously varying frequencies, the product of the uncertainties comes to

$$\Delta x \cdot \Delta p \geq \frac{1}{2}\hbar$$

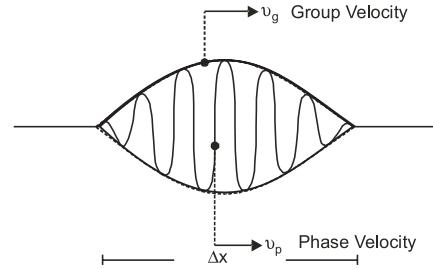


Fig. 20.10

20.13 IMPLICATION OF UNCERTAINTY PRINCIPLE

The uncertainty principle expresses a fundamental limitation in nature that also limits the precision of our measurements. According to classical mechanics the position and momentum

of a macro-particle can be determined exactly. But the uncertainty principle asserts that it is physically impossible to know simultaneously the exact position ($\Delta x = 0$) and exact momentum ($\Delta p_x = 0$) of a microparticle. According to it, the more precisely we know the position of the particle, the less precise is our information about its momentum. Thus, the momentum of a particle cannot be precisely specified without our loss of knowledge of the position of the particle at that time. Similarly, a particle cannot be precisely localized in a particular direction without our loss of knowledge of momentum in that particular direction. We can at best specify that certain momentum of the particle is more probable than the other or that the particle is more likely to be here than there. It means that our classical notions like coordinates and momentum derived from ordinary macroscopic experiences are inadequate to describe the atomic world. The uncertainty principle points out that in the microscopic world,

- (1) the dynamical variables of a particle are combined in sets of *simultaneously determined* quantities which are known as *complete sets* of quantities;
- (2) the coordinate and momentum components of a particle etc are **pairs of concepts** which are interrelated and fall in different complete sets of quantities. They cannot be defined simultaneously in a precise way.

Thus, the uncertainty principle implies that we can never define the path of an atomic particle with the absolute precision indicated in classical mechanics. Therefore, concepts such as velocity, position, and acceleration are of limited use in quantum world. To describe the quantum particle the concept of *energy* becomes important since it is related to the *state* of the system rather than to its path.

20.14 UNCERTAINTY PRINCIPLE IS NOT SIGNIFICANT IN CASE OF MACRO-BODIES

The Heisenberg Principle is of no practical importance for heavy bodies where the de Broglie wavelength is negligibly small.

For example, let us take the case of a cricket ball in flight. The indeterminacy in the position of the ball is, say, 1 mm. We can determine the indeterminacy of velocity of the ball from uncertainty principle.

$$\begin{aligned} \Delta x \cdot \Delta p &\approx h \\ \therefore \Delta x \cdot m\Delta v &\approx h \\ \Delta v &\approx \frac{h}{m\Delta x} = \frac{6.62 \times 10^{-34} J.s}{0.5 kg \times 10^{-3} m} \approx 10^{-30} m/s. \end{aligned}$$

The above inaccuracy is negligible and not detectable. It implies that the uncertainties are of no importance in case of macro bodies; and the position and velocity of a macro body can be simultaneously determined with a high degree of accuracy. As a result, macroscopic body follows a well defined trajectory.

In contrast if we take the example of an electron orbiting in a hydrogen atom, the inaccuracy in its position is $\pm 1\text{\AA}$. The uncertainty in its speed is

$$\Delta v = \frac{h}{m\Delta x} = \frac{6.62 \times 10^{-34} J.s}{9.11 \times 10^{-31} kg \times 2 \times 10^{-10} m} \approx 2 \times 10^5 m/s$$

which is of the same order as the velocity of the electron in the orbit. It means that it is not possible to determine the velocity and the position of a microparticle with certainty and as such we cannot talk of a specific trajectory. Instead we have to be content knowing only the probable values.

20.15 THOUGHT EXPERIMENTS

It is not possible to verify Heisenberg uncertainty principle in the laboratory. Therefore, we illustrate it with the help of two *thought experiments*.

1. Treating the electron as a wave

We assume that an electron has wave character. Suppose that we want to determine the y -coordinate of an electron moving along the x -axis. We place a slit of width ' d ' perpendicular to the direction of motion of the electron (Fig. 20.11). The precision of the position of electron, Δy , in y -direction is limited by the size of the slit. That is, $\Delta y \approx d$.

If the slit is narrow enough, it causes a change in the motion of the electron after going through the slit and brings out the wave character of electron as evidenced from the diffraction pattern observed on the screen. The uncertainty in the electron momentum parallel to y -axis depends on the diffraction angle θ . According to the theory of diffraction at a single slit, the angle θ is given by

$$\sin \theta = \frac{h}{\lambda d} \quad (20.33)$$

The uncertainty in the momentum of the electron parallel to y -axis is given by

$$\Delta p_y = p \sin \theta = \frac{h \lambda}{\lambda d} = \frac{h}{d} \approx \frac{h}{\Delta y}$$

$$\therefore \Delta y \Delta p_y \approx h \quad (20.34)$$

If we wish to determine the exact position of the electron along the y -axis, we have to use a very narrow slit. However, a very narrow slit produces a wider diffraction pattern, which leads to a larger uncertainty in our knowledge of the y -component of the momentum. Conversely, if we attempt to reduce the uncertainty in our knowledge of y -component of momentum, the diffraction pattern should be very narrow. Therefore, we have to use a very wide slit, which in turn results in a large uncertainty in the y -coordinate of the electron. Thus, our efforts to simultaneously reduce the uncertainties Δy and Δp will get frustrated.

2. Treating the Electron as a Particle

We now consider that electron is a particle and attempt to measure its position. For the sake of convenience, we assume that the electron is at rest and use a microscope to locate it. We cannot use an optical microscope for this purpose. The act of observing an object requires that light from a source be reflected by the object and enter a recording instrument such as eye. An object reflects a wave when the size of the object is sufficiently larger than the wavelength of the wave. In the present case, if we use light waves, they would pass on without getting reflected by the electron, since the size of the electron

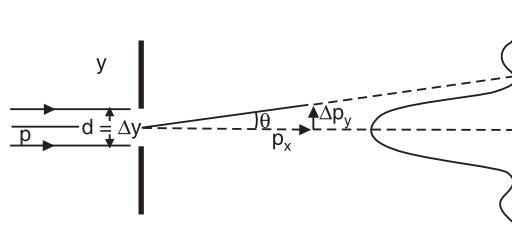


Fig. 20.11

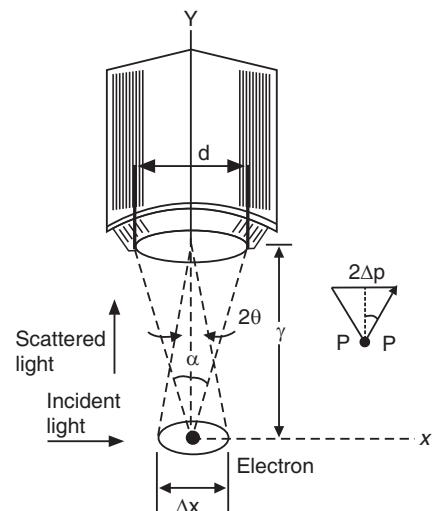


Fig. 20.12: Measurement of position and momentum of an electron by means of a γ -microscope.

is about 10^5 times smaller than the wavelength of light. Therefore, we use a γ -ray microscope to detect and locate an electron.

Let a free electron be directly beneath the center of the γ -ray microscope's lens. The circular lens forms a cone of angle 2α from the electron. The electron is illuminated from the left by γ -rays. The microscope can resolve objects to a size of Δx . Δx is given by the expression

$$\Delta x = \frac{\lambda}{2 \sin \alpha}$$

To be observed by the microscope, the γ -ray must be scattered into any angle within the cone of angle 2α . A γ -photon carries a very large momentum. When the γ -photon strikes the electron, part of the momentum and energy are transferred to the electron due to Compton scattering. Consequently, as the scattered photon enters the microscope, the electron has already moved away in a certain direction (Fig. 20.12). The total momentum p is related to the wavelength by the formula $p = \frac{h}{\lambda}$.

In the extreme case of diffraction of the gamma ray to the right edge of the lens, the total momentum in the x direction would be the sum of the electron's momentum p'_x in the X direction and the gamma ray's momentum in the x -direction:

$$p'_x + \left(\frac{h \sin \alpha}{\lambda'} \right) \quad (20.36)$$

where λ' is the wavelength of the deflected gamma ray. In the other extreme, the observed gamma ray recoils backward, just hitting the left edge of the lens. In this case, the total momentum in the X-direction is:

$$p''_x - \left(\frac{h \sin \alpha}{\lambda''} \right) \quad (20.37)$$

The final X-momentum in each case must equal the initial X-momentum, since momentum is *conserved*. Therefore, the final X-momenta are equal to each other:

$$p'_x + \left(\frac{h \sin \alpha}{\lambda'} \right) = p''_x - \left(\frac{h \sin \alpha}{\lambda''} \right)$$

If α is small, then the wavelengths are approximately the same, $\lambda' \approx \lambda'' \approx \lambda$. And we have

$$p''_x - p'_x = \frac{2h \sin \alpha}{\lambda}$$

or

$$\Delta p_x = \frac{2h \sin \alpha}{\lambda} \quad (20.38)$$

Using eq.(20.35) into the above equation, we obtain

$$\Delta p_x = \frac{h}{\Delta x}$$

\therefore

$$\Delta x \cdot \Delta p_x = h$$

20.16 APPLICATIONS OF UNCERTAINTY PRINCIPLE

We deal here with three simple examples to illustrate the application of uncertainty principle.

(a) Bohr's Orbit and Energy

Let us consider the electron in a hydrogen atom. We cannot know at any instant the position of the electron in its orbit. It might be on the left or right of the nucleus, as sketched in

Fig. 20.13. The electron position has an uncertainty $\pm r$. We cannot know likewise whether the electron is moving upward or downward. The uncertainty in its velocity therefore $\pm v$. Taking $\Delta x = r \approx 0.5 \times 10^{-10}$ m, the uncertainty in the electron speed is

$$\begin{aligned}\Delta v &= \frac{\hbar}{2\pi m \Delta x} \\ &= \frac{6.62 \times 10^{-34} \text{ J.s}}{2 \times 3.124 \times 9.11 \times 10^{-31} \text{ kg} \times 0.5 \times 10^{-10} \text{ m}} \\ &\approx 2 \times 10^6 \text{ m/s}\end{aligned}$$

The velocity 'v', of an electron in an atom is of the order of 1.0×10^6 m/s and is of the same order as the uncertainty Δv . Therefore, we conclude that the uncertainty in momentum is of the same order as the momentum. That is $\Delta p \approx p$. It means that sharp position and momentum do not exist simultaneously for the electron in an atom. Hence it is not possible to ascribe any specific trajectory to an electron in an atom. It can only be said that atomic electrons traverse the whole of the space about the nucleus, but however, they move most of the time at a distance corresponding to a permitted Bohr radius.

Now let us calculate the energy of the Bohr's first orbit. The total energy of the electron in the first orbit is given by

$$\begin{aligned}E &= \text{K.E.} + \text{P.E.} \\ &= \left[\frac{1}{2} mv^2 \right] - \frac{e^2}{4\pi\epsilon_0 r} = \left[\frac{p^2}{2m} \right] - \frac{e^2}{4\pi\epsilon_0 r} \quad (20.39)\end{aligned}$$

where p is the momentum of the electron.

$$\text{As } \Delta p \approx p, \text{ we can write } p \approx \frac{\hbar}{2\pi\Delta x} = \frac{\hbar}{2\pi r}. \quad (20.40)$$

$$\therefore E = \frac{\hbar^2}{8\pi^2 mr^2} - \frac{e^2}{4\pi\epsilon_0 r} \quad (20.41)$$

$$\text{But } r \text{ is given by } r = \frac{\epsilon_0 h^2}{\pi me^2}.$$

$$\therefore E = -\frac{me^4}{8\epsilon_0^2 h^2} \quad (20.42)$$

The above expression (20.42) is the same as that is given by Bohr theory.

(b) Particle in a Box:

Let us consider a particle confined to a box of length l . The uncertainty Δx in the position is l .

$$\begin{aligned}\Delta x \cdot \Delta p &\approx \hbar \\ \therefore \Delta p &= \frac{\hbar}{\Delta x} = \frac{\hbar}{l} \quad (20.43)\end{aligned}$$

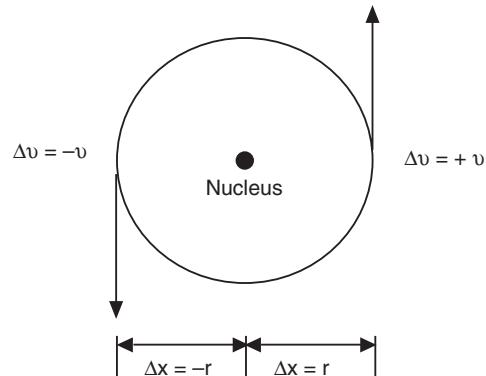


Fig. 20.13: The uncertainties in the position and velocity of electron in an atom.

Energy is given by $E = \frac{p^2}{2m} \approx \frac{(\Delta p)^2}{2m} = \frac{(\hbar/l)^2}{2m} = \frac{\hbar^2}{2ml^2}$ (20.44)

This result agrees with the result obtained from Schrödinger equation. Refer to § 20.95.

(c) Electrons cannot be present in the nucleus:

The radiation emitted by radioactive nuclei consists of α , β and γ -rays, out of which β -rays are identified to be electrons. We apply uncertainty principle to find whether electrons are coming out of the nucleus. The radius of the nucleus is of the order of 10^{-14} m. Therefore, if electrons were to be in the nucleus, the maximum uncertainty Δx in the position of the electron is equal to the diameter of the nucleus. Thus,

$$\Delta x = 2 \times 10^{-14} \text{ m.}$$

The minimum uncertainty in its momentum is then given by

$$\Delta p = \frac{\hbar}{\Delta x} = \frac{1.04 \times 10^{-34} \text{ J.s}}{2 \times 10^{-14} \text{ m}} = 5.2 \times 10^{-21} \text{ kg-m/s.}$$

The minimum uncertainty in momentum can be taken as the momentum of the electron. Thus,

$$p = 5.2 \times 10^{-21} \text{ kg-m/s.}$$

The minimum energy of the electron in the nucleus is then given by

$$E_{\min} = p_{\min} c = (5.2 \times 10^{-21} \text{ kg-m/s})(3 \times 10^8 \text{ m/s}) = 1.56 \times 10^{-12} \text{ J} = 9.7 \text{ MeV.}$$

It implies that if an electron exists within the nucleus, it must have a minimum energy of about 10 MeV. But the experimental measurements showed that the maximum kinetic energies of β -particles were of the order of 4 MeV only. Hence electrons are not present in the nucleus. It is subsequently established that emission of β -particles occurs due to transformations in the nucleus. The transformation of a neutron into a proton produces an electron.

Example 20.4: Uncertainty in time of an excited atom is about 10^{-8} s. What are the uncertainties in energy and in frequency of the radiation?

Solution:

$$\Delta E \Delta t \approx \frac{\hbar}{2\pi}$$

$$\therefore \Delta E = \frac{1.054 \times 10^{-34} \text{ J.s}}{10^{-8} \text{ s}} = \frac{1.054 \times 10^{-26}}{1.602 \times 10^{-19}} \text{ eV} = 6.58 \times 10^{-8} \text{ eV.}$$

$$\Delta v = \frac{\Delta E}{h} = \frac{1.054 \times 10^{-26} \text{ J}}{6.626 \times 10^{-34} \text{ J.s}} = 15.9 \text{ MHz.}$$

Example 20.5. An electron is confined to a potential well of width 10 nm. Calculate the minimum uncertainty in its velocity.

Solution.

$$\Delta x \Delta p \approx \frac{\hbar}{2\pi} \quad \text{or} \quad \Delta x \Delta p \approx \frac{\hbar}{2\pi}$$

$$\therefore \Delta v = \frac{\hbar}{2\pi m \cdot \Delta x}$$

$$\therefore \Delta v = \frac{6.63 \times 10^{-34} \text{ J.s}}{2 \times 3.143 \times 9.11 \times 10^{-31} \text{ kg} \times 10 \times 10^{-9} \text{ m}} = 12.1 \text{ km/s.}$$

Example 20.6: If the kinetic energy of an electron known to be about 1 eV, must be measured to within 0.0001 eV, what accuracy can its position be measured simultaneously?

Solution: $E = \frac{p^2}{2m}$ $\therefore \Delta E = \frac{2p \Delta p}{2m}$ $\therefore \Delta p = \frac{m}{p} \Delta E$

$$\Delta x \Delta p = \frac{\hbar}{2\pi}$$

$$\therefore \Delta x = \frac{h}{2\pi \cdot \Delta p} = \frac{h}{2\pi} \cdot \frac{p}{m \Delta E} = \frac{h \sqrt{2mE}}{2\pi m \Delta E} = \frac{h}{\pi \Delta E} \sqrt{\frac{E}{2m}}$$

$$\therefore \Delta x = \frac{6.63 \times 10^{-34} \text{ Js}}{3.143 \times 0.0001 \times 1.602 \times 10^{-19} \text{ J}} \cdot \sqrt{\frac{1.602 \times 10^{-19} \text{ J}}{2 \times 9.11 \times 10^{-31} \text{ kg}}}$$

$$= 1.95 \mu\text{m}.$$

Example 20.7: An electron and a 150 gm base ball are traveling at a velocity of 220 m/s, measured to an accuracy of 0.005 %. Calculate and compare uncertainty in position of each.

Solution: The uncertainty in the velocity is $\Delta v = v \times 0.065\% = (220 \text{ m/s}) \times \frac{0.065}{100} = 0.143 \text{ m/s}$.

(i) The uncertainty in the position of electron is

$$\Delta x_e = \frac{\hbar}{2 m \Delta v} = \frac{1.05 \times 10^{-34} \text{ J.s}}{2 \times 9.11 \times 10^{-31} \text{ kg} \times 0.143 \text{ m/s}} = 0.4 \text{ mm.}$$

(ii) The uncertainty in the position of baseball is

$$\Delta x_B = \frac{\hbar}{2 M \Delta v} = \frac{1.05 \times 10^{-34} \text{ J.s}}{2 \times 0.15 \text{ kg} \times 0.143 \text{ m/s}} = 2.5 \times 10^{-33} \text{ m.}$$

20.17 WAVE FUNCTION AND PROBABILITY INTERPRETATION

Waves represent the propagation of a disturbance in a medium. We are familiar with light waves, sound waves, and water waves. These waves are characterized by some quantity that varies with position and time. Light waves consist of variations of electric and magnetic fields in space, and sound waves consist of pressure variations. We cannot specify in a similar way what is actually varying in de Broglie waves. Since microparticles exhibit wave properties, it is assumed that a quantity ψ represents a de Broglie wave. This quantity ψ is called a **wave function**. ψ describes the wave as a function of position and time. However, it has no direct physical significance, as it is not an *observable* quantity. In general, ψ is a complex-valued function. According to Heisenberg uncertainty principle, we can only know the probable value in a measurement. The probability cannot be negative. Hence ψ cannot be a measure of the presence of the particle at the location (x, y, z) . But it is certain that it is in somehow an index of the presence of the particle at around (x, y, z, t) .

Probability Interpretation of Wave Function given by Max Born

A probability interpretation of the wave function was given by Max Born in 1926. He suggested that *the square of the magnitude of the wave function $|\psi|^2$ evaluated in a particular region represents the probability of finding the particle in that region*. In other words,

Probability, P , of finding the particle in an infinitesimal volume $dV (= dx dy dz)$ is proportional to $|\psi(x, y, z)|^2 dx dy dz$ at time t .

$$\text{or } P \propto |\psi(x, y, z)|^2 dV \quad (20.45)$$

$|\psi|^2$ is called the **probability density** and ψ is the **probability amplitude**.

Since the particle is certainly somewhere in the space, the probability $P = 1$ and the integral of $|\psi|^2 dV$ over the entire space must be equal to unity. That is

$$\int_{-\infty}^{+\infty} |\psi|^2 dV = 1 \quad (20.46)$$

The wave function ψ is in general a complex function. But the probability must be real. Therefore to make probability a real quantity, ψ is to be multiplied by its complex conjugate ψ^* .

$$\int_{-\infty}^{+\infty} \psi \psi^* dV = 1$$

Thus, ψ has no physical significance but $|\psi|^2$ gives the probability of finding the atomic particle in a particular region.

20.17.1 Normalization Condition

If at all the particle exists, it must be found somewhere in the universe. Since we are sure that the particle must be somewhere in space, the sum of the probabilities over all values of x, y, z must be unity. If $\Psi(x, y, z, t)$ is multiplied by a constant C such that $\Psi_N(x, y, z, t) = C \Psi(x, y, z, t)$, where $\Psi_N(x, y, z, t)$ satisfies the relation

$$\int_{-\infty}^{\infty} |\Psi_N(x, y, z, t)|^2 dx dy dz = |C|^2 \int_{-\infty}^{\infty} |\Psi(x, y, z, t)|^2 dx dy dz = 1 \quad (20.47)$$

then $\Psi_N(x, y, z, t)$ is said to be **normalized** wave function and C the **normalization constant**. This condition (20.47) is known as the *normalization condition*. From equ.(20.47), we have

$$|C|^2 = \frac{1}{\int_{-\infty}^{\infty} |\Psi(x, y, z, t)|^2 dx dy dz} \quad (20.48)$$

$|\Psi_N(x, y, z, t)|^2 dx dy dz$ is called the **probability density**.

Whenever wave functions are normalised, $|\psi|^2 dV$ equals the probability that a particle will be found in an elemental volume dV .

Thus,

$$\text{Probability } P = |\psi(x, y, z)|^2 dV \quad (20.49)$$

20.17.2 Well-behaved Wave Functions - Conditions to be satisfied by ψ -function

An acceptable wave function ψ must be normalized and fulfill the following requirements:

- (i) **ψ function must be finite:** The wave function must be finite everywhere. Even if $x \rightarrow \infty$ or $-\infty$, $y \rightarrow \infty$ or $-\infty$, $z \rightarrow \infty$ or $-\infty$, the wave function should not tend to infinity. It must remain finite for all values of x, y, z . If ψ is infinite, it would imply an infinitely large probability of finding the particle at that point. This would violate the uncertainty principle.

(ii) **ψ function must be single-valued:** Any physical quantity can have only one value at a point. For this reason, the function related to a physical quantity cannot have more than one value at that point. If it has more than one value at a point, it means that there is more than one value of probability of finding the particle at that point (Fig. 20.14).

(iii) **ψ function must be continuous:** ψ function should be continuous across any boundary. Since ψ is related to a physical quantity, it cannot have a discontinuity at any point. Therefore, the wave function ψ and its space derivatives $\frac{\partial\psi}{\partial x}$, $\frac{\partial\psi}{\partial y}$ and $\frac{\partial\psi}{\partial z}$ should be continuous across any boundary. Since ψ is related to a real particle, it cannot have a discontinuity at any boundary where potential changes.

Wave functions satisfying the above mathematical conditions are called *well-behaved wave functions*.

20.18 SCHRÖDINGER WAVE EQUATION

In 1926, Erwin Schrödinger, an Austrian physicist, reasoned that the de Broglie waves associated with electrons would resemble the classical waves of light and developed wave equation that describes the behaviour of matter waves. This equation, which has been named after him, defines the wave properties of electrons and also predicts particle-like behaviour.

Time dependent Schrödinger Wave Equation

Since the concept of de Broglie waves is not a result of previous physical theories, it is not possible to derive a wave equation for the wave associated with a particle. However, a wave equation can be developed. Let us assume that a particle of mass ' m ' is in motion along the x -direction. Let the wave function ψ be the dependent variable of the de Broglie wave which is a function of the coordinates x and t . Analogous to the classical wave, we may expect that, ψ will be a function of $(x - vt)$. As $v = \omega / k$, the wave function may be written as a function of $(kx - \omega t)$. Using the relation $p = \hbar k$ and $E = \hbar\omega$, we can write

$$\psi = f\left(\frac{px - Et}{\hbar}\right) \quad (20.50)$$

The more general wave would be a sum of a sine and cosine waves. Taking help of Euler's identity, we write the above equation in an exponential form as follows:

$$\psi = A \exp\left[\frac{In}{\hbar}(px - Et)\right] \quad (20.51)$$

We assume that the energy and momentum of the particle are constant. Differentiating the above equation with respect to x , we get

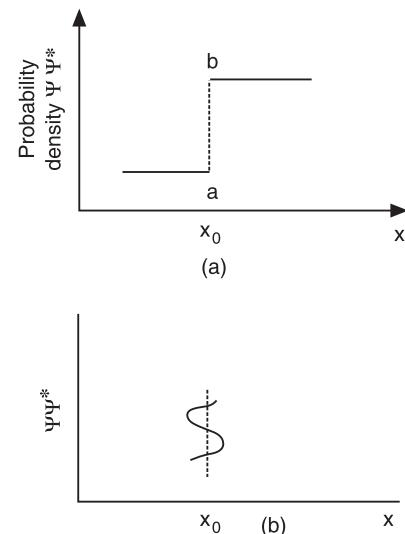


Fig. 20.14

$$\begin{aligned}\frac{\partial \psi}{\partial x} &= A \frac{ip}{\hbar} \exp \left[\frac{In}{\hbar}(px - Et) \right] \\ \therefore \frac{\partial \psi}{\partial x} &= \frac{ip}{\hbar} \psi\end{aligned}$$

Rearranging the terms in the above equation, we get

$$p\psi = \frac{\hbar}{i} \frac{\partial \psi}{\partial x} = -i\hbar \frac{\partial \psi}{\partial x} \quad (20.52)$$

Differentiating the equ.(20.51) with respect to t gives

$$\frac{\partial \psi}{\partial t} = -\frac{iE}{\hbar} A \exp \left[\frac{In}{\hbar}(px - Et) \right] = -\frac{iE}{\hbar} \psi$$

Rearranging the terms in the above equation, we get

$$E\psi = -\frac{\hbar}{i} \frac{\partial \psi}{\partial t} = i\hbar \frac{\partial \psi}{\partial t} \quad (20.53)$$

The partial derivatives with respect to x and t are connected by means of the relation between the energy and momentum. The classical expression for the kinetic energy in terms of the momentum is

$$E_k = \frac{mv^2}{2} = \frac{(mv)^2}{2m} = \frac{p^2}{2m} \quad (20.54)$$

The total energy and momentum are related by the expression

$$\frac{p^2}{2m} + V = E \quad (20.55)$$

where V is the potential energy of the particle. Multiplying the equ.(20.55) with ψ , we obtain

$$\frac{p^2}{2m} \psi + V \psi = E \psi$$

Using the relations (20.52) and (20.53) into the above equation, we get

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + V \psi = i\hbar \frac{\partial \psi}{\partial t}$$

The above equation may be rewritten as

$$i\hbar \frac{\partial \Psi}{\partial t} = \left[-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V \right] \Psi \quad (20.56)$$

The above equation is known as the **time-dependent Schrödinger wave equation**.

When extended to the three-dimensional case, we find that

$$-\frac{\hbar^2}{2m} \left(\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} + \frac{\partial^2 \Psi}{\partial z^2} \right) + V \Psi = i\hbar \frac{\partial \Psi}{\partial t}$$

$$-\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \Psi + V \Psi = i\hbar \frac{\partial \Psi}{\partial t}$$

or
$$-\frac{\hbar^2}{2m} \nabla^2 \Psi + V \Psi = i\hbar \frac{\partial \Psi}{\partial t}$$

where

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

Therefore, the Schrödinger equation for three dimensional motion may be written as

$$i\hbar \frac{\partial}{\partial t} \Psi(r, t) = \left[-\frac{\hbar^2}{2m} \nabla^2 + V(r) \right] \Psi(r, t) \quad (20.57)$$

Time independent Schrödinger Wave Equation

Knowing the form of V , eqn.(20.56) can be solved for the wave function ψ . In a number of cases the potential energy V of a particle does not depend on time; it varies with the position of the particle only and the field is said to be **stationary**. In the stationary problems Schrödinger equation can be simplified by separating out time and position-dependent parts. Accordingly, we can write the wave function as a product of x , $\psi(x)$ and a function of t , $\phi(t)$.

We, therefore, write that

$$\psi(x, t) = \psi(x)\phi(t) \quad (20.58)$$

Equation (20.56) may be written as

$$\begin{aligned} -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} (\psi, \phi) + V\psi\phi &= i\hbar \frac{\partial}{\partial t} (\psi, \phi) \\ -\frac{\hbar^2}{2m} \phi \frac{\partial^2}{\partial x^2} (\psi, \phi) + V\psi\phi &= i\hbar \frac{\partial}{\partial t} (\psi, \phi) \end{aligned}$$

Dividing the above equation with $\psi\phi$, we get

$$-\frac{\hbar^2}{2m} \frac{1}{\psi} \frac{d^2\psi}{dx^2} + V = i\hbar \frac{1}{\phi} \frac{d\phi}{dt} \quad (20.59)$$

If we assume that the potential energy V is a function of x only, the entire left hand side of equ.(20.59) is a function of x only while the right hand side is a function of t only. Since x and t are independent variables, both the function of x and t must be equal to a constant. The constant that each side must equal is called the separation constant E . Thus,

$$-\frac{\hbar^2}{2m} \frac{1}{\psi} \frac{d^2\psi}{dx^2} + V = E \quad (20.60)$$

and

$$i\hbar \frac{1}{\phi} \frac{d\phi}{dt} = E$$

Eq.(20.60) may be rewritten as

$$-\frac{\hbar^2}{2m} \frac{d^2\psi}{dx^2} + V\psi = E\psi \quad (20.61)$$

The above equation involves only the space coordinates and is called the **time-independent Schrödinger wave equation**.

Eq. (20.61) may be rewritten as

$$\begin{aligned} \frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2} (E - V)\psi &= 0 \\ \therefore \frac{d^2\psi}{dx^2} + \frac{8\pi^2 m}{\hbar^2} (E - V)\psi &= 0 \end{aligned} \quad (20.61a)$$

In three dimensions, equation (20.61) may be written as

$$-\frac{\hbar^2}{2m} \nabla^2 \psi + V\psi = 0$$

20.18.1 Allowed Wave Functions and Energies

The time-independent wave equation (20.61) is the pertinent equation for studying properties of atomic systems in stationary conditions. Wave mechanical methods of solving the problem of particle motion are essentially based on ψ functions. Appropriate wave equation is formulated by incorporating the particle mass ' m ' and potential energy function V for the region in which the particle is located. The next step consists of solving the differential equation for solutions, namely for the ψ functions which will satisfy the differential equation. By solving the Schrödinger equation, we obtain the possible set of ψ functions. In case of bound particles the acceptable solutions for the differential equation are possible only for certain specified values of energy. These energy values will be the only possible results of precise measurements of the total energy of the particle. These discrete values of energy E_1, E_2, \dots, E_n are called **eigenvalues** or **allowed values** of the energy of the particle. The solutions $\psi_1, \psi_2, \dots, \psi_n$ corresponding to the eigen energy values E_n are called the *eigen functions*. The quantization of energy thus appears as a natural element of the wave equation.

In Schrödinger's theory, the quantization of energy values follows directly from the mathematical formulation of the wave and particle nature of the electron. The existence of the stationary states has not been assumed and no assumptions have been made about orbits. The new theory yields all the results of the Bohr's theory without having any of the inconsistent hypotheses of the earlier theory. The new theory also accounts for the experimental information for which the Bohr's theory failed to account, such as the probability of an electron changing from one state to another.

We now apply the Schrödinger wave equation to different cases of particles confined to move in different enclosures and obtain information regarding the energy values it can take and its probability density etc.

20.19 THE FREE PARTICLE

A particle is said to be a **free particle** when it is moving in space without being subjected to any external force in any region of space, and its potential energy is constant [$V = \text{constant}$] everywhere. We study here the one dimensional motion of a particle under no forces.

Let us assume that the particle is moving in space along the positive x-direction. As the particle is not acted upon by a force, the potential energy of the particle is constant. For convenience, we take the constant potential to be zero. Then, the time-independent Schrödinger wave equation for the free particle is written as

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2 m}{h^2} E \psi = 0 \quad (20.62)$$

As the particle is moving freely with zero potential energy, its total energy will be kinetic energy which is given by

$$E = \frac{p_x^2}{2m} \quad (20.63)$$

where p_x is the momentum of the particle along the x-direction.

We may rewrite the equation (20.62) as

$$\frac{d^2\psi}{dx^2} + k^2\psi = 0 \quad (20.64)$$

where

$$\frac{8\pi^2 m E}{h^2} = k^2 \quad (20.65)$$

(a) Wave function

The general solution of the equation (20.65) is of the form

$$\psi(x) = Ae^{ikx} + Be^{-ikx} \quad (20.66)$$

where A and B are constants of integration. Eq.(20.66) gives the time-independent part of the wave function. The complete wave function is given by

$$\begin{aligned} \psi(x, t) &= \psi(x)e^{-i\omega t} \\ &= (Ae^{ikx} + Be^{-ikx}) e^{-i\omega t} \\ &= Ae^{-i(\omega t - kx)} + Be^{-i(\omega t + kx)} \end{aligned} \quad (20.67)$$

Equ.(20.67) represents a continuous plane simple harmonic wave. The first term represents the wave traveling along the positive x-direction while the second term represents the wave traveling along the negative x-direction. The particle traveling in the positive x-direction is represented by

$$\psi(x, t) = Ae^{-i(\omega t - kx)} \quad (20.68)$$

(b) Energy:

It is seen from eq. (20.65) that k has the following value.

$$k = \sqrt{\frac{8\pi^2 m E}{h^2}} = \sqrt{\frac{2mE}{\hbar^2}} \quad (20.69)$$

Therefore, the particle energy is given by

$$E = \frac{\hbar^2 k^2}{2m} \quad (20.70)$$

There are no boundary conditions to be applied to the particle motion and hence there are no restrictions placed on k . It follows from eq.(20.70) that the particle is permitted to have any value of energy. In other words, the energy is not quantized. It means that a freely moving particle possesses a continuous energy spectrum as shown in Fig. 20.15.

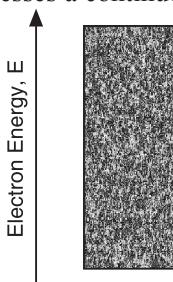


Fig. 20.15

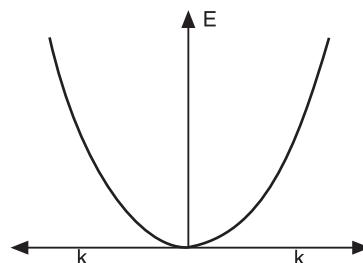


Fig. 20.16

The k -vector describes the wave properties of the particle. It may be seen from equ. (20.70) that $E \propto k^2$. The plot of E as a function of k gives a parabola, as illustrated in Fig. 20.16.

(c) Position of the Particle

The probability of finding the particle between x and $x + dx$ is given by

$$P dx = \psi^*(x,t)\psi(x,t)dx \quad (20.71)$$

or

$$P dx = A^2 dx \quad (20.72)$$

The probability density P for the position of the particle with the definite value of momentum is constant over the x -axis. It means that the particle is not located at a definite point but all positions are equally probable. This conclusion is in agreement with the Heisenberg uncertainty principle. As the momentum is well defined in this case, it is difficult to assign a position to the particle. The uncertainty in position will be infinity which means that the particle position is indeterminate.

20.20 POTENTIAL ENERGY STEP

A particle moving in a region of constant potential energy region may suddenly encounter a region of higher potential energy, which is also constant. Thus, the potential energy step is a square potential step as shown in Fig. 20.17. We will not discuss in detail the solutions but only give the outline of the salient steps and conclusions.

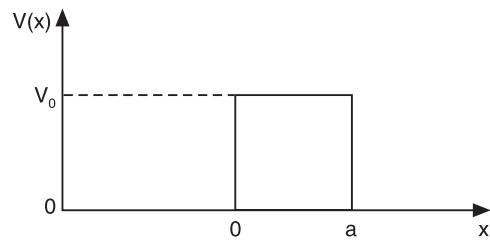


Fig. 20.17

(a) Potential energy step, $E > V_o$:

Let E be the fixed total energy of the particle and V be the value of the constant potential energy.

It is seen from Fig. 20.17 that $V = 0$ for $x < 0$
and $V = V_o$ for $x > 0$

(i) In the region $x < 0$ the time independent Schrödinger equation is given by

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2 m}{h^2} E \psi = 0$$

or

$$\frac{d^2\psi}{dx^2} + k_o^2 \psi = 0 \quad (20.73)$$

where

$$\frac{8\pi^2 m E}{h^2} = k_o^2.$$

(ii) In the region $x > 0$ the time independent Schrödinger equation is given by

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2 m}{h^2} (E - V_o) \psi = 0$$

or

$$\frac{d^2\psi}{dx^2} + k^2 \psi = 0 \quad (20.74)$$

where

$$k^2 = \frac{8\pi^2 m}{h^2} (E - V_o).$$

The solutions of equations (20.73) and (20.74) are

$$\psi(x) = A e^{ik_o x} + B e^{-ik_o x} \quad \text{for } x < 0 \quad (20.75 \text{ a})$$

$$\psi(x) = Ce^{ikx} + De^{-ikx} \quad \text{for } x > 0 \quad (20.75 \text{ b})$$

where A, B, C and D are constants of integration. The Schrödinger equation changes discontinuously at $x = 0$, as there is a discontinuity in V at $x = 0$. Relationships among the four coefficients A, B, C and D may be found by applying the condition that $\psi(x)$ and its derivative $\frac{d\psi}{dx}$ must be continuous at the boundary. Thus,

when $x = 0$,

$$\begin{aligned} \psi(x) &= A + B \text{ and } \psi(x) = C + D \\ \therefore A + B &= C + D \end{aligned} \quad (20.76 \text{ a})$$

At $x = 0$,

$$\frac{d\psi}{dx} = k_o(A - B)$$

and

$$\frac{d\psi}{dx} = k(C - D)$$

\therefore

$$k_o(A - B) = k(C - D)$$

or

$$(A - B) = \left(\frac{k}{k_o} \right)(C - D) \quad (20.76 \text{ b})$$

Solving the equations (20.76 a) and (20.76 b), we get

$$B = \left[\frac{k_o - k}{k_o + k} \right] A + \left[\frac{2k}{k_o + k} \right] D \quad (20.77 \text{ a})$$

and

$$C = \left[\frac{2k_o}{k_o + k} \right] A - \left[\frac{k_o - k}{k_o + k} \right] D \quad (20.77 \text{ b})$$

Substituting the value of B in equ.(20.75 a), we get

$$\begin{aligned} \psi(x) &= Ae^{ik_o x} + \left[\left(\frac{k_o - k}{k_o + k} \right) A + \left(\frac{2k}{k_o + k} \right) D \right] e^{-ik_o x} \\ \psi(x) &= A \left[e^{ik_o x} + \left(\frac{k_o - k}{k_o + k} \right) e^{-ik_o x} \right] + D \left[\left(\frac{2k}{k_o + k} \right) e^{-ik_o x} \right] \end{aligned}$$

or

$$\psi(x) = A\psi_1 + D\psi_2 \quad (20.78 \text{ a})$$

Substituting the value for C in equn.(20.75 b), we get

$$\psi(x) = \left[\left(\frac{2k_o}{k_o + k} \right) e^{ikx} \right] A + \left[e^{-ikx} - \left(\frac{k_o - k}{k_o + k} \right) e^{ikx} \right] D$$

or

$$\psi(x) = A\psi_1 + D\psi_2 \quad (20.78 \text{ b})$$

The equations (20.78 a) and (20.78 b) are the same and conditions of continuity at $x = 0$ are satisfied by both ψ_1 and ψ_2 .

The complete solution of equn.(20.73) is obtained when the equn.(20.75 a) is multiplied by the time dependent term $e^{-i\omega t}$. Thus,

$$\psi(x, t) = Ae^{-i(\omega t - k_o x)} + Be^{-i(\omega t + k_o x)} \quad (20.79)$$

The first term in the above equation represents a wave moving along the positive x-direction while the second term represents a wave moving along the negative x-direction.

Therefore, equ.(20.79) describes the superposition of a wave of intensity $|A|^2$ moving along the positive x-direction (from $-\infty$ to zero) and a wave of intensity $|B|^2$ moving along the negative x-direction. Therefore, when the particles are incident on the step from left on the potential energy step, $|A|^2$ gives the intensity of the **incident** wave and $|B|^2$ gives that of the **reflected** wave. The ratio $\frac{|B|^2}{|A|^2}$ gives us the reflected fraction of the incident wave intensity.

(b) Potential energy step, $E < V_o$:

Whenever E becomes less than V_o , a different solution results.

$$\psi(x) = Ae^{kx} + Be^{-kx} \quad (20.80 \text{ a})$$

If E were less than V_o , then the solution for ψ_o (for $x < 0$) would still be given by eqn. (20.75 a), but the solution for ψ (for $x > 0$) becomes

$$\psi_1(x) = Ce^{k_1 x} + De^{-k_1 x} \quad (20.80 \text{ b})$$

where $k_1 = \sqrt{\frac{8\pi^2 m}{h^2}(V_o - E)}$.

We set, $C = 0$ to keep $\psi_1(x)$ from becoming infinite as $x \rightarrow \infty$, and apply the boundary conditions on $\psi_1(x)$ at $x = 0$. The resulting solution is shown in Fig. 20.18 (b).

This illustrates an important difference between classical and quantum mechanics. Classically, the particle can never be found in the region $x > 0$, since its total energy is not sufficient to overcome the potential energy step. However, quantum mechanics allows the wave function, and therefore the particle, to penetrate into the classically forbidden region. The particle can never be observed in the forbidden region, but a particle can pass through a classically forbidden region and emerge into an allowed region where it can be observed.

Penetration into the forbidden region is associated with the wave nature of the particle. The penetration distance is consistent with the uncertainty in defining the location of the particle.

20.21 RECTANGULAR POTENTIAL BARRIER

According to classical ideas, a particle striking a hard wall has no chance of leaking through it. But the behaviour of a quantum particle is different owing to the wave nature associated with it. We know that when an electromagnetic wave strikes at the interface of two media, it is partly reflected and partly transmitted through the interface and enters the second medium. In a similar way the de Broglie wave also has a possibility of getting partly reflected from the boundary of the potential well and partly penetrating through the barrier. The penetration of a barrier by a quantum particle is called **tunneling**.

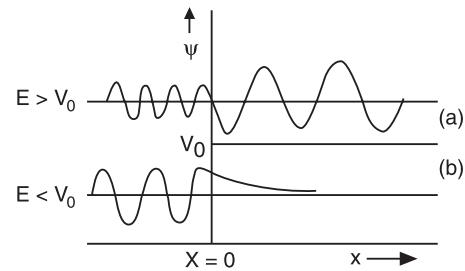


Fig. 20.18

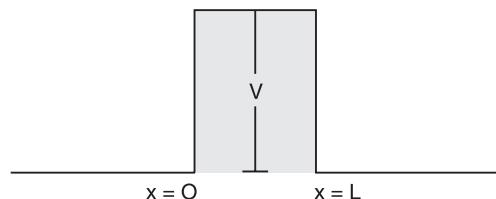


Fig. 20.19

Fig. 20.19 represents a potential barrier of height V_o and thickness L . The region around the barrier can be divided into three regions as shown in Fig. 20.20. The potential energy is zero in regions I and III, i.e., for $x < 0$ and $x > L$ and has a constant value V_o in the region II, i.e., for $0 < x < L$. Let us consider a particle of total energy E approaching the barrier from the left. In regions I and III the total energy of the particle is kinetic energy. In region II, the total energy is partly kinetic and partly potential energy. From the view-point of classical physics, the electron would be reflected from the barrier because its energy E is less than V . For the particle to overcome the potential barrier, it must have energy equal to or greater than V . Quantum mechanics leads to an entirely new result.

If $E < V$, according to quantum mechanics (20.20), there is a finite chance for the electron to leak to the other side of the barrier. We say that the electron tunneled through the potential barrier and hence in quantum mechanics, the phenomenon is called **tunneling**.

We write down the Schrödinger wave equation for the electron wave in the three regions and solve them. The Schrödinger wave equation for a particle constrained to move along x -axis is given by

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2m}{h^2}(E - V)\psi = 0 \quad (20.81)$$

In the region I, $V = 0$ and the equation (20.81) takes the form

$$\frac{d^2\psi_1}{dx^2} + \frac{8\pi^2mE}{h^2}\psi_1 = 0$$

or $\frac{d^2\psi_1}{dx^2} + k_o^2\psi_1 = 0$ (20.82)

\therefore In the region II, $E < V_o$, $V = V_o$ and the equation (20.81) takes the form

$$\frac{d^2\psi_2}{dx^2} + \frac{8\pi^2m}{h^2}(E - V_o)\psi_2 = 0$$

or $\frac{d^2\psi_2}{dx^2} - \frac{8\pi^2m}{h^2}(V_o - E)\psi_2 = 0$ (20.83)

or $\frac{d^2\psi_2}{dx^2} - k^2\psi_2 = 0$ (20.84)

where $k = \sqrt{\frac{8\pi^2m}{h^2}(V_o - E)}$.

In the region III, $V = 0$. Therefore,

$$\frac{d^2\psi_3}{dx^2} + k_o^2\psi_3 = 0$$

The solutions of equations (20.82), (20.83), and (20.84) are

$$\psi_1 = Ae^{ik_o x} + Be^{-ik_o x} \quad (20.85 \text{ a})$$

$$\psi_2 = Ce^{k x} + De^{-k x} \quad (20.85 \text{ b})$$

$$\psi_3 = Le^{ik_o x} + Me^{-ik_o x} \quad (20.85 \text{ c})$$

In the region III, no particle comes from right and hence $M = 0$.

$$\therefore \Psi_3 = Le^{ik_ox} \quad (20.85 \text{ d})$$

In eq. (20.85 a), Ae^{ik_ox} represents the de Broglie wave traveling along the x-direction in the region I, with amplitude A and Be^{ik_ox} represents the wave reflected along the negative x-direction with amplitude B. Since the probability that the particle is present in a region of the space is proportional to the square of the de Broglie wave amplitude, the ratio

$$R = \frac{|B|^2}{|A|^2} \quad (20.86)$$

is the **coefficient of reflection** of the particle from the barrier.

The wave function Ψ_2 is *not* zero inside the barrier (the region forbidden by classical mechanics), but decreases exponentially. De^{-kx} represents the exponentially decreasing wave in the barrier. The square of the amplitude of this wave defines the probability of penetration of the particle into the region II. The ratio is the **coefficient of penetration** of the interface. Ce^{kx} is the reflected wave within the barrier. Since such a wave is nonexistent, C should be equal to zero.

$$T = \frac{|D|^2}{|A|^2} \quad (20.87)$$

In eq.(20.85 d) Le^{ik_ox} represents the transmitted wave moving along the x-direction in the region III.

The form of the wave function in the region (I), (II) and (III) is also shown in the Fig. 20.20. The wave function Ψ_1 corresponds to the free electron with momentum $p = \sqrt{2mE}$. Since Ψ_3 is not equal to zero at $x = L$, there is a finite probability of finding the electron in the region III. That means the particle that is initially to the left of the barrier has some probability of being found to the *right* of the barrier. The wave function Ψ_3 represents the wave transmitted through the barrier and the free particle on the right side of the barrier. The particle has the same momentum as the incident particle but has smaller amplitude. Thus, it is possible for a particle to penetrate through the potential barrier even if its kinetic energy is less than the height of the potential barrier.

This probability is proportional to the square of the modulus of the wave function Ψ_2 .

$$P \propto |\Psi_2|^2 = |C|^2 \exp\left(-\frac{2x}{\hbar} \sqrt{2m(V_o - E)}\right) \quad (20.88)$$

This probability is indicative of the fact that the particle is able to penetrate the potential barrier of finite width L. Such a penetration is called the **tunnel effect**. The probability that the particle gets through the barrier is called the **transmission coefficient**. It is defined as

$$T = \frac{\text{Probability density of the transmitted wave}}{\text{Probability density of the incident wave}}$$

It is shown that the transmission coefficient is given approximately by

$$T = Ge^{-2kL} = \left[\frac{16E}{V} \left(1 - \frac{E}{V} \right) \right] e^{-2L\sqrt{8\pi^2 m(V-E)/h}} \quad (20.89)$$

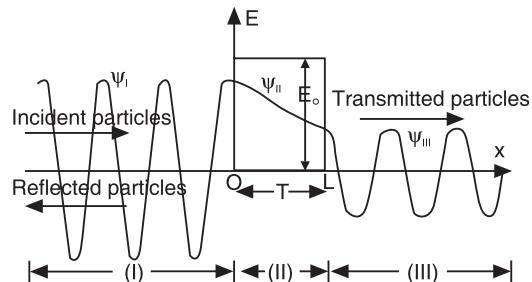


Fig. 20.20

where G is a constant close to unity. Equ.(20.89) shows that the probability of particle penetration through a potential barrier depends on the height, V and width, L of the barrier. The probability decreases rapidly with increasing barrier width, L .

Tunneling is significant in many areas of physics.

20.21.1 Application of Tunnelling

1. The **tunnel diode** is a semiconductor diode. The current in this device is largely due to tunneling of electrons through a potential barrier. The rate of tunneling or current can be controlled over a wide range by varying the height of the barrier, which is done by varying the applied voltage.
2. **α -decay:** The quantum possibility of penetrating potential barriers is the basis of the explanation of the α -decay of radioactive nuclei. In α -decay, an unstable nucleus disintegrates into a lighter nucleus and an α -particle. For example, a uranium nucleus ^{238}U undergoes α -decay and forms thorium nucleus ^{234}Th .

In 1928 George Gamow explained α -decay of unstable nuclei on the basis of quantum tunneling. The forces in the nucleus set up a potential barrier of height of the order of 30 MeV against α -particle emission. Classically, the α -particle would be trapped unless its energy exceeds 30 MeV. The α -particles have energies in the range of 4 to 9 MeV only. Therefore, it is impossible for a α -particle to jump the barrier. According to quantum mechanics, the α -particle tunnels through the potential barrier (Fig. 20.21). The experimental studies of Geiger-Nuttal on α - decay confirmed the predictions of Gamow's theory.

3. **Scanning tunnelling microscope (STM):** The instrument was invented in the early 1979 by Gerd Binnig and Heinrich Rohrer, who were awarded the 1986 Nobel prize in physics for their work. The scanning tunnelling electron microscope uses electron tunneling to produce images of surfaces down to the scale of individual atoms. If two conducting samples are brought in close proximity, with a small but finite distance between them, electrons from one sample flow into the other if the distance is of the order of the spread of the electronic wave into space. One says that the electrons "tunnel" through the barrier into the adjacent sample. For electrons, the barrier width which may be overcome via a tunneling process is of the order of nm, i.e. of the order of several atomic spacings. The probability of an electron to get through the tunneling barrier decreases exponentially with the barrier width, i.e. the so called tunneling current is extremely sensitive measure of the distance between two conducting samples. The STM makes use of this sensitivity.

Working: The schematic diagram of a scanning tunneling electron microscope is shown in Fig. 20.22 (a). In the scanning tunneling microscope the sample is scanned by a very fine metallic tip. The tip is mechanically connected to the scanner, an XYZ positioning device. The

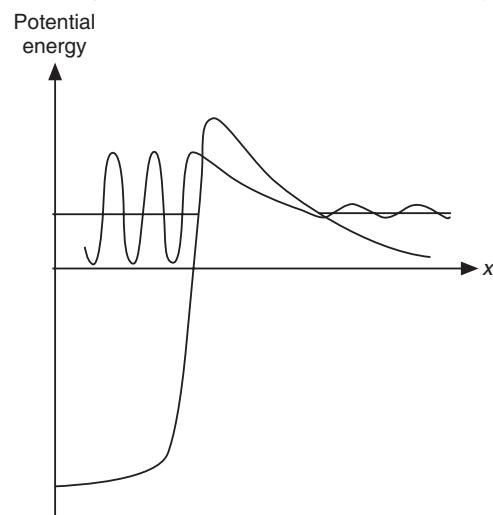


Fig. 20.21: Alpha particle having less energy than the height of the potential barrier tunnels through the barrier

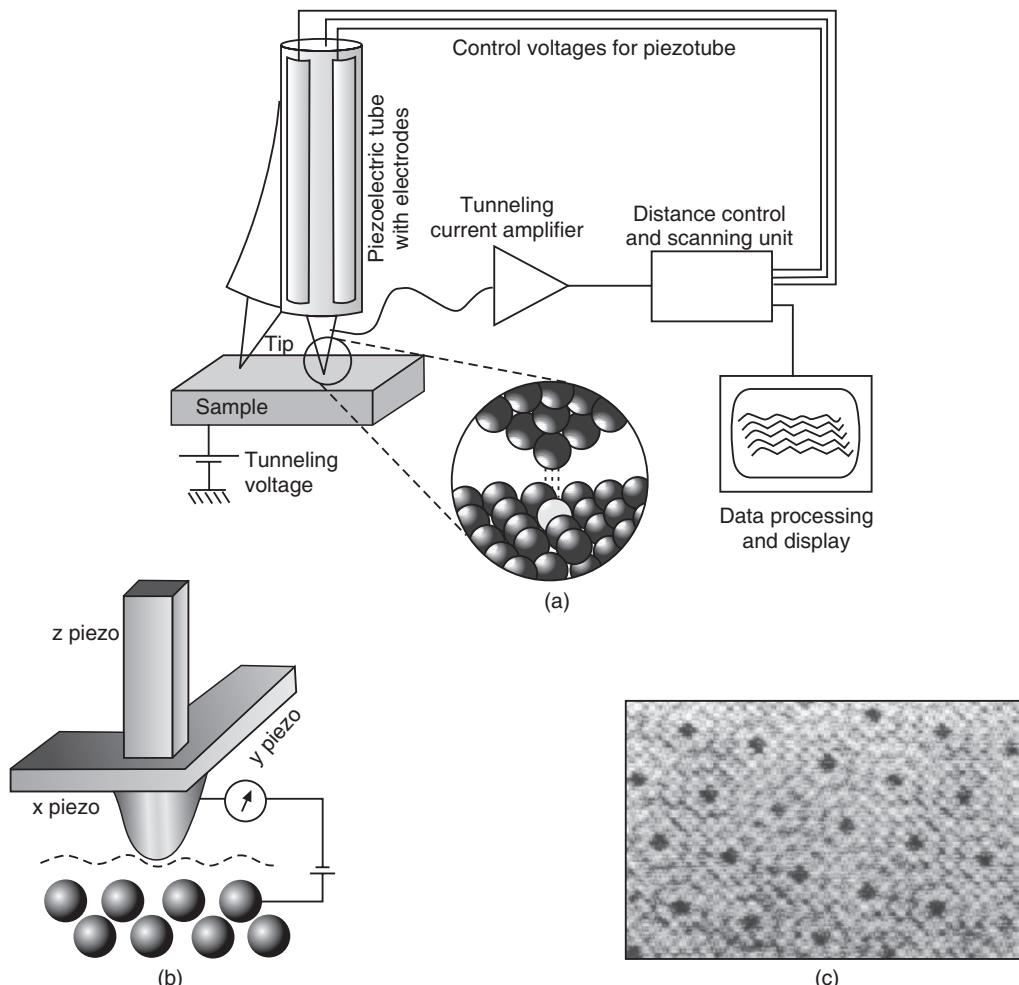


Fig. 20.22: (a) The schematic diagram of STM (b) The tunneling tip is scanned over the specimen, producing an image of the tunneling current (c) An image of a silicon crystal surface produced by a STM.

sharp metal needle is brought close to the surface to be imaged. The distance is of the order of a few angstroms. A bias voltage is applied between the sample and the tip. When the needle is at a positive potential with respect to the surface, electrons can tunnel through the gap and set up a small “tunneling current” in the needle. This feeble tunneling current is amplified and measured. With the help of the tunneling current the feedback electronics keeps the distance between tip and sample constant. The sensitivity of the STM is so large that electronic corrugation of surface atoms and the electron distribution around them can be detected. Fig. 20.22 (b) shows the tip of the sharp needle and Fig. 20.22 (c) shows the image of silicon crystal surface.

Example 20.8. A stream of electrons, each of energy $E = 3\text{eV}$, is incident on a potential barrier of height $V = 4\text{eV}$. The width of the barrier is 20\AA . Calculate the percentage transmission of the beam through this barrier.

Solution. The probability of transmission through a potential barrier is given by

$$\begin{aligned}
 T &= Ge^{-2kl} = \left[\frac{16E}{V} \left(1 - \frac{E}{V} \right) \right] e^{-2L\sqrt{8\pi^2m(V-E)/h}} \\
 \frac{2L}{h} \sqrt{8\pi^2m(V-E)} &= \frac{2 \times 20 \times 10^{-10} m}{6.63 \times 10^{-34} Js} \sqrt{8 \times (3.143)^2 \times 9.11 \times 10^{-31} kg (4-3) \times 1.602 \times 10^{-19} J} \\
 &= 20.49 \\
 \therefore T &= \left[\frac{16E}{V} \left(1 - \frac{E}{V} \right) \right] e^{-20.49} \\
 &= \frac{16 \times 3eV}{4eV} \left(1 - \frac{3eV}{4eV} \right) e^{-20.49} \\
 &= 3 \times e^{-20.49} = 3.8 \times 10^{-9}
 \end{aligned}$$

Therefore, the percentage of transmission = $3.8 \times 10^{-7}\%$.

20.22 INFINITE POTENTIAL WELL

A **potential well** is a potential energy function $V(x)$ that has a minimum (see Fig. 20.23). A potential well is the opposite of a potential barrier; it is a potential-energy function with a minimum. If a particle is left in the well and the total energy of the particle is less than the height of the potential well, we say that the particle is *trapped* in the well. In classical mechanics a particle trapped in a potential well can vibrate back and forth with periodic motion but cannot leave the well. In quantum mechanics, such a trapped state is called a **bound state**.

Let us consider a particle confined to the region $0 < x < L$. It can move freely within the region $0 < x < L$ but subject to strong forces at $x = 0$ and $x = L$. Therefore, it can never cross to the right to the region $x > L$ or to the left of 0. It means that $V = 0$ in the region $0 < x < L$ and rises to infinity ($V = \infty$) at $x = 0$ and $x = L$. This situation is called a **one-dimensional potential box**.

For a particle trapped in a one-dimensional potential box, $V = 0$ and the Schrodinger equation (20.61a) takes the following form

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2mE}{h^2}\psi = 0 \quad (20.90)$$

Putting $\frac{8\pi^2mE}{h^2} = k^2$, we rewrite the above equation as

$$\frac{d^2\psi}{dx^2} + k^2\psi = 0 \quad (20.91)$$

Because the particle can move back and forth freely between $x = 0$ and $x = L$, the solution of the equation (20.91) is of the form

$$\psi(x) = Ae^{ikx} + Be^{-ikx} \quad (20.92)$$

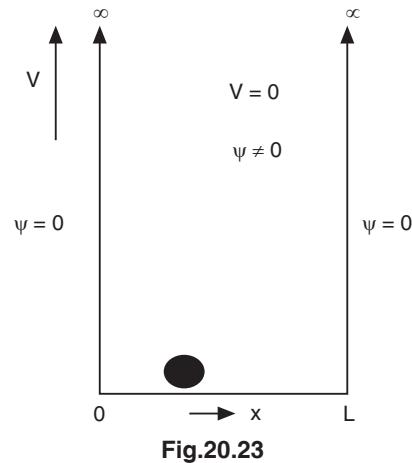


Fig.20.23

which contains motion in both directions. We can evaluate the constants A and B with the help of boundary conditions. The boundary conditions are as follows. The particle cannot jump over the walls and therefore the function does not exist outside the box. The particle is located within the box and therefore it exists within the box. It means that

$$\begin{aligned}\psi(x) &= 0 \text{ at } x = 0 \\ \psi(x) &= 0 \text{ at } x = L \\ \therefore \quad \psi_{x=0} &= A+B = 0 \\ \therefore \quad B &= -A\end{aligned}$$

Using this result into equ.(20.92), we get

$$\psi(x) = A(e^{ikx} - e^{-ikx})$$

or $\psi(x) = 2iA \sin kx$

Again $\psi(x) = 0 \text{ at } x = L$
 $\therefore \psi_{x=L} = 2iA \sin kL = 0$

The factor $2iA$ cannot be zero, because we would not then have the wave function. It means that

$$\begin{aligned}\sin kL &= 0 \\ \text{or} \quad kL &= n\pi \\ \text{As } k &= 2\pi/n, \quad \lambda = 2L/n\end{aligned}\tag{20.93}$$

where $n (=1,2,3, \dots)$ is an integer.

The above conclusion implies that the wave equation has solutions only when the particle wavelength is restricted to discrete values such that only a whole number of half-wavelengths are formed over the length L of the box. It means that particle waves form standing wave pattern within the potential box. We write (20.93) as

$$k = n\pi / L$$

(i) The possible values of the momentum of the particle are then given by

$$p = \hbar k = \frac{n\pi\hbar}{L} = \frac{nh}{2L}\tag{20.94}$$

(ii) The possible values of the energy are

$$E = \frac{p^2}{2m} = \frac{n^2\hbar^2}{8mL^2}\tag{20.95}$$

The above equation indicates that a particle confined in a certain region can have only certain values of energy. Other energy values are not allowed. In other words, **energy quantization** is a consequence of restricting a microparticle to a certain region.

The allowed wave functions which are the solutions of the Schrodinger equation are given by

$$\psi_n = 2iA \sin kx = C \sin kx$$

As $k = n\pi / L$, we write the above as

$$\psi_n = C \sin\left(\frac{n\pi x}{L}\right)\tag{20.96}$$

Applying normalization condition, we can determine the value of the constant C in the above equation. The normalisation condition is that

$$\int_0^L \psi_n \psi_n^* dx = 1$$

$$\therefore C^2 \int_0^L \sin^2 \frac{n\pi x}{L} dx = 1$$

$$\frac{C^2}{2} \int_0^L \left[1 - \cos \frac{2n\pi x}{L} \right] dx = 1$$

$$\frac{C^2}{2} \left[x - \frac{\sin \frac{2n\pi x}{L}}{\frac{2n\pi}{L}} \right]_0^L = 1$$

$$\therefore C^2 \frac{L}{2} = 1$$

$$\text{i.e., } C^2 = \frac{2}{L}$$

or

$$C = \sqrt{2/L} \quad (20.97)$$

Using the value of C into eq.(20.96), we obtain the wave function as

$$\psi_n = \sqrt{\frac{2}{L}} \sin\left(\frac{n\pi x}{L}\right) \quad (20.98)$$

Energy Levels:

The allowed energy states are given by

$$E_n = \frac{h^2}{8mL^2} n^2$$

where n is the quantum number given by $n = 1, 2, 3, \dots$

The energy states are obtained by inserting the value of n in the above equation. Thus

$$\text{First energy level} \quad E_1 = \frac{h^2}{8mL^2}$$

$$\text{Second energy level} \quad E_2 = \frac{h^2}{2mL^2}$$

$$\text{Third energy level} \quad E_3 = \frac{9h^2}{8mL^2} \text{ and so on.}$$

The energy levels are shown in Fig. 20.24.

It is important to note that the particle cannot have zero energy. The lowest energy allowed for the particle is $E_1 = h^2/8 mL^2$. It cannot possess energy less than this in the one-dimensional potential box.

The energy value E_1 is called **zero-point energy**. The zero point energy is a consequence of the uncertainty principle. If the energy of the particle is zero, its momentum also would be zero, and the uncertainty principle requires that the wavelength λ be infinite. In such case the particle cannot be confined to the box. Therefore, the particle must have a certain minimum amount of kinetic energy.

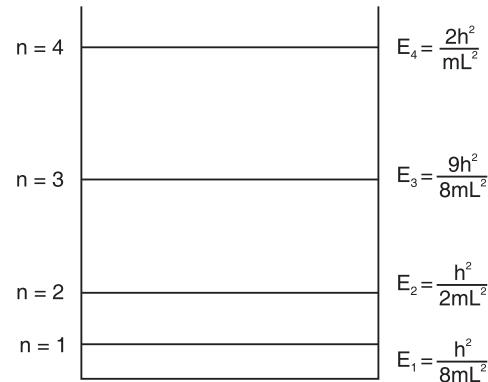


Fig. 20.24

Comparison with Classical Results:

- Classically, a particle in a box could have any positive energy. Quantum theory shows that the particle is restricted to take only certain discrete values.
- In classical theory, a particle in a box can have energy of zero value. The quantum theory shows that a particle in a box cannot have zero energy. The minimum energy possessed by the particle is $E_1 = \frac{h^2}{8mL^2}$.

Probability of locating the particle over an interval x and $x + dx$

The probability $P(x)dx$ of finding the particle over a infinitesimal distance dx at position x in the potential well is given by

$$P(x)dx = \int |\psi_n|^2 dx = \frac{2}{L} \sin^2\left(\frac{n\pi x}{L}\right) dx \quad (20.99)$$

The probability of finding the microparticle between the positions x_1 and x_2 is

$$\begin{aligned} P &= \frac{2}{L} \int_{x_1}^{x_2} \sin^2\left(\frac{n\pi x}{L}\right) dx \\ &= \frac{1}{L} \int_{x_1}^{x_2} \left(1 - \cos\frac{2n\pi x}{L}\right) dx \\ &= \frac{1}{L} [x]_{x_1}^{x_2} - \frac{1}{2\pi} \left[\sin\frac{2n\pi x}{L}\right]_{x_1}^{x_2} \end{aligned} \quad (20.100)$$

The probability density is given by

$$P(x) = \frac{2}{L} \sin^2\left(\frac{n\pi x}{L}\right) dx \quad (20.101)$$

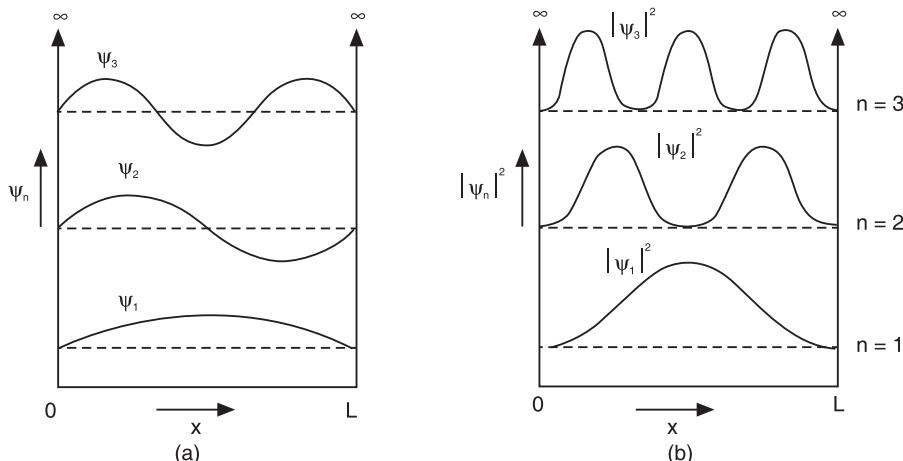


Fig. 20.25

The probability density is a maximum when $\frac{n\pi x}{L} = \frac{\pi}{2}, \frac{3\pi}{2}, \frac{5\pi}{2}, \dots$ or $x = \frac{L}{2n}, \frac{3L}{2n}, \frac{5L}{2n}, \dots$

The variation of the probability densities with x for $n = 1$, $n = 2$, and $n = 3$ are shown in Fig. 20.25.

It is easy to see that in the state ψ_1 the probability is seen to be largest at $x = L/2$, i.e., in the middle of the box (Fig. 20.25a) and it decreases towards the walls. It means that the particle will stay more in the centre of the box and avoids the region near walls. Similarly, in the state ψ_2 the most probable positions are at $x = L/4$ and $x = 3L/4$. Therefore, the particle will be found either in the right-half or in the left-half of the box but never in the middle. In classical theory, all positions within the box are equally probable for the particle. Classically, the particle passes back and forth between the walls and it has an equal probability of being found anywhere between $x = 0$ and $x = L$. The classical probability function is a constant with a value of $1/L$. The quantum results show peaks of magnitude $2/L$ and valleys where the probability is very small. The number of peaks is equal to the quantum number n . As n increases, the number of peaks increase and when n is very large, the distribution approaches the classical distribution.

Dependence of quantization on the width of the box:

Equation (20.95) suggests that the quantization of energy is dependent on the width of the potential box, L and on the quantum number n . Let us now study the effect of L on the quantization of particle energy. If we designate two adjacent energy levels by E_{n+1} and E_n , the separation between these levels is given by

$$\begin{aligned}\Delta E &= E_{n+1} - E_n \\ &= \frac{\hbar^2}{8mL^2} [(n+1)^2 - n^2] \\ &= \frac{\hbar^2}{8mL^2} (2n+1)\end{aligned}\quad (20.102)$$

As $2n > 1$, $2n+1 \approx 2n$

$$\therefore \Delta E = \frac{\hbar^2}{4mL^2} n \quad (20.103)$$

Case (i): Let us take the case of an electron moving in a box of side $L = 1 \text{ cm} = 10^{-2} \text{ m}$.

From equation (20.102) we get

$$\Delta E \approx n \times 10^{-15} \text{ eV}$$

The thermal energy possessed by the electron is

$$kT \approx 10^{-3} \text{ eV at } T = 1^\circ\text{K}$$

$$\Delta E \ll kT$$

It means that the electron can move from a lower energy level to an upper energy due to its own thermal energy and without the need of any energy input from external agency. As the levels are nearly continuous, the electron behaves as a classical particle without revealing the quantum aspects.

Case (ii) Let us next consider the case of an electron moving in a box of side $L = 10 \text{ \AA}$.

Then

$$\Delta E \approx 0.75 (n) \text{ eV.}$$

$$\therefore \Delta E \gg kT.$$

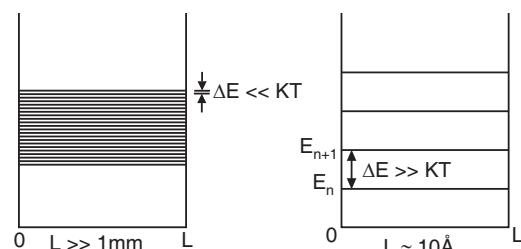


Fig. 20.26

This energy difference between the adjacent energy levels is very large and the electron cannot go to higher levels on its own, as the energy levels are separated by large energy values. Therefore, it becomes necessary to supply energy from an external agency.

The above example amply demonstrates that the quantization of energy assumes a great importance in the atomic world.

20.23 EXTENSION TO THREE-DIMENSIONAL CASE

The case can be generalized to the three-dimensional box with sides a , b and c . As the coordinates are orthogonal, the general time-independent Schrodinger equation (equ.20.61) can be separated into three equations and can be solved separately for the eigen values and eigen functions. The complete eigen function is

$$\begin{aligned}\psi(x, y, z) &= \psi_{nx}(x) \psi_{ny}(y) \psi_{nz}(z) \\ &= \sqrt{\frac{8}{abc}} \sin \frac{n_x \pi x}{a} \sin \frac{n_y \pi y}{b} \sin \frac{n_z \pi z}{c}\end{aligned}\quad (20.104)$$

The total energy of the particle is

$$\begin{aligned}E_n &= E_{nx} + E_{ny} + E_{nz} \\ &= \frac{h^2}{8m} \left[\frac{n_x^2}{a^2} + \frac{n_y^2}{b^2} + \frac{n_z^2}{c^2} \right]\end{aligned}\quad (20.105)$$

where n_x , n_y , and n_z are the three quantum numbers for this case.

In case of a cubical box, $a = b = c$ and we get

$$\psi(n_x, n_y, n_z) = \left[\frac{8}{a} \right]^{3/2} \sin \frac{n_x \pi x}{a} \sin \frac{n_y \pi y}{a} \sin \frac{n_z \pi z}{a} \quad (20.106)$$

and

$$E_n = \frac{h^2}{8m a^2} \left[n_x^2 + n_y^2 + n_z^2 \right] \quad (20.107)$$

or

$$E_n = \frac{n^2 h^2}{8ma^2} \quad (20.108)$$

where

$$n^2 = \left[n_x^2 + n_y^2 + n_z^2 \right]$$

The ground state is given by

$$E_1 = \frac{3h^2}{8ma^2}$$

and the first excited state is given by

$$E_2 = \frac{3h^2}{4 m a^2} = 2E_1.$$

Degeneracy:

For different combinations of quantum numbers, we may obtain the same energy value but the wave functions are different. Such quantum states having the same energy are called **degenerate**. Thus, for example, the wave functions

$$\psi_{112} = \sqrt{\frac{8}{a^3}} \sin \frac{\pi x}{a} \sin \frac{\pi y}{b} \sin \frac{2\pi z}{c}$$

$$\psi_{121} = \sqrt{\frac{8}{a^3}} \sin \frac{\pi x}{a} \sin \frac{2\pi y}{b} \sin \frac{\pi z}{c}$$

$$\psi_{211} = \sqrt{\frac{8}{a^3}} \sin \frac{2\pi x}{a} \sin \frac{\pi y}{b} \sin \frac{\pi z}{c}$$

are different. But the corresponding energies are the same. The quantum numbers corresponding to the first wave function are

$$\text{and } n_x = 1, n_y = 1, \text{ and } n_z = 2$$

$$n^2 = [n_x^2 + n_y^2 + n_z^2] = 6$$

$$\therefore E_{112} = \frac{6h^2}{8ma^2}$$

Similarly, for the other two wave functions, we get $E_{121} = E_{211} = \frac{6h^2}{8ma^2}$

$$\therefore E_{112} = E_{121} = E_{211} = \frac{6h^2}{8ma^2}$$

Thus, the first excited state is degenerate, since the same n value is given by three sets $(1,1,2)$, $(1,2,1)$, and $(2,1,1)$. The number of different states with a certain value of the energy is known as the **degree of degeneracy**. Thus, we say that the first excited state is three-fold degenerate.

Example 20.9. Find the probability that a particle trapped in a box L wide can be found between $0.45 L$ and $0.55 L$ for the ground state and first excited state.

Solution.

$$\begin{aligned} P &= \frac{2}{L} \int_{x_1}^{x_2} \sin^2 \left(\frac{n\pi x}{L} \right) dx \\ &= \frac{1}{L} \int_{x_1}^{x_2} \left(1 - \cos \frac{2n\pi x}{L} \right) dx \\ &= \frac{1}{L} [x]_{x_1}^{x_2} - \frac{1}{2n\pi} \left[\sin \frac{2n\pi x}{L} \right]_{x_1}^{x_2} \end{aligned}$$

For the ground state, $n = 1$ and

$$\begin{aligned} P &= \frac{1}{L} [x]_{x_1}^{x_2} - \frac{1}{2\pi} \left[\sin \frac{2\pi x}{L} \right]_{x_1}^{x_2} \\ &= \frac{1}{L} [x]_{0.45L}^{0.55L} - \frac{1}{2\pi} \left[\sin \frac{2\pi x}{L} \right]_{0.45L}^{0.55L} \end{aligned}$$

or

$$P = \frac{1}{L} [x_2 - x_1] - \frac{1}{2\pi} \left\{ \left[\sin \frac{2\pi x_2}{L} \right] - \left[\sin \frac{2\pi x_1}{L} \right] \right\}$$

$$P = \frac{1}{L} [0.55L - 0.45L] - \frac{1}{2\pi} \left\{ \left[\sin \frac{2 \times 180^\circ \times 0.55L}{L} \right] - \left[\sin \frac{2 \times 180^\circ \times 0.45L}{L} \right] \right\}$$

$$= 0.1 - \frac{1}{2 \times 3.143} (-0.309 - 0.309) = \mathbf{0.198}$$

For the 1st excited state, $n = 2$ and

$$P = \frac{1}{L} [0.55L - 0.45L] - \frac{1}{4\pi} \left\{ \left[\sin \frac{4 \times 180^\circ \times 0.55L}{L} \right] - \left[\sin \frac{4 \times 180^\circ \times 0.45L}{L} \right] \right\}$$

$$= 0.1 - \frac{1}{4 \times 3.143} (0.5878 + 0.5878) = \mathbf{0.0065}.$$

Example 20.10. An electron is confined to move in a one dimensional potential well of length 5\AA . Find the quantized energy values for the three lowest energy states.

Solution. The quantized energy values of an electron confined in a potential well are given by

$$E_n = \frac{h^2}{8mL^2} n^2$$

$$\text{Energy of the ground state } E_1 = \frac{h^2}{8mL^2} 1^2 = \frac{h^2}{8mL^2} = \frac{(6.63 \times 10^{-34} \text{ J.s})^2}{8 \times 9.11 \times 10^{-31} \text{ kg} \times (5 \times 10^{-10} \text{ m})^2}$$

$$= 2.41 \times 10^{-19} \text{ J} = \mathbf{1.5 \text{ eV}.}$$

$$\text{Energy of the 1}^{\text{st}} \text{ excited state } E_2 = \frac{h^2}{8mL^2} 2^2 = \frac{4h^2}{8mL^2} = 4 \times 1.5 \text{ eV} = \mathbf{6 \text{ eV}.}$$

$$\text{Energy of the 2}^{\text{nd}} \text{ excited state } E_3 = \frac{h^2}{8mL^2} 3^2 = \frac{9h^2}{8mL^2} = 9 \times 1.5 \text{ eV} = \mathbf{13.5 \text{ eV}.}$$

Example 20.11: Calculate the energy required for an electron to jump from ground state to the second excited state in a potential well of width L .

Solution:

$$E_n = \frac{h^2}{8mL^2} n^2$$

$$\text{Energy of the ground state } E_1 = \frac{h^2}{8mL^2} 1^2 = \frac{h^2}{8mL^2}$$

$$\text{Energy of the 2}^{\text{nd}} \text{ excited state } E_3 = \frac{h^2}{8mL^2} 3^2 = \frac{9h^2}{8mL^2}$$

$$\text{Energy required for the transition } E = E_3 - E_1 = \frac{9h^2}{8mL^2} - \frac{h^2}{8mL^2} = \frac{8h^2}{8mL^2} = \frac{h^2}{mL^2}.$$

Example 20.12. An electron is trapped in a one-dimensional box of length 0.1 nm . Calculate the energy required to excite the electron from its ground state to the fourth excited state.

Solution.

$$E_n = \frac{h^2}{8mL^2} n^2$$

$$\therefore E_1 = \frac{1^2 \times (6.63 \times 10^{-34} \text{ J.s})^2}{8 \times 9.11 \times 10^{-31} \text{ kg} \times (0.1 \times 10^{-9} \text{ m})^2}$$

$$= 60.3 \times 10^{-19} \text{ J} = 37.6 \text{ eV.}$$

and

$$E_4 = \frac{4^2 \times (6.63 \times 10^{-34} \text{ J.s})^2}{8 \times 9.11 \times 10^{-31} \text{ kg} \times (0.1 \times 10^{-9} \text{ m})^2} = 601.6 \text{ eV}$$

The energy required to excite the electron from its ground state to the fourth excited state is

$$E_4 - E_1 = 601.6 \text{ eV} - 37.6 \text{ eV} = 564 \text{ eV.}$$

Example 20.13. An electron is confined to move between two rigid walls separated by 1 nm. Find the de Broglie wavelength representing the first two allowed energy states of the electron and the corresponding energies.

Solution. The eigen values of electron energy confined in a potential well are given by

$$E_n = \frac{h^2}{8mL^2} n^2$$

Energy of the 1st allowed state $E_1 = \frac{h^2}{8mL^2} 1^2 = \frac{h^2}{8mL^2}$

$$\therefore E_1 = \frac{1^2 \times (6.63 \times 10^{-34} \text{ J.s})^2}{8 \times 9.11 \times 10^{-31} \text{ kg} \times (1 \times 10^{-9} \text{ m})^2} = 6.03 \times 10^{-20} \text{ J}$$

Energy of the 2nd allowed state $E_1 = \frac{h^2}{8mL^2} 2^2 = 4 \times \frac{h^2}{8mL^2} = 4 \times 6.03 \times 10^{-20} \text{ J} = 2.4 \times 10^{-19} \text{ J}$

De Broglie wavelength of the electron at the 1st energy level, $\lambda_1 = \frac{h}{\sqrt{2mE_1}}$

$$= \frac{6.63 \times 10^{-34} \text{ J.s}}{\sqrt{2 \times 9.11 \times 10^{-31} \text{ kg} \times 6.03 \times 10^{-20} \text{ J}}} = 2 \text{ nm.}$$

De Broglie wavelength of the electron at the 2nd energy level, $\lambda_2 = \frac{h}{\sqrt{2mE_2}}$

$$= \frac{6.63 \times 10^{-34} \text{ J.s}}{\sqrt{2 \times 9.11 \times 10^{-31} \text{ kg} \times 2.4 \times 10^{-19} \text{ J}}} = 1 \text{ nm.}$$

Example 20.14. Find the lowest energy of an electron confined to move in a one-dimensional box of length 1 Å. Express the result in electron volts.

Solution. The energy values of electron in a one-dimensional box are given by

$$E_n = \frac{n^2 h^2}{8mL^2}$$

For the lowest energy level $n = 1$.

$$\begin{aligned} \therefore E_1 &= \frac{(6.626 \times 10^{-34} \text{ J.s})^2}{8 \times 9.11 \times 10^{-31} \text{ kg} \times (10^{-10} \text{ m})^2} \\ &= 6 \times 10^{-18} \text{ J} = (6 \times 10^{-18})(6.242 \times 10^{18} \text{ eV}) = 37.4 \text{ eV}. \end{aligned}$$

Example 20.15: Compare the lowest three energy states for (i) an electron confined in the infinite potential well of width 10 \AA and (ii) a grain of dust ($m = 10^{-6} \text{ gm}$) moving with a speed of 10^6 m/s in an angle potential well of width 0.1 mm . What can you conclude from this comparison?

Solution: $E_n = \frac{n^2 h^2}{8mL^2}; \quad n = 1, 2, 3, \dots$

(i) Electron

$$E_1 = \frac{(1)^2 (6.626 \times 10^{-34} \text{ J.s})^2}{8(9.11 \times 10^{-31} \text{ kg})(10^{-9} \text{ m})^2}$$

$$E_1 = 6.03 \times 10^{-20} \text{ J} = 0.37 \text{ eV}$$

$$E_2 = 4 \times 6.03 \times 10^{-20} \text{ J} = 24 \times 10^{-20} \text{ J} = 1.48 \text{ eV}$$

$$E_3 = 9 \times 6.03 \times 10^{-20} \text{ J} = 54 \times 10^{-20} \text{ J} = 3.33 \text{ eV}$$

(ii) Grain of dust,

$$E_1 = \frac{(1)^2 (6.626 \times 10^{-34} \text{ J.s})^2}{8(10^{-9} \text{ kg})(10^{-4} \text{ m})^2} = 5.49 \times 10^{-51} \text{ J}$$

$$E_1 = 5.49 \times 10^{-51} \text{ J}$$

$$E_2 = 4 \times 5.49 \times 10^{-51} \text{ J} = 2 \times 10^{-51} \text{ J}$$

$$E_3 = 9 \times 5.49 \times 10^{-51} \text{ J} = 48.7 \times 10^{-51} \text{ J}$$

$$K.E. = \frac{1}{2}mv^2 = \frac{1}{2} \times 10^{-9} \text{ kg} \times (10^6 \text{ m/s})^2 = 500 \text{ J}$$

It may be inferred from the above calculations that

- (i) Energy levels of an electron in an infinite potential well are quantized and the energy difference between the successive levels is quite large. The electron cannot jump from one level to another level on the strength of the thermal energy. Hence quantization of energy plays a significant role in case of an electron.
- (ii) The energy levels of a grain of dust are so near to each other that they constitute a continuum. These energy values are far smaller than the kinetic energy ($= 500 \text{ J}$) possessed by the grain of dust. It can move through all these energy levels without an external supply of energy. Thus quantization of energy levels is not at all significant in case of macroscopic bodies.

20.24 HARMONIC OSCILLATOR

Atoms in solids execute harmonic vibrations about their equilibrium positions and hence can be treated as simple harmonic oscillators. As the atom displaces through a distance x , from the equilibrium position ($x = 0$), a restoring force proportional to x appears. It is given by

$$F = -kx$$

where k is the force constant. The potential energy of a harmonic oscillator is given by

$$V = \frac{1}{2}kx^2$$

The Schrödinger equation for this case may be written as

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2m}{h^2}(E - \frac{1}{2}kx^2)\psi = 0 \quad (20.109)$$

This equation has a solution only at the following values of energy

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega \quad (20.110)$$

where $n = 0, 1, 2, 3, \dots$

Thus the energy levels of a harmonic oscillator are quantized. The lowest value of the energy which the oscillator can take, corresponds to $n = 0$.

∴

$$E_0 = \frac{1}{2}\hbar\omega \quad (20.111)$$

This level is the ground state. The energy E_0 is called **zero-point energy**, as it does not vanish even at absolute zero temperature. The energy levels of the harmonic oscillator are shown in Fig. 20.27.

The energy levels are equidistant and are separated by an energy value

$$\Delta E = \hbar\omega \quad (20.112)$$

The existence of zero-point energy is the consequence of uncertainty principle.

At $T = 0K$, if the atomic vibrations come to a halt, it would imply that the position and momentum of the atoms can be determined precisely. But such a determination is not possible in the microworld, since the behaviour of micro particles obeys uncertainty principle.

20.25 THE WAVE MECHANICAL MODEL OF ATOM

We now apply the time-independent Schrödinger equation to the hydrogen atom and obtain the eigen values of energy and eigen functions related to the electronic states of the hydrogen atom. The Schrödinger equation for the electron in three dimensions is

$$\nabla^2\psi + \frac{8\pi^2m}{h^2}(E - V)\psi = 0$$

The potential V is that of a positive charge and negative charge separated by a distance ' r ' which is given by

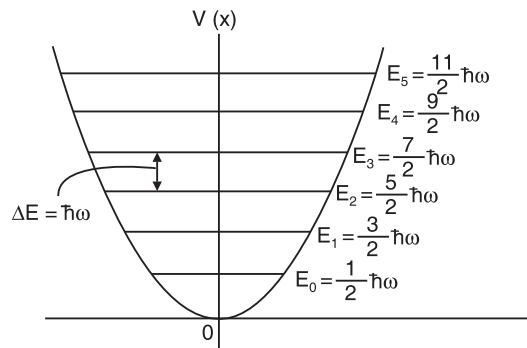


Fig. 20.27

$$V = -\frac{e^2}{4\pi\epsilon_0 r} \quad (20.113)$$

$$\therefore \nabla^2 \psi + \frac{8\pi^2 m}{h^2} \left(E + \frac{e^2}{4\pi\epsilon_0 r} \right) \psi = 0 \quad (20.114)$$

It is difficult to solve this partial differential equation. It can be easily solved if it is expressed in spherical polar coordinates. The equation takes the following form in spherical polar coordinates.

$$\frac{1}{r^2} \left[\frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right] \psi + \frac{8\pi^2 m}{h^2} \left(E + \frac{e^2}{4\pi\epsilon_0 r} \right) \psi = 0 \quad (20.115)$$

The above equation can be easily separated into three independent equations, each involving only one variable. After carrying out appropriate substitutions and separating the variables, we obtain the following three total differential equations.

Azimuthal wave equation:

$$\frac{d^2 \Phi}{d\phi^2} + m_l^2 \Phi = 0 \quad (20.116)$$

Polar wave equation:

$$\frac{1}{\sin \theta} \frac{d}{d\theta} \left(\sin \theta \frac{d\Theta}{d\theta} \right) + \left[l(l+1) - \frac{m_l^2}{\sin^2 \theta} \right] \Theta = 0 \quad (20.117)$$

Radial wave equation:

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) + \left[\frac{8\pi^2 m}{h^2} \left(E + \frac{e^2}{4\pi\epsilon_0 r} \right) - \frac{l(l+1)}{r^2} \right] R = 0 \quad (20.118)$$

The azimuthal wave equation (20.116) is a simple differential equation having the following solutions.

$$\Phi(\phi) = A e^{i m_l \phi}$$

The wave function Φ must have a single value at a given point in space. It can happen only when

$$m_l = 0, \pm 1, \pm 2, \pm 3, \dots \quad (20.119)$$

The polar equation (20.117) has a complicated solution in terms of associated Legendre functions. The solution exists only when the constant l is an integer equal to or greater than $|m_l|$. This condition implies that for a given value of l , m_l can have values

$$m_l = 0, \pm 1, \pm 2, \pm 3, \dots, \pm l. \quad (20.120)$$

The radial wave equation (20.118) also has a complicated solution in terms of polynomials called the associated Laguerre functions. In case of the electron bound to atom, the radial wave equation can be solved only under the following conditions.

1. E has one of the negative values E_n given by

$$E_n = -\frac{me^4}{8\epsilon_0^2 h^2} \cdot \frac{1}{n^2} \quad (20.121)$$

2. n is an integer equal to or greater than $(l + 1)$. This condition means that

$$l = 0, 1, 2, 3, \dots, (n - 1). \quad (20.122)$$

The energy eigen values specified by eq.(20.121) are precisely the same as those obtained by Bohr on the basis of the semi-classical model. The three quantum numbers needed to describe the three-dimensional motion of the electron in the hydrogen atom are n , l , and m_l . They must satisfy the following conditions:

- (i) Principal quantum number, $n \geq 1$; $n = 1, 2, 3, \dots$
- (ii) Orbital quantum number, $l \leq (n - 1)$; $l = 0, 1, 2, 3, \dots, (n - 1)$
- (iii) Orbital magnetic quantum number, $|m_l| \leq l$; $m_l = 0, \pm 1, \pm 2, \pm 3, \dots, \pm l$.

Using the above conditions, we can write down all the allowed eigen functions describing the various possible quantum states for the hydrogen atom by using the notation $\psi(n, l, m)$ for each state.

20.25.1 Wave Functions

The wave function ψ_1 of the ground state ($n = 1$) of the hydrogen atom is relatively simple to find out. As ψ_1 has to be completely spherically symmetric, we assume the solution for it.

$$\psi_1 = C_1 e^{-\alpha r} \quad (20.123)$$

is a constant which turns out to be the reciprocal of the value of the first Bohr radius. For a spherically symmetric volume element $4\pi r^2 dr$, the normalization constant C_1 has a value

$$C_1 = \frac{1}{\pi^{1/2} r_o^{3/2}}$$

Therefore, the wave function for the ground state of hydrogen is

$$\psi_1 = \left(\frac{1}{\sqrt{\pi r_o^{3/2}}} \right) e^{-r/r_o} \quad (20.124)$$

The wave function does not tell us where we actually would find the electron. The probability $P(r)dr$ of finding an electron at any radial distance within a spherically symmetric shell of thickness dr and volume $4\pi r^2 dr$ is given by

$$\psi_1^* \psi_1 4\pi r^2 dr = 4\pi r^2 C_1^2 e^{-2r/r_o} dr$$

or

$$P(r)dr = \left[\frac{1}{\pi r_o^3} \exp\left(-\frac{2r}{r_o}\right) \right] 4\pi r^2 dr \quad (20.125)$$

The probability $P(r)dr$ of finding the electron at a given distance from the nucleus is plotted in Fig. 20.28. The curve for $n = 1$ shows a maximum at a distance r_o . The total probability of finding the electron at all points of distance r from the nucleus is greatest when $r = r_o$. This value is the same as the value determined by the Bohr theory for the radius of the first orbit. In wave mechanics, r_o is the distance at which the electron is likely to be found most often. It is not possible to draw a shape that bounds a region in which the probability of

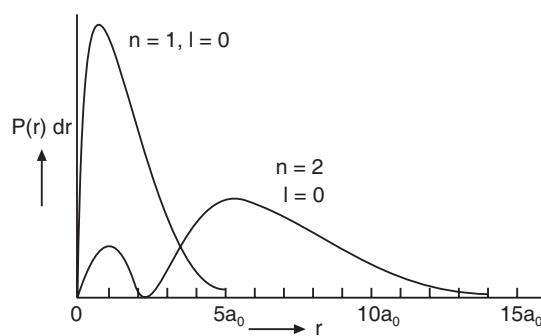


Fig. 20.28: The probability distribution as function of r . In the figure, the x-axis represents the radial distance r in units of a_0 , with major ticks at $0, 5a_0, 10a_0, 15a_0$. The y-axis represents the probability density $P(r)dr$. Two curves are shown: one for the $n=1, l=0$ state, which is a sharp peak centered at $r=0$, and another for the $n=2, l=0$ state, which is a broader peak centered at $r=7.5a_0$.

finding the electron is 100%. The electron probability is distributed over the entire volume of the atom. However, a surface can be drawn that connects points of equal probability and that encloses a volume in which the probability of finding the electron is high. An electron in an atom may therefore be visualized as an electron cloud. There is a region where the electron probability density falls steeply to a low value and it corresponds to the boundary of the atom. The concept of orbit is replaced with that of an orbital. An **orbital** is the region around the nucleus in which the probability of finding the electron is the highest. Depending upon the configuration and size of the boundary surface, the shape and size of the orbital is determined.

20.25.2 Orbital Angular Momentum

It is found that like total energy E , the angular momentum of electron is both conserved and quantized. Quantum theory shows that the magnitude L of the electron orbital angular momentum L is given by

$$L = \sqrt{l(l+1)}\hbar \quad (20.126)$$

The law of conservation angular momentum implies that an electron in a definite stationary state can have a definite angular momentum. However ($L = mvr$) contains simultaneously both $p (= mv)$ and r . The uncertainty principle forbids a particle to have a definite momentum p and a definite coordinate r at the same time. The quantum theory shows that the conserved angular momentum of the electron can be characterized by the magnitude L and one of its rectangular components, L_z . The component L_z is also quantized which implies the quantization of direction of L . Thus, space quantization emerges as a natural element of quantum mechanical solution of electron motion in the atom.

$$L_z = m_l \hbar \quad (20.127)$$

The choice of L_z among the three components L_x, L_y, L_z is simply a matter of convention where a magnetic field is assumed parallel to z-direction for the sake of reference. When L_z is chosen to have a definite value, then the other components L_x and L_y do not have well defined values; one can only determine the probabilities of specific values of L_x and L_y .

The simultaneous quantization of L and L_z implies that the vector L can never be fixed in space pointing any specific direction. In fact, the direction of L constantly changes and it precesses around the z-axis tracing out a cone in space. The tilt angle of L is determined by m_l .

$$\cos \theta = \frac{L_z}{L} = \frac{m_l}{\sqrt{l(l+1)}} \quad (20.128)$$

20.25.3 Intrinsic Angular Momentum

Electron is known to possess spin angular momentum. Schrödinger equation does not give any hint regarding electron spin. When relativity is included into the quantum mechanical treatment of electron motion in an atom, the intrinsic non-orbital angular momentum of the electron emerges in a natural way. The relativistic wave equation was first formulated in 1928 by P.A.M. Dirac (1902–1984). According to Dirac's theory, the only value that the spin

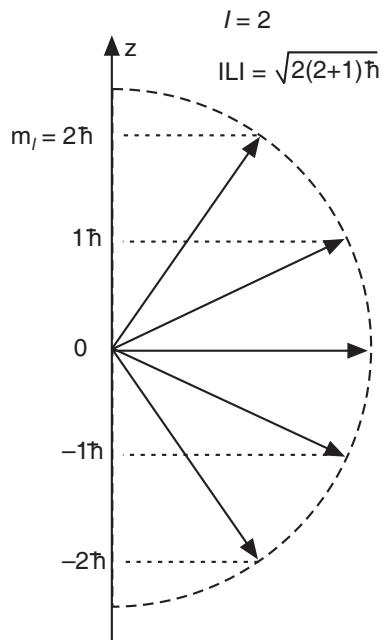


Fig. 20.29

quantum numbers can have is $\frac{1}{2}$. The magnitude S of the intrinsic angular momentum of the electron is given by

$$S = \sqrt{s(s+1)}\hbar \quad (20.129)$$

The direction of this intrinsic angular momentum is also quantized. The space quantization of intrinsic angular momentum is described by the spin magnetic quantum number m_s . The components S_z along a magnetic field in the z-direction is given by

$$S_z = m_s\hbar \quad (20.130)$$

20.25.4 Selection Rules

The fact that each level in a hydrogen-like atom is composed of several angular momentum states is important from the point of view of light emission. The electron transitions are restricted by **selection rules**. The selection rules for electric dipole transitions are

$$\Delta l = \pm 1 \text{ and } \Delta m_l = 0, \pm 1$$

These selection rules are imposed by the law of conservation of angular momentum since the emitted or absorbed photon carries a spin angular momentum of 1. Therefore, the angular momentum of the atom must change by one unit to compensate for the angular momentum carried by the emitted or absorbed photon.

20.26 THE TRANSITION FROM DETERMINISTIC TO PROBABILISTIC NATURE

The probabilistic interpretation is a fundamental feature of quantum mechanics. Every particle has a wave function associated with it. The wave function determines the corpuscular characteristics like position, momentum etc in a statistical sense. For instance a particle with a wave function $\psi(x, t)$ has a probability proportional to $|\psi(x, t)|^2$ for being found in the neighbourhood of the point x . *Nature does not permit any more precise specification of the state of a particle than what is provided by the wave function and the probability distributions obtainable from it.*

Quantum mechanics yields probability functions, not precise trajectories.

It should be noted that due to uncertainty principle, our idea regarding causality needs revision. The classical concept of cause and effect applies to systems which are left undisturbed. Since a small system cannot be observed without producing serious disturbances in it, we cannot expect to find causal connection between the results of our observations. There is an unavoidable indeterminacy in the calculation of the observational results. Quantum mechanics enables us only to calculate the probability of our obtaining a particular result when we make an observation.

In classical mechanics the motion of a body under the action of a force is described by Newton's second law. If the initial position and the initial velocity are specified, the position and velocity of the body at any next instant can be found by Newton's second law. Thus, giving the position and velocity of a body at any instant chosen as reference ($t = 0$) will fully

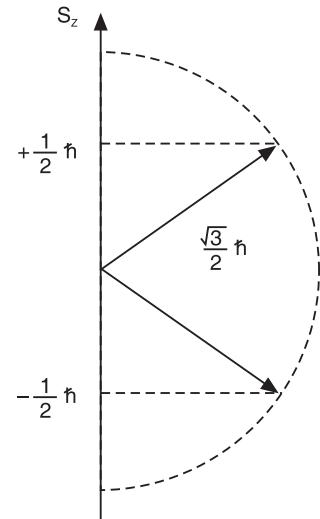


Fig.20.30

specify the state of the body. If one knows the state of the system at some instant, one can accurately predict the state of the system at any other instant. This is known as **mechanical determinism**. It is assumed that the accuracy with which the initial position and velocity of bodies can be specified is solely dependent on the quality of the instruments measuring position and velocity.

In contrast, since the position and velocity of quantum mechanical particles can be determined simultaneously only with uncertainties given by Heisenberg relation, one is prompted to conclude that the state of a system cannot be accurately specified at the initial instant, t_0 , or at any other succeeding instants. The probabilistic interpretation of the wave function makes quantum mechanics an inherently statistical theory.

20.27 SUPERPOSITION PRINCIPLE

One of the most fundamental ideas of quantum mechanics is that its equations are linear with respect to the wave function ψ . Thus sum of two solutions of a linear equation again satisfies the same equation. It follows from this that any solution of a wave equation can be represented in the form of a certain set of standard solutions. This statement concerning the possibility of representing a single wave function in terms of the sum of other wave functions is called the **superposition principle**. The *principle of superposition of states* is one of the important principles of quantum mechanics.

According to this principle, if a certain quantum-mechanical system has an allowed state ψ' and also state ψ'' , then the system can also exist in a state described by the function

$$\psi = c' \psi' + c'' \psi'' \quad (20.131)$$

where c' and c'' are arbitrary complex numbers.

Given several solutions of the time-independent Schrödinger equation with different values E_n of the energy, any linear superposition of the form

$$\psi = \sum_n c_n \psi_n$$

will also be a solution of the Schrödinger wave equation.

Let ψ_1 and ψ_2 be two wave functions representing two physically allowed states of a given object. Then any linear superposition of the form

$$\psi = a_1 \psi_1 + a_2 \psi_2 \quad (20.132)$$

also describes the state of the object. a_1 and a_2 are arbitrary complex coefficients. Repeated application of the superposition principle shows that if the principle holds for the superposition of two wave functions, it also holds for the superposition of an arbitrary number of wave functions.

20.28 OBSERVABLES AND OPERATORS

An **observable** is a quantity obtained through the process of measurement on a physical system. The measurement of an observable in a system is expressed as a number. An observable is always a *real quantity* as it is the result of an actual measurement.

An **operator** is an entity which when applied to a function, transforms it into a new function.

An operator is denoted by the symbol \hat{L} . The equation

$$\hat{L}\psi(x) = \phi(x) \quad (20.133)$$

means that the operator denoted by \hat{L} acts on the function $\psi(x)$ and as a result we get the function $\phi(x)$. The action could be a *mathematical operation* in algebra such as to multiply the input function by a constant, add something else to it, or an operation in calculus such as to differentiate it, integrate it, or so on. A quantum mechanical operator \hat{L} does not work on an algebraic function but on a state vector like $|\psi\rangle$, which is an abstract description of a physical situation. An operator performs an operation on a state $|\psi\rangle$ and produces a new state $|\phi\rangle$.

The operator \hat{L} is called a **linear operator** if it satisfies the following conditions:

$$\hat{L}(\psi_1 + \psi_2) = \hat{L}\psi_1 + \hat{L}\psi_2 \text{ and } \hat{L}(a\psi) = a\hat{L}\psi \quad (20.134)$$

where a is some number. We are concerned here with linear operators only.

The effect of an operator acting on a function may be represented as a definite or an improper integral.

$$\hat{L}\psi(x) = \int L(x, y)\psi(y)dy$$

If the variable is discrete, we will have

$$\hat{L}\psi_n = \sum_m L_{nm}\psi_m \quad (20.135)$$

The totality of coefficients L_{nm} forms the *matrix of the operator* \hat{L} and we speak of the matrix representation of the operator.

Suppose $\hat{L}\psi = \varphi$; the operator \hat{L}^* is called the *complex conjugate* of the operator \hat{L} , if by the action of this operation on the function ψ^* we get the function φ^* :

$$\hat{L}^*\psi^*(x) = \varphi^*(x)$$

The expectation value of a dynamical variable gives the expected average of the result of measurement of the dynamical variable. So it must be a real number since it is a physically measurable quantity. The operators representing dynamical variables whose expectation values are real are said to be **Hermitian operators**. If \hat{A} is the operator, then the expectation value of A is given by

$$\langle A \rangle = \int \psi^* \hat{A}\psi dx \quad (20.136)$$

where ψ is a normalised wave function specifying the state of the system. The complex conjugate of the expectation value of A is

$$\langle A \rangle^* = \left(\int \psi^* \hat{A}\psi dx \right)^* = \int \psi \hat{A}^* \psi^* dx$$

For the expectation value to be real, we must have $\langle A \rangle = \langle A \rangle^*$.

or
$$\int \psi^* \hat{A}\psi dx = \int \psi \hat{A}^* \psi^* dx \quad (20.137)$$

An operator is said to be Hermitian if it satisfies the above condition.

A linear operator commutes with constants and obeys distributive and associative laws. The sum of two linear operators is commutative, but their product may or may not be commutative. Thus, if \hat{A} and \hat{B} are two linear operators, then these may or may not commute. It means that the product operator $\hat{A}\hat{B}$ may or may not be equal to the operator $\hat{B}\hat{A}$. The **commutator** of two operators \hat{A} and \hat{B} corresponding to two physical quantities A and B is defined as

$$[\hat{A}, \hat{B}] = \hat{A}\hat{B} - \hat{B}\hat{A} \quad (20.138)$$

When \hat{A} and \hat{B} commute, the commutator vanishes. That is,

$$[\hat{A}, \hat{B}] = 0$$

If $[\hat{A}, \hat{B}]\psi = 0$, then the eigen values of A and B can be measured simultaneously.

If \hat{A} and \hat{B} do not commute, then the commutator $[\hat{A}, \hat{B}] \neq 0$. If $[\hat{A}, \hat{B}]$ when operated on some wave function ψ yields a non-zero result, then there will be an uncertainty relation limiting A and B of the form

$$(\Delta A)(\Delta B) \geq \frac{1}{2}\sqrt{-[A, B]^2}$$

Pairs of physical variables that are linked by an uncertainty relation are known as **conjugate or complementary variables**.

20.29 IMPORTANT OPERATORS OF QUANTUM MECHANICS

In quantum mechanics an operator when applied to a wave function gives the corresponding observable quantity of the system multiplied by the wave function. The most important operators of quantum mechanics are position operator, momentum operator, energy operator and Hamiltonian operator.

Let us consider the wave function $\psi = A \exp\left[\frac{i}{\hbar}(p_x x - Et)\right]$. (20.139)

Differentiating the above expression with respect to x , we obtain

$$\begin{aligned} \frac{\partial \psi}{\partial x} &= A \frac{ip_x}{\hbar} \exp\left[\frac{i}{\hbar}(p_x x - Et)\right] \\ \text{or } \frac{\hbar}{i} \frac{\partial}{\partial x} \Psi &= p_x \Psi \end{aligned}$$

From the above we see that the operator corresponding to the observable quantity momentum p_x can be represented by

$$\hat{p}_x = \frac{\hbar}{i} \frac{\partial}{\partial x} = -i\hbar \frac{\partial}{\partial x} \quad (20.140)$$

In three dimensions the operator is given by $\hat{p} \rightarrow -i\hbar\nabla$.

Now, the linear momentum operator \hat{p}_x when applied to the wave function ψ gives the corresponding observable quantity, linear momentum p_x multiplied by the wave function.

Similarly, when the eqn.(20.139) is differentiated with respect to t , we get

$$\begin{aligned} \frac{\partial \psi}{\partial t} &= -\frac{iE}{\hbar} A \exp\left[\frac{i}{\hbar}(p_x x - Et)\right] = -\frac{iE}{\hbar} \Psi \\ \text{or } i\hbar \frac{\partial}{\partial t} \Psi &= E\Psi \end{aligned}$$

We denote the operator E by \hat{E} . When \hat{E} operates on the wave function ψ , we get the total observable property ‘total energy’ E multiplied by the wave function.

$$\hat{E}_\psi = i\hbar \frac{\partial}{\partial t} \Psi = E\Psi \quad (20.141)$$

The operator corresponding to energy is thus $\hat{E} \rightarrow i\hbar \frac{\partial}{\partial t}$. Thus, the dynamical variables momentum and energy of a quantum particle can be represented by the two mathematical

operators $-i\hbar \frac{\partial}{\partial x}$ and $i\hbar \frac{\partial}{\partial t}$ respectively. In addition since the coordinate x is a multiplying operator, we get $x \rightarrow \hat{x} = x$.

Some useful physical operators are given in the following table.

Dynamical variable	Quantum Mechanical Operator	Result (Multiplication of ψ by)
Position, x	\hat{x}	x
Linear momentum, p	$-i\hbar \nabla$	p
Kinetic Energy, E	$-\frac{\hbar^2}{2m} \nabla^2$	E
Potential Energy	$V(r)$	V

In classical mechanics, Hamiltonian is defined as $H = \frac{p^2}{2m} + V$ and it gives the total energy E of the system. Hence, the total energy operator for a non-relativistic particle is represented by Hamiltonian operator \hat{H} . As $\hat{p} = i\hbar \nabla$, the Hamiltonian operator is

$$\hat{H} = -\frac{\hbar^2}{2m} \nabla^2 + V \quad (20.142)$$

When the Hamiltonian operator operates on the function ψ , the result is the function multiplied by total energy E . Thus,

$$\hat{H}\Psi = \left[-\frac{\hbar^2}{2m} \nabla^2 + V(r) \right] \Psi = E \Psi$$

It may be noted here that the quantum mechanical operators for x and p_x satisfy the following commutation relation.

$$[\hat{x}, p_x] = \left[\hat{x}, \frac{\hbar}{i} \frac{\partial}{\partial x} \right] = -i\hbar \left(x \frac{\partial}{\partial x} - \frac{\partial}{\partial x} x \right) = i\hbar$$

Thus, the commutator $[\hat{x}, \hat{p}_x] \neq 0$. Therefore, the precise simultaneous measurement of these variables is not possible. If the operators corresponding to variables are commutative, that is, $[\hat{A}, \hat{B}] = 0$, then the variables can be determined to a high precision simultaneously.

20.30 EXPECTATION VALUES

Some questions in quantum contexts will yield exact answers whereas others can only be answered in terms of a probability distribution. For example, if we ask about the energy of an electron in 1s orbit of the hydrogen atom, the answer is definite and is -13.6 eV. But if we ask what its position is, our answer is not definite. If we take a large number of hydrogen atoms and somehow measure the positions of electrons in them simultaneously, and average the results, we get the *average* position of the electron from the nucleus. The theoretically predicted value of this average is known as the **expectation value** of the position. It is designated by $\langle x \rangle$.

$$\text{Average value of } x \text{ is } \langle x \rangle = \frac{x_1 + x_2 + x_3 + \dots + x_i + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Thus, if a variable takes n discrete values, we add all these values of the variable and divide the sum by the total number of values to obtain the average value of that variable. On the other hand if the variable takes the value of $x_1 m_1$ times, the value $x_2 m_2$, the value $x_3 m_3$ and so on, then

$$\begin{aligned}\text{Average value of } x \text{ is } \langle x \rangle &= \frac{m_1 x_1 + m_2 x_2 + m_3 x_3 + \dots + m_i x_i + \dots + m_n x_n}{m_1 + m_2 + \dots + m_i + \dots + m_n} \\ &= \frac{\sum_{i=1}^n m_i x_i}{\sum_{i=1}^n m_i} = \frac{\sum_{i=1}^n m_i x_i}{n}\end{aligned}$$

or

$$\langle x \rangle = \sum_{i=1}^n x_i \frac{m_i}{n}$$

But $\frac{m_i}{n} = P_i$, the probability of occurrence of the value x_i .

$$\therefore \langle x \rangle = \sum_{i=1}^n x_i P_i$$

If the variable is continuous instead of discrete, then in place of summation we use the integral and get

$$\langle x \rangle = \int x P dx \quad (20.143)$$

Given a probability distribution for any measurable physical quantity A, the simplest and most important property of that distribution is the average value of A, designated as $\langle A \rangle$ and understood as an average taken over many individual measurements under identical conditions on the same initial state.

20.30.1 Expectation Value in Operator Notation

When a given system is in a state ψ , the average of a large number of independent measurements of an observable corresponding to the operator \hat{L} is given by

$$\langle L \rangle = \frac{\int_{\text{Available region}} \psi^* \hat{L} \psi dr}{\int_{\text{Available region}} \psi^* \psi dr} \quad (20.144)$$

$\langle L \rangle$ is called the expectation value of the observable. For a normalised function, $\int \psi^* \psi dr = 1$.

$$\therefore \langle L \rangle = \int_{\text{Available region}} \psi^* \hat{L} \psi dr \quad (20.145)$$

The general rule for calculating expectation values is

$$\langle \text{Observable} \rangle = \int \psi^*(x) \{(\text{operator})\} \psi(x) dx \quad (20.146)$$

where operator denotes the operator corresponding to the observable of interest. The order of computation is important here. First $\{\text{operator}\psi(x)\}$ is computed. Then, the result is multiplied by ψ^* and finally the integral is carried out.

Position: The simplest example is the expectation value of the position \mathbf{r} of an object. From eq. (20.145), it follows that

$$\langle r \rangle = \int_{-\infty}^{+\infty} \psi^* \hat{r} \psi \, dr = \int_{-\infty}^{+\infty} \psi^* \mathbf{r} \psi \, dr = \int_{-\infty}^{+\infty} \mathbf{r} |\psi|^2 \, dr \quad (20.147)$$

The operator $\hat{r} \equiv r$ operates on the wave function ψ to its right, the operation being ordinary multiplication with the vector \mathbf{r} . Thus, \hat{r} is the *position operator*. It is an **algebraic operator**.

Momentum: The expectation value for the momentum \mathbf{p} is given by

$$\langle p \rangle = \int \psi^* \hat{p} \psi \, dr = \int \psi^* (-i \hbar \nabla) \psi \, dr$$

In one-dimensional case $p_x = -i \hbar \frac{\partial}{\partial x}$.

$$\therefore \langle p_x \rangle = -i \hbar \int \psi^*(x) \frac{\partial}{\partial x} \psi(x) \quad (20.148)$$

Integrals of the general form of equation (20.148) arise frequently in quantum mechanics. To reduce the tedium of writing these out, it is convenient to adopt a short hand notation introduced by Dirac and known as Dirac “**bra-ket**” notation. In this notation, the expectation value of $f(x)$ is written simply as

$$\langle \Psi^*(x) | f(x) | \Psi(x) \rangle$$

$$\text{Thus, } \langle \Psi^*(x) | f(x) | \Psi(x) \rangle = \int \psi^*(x) \{ \text{Op}[f(x)] \} \psi(x) dx$$

where $|f(x)|$ denotes $\text{Op}[f(x)]$ which is the operator corresponding to $f(x)$.

Example 20.16: Find the expectation value $\langle x \rangle$ of the position of a particle trapped in a box of width L .

Solution:

$$\begin{aligned} \langle x \rangle &= \int_{-\infty}^{\infty} x |\psi|^2 dx \\ &= \frac{2}{L} \int_0^L x \sin^2 \frac{n\pi x}{L} dx \\ &= \frac{2}{L} \left[\frac{x^2}{4} - \frac{x \sin(2n\pi x/L)}{4n\pi/L} - \frac{\cos(2n\pi x/L)}{8(n\pi/L)^2} \right]_0^L \end{aligned}$$

or

$$\langle x \rangle = \frac{2}{L} \left[\frac{L^2}{4} \right] = \frac{L}{2}.$$

Thus, the average position of the particle is the middle of the box in all quantum states.

20.31 FUNDAMENTAL POSTULATES OF QUANTUM MECHANICS

We summarize here the fundamental postulates which we have introduced in the course of our above study.

Postulate I

- (a) The state of a quantum mechanical system is described or represented by a wave function $\psi(x, t)$. All constant multiples of a given ψ describe one and the same state.
- (b) The state ψ of the system can be built up by applying the principle of superposition. Thus,

$$\psi = \sum_n c_n \psi_n$$

where c_n are complex numbers.

Postulate II

Each dynamical variable $A(x, p)$ is represented by a linear operator in quantum mechanics. Dynamical variables in general do not commute with each other.

Postulate III

If large number of measurements of a dynamical variable A are made on a system which is prepared to be in one and the same state before each measurement, the results of the measurements are distributed over an average value known as expectation value $\langle A \rangle$. The expectation value of the dynamical variable is given by

$$\langle A \rangle = \frac{\int \psi^* \hat{A} \psi \, dt}{\int \psi \psi^* \, dt}$$

Postulate IV

The only possible results of measurement of a dynamical variable A are the eigen values of the operator \hat{A} satisfying the eigen value equation

$$\hat{A} u_n = a_n u_n$$

where u_n is the eigenfunction of the operator \hat{A} belonging to the eigen value a_n .

The eigen value equation satisfied by the Hamiltonian operator associated with the Hamiltonian function H can be written as

$$\hat{H} \psi_n = E_n \psi_n$$

where E_n is the energy of the system. The above equation is the Schrödinger equation.

QUESTIONS

1. What is de Broglie hypothesis?
2. What is a matter wave? Explain its significance with change in mass of particle. (C.S.V.T.U.,2007)
3. Show that the de Broglie wavelength for electron is found to be equal to $\frac{12.26}{\sqrt{V}}$ Å.
4. What are matter waves? Describe the experiment that supports the existence of matter waves. (R.T.M.N.U.,2005)
5. Why is the wave nature of matter not apparent to our daily observation? Give a suitable example to illustrate the point.
6. Describe an experiment which proves the validity of de-Broglie hypothesis regarding wave nature of matter.

7. Describe Davisson and Germer experiment. What does it confirm? **(R.T.M.N.U.,2007)**
8. Calculate the wavelengths of electron using Bragg law and de Broglie hypothesis using the data as obtained in the Davisson and Germer experiment. Explain the importance of results obtained. **(R.T.M.N.U., 2006)**
9. What are matter waves? Why does a single monochromatic wave not represent a localized particle? Explain synthesis of a wave packet.
10. Explain in detail the experiment which establishes the relationship between momentum and wavelength of a moving particle. **(RGPV, 2008)**
11. Explain the duality of matter waves from the inferences drawn from photoelectric effect and Davisson-Germer effect. **(VTU, 2007)**
12. Explain the de Broglie concept of matter waves. How is the wave nature of electron demonstrated experimentally? Explain with the help of an experiment. **(R.T.M.N.U.,2006)**
13. Discuss de-Broglie's concept of matter waves in light of wave-particle dualism of radiation.
14. Distinguish between phase and group velocities. Show that the de Broglie wave group (wave packet) associated with a moving particle travels with same velocity as that of particle.
15. Explain the terms phase velocity, group velocity and wave packet. **(RGPV, 2008)**
16. Define phase velocity and group velocity. Derive an expression for de-Broglie wavelength from group velocity. **(VTU, 2008)**
17. Derive a relationship between group and phase velocity of matter waves. **(RGPV,2007)**
18. Show that the phase velocity of de Broglie wave is greater than the velocity of light, but the group velocity is equal to the velocity of the particle with which the waves are associated. **(UPTU, Lucknow)**
19. Define group velocity and obtain an expression for the same. **(VTU, 2007)**
20. How could Davison and Germer be sure that the peak obtained for 54 volts electron was a first order diffraction peak?
21. What do you understand by a wave packet? Using the concept of matte waves, obtain the Bohr's condition for quantization of angular momentum.
22. Show how the quantization of angular momentum follows from the concept of matter waves.
23. Describe an experiment which supports the existence of matter waves.
24. Explain de Broglie's concept of matter waves. Give an account of Davisson and Germer experiment to show the wavelike character of a beam of electrons. **(C.S.V.T.U.,2007)**
25. How Bohr's condition of stationary orbits of an atom can be obtained from concept of matter wave?
26. Discuss similarities and differences between a matter wave and an electromagnetic wave. Why is the wave nature of matter less apparent in our daily observations?
27. Explain how de Broglie hypothesis regarding wave nature of matter leads to Bohr's quantization condition for angular momentum of an electron. **(R.T.M.N.U.,2007)**
28. Describe the construction, and working of an electron microscope. **(M.G.Univ.,2005, 2006)**
29. With a neat sketch, explain the working of an electron microscope. Which are the features that differentiate it from an optical microscope? **(M.G.Univ.,2005)**
30. Describe the construction, principle, and working of a scanning electron microscope.
31. (a) An electron microscope is made by using short wavelength electrons to provide high resolution. Then is it possible to make a proton microscope and a neutron microscope? Explain your answer in each case.
 (b) If the particles listed below all have the same energy, which has the shortest wavelength: electron, α -particle, neutron, proton?
 (c) "If an electron is localized in space, its momentum becomes uncertain. If it is localized in time, its energy becomes uncertain."Explain this statement. **(Bombay Univ.)**
32. Write the statement of Heisenberg uncertainty principle. **(C.S.V.T.U.,2008)**
33. State Heisenberg uncertainty principle. Show that electrons cannot exist within the nucleus on the basis of the above principle.

34. Arrive at Heisenberg's uncertainty principle with the help of a thought experiment.
35. Show that the uncertainty in measurement of electron momentum is equal to its momentum itself; if the uncertainty in measurement of location of electron is equal to its de Broglie wavelength.
36. State uncertainty principle. Write its mathematical form for the following pairs of variables:
 (a) Position and momentum
 (b) Energy and time
 (c) Angular position and angular momentum.
37. What is Heisenberg Uncertainty principle? Explain how it is the outcome of the wave description of a particle. Arrive at Heisenberg uncertainty principle with the help of a simple thought experiment. **(C.S.V.T.U., 2006, 2007)**
38. What is Heisenberg uncertainty principle? Describe a suitable thought experiment to support uncertainty principle. **(R.T.M.N.U., 2005, 2007)**
39. Explain Heisenberg uncertainty principle. Based on this, show the non-existence of electrons inside the nucleus. **(VTU, 2008)**
40. Can a wave given by an equation $Y = A \sin(wt - kx)$ represent a particle? Explain the concept of a wave packet. How does this concept lead to Heisenberg's uncertainty principle?
41. Explain why a single monochromatic wave cannot represent a particle. **(R.T.M.N.U., 2006)**
42. Explain the physical significance of wave function ψ . **(Amaravati Univ., 2008), (C.S.V.T.U., 2007), (R.T.M.N.U., 2006)**
43. What is the physical significance of wave function Ψ ? **(R.T.M.N.U., 2005, 2007)**
44. Define a wave function. Show that it represents the probability density of finding a particle at a given position and given time.
45. What is a wave function? What are the necessary conditions of physically acceptable wave function? **(RGPV, 2008)**
46. Define wave function. What is meant by normalized wave function?
47. Explain concept of electron wave. Give physical significance of wave function. **(Amaravati Univ., 2006)**
48. Write down Schrödinger's time dependent and time independent wave equations of matter waves. Explain, why: (i) the wave function Ψ must be single valued and continuous function of position. (ii) The integral of $|\Psi|^2$ overall space must be equal to unity. **(R.T.M.N.U., 2007)**
49. Derive time dependent Schrödinger wave equation. **(UPTU, Lucknow)**
50. (a) Derive the time independent Schrödinger's wave equation.
 (b) What is meant by expectation value in quantum mechanics? **(Calicut Univ., 2006)**
51. State time independent Schrödinger equation. Find out the wave function associated with free electron and discuss the relationship between energy and wave vector in case of free electrons. **(RGPV, 2008)**
52. Write down:
 (a) time independent and
 (b) time dependent Schrödinger equations. **(R.T.M.N.U., 2005)**
53. (a) Derive time independent Schrodinger equation.
 (b) What is the physical significance of wave equation ψ ?
 (c) State and explain Heisenberg's uncertainty principle. **(Calicut Univ., 2007)**
54. Derive the time dependent Schrodinger equation for a free particle.
55. Distinguish between Newtonian and quantum mechanics. **(Calicut Univ., 2005)**
56. (a) Derive time independent Schrödinger's wave equation.
 (b) Apply the above equation for a particle confined to a rigid box and discuss its wave functions and energy levels. **(Andhra Univ.)**
57. Show that the energy of an electron confined in a 1-D potential well of length L and infinite depth is quantized. Is the electron trapped in potential well allowed to take zero energy? If not, why?

58. Show that the solution of Schrödinger's equation for a particle in an infinite potential well leads to the concept of quantization of energy. **(R.T.M.N.U.,2006)**
59. Assuming the time independent Schrödinger wave equation, discuss the solution for a particle in one dimensional potential well of infinite height. Hence obtain the normalized wave function. **(VTU, 2008)**
60. Show that the wave function for a particle confined in an infinite one-dimensional potential well of length ' l ' is given by $\psi_n(x) = \sqrt{\frac{2}{l}} \sin\left(\frac{n\pi x}{l}\right)$. Hence discuss the energy levels and their discreteness.
61. Show that the energy of a microparticle confined in an infinite one-dimensional potential well of length ' l ' is given by

$$E_n = \frac{n^2 h^2}{8mL^2}$$

where the symbols have their usual meaning. In the above situation the particle cannot have zero energy. Explain, why? **(R.T.M.N.U.,2007)**

62. Show that the wave function for a particle confined in an infinite one-dimensional potential well of length ' l ' is given by $\psi_n(x) = A \sin\left(\frac{n\pi x}{l}\right)$. Hence using normalization condition on ψ show that A is given by $\sqrt{\frac{2}{l}}$. **(R.T.M.N.U., 2007)**

63. Derive the eigen values and eigen functions for a particle in a one dimensional box. **(Calicut Univ.,2005)**
64. Explain the terms potential well and potential barrier. How does a particle with energy lower than the barrier height, tunnel through it? Give one example.
65. Explain in short the phenomenon of tunneling that occurs when a beam of particles are incident on a potential barrier of finite width. **(R.T.M.N.U., 2006)**
66. Explain the barrier tunneling of electron on the basis of quantum mechanics. **(R.T.M.N.U.,2006)**
67. Formulate Schrödinger wave equation for a linear harmonic oscillator.
68. Write the Schrödinger wave equation of the hydrogen atom in polar coordinates. Explain the origin and significance of the quantum numbers n , l and m_l .
69. (a) Mention the ideas which prompted de Broglie to propose his concept of matter waves.
 (b) Derive an expression for the de Broglie wavelength of an electron.
 (c) Describe the experimental verification of matter waves using Davisson-Germer experiment. **(JNTU, 2010)**

70. (a) Write the statement of Heisenberg Uncertainty Principle. Why is it significant only for sub-atomic particles, and not for heavy bodies ?
 (b) Distinguish between phase velocity and group velocity. **(RTMNU, 2010)**
71. (a) Starting from Schrodinger's time independent equation, show that the energy of a particle in one-dimensional potential well of infinite height is quantized. Hence obtain eigen function for the particle. Show necessary wave forms.
 (b) Explain the tunneling effect **(RTMNU, 2010)**
72. Explain the concept of group waves. Define group velocity and particle velocity in reference of a group wave and establish relation between group and particle velocity. **(RTMNU, 2010)**

PROBLEMS

1. Calculate the wavelength associated with 1 MeV electron. [Ans: 1.23×10^{-12} m]
2. Calculate the energy of the neutron in eV if its de Broglie wavelength is 3×10^{-10} m. [Ans: 9×10^{-3} eV]
3. Calculate the velocity and kinetic energy of an electron of wavelength 1.66 Å. [Ans: 4386 km/s, 54.7 eV]
4. Calculate de Broglie wavelength of an electron moving with velocity 10^7 m/s. [Ans: 0.72 Å]
5. Calculate the de Broglie wavelength of an α -particle accelerated through a potential difference of 400 volts. [Ans: 1.6×10^{-13} m]
6. Calculate the de Broglie wavelength of neutron of energy 12.8 MeV. [Ans: 8×10^{-5} Å]
7. Determine the velocity and kinetic energy of a neutron having de Broglie wavelength of 1 Å. [Ans: 3.97×10^3 m/s, 0.0825 eV]
8. An electron is confined to a spherical box of diameter 10^{-8} m. Calculate the minimum uncertainty in its velocity. [Ans: 1.16×10^4 m/s]
9. If the uncertainty in the position of an electron is 4×10^{-10} m, calculate the uncertainty in its momentum. [Ans: 26.4×10^{-26} kg.m/s]
10. An electron has a speed of 800 m/s with an accuracy of 0.004%. Calculate the certainty with which we can locate the position of the electron. [Ans: 3.6×10^{-3} m]
11. Compute the minimum uncertainty in the location of a mass of 2.0 gm moving with a speed of 1.5 m/s and the minimum uncertainty in the location of an electron moving with a speed of 0.5×10^8 m/s given that the uncertainty in the momentum p for both is $\Delta p = 10^{-3} p$. [Ans: 23.2 Å]
12. A meson has a life time of 10^{-23} s. To what accuracy can its mass be known? [Ans: 1.17×10^{-28} kg]
13. An electron is trapped in an infinite potential well of length 2 Å. In the ground state, evaluate the probability of finding the electron in the region $x = 0$ to 0.25 Å. [Ans: 0.0125]
14. An electron is moving in a one-dimensional infinite well of width 2×10^{-10} m. What is the probability of finding the electron between $x = 0$ and $x = 10^{-10}$ m in the first excited state? [Ans: 0.5]
15. A particle is moving in one dimension infinite potential box of width 25 Å. Calculate the probability of finding the particle within a small interval of 0.05 Å at the centre of the box when it is in its state of least energy. [Ans: 0.022]
16. The position and momentum of a 1 keV electron are simultaneously determined. If its position is located to within 1 Å, what is the percentage of uncertainty in its momentum? [Ans: 6.17%]
17. An electron is confined to a box of length 100 Å. Calculate the minimum uncertainty in its velocity. [Ans: 72.8 km/s]
18. Compare the uncertainties in the velocities of an electron and a proton confined in a 10 Å box. [Ans: 1.16×10^5 m/s; 63 m/s]
19. Calculate the probability of transmission for a proton of energy 1 MeV through a 4 MeV high rectangular potential energy barrier of width 10^{-2} cm. [Ans: 0.0015]
20. Calculate the probability of transmission for an electron of energy 2 eV incident upon a rectangular potential energy barrier of height 4 eV and width 10^{-9} m. [Ans: 2×10^{-6}]
21. Calculate the de Broglie wavelength of the orbital electron of hydrogen atom, given that its energy is 13.6 eV. (RTMNU, 2010)

CHAPTER

21

Atomic Nucleus and Nuclear Energy

21.1 INTRODUCTION

The nucleus of an atom occupies an astonishingly small volume at the center of the atom and is most densely packed with protons and neutrons. Almost all the entire mass of the atom is accounted by the nucleus alone. The particles in the nucleus are held together by nuclear forces which are a fundamental force like gravitational force. The heavier nuclei exhibit instability and transform into stable nuclei through radioactive disintegration. The natural radioactivity of nuclei is effectively used in the determination of ages of organic as well as inorganic substances. Radioactivity can be induced in nuclei through nuclear reactions, which assist mankind in different fields such as medicine, industry and agriculture. The neutron induced nuclear reactions occupy special place. The neutrons have the ability of causing fission of uranium and plutonium nuclei which liberates tremendous energy. The nucleus is thus a storehouse of enormous power, which can be utilized in wars as well as in production of electrical power. For the past century fossil fuels namely, coal, oil and natural gas have supplied the major portion of our energy requirements. These sources will be nearly exhausted in the near future and it is imminent that alternative sources of power are to be searched for. Nuclear power is one of the alternative sources. The energy reserve in the form of uranium is many times greater than that of fossil fuels. The generation of power through nuclear fission, however, poses its own threats and problems. Nuclear fusion power is believed to be an inexhaustible source of energy and is without any attendant hazards. Efforts are in progress for producing electrical power through nuclear fusion on commercial scale.

21.2 THE ATOMIC NUCLEUS

The British physicist Ernest Rutherford proposed the nuclear model of atom in 1911 in an attempt to explain the large angle scattering of α -particles which were directed at a thin gold foil. He postulated that all the positive charges of an atom were concentrated in a central massive core and named it nucleus. A nucleus is the centre of the atom where most of the mass of the atom and all of the positive charge are concentrated.

There is a vast difference between the size of an atom and its nucleus. The nucleus is a very small part of an atom. The radius of nucleus is of the order of 10^{-14} m while that of an atom is of the order of a few angstroms (10^{-10} m). Therefore, an atom occupies about a million times more space than does a nucleus. In between the atomic electrons and the nucleus, there is a lot of void of space.

An atomic nucleus is not a single indivisible point mass. It is composed of smaller particles. It is obvious that nucleus contains protons. However, all of the nuclear mass could not be accounted for by protons. In 1920, Rutherford suggested the existence of a neutral particle in the nucleus which he called a **neutron**. Neutron was discovered in 1932 by James Chadwick. All nuclei, with the exception of hydrogen nucleus, contain neutrons. Protons and neutrons are about 2000 times more massive than electrons. They are collectively known as **nucleons**. Since a stable atom is electrically neutral, the number of positively charged protons in the nucleus is always equal to the number of negatively charged electrons around the nucleus. The number of protons or electrons in an atom is known as the **atomic number** and is denoted by Z . Total number of protons and neutrons in an atomic nucleus is called the **mass number** and is denoted by A . A special notation is used to designate a particular nucleus. The nucleus of an atom is denoted by the chemical symbol of the atom subscripted and superscripted respectively by atomic number Z and mass number A . Thus, if the chemical of an atom is X , its nucleus is denoted by ${}_Z^AX$. For example, ${}_{\text{6}}^{12}\text{C}$ (read as carbon-six-twelve) denotes the nucleus of carbon atom, containing 6 protons and $12 - 6 = 6$ neutrons. There are 92 stable species of atom available in nature.

21.3 ISOTOPES

A particular kind of atom of any element is called a **nuclide**. A nuclide is distinguished from other nuclides by the number of protons and neutrons it contains. The atomic number Z determines the chemical nature of an element. Although for a particular element the number of electrons and protons is fixed, the number of neutrons in the nucleus may vary. It implies that the mass number A may differ though the atomic number Z remains the same. Such atoms will be chemically identical but their nuclei show marked differences in stability. Nuclei of the same element having different numbers of neutrons are called isotopes. Thus, **isotopes** are atoms of a given element that have different masses. To cite an example, hydrogen has three isotopes. ${}_{\text{1}}^1\text{H}$ is the most common isotope. It is just a proton. The other isotopes ${}_{\text{1}}^2\text{H}$ and ${}_{\text{1}}^3\text{H}$ are called **deuteron** and **triton** respectively. In atomic form they are known as deuterium and tritium respectively. The deuteron denoted by D is made of one proton and one neutron. For about every 6500 atoms of ordinary hydrogen in water, there is one atom of deuterium. Triton denoted by T is made of one proton and two neutrons. It is radioactive with a half-life period of 12.26 years.

The uranium element which plays a very important role in the production of nuclear energy exists in three isotopic forms.

Isotope	Relative abundance	Half-life
${}_{\text{92}}^{238}\text{U}$	99.28%	4.5×10^9 years
${}_{\text{92}}^{235}\text{U}$	0.714%	7.1×10^8 years
${}_{\text{92}}^{234}\text{U}$	0.006%	2.5×10^5 years

21.4 THE NUCLEAR FORCE

The nucleons are clustered together within the small volume of the nucleus. Large repulsive electrical forces operate between the positively charged protons, which tend to tear the nucleus apart. Yet the nuclei of most atoms are stable. It means that there must be some strong attractive force which more than balances the electrostatic repulsion and holds the nucleus

together. This strong attractive force is called the **nuclear force**. The nuclear interaction between nucleons is, therefore, a **strong interaction**. The nuclear force is a fundamental force like the gravitational and electrical forces, but is more complicated and not completely understood. The nuclear forces have the following properties.

Salient Features of the Nuclear Forces:

1. The nuclear force is strongly attractive and much larger in magnitude than either the electrostatic force or gravitational force. They act between any pair of nucleons – proton-proton, proton-neutron, and neutron-neutron.
2. Nuclear forces are very **short-ranged** forces. A nucleon interacts only with its nearest neighbours, over distances of the order of 10^{-15} m.
3. Nuclear forces are **charge-independent**. They do not depend on the charge of nucleons. The nuclear forces acting between two protons, between a proton and a neutron or between two neutrons have the same magnitude. That is $n-n$ force is no different from the $p-p$ and $n-p$ force.
4. Nuclear forces are **spin-dependent** and depend on the mutual orientation of the spins of the nucleons. For example, a neutron and a proton are kept together forming a nucleus of deuteron, only if their spins are parallel to each other.
5. Nuclear forces are **not central** forces. They cannot be represented as directed along the straight line connecting the centers of the interacting nucleons.
6. Nuclear forces have the property of **saturation**. It means that each nucleon in a nucleus interacts with a limited number of neighbours. Saturation manifests itself in that the binding energy per nucleon does not grow with an increase in the number of nucleons, but remains approximately constant. Further, the saturation of the nuclear forces is indicated by the volume of a nucleus being proportional to the number of nucleons forming it.
7. The magnitude of nuclear force is so high that the work required to divide a nucleus into its constituents is about 8 MeV in contrast to a few eV required to separate the extra-nuclear electrons from its atom.

21.4.1 Proton-neutron Theory

For elements of low mass numbers, the atomic number Z is nearly half the mass number A . Thus the number of protons is nearly equal to the number of neutrons. With increasing mass number, the value of Z becomes less than half of A which means that the number of neutrons exceeds that of protons. In case of ^{238}U , $Z = 92$ and $A = 238$. Therefore, $(A-Z) = 146$. Thus, uranium nucleus consists of 92 protons and 146 neutrons. For elements of low mass numbers, the neutron-to-proton ratio is close to unity. But as the number of protons in a nucleus exceeds 20, the neutron to proton ratio is greater than unity in stable nuclides. It appears that in order to maintain nuclear stability, the neutrons must exceed protons in number (Fig. 21.1). Further, the neutron excess increases with the increasing atomic number. For the heaviest stable nuclides such as $^{208}_{82}Pb$ the neutron-to-proton ratio

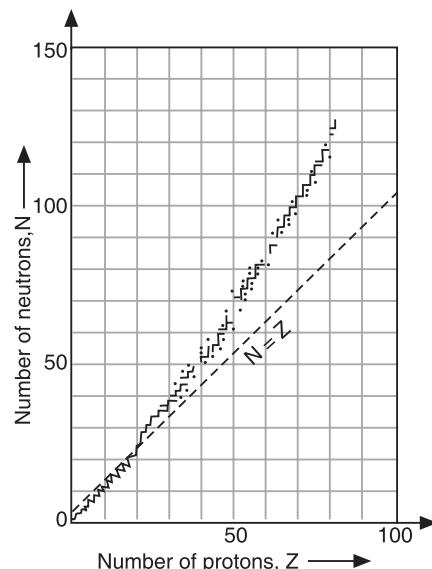
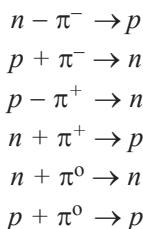


Fig. 21.1: A Plot of neutron number N versus proton number Z in nuclides. Nuclides with proton numbers greater than 20 have more neutrons than protons. The dashed line corresponding to the condition $N = Z$ is called the line of stability.

exceeds 1.5. It is inferred from the above facts that neutrons act as a nuclear “glue” to hold the nucleus together. As the number of protons increases for larger stable nucleus, the number of neutrons also increases, so that the short range nuclear forces are greater than the long range repulsive electrical forces. Neutrons provide attractive nuclear forces between both protons and other neutrons. The proton-neutron theory was put forward by Heisenberg in 1932.

Origin of the Strong Interaction

In 1935, H.Yukawa suggested a successful theoretical model to explain the origin of nuclear force. He proposed that the nuclear force is an **exchange force** acting through virtual particles. The virtual particles responsible for nuclear binding forces are called π **mesons** (or pions). The existence of mesons was discovered experimentally in 1947. Three different types of pions, namely positive, negative and neutral pions exist. The mass of a π meson is about 273 times the mass of an electron. It is postulated that nucleons consist of some sort of common core surrounded by a pulsating cloud of pions. The pions are supposed to rapidly jump back and forth between the nucleons, thereby changing their identity equally fast and at the same time keeping them bound together. When neutral pions are exchanged, the nucleons do not undergo any change. Thus, neutral π^0 mesons are associated with the forces that exist between neutrons and neutrons. When a negative pion is exchanged between a neutron and proton, the neutron emits a negative pion and turns into a proton and the proton absorbing the negative pion turns into a neutron. Thus, negative π^- mesons are associated with the forces that exist between neutrons and protons. When a positive pion is exchanged between a neutron and proton, the proton emits a positive pion and turns into a neutron and the neutron absorbing the positive pion turns into a proton. No particular proton will remain a proton in the nucleus and no particular neutron remains a neutron in the nucleus. If a proton is in the field of the negative meson, it is converted into a neutron. When a neutron is in the field of a positive meson, it is converted into a proton. Thus, the nucleus is seen to be an ever-changing complex structure. Thus, according to Yukawa’s theory, the protons and neutrons in the nucleus are held together by the continual exchange of π mesons. The interactions are summarized as follows.



21.5 STATIC PROPERTIES OF NUCLEUS

21.5.1 Nuclear Mass

Mass of an atom, a nucleus or an elementary particle is extremely small. Expressing their masses in conventional units involves cumbersome negative powers of ten. Therefore, a separate unit called **atomic mass unit** has been devised for expressing the masses of atomic particles. Nuclear mass means the mass or weight of the nucleus alone. Atomic mass unit (a.m.u.) is defined as 1/12th the mass of the $^{12}_6C$ atom.

$$1 \text{ a.m.u.} = \frac{1}{12} \times \text{Mass of one carbon atom} = \frac{1}{12} \times \frac{12 \text{ kg}}{6.02 \times 10^{26}} = 1.66 \times 10^{-27} \text{ kg.}$$

The uranium nucleus $U-238$ has a mass of 238 a.m.u. As the mass of the electron is negligibly small in comparison to that of protons and neutrons, the mass of the nucleus is taken as the mass of the corresponding atom also. Thus, the mass of the uranium atom is 238 a.m.u.

The a.m.u. is now called **unified atomic mass** and is denoted by the letter “ u ”.

21.5.2 Nuclear Radius

Nuclear radii are estimated from the measurements on the maximum scattering angles of α -particles when they approached the target nuclei. The nucleus is assumed to be spherical and the following empirical relation gives its radius.

$$R = R_o A^{1/3} \quad (21.1)$$

where R_o has an average value of 1.4×10^{-15} m.

\therefore The radius of uranium nucleus $= 1.4 \times 10^{-15}$ m $\times 238^{1/3} = 8.68 \times 10^{-15}$ m.

21.5.3 Nuclear Density

The density of nuclear matter is tremendous. All nuclei have nearly the same density. The density is calculated as follows.

$$\begin{aligned} \rho &= \frac{\text{Mass of nucleus}}{\text{Volume of the nucleus}} = \frac{M_N}{\frac{4}{3}\pi R^3} = \frac{m_p A}{\frac{4}{3}\pi R^3} \\ &= \frac{1.673 \times 10^{-27} \text{ kg}}{11.5 \times 10^{-45} \text{ m}^3} = 1.45 \times 10^{17} \text{ kg/m}^3. \end{aligned} \quad (21.2)$$

21.5.4 Nuclear Charge

Nucleus is electrically positive and the magnitude of its charge is equal to the number of protons in the nucleus. Thus, the charge on a nucleus is equal to its atomic number, Z . Thus, a uranium nucleus carries a charge of 92 units.

21.5.5 Nuclear Quantum States

The studies of α - and γ -ray spectra show that every nucleus possesses a set of quantum states. Transitions between different nuclear states cause emission of γ -rays.

21.5.6 Spin and Magnetic Moment

The hyperfine structures observed in atomic spectra indicated that the nucleus has spin motion. It is concluded on the basis of experimental evidence that the protons and neutrons are in continuous motion in discrete quantized orbits. Because of this motion, the nucleus possesses angular momentum and magnetic moment.

The magnetic moment of nuclei is given by

$$\mu_l = g \frac{h}{2\pi} \cdot \frac{e}{2M} \quad (21.3)$$

The product $\frac{h}{2\pi} \cdot \frac{e}{2M}$, often written as $\frac{e\hbar}{2M}$ is known as **nuclear magneton**.

21.6 MASS DEFECT

The nucleus is formed by bringing protons and neutrons together. The mass of the resulting nucleus is less than the sum of the masses of the constituent protons and neutrons. This mass difference is called **mass defect** and is denoted by Δm .

If Z is the number of protons in the nucleus, then the number of neutrons in the nucleus is $(A-Z)$. If m_p is the mass of the proton and m_n is that of the neutron, then the sum of the masses of the protons and neutrons

$$= Zm_p + (A - Z)m_n$$

If M is the actual mass of the nucleus, then mass defect is

$$\Delta m = Zm_p + (A - Z)m_n - M \quad (21.4)$$

Because the nucleus is more stable than the separated neutrons and protons, the nucleus is in a lower energy state. It implies that the mass, which disappears, is released in the form of energy when nucleons are bound together in a nucleus.

For example, a ${}_2^4He^4$ nucleus is formed from two neutrons and two protons.

Mass of two protons	$= 2 \times 1.007826 = 2.015652$ amu
Mass of two neutrons	$= 2 \times 1.008665 = 2.017330$ amu
Total mass	$= 4.032982$ amu

Measured mass of ${}_2^4He^4 = 4.002604$ amu

$$\therefore \text{Mass defect } \Delta m = (4.032982 - 4.002604) \text{ amu} = 0.030378 \text{ amu}$$

Atoms with atomic numbers between 30 and 63 have a greater mass defect per nuclear particle than very light elements or very heavy ones, as seen in Fig. 21.2. The most stable nuclei are in the atomic number range from 30 to 63.

21.7 BINDING ENERGY

The energy required to remove any nucleon from the nucleus is called the **binding energy** of that nucleon in the nucleus. It would be equal to the work that must be done in order to remove the nucleon from the nucleus without imparting it any kinetic energy. The total binding energy of a nucleus is defined as the energy required to break up the nucleus into its constituent protons and neutrons and place them at rest at infinite distances from one another.

It means that the binding energy is the energy equivalent of the mass defect. Thus,

$$\Delta E_b = (\Delta m)c^2$$

$$\Delta E_b = [Zm_p + (A-Z)m_n - M]c^2$$

The binding energy of the nucleus in MeV is expressed as

$$\Delta E_b = 931.4 [Zm_p + (A-Z)m_n - M] \text{ MeV/u} \quad (21.5)$$

The average binding energy per nucleon is given by

$$\bar{\Delta E}_b = \Delta E_b/A \quad (21.6)$$

On the average, the binding energy per nucleon is found to be about 8 MeV. The nuclear binding energy ΔE_b depends mainly on the total number of nucleons in the nucleus. To a first approximation, it rises linearly with an increase in mass number A . It implies that each nucleon addition to a nucleus causes the liberation of about the same amount of energy. A plot of the average binding energy per nucleon as a function of mass number A is shown in Fig. 21.3.

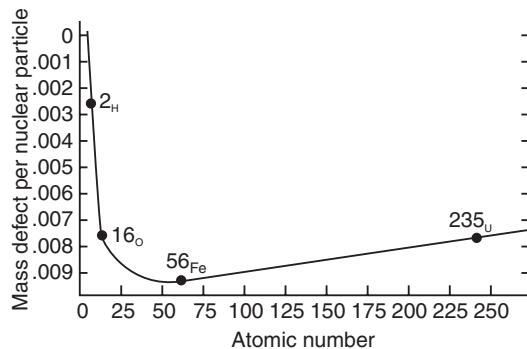


Fig. 21.2: Mass defect for different nuclides. The most stable nuclei center around Fe-56 which has the largest mass defect per nucleon.

The following important features are seen from the binding energy curve.

1. The binding energies of very light nuclei such as ${}_1H^2$ are very small. In case of light nuclei, most nucleons will be in the nuclear surface and as such they will not be in a position to use all their bonds. Consequently, $\Delta E_b/A$ is lower in these cases.
2. The binding energy is high in the middle of the periodic table, for elements whose A is in the range $28 < A < 138$, i.e. from ${}_{14}Si^{28}$ to ${}_{56}Ba^{138}$. For these nuclei, the binding energy per nucleon is nearly 8.7 MeV.
3. The binding energy per nucleon decreases in case of elements having $A > 138$. It is found to decrease to 7.6 MeV for uranium. The decrease may be attributed to the repulsive forces between protons, whose number increases in heavy nuclei.

Two most important conclusions may be drawn from the binding energy curve:

- (i) If a heavy nucleus ($A \approx 240$) is divided into two intermediate nuclei ($A \approx 120$), the resulting nuclei will be more stable than the initial heavy nucleus.
- (ii) If a single nucleus is synthesized from two light nuclei, again the resulting nucleus will be more stable than the initial light nuclei.

The first one points out the possibility of nuclear fission while the second one points out the possibility of nuclear fusion. Both the fission and fusion processes will be accompanied by the release of great amounts of energy.

Example 21.1. Calculate the binding energy of a nitrogen nucleus in MeV from the following data:

$$m_H = 1.00783 \text{ u}, \text{ and } m_n = 1.00867 \text{ u} \text{ and } m({}_{7}^{14}N) = 14.00307 \text{ u}$$

Solution. Mass defect $\Delta M = 7 \times 1.00783 + 7 \times 1.00867 - 14.00307 \text{ u} = 0.11243 \text{ u}$

$$\therefore \text{Binding energy} = 0.11243 \text{ u} \times 931.4 \text{ MeV/u} = \mathbf{104.7 \text{ MeV}}$$

Example 21.2. What is the binding energy per nucleon in ${}_{3}^{7}Li$ nuclide?

$$\text{Proton mass, } m_p = 1.00814 \text{ amu}$$

$$\text{Neutron mass, } m_n = 1.008665 \text{ amu}$$

$$\text{Mass of lithium nucleus. } M = 7.01822 \text{ amu}$$

$$\text{and } 1 \text{ amu} = 931 \text{ MeV}$$

Solution. Number of neutrons in lithium nucleus = $7 - 3 = 4$

$$\text{Total binding energy } \Delta E_b = [3 \times 1.00814 + 4 \times 1.008665 - 7.01822]931 \text{ MeV} = \mathbf{38.041 \text{ MeV}}$$

$$\text{Binding energy per nucleon } \Delta\xi = \frac{\Delta E_b}{A} = \frac{38.041}{7} \text{ MeV} = \mathbf{5.43 \text{ MeV}}$$

Example 21.3. The atomic masses of ${}_{7}^{15}N$, ${}_{8}^{16}O$ and O are 15.0001u, 15.0030u and 15.9949u respectively. How much energy is needed to remove one proton from ${}_{8}^{16}O$? How much energy

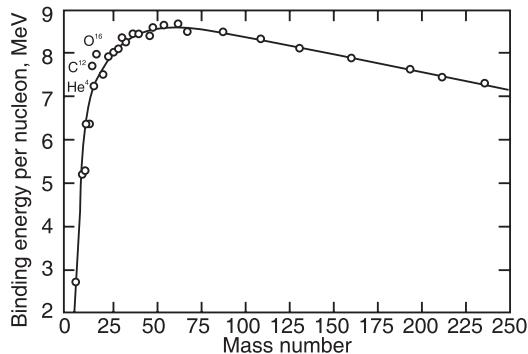
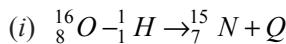


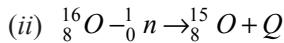
Fig. 21.3

is needed to remove one neutron from ${}_{8}^{16}O$? Why are these figures different from each other?
(Proton rest mass = 1.0072766 u, Neutron rest mass = 1.0086654 u)

Solution.



$$Q = [15.9949 - 1.0072766 - 15.0001]931 \text{ MeV} = [-0.0124766]931 \text{ MeV} = -11.62 \text{ MeV}$$



$$Q = [15.9949 - 1.0086654 - 15.0030]931 \text{ MeV} = [-0.0167654]931 \text{ MeV} = -15.62 \text{ MeV}$$

Example 21.4. Find the binding energy of an α -particle in MeV and Joule. Given : Mass of proton = 1.00758 amu, mass of neutron = 1.00897 amu and mass of helium nucleus = 1.0028 amu.

Solution. $\Delta E_b = [2 \times 1.00758 - 2 \times 1.00897 - 4.0028]931 \text{ MeV}$
 $= 28.22 \text{ MeV} = 28.22 \times 10^6 \times 1.602 \times 10^{-19} \text{ J} = 4.5 \times 10^{-12} \text{ J}$

21.8 NUCLEAR MODELS

Models are devised to account for the properties and behaviour of a system. In case of an atom, Thomson's model, Rutherford model, Bohr model, de Broglie model, vector atom model and quantum mechanical model are devised which successively incorporated refinements one over the other. Ultimately, the quantum-mechanical model succeeded in interpreting all the properties and behaviour of an atom. In the same way efforts are made to develop a model, which can successfully explain the properties of the nucleus such as stability, spin, magnetic moment, etc. Various models have been proposed for nucleus. However, each of the models can explain only a limited number of properties of the nucleus. One of the earliest models was α - particle model proposed by Gamow.

21.8.1 Gamow Model

It is also known as α -particle model. According to this model, nucleus is assumed to have sub-groups in the form of α -particles. Each sub group has two protons and two neutrons. Hydrogen and deuterium nuclei are exceptions. The model was successful in explaining the emission of α -particles by the radioactive nuclei. The model is discarded subsequently.

Out of the other different models, two models namely *shell model* and *liquid drop model* are of importance and we study them here.

21.8.2 Nuclear Shell Model

Nuclear shell model is similar to the Bohr's model for the atom. By analogy with the closed sub-shells and shells in the case of atoms, it is assumed that nucleons also form similar closed sub-shells and shells within the nucleus. The electrons in an atom are supposed to revolve in the Coulomb electrostatic field of the nucleus in allowed orbits. In a similar manner, it is assumed in the shell model that each nucleon moves inside the nucleus in a fixed orbit under the influence of a central field of force produced by the average interaction between all the remaining nucleons.

It is found that the nuclear properties vary periodically with Z and N , in a way similar to the periodic variation of atomic properties with Z . Secondly, a nucleus is stable if it has a certain definite number of either protons or neutrons. Nuclei containing the following numbers of protons and neutrons exhibit high stability.

Z	2	8	20	50	82
N	2	8	20	50	82 126

These values of Z and N are commonly called **magic numbers**. These are some features of the nucleus which indicate that the nucleus exhibits a shell structure. To interpret the existence of magic number and periodic variation of nuclear properties, one may assume that the protons and neutrons are arranged in shells in the nucleus just as extra nuclear electrons are distributed in various shells outside the nucleus. It is assumed that nucleons are arranged in shells, each shell containing a restricted number of nucleons in accordance with the Pauli principle. With increase of the number of nucleons, the first shell is filled, then the next nucleon shell etc. Since there are two classes of particles in a nucleus, there is a double shell arrangement, one for protons and another for neutrons. We may characterize the nuclear energy states by quantum number n and l , specifying the energy level and the orbital angular momentum. The shells are regarded as ‘filled’ when they have their full quota of nucleons. The magic properties are observed when the outer nucleon shell is completely filled. The first shell is filled in ${}_2He^4$ which consists of two protons and two neutrons; the next shell is filled in ${}_8O^{16}$ etc. The enhanced stability of magic nuclei resembles the chemical inertness of helium, neon, etc.

Even Z -even N nucleus has zero total nuclear spin and even-odd and odd-even nuclei have half integral nuclear spin. It means that the protons and neutrons fill their levels in pairs with each pair having antiparallel spins. When the number of both the nucleons is even, they yield zero net spin; if one of them is even and the other is odd, the net spin is obviously half-integral spin. In case of odd-odd nucleus the proton, left over after the other protons are paired off, has its half spin added to the half spin of the left over neutron resulting in integral nuclear spin.

On excitation of the nucleus, a nucleon or several nucleons move into excited levels. Transition to the ground state is accompanied by the emission of γ -quanta.

21.8.3 Liquid Drop Model

The liquid drop model was proposed in 1936 by Niels Bohr basing on the external analogy between the nucleus and a liquid drop. The nucleus of an atom is similar to a drop of a liquid in many ways.

The following are the similarities between a small drop of liquid and the nucleus.

1. Molecules at the surface of liquid are attracted to the inner molecules and the result is surface tension because of which the liquid drop takes spherical shape. Nucleus is also assumed to be spherical in shape.
2. The density of the nucleus is independent of its volume and the density of liquid is also independent of the volume.
3. In a liquid the forces are short-range forces; each molecule interacts only with its immediate neighbours. The nuclear force is also a short-range force and would primarily act between a nucleon and its nearest neighbours.
4. The molecules in the liquid drop are in random motion and frequently collide with each other. Similarly, the nucleons in a nucleus are in random motion and undergo frequent collisions.
5. Each collision between nucleons involves an exchange of energy and momentum between them. If a certain amount of energy is imparted to the nucleus, it gets excited. The energy gets redistributed rapidly between the nucleons due to their mutual collisions. Some of the energy may be transferred to a surface nucleon. If the energy of the nucleon exceeds its binding energy in the nucleus, the nucleon overcomes the nuclear forces and leaves the nucleus. The expulsion of a nuclear particle from the nucleus is similar to the evaporation of a molecule from a liquid drop.

The liquid drop model successfully explained the process of nuclear fission and α -decay.

21.9 NATURAL RADIOACTIVITY

In 1896, Henry Becquerel discovered that uranium salts emit invisible radiation.

The phenomenon of emission of radiation by elements is called radioactivity. The elements, which emit radiations, are called radioactive elements.

It was found that radioactivity is the property

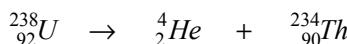
of certain nuclei and hence such nuclei are called **radioactive nuclides**. Investigations on radioactive materials (Fig. 21.4) showed that the emitted radiations consist of three types of radiations, namely, α -rays, β -rays and γ -rays. Soon it was established that α -rays are streams of 4_2He nuclei; β -rays are streams of very fast electrons and γ -rays are electromagnetic radiations which are much more penetrating than X-rays. During the emission of α -rays and β -rays, the composition of the nucleus changes, whereas in γ -rays emission the nuclear composition remains unaltered. γ -rays are emitted whenever the nucleus goes over from a higher energy state to a lower energy state.

Radioactivity may be natural or artificial. The **natural radioactivity** is the radioactivity found in nature and is exhibited by only a very small number of nuclei. Radioactivity exhibited by certain nuclei produced in the laboratory through nuclear reactions is called **induced radioactivity** or **artificial radioactivity**.

21.10 RADIOACTIVE DECAY

The spontaneous transformation occurring in the constitution of radioactive nuclides due to their radioactive property is called **radioactive decay** or **radioactive disintegration**. The transformation accompanied by the emission of α -rays is called **α -decay** and that accompanied by the emission of β -rays is called **β -decay**. The nucleus that undergoes radioactive decay is called the **parent**, the intermediate products are called **daughters** and the final stable nonradioactive nucleus is called the **end product**.

α -decay: An α -particle is the same as a helium nucleus and consists of two protons and two neutrons. It carries a positive charge of two units. In α -decay, the mass number of the nucleus decreases by four units and the charge on the nucleus by two units. Therefore, the original element is transformed into an element two steps down in the periodic table. For example, α -decay of uranium produces thorium. Thus,



β -decay: A beta particle is simply an electron. β -decay of a nucleus leads to a decrease

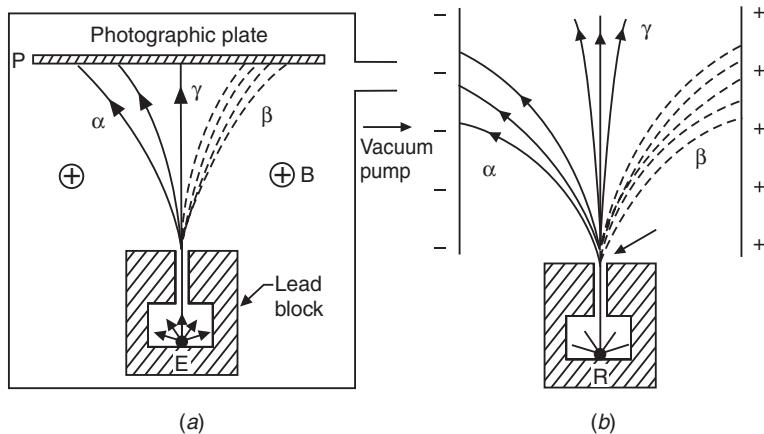
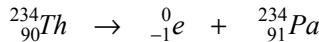


Fig. 21.4: Passing the radiation from a radioactive source through (a) a magnetic field or (b) an electric field shows that there are three components

of one negative charge and no change in mass. It produces an element one step higher in the periodic table. For example, Th-234 emits β -particles and transforms into protactinium (Pa).



γ -decay: A γ -particle is a quantum of energy. γ -radiation may or may not be emitted simultaneously with α - or β -rays. The nucleus has energy levels similar to the atomic energy levels. A γ -ray is emitted when an excited nucleus jumps to a lower energy level. For example,



Since γ -rays do not involve mass or charge, the emission of a gamma photon does not cause a transformation of species of nucleus.

21.11 RADIOACTIVE SERIES

All the naturally occurring radioactive elements lie in the range of atomic numbers from $Z = 81$ to $Z = 92$. The nuclei of these elements are unstable and disintegrate by emitting either α or β particles. Sometimes γ -rays accompany the emission of these particles. A radioactive nucleus often decays to another radioactive nucleus. This daughter nucleus decays to a third nucleus, which is also radioactive. The chain of radioactive decays continues until finally a stable nucleus forms. It is known that all natural disintegration processes end with the formation of stable lead atoms. The chain of successive radioactive decays is said to form a **radioactive series**.

There are three series of naturally occurring radioactive elements and a fourth one produced artificially. The three series of naturally occurring radioactive elements are

1. Uranium series
2. Actinium series
3. Thorium series

The series of artificially produced radio-active elements is Neptunium series.

Each series is named after the name of the parent nucleus from which the series starts. These series follow a decay pattern similar to that of the uranium series. In each radioactive series, each nuclide transforms into the

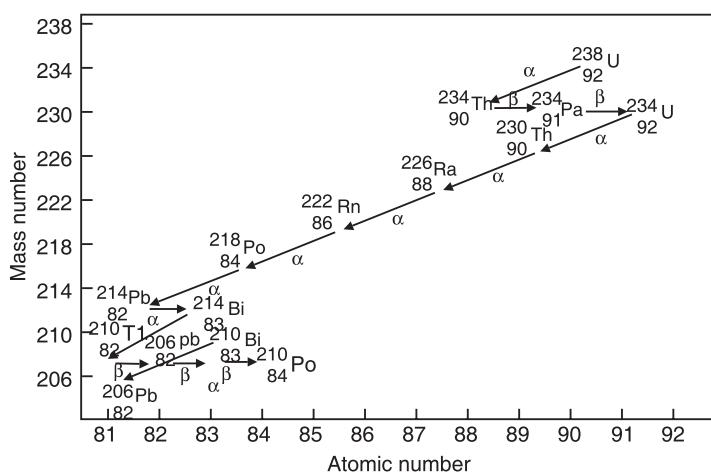


Fig. 21.5

next through a chain of α - and β -decays and ultimately a stable nucleus forms at the end. Thus, the uranium series terminates in the ${}^{206}_{82}Pb$ nucleus, the actinium series in ${}^{207}_{82}Pb$ nucleus and the thorium series in ${}^{208}_{82}Pb$ nucleus. The uranium radioactive decay series is shown in Fig. 21.5. The neptunium series terminates in the ${}^{209}_{83}Bi$ nucleus.

21.12 LAW OF RADIOACTIVE DECAY

The number of parent nuclei in a radioactive material decreases with time because of radioactive disintegration. The disintegration is a statistical and random process. Which

nucleus in the material disintegrates first is only a matter of chance. Assuming that each nucleus has the same probability of decaying in one second, we can determine how many nuclei in a sample will decay over a given period of time.

The number of nuclei that disintegrate per second is called *rate of radioactive decay*. The rate of decay is proportional to the number of nuclei that have not yet disintegrated at any instant.

Let there be N untransformed nuclei present in a radioactive sample at time t and let dN be the number of decays in a short duration between $t+dt$. Then, dN , the number of nuclei disintegrating in the interval dt will be proportional to N and dt . That is,

$$dN \propto N dt$$

or

$$dN = -\lambda N dt \quad (21.7)$$

where λ is constant of proportionality and is known as **decay constant** or **disintegration constant**. It is characteristic of the nuclear species. The minus sign in the above equation (21.7) indicates that N decreases with time. The above relation implies that during a longer interval of time, a greater number of nuclei disintegrate and the number of nuclei undergoing decay per unit time will be greater with a larger sample.

The fraction of nuclei decaying in a time dt is given by

$$\frac{dN}{N} = -\lambda dt$$

Assuming that there were N_o nuclei at time $t = 0$, and integrating the above equation, we can find out the nature of the decay.

$$\int_{N_o}^N \frac{dN}{N} = -\lambda \int_0^t dt$$

$$\ln[N/N_o] = -\lambda t$$

Taking exponential on both the sides of the above equation, we obtain

$$\frac{N}{N_o} = e^{-\lambda t}$$

or

$$N = N_o e^{-\lambda t} \quad (21.8)$$

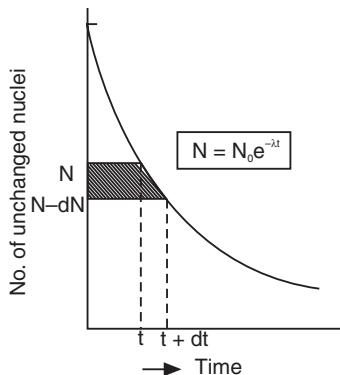


Fig. 21.6

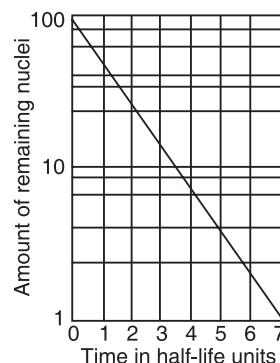


Fig. 21.7

The above equation (21.8) is known as the **law of radioactive decay**. This relation shows that the number of nuclei of a given species decreases exponentially with time provided no new nuclei are added. The decay occurs rapidly initially and then becomes slower and slower as depicted in Fig. 21.6.

A plot of $\ln(N/N_o)$ versus time is plotted in Fig. 21.7. It is a straight line and the slope of the line gives the value of λ . Eqn.(21.7) may be rearranged as

$$\lambda = -\frac{(dN / N)}{dt} \quad (21.9)$$

λ may be defined as the fractional decrease in the number of nuclei decaying per unit time.

21.13 ACTIVITY

In the decay process, we are generally interested in the number of disintegrations per second. This is called the **activity**, A of the sample.

Thus,

$$A = \left| \frac{dN}{dt} \right| = \lambda N_o e^{-\lambda t} = \lambda N \quad (21.10)$$

It is obvious from eqn.(21.10) that the activity at $t = 0$ is given by $A_o = \lambda N_o$. Therefore,

$$A = A_o e^{-\lambda t} \quad (21.11)$$

The plot of activity as a function of time is shown in Fig. 21.8.

21.14 HALF-LIFE

The rate of decay or activity of a radioactive element is measured in terms of a characteristic time called the half-life. The **half-life** of an element is defined as the time taken by half of the original quantity to undergo decay. If N_o are the number of nuclei present at $t = 0$, then the time in which $N_o/2$ nuclei decay will be the half-life. In other words, half-life is the time after which half of the original number of nuclei remains untransformed. Half-life can also be defined as the time required for the initial activity A_o to decrease to $A_o/2$. Half-life is denoted by $T_{1/2}$.

$$N = \frac{N_o}{2} \quad \text{at } t = T_{1/2}$$

Using these values into the exponential decay law (21.8), we get

$$\frac{N_o}{2} = N_o e^{-\lambda T_{1/2}}$$

or

$$e^{\lambda T_{1/2}} = 2$$

∴

$$\lambda T_{1/2} = \ln 2 = 0.693$$

∴

$$T_{1/2} = \frac{0.693}{\lambda} \quad (21.12)$$

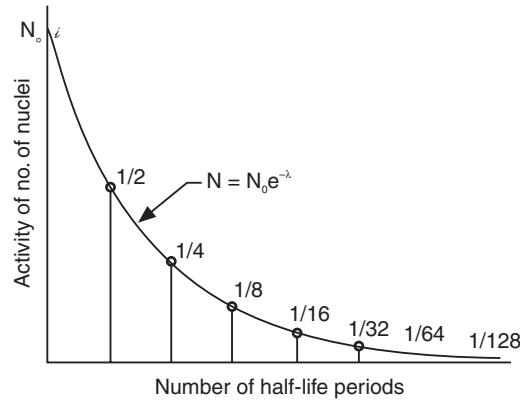


Fig. 21.8

Eq.(21.12) suggests that the longer the half-life of an element, the slower it decays. The half-lives of natural radioactive nuclei vary between wide limits. For uranium, it is 4500 million years, for radium 1620 years, and for radon it is 3.8 days only.

21.15 AVERAGE LIFE TIME

The atoms of radioactive substances are continuously disintegrating. Therefore, the life of each nucleus is different. The nuclei, which disintegrate earlier, have shorter life whereas those that disintegrate at the end have a longer life. It is difficult to specify why some atoms have shorter existence while others remain untransformed for a long time. Therefore, it becomes necessary to specify the **mean lifetime** or **average lifetime** of a radioactive species. The average lifetime, τ of a nuclide of a radioactive substance is defined as the ratio of the sum of lifetimes of all nuclei present in the substance to the total number of nuclei. That is,

$$\tau = \frac{\text{Sum of the life times of all nuclei}}{\text{Total number of nuclei}}$$

Out of the original N_o nuclei, let dN_1 nuclei live for time t_1 , dN_2 nuclei live for time t_2 , and so on. Then,

$$\tau = \frac{t_1 dN_1 + t_2 dN_2 + \dots}{dN_1 + dN_2 + \dots}$$

The above equation may be written in the integral form as

$$\tau = \frac{\int_0^{N_o} t \cdot dN}{\int_{N_o}^0 dN} = \frac{\int_0^{N_o} t \cdot dN}{\int_0^{N_o} dN}$$

As, $N = N_o e^{-\lambda t}$, $dN = -\lambda N_o e^{-\lambda t} dt$.

$$\begin{aligned} \therefore \tau &= \frac{\int_0^{\infty} (-\lambda) t N_o e^{-\lambda t} dt}{N_o} \quad : \quad \int_0^{N_o} dN = N_o \\ &= \lambda \int_0^{\infty} t e^{-\lambda t} dt \\ &= \lambda \left[\frac{e^{-\lambda t}}{-\lambda} t - \frac{e^{-\lambda t}}{-\lambda^2} \right]_0^{\infty} = -\frac{1}{\lambda} \left[(\lambda t + 1) e^{-\lambda t} \right]_0^{\infty} = \frac{1}{\lambda} \end{aligned} \tag{21.13}$$

Thus, the mean lifetime of a radioactive species is equal to the reciprocal of the decay constant.

Using equ. (21.13) into equ. (21.12), we obtain

$$T_{1/2} = 0.693 \tau \tag{21.14}$$

21.16 UNITS OF ACTIVITY

The **activity** of a radioactive source is defined as the number of disintegrations that occur per second. The traditional unit of activity is **curie** (Ci). It is defined as

$$1 \text{ Ci} = 3.7 \times 10^{10} \text{ disintegrations per second.}$$

This definition is based on the activity of 1 gm of radium. Curie is rather a large unit, so millicuries ($m \text{ Ci}$) and microcuries ($\mu \text{ Ci}$) are commonly used.

The *SI* unit of radioactivity is the Becquerel (Bq). It is defined as

$$1 \text{ Bq} = 1 \text{ decay per second}$$

Example 21.5: A certain radioactive substance has a disintegration constant $\lambda = 1.44 \times 10^{-3}$ per hour. In what time will 75% of the initial number of atoms disintegrate?

Solution.

$$N = N_o - \frac{3}{4} N_o = \frac{1}{4} N_o$$

But

$$N = N_o e^{-\lambda t} \quad \therefore \quad \frac{1}{4} = e^{-\lambda t} \quad \text{or} \quad e^{\lambda t} = 4$$

$$\therefore t = \frac{\log_e 4}{\lambda} = \frac{2.3026 \times 0.6021}{1.44 \times 10^{-3}} \text{ hours} = 962.9 \text{ hrs.}$$

Example 21.6: Calculate the activity of 1 mg sample of ^{90}Sr whose half life is 28 years.

Solution:

$$\begin{aligned} \lambda &= \frac{0.693}{T_{1/2}} = \frac{0.693}{28 \text{ yrs}} = \frac{0.693}{28 \times 365 \times 24 \times 3600 \text{ s}} \\ &= 7.85 \times 10^{-10} / \text{s.} \end{aligned}$$

Number of nuclei in one mg sample,

$$N = \frac{m N_A}{M} = \frac{10^{-3} \times 6.02 \times 10^{26}}{90} = 6.69 \times 10^{21}$$

$$\begin{aligned} \text{Activity of the sample} \quad A &= \lambda N = 7.85 \times 10^{-10} / \text{s} \times 6.69 \times 10^{21} \\ &= 5.25 \times 10^{12} \text{ disintegrations/s.} \end{aligned}$$

Example 21.7: One gm of a radioactive material having a half-life of 2 years is kept in store for a duration of 4 years. Calculate how much of the material remains unchanged.

Solution:

$$T_{1/2} = \frac{0.6931}{\lambda} = 2 \text{ yrs}$$

\therefore

$$\lambda = \frac{0.6931}{2 \text{ yrs.}}$$

As

$$N = N_0 e^{-\lambda t}, t = \frac{1}{\lambda} \ln \frac{N_0}{N}$$

\therefore

$$4 \text{ yrs.} = \left(\frac{2 \text{ yrs.}}{0.6931} \right) \ln \frac{N_0}{N}$$

\therefore

$$\ln \frac{N_0}{N} = \frac{4 \times 0.6931}{2} = 1.386$$

\therefore

$$\frac{N_0}{N} = 4$$

or

$$N = \frac{N_0}{4} = \frac{1\text{ gm}}{4} = 0.25 \text{ gm}$$

The material that remains unchanged after 4 years is 0.25 gm.

Example 21.8. 5 gm of radium is reduced by 10.5 mg in 5 years. Calculate the half-life of radium.

Solution. The initial quantity of radium $N_0 = 5 \text{ gm}$.

Quantity of radium present $N = 5 - 10.5 \times 10^3 \text{ gm} = 4.9895 \text{ gm}$

$$\frac{N}{N_0} = e^{-\lambda t}$$

$$\therefore \frac{4.9895 \text{ gm}}{5 \text{ gm}} = e^{-\lambda t}$$

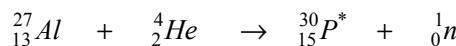
$$\therefore \lambda t = \ln \left[\frac{5}{4.9895} \right]$$

$$\text{As } t = 5 \text{ years, } \lambda = \frac{\ln \left(\frac{5}{4.9895} \right)}{5 \text{ years}} = 4.2 \times 10^{-4} \text{ dis./year}$$

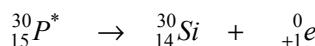
$$\frac{T_1}{2} = \frac{0.693}{\lambda} = \frac{0.693}{4.2 \times 10^{-4}} \text{ yrs.} = 1650 \text{ years}$$

21.17 INDUCED RADIOACTIVITY

Joliot-Curie and Frederic discovered in 1934 the phenomenon of **induced** or **artificial radioactivity**. They discovered that aluminium, boron and magnesium become radioactive when irradiated with α -particles. The irradiation of aluminium induces the nuclear reaction



The phosphorous isotope $^{30}_{15}P^*$ is radioactive. It turns into a stable silicon isotope after emitting a positron.



The discovery of induced radioactivity is important on two counts. It was for the first time that radioactive materials were synthesized and secondly, it proved that not only heavy nuclei have radioactive isotopes but light nuclei also have radioactive isotopes. Subsequent works demonstrated that radioactive isotopes of all elements could be synthesized. Radioactive isotopes are produced by irradiating the nuclei with α -particles, protons, deuterons or high energy γ -rays.

Fermi studied radioactivity induced by neutrons. At present the most widely used method for producing radioactive isotopes is by neutron irradiation. All nuclei except 4_2He absorb neutrons and transform into β -active isotopes.

21.18 APPLICATIONS OF RADIOACTIVITY

Radioactivity is widely used in many areas of science and technology.

- (i) Wide use is made of high penetration power of γ -rays. Since the absorption of radiation increases with the distance it travels through an object, the variations in intensity of radiation can be used to measure the thickness of an object or to detect internal defects. Metal castings in industry are tested using this method. The radioisotope ^{60}Co which emits γ -rays is housed in an aluminium thimble and is placed inside the casting under test. A photographic film is positioned outside the object. The gamma rays penetrate the metal part and make structural flaws such as cavities in the metal observable on the photographic film. The structural members of big boilers in aircrafts are tested by this method. The thickness measurement and monitoring is carried out in metal sheet and foil manufacturing with the help of this method.
- (ii) Radioactive isotopes behave chemically just like nonradioactive isotopes of the same element. Therefore, artificial radioactive isotopes are widely used as **tracers**.
- (iii) In medicine radioactive tracers are used for diagnostic purpose and treatment of sickness. For example, the extent of affliction of thyroid gland can be determined using radioactive iodine. The minimum daily requirement of iodine is 150 micrograms. If a patient is given radioactive iodine-131, the thyroid gland does not know the difference. By measuring the counting rate near the gland, we can know how the radioactive iodine is distributed in the gland. The count rate is affected if there is some abnormality in the functioning of the gland.
- (iv) The ionizing effect of nuclear radiation is used in medicine to destroy malignant tumours. The heating effect caused by nuclear radiation is used to generate electricity which in turn powers a cardiac pacemaker. The electricity generated by plutonium is sent as a pulse directly to the ventricles of the heart at a predetermined rate.
- (v) Tracer atoms are used to study photosynthesis in plants. Tracer atoms help determine the effect of fertilizers on plants. This type of research leads to better yields per acre, and more food at less expense.
- (vi) Nuclear radiation is used to preserve foods. Foods are pasteurized by irradiation to retard the growth of organisms such as bacteria, molds and yeasts. The irradiation prolongs shelf life under refrigeration.
- (vii) Similarly the location of an underground pipe or a clog can be determined by using a liquid radioactive tracer.
- (viii) The wearing out of piston rings etc is studied by introducing radioactive iron isotopes into the parts. During operation of an engine, the radioactive atoms appear in the lubricating oil. By periodically determining the activity of the oil, the wear of a particular component can be determined.
- (ix) Radioactive ^{131}I is also used for therapy, in the treatment of thyroid cancer. A patient is given a solution of ^{131}I as potassium iodide. The iodine makes its way to the thyroid gland and the beta rays emitted by iodine-131 destroy the cancerous thyroid cells.

21.18.1 Radioactive Dating

One of the most important applications of radioactivity lies in the determination of the age of rocks, planets and solar system, and the age of fossils and objects discovered in excavations. The exponential character of disintegration of radioactive nuclei provides time scale for these determinations. From equations (21.8) and (21.12), the time interval between the instants when the number of radioactive nuclei is N_o and N may be given by

$$t = \frac{1}{\lambda} \ln \frac{N_o}{N} = 1.44 T_{1/2} \ln \frac{N_o}{N} \quad (21.15)$$

If N is the number of unchanged nuclei at the present time, the above equation gives the age of the given species of radioactive nuclei.

Each application requires a different time scale in practice. In geology a sufficiently slow radioactive time scale is required and uranium is the most suitable scale for this purpose. The age of uranium ore can be determined from the ratio of uranium to the final product of lead contained in it. Age determinations for various rocks taken from different parts of the world indicate their ages to be in the neighbourhood of 3 to 4 billion years. It suggests that the crust of the earth was formed about 4 billion years ago. The age of some meteorites is found to be 4.5 billion years. Therefore, the age of the planets in the solar system may be thought to be 4.5 billion years. In archaeology, a time scale with a half-life of only a few centuries is required. Uranium time is very long and also many of the excavated objects do not contain uranium, making the uranium dating unsuitable. Carbon-14 dating is employed in these studies. The interaction of cosmic rays with the nuclei of nitrogen in the atmosphere turns them into radioactive carbon isotope ^{14}C . The half-life of carbon-14 is 5730 years. The radioactive carbon produces radioactive carbon dioxide which mixes with the ordinary carbon dioxide in the atmosphere and, in turn, is assimilated by all living matter through natural food chains. Upon the death of the organism, the intake of food ceases and the natural level of radioactive carbon present within the structure begins to decrease. Therefore, radioactive carbon-14 can be used to determine the date of ancient artifacts derived from the living matter.

Example 21.9: In a uranium mineral, lead -206 is found predominantly. The mineral contains 0.093 of lead in 1 gm of uranium. Calculate the age of the mineral if half-life of uranium is 4.3×10^9 years.

Solution: As $t=0$, let N_0 be the uranium nuclei. Number of uranium nuclei at t be N .

$$N = N_0 e^{-\lambda t}$$

Number of uranium nuclei disintegrated in time $t = N' =$ Number of lead nuclei at time t .

$$\begin{aligned} N' &= N_0 - N_0 e^{-\lambda t} = N_0 (1 - e^{-\lambda t}) \\ &= \frac{\text{Number of lead nuclei at } t}{\text{Number of uranium nuclei at } t} = \frac{N_0 (1 - e^{-\lambda t})}{N_0 e^{-\lambda t}} = \frac{N'}{N} \\ \therefore e^{\lambda t} - 1 &= \frac{N'}{N} \quad \text{or } e^{\lambda t} = 1 + \frac{N'}{N} \\ \lambda t &= \ln \left(1 + \frac{N'}{N} \right) \\ t &= \frac{1}{\lambda} \ln \left(1 + \frac{N'}{N} \right) \end{aligned}$$

$$\text{Number of nuclei in 1 gm of uranium, } N = \frac{N_A}{238}$$

$$\text{Number of nuclei in 0.093 gm of lead, } N' = \frac{0.093 N_A}{206}$$

$$\frac{N'}{N} = \frac{0.093 \times 238}{206} = \frac{22.134}{206}$$

$$1 + \frac{N'}{N} = 1 + \frac{22.134}{206} = \frac{228.134}{206} = 1.107$$

$$t = \frac{1}{\lambda} \ln(1.107) = \frac{4.5 \times 10^9 \text{ yrs.}}{0.693} \ln(1.107)$$

$$= 6.6 \times 10^8 \text{ years} = \mathbf{660 \text{ Million Years}}$$

Example 21.10: A wooden piece of antiquity weighs 50 gm and shows ^{14}C activity of 320 disintegrations per minute. Estimate its age, assuming that the living tree, of which the wooden piece was a part, shows ^{14}C activity of 12 disintegrations per minute per gm. The half-life of ^{14}C is 5730 years.

Solution: Let there be N_0 radioactive C-14 atoms in the tree just before it died. Its activity is

$$A_0 = \lambda N_0$$

After its death, the activity decreases exponentially

Thus

$$A = -\frac{dN}{dt} = \lambda N = \lambda N_0 e^{-\lambda t}$$

$$\therefore \frac{A}{A_0} = e^{-\lambda t}$$

$$A_0 = 12 \text{ disintegrations/min/gm}$$

$$A = \frac{320}{50} \text{ disintegrations/min/gm}$$

$$\lambda = \frac{0.693}{5730} (\text{year})^{-1}$$

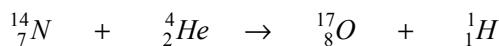
$$\ln\left(\frac{A}{A_0}\right) = -\lambda t$$

$$t = \frac{1}{\lambda} \ln\left(\frac{A_0}{A}\right) = \left(\frac{5730 \text{ yrs.}}{0.693}\right) \ln\left(\frac{50 \times 12}{320}\right) = \mathbf{5197.5 \text{ yrs.}}$$

21.19 NUCLEAR REACTIONS

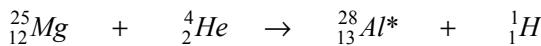
Natural radioactivity is a nuclear reaction in which a radioactive nucleus of one element spontaneously changes into the nucleus of another element. It is also possible to induce a nuclear reaction artificially, by bombarding a nucleus with high-energy particles. The reaction, in which an external bombarding particle successfully changes the identity of a target nucleus, is known as an **artificial nuclear reaction** or simply **nuclear reaction**.

The first artificial nuclear reaction was produced by Rutherford in 1919 by bombarding nitrogen gas with α -particles emitted by a natural bismuth-214 source.



The bombardment resulted in conversion of nitrogen nucleus into oxygen nucleus. Such transformation of one nucleus into another nucleus is called **transmutation**. The important implication of this reaction is that one element can be changed into another by use of bombardment process. Subsequent to Rutherford's experiment, many isotopes were subjected to beams of high energy particles and numerous reactions were found. Not all nuclear

reactions produce stable isotopes. If $^{25}_{12}Mg$ is bombarded with an alpha source, a radioactive isotope of aluminium $^{28}_{13}Al$ is produced which does not exist in nature.

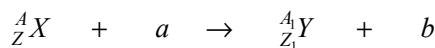


The asterisk (*) is used to denote a **radioactive isotope**. Artificial radioactive isotopes have characteristic half-life as do the naturally occurring ones.

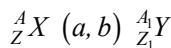
In a typical nuclear reaction experiment, a target material is bombarded by high-energy particles such as protons, neutrons, deuterons, electrons or α -particles. The interaction between the target material and the incident particle depends on the nature and the energy of the particles and the target material.

Most of the alpha particles emitted from natural radioactive decay do not have enough energy to penetrate a heavy positively charged nucleus to induce a nuclear reaction. If the artificial nuclear reactions are to be studied for heavier elements, the kinetic energy of projectile particles must be increased.

A nuclear reaction is represented in a form similar to that of a chemical reaction. Thus, the reaction equation is shown as



where a is the projectile particle that bombards the target nucleus X resulting in the product nucleus Y and an outgoing particle b . The reaction is represented in an abbreviated form as



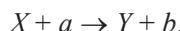
Thus, the first artificial nuclear reaction is written as $^{14}_7N (\alpha, p) ^{17}_8O$. The type of (a, b) nuclear reaction is determined by the nature of the projectile and the emitted particles.

In any nuclear reaction, mass and charge are conserved. Conservation of mass requires that the sums of the superscript mass numbers on each side of the equation are equal. Conservation of charge requires that the sums of the subscript protons on each side of the equation are equal.

21.20 Q-VALUE

Nuclear reactions resemble ordinary chemical reactions and are accompanied by energy changes. The energy liberated or absorbed during a nuclear reaction is called **nuclear reaction energy**. This energy is denoted by Q in the reaction equation and is called the *energy balance* of the reaction or more commonly, its **Q-value**. The Q-value of a reaction can be positive or negative depending on the nature of the reaction. According to Einstein's mass-energy equivalence principle, Q-value must be balanced by the changes in mass associated with the nuclear reaction.

Let us consider a general nuclear reaction



Let M_0 be the mass of the stationary target nucleus X ,

M_1, E_1 be the mass and energy of the projectile a ,

M_2, E_2 be the mass and energy of the product nucleus Y , and

M_3, E_3 be the mass and energy of the emitted particle b .

The nuclear reaction may be written as

$$M_0 + (M_1 + E_1) = (M_2 + E_2) + (M_3 + E_3)$$

or
$$(M_0 + M_1) - (M_2 + M_3) = (E_2 + E_3) - E_1$$

As Q-value is the energy balance, it may be expressed as

$$Q = (E_2 + E_3) - E_1 \quad (21.16 \text{ a})$$

or

$$Q = (M_0 + M_1) - (M_2 + M_3) \quad (21.16 \text{ b})$$

The above equations show that the Q-value of a nuclear reaction may be determined either from the known kinetic energies of the particles involved or from the known masses of the reactants and the product nuclei. However, we define *Q-value as the difference in masses of the reactants and the products.*

(i) If $Q > 0$, the reaction is said to be **exothermic** or **exoergic**. That is, energy is released in the reaction. In this case, $(M_0 + M_1) > (M_2 + M_3)$. Hence, the total mass of the products is less than that of the reactants. The difference in masses (mass defect) is converted into energy.

(ii) If $Q < 0$, the reaction is said to be **endothermic** or **endoergic**. In this case $(M_0 + M_1) < (M_2 + M_3)$. That is, the total mass of the reactants is less than that of the products. It means that there is a gain of mass in the reaction, which could happen when there is absorption of energy. Therefore, energy is to be supplied in the form of kinetic energy of the projectile *a*. The kinetic energy of the projectile particle must have some minimum value, below which the reaction cannot occur. The minimum energy necessary for an endothermic reaction to occur is called the **threshold energy**.

1. Example of Exothermic Nuclear Reaction:

${}^7_3Li(p, \alpha){}^4_2He$ is an example of exothermic reaction. Let us calculate the Q-value of this reaction.

Reactants total mass	Products total mass
$M_0 = 7.01822$ amu	$M_2 = 4.00387$ amu
$M_1 = 1.00814$ amu	$M_3 = 4.00387$ amu
$M_0 + M_1 = 8.02636$ amu	$M_2 + M_3 = 8.00744$ amu

$$\Delta M = (M_0 + M_1) - (M_2 + M_3) = 0.01862 \text{ amu}$$

$$Q = c^2 \Delta M = 931.4 (\Delta M) \text{ MeV} = 931.4 (0.01862) \text{ MeV}$$

or

$$Q = 17.34 \text{ MeV.}$$

2. Example of Endothermic Nuclear Reaction:

${}^{14}_7N(\alpha, p){}^{17}_8O$ is an example of endothermic reaction. Let us calculate the Q-value of this reaction.

Reactants total mass	Products total mass
$M_0 = 14.00753$ amu	$M_2 = 17.00450$ amu
$M_1 = 4.00387$ amu	$M_3 = 1.00814$ amu
$M_0 + M_1 = 18.01140$ amu	$M_2 + M_3 = 18.01264$ amu

$$\Delta M = (M_0 + M_1) - (M_2 + M_3) = -0.00124 \text{ amu}$$

$$Q = c^2 \Delta M = -931.4 (\Delta M) \text{ MeV} = -931.4 (0.00124) \text{ MeV}$$

or

$$Q = -1.15 \text{ MeV.}$$

Example 21.11. A nuclear reaction is given by ${}^{10}_5B + {}^4_2He \rightarrow {}^{13}_6C + {}^1_1H$

Given that: ${}^{10}_5B = 10.016125$ amu, ${}^4_2He = 4.003874$ amu
 ${}^{13}_6C = 13.007440$ amu and ${}^1_1H = 1.008146$ amu

Compute the energy released.

Solution. $Q = [10.016125 + 4.003874 - (13.007440 + 1.008146)] 931.4 \text{ MeV}$
 $= 4.11 \text{ MeV}$

21.21 NUCLEAR REACTION CROSS-SECTION

All particles that are incident on a target material do not produce nuclear reactions; only a small fraction of them interact with the nuclei of the target material. **Nuclear cross-section** is a convenient way to express the probability that a bombarding particle will interact in a certain way with a target particle. It is supposed that each target particle presents a certain area known as its *cross-section* to the incident particles. Any incident particle that is directed at this area interacts with the target particle. Hence, the greater the cross-section, the greater is the likelihood of interaction. The nuclear cross-section is usually denoted by σ .

The nuclear reaction cross-section may be defined as

- (i) the probability that an event may occur when a single nucleus is exposed to a beam of particles of total flux of one particle per unit area, or
- (ii) the probability that an event may occur when a single particle is shot perpendicularly at a target consisting of one nucleus per unit area.

21.21.1 Calculation of Microscopic Cross-section

Let us suppose that we have a slab of some material whose area is A and thickness is dx . We assume here that the target is thin and there is no overlapping of nuclei.

The volume of the slab = $A dx$.

Let the material contain n atoms per unit volume,
then the total number of nuclei in the slab = $n A dx$.

If each nucleus has a cross-section σ for some particular interaction,
then the aggregate cross-section for all the nuclei in the slab = $\sigma n A dx$.

If N is the number of incident particles in a beam, and dN is the number of particles that interact out of them with the nuclei in the slab, then

$$\frac{\text{Number of interacting particles}}{\text{Number of incident particles}} = \frac{\text{Aggregate cross-section}}{\text{Target area}}$$

$$\frac{dN}{N} = \frac{\sigma n A dx}{A} = \sigma n dx$$

or

$$\sigma = \frac{dN / N}{n dx} \quad (21.17)$$

The above equation gives the cross-section per nucleus. Here, σ is known as the *microscopic cross-section*.

21.21.2 Calculation of Macroscopic Cross-section

Now we consider the case where a beam of particles is incident on a slab of finite thickness x (Fig. 21.9). If each incident particle can interact with nuclei only once, then dN represents the number

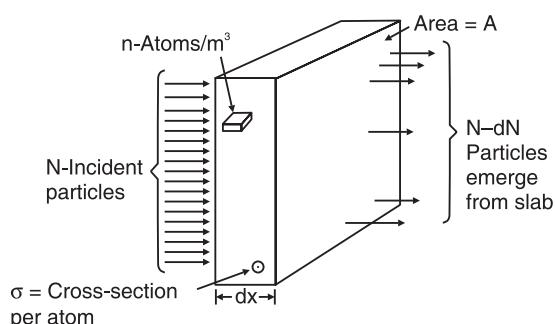


Fig. 21.9

of particles removed from the beam in passing through the first dx thickness of the slab. Hence,

$$\frac{-dN}{N} = \sigma n dx$$

The negative sign in the above equation indicates that the number of particles decrease as the beam passes through the slab. Let N_0 be the initial number of incident particles. Then

$$\int_{N_0}^N \frac{dN}{N} = -\sigma n \int_0^x dx$$

$$\ln \frac{N}{N_0} = -\sigma nx$$

$$\therefore N = N_0 e^{-\sigma nx}$$

Thus, N the number of surviving particles decreases exponentially with increasing thickness of the slab.

The product of microscopic cross-section and the number of nuclei per unit volume, i.e. σn is known as the macroscopic cross-section. It is denoted by Σ .

σ has the dimensions of area and Σ has the dimension of L^{-1} . It is important to note that the effective cross-section of a nucleus is not equal to its geometric cross-section. It is only a convenient artifice to describe the probability of occurrence of nuclear reactions. In fact, the cross-sections of target nuclei depend upon the type of interaction, type of incident particle and its energy.

The unit used for nuclear cross-section is barn.

$$1 \text{ barn} = 10^{-28} \text{ m}^2$$

21.21.3 Differential Cross-section

In many nuclear reactions, the particles are not produced in an isotropic manner. In such a case, if dN is the number of product nuclei emitted per unit time in a small solid angle $d\Omega$ at some angle θ with the direction of the incident beam (Fig. 21.10), then from equ.(21.16) we can write

$$\frac{dN}{N} = n dx d\sigma$$

where $d\sigma$ is the small cross-section corresponding to the solid angle $d\Omega$. the above equation may be rewritten as

$$\frac{1}{N} \frac{dN}{d\Omega} = n dx \frac{d\sigma}{d\Omega}$$

$$\text{or } \frac{d\sigma}{d\Omega} = \frac{1}{N} \frac{dN/d\Omega}{n dx} \quad (21.18)$$

$\frac{d\sigma}{d\Omega}$ is called the **differential cross-section** of the nuclear reaction and represents the probability that the out-going product particles are confined to an element of solid angle $d\Omega$.

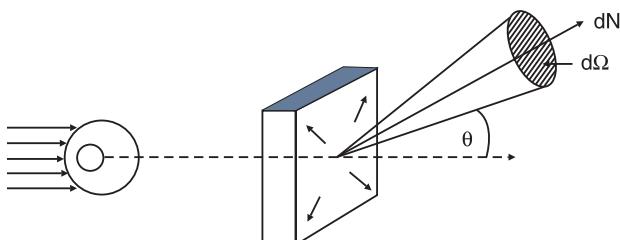


Fig. 21.10

Example 21.12. A 0.01 mm thick 7_3Li target is bombarded with 10^{13} protons/s. As a result 10^8 neutrons/s are produced. What would be the cross-section for this reaction? The density of lithium is 500 kg/m^3 .

Solution.

$$\sigma = \frac{N}{N_0 N_t}$$

$$\text{The number of target nuclei per unit volume} = \frac{\rho N_A}{M} = \frac{(500 \text{ kg/m}^3)(6.02 \times 10^{26})}{7 \text{ kg}} = 4.3 \times 10^{28}/\text{m}^3.$$

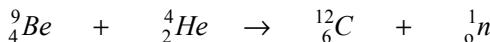
$$\begin{aligned}\text{The number of target nuclei/unit area } N_t &= \left(\frac{\rho N_A}{M} \right) t = (4.3 \times 10^{28}/\text{m}^3) (0.01 \times 10^{-3} \text{ m}) \\ &= 4.3 \times 10^{23}/\text{m}^2.\end{aligned}$$

$$\begin{aligned}\text{Number of nuclei undergoing interaction per second} &= \text{Number of neutrons produced / s} \\ &= 10^8.\end{aligned}$$

$$\therefore \sigma = \frac{10^8}{10^{13} \times 4.3 \times 10^{23}/\text{m}^2} = 2.3 \times 10^{-29} \text{ m}^2.$$

21.22 NEUTRONS AND NEUTRON INDUCED REACTIONS

The neutron was discovered by the English physicist James Chadwick in 1932. He found that when beryllium is bombarded with α -particles, it emits neutral particles with a mass close to that of the proton.



The neutron produced in this reaction carries sufficient energy to cause additional nuclear reactions in nuclei with which the neutron collides. Chadwick received Nobel Prize in physics in 1935 for his discovery of neutron.

Neutrons are very effective in initiating nuclear reactions because they do not possess electric charge. As such they penetrate nuclei more deeply than any other particle. Neutrons interact with nucleus differently, according to whether they are fast or slow. **Fast neutrons** are those having energies in the range of 100 keV to 50 MeV. **Slow neutrons** are those which have energies not exceeding 100 keV.

The heaviest naturally available element is uranium having an atomic number 92. Enrico Fermi, an Italian physicist, who undertook a systematic study of neutron-induced nuclear reactions, speculated that neutron bombardment of uranium would yield new elements that would be more massive than uranium. In 1934 Fermi found that uranium bombarded with neutrons yielded radioactive products which were assumed to be trans-uranium elements. However, in 1938 Otto Hahn and Fritz Strassmann, the German chemists, using precise radiochemical identification established beyond doubt that the neutron bombardment of uranium produced an isotope of barium (${}^{139}_{56}Ba$). More similar identifications followed. In 1939, two Austrian physicists Lise Meitner and Otto R. Frisch suggested that the neutron caused a division of the uranium nucleus into 'two nuclei of roughly equal size'. They called the process **nuclear fission**, by analogy to the biological fission of a living cell into two parts. Shortly afterwards, it was found that transuranium elements may also form when uranium is bombarded with neutrons. Thus, neutron bombardment of uranium leads sometimes to fission

and sometimes to formation of transuranium elements neptunium ($^{239}_{93}Np$) and plutonium ($^{239}_{94}Pu$).

21.23 NUCLEAR FISSION

Nuclear fission is a neutron-induced nuclear reaction in which a heavy nucleus such as uranium splits into two intermediate lighter nuclei.

In a nucleus there is a competition between the nuclear force, which holds the nucleus together, and the electrostatic repulsion of the protons which tries to tear the nucleus apart. In case of heavy nuclei there is a delicate balance between the nuclear and electric forces. This balance can be easily upset. The energy necessary to cause fission is about 6 MeV.

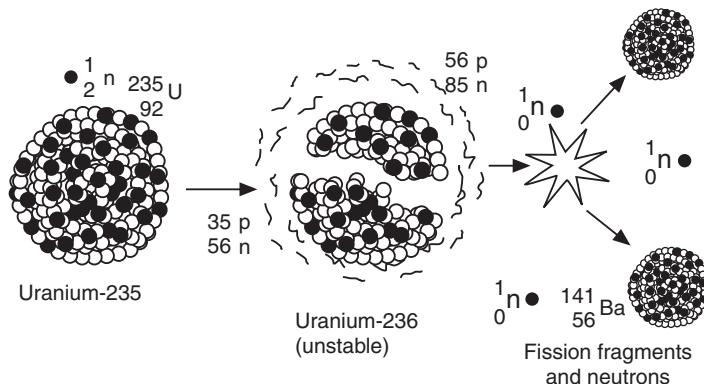
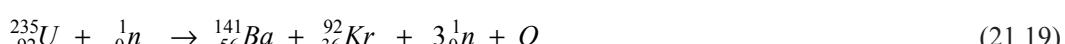


Fig. 21.11

Only certain heavy nuclei can undergo fission and the probability of fission reaction for a particular nucleus depends on the energy of the incident neutrons. Nuclei with odd number of neutrons, $^{233}_{92}U$, $^{235}_{92}U$ and $^{239}_{94}Pu$ undergo fission with slow neutrons. On the other hand, the nucleus $^{232}_{90}Th$ with an even number of neutrons requires fast neutrons with energies of 1 MeV or more.

Nuclear fission is the phenomenon of breaking up the nucleus of a heavy atom into two more or less equal segments with the release of a large amount of energy.

Hahn discovered in 1938 that when uranium was bombarded with thermal neutrons, the uranium nucleus broke up into barium and krypton nuclei of atomic numbers 56 and 36 and liberated 3 neutrons accompanied by a tremendous amount of energy (Fig. 21.11). The nuclear reaction is represented as follows:



The U^{235} nuclei do not all split up into those of Ba and Kr . The fission reaction can proceed in about 40 different ways, yielding different final products. The heavy nuclei may split up into nuclei of several pairs of elements lying in the central region of the periodic table with slightly unequal masses. These are known as *fission fragments*.

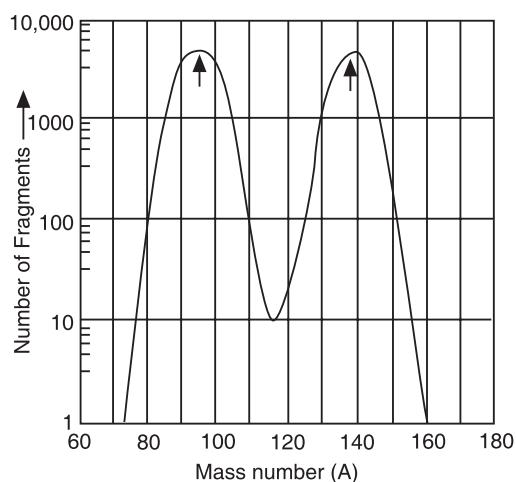


Fig. 21.12

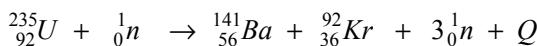
The fission products such as $^{141}_{56}Ba$ and $^{92}_{36}Kr$ are radioactive and undergo disintegrations until stable isotopes are formed. After several steps they form the stable isotopes $^{141}_{59}Pr$ and $^{90}_{40}Zr$ respectively. Radio-chemical analysis showed that nuclides resulting from fission have atomic numbers between 30 and 63 and mass numbers between 72 and 158. Fig. 21.12 shows the mass distribution of the fission fragments from the fission of U-235 nuclei. It is most likely that one fragment will have a mass number of about 95 and the other about 140. Most of the energy (about 80%) released in fission goes into kinetic energy of the two fission fragments, and the remaining (20%) appears as decay products (β and γ rays) and kinetic energy of neutrons emitted in the fission process. The neutrons typically have energies of one to several MeV.

The most important features of nuclear fission is that the process is accompanied by

- (i) the release of a large amount of energy and
- (ii) the emission of two or more energetic neutrons which under appropriate conditions cause fission in neighbouring nuclei.

21.23.1 Energy Released during Nuclear Fission

The energy released during the process of fission is known as *nuclear energy or atomic energy*. The amount of energy released during fission may be estimated using mass defect method. We illustrate the method by taking the fission reaction (21.19) as an example.



Actual mass before the fission reaction

Mass of U-235 nucleus = 235.125 amu

Mass of the neutron = 1.009 amu

Total mass = 236.134 amu

Actual mass after the fission reaction

Mass of Ba-141 nucleus = 140.958 amu

Mass of Kr-92 nucleus = 91.926 amu

Mass of three neutrons = 3.027 amu

Total mass = 235.911 amu

Mass decrease during the reaction = $(236.134 - 235.911)$ amu = 0.223 amu

$$= 0.223 \text{ amu} \times 931.4 \text{ MeV/amu} = 207 \text{ MeV}$$

Therefore, each fission event produces about 200 MeV of energy. The energy converted per atom is roughly 10^8 times greater in nuclear reactions than in chemical reactions. The energy, $200 \text{ MeV} = 200 \times 10^6 \times 1.6 \times 10^{-19} \text{ J} = 3.2 \times 10^{-11} \text{ J}$, is in fact very small. We spend about 5J when we do a simple job, say picking up an object from a table. However, as there are billions of nuclei even in a small sample of uranium, the total energy that could be produced will be enormous.

As an example, let us compute the energy produced when nuclei contained in 1 gram of U-235 undergo fission.

Atomic weight of U-235 = 235

There are 6.023×10^{26} atoms in 235 kg of Uranium.

$$\therefore \text{Number of nuclei in one gram of } U-235 = \frac{6.023 \times 10^6}{235 \times 1000} = 2.56 \times 10^{21}$$

$$\begin{aligned}\text{Energy produced by 1 gm of U-235} &= (2.56 \times 10^{21}) \times 200 \text{ MeV} \\ &= 2.56 \times 10^{21} \times 3.2 \times 10^{-11} \text{ J} \\ &= 8.2 \times 10^{10} \text{ J.}\end{aligned}$$

$$1 \text{ kWh} = (1 \times 10^3 \text{ J/s}) (3600 \text{ s}) = 3.6 \times 10^6 \text{ J}$$

$$\therefore \text{Energy produced by 1 gm of U-235} = \frac{8.2 \times 10^{10} \text{ J}}{3.6 \times 10^6 \text{ J/kWh}} = 22.8 \times 10^3 \text{ kWh} = 22.8 \text{ MWh.}$$

We can as well express the energy in calories.

$$1 \text{ cal} = 4.187 \text{ J}$$

$$\text{Energy produced by 1 gm of U-235} = \frac{8.2 \times 10^{10} \text{ J}}{4.187 \text{ cal}} = 2 \times 10^{10} \text{ cal}$$

The quantity of coal required to produce an energy equivalent to the above figure may be calculated assuming that the particular grade of coal used in thermal power plant yields 7×10^3 kcal energy per kg.

$$\text{Quantity of coal required} = \frac{2 \times 10^{10} \text{ cal}}{7 \times 10^3 \text{ kcal/kg}} = \frac{2 \times 10^{10}}{7 \times 10^6} \text{ kg} \approx 3000 \text{ kg.}$$

It means that about 3 tonnes of coal is required to produce as much energy as 1 gm of U-235 produces.

Enrico Fermi, the Italian physicist was honoured in 1938 with the Nobel prize in physics for his discovery of nuclear reactions brought about by slow neutrons.

21.23.2 Theory of Nuclear Fission

Niels Bohr and John A. Wheeler explained in 1939 the process of nuclear fission using the liquid drop model. According to this model the stable nucleus may be compared to a spherical liquid drop. The shape of the drop depends on the balance between the short range forces and coulomb repulsion forces. When a nucleus captures a neutron, a compound nucleus is formed. The nucleus is excited by being given mechanical energy and is in a state of higher energy. Just as an excited liquid drop oscillates, the nucleus will be in a state of oscillations. The oscillations tend to distort the spherical shape so that the nucleus assumes an ellipsoid shape. The restoring forces arising from short-range nuclear forces tend to make the nucleus return to its original spherical shape. If the excitation energy is sufficiently large, the ellipsoid deforms further into a dumb bell shape. The coulomb force of repulsion between the two parts of the deformed nucleus can overcome the short-range attractive forces causing the nucleus to split and the fragments to separate with higher speeds. Each of the fragments will then quickly take spherical form because within it the attractive nuclear forces predominate again. A possible sequence of stages of fission is shown in Fig. 21.13. Generally, the two fission fragments are not of the same size. As neutron/proton ratio for heavy nuclei is higher than for lighter nuclei, the fragments contain excess neutrons. To reduce the excess, the fragments release two or three neutrons as soon as they are formed.

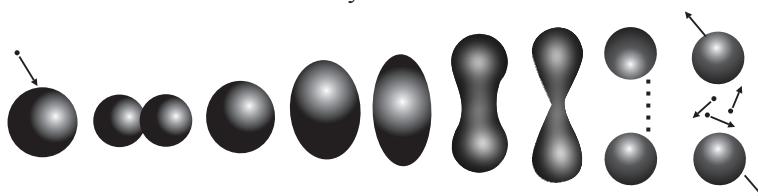


Fig. 21.13

Obviously, a certain minimum amount of energy must be available to the nucleus to deform it enough and cause fission. The energy is called the **threshold** or **critical energy**, can be calculated. Basing on such calculations, Bohr and Wheeler predicted that U-235 nucleus undergoes fission with slow neutrons. They also showed that U-238 does not undergo fission unless it is bombarded with fast neutrons having energies of 0.9 MeV or more.

21.24 NUCLEAR CHAIN REACTION

A nuclear chain reaction is a self-propagating process in which the number of neutrons goes on multiplying rapidly almost in geometrical progression till the total fissionable nuclei in the material are fissioned.

A neutron striking a large fissionable nucleus initiates the fission process. The neutrons released in the process strike other fissionable nuclei in turn and induce fission in them. The average number of neutrons emitted per one thermal neutron absorbed is about 2.5. This abundant emission of neutrons enhances the possibility of fission occurring in more nuclei. If each of the newly emitted neutrons induces fission in a neighbour nucleus, the

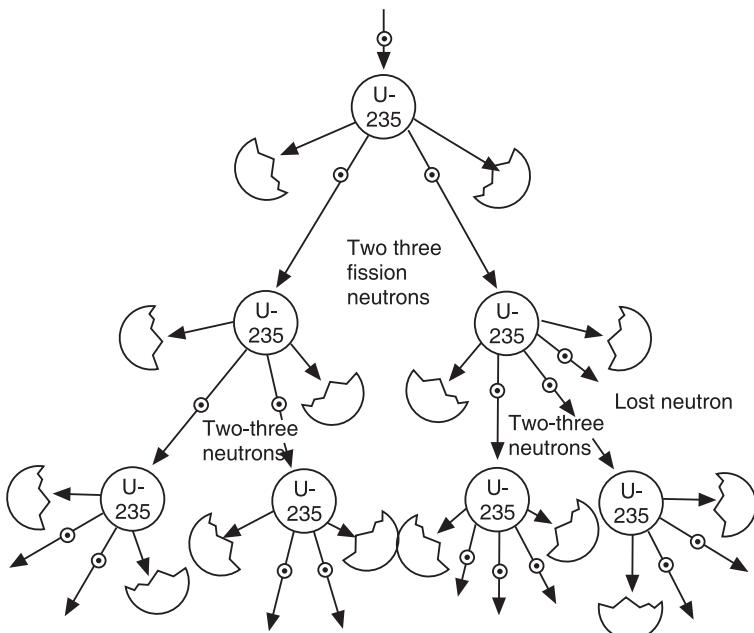


Fig. 21.14

number of fission events grows cumulatively at a rapid rate. For example, let us say that three neutrons are emitted in a uranium fission act. The three neutrons could produce fission in three more uranium nuclei, the 3^2 neutrons emitted by the nuclei could produce the nine fissions, the 3^3 neutrons produce 27 fissions, the resulting 3^4 ($= 81$) neutrons produce 81 fissions and 3^5 ($= 243$) neutrons could produce 243 fissions and 3^6 ($= 729$) neutrons and so on. This process is called a **chain reaction**. It is shown in Fig. 21.14. *A nuclear chain reaction is defined as the process in which the number of neutrons goes on multiplying rapidly in a geometrical progression till the total fissionable nuclei are fissioned.*

The neutrons released in the process strike other fissionable nuclei in turn and induce fission in them. The average number of neutrons emitted per one thermal neutron absorbed is about 2.5. This abundant emission of neutrons enhances the possibility of fission occurring in more nuclei. If each of the new neutrons induces fission in neighbouring nuclei, the number of fissions will grow cumulatively, at each successive stage at a rapid rate.

In practice, not all neutrons produced in fission participate in further fission acts. Some of the neutrons are lost as a result of the following processes.

1. capture of neutrons by uranium, which may not lead to fission.
2. capture of neutrons by other nuclei in the sample and
3. escape out of the sample without being captured.

These losses affect the course of chain reaction.

21.24.1 Controlled Chain Reaction

The course of a chain reaction is determined by the availability of neutrons to continue fission events in the sample. The condition required is conveniently expressed in terms of a **multiplication factor**, k , of the system, which is defined as

$$k = \frac{\text{Number of neutrons in a particular generation}}{\text{Number of neutrons in the preceding generation}} \quad (21.20)$$

When $k = 1$ production of neutrons by fission is equal to their loss by leakage and the neutron population is constant. The chain reaction is just possible and will be **self-sustaining**. When the above criteria are satisfied, there would be a favourable balance between the net production of neutrons by fission acts and the loss of neutrons due to the various reasons and the chain reaction occurs at an even rate. If the fission reaction is controlled, the corresponding energy release is also under control. If the reaction is controlled over an extended period, then the energy produced can be utilized for the benefit of humanity. Nuclear reactors utilize controlled chain reactions in their operation.

21.24.2 Critical Mass or Critical Size

Fission occurs throughout the volume of the reacting body, and neutron leakage takes place through the surface of the body. As the size of a body is changed, the volume changes according to the cube of its dimensions, while its area changes as the square of its dimensions. As neutrons production is proportional to volume and neutron leakage to surface area, uranium is taken in the form of a sphere because for a given volume, the sphere has the smallest surface area. Let us consider uranium sample taken in the form of a sphere of radius, r . Then

$$\text{Neutron production rate, } N_1 \propto \frac{4}{3}\pi r^3 = C_1 r^3$$

$$\text{Neutron loss rate (in non-fission events), } N_2 \propto \frac{4}{3}\pi r^2 = C_2 r^2$$

$$\text{and Neutron leakage rate, } N_3 \propto \frac{4}{3}\pi r^2 = C_3 r^2$$

where C_1 , C_2 , and C_3 are proportionality constants.

$$\text{When } k = 1 \quad N_1 > N_2 + N_3,$$

$$\text{i.e.,} \quad C_1 r^3 > C_2 r^2 + C_3 r^2$$

$$\text{or} \quad (C_1 - C_2)r > C_3$$

$$\therefore r > \frac{C_3}{(C_1 - C_2)} = C \text{ (say)} \quad (21.21)$$

The above equation implies that the larger the size of the body, lesser the escape rate as compared to the production rate. Therefore, there is always a certain size called **critical size**, C of the material. For a self-sustained reaction to take place, the sample size must be greater than the critical size.

If the size of the body is less than the critical size, i.e. it is **subcritical** so that $k < 1$, then loss of neutrons is greater than their production. Each new generation of neutrons produce a decreasing number of fission acts and the reaction terminates soon for want of neutrons. Consequently, chain reaction is not possible under these conditions.

21.24.3 Uncontrolled Chain Reaction

If $k > 1$, the size is greater than the critical size and is called **supercritical** size. Each new generation of neutrons produces an increasing number of fission acts and chain reaction becomes uncontrollable. The neutron population increases exponentially with time and the chain reaction builds up at an alarming rate leading to explosion, accompanied by the liberation of enormous energy and a rise of temperature of the surrounding medium to several million degrees. Atomic bombs utilize uncontrolled chain reactions for their operation.

Atom Bomb

An atom bomb is an example where uncontrolled chain reaction takes place leading to an explosion. Natural uranium is composed of 99.3% $U-238$ and only 0.7% of the fissionable $U-235$. The material used in weapons is enriched to 99% or more $U-235$. The critical mass of such highly enriched uranium is kept separated into several subcritical masses, so that chain reaction is not initiated (Fig. 21.15). The masses are brought together by suddenly firing explosives like TNT. The segments form one large mass which becomes supercritical. Neutrons entering this supercritical mass from a Ra-Be source initiate a chain reaction which develops into an uncontrollable reaction.

It results in a violent explosion with the liberation of tremendous energy. Everything in the immediate vicinity will be heated to temperatures of 5 to 10 million degrees. The atmosphere gets heated up and expands suddenly exploding every thing near by.

There is no danger of atomic explosion in the uranium mineral deposits in the earth. First, only 0.7% of natural uranium contains fissionable $U-235$. Secondly, uranium occurs only in compound form and is not found in pure form in nature.

21.25 NUCLEAR ENERGY

Energy requirements of mankind are at present met by fossil fuels, namely coal, oil and natural gas. These sources are being fast depleted and will be nearly exhausted in the near future. It is necessary that alternative sources of power are to be searched for. Nuclear power is one of the alternative sources. The energy reserve in the form of uranium is many times greater than that of fossil fuels.

The nucleus is a store-house of enormous power. Heavier nuclei, uranium and plutonium, can be split into fragments by neutrons and such a fission process liberates tremendous energy. Nuclear fusion power is believed to be an inexhaustible source of energy and is without the radioactive hazards and pollution threats.

Example 21.13. A reactor is developing energy at the rate of 32 MW. How many atoms of $U-235$ undergo fission per sec.? Assume that on the average an energy of 200 MeV is released per fission.

Solution. Energy released per second $E_1 = 32 \times 10^6 \text{ J}$

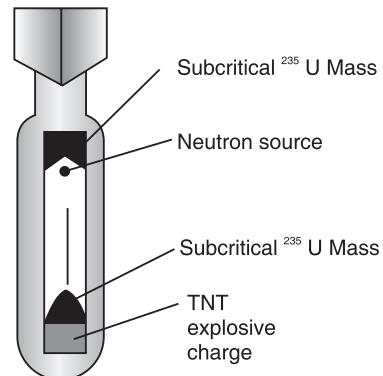


Fig. 21.15

Energy released per fission $E_2 = 200 \times 10^6 \text{ eV} = 200 \times 10^6 \times 1.6 \times 10^{-19} \text{ J}$

$$\text{Number of atoms undergoing fission /second } N = \frac{E_1}{E_2} = \frac{32 \times 10^6 \text{ J}}{200 \times 10^6 \times 1.6 \times 10^{-19} \text{ J}} = 10^{18}.$$

Example 21.14. A city requires 100 MW of electrical power on the average. This is to be supplied by a nuclear reactor of efficiency 20% using U-235 as nuclear fuel. Calculate of the fuel required for one-day operation. Given that the energy released per fission of U-235 nuclide is 200 MeV.

Solution. Energy required per day $E_1 = 100 \times 10^6 \times 86400 \text{ J} = 8.64 \times 10^{12} \text{ J}$

$$\text{Let the fuel required be } m \text{ kg. Number of atoms in } m \text{ kg.} = \frac{m \times 6.023 \times 10^{26}}{235}$$

Energy released per fission = 200 MeV = $3.2 \times 10^{-11} \text{ J}$

$$\text{Total energy released} = \frac{m \times 6.023 \times 10^{26} \times 3.2 \times 10^{-11}}{235} \text{ J}$$

$$\text{Useful energy released } E_2 = \frac{m \times 6.023 \times 10^{26} \times 3.2 \times 10^{-11}}{235} \text{ J} \times \frac{20}{100} = m(16.403 \times 10^{12} \text{ J})$$

$$\text{Therefore, } m = \frac{8.64 \times 10^{12} \text{ J}}{16.403 \times 10^{12} \text{ J}} = 0.5267 \text{ kg/day.}$$

Example 21.15. Calculate the power output of a nuclear reactor which consumes 25 gm of U-235 per day. Assume 5% reactor efficiency. Energy released per fission is 200 MeV.

Solution. Number of nuclei in 235 kg of U-235 = N_A

$$\text{Number of nuclei in 25 gm} = \frac{6.02 \times 10^{26}}{235 \times 40} = 6.4 \times 10^{22} \left(\because 25 \text{ gm} = \frac{1}{40} \text{ kg} \right)$$

Energy released per fission = 200 MeV = $3.2 \times 10^{-11} \text{ J}$

Energy produced by 6.4×10^{22} nuclei = $6.4 \times 10^{22} \times 3.2 \times 10^{-11} \text{ J} = 2.05 \times 10^{12} \text{ J}$

Efficiency = 5%

$$\text{Energy converted to electrical power} = \frac{2.05 \times 10^{12} \times 5}{100} \text{ J} = 1.025 \times 10^{11} \text{ J}$$

Time taken to consume 25 gm of U-235 = 1 day = $24 \times 3600 \text{ sec} = 8.64 \times 10^4 \text{ secs.}$

$$\text{Power output} = \frac{\text{Energy}}{\text{time}} = \frac{1.025 \times 10^{11}}{8.64 \times 10^4} \text{ J/s} = 1.19 \text{ MW}$$

Example 21.16. A nuclear reactor consumes 20.4 kg of U-235 in 1000 hrs. of operation. Assuming that on an average 200 MeV of energy is released per fission of a single U-235 nucleus, determine the power developed by the reactor.

$$\text{Solution. Number of nuclei in 20.4 kg of U-235} = \frac{N_A m}{M} = \frac{6.02 \times 10^{26} \times 20.4 \text{ kg}}{235 \text{ kg}}$$

$$= 523 \times 10^{25}$$

Energy released per fission = 200 MeV = $3.2 \times 10^{-11} \text{ J}$

$$\text{Total energy released} = 5.3 \times 10^{25} \times 3.2 \times 10^{-11} \text{ J} = 1.67 \times 10^{15} \text{ J}$$

$$\text{Power} = \frac{\text{Energy}}{\text{time}} = \frac{E}{t} = \frac{1.67 \times 10^{15} \text{ J}}{1000 \text{ hrs.}} = \frac{1.67 \times 10^{15} \text{ J}}{3.6 \times 10^6 \text{ s}} 4.65 \times 10^8 \text{ W} = 465 \text{ MW}$$

21.26 NUCLEAR REACTORS

We can sum up the three important features of nuclear fission reaction as follows:

- (a) **Energy Emission:** Each fission reaction produces about 200 MeV of energy and the total energy produced by a small sample of uranium, say 1 gm, is as enormous as $8.2 \times 10^{10} \text{ J} = 22.8 \text{ MWh}$. Most of the energy is carried by fission fragments in the form of kinetic energy.
- (b) **Neutron Multiplicity:** Each fission event is accompanied by the emission of a number of neutrons which in turn cause fission of other nuclei.
- (c) **Delayed Neutrons:** Many of the neutrons are emitted at the instant of fission. They are known as *prompt neutrons*. About 1% of the neutrons are emitted due to decays of fission fragments and they are called *delayed neutrons*.

The last feature enables mechanical control of the reaction rate and keeps the reaction from proceeding too rapidly. The second feature can be used to build a self-sustaining chain of nuclear fissions. The energy produced in the reactions can be extracted as heat and used to boil water, the resulting steam can then be used to drive a turbine to generate electrical power.

*A system in which nuclear fission is produced in a controlled, self-sustaining chain reaction is known as **nuclear reactor**.* Enrico Fermi built the first nuclear reactor in 1942. A very large number of nuclear reactors are in operation in different countries as on today. They differ widely in design and construction depending on their and use. A nuclear reactor essentially consists of seven components arranged in different zones, each of which serves a definite purpose.

Fuel

Nuclear reactors can use pure fissile materials but it is easier and cheaper to use mixtures of isotopes. Naturally occurring uranium consists of 99.3% *U-238* and only 0.7% *U-235*. *U-238* is for all practical purposes not fissionable. Often natural uranium in which there is one *U-235* atom per 140 *U-238* atoms is used as nuclear fuel. In order to be useful as a fuel, the concentration of *U-235* must be substantially increased (upto 3%) in naturally available uranium. This process is called *enrichment*. In many reactors, uranium enriched in *U-235* is used as fuel. Another fissionable material is *Pu-239*. This does not occur in nature. It is produced through neutron capture by the non-fissionable *U-238*. The process of plutonium fuel production from uranium is known as *breeding*.

Moderators

The kinetic energy of neutrons emitted in fission process is of the order of a few MeV. Such high energy neutrons are known as *fast neutrons*. The fast neutrons have relatively low probability of inducing fission. Nuclear fission is much more probable with *slow neutrons*. The fast neutrons must be therefore, slowed down in order to increase their chances of initiating fission events. If a neutron is scattered from a heavy nucleus like uranium, the energy of the neutron is not changed. On the other hand in a collision with a very light nucleus, the neutron can lose substantial energy. Materials consisting of atoms of lower atomic mass are used for slowing down neutrons. Such materials are called **moderators**.

The commonly used moderators are water, heavy water, graphite and beryllium. Heavy water is considered to be the best moderator, since it does not absorb neutrons. When a solid moderator is required, graphite is commonly used. Ordinary water is a good moderator, but has high neutron absorption cross-section. Beryllium and its oxide are also good moderators, but they are expensive, toxic and have poor mechanical property.

Reflectors

Some of the neutrons generated in the fission process may leak away without being absorbed. It is necessary to conserve neutrons so that we minimize the consumption of fissile material and keep the size of the reactors small. To reduce the neutron loss due to leakage the inner surface of the reactor is surrounded by a material which reflects the neutrons back into the core. Such materials are called **reflectors**. Reflectors are made of nickel, thorium or other suitable materials.

Coolants

Intense heat is generated within the reactor core due to nuclear fission reactions. This heat must be removed using coolants for the safe operation of reactor. Ordinary water, heavy water and liquid metals are used as coolants. The **coolant** keeps the fuel assembly at a safe temperature; at the same time the heat carried away by the coolant is used in the heat exchange for utilization in power generation.

Control Rods

The rate of reactions in a nuclear reactor is controlled by **control rods**. Since the neutrons are responsible for the progress of the chain reactions, suitable neutron absorbers are employed to achieve control of reaction rate. Cadmium and boron are the most frequently used materials. The *control* procedure involves the insertion or withdrawal of these materials, taken in the form of rods, into or from the reactor core. With the control rod fully inserted, enough neutrons are absorbed so that the average number of neutrons available to cause new fissions is less than one per fission reaction. As the rod is slowly withdrawn, the average number of available neutrons increases until it is just equal to one per reaction. At that time the reactor is said to be *critical*. During the operation, the position of the control rod is continually adjusted so that energy is released at a steady rate.

Structural and Cladding materials

In a reactor, structural materials have to be used in the form of mechanical support for the various components. They are also used for holding the fuel, coolants, control rods and measuring instruments. Uranium readily reacts with air, water and other fluids. Hence **cladding** is required to pack fuel elements to isolate them. The cladding also prevents escape of the fission fragments. Zirconium has been found to be an excellent structural and cladding material. Titanium is also a good material but it is very expensive.

Reactor Shielding

A nuclear reactor is a powerful source of highly penetrating neutrons and γ -radiation, which are very harmful to life. Therefore, a thermal shield and a biological shield are used to minimize the effects of the harmful radiation. The thermal shield, usually a wall of steel, is placed between the reactor and the biological shield to protect the latter from excessive heating and damage. The biological shield is usually made of concrete. Concrete wall of about 2 m thick surrounding the reactor serves as an adequate shield.

21.26.1 Nuclear Reactor

A schematic of a nuclear reactor is shown in Fig. 21.16.

It consists of several zones each of which serves a definite purpose. The central part is called the **active zone** or **core**. The core is made up of moderator blocks with slots in which fuel channels are housed. A **fuel channel** is a metal tube containing the fuel elements. The **fuel element** consists of uranium containing slug clad in a stainless steel or zirconium metal casing. The fuel elements are grouped lengthwise in the fuel channel and form a common construction called the **fuel assembly**. Inlet and outlet pipes are incorporated in the fuel channel for circulation of coolant. The active zone is filled with a moderator. The fuel assembly and moderator are cooled by the coolant passing through the fuel channels. The heated coolant then passes through an outlet pipe to the collecting tank. The heat can be used to generate steam and drive the turbine.

21.26.2 Types of Reactors

Reactors are designed and fabricated for different purposes, mainly for generation of power and for use in research.

Power Reactors

In a power reactor, heat produced in the nuclear reactor fuel is extracted to drive a turbine connected to an electrical generator. In practice two different types of reactors are used for producing power.

- (i) **Boiling – Water Reactor:** In this type of reactors, a stream of water circulates through the core. The heat turns the water into steam, which is then directly fed to the turbine. Therefore, an external steam generators is not required. There is a considerable financial saving due to this. The disadvantage of this type of reactor is that the water can become radioactive; and a rupture of the pipes near the turbines could result in a serious accident with the spread of radioactive materials.
- (ii) **Pressurized – Water Reactor:** In this type of reactor, heat is extracted in two steps. Water is circulated through the core under high pressure, which prevents the water turning into steam. This hot water in turn heats a secondary water system, which delivers steam to the turbine. Since steam is not taken from the reactor core, it is not radioactive. The main disadvantages of these reactors are the necessity of using highly enriched fuel and expensive structural materials.

Research Reactors

Research reactors produce high neutron fluxes for research. These neutrons are used as projectiles in nuclear reactions to produce new nuclides. These nuclides are used as tracers and for medical purposes.

21.26.3 Site Selection

The main considerations for site selection for erecting nuclear reactors are: (i) availability of fuel (ii) availability of water resources (iii) facilities for removal and disposal of radioactive waste.

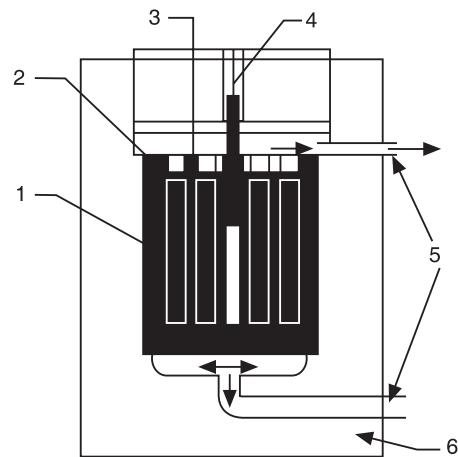


Fig. 21.16: Nuclear Reactor—1. Nuclear Fuel 2. Nuclear Reflector 3. Moderator 4. Control Rod 5. Coolant 6. Shield.

21.27 NUCLEAR POWER PLANT

A schematic diagram of nuclear power plant is shown in Fig. 21.17. A nuclear power plant consists of a nuclear reactor, heat exchanger (i.e., steam generator), turbine, electric generator and condenser.

The heat liberated in the reactor due to the nuclear fission of the fuel is received by the coolant circulating through the reactor core. Hot coolant leaves the reactor at top end and then flows out into tubes of steam generator (boiler) and passes on its heat to the feed water. The steam produced is passed through the turbine of special design and drives it. The turbine in turn drives an electric generator developing electrical power. From the turbine, after work is performed, steam flows to condenser. Pumps are provided to maintain the flow of coolant and feed water.

21.28 NUCLEAR FUSION

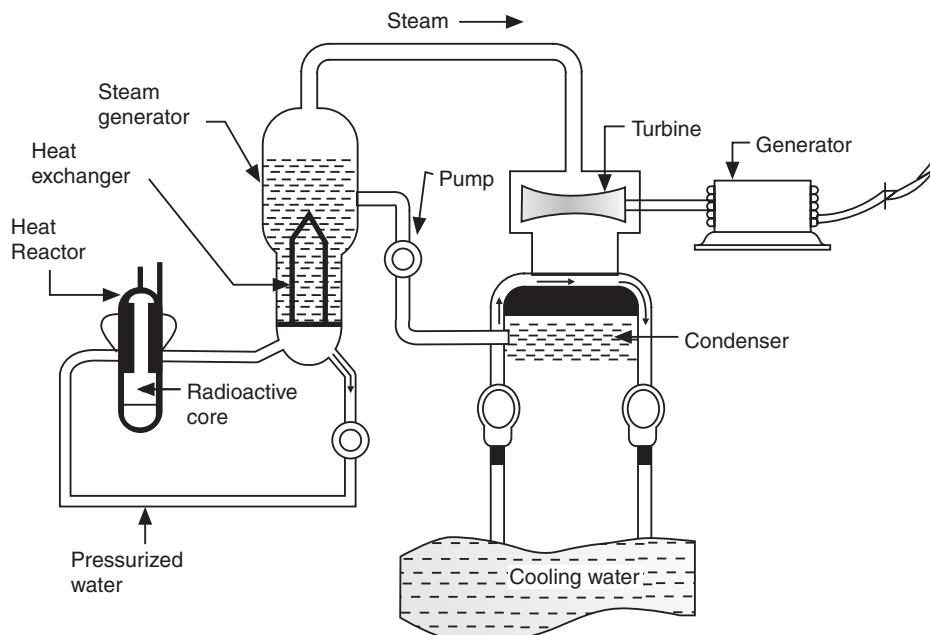


Fig. 21.17: Schematic of Nuclear Power Plant.

Nuclear fusion is the process of combining two lighter nuclei into a heavier nucleus. Since the mass of final nucleus is less than the rest masses of the original nuclei, there is a loss of mass, which gets converted into large amounts of energy. Fusion is the source of energy for the sun and the stars.

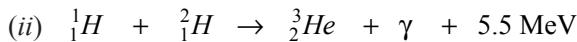
21.29 FUSION REACTION IN STARS

Sun is an average sized star and has been emitting enormous energy at the rate of 4×10^{26} J/s for the past 4×10^9 years. Such huge amount of energy cannot be liberated through ordinary sources of energy such as chemical reactions. Therefore, a series of nuclear fusion reactions are the only possible source of energy being produced by stars and the sun. The sun is mostly made up of hydrogen and helium. It was suggested that four hydrogen nuclei combine, through suitable series of nuclear reactions, and produce one helium nucleus releasing energy in these processes. The series of nuclear reactions that convert four protons into a helium nucleus are

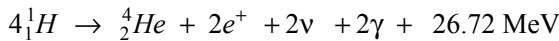
known as *thermonuclear reaction cycles*. Though many nuclear reactions are possible in sun, the two most common ones are the proton-proton cycle and the carbon cycle.

Proton-proton cycle

The simplest process of fusion is known as *proton-proton cycle*. It consists of fusing two hydrogen nuclei, which finally give a helium nucleus. The sequence of the fusion reactions is



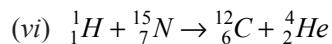
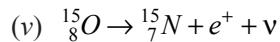
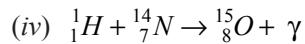
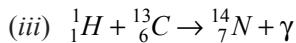
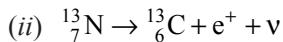
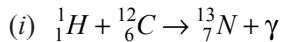
In the first stage, two hydrogen nuclei combine to form a deuteron with release of 0.42 Mev of energy, a positron and a neutrino. In the second stage, the deuteron nucleus combines with another hydrogen nucleus to form a ${}_2^3He$ nucleus with release of 5.5 MeV energy and γ -radiation. In the final state, two ${}_2^3He$ nuclei combine to give a helium nucleus and two hydrogen nuclei. The net effect of the sequence is



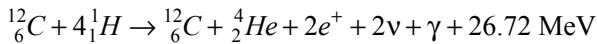
Thus, the result is the gradual depletion of protons and the building up of helium concentration. The production of energy in sun is mainly due to proton-proton cycle.

Carbon cycle

Another self-sustaining process is known as *carbon-nitrogen cycle*. In this process, carbon acts as a catalyst and four hydrogen nuclei fuse to yield a helium nucleus, positrons, neutrinos and energy of 26.72 MeV. The steps in the process are as follows:



The net result is



In case of medium and heavy stars the main part of the energy production is due to the carbon cycle. On the other hand, the main part of the energy production in lighter stars is due to the proton cycle.

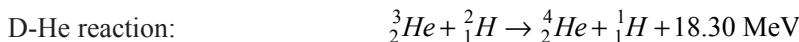
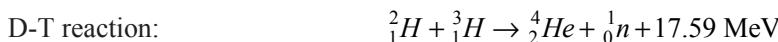
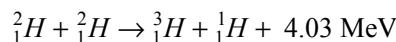
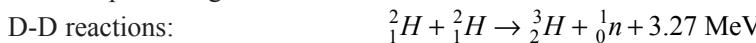
When all the available hydrogen has been converted to helium, the star (or sun) will contract and its temperature will increase till helium starts burning. Helium fusion requires more thermal energy than that required for hydrogen fusion. When helium is exhausted, the star will again contract and its temperature will increase, till carbon burning occurs. Such processes will continue until ${}^{56}Fe$ is reached. The star then collapses under its own gravitational forces to become a white dwarf, or a neutron star or a black hole. It may even undergo supernova explosion.

21.30 CONTROLLED THERMONUCLEAR REACTIONS

Fusion can take place at a very high pressure and temperature of the order of 10^6 K because it is only under these conditions that nuclei are able to overcome their mutual coulomb repulsion. The energy released in nuclear fusion is known as **thermal nuclear energy** and it is possible to produce nuclear fusion under controlled conditions.

For achieving the fusion the two nuclei should have enough kinetic energies to overcome the mutual Coulomb repulsion. The force of repulsion will go on increasing as the nuclei approach each other. It becomes therefore necessary that the nuclei approach each other at velocities high enough to overcome the Coulomb repulsion. It is calculated that a deuterium atom should have about 0.25 MeV energy to overcome Coulomb repulsion. We can estimate the temperature to which the deuterium gas is to be heated to attain the required energy. Using $(3/2) kT = 0.25$ MeV, we find that it corresponds to a temperature of the order of 10^9 K. Therefore, fusion reaction can be accomplished by heating the fuel to extremely high temperature. Because of the requirement of such high temperatures, the fusion reactions are called *thermonuclear reactions*. If a gas is heated to such a very high temperature, the electrons are stripped off the atoms; and the bare positive nuclei and the negative electrons move about freely. This mixture is known as **plasma**, which is regarded as the fourth state of matter. The plasma as a whole is electrically neutral. In the state of plasma and at temperatures of the order of 10^8 K, the nuclei will have energies of the order of a few keV and the bare nuclei have a probability to overcome Coulomb repulsion and produce fusion reactions.

Hydrogen and its isotopes are more suitable for controlled thermonuclear reactions as their nuclei carry smaller charge and hence the smaller is the repulsive force. The following are the more promising fusion reactions.



Out of the above four reactions, the D-T (deuterium-tritium) reaction has large energy release and perhaps the best one. This reaction liberates energy about 5.9 MeV/nucleon which is about six times as that produced in a fission reaction. Such enormous amount of energy release in fusion reaction makes it an attractive process for generating electrical power.

21.30.1 Ignition Temperature

Plasma emits energy as radiation due to the electrical interactions between the nuclei and electrons. This constitutes a loss of energy. At moderately high temperatures, more energy is radiated away from the plasma than is produced by fusion. As the temperature is further increased, production of fusion energy increases more rapidly than emission of radiation. At a sufficiently high temperature, the fusion becomes sustained. The temperature at which the rate of energy production becomes greater than the rate of energy loss is called the *critical ignition temperature*. It has been estimated that for D-T reaction $T_c \approx 4$ KeV while it is 10 times higher for D-D reaction. (In fusion studies, temperatures are expressed in KeV; 1 KeV = 11×10^6 K). This is why the D-T reaction is considered to be the most promising reaction for power generation.

21.30.2 Lawson Criterion

Apart from high temperature requirement, there are two other essential requirements to increase the probability of collisions, which would result in fusion of nuclei.

- (i) high density of nuclei, n and
- (ii) a long *confinement time*, τ during which the high temperature and density must be maintained.

In 1957 J.D. Lawson showed that for the condition of self-sustained fusion reaction the product $n\tau$ must exceed a certain minimum value. This is known as *Lawson criterion*.

In case of *D-T* reaction, calculations showed that

$$n\tau \geq 10^{20} \text{ s/m}^3 \quad (21.22)$$

Two techniques are under development for heating the fusion fuel to the required high temperature and to confine the heated plasma long enough for the fusion to produce power. They are known as magnetic confinement and inertial confinement.

21.30.3 Magnetic Confinement

The plasma can be heated by passing electric current through the plasma. We may regard the plasma column as made up of a large number of plasma threads, each carrying an electric current in the same direction. They attract one another as a result of which the whole plasma column is compressed perpendicular to its axis. This is known as the *pinch effect*. The sudden adiabatic compression results in the heating of the plasma column.

The hot plasma may be confined by the application of a magnetic field. The particles in the plasma must be prevented from reaching the container wall for adequately long time, since contact with the wall will immediately cool them. Fig. 21.18 shows toroidal magnetic confinement geometry. The device is called a *tokomak*.

The tokomak is a doughnut-shaped reaction chamber. The chamber contains plasma with a circular cross-section where a toroidal magnetic field is provided for confinement and a toroidal electric field to maintain current through the plasma. The electric current generates a circular magnetic field in the cross-section of the toroid. The combination of this field and the toroidal field produces helical magnetic field lines which confine the plasma.

21.30.4 Inertial Confinement

An alternative approach to achieve fusion in a self-sustained manner is known as the inertial confinement.

In case of plasma, particle density is of the order of 10^{20} per m^3 . Therefore, Lawson criterion can be satisfied if the confinement time is of the order of 1 sec or more. If the fusion fuel could be used in the condensed state where the density is of the order of 10^{28} per m^3 , then the confinement time would be of the order of 10^{-8} s. Therefore in inertial confinement the fuel is compressed to high densities for very short confinement times.

A small D-T fuel pellet of about 1 mm diameter is injected into the spherical cavity of the reaction chamber. When the pellet reaches the center, several symmetrically located high-energy laser beams are focused simultaneously on the pellet for about 10^{-9} s. This causes

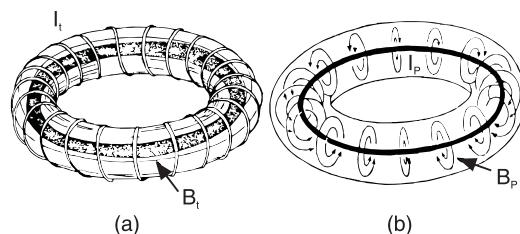


Fig. 21.18

the surface of the pellet to evaporate. The particles coming out produce a reaction force on the core of the pellet causing an implosion. The implosion causes an increase in temperature and also compresses the material to an extremely high density. When the temperature of the core reaches ignition temperature, fusion reactions cause the pellet

to explode. Fusion could occur if the pellet stays together for a sufficient time. It depends on the inertia. If the confinement time is made very short ($\leq 10^{-9}$ s), the particles do not move appreciably from their initial positions because of their own inertia. Hence this technique is called *inertial confinement*. After an interval of a second or more, another pellet is injected and the process is repeated.

Example 21.17: The fusion of ${}_1^2\text{H} + {}_1^2\text{H} \rightarrow {}_2^4\text{He} + Q$ is proposed to be used for the production of industrial power. Assuming the efficiency of the process to be 30%, find how many kg of deuterium will be consumed in a day for an output of 50 MW. Given: mass of ${}_1^2\text{H} = 2.01478$ u and mass of ${}_2^4\text{He} = 4.00388$ u.

Solution: Energy released per D-D reaction

$$Q = [2(2.01478) - 4.00388]931.4 \text{ MeV} = 23.92 \text{ MeV}$$

$$\text{Actual output per two atoms} = 30\%Q = (0.3)(23.92) \text{ MeV} = 7.175 \text{ MeV}$$

$$\text{Actual output per one atom} = 3.5875 \text{ MeV.}$$

$$\text{Number of deuterium atoms required} = \frac{50 \times 10^6 \text{ J}}{3.5875 \times 10^6 \times 1.602 \times 10^{-19} \text{ J}} = 8.7 \times 10^{19}.$$

$$\text{Equivalent mass of deuterium required per second} = \frac{8.7 \times 10^{19} \times 2.01478}{6.02 \times 10^{26}} \text{ kg/s} = 2.9 \times 10^{-7} \text{ kg/s}$$

$$\therefore \text{Deuterium requirement per day} = (2.9 \times 10^{-7} \text{ kg/s})(8.64 \times 10^4 \text{ s}) = 0.025 \text{ kg.}$$

21.31 FUSION REACTOR

A fusion power reactor is shown in Fig. 21.20.

The fusion energy of the nuclei is conveyed to the neutrons that are released in the reaction. The energy is extracted by a lithium blanket surrounding the reactor core. Lithium transfers the energy to the turbogenerator through a heat exchanger.

The chief advantage of fusion over fission and other processes is that it offers a low cost and pollution-free energy.

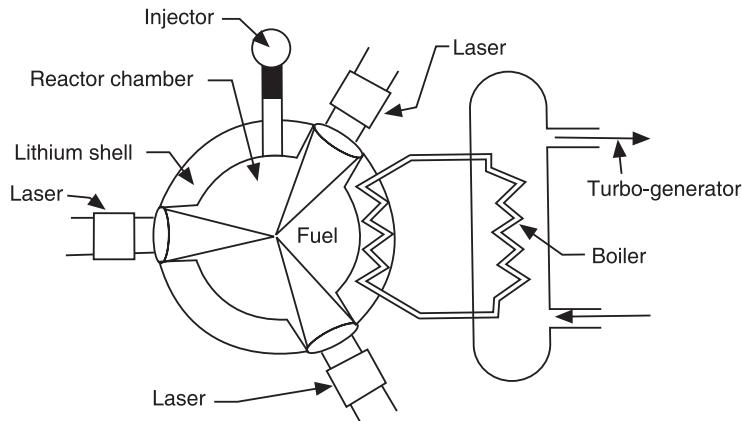


Fig. 21.19

Fusion is believed to be the ultimate and inexhaustible energy source because the fuel is abundantly available and at a cheaper price. There is one atom of deuterium for every 6500 atoms of ordinary hydrogen in water. The amount of deuterium in the oceans would be enough to meet the world energy needs for about 100 billion years, which is 10 times the estimated age of our universe.

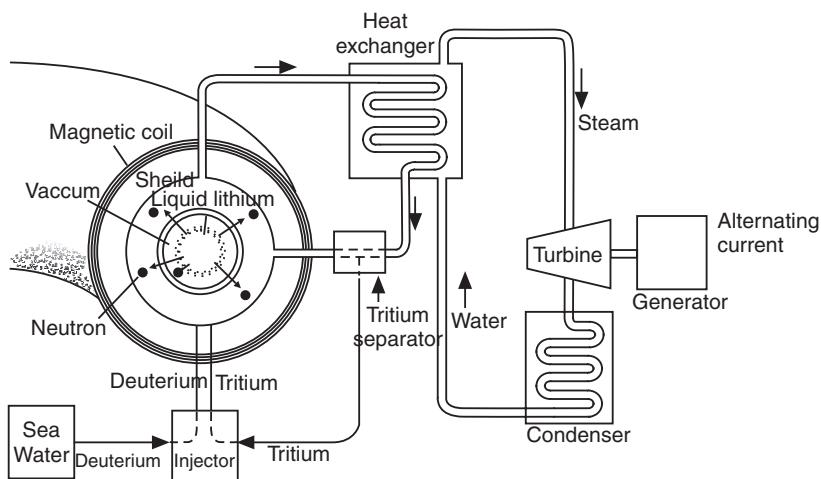


Fig. 21.20: Thermal energy conversion from nuclear fusion.

QUESTIONS

- Define binding energy of a nucleus.
- Why is $Z < N$ for the heavier nuclei?
- What do you understand by mass defect of a nuclide? Write an expression for the mass defect.
(Bombay Univ.)
- What do you understand by binding energy of a nucleus? Draw a curve showing the variation of binding energy per nucleon against the mass number. What information do you derive from such a curve?
(Shivaji Univ.)
- What is meant by binding energy of a nucleus? How does the binding energy per nucleon vary with mass numbers for different elements in the periodic table.? Explain its significance.
(Bombay Univ.)
- What are the properties of nuclear forces?
(Calicut Univ., 2007)
- Explain briefly the shell model of the nucleus.
- Explain salient features of nuclear shell model. What are magic numbers given for nucleus of an atom?
(R.G.P.V.-2008)
- Explain briefly the liquid drop model of the nucleus.
- Write a short note on the structure of the nucleus.
- Sketch the plot of binding energy per nucleon versus A and mention its salient features.
- Define activity of a radioactive substance.
- What is radioactive decay? Describe three types of decay and the law of radioactive decay with mathematical expression. Hence define and derive expression for half-life.
(C.S.V.T.U., 2005)
- What is decay constant? How is it related to the decay probability per nucleus per second?
- What is mean life of a radioactive isotope? Show that the mean life is the time for nuclei to decay to $1/e$ times their original number.
- What is radioactivity? Explain natural and artificial radioactivity.
(C.S.V.T.U., 2009)

17. What is nuclear reaction? Define Q-value of nuclear reaction. Describe controlled and uncontrolled chain reactions and criteria of critical mass. **(C.S.V.T.U., 2007)**
18. Define the Q-value of a nuclear reaction in terms of the rest mass of the constituents.
19. Define threshold energy for an endothermic nuclear reaction.
20. Explain liquid drop model of nuclear fission. What is nuclear energy? Calculate energy released in nuclear fission by mass-defect method. **(R.G.P.V., 2003)**
21. Explain nuclear fission reaction with suitable illustration. Hence explain the mechanism of nuclear fission reaction on the basis of Bohr and Wheeler's liquid drop model. **(Univ. of Pune, 2007)**
22. What is fission? What is the fundamental difference between controlled and uncontrolled fission chain reaction? How do you use controlled fission energy in a constructive way? **(Bombay Univ.)**
23. What do you understand by the multiplication factor of a nuclear chain reaction? Discuss the factors on which it depends.
24. Classify nuclear reactions.
25. What is fission? Explain with neat diagram the chain fission reaction. What are the difficulties encountered in the reaction? Suggest suitable remedy for their removal. **(Bombay Univ.)**
26. When do you say a nuclear fission reaction is (i) critical (ii) supercritical and (iii) subcritical?
27. What is chain reaction in nuclear fission? Discuss the factors on which the chain reaction depends.
28. Explain nuclear chain reaction and construction and working of a nuclear reactor. **(RGPV, 2007)**
29. Define multiplication factor. How is this related with size of fissionable material? **(Shivaji Univ.)**
30. Explain the terms: (i) Explosive chain reaction (ii) Binding energy curve. **(Shivaji Univ.)**
31. Explain the main features of the design and working of a nuclear fission reactor. **(Shivaji Univ.)**
32. Explain the function of moderator, reflector, shield, coolant and control rod for a nuclear reactor. **(Shivaji Univ.)**
33. Discuss how nuclear chain reaction is used in a constructive way in a nuclear reactor. Give details of different parts of the reactor along with a neat diagram. **(Bombay Univ.)**
34. Explain chain reaction. How can the chain reaction be controlled? **(Bombay Univ.)**
35. What are the essential components of a nuclear reactor? Describe the function of each of these components.
36. What do you mean by nuclear reactor? Give the sketch of it. What is the function of control rod? Write two factors which should be considered while selecting the site for nuclear reactor. **(C.S.V.T.U., 2009)**
37. Explain the function of moderator in a fission reactor.
38. Explain the following:
- (i) Q value
 - (ii) Critical size
 - (iii) Nuclear cross-section
 - (iv) Chain reaction
- (RGPV, 2008)**
39. Explain the function of control rods in a fission reactor.
40. Why is enriched U-235 used as fuel in a nuclear reactor?
41. Explain nuclear fusion. What is ignition temperature?
42. (i) Define nuclear fission
(ii) Define thermonuclear reaction. **(C.S.V.T.U., 2008)**
43. What do you mean by nuclear fusion? Give an account of proton-proton cycle as the cause of stellar energy. State the conditions required to initiate a self sustaining fusion reaction. **(Univ. of Pune, 2008)**

44. Explain what do you understand by nuclear fusion. Describe the necessary conditions to bring about fusion process. **(Bombay Univ.)**
45. Explain the proton-proton cycle, which is the main source of energy from sun.
46. Show that in the carbon cycle, carbon is not consumed and the net effect is the same as for the proton-proton cycle.
47. What do you mean by thermonuclear reaction? **(Shivaji Univ.)**
48. Describe proton-proton cycle in nuclear fusion. **(Shivaji Univ.)**
49. Define critical ignition temperature, confinement time and Lawson criterion. **(Shivaji Univ.)**
50. What do you mean by inertial confinement in fusion? **(Shivaji Univ.)**
51. Distinguish between the nuclear fission and fusion reactions. Explain the release of energy during fission and fusion reactions with example. **(C.S.V.T.U.,2006)**

PROBLEMS

- Calculate the binding energy and average binding energy per nucleon of $^{12}_6C$.
Given: Mass of proton = 1.007825 amu
Mass of neutron = 1.008665 amu **[Ans: 92.16 MeV; 7.68 MeV]**
- The disintegration constant λ of a radioactive element is 0.00231 per day. Calculate its half-life and average life. **[Ans: 300 days; 432.9 days]**
- The activity of certain radionuclide decreases to 15% of its original value in 10 days. Find its half-life. **[Ans: 3.65 days]**
- Compute the Q-value of the reaction $Be^9(d, n)B^{10}$. Given: Mass of $Be^9 = 9.012182$ u,
Mass of $B^{10} = 10.012938$ u, $m_d = 2.014102$ u and $m_n = 1.008665$ u. **[Ans: 4.36 MeV]**
- Determine the quantity of coal required to produce an energy equivalent to 1 gramme of U-235 if the heat of combustion of coal is 7000 k.cal/kg. **[Ans: 2.7 tons]**
- Assuming the energy released per fission of U-235 is 200 MeV, calculate the rate in kg per year at which fission should occur in a nuclear reactor in order to operate at a power level of 1 watt. **[Ans: 1.64×10^{-9} kg]**
- Fission of one atom of U-235 releases 200 MeV energy. Calculate the amount of energy released by 10^{-3} kg of uranium. **[Ans: 5.12×10^{23} MeV]**
- A nuclear reactor consumes 20.4 kg of U-235 in 1000 hrs operation. Determine the power developed by the reactor, assuming that on an average 200 MeV of energy is released per fission of a single U-235 nucleus. **[Ans: 465 MW]**
- Calculate the energy release per gram of fuel for the reaction

$${}_{1}^{2}H + {}_{1}^{2}H \rightarrow {}_{1}^{3}H + {}_{1}^{2}H$$

Given: Mass of ${}_{1}^{2}H = 1.007825$ u and mass of ${}_{1}^{3}H = 3.016090$ u. **[Ans: 6.05 MeV]**
- Calculate the energy released if 1 kg deuterium undergoes fusion. Assume the energy released per deuterium-tritium event is 17.6 MeV. **[Ans: 2.4×10^8 kWh]**

CHAPTER

22

Cosmic Rays and Elementary Particles

22.1 INTRODUCTION

Cosmic rays are highly penetrating radiations which are continually entering the earth's atmosphere in all directions from outer space. Cosmic rays consist of high energy particles. Most of the particles have energy of the order of 15 GeV.

Cosmic rays were first detected by Elster and Geitel and by C.T.R.Wilson. They found that the charge of a well insulated electroscope leaked away slowly even in the absence of any ionizing agent. At first, it was thought that the loss of charge of electroscope is due to ionizing radiations coming from radioactive substances present in the earth. In that case, the rate of discharge must decrease when the electroscope is taken to higher altitudes. Hess in 1911, recorded ionization of air at different altitudes by sending electroscope in balloons and found that the intensity of ionizing radiation increases as we go to higher altitudes. On the other hand, the intensity of ionization decreases as we go down into mines or the sea. This means that the source of the radiation is not on the earth but somewhere in space. Hess suggested that some kind of penetrating rays were entering the earth's atmosphere from outer space in all directions. These rays were named as **cosmic rays** by Millikan. Hess is credited with the discovery of cosmic rays and was awarded the Nobel Prize in physics for the year 1936.

The fundamental building blocks of matter are known as the elementary particles. Various types of elementary particles were discovered in the study of cosmic rays. With advent of sophisticated particle accelerators and detectors, more and more elementary particles were subsequently discovered. The study of elementary particles helps us understand about matter and the basic forces that hold matter together.

22.2 PRIMARY COSMIC RAYS

The cosmic rays which are just entering the earth's atmosphere are known as **primary cosmic rays**. The primary cosmic rays consist of stable high-energy particles flying in space in all possible directions. The intensity of cosmic radiation in the solar system is on the average 2 to 4 particles per square centimeter per second. It was established that the intensity of the cosmic rays near the poles of the earth is about 1.5 times greater than at its equator. Primary cosmic radiation, on account of its deflection by the earth's magnetic field, was identified as *consisting of positively charged particles*. The radiation mainly consists of protons (about 91%), α -particles (about 6.6%) and small percentages of nuclei of other elements (under 1%).

and electrons (about 1.5%). The energies of primary cosmic rays range from 1 MeV to 10^{12} MeV.

22.3 SECONDARY COSMIC RAYS

When primary cosmic rays coming out of space interact with the nuclei of atoms in the upper layers of the atmosphere, **secondary cosmic rays** are produced. Below an altitude of 20 km, all cosmic radiation is secondary. On entering the atmosphere, the primary cosmic rays collide with air nuclei and produce mostly π -mesons and some hyperons. The π -mesons carry sufficient energy and decay into lighter particles, namely μ -mesons, electrons, positrons, neutrinos and photons. All such particles constitute the secondary cosmic rays. The μ -mesons do not take part in nuclear interactions and spend their energy only on ionization. Because of this they have great penetration power. At sea level the secondary cosmic rays contain nearly 70% μ -mesons, 29% electron-positron pairs and 1% heavy particles. μ -mesons penetrate the atmosphere and are observed even deep below the surface of the earth. Mesons constitute the **hard component** of cosmic radiation and the electrons, positrons and photons constitute the **soft component**.

The π -mesons and μ -mesons in cosmic rays fly at speeds close to the speed of light and because of the relativistic time dilatation they are able to cover large distances before decaying.

22.4 ORIGIN OF COSMIC RAYS

It is, as yet, a mystery where the cosmic ray particles originate and where they are accelerated to such enormous energies. Different theories have been put forward to explain the possible sources of cosmic rays.

- i. There is enough matter in the interstellar spaces in the galaxy in the form of hydrogen clouds out of which stars evolve in the course of time due to condensation. In the process, lighter elements are converted into heavier elements. The temperature of the star rises due to more and more gravitational contraction. At the end, the star exhausts most of its hydrogen and develops so much compression that it suffers violent explosion. Such exploding stars are called Nova and Supernova. It is presumed that cosmic ray particles are ejected in the outbursts of nova and supernova and are accelerated in the non-homogeneous magnetic fields of interstellar space.
- ii. Another view is that the sun may be the source of atleast some of the cosmic rays. At the time of solar activity violent eruptions occur and ionized gases shoot out from the sun. Thus, some of the protons and α -particles in the sun are thrown out into interplanetary space. The points in favour of this view are that the cosmic ray intensity increases during solar flares. But since the cosmic ray intensity remains uniform at all hours of the day and night, the sun cannot be the source for the majority of the primary cosmic rays. It may be the source of a small fraction of the low energy primary particles.

22.5 ALTITUDE EFFECT

The intensity of cosmic rays was determined by measuring the ionization produced by cosmic rays at different altitudes. The variation of cosmic ray intensity with altitude is shown in Fig. 22.1.

The intensity of cosmic rays rises slowly up to a height of about 8 km, after which the rise becomes faster up to about 19 to 24 km. At heights above 24 km the intensity decreases gradually. The experiments were conducted at 3° , 38° 51° and 60° N and the results are similar.

The maximum of intensity is not at the top of the atmosphere but well below it. The reason is that at this height the primary cosmic rays produce quite a good number of secondary particles due to interaction with nuclei of atmospheric gases. Thus, both primary and secondary rays are present in abundance at that height. With decrease of altitude the absorption increases and therefore, intensity falls.

22.6 LATITUDE EFFECT

The variation of cosmic ray intensity with latitude is shown in Fig. 22.2. The curve shows that as one proceeds from magnetic north to magnetic south, at sea level, the cosmic ray intensity remains constant until a magnetic latitude of about 42° is reached. In this region the intensity begins to drop appreciably, reaches a minimum at the equator and rises again to symmetrical intensities in the southern Hemisphere. Thus, the cosmic ray intensity is maximum at the geomagnetic poles ($\lambda = 90^\circ$) and minimum at the geomagnetic equator ($\lambda = 0^\circ$). The intensity remains constant between 42° and 90° . This variation of cosmic ray intensity with geomagnetic latitude is called **latitude effect**.

The latitude effect may be explained on the basis of effect of earth's magnetic field on cosmic rays. The presence of such an effect indicates that the cosmic rays are charged particles. The earth's magnetic field is directed from south to North Pole. The earth's magnetic field at the equator is directed perpendicular to the direction of travel of charged cosmic particles. Therefore, it exerts maximum force upon the particles which consequently suffer maximum deflection. They are deflected away from the earth. The intensity of cosmic rays is therefore a minimum at the equator. At poles, cosmic particles move parallel to the earth's magnetic field. They suffer minimum deflection. Hence, the intensity of cosmic rays is a maximum at the poles.

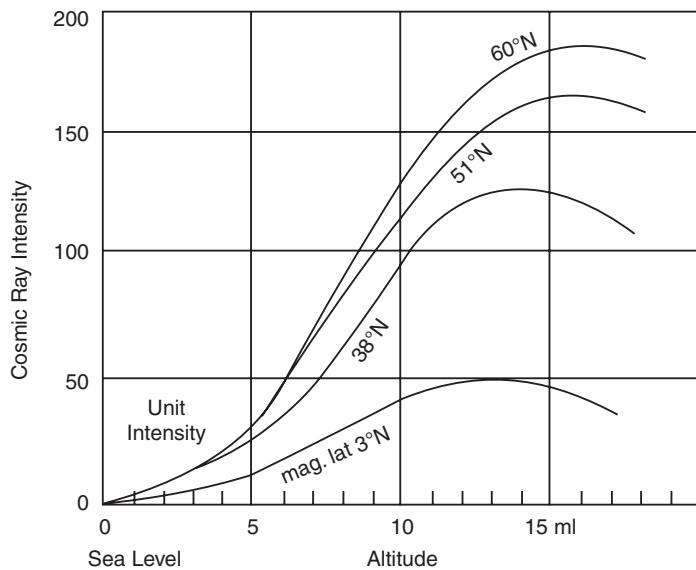


Fig. 22.1

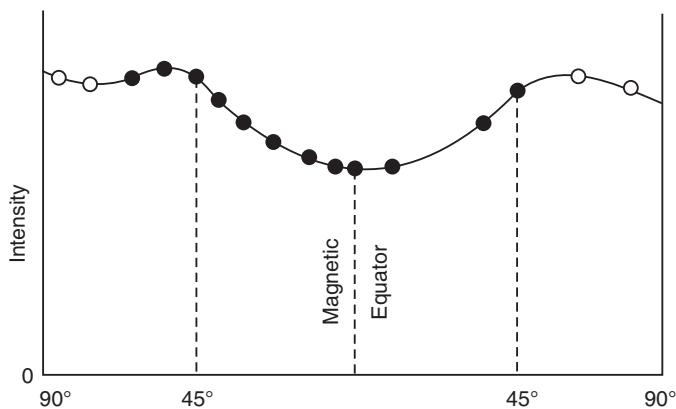


Fig. 22.2

22.7 LONGITUDE EFFECT

The intensity of cosmic rays also depends upon the longitude of the point of observation. It is called **longitude effect**. The intensity of cosmic radiation along the equator varies at different longitudes. The equatorial variation is attributed to the fact that the earth's magnetic field is not symmetrical about its axis.

22.8 EAST-WEST EFFECT

It is found that the number of cosmic ray particles coming from the west direction is greater than those coming from the east direction. This effect is called **east-west effect** and is a maximum at the equator. At the equator, the number of particles coming from west is 14% more than the particles coming from east. This phenomenon gave evidence to the fact that cosmic rays are composed predominantly of positively charged particles.

The charged particles approaching the earth's atmosphere are deflected by the earth's magnetic field in a direction perpendicular to the magnetic field and to the direction of their motion. The positively charged particles are deflected towards the east by earth's magnetic field and thus appear to come from the west of the vertical. Thus at any azimuth, more particles approach the earth from the west than from the east.

22.9 THE POSITRON

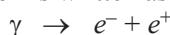
In 1928 the British physicist Paul A.M. Dirac predicted the existence of a particle similar to the electron but carrying a positive charge. These particles are named as *positrons*.

In 1932 Anderson discovered positron by photographing the tracks of cosmic rays in a Wilson cloud chamber. If a strong magnetic field is applied perpendicular to the face of the cloud chamber, positively charged particles should bend to the right and negatively charged particles bend to the left. Anderson set up a strong magnetic field in Wilson cloud chamber and observed tracks of small curvature which could be attributed to a high-energy charged particle. Anderson inserted a block of lead in the chamber to slow down the particles. Passing through the plate the particle would lose its speed and move in a more curved path. On one of the photographs he again observed the track of this particle. Knowing the direction of motion, the direction of the field, and the direction of bending, Anderson concluded that such a particle had a positive charge. Calculations proved the mass and the magnitude of the charge of the new particle to be identical to those of the electron.

22.10 PAIR PRODUCTION

An extension of the quantum theory of the electron proposed by P. Dirac led scientists to the prediction that if a high energy photon (γ -ray) were to come close enough to the nucleus of an atom, the electric field of the nucleus would annihilate the γ -ray and create in its place a pair of particles, an electron and a positron. According to the theory, these two particles should have the same mass and equal but opposite charges. Blackett, Anderson and others discovered such pairs in a cloud chamber.

The reaction for pair production is written as



If $h\nu$ is the energy of the γ -ray, then

$$h\nu = 2m_0c^2 + E^- + E^+$$

where m_0c^2 is the rest energy of electron (or positron), E^- is the kinetic energy of electron and E^+ is the kinetic energy of positron. Pair production is materialization of radiant energy. From

in the above equation, it is clear that the *threshold energy* for electron-positron pair production is $2m_0c^2$.

The converse of materialization of energy is the annihilation of matter. When a positron combines with an electron, both the particles disappear and in their place two γ -rays are produced. Thus,



This process is called **annihilation of matter**.

Positron is known as an **antiparticle** of electron.

22.11 COSMIC RAY SHOWERS

During the study of cosmic rays with the help of Wilson cloud chamber, the experimenter occasionally finds a picture of a group of tracks, instead of the usual one or two tracks. The group of tracks is termed as *cosmic ray showers*. An extensive study of showers has led to the conclusion that each shower is produced by a single, high-energy cosmic ray. Rossi investigated the phenomenon extensively and the results were explained by Homi J. Bhabha and Heitler on the basis of cascade theory.

According to them, a high energy electron

(or positron) present in the cosmic rays loses energy when it encounters the atomic nuclei in the earth's atmosphere. The energy appears as high energy photon. The photon interacts with the electric field of an atomic nucleus and is completely absorbed resulting in the production of electron-positron pair. The energy required for pair production is more than 1 MeV. The electron and positron so produced have sufficient energy to produce more photons on interaction with nuclei. These photons are further capable of bringing about pair production. The result is the generation of a large number of photons, electrons and positrons having a common origin (Fig. 22.3). The multiplication continues until the initial energy becomes divided between a large number of pairs and the individual energies of the particles fall below the critical energy when photon emission and pair production can no longer occur.

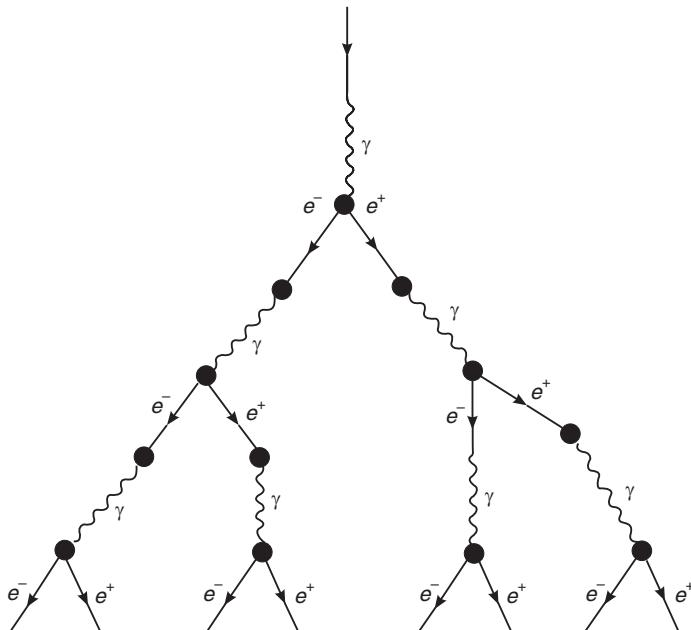


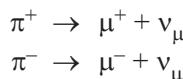
Fig. 22.3

22.12 THE MESONS

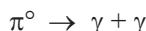
In 1938 Anderson and Nedermeyer discovered the presence, in cosmic rays, of charged particles having a mass greater than that of an electron but considerably smaller than that of proton. These particles are called *mesons*. They are produced high in the atmosphere by

the collisions of cosmic rays with atomic nuclei. When a high energy proton collides with a nucleus it knocks out protons, neutrons and **π -mesons** or **pions**. Some pions are positively charged (π^+), some are negatively charged (π^-) while others are neutral (π^0). The π^+ and π^- are antiparticles to each other. They have a rest mass of $273\ m_e$. The rest mass of π^0 meson is $264\ m_e$. All charged π -mesons are unstable and have a half-life of 2×10^{-8} s and decay into a charged μ particle and a lightweight neutral particle called a *neutrino*.

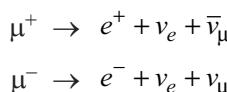
The decay of pions may be written as interactions as follows.



The neutral pion has a half-life of 8.7×10^{-17} s and decays into γ -rays.



The μ -mesons are known as *muons*. μ -mesons are unstable and have a half-life of 2×10^{-6} s and decay into electrons and neutrinos. The decay of the muons may be written as



The particle ν_μ is known as muonic neutrino and ν_e is the electronic neutrino. The particles $\bar{\nu}_\mu$ and $\bar{\nu}_e$ are the antiparticles corresponding to the muonic neutrino and the electronic neutrino respectively. The properties of the muonic neutrino and the corresponding antineutrino are quite similar to those of the electronic neutrino and the corresponding antineutrino, but experiments proved them to be different particles.

A wonderful property of the muon is its absolute similarity with the electron in everything but its mass; the muon is 207 times heavier than the electron. It can also temporarily occupy the place of the electron in an atom, with an orbit very close to the nucleus. Such an atom is known as *mesoatom*.

22.13 ELEMENTARY PARTICLES

The fundamental building blocks of matter are known as the **elementary particles**. Molecules are built up from the *atom*, which is the basic unit of any chemical. Early in 1930, atoms were found to consist of the *electron* and *proton*. In 1932, the *neutron* was discovered and these three particles, namely the electron, proton and neutron were recognized to be the constituents of the atom. Hence, they were thought to be the elementary particles. Subsequently, the positron and neutrino were discovered. Various types of mesons were discovered in the study of cosmic rays. With the advent of sophisticated particle accelerators and detectors, more and more elementary particles were discovered.

22.14 CLASSIFICATION OF ELEMENTARY PARTICLES

Several systems classify the elementary particles on the basis of their various properties. Particles having their half-life of the order of 10^{-16} s are known as **stable particles** and those having half-life of the order of 10^{-22} s are known as **resonances**. One of the classifications of stable elementary particles is based on the basic nuclear force interactions.

2.15 BASIC FORCES IN NATURE

Four basic forces are recognized which are responsible for interaction of elementary particles.

They are

1. Weak nuclear interaction
2. Strong nuclear interaction
3. Electromagnetic interaction and
4. Gravitational interaction

Each force is carried by an elementary particle. The electromagnetic force, for instance, is mediated by the photon, the basic quantum of electromagnetic radiation. The strong force is mediated by the **gluon**, the weak force by the W and Z particles, and gravity is thought to be mediated by the graviton.

Gravitational interaction is the weakest nuclear force. The particle that mediates the gravitational interaction is called the **graviton**. It has not been detected so far but is assumed to be massless and possess a spin of 2. It travels at the speed of light and interacts with all particles that have mass.

Electromagnetic interaction operates between charged particles and particles having electric and magnetic moments. It is responsible for the binding of electrons in the atoms, for formation of molecules, and for all chemical and biological processes. Almost all non-gravitational interactions are due to electromagnetic interaction. The particle that mediates the electromagnetic interaction is the **photon**. Photon is mass-less and has a spin of 1.

Strong nuclear interaction operates between the nucleons in a nucleus. It is the strongest of all the forces. It keeps a number of protons together in a nucleus against the strong Coulomb force of repulsion. It is a short-range force. The nucleons and mesons are part of a general group of particles called **quarks**. The particle that mediates the strong interaction between quarks is called a **gluon**. Its mass is zero and has a spin 1. A free quark or a gluon is not observed since they remain hidden within particles confined by the strong force.

Weak nuclear interaction is also a short-range force operating up to a distance of 0.001 fm. It is responsible for the decay of some heavier particles.

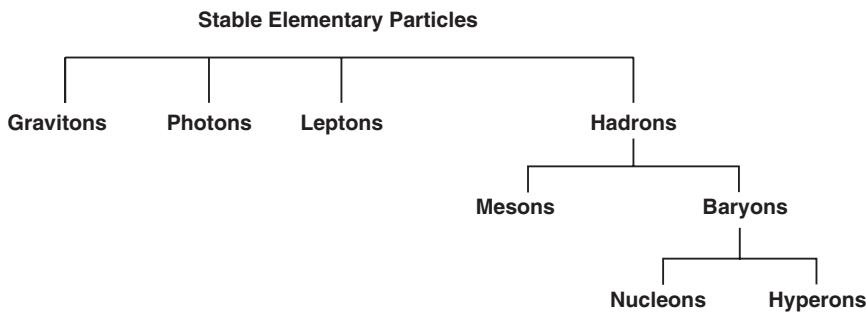
The electromagnetic and weak interactions are unified in the electroweak theory. According to this theory, the particle responsible for the weak interaction is called the **W particle**. There are three types of the W particle: W^+ , W^- and Z° . All these particles have spin one.

22.16 CLASSIFICATION OF ELEMENTARY PARTICLES BASING ON THE BASIC FORCES

The stable elementary particles are divided into four groups basing on the above basic forces.

1. **Gravitons:** They mediate the gravitational interaction and are yet to be discovered. They have spin 2 and therefore, they are **bosons**.
2. **Photons:** They mediate the electromagnetic interaction and obey Bose-Einstein statistics. Therefore, they are also **bosons**.
3. **Leptons:** Particles that interact via the weaker nuclear force but not the strong force are called **leptons** ("light ones"). The lepton family includes the electron, muon, and the neutrinos. They obey Fermi-Dirac statistics. Hence they are called **fermions**.
4. **Hadrons:** Particles that interact via the strong nuclear force are called **hadrons**. These include the proton, neutron, and pion. These are also **bosons**. Hadrons are further subdivided into two groups, namely **mesons** and **baryons**.

The classification may be depicted as follows in a chart.



22.17 ANTIPARTICLES

An *antiparticle* is a particle having the same mass, spin and lifetime, but its charge having opposite sign to that of the particle. For example, positron e^+ is an antiparticle to electron e^- .

22.18 LEPTONS

There are only six leptons and six of their antiparticles. They are

- (i) electron e^- and its antiparticle positron e^+
- (ii) negative muon μ^- and its antiparticle positive muon μ^+
- (iii) tau τ^- and its antiparticle τ^+ .
- (iv) electron-neutrino and its antiparticle
- (v) muon neutrino and its antiparticle
- (vi) tau neutrino and its antiparticle

The electron, muon and tau are negatively charged particles and each of them has an associated neutrino.

(i) Electron and positron

The most familiar lepton is the **electron**. It is a stable particle and the only lepton that exists naturally in atoms. It has spin $\frac{1}{2}$ and obeys Fermi-Dirac statistics. Therefore, it is a **fermion**.

There is no evidence of any internal structure, at least down to the measurement limit of 10^{-17} m. Thus, at present the electron is considered to be a point particle.

Positron is the antiparticle of the electron. It is a positively charged particle having the charge, mass, and spin equal to that of electron and it is a fermion.

(ii) Muons

Muons were first observed in cosmic rays. Muon is electrically charged and 207 times more massive than an electron. Its spin is $\frac{1}{2}$ and it is a fermion. Muons are unstable and emit electrons. Since muons also appear not to have any internal structure, they are sometimes called *heavy electrons*. Muons are unstable and decay in about 2.22×10^{-6} s, according to the following schemes:

$$\begin{aligned}\mu^- &\rightarrow e^- + \nu_e + \nu_\mu \\ \mu^+ &\rightarrow e^+ + \nu_e + \nu_\mu\end{aligned}$$

Thus, for each muon decay, the neutrino and antineutrino particles are of two different types: one is electron-neutrino and the other is muon-neutrino.

(iii) Tauon

A third charged lepton is known as a tau (τ^-) particle, or **tauon**. It is 3000 times heavier than the electron.

The electron, muon, and tauon are negatively charged. They have no apparent internal structure, and have positively charged antiparticles. The remaining leptons are neutrinos, which are electrically neutral.

(iv) Neutrinos and antineutrinos

Neutrinos are present in cosmic rays and are emitted by the Sun and by some radioactive decays. Neutrinos have very little or no mass and thus travel at the speed of light. They are not influenced either by the electromagnetic force or the strong force and so pass through matter as if it were not there. They are electrically neutral and stable particles with spin $\frac{1}{2}$.

There are three types of neutrinos, each associated with a different charged lepton $\begin{pmatrix} + & + & + \\ - & - & - \\ e, \mu, \tau \end{pmatrix}$. They are known as the electron neutrino (ν_e), the muon neutrino (ν_μ) and the tau neutrino (ν_τ). There is an antineutrino for each of these, for a total six different neutrinos. Thus, there are twelve different leptons in all including six leptons and their antiparticles. The neutrino and antineutrino pair annihilates to give 2 γ -ray photons. For example,

$$\nu_e + \bar{\nu}_e \rightarrow 2\gamma \text{ or } \nu_\mu + \bar{\nu}_\mu \rightarrow 2\gamma$$

22.19 HADRONS

Hadrons are said to be strongly interacting particles. They interact via the strong force, the weak force, and gravity. The electrically charged members can also interact by the electromagnetic force. The hadrons are divided into two subgroups - *baryons* and *mesons*.

Baryons

Baryons have masses at least as large as the proton mass. The proton and neutron also belong to this group. They have half- integer spins $\left(\frac{1}{2} \text{ or } \frac{3}{2}\right)$. Except for the stable proton, baryons decay into products that eventually include a proton.

Mesons

Mesons, which include pions, have integer spin values (0 or 1) and eventually decay into leptons and photons. For example the neutral pion decays into two gamma rays. Their masses are greater than the muon mass and have fairly short lifetime. They strongly interact with matter.

Hyperons

Hyperons are unstable particles having masses greater than that of nucleon. They were found in cosmic rays. Three groups of hyperons, called lambda (Λ), sigma (Σ) and Xi (Ξ) are now

known. The hyperons have a half-life of the order of 10^{-10} s and generally decay by several alternative modes to nucleons and pions.

22.20 RESONANCES

The term stable particle is applied to those particles which have half-lives far greater than the time required for light to travel a distance equal to the diameter of the elementary particle. The diameter is taken to be around 10^{-15} m and the time required to travel this distance is of the order of 10^{-23} s. Decays caused by weak forces have lifetimes of the order of 10^{-10} s. In the meson and baryon classes, a number of particles are there which cannot be detected directly because their lifetimes are so short that they cannot be detected before decaying. These particles are known as **resonances**, or **resonance states**, because of an analogy between their manner of creation and the resonance of an electrical circuit.

The first such particle was observed by Fermi when he bombarded protons with a beam of π^+ pions. The graph drawn between the number of interactions and the kinetic energy of pions showed a large peak around 200 MeV. Fermi concluded that the proton and pion combined momentarily to form a short-lived particle before separating apart again. Subsequently, a large number of resonance particles were discovered.

22.21 THE QUARK MODEL

The large numbers of hadrons suggests that they may be composites of other truly elementary particles. In 1963 Murray Gell – Mann and George Zweig put forth the *quark theory*. They proposed that hadrons are not elementary particles in the sense that they are not fundamental building blocks, but are composite particles

composed of truly elementary (fundamental) particles. They named these particles quarks. Originally quark model consisted of three different quarks (with fractional charges) and their antiquarks (antiquarks). These quarks were named the up quark (u), the down quark (d), and the strange quark (s). By combining three quarks, the relatively heavy hadrons, the baryons could be built. Quark-antiquark pairs could account for the lighter hadrons, the mesons.

Quarks are the fundamental particles of the hadron family. Since several quarks have to combine to give the charge on the hadron, it was clear that they have fractions of an electron charge, e . Thus the theory proposed that u , d , and s quarks have charges of $+\frac{2}{3}e$, $-\frac{1}{3}e$ and $-\frac{1}{3}e$, respectively. The antiquarks, designated by overbars, such as \bar{u} , have opposite charges (for example, the charge on the \bar{u} is $-\frac{2}{3}e$). Thus three quark combinations could produce any baryons. A proton, for instance, consists of two up quarks and a down quark, and a neutron consists of two down quarks and an up quark. Thus, the quark composition of the proton and neutron would be uud and udd , respectively (see Fig. 22.4).

A meson is made up of one quark and one antiquark. For example, the π^+ meson is the combination of u quark and u antiquark ($u\bar{d}$).

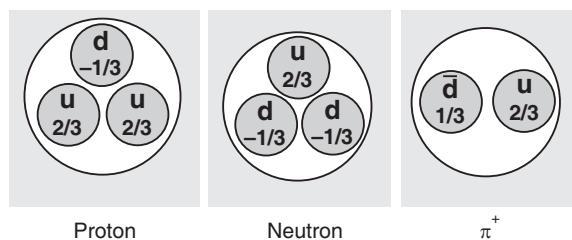


Fig. 22.4

The discovery of new elementary particles in the 1970s led to the addition of three more quark types: charm (c); top, or truth (t); and bottom or beauty (b). Today there is firm experimental evidence of the existence of all six quarks and their antiparticles.

According to the quark model, the truly elementary particles may be grouped as: leptons, quarks, photons and gravitons.

22.22 OTHER MODELS

The behavior of all known subatomic particles can be described within a single theoretical framework called the Standard Model. This model incorporates the quarks and leptons as well as their interactions through the strong, weak and electromagnetic forces. Only gravity remains outside the Standard Model. Each quark is assumed to carry a **colour charge** of the strong nuclear force, which is analogous to the electric charge in electromagnetism; antiquarks similarly carry anticolor. There are three colors for each quark-red (R), green (G) and blue (B). Color charged particles interact via **gluon** exchange in the same way that charged particles interact via photon exchange. However, gluons are themselves color charged, resulting in an amplification of the strong force as color charged particles are separated. Unlike the electromagnetic force which diminishes as charged particles separate, color charged particles feel increasing force; effectively, they can never separate from one another.

According to another theory, known as **string theory**, each kind of fundamental particle corresponds to a different pattern of fundamental string. All strings are essentially the same, although they may be open (lines) or closed (loops). Different particles differ in the coordination of their strings. One particular prediction of string theory is the existence of extremely massive counterparts of ordinary particles due to vibrational excitations of the fundamental string. Another important prediction of string theory is the existence of a mass-less spin-2 particle behaving like the *graviton*.

QUESTIONS

1. Why is a free quark not observed?
2. What are cosmic rays? How are they affected with latitude and altitude?
3. What are primary and secondary cosmic rays?
4. Given an account of the production and properties of π and μ mesons.
5. What are cosmic showers?
6. What are fundamental interactions in nature? Explain their relative range and strength.
7. Briefly explain the mediating particles in the different fundamental interactions.
8. Write a brief note on the elementary particles as identified today.
9. What are leptons? Name them.
10. Discuss the properties of leptons in brief.
11. Compare the properties of leptons and baryons.
12. What is a resonance particle?
13. Briefly explain the quark model.

CHAPTER

23

Nuclear Instruments

23.1 INTRODUCTION

Radioactive decay and nuclear reactions are accompanied by the emission of charged particles like α -particles, β -particles, protons and γ -rays. Our senses cannot detect these products directly, and detection must be done by indirect means. More immediate and quantitative methods of detection are desirable and a variety of instruments have been developed for this purpose. As a result of studies of radioactive transmutations, it was found that there are atomic nuclei with equal atomic numbers but with different mass numbers. Such nuclei are called isotopes. Isotopes frequently appear in radioactive transmutations. Radium, radon, uranium, thorium have several isotopes each, and so on. Mass spectrometers are utilized to analyze the isotopes. The mass spectrometer helped in the discovery of isotopes of uranium. Investigations on nuclear reactions require charged particles accelerated to very high energies. The high-energy particles are used in producing radio-active isotopes required in therapy and agriculture. They are also used in discovering new elementary particles. Particle accelerators are used to accelerate the charged particles. Some of the common radiation detectors, mass spectrometers and particle accelerators are described in this chapter.

23.2 GEIGER-MULLER COUNTER

Nuclear radiations emitted by disintegrating nuclei cannot be sensed directly. Indirect methods are to be employed to detect them. Alpha, beta and gamma rays have the ability to ionize neutral atoms. This property is used in radiation detecting instruments.

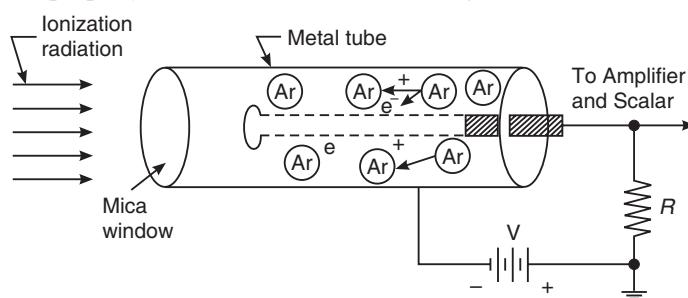


Fig. 23.1. A Schematic of Geiger -Muller counter. Radiation ionizes the argon gas molecules in the tube giving rise to electrical pulse, which is counted.

The Geiger-Muller counter is a radiation detector. It is a modified cathode ray tube with electrical circuits needed to amplify the current and detect it. The Geiger-Muller (G.M.) tube (Fig. 23.1) consists of a rugged metal case enclosed in a thin glass tube. The hollow metal case acts as cathode. A fine wire, usually of tungsten, runs through the center of the tube and is insulated from the metal. It acts as anode. The tube is evacuated and then partially filled with a mixture of 90% argon at 10 cm pressure and 10% ethyl alcohol vapour at 1 cm pressure. At one end of the tube a thin window of mica is arranged to allow the entry of radiation into the tube.

A dc potential of about 1200 volts is applied between the cathode and the wire. The value of the voltage is adjusted to be somewhat below the breakdown voltage of the gaseous mixture. A high resistance R is connected in series with battery.

A high energy particle entering through the mica window will cause one or more of the argon atoms to ionize. The electrons and ions of argon thus produced cause other argon atoms to ionize in a cascade effect. The result of this one event is a sudden, massive electrical discharge that causes a current pulse. The current through R produces a voltage pulse of the order of $10 \mu\text{V}$. An electronic pulse amplifier accepts the small pulse voltages and amplifies them to about 5 to 50 volts. The amplified output is then applied to a counter. As each incoming particle produces a pulse, the number of incoming particles can be counted.

The number of secondary electrons is independent of the number of the primary ions produced by incoming particle. The incoming particle acts as a trigger to release an avalanche of secondary electrons. The electrons reach the anode and cause ionization current in the circuit, whereas the positive ions move slowly and form a sheath around the anode for a short time. They reduce the potential difference to such a low value that the current in the circuit is stopped. Therefore, a brief pulse of current is produced by each incoming particle.

Fig. 23.2 shows a plot of counts per minute as a function of voltage. For voltages less than 1000 volts there is no discharge and hence no counts. Between 1000 to 1200 volts the number of pulses increases with the applied voltage almost linearly. Above 1200 volts, the number of counts remains constant over a certain region known as *plateau*. In this region, the magnitude of pulses becomes independent of the amount of original ionization. This plateau region is used for G.M. counter operation. If the voltage is increased above this region, a continuous discharge will take place, which is undesirable and is hence avoided.

Quenching

When the positive ions reach the cathode, they dislodge secondary electrons from the cathode because of their high energies. These electrons move toward anode and produce unwanted avalanches. As a result the counter goes into a state of continuous avalanching. During the measurements, the counter fails to distinguish between the two types of pulses, one that is due to an incoming particle and the other due to unwanted avalanching. The process of preventing the undesirable continuous avalanching is known as *quenching*. In other words, quenching is the elimination of sheath of positive ions around the cathode.

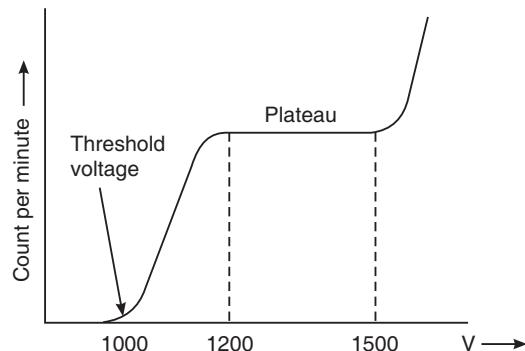


Fig. 23.2. A plot of potential difference applied across the electrodes in a G.M. counter versus count rate. G.M. counter is operated in the plateau region.

Self-quenching

To cause internal automatic quenching, a small percentage of ethyl alcohol vapour is added to the argon gas in the tube which prevents undesirable continuous avalanching.

Counting rate

The G.M. counter can count about 5000 particles per second. The counting rate depends upon the dead time and recovery time of the GM counter.

Dead time

Dead time refers to the time taken by the tube to recover between counts. In the counter, the slowly moving positive ions take about 100 μs to reach the cathode. If a second particle enters the tube during this time, it will not be registered, as the potential difference across the electrodes is very low. Hence, the time interval is known as the dead time.

Recovery time

After dead time, the tube takes approximately 100 μs before it regains the original working conditions. This time interval is known as recovery time. Thus, *recovery time* is the time after which the original pulse levels are restored.

Paralysis time

The sum of dead time and recovery time is known as paralysis time, which is 200 μs . The tube can respond to the second incoming particle only after 200 μs .

Applications

The G.M. counter is very useful for detecting nuclear radiations and charged particles. It is largely used for recording cosmic ray events and measuring cosmic ray intensities.

Limitations

1. G.M. counter has a large ‘dead time’ and recovery time of the order of 200 μs . If a large number of particles enter the G.M. tube at a rapid rate, the tube will not have time to recover and some particles may not be counted.
2. It has very efficiency for detection of γ -radiation.
3. It cannot detect neutral particles.
4. It cannot provide information regarding the nature of the ionizing particle.
5. It has limited lifetime.

Example 23.1. A G.M. counter wire collects 10^8 electrons per discharge. When the counting rate is 500 counts/min, what will be the average current in the circuit?

Solution: Number of electrons collected in one minute, $n = 10^8 \times 500$

$$\text{Charge per minute, } Q = n e = (10^8 \times 500) (1.602 \times 10^{-19} \text{ C}) / \text{min.}$$

$$\text{Average current, } I = \frac{Q}{60 \text{ s}} = \frac{(10^8 \times 500) (1.602 \times 10^{-19} \text{ C})}{60 \text{ s}} = 1.3 \times 10^{-10} \text{ A.}$$

23.3 THE WILSON CLOUD CHAMBER

In 1911, C.T.R. Wilson devised an instrument known as cloud chamber using which, charged particles can be detected and their paths can be photographed. It was extensively used in the study of cosmic rays and played an important role in the discovery of new elementary particles such as positron, meson etc.

Principle:

The working of the cloud chamber is based on the fact that supercooled vapour in a chamber condenses when dust particles or ions are present in the chamber. In case of completely dust-free and ion-free supercooled vapour, condensation does not take place. But

if ions are available in the chamber, they serve as nuclei for condensation. If saturated vapour in the chamber is suddenly subjected to an adiabatic expansion, there will be an increase in volume which produces cooling rendering the saturated vapour to a supersaturated unstable state. If an ionizing particle, such as an α -particle or β -particle passes through the chamber, ions are produced along its path and droplets condense on these ions. Consequently, the *track* of the particle becomes visible.

Construction:

The apparatus (Fig. 23.3) consists of an air-tight cylindrical chamber A, with walls and ceiling made of glass. It is provided with a movable piston P. The chamber contains a dust-free mixture of alcohol vapour and air. A small amount of water and alcohol is kept in a trough at the bottom of the chamber to ensure that the vapour is in saturated state. When the piston moves down rapidly, adiabatic expansion of the vapour inside the chamber takes place. The wooden blocks WW reduce the air space inside the piston. The piston is connected to a large evacuated vessel F through a valve. The chamber is illuminated by a mercury vapour lamp L. Two cameras CC are used to photograph the tracks.

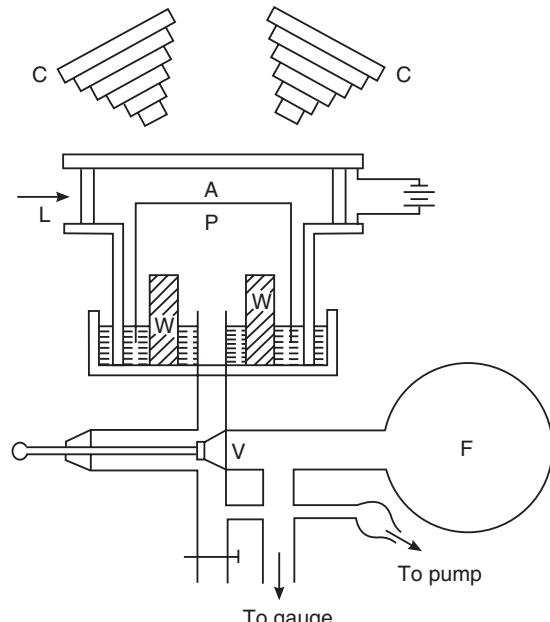


Fig. 23.3

Working:

Initially, the chamber A is filled with the saturated vapour. When the valve is opened, the air under the piston rushes into the evacuated vessel F, and causes the piston to drop suddenly. As a result, the gaseous mixture in the expansion chamber is subjected to adiabatic expansion and is supercooled. When ionizing particles pass through, they ionize the gas. Negative and positive ions are formed all along the path of the ionizing particle. A large number of extremely fine droplets are formed on the newly produced ions. These droplets form a **track** of the moving ionizing particles. The droplets are clearly visible when the chamber is illuminated by light. The tracks can be photographed with the help of cameras.

In fact, the operation of the cloud chamber is automatic. The process of expansion, entry of the ionizing particles into the expansion chamber, illuminating the chamber and clicking the camera are all carried out in rapid succession automatically. Immediately after photographing of the track, the particle track is removed by applying an electric field across the air gap. Then, the chamber gets ready for taking the next photograph.

Different particles produce different types of tracks. The ionizing particle can be easily identified from its path in the cloud chamber. α -particles, being highly ionizing, produce short, broad, thick, and straight line tracks. β -particles being very light and less ionizing, produce thin, beaded and crooked tracks. As the ionizing power of γ -rays is very low, they are never observed in a cloud chamber.

If the cloud chamber is placed between the pole pieces of a strong electromagnet, the paths of the charged particles are deflected. From the direction of curvature, the positive

and negative charged particles can be distinguished. The momentum of the particle can be estimated from the measurement of the radius of the curved track as follows.

The radius R , of the curved path of a charged particle moving in a transverse uniform magnetic field is given by

$$R = \frac{mv}{qB}$$

$$\therefore \text{The momentum} \quad p = mv = qBR. \quad (23.1)$$

23.4 BUBBLE CHAMBER

The bubble chamber was invented by Donald Glaser in 1952.

Principle:

The principle of bubble chamber is based on the property of superheated liquid. Normally, a liquid boils with the evolution of bubbles of vapour at its boiling point. If the liquid is heated above its normal boiling point by increasing the pressure over it, it is known as a *superheated liquid*. When the pressure on the superheated liquid is suddenly reduced to atmospheric pressure, the liquid does not start boiling immediately but remains in its unstable superheated state for some time. This liquid is very sensitive to the passage of charged particles, which initiate boiling. As a result of the energy they deposit by ionizing the atoms as they force their way through the liquid. The newly created ions act as centers of condensation and form a track of vapour bubbles.

Construction:

A simplified diagram of a bubble chamber is shown in Fig. 23.4. A box with glass walls is filled with a liquid such as liquid hydrogen, neon or a mixture of these. The box is connected to a pressure system. The ionizing rays enter the liquid through a thin window. The box can be illuminated by a flash of flood lights. A camera is attached to photograph the track of the particles.

Working:

The liquid, a roughly 2:1 neon-hydrogen mix, is prepared and held under a pressure of about 5 atmospheres. Just before the ionizing particles arrive, the pressure is reduced to about 2 atmospheres making the liquid superheated. As charged beam particles pass through the liquid they deposit energy by ionizing atoms and this causes the liquid to boil along their paths. The bubbles formed are allowed to grow for a few milliseconds, and when they have reached a diameter of about 1 mm, a flash photograph is taken. The pressure is then increased again to clear the bubbles and await the arrival of the next burst of particles.

Advantages: The bubble chamber has the following advantages over the cloud chamber.

- (i) The tracks obtained in a bubble chamber are sharper.
- (ii) Highly energetic particles can be stopped within the liquid and hence their ranges can be measured.
- (iii) Particles of low ionizing ability produce good tracks in bubble chamber.

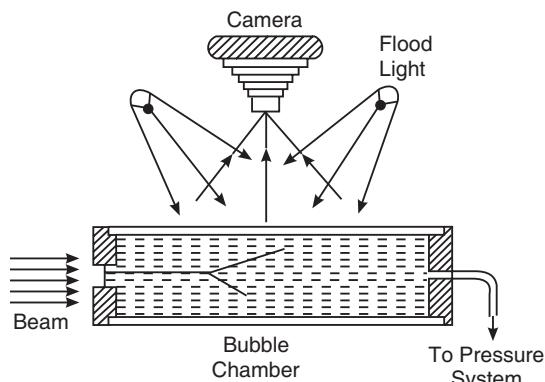


Fig. 23.4

23.5 SPARK CHAMBER

Spark chamber is one of the important detectors developed for use in the study of high-energy nuclear reactions.

Construction:

A spark chamber consists of a stack of equally spaced conducting plates (Fig. 23.5). Typically, there are about 25 to 100 plates, each about 1 mm thick and 1 m square. They are accurately spaced about 6 mm apart. Alternate plates are connected together and the two sets are connected to a 10 kV to 15 kV d.c. voltage source. The chamber is filled with pure helium or a mixture of 90% neon and 10% helium at a pressure of 1 atmosphere.

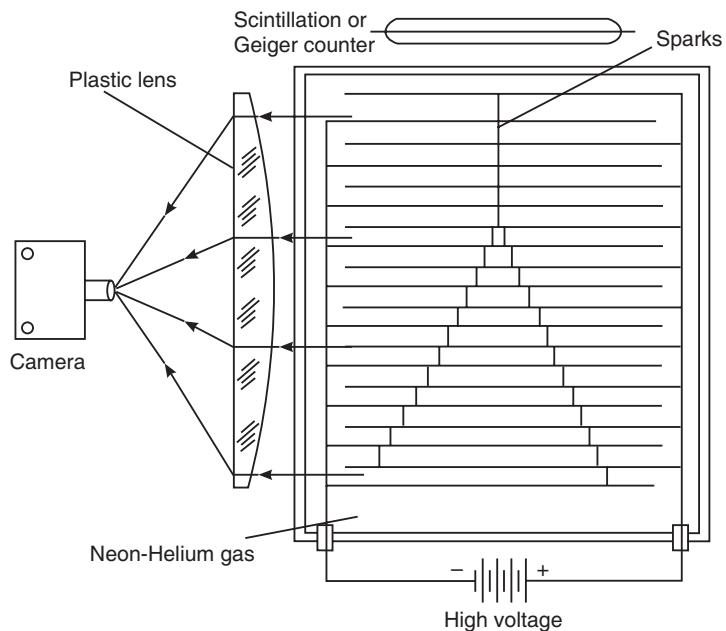


Fig. 23.5

Working:

When a high-energy charged particle enters the spark chamber, a detector triggers the high voltage to the plates and opens the camera shutter. Traversing the chamber the charged particle produces many ion pairs along its path and sparks jump between the pairs of oppositely charged plates. A large plastic lens makes it possible to photograph the light from sparks between the plates. A second large lens and camera looking into an adjacent side of the chamber permits the simultaneous recording of a second photograph, thereby recording stereoscopic photos of each event. A series of sparks marks the track of the particle.

23.6 SCINTILLATION COUNTER

Fig. 23.6 shows a schematic of a scintillation counter. It is a combination of a scintillator and a photomultiplier tube. The passage of a charged particle through the scintillating material causes some of the scintillator atoms to get excited. As the electrons in the excited atoms return to their ground states, photons are emitted. The photons strike the semitransparent photocathode in the photomultiplier tube and cause emission of electrons through

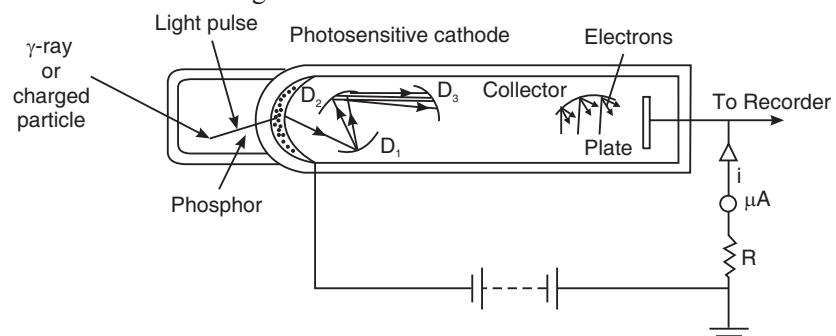


Fig. 23.6

photoelectric effect. Ideally each photon would produce one electron at the photocathode. In practice the efficiency of electron production is about 10%. Electrons produced at the photocathode are accelerated towards the first *dynode* D_1 which is held positive with respect to the cathode. The number of electrons emitted at the photocathode will be proportional to the number of photons incident on it and therefore, to the energy of the original incoming particle. The dynode D_1 has a specially prepared surface favourable for secondary emission. In the process of secondary emission, an electron hitting the surface results in ejection of more than one electron from the surface. The electrons from D_1 are accelerated towards the second dynode D_2 where secondary emission again occurs. A photomultiplier tube contains ten to eleven dynodes. Therefore, a large multiplication of the electrons takes place. Therefore, a single electron emitted by the photocathode can result in a detectable current pulse at the last dynode which is usually called the *collector*.

Scintillation counters are capable of detecting particles whose times of arrival are separated by less than a microsecond. They are therefore much faster. Secondly, the voltage pulse produced by the photomultiplier tube has amplitude proportional to the energy of the incident particle and therefore, it is possible to determine the energies of incoming particles.

23.7 SOLID STATE DETECTORS

Fig. 23.7 shows a semiconductor diode used as a particle detector. The detector consists of a p-n junction formed between p-type and n-type silicon. When a p-n junction is made, a depletion region arises between the p- and n-type semiconductor regions. The depletion region does not contain any mobile charges. The p-n junction is reverse biased with the help of battery B. The reverse bias causes a widening of the depletion region. It also minimizes the current flowing in the detector when no radiation is incident on it.

When a charged particle travels the depletion region, it interacts with the atoms in the depletion region and produces electron-hole pairs. The electric field acting across the depletion region separates the electrons and holes which move through the external circuit. As the electrons and holes are oppositely charged, they cause current in the same direction in the external circuit. As they pass through the resistor R, they produce a voltage pulse. The output pulse is then amplified. The strength of the output pulse depends on the number of carriers collected.

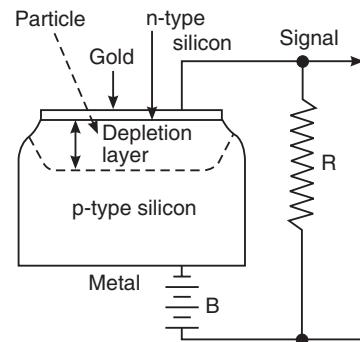


Fig. 23.7

23.8 CERENKOV DETECTOR

When a charged particle moves through a transparent dielectric medium with a velocity greater than the velocity of light in that medium, a cone of light waves is emitted. The light waves are known as Cerenkov radiation. In a dielectric medium of refractive index ν , photons move with a velocity c / μ . Let us consider a charged particle moving with a velocity v through the dielectric medium. Fig. 23.8 (a) shows the interactions when $v > c / \nu$. The envelope of the radiation is a cone of half-angle θ with the particle at its apex. θ is given by the following relation.

$$\sin \theta = \frac{(c / \mu)t}{vt} = \frac{c}{\mu v} \quad (23.2)$$

The angle θ of the cone of radiation depends on the speed v of the particle. When a beam of fast charged particles moves through a medium such as glass or transparent plastic, the radiation can be detected and the angle θ can be measured. Thus, the speed of the particle can be determined.

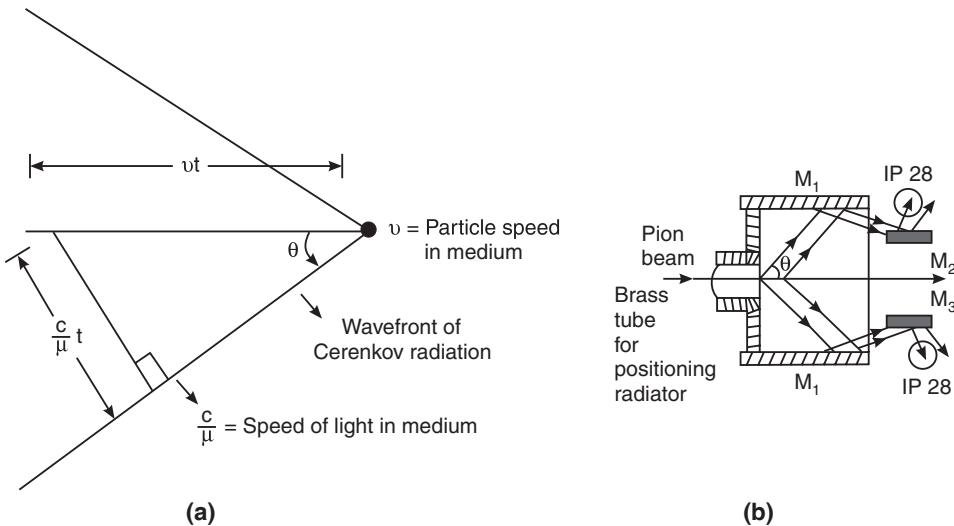


Fig. 23.8

Fig. 23.8 (b) shows a Cerenkov detector designed by Marshall. A collimated beam of pions is allowed to pass along the axis of a large hemispherical Perspex lens. The Cerenkov radiation radiated in this lens is reflected by the cylindrical mirror M_1 on to one of the plane mirrors M_2 , M_3 and then on to the cathode of photomultiplier tubes (IP 28). The two photomultipliers operate in coincidence. The position of the radiator lens is adjusted so that Cerenkov light cone is correctly focused on to the phototubes to give maximum counting rate.

23.9 MASS SPECTROGRAPHS

Atoms that are chemically identical but differ in masses are called *isotopes*. **Mass spectrograph** is an instrument that separates different isotopes from a stream of positive ions of an element by using electric and magnetic fields and measures their individual masses as well as their relative abundance. By analogy with the instrument that separates light of different wavelengths, the instrument that separates ions of different masses is called a *mass spectrograph*. In the mass spectrograph positive ions having the same q/M value are brought to a common focus and produce a line on a photographic plate. Ions having different q/M values produce different lines of different intensity on the photographic plate. Owing to its similarity to an optical line spectrum, the image obtained on the plate is called a **mass spectrum**.

23.10 ASTON MASS SPECTROGRAPH

Principle:

A schematic of Aston's mass spectrograph is shown in Fig. 23.9. A beam of positive ions issuing out of a discharge tube is collimated by a system of slits and is allowed to pass through a uniform electric field first. The electric field produces a dispersion of the ions with respect to the velocity. It is because ions with a given value of q / M traveling slower spend more time in the electric field and experience more deflection, while faster ions are deflected less. Consequently, the ions leave the electric field in the form of a diverging beam.

The beam passes next through a magnetic field whose direction is perpendicular to that of the electric field. In the magnetic field, the faster ions are bent less (large R) whereas the slower ions are bent more (smaller R). It results in focusing of the ions of same q / M value to the same point. Thus, ions of different q / M values get focused at different points on the photographic plate.

Focussing by electric and magnetic fields

The positive ions produced in a discharge tube possess a wide range of velocities. They are collimated by two narrow slits S_1 and S_2 in the form of a very thin ribbon. Let AO be the direction in which the well-collimated stream of positive ions enters the electric field region (see Fig. 23.10). The electric field, E , acts vertically downward from plate P to plate Q and hence the positive ions are deflected downward. As the positive ion beam contains ions having a range of velocities spread between v and $v + dv$, they will be dispersed through an angle $d\theta$. Thus, let θ be the angle of deflection of the beam with respect to the axis of the electric field and $d\theta$ be the angle of dispersion of the beam. Let a (OO') be the distance between the centers of the electric and magnetic fields and b be the distance from the center of the magnetic field to the photographic plate. The magnetic field, B , acts into the plane of the page and under its action the positive ion beam deflects upward. Because of the velocity focusing action of the magnetic field the ion beam converges on the plane of the photographic plate. Let ϕ be the angle of deflection due to magnetic field and $d\phi$ the angle of refocusing.

The deflection of an ion of mass M and charge e in the electric field is inversely proportional to the square of the ion velocity and is given by

$$x = k_1 \frac{eE}{mv^2}$$

But

$$x \propto \theta$$

\therefore

$$\theta \propto k_1 \frac{eE}{mv^2}$$

or

$$\theta = L \frac{e}{mv^2} \quad (23.3)$$

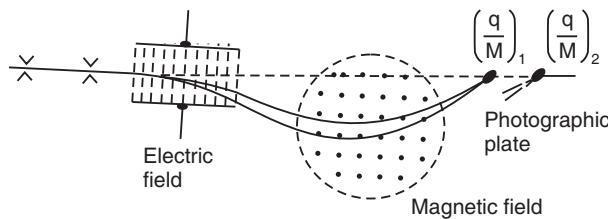


Fig. 23.9

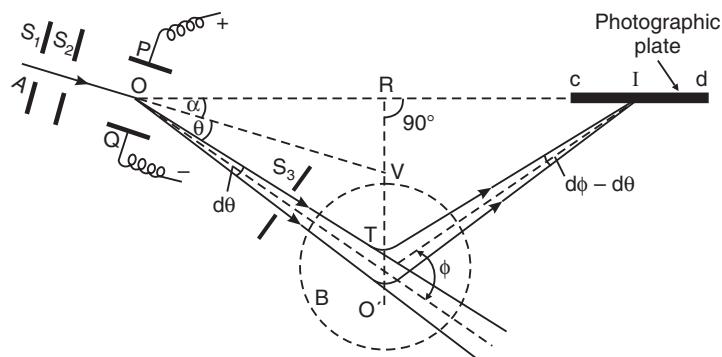


Fig. 23.10. Aston's Mass Spectrograph

where L is a constant depending on the values of E and the distance traveled in the field.

Differentiating equation (23.3), we get

$$\begin{aligned} d\theta &= -\frac{2Le}{mv^3} dv \\ \therefore \frac{d\theta}{\theta} &= -\frac{2Le dv}{mv^3} \cdot \frac{mv^2}{Le} = -\frac{2dv}{v} \end{aligned} \quad (23.4)$$

After emerging from the electric field, the ion beam enters a uniform magnetic field at a mean distance ' a ' from the center of the electric field. The magnetic field causes deflection of the ion beam in the same plane as that of electric deflection. The magnetic deflection is given by

$$y = k_2 \frac{eB}{mv}$$

But

$$y \propto \phi$$

Therefore,

$$\phi \propto k_2 \frac{eB}{mv}$$

or

$$\phi = M \frac{e}{mv} \quad (23.5)$$

where M is a constant depending on the values of B and the distance traveled in the field.

Differentiating equation (23.5), we get

$$\begin{aligned} d\phi &= -\frac{Me}{mv^2} dv \\ \therefore \frac{d\phi}{\phi} &= \frac{Medv}{mv^2} \cdot \frac{mv}{Me} = -\frac{dv}{v} \end{aligned} \quad (23.6)$$

It follows from equs.(23.4) and (23.6) that

$$\frac{d\theta}{\theta} = \frac{2d\phi}{\phi} \quad \text{or} \quad \frac{d\phi}{d\theta} = \frac{\phi}{2\theta} \quad (23.7)$$

In the absence of the magnetic field, the dispersion produced in the beam for a distance $(a + b)$ is equal to $(a + b)d\theta$. The magnetic field acting in a direction perpendicular to the electric field compensates the electric dispersion $d\theta$ completely and refocuses the ion beam at some distance b . Dispersion produced by the magnetic field in a distance b equals $b d\phi$. The ions are all focused to the same position, under the condition that the linear electric dispersion $(a + b)d\theta$ is equal and opposite to the linear magnetic dispersion $b d\phi$. Thus,

$$\begin{aligned} (a + b)d\theta &= b d\phi \\ \text{or} \quad \frac{d\phi}{d\theta} &= \frac{a + b}{b} \end{aligned} \quad (23.8)$$

Comparing equs.(23.7) and (23.8), we get

$$\begin{aligned} \frac{a + b}{b} &= \frac{\phi}{2\theta} \\ \text{or} \quad 2a\theta &= b(\phi - 2\theta) \\ \text{or} \quad \frac{b}{a} &= \frac{2\theta}{\phi - 2\theta} \end{aligned} \quad (23.9)$$

The above equation gives the ion-focusing distance b from the magnetic field.

If a photographic plate is placed along the locus of the ion focusing points, the ions register at different points on the plate.

Now, let $O'R$ be perpendicular to the line cd produced backward. Then in the $\Delta^{le}ROO'$

$$RO' = OO' \sin(\alpha + \theta) = a \sin(\alpha + \theta) \quad (23.10)$$

$$\text{In the } \Delta^{le}RIO', RO' = IO' \sin RIO' = b \sin[180 - (\phi - \alpha - \theta)] = b \sin(\phi - \alpha - \theta) \quad (23.11)$$

Equating (23.10) and (23.11), we obtain

$$a \sin(\alpha + \theta) = b \sin(\phi - \alpha - \theta)$$

For small angles, we can write the above equation as

$$a(\alpha + \theta) = b(\phi - \alpha - \theta) \quad (23.12)$$

Comparing eqns.(23.9) and (23.12), it is clear that the two equations are identical if $\alpha = 0$. Thus, the condition for focusing is that the photographic plate must be placed at an angle θ with the direction of the incident positive ion beam such that $\alpha = \theta$, where θ is the deviation produced by the electric field.

23.11 DEMPSTER MASS SPECTROGRAPH

In 1918 A.J.Dempster, the American physicist built a mass spectrograph which was further improved in 1922.

Principle:

When all the positive ions produced are accelerated by the same high potential difference, it may be assumed that all of them have the same velocity and a homogeneous magnetic field acting normal to the path of the ions will separate the ions of different masses.

Construction:

A schematic diagram of one of the early mass spectrometers developed by Dempster is shown in Fig. 23.11. The material under investigation is fed into the spectrometer in the form of gas. It is ionized with the help of high voltage applied across the spark gap. The positively charged ions of the element are collimated by slit S_1 and are accelerated toward the slit S_2 by a high negative potential of about 1000V. A

stream of positive ions which may be regarded as emerging with uniform energy, spreads at high velocity through the slit S_2 into the semicircular space M . A magnetic field of sufficient strength is applied normal to the path of the ions. The magnetic field causes the stream of ions to follow semicircular path. The ions pass through a third slit S_3 and impinge on a plate P connected to an electrometer or a similar device for measuring the ion current. The instrument is called a *mass spectrometer* because the ion current is measured in the instrument rather than recording the ions on a photographic plate.

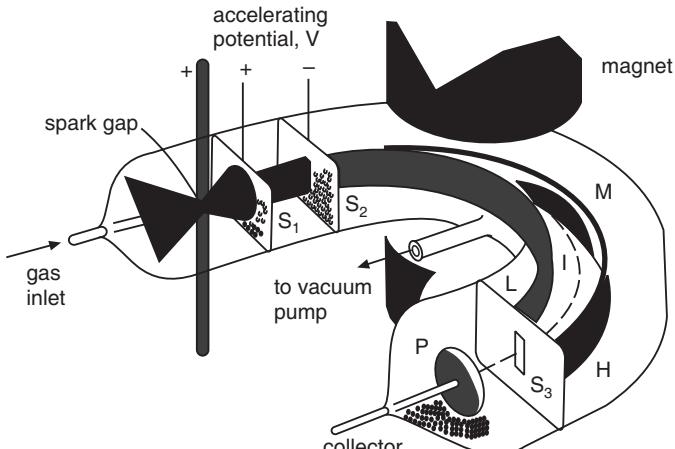


Fig. 23.11

Working:

If V is the accelerating potential through which the positive ions are accelerated in the region between the slits S_1 and S_2 , the velocity of the ions which may be assumed to be the same for all ions, is given by

$$v = \left[\frac{2qV}{M} \right]^{\frac{1}{2}} \quad (23.13)$$

In the magnetic field the ions describe a semicircular path whose radius is given by

$$R = \frac{Mv}{qB} \quad (23.14)$$

Using equ.(23.13) into equ.(23.14), we get

$$\begin{aligned} R &= \frac{M}{qB} \left[\frac{2qV}{M} \right]^{\frac{1}{2}} \\ \text{or } R^2 &= \frac{2MV}{qB^2} \\ \therefore \frac{q}{M} &= \frac{2V}{B^2 R^2} \end{aligned} \quad (23.15)$$

In the mass spectrometer only positive ions describing a definite value of R can pass through the third slit S_3 and reach the plate P . The radius R is defined by slits S_2 and S_3 ; and the q/M values of ions that describe a path of this radius are determined by the accelerating potential V and the magnetic field induction B . In practice, the magnetic field B is kept constant and the potential V is varied steadily. The corresponding ion current is measured for each value of V . Since each accelerating potential corresponds to a definite mass of ions reaching the electrometer, the current can be plotted against the mass of the ions. The resulting graph shows a series of maxima. Each maximum corresponds to a set of ions having a definite q/M value. Knowing the value of the potential V at which a maximum occurs, it is possible to calculate the value of q/M using equ.(23.15).

With the element potassium, Dempster obtained the results that are shown in Fig. 23.12. Ions of potassium passed through the slit S_3 at $V_1 = 866.5$ V and $V_2 = 911$ V. From equ. (23.15) we obtain

$$\begin{aligned} M_1 V_1 &= \text{constant} = M_2 V_2 \\ \therefore M_1 &= \frac{M_2 V_2}{V_1} \end{aligned} \quad (23.16)$$

Dempster knew from other measurements that most potassium atoms have a mass of about 39 amu. Hence he assigned the maximum at 911 V to ^{39}K . Then using the values into equ.(23.16), he found that

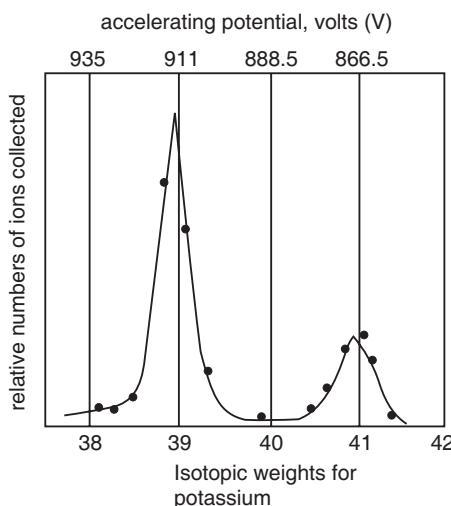


Fig. 23.12

$$M_1 = 39 \text{ amu} \quad \frac{911V}{866.5V} = 41 \text{ amu}$$

The chief advantage of the mass spectrometer is that the relative amounts of the various particles can be estimated from the magnitude of the ion currents at the corresponding maxima. It is easy to see from Fig. 23.12 that the two isotopes of potassium have relative abundances in the ratio 18 : 1.

23.12 BAINBRIDGE MASS SPECTROGRAPH

Principle:

The Bainbridge mass spectrograph is based on the principle that uniform magnetic field acting perpendicular to the path of ions deflects them along circular paths. Ions having the same velocity but different masses are deflected along circular paths of different radii. It is necessary that all the ions in the beam entering the transverse magnetic field must have a single velocity. However, the beam of positive ions produced in a discharge tube and entering the spectrograph will have a wide range of velocities. Therefore, a velocity selector is used to form a single-velocity ion beam.

Construction:

Fig. 23.13 (a) shows the schematic of Bainbridge mass spectrograph. It is essentially a vacuum chamber placed in uniform magnetic field. A discharge tube produces positive ions of the element under investigation. The slits S_1 and S_2 accelerate and also collimate the incoming ion beam. The ion beam passes through a velocity selector. Beyond the velocity filter, another slit S_3 is arranged which further collimates the mono-velocity beam. A photographic plate is mounted in the analyzing chamber in line with the slit S_3 .

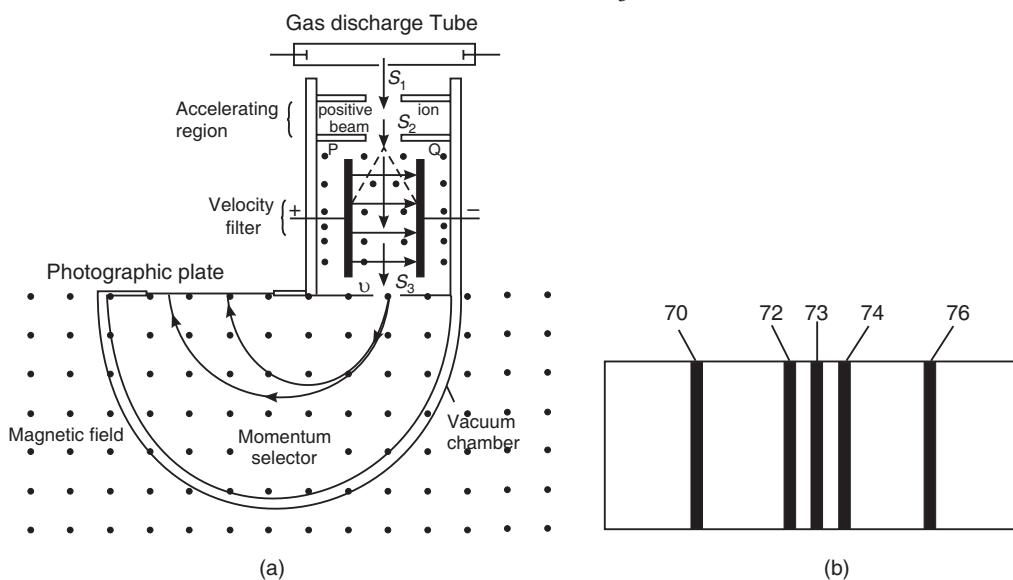


Fig. 23.13. (a) Schematic of Bainbridge Mass Spectrograph **(b)** Mass spectrum of Germanium

Working:

The element under study is taken in the form of gas and introduced into a discharge tube. The gas is ionized under the action of the applied voltage and the positive ions, which are formed in the discharge tube, are accelerated and conducted into the spectrograph through

slits S_1 and S_2 . Most of the positive ions formed will carry a charge of +1 because it is much more difficult to remove further electrons from an already positive ion. The ion beam, thus consisting of ions carrying a positive charge of +1, possesses a wide range of velocities $v \pm \Delta v$. It is then passed through the velocity selector.

A velocity selector is an electro-optic device, which selects a stream of single velocity ions from a beam of ions having a wide range of velocities. It consists of a combination of uniform electric and magnetic fields in crossed configuration.

Referring to Fig. 23.13 (a), the electric field E is produced in the horizontal direction by a set of charged parallel plates PQ and the uniform magnetic field B is applied normal to E in the same region. Both E and B act normal to the ion path. The electric field produces a dispersion of the ions with respect to the velocity. The ions travelling with a velocity $v (=E/B)$ do not experience any net force and they continue to travel along the initial direction. Ions travelling with velocities differing from v are deflected sideways and are absorbed by the metal plates. Therefore, a strictly mono-velocity ion stream, having a velocity $v = E / B$, issues out of velocity selector. This beam, of ribbon shape, then passes through the slit S_3 and enters the analyzing chamber. The uniform magnetic field B acting on the chamber plays the role of a *momentum selector*. Ions of different masses are deflected along circular paths of different radii depending on their momentum and strike the photographic plate after completion of a semicircular movement. They leave vertical lines on the plate, which resemble spectral lines formed by light. Hence, it is called **mass spectrum**. The number of lines on the plate corresponds to the number of isotopes of the element. Fig. 23.13 (b) shows the mass spectrum of germanium.

Considering the isotope with a mass M_1 , the circular path described by it has the radius R .

$$\text{Now } R = \frac{Mv}{qB}$$

$$\text{But as } v = \frac{E}{B}, \quad R = \frac{ME}{qB^2} \quad (23.17)$$

The distance of the line formed by the isotope on the photographic plate can be measured from a reference line. If x is the distance of the line from the slit, then

$$x = 2R$$

$$\therefore M = \frac{qB^2}{2E} x \quad (23.18)$$

or $M = kx$

where k is a constant because q, B and E are constant parameters.

$$\therefore M \propto x$$

The above relation is linear in x and hence the mass scale is *linear*.

As the relation between M and x is linear, the mass of an ion may be obtained from the measurements of E, B and x by proper calibration of the photographic plate. The relative masses of two isotopes involve only measurement of x and therefore, it can be obtained with high precision. If M_1 and M_2 are the masses of two isotopes and if x_1 and x_2 are the distances of the corresponding lines respectively from S_3 , the line separation is given by

$$\Delta x = (x_2 - x_1) = \frac{2E}{qB^2} (M_2 - M_1) \quad (23.19)$$

In practice, an ion with known mass is used as a standard and the mass of the ion under study is compared with it by means of equation (23.19).

Example 23.2. In a Bainbridge mass spectrograph, the magnetic field in the velocity selector is 1.0 T and the ions having a speed of 4×10^6 m/s pass through it undeflected.

- What should be the electric field between the plates?
- If the separation of the plates is 0.5 cm, what is the potential difference between the plates?

Solution. The velocity of the undeflected ions is given by $v = E/B$.

$$E = vB = (4 \times 10^6 \text{ m/s})(1.0 \text{ T}) = 4 \times 10^6 \text{ V/m.}$$

The potential difference between the plates is given by $V = Ed$.

$$\therefore V = (4 \times 10^6 \text{ V/m})(5 \times 10^{-3} \text{ m}) = 20 \text{ kV.}$$

Example 23.3. The electric field between the plates of velocity selector is 150 V/cm and the magnetic field is 0.5 T. If the source contains the three isotopes of magnesium Mg^{24} , Mg^{25} and Mg^{26} and the ions are singly charged, find the distance between the lines formed by these isotopes on the photographic plate. (RTMNU, 2010)

Solution. The separation between the successive lines is given by $\Delta x = \frac{2E}{qB^2}(M_2 - M_1)$.

$$\therefore \Delta x = \frac{2(1500 \text{ V/m})}{(1.602 \times 10^{-19} \text{ C})(0.5 \text{ T})^2} (25 \text{ u} - 24 \text{ u}) \frac{3000 \text{ V/m}}{(1.602 \times 10^{-19} \text{ C})(0.5 \text{ T})^2} (1.66 \times 10^{-27} \text{ kg}) \\ = 1.2 \text{ mm}$$

Example 23.4. In a Bainbridge mass spectrograph, the electric field used is 8×10^4 V/m and the magnetic field common to both places is 0.55 wb/m². If the ion source consists of singly ionized neon isotopes of atomic masses 20 and 22, calculate linear separation of lines formed on photographic plate.

Solution. The separation between the successive lines is given by $\Delta x = \frac{2E}{qB^2}(M_2 - M_1)$.

$$\therefore \Delta x = \frac{2(8 \times 10^4 \text{ V/m})}{(1.602 \times 10^{-19} \text{ C})(0.55 \text{ wb/m}^2)^2} (22 \text{ u} - 20 \text{ u}) \\ = \frac{16 \times 10^4 \text{ V/m}}{(1.602 \times 10^{-19} \text{ C})(0.55 \text{ T})^2} (2 \times 1.66 \times 10^{-27} \text{ kg}) = 11 \text{ mm}$$

23.13 PARTICLE ACCELERATORS

Particles are accelerated to very high energies using electric and magnetic fields. A charged particle acquires energy 'qV' when it is accelerated by a potential difference V. Choosing V to be very large, it should be possible to achieve any desired energy. In practice, the maximum voltage that one can produce limits the maximum energy of the particles. The production of maximum voltage in fact depends on the problem of providing adequate insulation. The Cockcroft-Walton and the Van de Graaff accelerators are examples of single-step high voltage accelerators. We can bypass the problem of insulation if we use small voltages and provide acceleration through small steps. By making the number of steps large, we can produce very high-energy charged particle beam. Two different types of small voltage accelerators are built which differ in their principle of operation. They are known as linear accelerators and cyclic accelerators.

(i) Linear accelerator: A linear accelerator (LINAC) accelerates charged particles to high energies without the need for very high voltages. In a linear accelerator, there is a succession of electrodes to which an alternating voltage is applied. Successive tubes have opposite voltages, but the voltage alternates with the frequency of the applied voltage. If the lengths of the tubes are correctly chosen, the motion of the charged particles is synchronised with the alteration of the voltage so that they cross the gap between the successive tubes at the right time to receive a push that increases their energy.

The largest proton linear accelerator is at the University of Minnesota, U.S.A., which can accelerate protons to 68 MeV. The largest electron linear accelerator is at Stanford University, U.S.A. It is 3.2 km long and can accelerate electrons to 25 GeV.

(ii) Cyclic accelerator: In cyclic accelerators, the particles are forced by a magnetic field to describe a curved path along which they receive increase in energy from electric fields at certain points on its path. Cyclotron is the first cyclic accelerator. The maximum energy that can be reached with a cyclotron is limited to several MeV. As demand arose for more energetic particles, cyclotrons were replaced by synchrotrons. The most powerful of these machines is the super proton synchrotron (SPS) at the European Centre for Nuclear Research (CERN), Switzerland, which can accelerate protons to an energy of 26 GeV.

Examples: Van de Graff generator and LINAC are examples of linear accelerators, while cyclotron and betatron are examples of cyclic accelerators.

23.14 DRIFT TUBE ACCELERATOR

A **linear accelerator** (LINAC) accelerates charged particles to high energies without the need for very high voltages. In linear accelerators, there are a series of coaxial hollow cylindrical electrodes, known as *drift tubes* to which an alternating voltage is applied. Successive tubes have opposite voltages, but the voltage alternates with the frequency of the applied voltage. If the lengths of the tubes are correctly chosen, the motion of the charged particles is synchronized with the alteration of the voltage so that they cross the gap between the successive tubes at the right time to receive a push that increases their energy.

Construction and working

The LINAC consists of a set of drift tubes of increasing lengths arranged linearly in an evacuated glass chamber (see Fig. 23.14). Alternate tubes are connected together. The odd numbered tubes are joined to one terminal and the even numbered tubes are joined to the other terminal of a high frequency power supply, which is in fact an r.f. oscillator. Thus, when tubes C_1 , C_3 , C_5 are positive, the other tubes C_2 , C_4 and C_6 are negative and reversal of potential takes place in step with the frequency of the power supply. The positive ions produced by the source S travel along the axis of the tubes and are accelerated on crossing the gaps between the drift tubes. The ions do not receive acceleration while traveling inside the tubes because electric field does not exist there.

Initially, the positive ions move towards C_1 when it is negative. Then they travel through it with constant velocity. Length of C_1 is so selected in relation to the frequency of the oscillator that these ions arrive at the gap between C_1 and C_2 at the instant when C_2 has just become

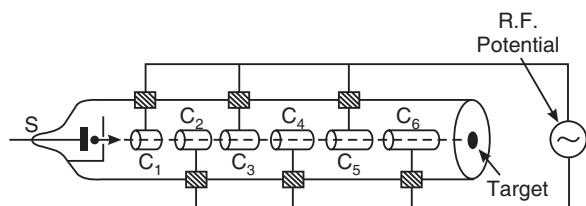


Fig. 23.14

negative. It will happen only when the time taken by ions to travel through C_1 is exactly equal to half the time period of the oscillator. The ions are further accelerated across the gap and travel through C_2 at a higher uniform velocity. The length of C_2 is such that the ions arrive in the gap at the instant when C_3 has just become negative and are further accelerated. Thus, the ions are accelerated in the successive gaps and emerge from the final drift tube with extremely high velocities.

If N is the number of tubes, V is the maximum voltage of the oscillator, then the energy acquired by an ion of charge q is given by

$$E = \frac{1}{2} m v_N^2 = N q V \quad (23.20)$$

The time required to travel through any tube is equal to half time period of the applied r.f. voltage. Hence,

$$t = \frac{T}{2} = \frac{1}{2v}$$

If l_N is the length of the N^{th} tube and v_N the velocity of the ion while traveling through it, then

$$t_N = \frac{l_N}{v_N}$$

$$\therefore \frac{l_N}{v_N} = \frac{1}{2v}$$

or

$$l_N = \frac{v_N}{2v} \quad (23.21)$$

Substituting the value of v_N from equ. (23.20) into the above equation, we get

$$l_N = \sqrt{\frac{NqV}{2mv^2}} \quad (23.22)$$

The largest proton linear accelerator is at the University of Minnesota, USA, which can accelerate protons to 68 MeV. The largest electron linear accelerator is at Stanford University, USA. It is 3.2 km long and can accelerate electrons to 25 GeV.

Example 23.5. In a linear accelerator, proton accelerated thrice by a potential of 40 kV leaves a tube and enters an accelerating space of length 30 cm before entering the next tube. Calculate the frequency of the r.f. voltage and the length of the tube entered by the proton.

Solution. Let v_1 and v_2 be the velocities of the proton entering and leaving the accelerating space. Let q and m be the mass and charge of the proton respectively. Then

$$\frac{1}{2} m v_1^2 = 3 \times 1.602 \times 10^{-19} \text{ C} \times 4 \times 10^4 \text{ V}$$

$$\therefore v_1 = \left[\frac{2 \times 3 \times 1.602 \times 10^{-19} \text{ C} \times 4 \times 10^4 \text{ V}}{1.67 \times 10^{-27} \text{ kg}} \right]^{\frac{1}{2}} = 4.8 \times 10^6 \text{ m/s}$$

$$\text{Similarly } v_2 = \left[\frac{2 \times 4 \times 1.602 \times 10^{-19} \text{ C} \times 4 \times 10^4 \text{ V}}{1.67 \times 10^{-27} \text{ kg}} \right]^{\frac{1}{2}} = 5.5 \times 10^6 \text{ m/s}$$

$$\text{Mean velocity } v = 5.165 \times 10^6 \text{ m/s.}$$

The time taken to travel 30 cm = 0.3 m equals the half-period of the r.f. voltage.

$$\therefore \frac{T}{2} = \frac{0.3 \text{ m}}{5.165 \times 10^6 \text{ m/s}}$$

$$\therefore \text{Frequency of the r.f. voltage } v = \frac{1}{T} = \frac{5.165 \times 10^6 \text{ m/s}}{2 \times 0.3 \text{ m}} = 8.6 \text{ MHz.}$$

$$\text{Length of the next tube entered by the protons } l = \frac{v_2}{2v} = \frac{5.5 \times 10^6 \text{ m/s}}{2(8.6 \times 10^6 \text{ Hz})} = 0.32 \text{ m.}$$

23.15 CYCLOTRON

Cyclotron is the first **cyclic accelerator** built by E.O.Lawrence and M.S.Livingston in 1932.

Principle:

A moving charged particle describes a circular path in the presence of a transverse uniform magnetic field. The frequency of revolution of the charged particle is given by

$$v = \frac{qB}{2\pi m} \quad (23.23)$$

which is independent of the particle velocity. It means that a faster particle moving in a circle of bigger radius and a slower particle moving in a smaller circle take the same time for completing one revolution in a given uniform magnetic field. Hence, charged particles having different initial velocities can be uniformly accelerated to produce high-energy particle beam using a combination of crossed electric and magnetic fields.

Construction:

The schematic of a cyclotron is shown in Fig. 23.15. A cyclotron consists of two hollow metal dees formed by cutting a short, cylindrical box along its diameter. The dees are separated by a few centimetres from each other. They are insulated from each other and are placed in a vacuum chamber located between the pole pieces of an electromagnet. The magnet produces a uniform magnetic field in a direction perpendicular to the semicircular faces of dees. A high frequency oscillator is connected to the dees, which produces a r.f. electric field in the gap between the dees. A source of charged particles is located at the centre of the gap between the dees.

Working:

Charged particles, say protons, are injected by the ion source S into the gap between the dees. The protons are accelerated by the r.f. electric field existing in the gap toward the dee, which is at a negative potential at that instant. However, the magnetic field that is acting perpendicular to the protons deflects them along a circular path. The protons travel in the hollow region of the dee and come back into the gap after completion of a half-revolution. The time taken for a half-revolution is given by

$$\frac{T}{2} = \frac{\pi m}{qB} \quad (23.24)$$

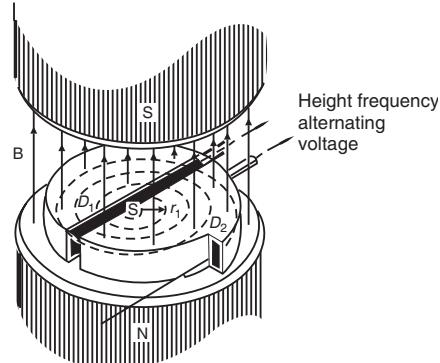


Fig. 23.15

where T is the time period for the circular path in the magnetic field. The protons will be further accelerated if the dees reverse their polarity at the instant when the protons emerge

into the gap. In such a case, the protons travel further in a semicircular path in the other dee and reach the gap in a time $T/2$. If, again at the same instant, the dees reverse their polarity, the protons receive another dose of acceleration. The process is repeated over and again many times. As the velocity increases with each dose of acceleration, the protons describe a spiral path in the dees, as shown in Fig. 23.15. During each revolution, each proton receives energy of $2qV$ electron volts and after about a hundred or more revolutions, the protons acquire energies of the order of several million electron volts. At the end of the journey, the proton beam is pulled out of its circular path by a negatively charged deflector plate and emerges out of the chamber through a narrow aperture.

Condition of Resonance:

In a cyclotron the protons are progressively accelerated provided that the time period T_0 of the r.f. electric field equals the time period T of revolution of protons in the magnetic field, B . Thus, it is required that the condition $T_0 = T$ is fulfilled. It means that

$$T_0 = \frac{2\pi m}{qB} \quad (23.25)$$

or

$$v_0 = \frac{qB}{2\pi m} \quad (23.26)$$

The above relation is known as the **condition of resonance**.

Energy Acquired by the Charged Particles:

A proton makes N revolutions, receiving energy of $2qV$ during each revolution. The total kinetic energy acquired is

$$E = 2NqV \quad (23.27)$$

The radius R of the final orbit is given by $R = \frac{mv_{\max}}{qB}$

The velocity in the final orbit is therefore $v_{\max} = \frac{qBR}{m}$

$$E = \frac{1}{2}mv_{\max}^2 = \frac{B^2q^2R^2}{2m} \quad (23.28)$$

It is seen from the above expression that ***the final energy acquired by the particles in a cyclotron does not depend on the magnitude of the voltage applied across the dees***. A comparison of the two equations for final energy of the particles suggests that ***the particles will have to execute a larger number of revolutions if the applied voltage is low or to make smaller number revolutions if the applied voltage is higher, to gain the same amount of kinetic energy***.

Role of Electric and Magnetic Fields:

The primary function of the r.f. electric field consists in imparting high kinetic energy to the charged particles. The primary function of magnetic field is to deflect the charged particles along circular path so that they repeatedly pass through the r.f. electric field region, and acquire more and more energy.

The second function of r.f. electric field is to focus the charged particles into a sharp beam. The non-uniform electric field in the gap between the dees plays the role of electron lens (Fig. 23.16 a.) and causes focussing of charged particles. The second function of magnetic field is to correct the paths of charged particles revolving nearer to periphery. Particles, which tend

to stray from the median plane, are brought back to the median plane due to the Lorentz force component produced by non-uniform field at the outer edges of the magnet poles (Fig. 23.16b).

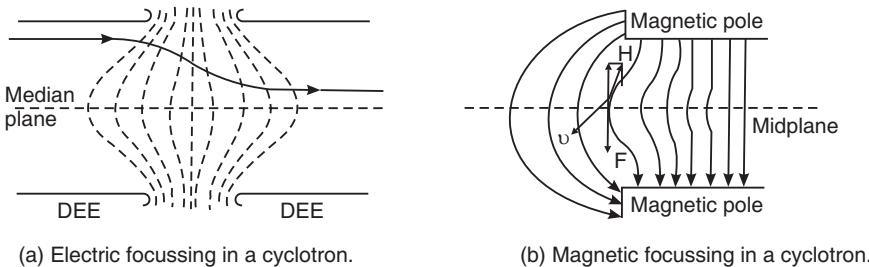


Fig. 23.16

Limitation of Cyclotron:

The kinetic energy acquired by charged particles in a cyclotron is given by

$$E = \frac{B^2 R^2 q^2}{2m}$$

According to this relation, it appears that the maximum energy of the particle beam is limited by the magnetic field B . Increasing the size of the magnet can increase B . However, there exists an ultimate limit for the size of magnet and the electric current to drive the electromagnet.

There arises a more basic limitation due to the relativistic variation of particle mass at velocity $v \approx c$. It is known that the mass ' m ' of a particle increases with velocity according to the relation

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}} \quad (23.29)$$

where m_0 is known as the rest mass of the particle.

Therefore, the time taken by the particle to complete a revolution increases from

$$T_0 = \frac{2\pi m_0}{qB}$$

to

$$\begin{aligned} T &= \frac{2\pi m}{qB}, \text{ at velocities comparable to } c. \\ &= \frac{2\pi m_0}{qB} \left[\frac{1}{\sqrt{1 - v^2/c^2}} \right] \end{aligned} \quad (23.30)$$

We can express T as

$$T = T_0 \left(\frac{m}{m_0} \right) = T_0 \left(\frac{mc^2}{m_0 c^2} \right) = T_0 \left(\frac{E}{E_0} \right)$$

where E_0 is the energy of the particle at $v \ll c$ and E is the energy of the particle at $v \approx c$.

Writing

$$E = E_0 + K,$$

$$T = \left(1 + \frac{K}{E_0} \right) \quad (23.31)$$

As K increases with increasing particle velocity, $T > T_0$, the above inequality shows that the period of particle revolution significantly increases with the increasing velocity. The result is that the particle fails to reach the gap at the right moment when the electric field reversal occurs. Because of the failure, it gets decelerated instead of being accelerated.

In case of electrons, for example, the rest mass energy

$$E_0 = m_0 c^2 = (9.1 \times 10^{-31} \text{ kg}) (3 \times 10^8 \text{ m/s})^2 = 0.51 \text{ MeV}$$

It implies that even at low energies of the order of less than 1 MeV the period of revolution of electrons is doubled and hence they cannot show up in the gap at the required instant. Therefore, they drop out of synchronism and cease to be accelerated further. Hence, electrons cannot be accelerated to high energies in a cyclotron.

Example 23.6. A cyclotron with its dees of radius 2 m has a magnetic field of 0.75 wb/m^2 . Calculate the maximum energies to which (i) protons and (ii) deuterons can be accelerated.

Solution.

The maximum energy to which particles are accelerated in a cyclotron is given by

$$E_{\max} = \frac{B^2 q^2 R^2}{2m}$$

(i) In case of protons

$$\begin{aligned} E_{\max} &= \frac{(0.75 \text{ wb/m}^2)^2 (1.602 \times 10^{-19} \text{ C})^2 (2\text{m})^2}{2(1.67 \times 10^{-27} \text{ kg})} \\ &= 1.73 \times 10^{-11} \frac{\text{wb}^2 \cdot \text{C}^2 \cdot \text{m}^2}{\text{m}^4 \cdot \text{kg}} = 1.73 \times 10^{-11} \frac{\text{V}^2 \cdot \text{s}^2 \cdot \text{C}^2}{\text{m}^2 \cdot \text{kg}} \\ &= 1.73 \times 10^{-11} \text{ J} = (1.73 \times 10^{-11})(6.24 \times 10^{18} \text{ eV}) \\ &= 107.9 \times 10^6 \text{ eV} \end{aligned}$$

∴

$$E_{\max} = 108 \text{ MeV}$$

(ii) In case of deuterons

$$\begin{aligned} E_{\max} &= \frac{(0.75 \text{ wb/m}^2)^2 (1.602 \times 10^{-19} \text{ C})^2 (2\text{m})^2}{2(3.34 \times 10^{-27} \text{ kg})} \\ &= 8.64 \times 10^{-12} \text{ J} = (8.64 \times 10^{-12})(6.24 \times 10^{18} \text{ eV}) \\ &= 53.91 \text{ MeV.} \end{aligned}$$

Example 23.7. Protons are accelerated in a cyclotron. The magnetic field strength is 1.3 wb/m^2 and the radius of the last semicircle is 0.5 m.

(i) What must be the frequency of the oscillator supplying power to the dees?

(ii) What is the final energy acquired by the proton beam?

(iii) If the total transit time of a proton is $3.3 \mu\text{s}$, how much energy is imparted to protons in each passage from one dee to the other?

Solution. (i) Frequency $v = \frac{Bq}{2\pi m} = \frac{1.3 \text{ wb/m}^2 \times 1.602 \times 10^{-19} \text{ C}}{2 \times 3.143 \times 1.67 \times 10^{-27} \text{ kg}} = 19.85 \text{ MHz.}$

(ii) Final energy $E_{\max} = \frac{B^2 q^2 R^2}{2m} = \frac{(1.3 \text{ wb/m}^2)^2 (1.602 \times 10^{-19} \text{ C})^2 (0.5\text{m})^2}{2(1.67 \times 10^{-27} \text{ kg})}$
 $= 20.28 \text{ MeV.}$

(iii) Number of revolutions $N = 2vT = 2 \times 19.85 \times 10^6 \text{ Hz} \times 3.3 \times 10^{-6} \text{ s} = 131$

$$\text{Energy gained by proton during one transit} = \frac{2NeV}{N} = \frac{20.28 \text{ MeV}}{131} = 155 \text{ keV.}$$

Example 23.8. The magnetic field strength in a certain cyclotron is 0.9 Wb/m². If light hydrogen ions (protons) are accelerated in the cyclotron

(i) What must be the frequency of the oscillator supplying power to the dees?

(ii) If each passage of ions across the accelerating gap increases the energy of the ion by 60,000 eV, how long does it take for the ion introduced at the centre of the dees to emerge at the rim of the dee with energy of 6 MeV?

(iii) Calculate the radius of the last semicircle before the ion emerges from the cyclotron.

$$\text{Solution. (i) Frequency } v = \frac{Bq}{2\pi m} = \frac{0.9 \text{ wb / m}^2 \times 1.602 \times 10^{-19} \text{ C}}{2 \times 3.143 \times 1.67 \times 10^{-27} \text{ kg}} = 13.72 \text{ MHz.}$$

$$\text{(ii) Transit time } T = \frac{N}{2v} = \frac{E_{\max} / E_1}{2v} = \frac{6 \times 10^6 (1.602 \times 10^{-19}) \text{ J}}{6 \times 10^4 (1.602 \times 10^{-19}) \text{ J} \times 2 \times 13.72 \times 10^6 \text{ Hz}}$$

$$= 3.64 \mu\text{s.}$$

$$\text{(iii) Maximum Energy } E_{\max} = \frac{B^2 q^2 R^2}{2m} \quad \therefore \quad R = \frac{\sqrt{2mE_{\max}}}{Bq}$$

$$\therefore R = \frac{(2 \times 1.67 \times 10^{-27} \text{ kg} \times 6 \times 10^6 \text{ eV} \times 1.602 \times 10^{-19} \text{ J / eV})^{1/2}}{(0.9 \text{ wb / m}^2)(1.602 \times 10^{-19} \text{ C})} = 0.39 \text{ m.}$$

Example 23.9. Deuterons are accelerated in a fixed frequency cyclotron to a maximum dee orbit radius of 0.88 m. The magnetic field is 1.4 T. Calculate the energy of the emerging deuteron beam and the frequency of the dee voltage. What change in magnetic flux density is necessary if doubly charged helium ions are accelerated?

Given atomic masses: H² = 2.014102 amu and He⁴ = 4.002603 amu

$$\text{Solution. (i) Frequency } v = \frac{Bq}{2\pi m} = \frac{1.4 \text{ wb / m}^2 \times 1.602 \times 10^{-19} \text{ C}}{2 \times 3.143 \times 2.014102 \times 1.67 \times 10^{-27} \text{ kg}} = 10.61 \text{ MHz.}$$

$$\text{(ii) Energy } E_{\max} = \frac{B^2 q^2 R^2}{2m}$$

$$= \frac{(1.4 \text{ wb / m}^2)^2 (1.602 \times 10^{-19} \text{ C})^2 (0.88 \text{ m})^2}{2 \times 2.014102 \times 1.67 \times 10^{-27} \text{ kg}}$$

$$= 36.19 \text{ MeV.}$$

$$\text{(iii) Magnetic flux density } B_{\text{He}} = \frac{2\pi mv}{q}$$

$$= \frac{2 \times 3.143 \times 4.002603 \times 1.67 \times 10^{-27} \text{ kg} \times 10.61 \times 10^6 \text{ Hz}}{2 \times 1.602 \times 10^{-19} \text{ C}}$$

$$= 1.39 \text{ T}$$

Example 23.10. A particle cyclotron is designed with dees of radius 75 cm and with magnets that can provide a field of 1.5 T.

- To what frequency should the oscillator be set if deuterons are to be accelerated?
- What is the maximum energy of deuterons that can be obtained?

Solution. (i) Frequency $v = \frac{Bq}{2\pi m} = \frac{1.5 \text{ wb/m}^2 \times 1.602 \times 10^{-19} \text{ C}}{2 \times 3.143 \times 2 \times 1.67 \times 10^{-27} \text{ kg}} = 11.45 \text{ MHz.}$

(ii) Energy $E_{\max} = \frac{B^2 q^2 R^2}{2m} = \frac{(1.5 \text{ wb/m}^2)^2 (1.602 \times 10^{-19} \text{ C})^2 (0.75 \text{ m})^2}{2 \times 2 \times (1.67 \times 10^{-27} \text{ kg})} = 30.26 \text{ MeV.}$

23.16 SYNCHROCYCLOTRON

Synchrocyclotron is a modified form of cyclotron. In a cyclotron the loss of resonance between the applied r.f. voltage and the moving ion is caused by the relativistic increase in mass of the ion. With growth of mass, the frequency of revolution of the ions in the cyclotron decreases and the ion gets out of phase with the h.f. voltage of fixed frequency applied across the dees and as a result cannot be accelerated. The decrease in ion frequency may be compensated and the ion may be kept in phase with the h.f. voltage in two ways. One of the ways is to keep the value of the magnetic field B constant and decreasing the frequency of h.f. generator in step with the decrease in the frequency of revolution of the ion. The frequency of the applied alternating electric field is gradually changed at such a rate that as the ion lags a little due to increase in mass, the electric field frequency also automatically decreases and the ion always enters the dee at the right moment when it experiences maximum acceleration. Accelerators based on this principle are known as **synchrocyclotrons**.

A synchrocyclotron (Fig. 23.17) consists of only one dee enclosed in a vacuum chamber. The entire unit is kept between the pole pieces of a huge electromagnet. Instead of the second dee, a metal plate (ejector) is held opposite the opening of the dee. The output of a h.f. generator is connected between the single dee and the earthed chamber. The frequency of the h.f. generator is modulated with a low frequency of 120 Hz. The high-energy ions are extracted from the chamber with the help of a high positive pulse applied to the ejector plate. A synchrocyclotron is also known as a *frequency modulated cyclotron*.

The following are the basic differences between a cyclotron and a synchrocyclotron.

1. A cyclotron uses two dees, whereas a synchrocyclotron uses only one dee.

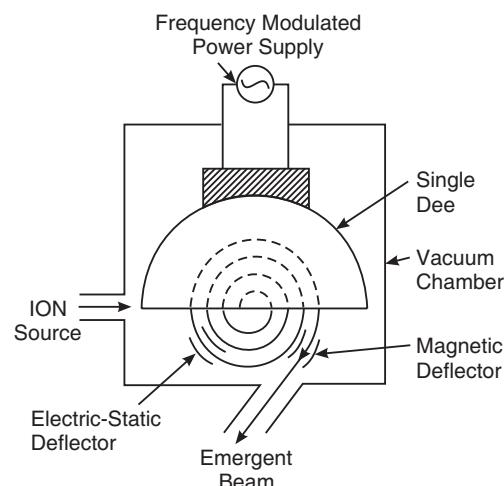


Fig. 23.17. A Schematic diagram of a synchrocyclotron. The spiral illustrates the typical path of a charged particle accelerated in the synchrocyclotron

2. The frequency of the h.f. generator is fixed in a cyclotron operation, whereas the h.f. generator is modulated and hence is variable in a synchrocyclotron.
3. A continuous beam of high-energy ions is produced in a cyclotron. In case of synchrocyclotron ions come out in bursts of a few hundred per sec. Each burst lasts for about 100 μ s.

23.17 BETATRON

Beta-particles are fast moving electrons. The accelerator that produces fast moving electrons is called a **betatron**. D.W.Kerst constructed the first betatron in 1941. With a betatron it is possible to obtain electrons of 300 MeV energy.

Principle of Betatron

The principle of betatron is very much similar to that of transformer. In the transformer, if an alternating electric current is passed through the primary coil an alternating field will appear in the core. This field induces an e.m.f. in the secondary coil.

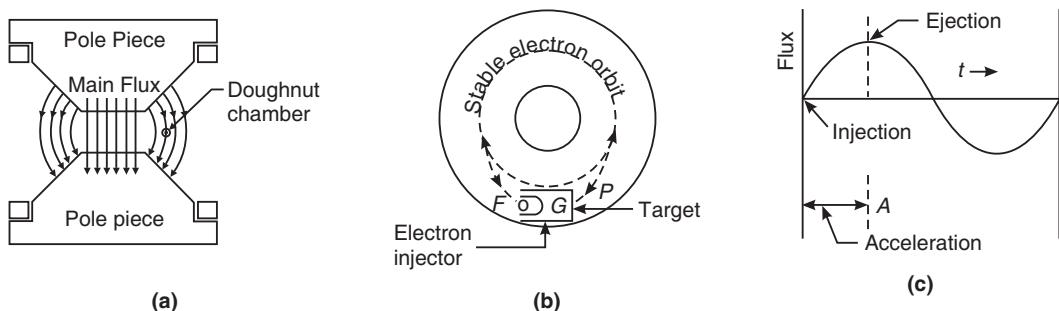


Fig. 23.18

In a betatron, the electrons are accelerated with the help of an electric field produced by a changing magnetic field. The electrons are held in an orbit of fixed radius by varying transverse magnetic field acting on the orbit and simultaneously these electrons are imparted energy by an induced e.m.f. resulting from an increase in the magnetic field.

Construction

A schematic of a betatron is shown in Fig. 23.18 (a). The betatron consists of a highly evacuated chamber made of glass or porcelain and has the shape of a doughnut. The chamber is mounted between the pole pieces of an electromagnet fed by an alternating current of 50 Hz. The poles of electromagnet are tapered (Fig. 23.18 a) so that the doughnut shaped chamber lies in the region of the magnetic field where flux density is half the main flux density. It maintains electrons in an orbit of constant radius in the annular space of doughnut chamber. To introduce electrons into the stable orbit, an electron gun is used. It consists of a filament F that gives out thermionic electrons, a focusing grid G and the positive plate P (Fig. 23.18 b).

The electrons are accelerated for a time 1/200 second at intervals of 1/50 seconds.

Working

The electron gun injects electrons into annular tube at the instant when the magnetic flux just starts increasing. An increasing magnetic flux in a given direction is only obtained during the quarter cycle in which the current increases from zero to its maximum value (Fig. 23.18 c). The electrons injected in a plane close to the stable orbit position will be accelerated by the magnetic field and execute a damped oscillatory motion in the beginning but will finally

settle in the orbit. During $1/200$ s, when the magnetic field is increasing the electron will make several hundred thousand revolutions in the stable orbit gaining energy continuously. The electrons must be ejected from the betatron when the magnetic field reaches its maximum value; otherwise the electrons would slow down as the magnetic flux decreases and finally will reverse their direction. The electrons are deflected from their stable orbit by sending a pulse of current through an auxiliary coil. The high energy electron beam can be made to strike a target T within the tube, thus producing an intense x-ray beam (more correctly γ -rays) or the electron beam can be guided out through a window.

Betatron is like a transformer

The action of the betatron depends upon the same principle as that of the transformer in which an ac current applied to a primary coil induces a similar current in the secondary windings. The primary current produces an oscillating magnetic field which in turn induces an oscillating emf in the secondary coil. The betatron is also like a transformer in which a cloud of electrons located inside a dough-nut shaped vacuum chamber takes the place of the secondary windings. The chamber is placed within the pole pieces of an electromagnet energized by an alternating pulsed current and the magnet produces a strong varying field in the central place or hole of the dough-nut. The electrons move in a circular orbit of constant radius within the vacuum chamber and gain energy by induction because of the change with time of the magnetic flux linking the orbit. Thus the electromagnet plays the role of the primary coil.

Betatron Condition

In a betatron, electrons are accelerated with the help of an electric field produced by a changing magnetic field. The electrons are maintained in a circular orbit by the magnetic field and at the same time they are accelerated by an induced emf resulting from an increase in the magnetic field. Let us consider an electron moving in an orbit of radius ' r ' where the total magnetic flux through the orbit is ϕ (Fig. 23.18 d). The flux increases at a rate $d\phi/dt$ and the electric field induced in the electron orbit may be written from Faraday's laws as

$$V = - \frac{d\phi}{dt} \quad (23.32)$$

The work done on the electron in one revolution = $V e$.

The force, F , on the electron acts along the tangent to the circular path at any point.

$$\text{Work done} = \text{Force} \times \text{distance} = F \cdot 2\pi r$$

∴

$$Ve = F \cdot 2\pi r$$

∴

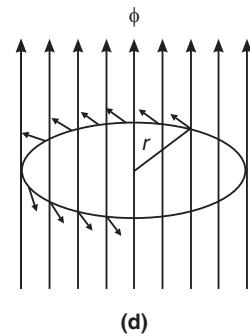
$$F = \frac{Ve}{2\pi r}$$

or

$$F = - \frac{e}{2\pi r} \cdot \frac{d\phi}{dt} \quad (23.33)$$

Under the influence of the induced electric field, the energy of each electron increases as it goes round the orbit. Consequently, its radius will tend to increase causing the electron to spiral outward. The condition for confining the electron to a circular orbit of a constant radius can be obtained as follows.

The momentum of the electron in the circular orbit under the action of magnetic field is given by



$$m\mathbf{v} = Ber$$

$$\therefore F = \frac{d}{dt}(m\mathbf{v}) = e \frac{d}{dt}(Br)$$

When r is constant, the tangential force becomes

$$\frac{d}{dt}(m\mathbf{v}) = er \frac{dB}{dt} \quad (23.34)$$

To maintain a constant radius of the orbit, the net tangential force on the electron must be zero, that is, the values of F given by relations (23.33) and (23.34) must be numerically equal.

Equating (23.33) and (23.34), we get

$$\frac{e}{2\pi r} \frac{d\phi}{dt} = er \frac{dB}{dt}$$

or

$$d\phi = 2\pi r^2 \times dB$$

On integration of the above equation, we obtain

$$\phi = 2\pi r^2 B \quad (23.35)$$

The above condition is known as the **betatron condition**.

If a uniform magnetic field B acts over a ring of area πr^2 , the magnetic flux

$$\phi = \pi r^2 B \quad (23.36)$$

Eq. (23.35) and (23.36) imply that *the electron moves in an orbit of constant radius in an increasing magnetic field if the field in the orbit itself is half of the field over the whole area of the orbit*.

Number of revolutions in the fixed orbits

Electrons are accelerated in betatron for a time interval equal to a quarter of time period. i.e. $T/4$ s. Assuming the electron velocity in the orbit to be equal to the velocity of light ' c ', the total distance traveled by the electron is

$$L = c \times \frac{T}{4} = \frac{c}{4v} = \frac{\pi c}{2\omega} \quad (23.37)$$

$$\therefore \text{Number of revolutions, } N = \frac{L}{2\pi r} = \frac{c}{4\omega r} \quad (23.38)$$

Energy acquired by the electrons

The momentum of the electron, $m\mathbf{v} = \frac{E}{c}$

Also,

$$m\mathbf{v} = Ber$$

$$\therefore \frac{E}{c} = Ber$$

The energy acquired by the electrons, $E = Berc$

(23.39)

Alternative Derivation

Let us suppose that magnetic flux in the betatron is given by the relation

$$\phi = \phi_0 \sin \omega t$$

As an increasing magnetic field in a given direction is only obtained during the quarter cycle in which the current in the electromagnet increases from zero to maximum value.

$$\therefore \text{The time of acceleration} = \frac{T}{4} = \frac{1}{4} \times \frac{2\pi}{\omega} = \frac{\pi}{2\omega}$$

where T is time period of the changing magnetic flux and ω the corresponding angular frequency.

Energy gained by the electron per turn = $V e$

$$= e \frac{d\phi}{dt} = e \frac{d}{dt}(\phi_0 \sin \omega t) = e\phi_0 \frac{d}{dt}(\sin \omega t)$$

$$\text{This energy is gained in a time } \frac{T}{4} = \frac{\pi}{2\omega}.$$

$$\text{Average value of energy per turn} = \frac{e\phi_0}{\pi/2\omega} \int_0^{\pi/2} \frac{d}{dt}(\sin \omega t) dt = \frac{2e\omega\phi_0}{\pi}$$

Substituting the value of $\phi_0 = 2\pi r^2 B$ from equ.(23.35), we get

$$\text{Average energy per turn} = \frac{2e\omega}{\pi} \times 2\pi r^2 B = 4e\omega r^2 B$$

Total (final) energy, $E = \text{Number of revolutions made} \times \text{Average energy per revolution}$

$$= \frac{c}{4\omega r} \times 4e\omega r^2 B = cer B$$

Example 23.11. In a certain betatron the maximum magnetic field at the electron orbit is 0.5 wb/m^2 . The diameter of the stable orbit is 1.5 m . If the frequency of the alternating current through electromagnet coils is 59 Hz , what is the (i) final energy gained by the electrons and (ii) the number of revolutions taken by the electrons? (R.G.P.V.-2007)

$$\text{Solution. Final energy } E = B eR c = \frac{0.5 \text{ wb/m}^2 \times 1.6 \times 10^{-19} \text{ C} \times 0.75 \text{ m} \times 3 \times 10^8 \text{ m/s}}{1.6 \times 10^{-19} \text{ J/eV}}$$

$$= 112.5 \text{ MeV.}$$

$$\text{Number of revolutions } N = \frac{c}{4\omega r} = \frac{3 \times 10^8 \text{ m/s}}{4(2\pi \times 59 \text{ Hz}) \times 0.75 \text{ m}} = 2.7 \times 10^5.$$

Example 23.12. A betatron working on an operating frequency of 60 Hz has a stable orbit of diameter 1.6 m . Find the energy gained per turn as also the final energy if the magnetic field at the orbit is 0.5 T .

Solution. Average energy per turn = $4 e \omega r^2 B$

$$= \frac{4(1.602 \times 10^{-19} \text{ C}) \times 2\pi \times 60 \times (0.8 \text{ m})^2 \times 0.5 \text{ T}}{1.602 \times 10^{-19} \text{ J/eV}} = 482.6 \text{ eV.}$$

$$\text{Final energy} = E = B eR c = \frac{0.5 \text{ T} \times 1.602 \times 10^{-19} \text{ C} \times 0.8 \text{ m} \times 3 \times 10^8 \text{ m/s}}{1.602 \times 10^{-19} \text{ J/MeV}} = 120 \text{ MeV.}$$

Example 23.13. In a betatron, the magnetic flux at a stable orbit changes at a constant rate of 15 Wb/s . What would be the energy of an electron which undergoes 10^6 revolutions?

Solution. The increase in electron energy per revolution = $e \frac{d\phi}{dt} = 15 \text{ eV}$

Final energy of electron after 10^6 revolutions = $15 \text{ eV} \times 10^6 = 15 \text{ MeV}$.

23.18 ELECTRON SYNCHROTRON

The electron synchrotron is based on the principle of the combined working of betatron and cyclotron. In the electron synchrotron, the electrons are first accelerated by using the action of the betatron to energy of about 2 MeV. Then they have a velocity of $0.98 c$. Subsequently, the electrons travel at practically constant speed, but increase in mass as energy is imparted to them. For an electron traveling with an angular velocity ω in a circular orbit of radius r

$$m\omega^2 r = Be \omega r$$

or

$$\omega = \frac{Be}{m} \quad (23.40)$$

where B is the magnetic flux density at the orbit. If ω is to remain constant, B must increase in the same ratio as ' m '. In order to keep the electrons in a stable orbit, a small magnet is used inside the dough-nut tube. The magnet is less massive as the acceleration of the electrons beyond 2 MeV energy is achieved by radio-frequency (r.f.) electric field. This r.f. electric field is obtained between silver electrodes deposited over a short length along the arc of the tube (see Fig. 23.19). The silver coating has a short gap in it across which the output of an r.f. oscillator is connected. The frequency of the r.f. supply is adjusted to be equal to the time of one revolution of the electron in the circular orbit. Thus, the electrons are accelerated each time they cross the gap and gain additional energy. The r.f. supply is kept on while the magnetic flux is increasing and is automatically cut off when the electrons attain the required energy.

Maximum energy

In an electron synchrotron the maximum energy of electrons depends upon the radius r of the orbit and on maximum magnetic field strength B and is given by

$$E = Berc \quad (23.41)$$

or

$$E = \frac{(3 \times 10^8 \text{ m/s})(1.602 \times 10^{-19} \text{ C})}{1.602 \times 10^{-13}} rB \text{ MeV} = 300 rB \text{ MeV} \quad (23.42)$$

Frequency of the r.f. electric field

As the synchrotron acceleration starts when the velocity of electrons are very close to c , the frequency is given by

$$v = \frac{c}{2\pi r} = \frac{4.7}{r} \text{ MHz} \quad (23.43)$$

The r.f. accelerating electric field must have a frequency equal to the above frequency.

Electrons can be accelerated upto 1 BeV using the electron-synchrotron.

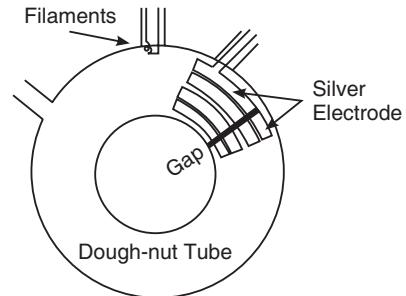


Fig. 23.19

23.19 PROTON SYNCHROTRON

The principle of proton synchrotron is essentially the same as that of the electron synchrotron. A proton synchrotron consists of a doughnut shaped vacuum chamber which has the form of a race-track (Fig. 23.20).

It is made of stainless steel, porcelain or plastic supported in the gap of an annular magnet. The annular magnet is made of four quadrants separated by straight gaps. The straight sections are free from magnetic field and are used for injecting, accelerating and ejecting the protons. Thus, the synchrotron consists of four sections joined up by arc shaped segments.

The protons are first accelerated up to 10 MeV in a Van de Graff accelerator and then fed into the synchrotron. Protons at low energy are injected in periodic pulses into the orbit. The protons are made to go in circular orbit by the magnetic field. They are accelerated once in each revolution when they pass between the electrodes connected to an r.f. oscillator. The magnet is excited periodically from 300 gauss to 15,000 gauss and the protons are accelerated during the time the magnetic field is increasing. Simultaneous control over the variation of magnetic field strength and the frequency of the r.f. oscillator is maintained in such a way that the protons travel in an orbit of constant radius and arrive at the electrodes when the applied r.f. voltage is in phase of acceleration.

When the protons have attained maximum energy, the frequency is distorted so that the orbit does not remain stable. By a suitable adjustment, the high energy protons are allowed to strike the target.

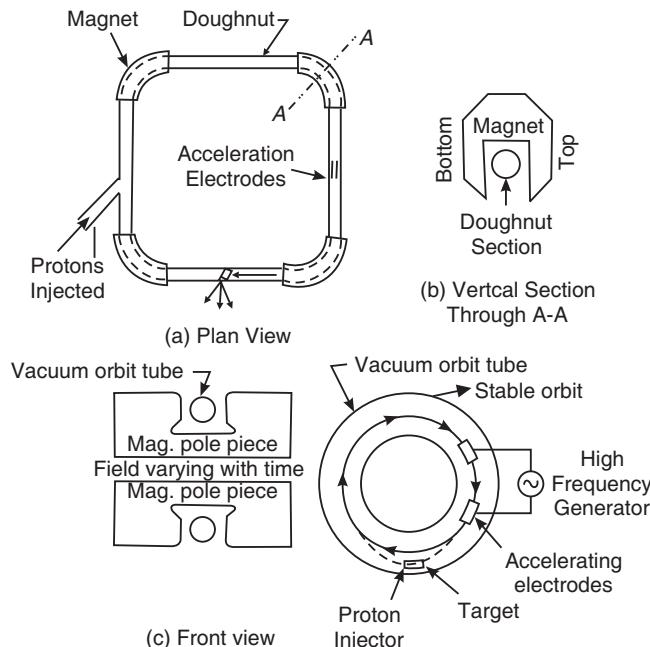


Fig. 23.20

QUESTIONS

1. Describe the construction and working of Geiger-Muller counter. What are its limitations?
2. Describe the principle, construction and working of G-M counter. (Shivaji Univ.)
3. Describe the construction and working of G-M counter. (RGPV,2007)
4. Explain the construction and working of Wilson's cloud chamber. What are the limitations of this apparatus?
5. Describe the principle and working of a bubble chamber.
6. Describe the principle, construction and working of a Wilson Cloud Chamber. (Shivaji Univ.)
7. What is the basic principle involved in the working of Wilson cloud chamber? (Bombay Univ.)
8. Describe the construction and working of a scintillation counter.
9. Write a short note on: (i) solid state detector (ii) Cerenkov detector

10. Give the construction and theory of Aston's mass spectrograph. Discuss the experimental results obtained.
11. What is mass spectrograph? Explain construction and working of Bainbridge mass spectrograph. How the relative proportions of isotopes are determined? **(C.S.V.T.U., 2009)**
12. Explain construction and working of Aston's Mass Spectrograph. Discuss the condition of focusing of ion beam. **(RGPV, 2008)**
13. Explain the construction and theory of Dempster's mass spectrograph.
14. What are isotopes? Explain the principle of their detection, in case of Bainbridge mass spectrograph. **(Amaravati Univ., 2006)**
15. Explain with a neat diagram the principle, construction and working of Bainbridge's Mass spectrograph. **(Univ. of Pune, 2008)**
16. Explain the function of velocity selector in Bainbridge mass spectrograph. Obtain an expression for linear separation in it.
17. Describe the construction and working of Bainbridge mass spectrograph and show that it has linear mass scale. **(Amaravati Univ., 2008)**
18. Explain the working of velocity selector. **(R.T.M.N.U., 2006)**
19. What is a mass spectrograph? Explain the working of velocity selector arrangement in Bainbridge mass spectrograph. Show that the mass scale is linear. **(R.T.M.N.U., 2005)**
20. Describe construction, principle and working of Bainbridge mass spectrograph using a well labeled diagram. **(C.S.V.T.U., 2006)**
21. What are particle accelerators? Explain construction and working of Linac and cyclotron and the difference between them. **(RGPV, 2007)**
22. Describe the theory, construction and working of cyclotron. What are the limitations of a cyclotron? **(C.S.V.T.U., 2007)**
23. In a cyclotron, show that the time spent by the particle in a dee is independent of its speed and radius of its circular path. Discuss whether a cyclotron can be used to accelerate electrons.
24. Explain the principle and working of cyclotron. Obtain necessary equation for resonance frequency. **(Amaravati Univ., 2006)**
25. Explain the principle and working of cyclotron. Show that all the particles of the same mass and charge take the same time 't' to describe a semicircle and 't' is independent of the radius of the path and velocity of the particles.
26. With the help of neat diagram explain the principle, construction and working of cyclotron. Obtain the expression for the cyclotron frequency and maximum energy of the particle. **(Univ. of Pune, 2007), (R.T.M.N.U., 2006)**
27. Explain construction and working of cyclotron. Also find out the expression for frequency of cyclotron. **(Amaravati Univ., 2004, 2005, 2006, 2008)**
28. What is cyclotron ? State resonance condition. Why electrons cannot be accelerated to high energies in cyclotron? **(R.T.M.N.U., 2010)**
29. What is the primary function of electric and magnetic fields in a cyclotron?
30. Describe construction and working of a cyclotron with a neat diagram. Derive and explain the resonance condition. Find the energy acquired by a charged particle after 'N' complete rounds.
31. How do electric and magnetic fields operate in a cyclotron? Why electrons cannot be accelerated to high energies in a cyclotron?
32. Discuss the construction and working of a cyclotron.
33. Give the construction and working of the following:
 - (i) Bainbridge mass spectrograph
 - (ii) Cyclotron **(RGPV, 2007), (R.T.M.N.U., 2006)**
34. What is a synchrocyclotron? How does it work?
35. Discuss betatron condition. How does it help in maintaining circular orbit?
36. Explain construction, principle and working of betatron and what is betatron condition? **(R.G.P.V.-2007)**

37. Describe the construction of a betatron. How is energy gained in it? Obtain betatron condition. How is it achieved?
38. Give the construction and working of betatron with neat diagram. Obtain the betatron condition. **(Univ. of Pune, 2007, 2008)**
39. Explain the working of Betatron. Obtain the betatron condition. **(Shivaji Univ.)**
40. Give the principle and working of a Betatron. What is Betatron condition? **(RGPV, 2008)**
41. Give the principle and working of a synchrotron. Explain how the magnetic field is modulated in an electron-synchrotron.
42. Give the principle and working of proton synchrotron.

PROBLEMS

- Calculate the frequency of ac potential that must be applied to cyclotron dees in which protons are accelerated. Given magnetic flux density = $3\text{wb}/\text{m}^2$. **[Ans: f = 45.8 MHz]**
- The voltage across the dees of a cyclotron is 50 kV. How many revolutions do protons make to reach a kinetic energy of 20 MeV? **[Ans: 200]**
- What strength of magnetic field is used in a cyclotron in which protons make 1.9×10^7 revolutions per second? **[Ans: 1.2 T]**
- A cyclotron with dee radius 40 cm is to be used to accelerate protons using an oscillator of frequency 8 MHz. Calculate the magnetic field needed to maintain resonance and the final energy of the protons. **[Ans: 0.52 T, 2.07 MeV]**
- Deuterons are accelerated in a cyclotron with dees of radius 0.3 m and magnetic field of flux density $0.7 \text{ Wb}/\text{m}^2$. Calculate (i) the velocity of deuterons (ii) the energy and (iii) the frequency of the field between the dees. **[Ans: 107 m/s, 1.08 MeV, 3.3 MHz]**
- A cyclotron accelerating protons employs a magnetic flux density $0.6 \text{ wb}/\text{m}^2$. How rapidly should the electric field between the metal dees be reversed? Also obtain the energy of the emerging proton. **[Ans: $1.1 \times 10^{-7}\text{s}$, 1.08 MeV]**
- A proton is accelerated by cyclotron and acquires energy of 1.0 MeV. The voltage between dees is 1000 volts. How many rotations the proton shall have to make to achieve this energy? **[Ans: 500]**
- Singly ionized Ne^{20} and Ne^{22} ions enter a Bainbridge mass spectrograph with a velocity of 10^5 m/s . If they are deflected by a magnetic field of flux density $0.08 \text{ Wb}/\text{m}^2$, calculate the radii of their paths. **[Ans: 0.26 m, 0.29 m]**
- Singly charged neon ions with mass numbers 20 and 22 and having kinetic energy $6.2 \times 10^{-16}\text{J}$ come into a uniform magnetic field of induction 0.24 T in vacuum at right angles to the induction lines and traveled in a semicircle separate into two beams. What is the separation between the beams? **[Ans: 33 mm]**
- The element tin is being analyzed in a Bainbridge mass spectrometer. Among the isotopes present are those of masses 116, 117, 118, 119 and 120 u. The electric and magnetic fields are $E = 20\text{kV/m}$ and $B = 0.25 \text{ T}$. What is the spacing between the marks produced on the photographic plate by the ions of tin-116 and tin-120? **[Ans: 27 mm]**
- In a betatron, the maximum magnetic field $0.4 \text{ wb}/\text{m}^2$ is operating at 50 Hz. The diameter of the stable orbit is 1.524 m. Determine the final energy. **[Ans: 91.1 MeV]**
- A betatron of 100 MeV energy has a stable radius of 0.84 m. Calculate (i) the value of magnetic field intensity at the orbit for this energy and (ii) the frequency of the applied electric field if average energy gain per turn is 420 eV. **[Ans: 0.4 T, 60 Hz]**
- A G.M. counter wire collects 10^8 electrons per discharge when the counting rate is 500 counts/min. What will be the average current in the circuit? **(RGPV, 2010)**
- In a certain cyclotron, the maximum radius that the path of a deuteron may have before it is deflected out of the magnetic field is 20 cm. Calculate the velocity of the deuteron at this radius.
[Given: deuteron charge $q = 1.6 \times 10^{-19} \text{ C}$ and mass = $3.34 \times 10^{-27} \text{ kg.}$] **(RGPV, 2010)**

CHAPTER

24

Lasers

24.1 INTRODUCTION

Laser is one of the outstanding inventions of the 20th century. The word ‘LASER’ is the acronym for Light Amplification through Stimulated Emission of Radiation. However, laser is not a simple amplifier of light but is a generator of light. It is an artificial light source that differs vastly from the traditional light sources. Laser is more akin to radio and microwave transmitters and produces a highly directional coherent monochromatic polarized light beam. Einstein gave the theoretical basis for the development of laser in 1916, when he predicted the possibility of stimulated emission. In 1954, C.H.Townes and his co-workers put Einstein’s prediction for practical realization. They developed a microwave amplifier based on stimulated emission of radiation. It was called a **maser**. Shortly thereafter, in 1958, A.Schawlow and C.H.Townes extended the principle of masers to light and T.H.Maiman built the first laser device in 1960. In 1961, A.Javan and associates developed the first gas laser, the helium-neon laser. Laser is a high technology device and is the most sought after tool in a wide variety of fields such as metalworking, entertainment, communications, surgery, ophthalmology and weapon guidance in wars.

24.2 INTERACTION OF LIGHT WITH MATTER AND THE THREE QUANTUM PROCESSES

It is familiar to us that when light travels through a medium, it undergoes absorption and scattering processes. Light absorption means the transfer of energy from light to atoms and light scattering involves change in the direction of travel of waves. As a result of these two processes, light intensity decreases with distance in the medium. Transfer of energy from atom to light is not conceivable from classical point of view. However, it is found to be possible when the interaction of light with medium is considered from the point of view of quantum mechanics. The transfer of energy from atom to light results in **light amplification**. A light amplifier can be further converted into a source of light having superior characteristics compared to traditional light sources. A **laser is a monochromatic coherent light source** that depends on quantum processes for its operation. It is therefore necessary to first understand the quantum processes in order to understand the operation of a laser.

Let us consider a material medium, which is composed of *identical atoms*. Atoms are characterized by many energy levels but for the sake of simplicity to understand, let us

assume that the atoms of the material medium under consideration be characterized by only two energy levels, namely energy level E_1 and energy level E_2 . E_1 is the ground state while E_2 is the excited state. As the atoms of the material are identical, majority of them occupy the energy level E_1 and the others the energy level E_2 . The number of atoms per unit volume at an energy level is called the **population density**. Let the populations at the two energy levels E_1 and E_2 be N_1 and N_2 respectively. Under normal conditions higher the energy, lesser is its population. Hence, $N_1 > N_2$. Now, let light radiation be incident on the material and we assume that the radiation and the medium are in thermal equilibrium. The incident radiation may be viewed as a stream of photons, and let the photon density be $\rho(v)$. Let each photon carry an energy E , where $E = E_2 - E_1 = hv$. When photons travel through the medium, they are likely to cause three different processes. They are absorption, spontaneous emission and stimulated emission.

24.2.1 Absorption

Suppose an atom is in the lower energy level E_1 . If a photon of energy $(E_2 - E_1)$ is incident on the atom, it imparts its energy to the atom and disappears. Then we say that the atom absorbed the incident photon. As a result of absorption of adequate energy, the atom jumps to the excited state E_2 (Fig. 24.1). Such a transition is called an **absorption transition**.

In each absorption transition event, an atom in the medium is excited and one photon is subtracted from the incident light beam. As the absorption process is induced by a photon, it is also called as **induced absorption**. We may express the process as

$$A + hv = A^*$$

where A is an atom in the lower state and A^* is an excited atom.

The probability that an absorption transition occurs is proportional to the photon density $\rho(v)$.

$$P_{12} \propto \rho(v)$$

or

$$P_{12} = B_{12}\rho(v) \quad (24.1)$$

where B_{12} is the constant of proportionality. B_{12} is known as the *Einstein coefficient for induced absorption*. It is a constant characteristic of the atom and represents the properties of the energy states E_1 and E_2 .

The number of absorption transitions occurring in the material at any instant will be equal to the product of the number of atoms at the energy level E_1 and the probability P_{12} for the absorption transition. Thus, the number of atoms, N_{ab} , excited during the time Δt is

$$N_{ab} = B_{12}N_1\rho(v)\Delta t \quad (24.2)$$

where N_1 is the population of atoms at E_1 . When the atoms are more at the lower energy level, then more atoms can jump into the excited state. Similarly, when more photons are incident on the assembly of atoms, then more atoms can get excited to the higher energy level. Induced absorption involves the excitation of the atom to the fixed higher level only. As a result of this absorption, N_1 decreases while N_2 increases. But under normal conditions N_2 cannot be greater than N_1 .

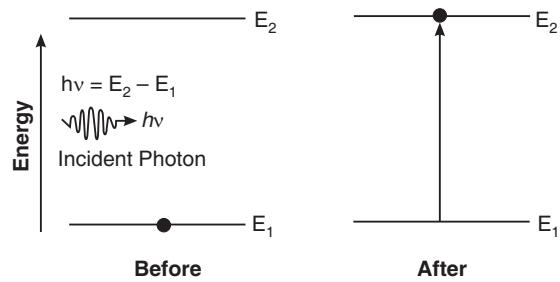


Fig. 24.1. Absorption Process

As the number of atoms is greater in the lower energy state, the absorption process leads to attenuation of radiation as light travels through the medium.

24.2.2 Spontaneous Emission

When an atom at lower energy level is excited to a higher energy level, it cannot stay in the excited state for a relatively longer time. In a time of about 10^{-8} s, the atom reverts to the lower energy state by releasing a photon of energy $h\nu$ where $h\nu = E_2 - E_1$. The emission of photon occurs *on its own* and without any external impetus given to the excited atom (Fig. 24.2). Emission of a photon by an atom *without any external impetus* is called **spontaneous emission**. We may write the process as



The probability that a spontaneous transition occurs depends only on the properties of energy states E_2 and E_1 and is independent of the photon density. It is equal to the lifetime of level E_2 . Thus,

$$(P_{21})_{\text{Spont.}} = A_{21} \quad (24.3)$$

where A_{21} is a constant and known as the *Einstein coefficient for spontaneous emission*. A_{21} is a constant characteristic of the atom. $1/A_{21}$ is a measure of the lifetime of the upper state against spontaneous transition to the lower state.

The number of spontaneous transitions, N_{sp} , taking place during the time Δt depends *only* on the number of atoms N_2 staying at the excited state E_2 . Thus,

$$N_{\text{sp}} = A_{21} N_2 \Delta t \quad (24.4)$$

It is the process of spontaneous emission that dominates in conventional light sources.

The process of spontaneous emission is essentially probabilistic, that is, the atom has some probability for making the transition, and it is not amenable for control from outside. The instant of the transition, direction of emission of photon, the phase of the photon, the polarization state of the photon are all random quantities. There will not exist any correlation among the parameters of the innumerable photons emitted spontaneously by the assembly of atoms of the medium. Therefore, the light generated by the medium will be **incoherent**. The light from the conventional sources originates in spontaneous emission process and is **incoherent**. It contains a superposition of many waves of random phases. The net intensity of such incoherent waves is proportional to the number of radiating atoms. The light is radiated in the form of short duration wave trains emitted in all directions and the intensity goes on decreasing as the wave trains travel away from the source. Each of them bears no consistent phase relationship with each other nor do they share a common polarization plane. As a consequence, there is no compounding of the individual waves. The light is also not monochromatic because of various line broadening processes that take place in the medium.

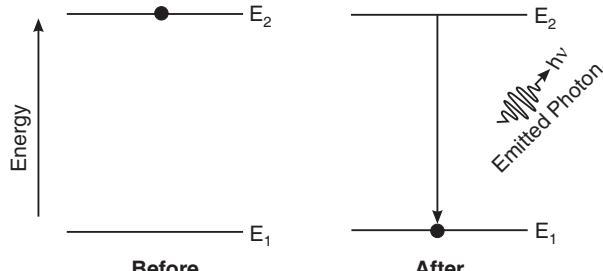
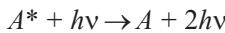


Fig. 24.2. Spontaneous emission process

24.2.3 Stimulated Emission

In 1916, Einstein showed the existence of equilibrium between matter and radiation required a new radiation process called stimulated radiation. It requires the presence of external radiation. If an atom in the excited state interacts with a photon with energy $h\nu = E_2 - E_1$, the photon induces the excited atom to make a downward transition well before the atom can make a spontaneous transition. The atom emits the excess energy in the form of a photon, $h\nu = E_2 - E_1$, as it drops to the lower energy state. The passing photon is not affected while the excited atom emits a photon (Fig. 24.3). The phenomenon of forced photon emission by an excited atom due to the action of an external agency is called **stimulated emission**. It is also known as *induced emission*. The process may be expressed as



The probability that a stimulated transition occurs is given by

$$(P_{21})_{\text{stimulated}} \propto \rho(v)$$

or

$$(P_{21})_{\text{stimulated}} = B_{21}\rho(v) \quad (24.5)$$

where B_{21} is the constant of proportionality and is known as the *Einstein coefficient for stimulated emission*. It is a constant characteristic of the atom and represents the properties of the energy states E_1 and E_2 .

The number of stimulated transitions occurring in the material at any instant will be equal to the product of the number of atoms at the energy level E_2 and the probability P_{21} for the stimulated transition. Thus, the number of atoms, N_{st} , that undergo downward transition during the time Δt is

$$N_{st} = B_{21}N_2\rho(v)\Delta t \quad (24.6)$$

Multiplication of Stimulated Photons

The photon induced in this process propagates in the same direction as that of stimulating photon. The induced photon has features identical to that of the inducing photon. It has the same frequency, phase and plane of polarization as that of the stimulating photon. The outstanding feature of this process is the *multiplication of photons*. For one photon interacting with an excited atom, there are two photons emerging. The two photons travelling in the same direction interact with two more excited atoms and generate two more photons and produce a total of four photons. These four photons in turn stimulate four excited atoms and generate eight photons, and so on. The number of photons builds up in an avalanche like manner, as shown in Fig. 24.4.

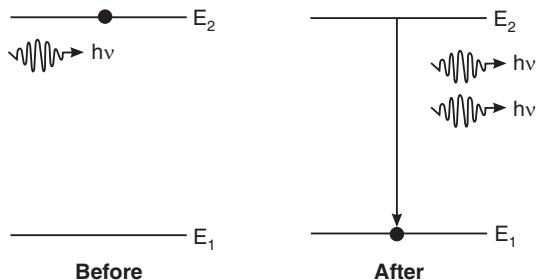


Fig. 24.3. Process of Stimulated Emission.

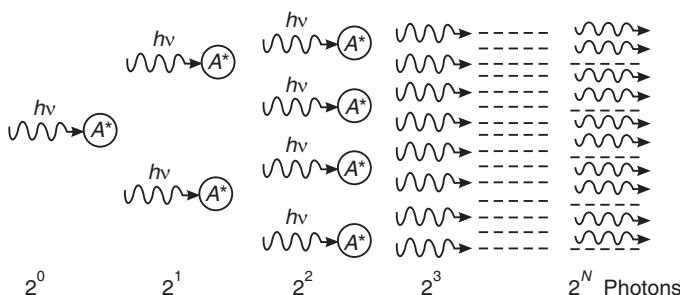


Fig. 24.4. Multiplication of Stimulated photons into an avalanche

All the light waves generated in the medium are due to one initial wave and all of the waves are in phase. Thus, the waves are coherent and *interfere constructively*. The net intensity of light will be proportional to the square of the number of atoms radiating light. Thus,

$$I_{\text{total}} = N^2 I$$

Since the number of atoms in the material medium is very large, coherent emission leads to an enormously high intense light and we say that the incident *light is amplified*. Therefore, the process of stimulated emission is the key to the operation of a laser.

24.2.4 Distinction between Spontaneous and Stimulated Emission

Sl. No.	Spontaneous emission	Stimulated emission
1.	Spontaneous emission is a random and probabilistic process.	Not a random process.
2.	Not amenable for control from outside.	Amenable for control from outside.
3.	The photons are emitted haphazardly. The instant of emission, direction of emission, phase, polarization state of photon are all random quantities and cannot be controlled.	The stimulating photon imposes its characteristics on the photon emitted.
4.	Photons are emitted uniformly in all directions from an assembly of atoms. As a result, the light is non-directional .	The photons emitted in the process travel in the same direction as that of stimulating photon. The light produced by the process is essentially directional .
5.	Photons of slightly different frequencies are generated. As a result, the light is not monochromatic .	The spread of photon frequencies is relatively very narrow. As such the light is nearly monochromatic .
6.	Photons do not have any correlation in their phases, which fluctuate randomly. Therefore, the light produced by this process is incoherent .	The photons emitted by this process are all in phase and therefore, the light is coherent .
7.	In this process multiplication of photons does not take place. Hence there is no amplification of light due to the process.	One stimulating photon causes emission of two more photons. These two produce four photons, which in turn generate eight photons and so on. Thus, if there are N excited atoms, 2^N photons will be produced. Light amplification occurs due to such multiplication of photons.
8.	The net intensity of the generated light is given by $I_T = N I$ where N is the number of atoms emitting photons and I is the intensity of each photon.	As all the photons are in phase, they constructively interfere and produce an intensity $I_T = N^2 I$
9.	The planes of polarization of the photons are oriented randomly. Hence, light from the source is unpolarized .	The planes of polarization are identical for all photons. Consequently, light is polarized .

24.2.5 Steady State

The three processes described above occur simultaneously (Fig. 24.5). Under steady state condition the absorption and emission balance each other. Thus,

$$N_{\text{absorption}} = N_{\text{spont.emission}} + N_{\text{stim.emission}} \quad (24.7)$$

$$B_{12} N_1 \rho(v) \Delta t = A_{21} N_2 \Delta t + B_{21} N_2 \rho(v) \Delta t$$

$$B_{12} N_1 \rho(v) = A_{21} N_2 + B_{21} N_2 \rho(v) \quad (24.8)$$

If we consider a medium in thermal equilibrium, there would be more atoms in the lower level than at higher level. That is $N_1 \gg N_2$. As the probability for absorption transition is equal to the probability for stimulated transition, a photon traveling through the medium is *more likely to get absorbed* than to stimulate an excited atom to emit a photon. Therefore, usually the process of absorption dominates the process of stimulated emission. Similarly, an atom that is at the excited state is more likely to jump to the lower level on its own than being stimulated by a photon. It is due to the fact that the photon density in the incident beam is not sufficient to interact with the excited atoms; and the photons instead interact with atoms at lower level because of the large population available at that level. Owing to this, the spontaneous emission dominates the stimulated emission.

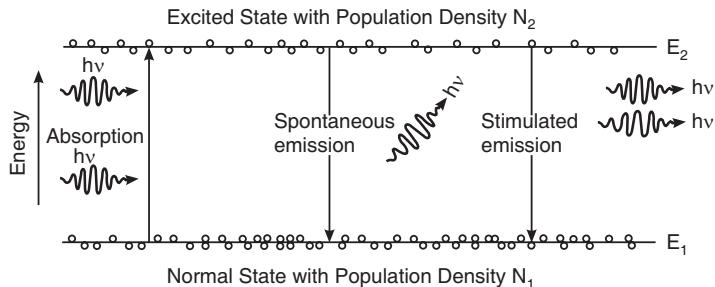


Fig. 24.5. Absorption and Emission Processes in steady state

24.3 EINSTEIN COEFFICIENTS AND THEIR RELATIONS

24.3.1 Einstein Coefficients

We summarize here the three Einstein coefficients, which are the proportionality constants introduced in the above discussions.

(i) The probability that an absorption transition occurs is given by

$$P_{12} = B_{12}\rho(v)$$

where B_{12} is the constant of proportionality known as the *Einstein coefficient for induced absorption*. It is a constant characteristic of the atom and represents the properties of the energy states E_1 and E_2 .

(ii) The probability that a spontaneous transition occurs is given by

$$(P_{21})_{\text{Spontaneous}} = A_{21}$$

where A_{21} is a constant known as the *Einstein coefficient for spontaneous emission*. A_{21} is a constant characteristic of the atom and is known as the *radiative rate* measured in units of s^{-1} . $1/A_{21}$ is the lifetime of the upper state against spontaneous decay to the lower state.

(iii) The probability that a stimulated transition occurs is given by

$$(P_{21})_{\text{stimulated}} = B_{21}\rho(v)$$

where B_{21} is the constant of proportionality known as the *Einstein coefficient for stimulated emission*. It is a constant characteristic of the atom and represents the properties of the energy states E_1 and E_2 .

We may note the following points here regarding the Einstein coefficients.

- The coefficients indicated by B are related to the induced transitions, i.e., transitions induced by external photons. Thus, B_{12} represents the transition induced by a photon from lower energy level E_1 to the higher energy level E_2 , whereas B_{21} denotes the transition induced by a photon from higher energy level E_2 to the lower energy

level E_1 . It turns out that B_{12} and B_{21} are equal under the special condition that the quantum states E_1 and E_2 are single energy levels (i.e., nondegenerate levels).

- The coefficient indicated by A is related to the spontaneous transition, i.e., transition occurred on its own without the assistance of external agent. Since a spontaneous transition cannot take place from lower energy state E_1 to the higher energy state E_2 , we do not have the coefficient A_{12} . In other words, $A_{12} = 0$.

24.3.2 Relation between the Einstein Coefficients

The Einstein coefficients A_{21} , B_{12} and B_{21} are interrelated. To find out the relation, we assume that

- (i) The atoms and the radiation are in thermal equilibrium.
- (ii) The radiation is identical with black body radiation and consistent with Planck's radiation law for any value of T .
- (iii) The population densities N_1 and N_2 at the lower and upper energy levels respectively are constant in time and are distributed according to Boltzmann law in the energy levels.

The above conditions require that the rate of change of atoms at the level E_2 must equal to zero. It means that the number of transitions from E_2 to E_1 must be equal to the number of transitions from E_1 to E_2 (see Fig. 24.5).

Thus,

$$\left. \begin{array}{l} \text{The number of atoms absorbing} \\ \text{photons per second per unit volume} \end{array} \right\} = \left. \begin{array}{l} \text{The number of atoms emitting} \\ \text{photons per second per unit volume} \end{array} \right\}$$

$$\left. \begin{array}{l} \text{The number of atoms absorbing} \\ \text{photons per second per unit volume} \end{array} \right\} = B_{12} \rho(v) N_1$$

$$\left. \begin{array}{l} \text{The number of atoms emitting} \\ \text{photons per second per unit volume} \end{array} \right\} = A_{21} N_2 + B_{21} \rho(v) N_2$$

As the number of transitions from E_1 to E_2 must equal the number of transitions from E_2 to E_1 , we have

$$B_{12} \rho(v) N_1 = A_{21} N_2 + B_{21} \rho(v) N_2 \quad (24.9)$$

$$\rho(v) [B_{12} N_1 - B_{21} N_2] = A_{21} N_2$$

$$\therefore \rho(v) = \frac{A_{21} N_2}{[B_{12} N_1 - B_{21} N_2]} \quad (24.10)$$

By dividing both the numerator and denominator on the right hand side of the above equation with $B_{12} N_2$, we obtain

$$\rho(v) = \frac{A_{21} / B_{12}}{\left[\frac{N_1}{N_2} - \frac{B_{21}}{B_{12}} \right]} \quad (24.11)$$

But

$$\frac{N_2}{N_1} = e^{-(E_2 - E_1)/kT}$$

As $E_2 - E_1 = hv$,

$$\frac{N_2}{N_1} = e^{-hv/kT} \quad \text{or} \quad \frac{N_1}{N_2} = e^{hv/kT}$$

$$\therefore \rho(v) = \frac{A_{21}}{B_{12}} \left[\frac{1}{e^{hv/kT} - B_{21}/B_{12}} \right] \quad (24.12)$$

To maintain thermal equilibrium, the system must release energy in the form of electromagnetic radiation. It is required that the radiation be identical with black body radiation and be consistent with Planck's radiation law for any value of T . According to Planck's law

$$\rho(v) = \left(\frac{8\pi h v^3 \mu^3}{c^3} \right) \left[\frac{1}{e^{hv/kT} - 1} \right] \quad (24.13)$$

where μ is the refractive index of the medium and c is the velocity of light in free space.

Energy density $\rho(v)$ given by equ.(24.12) will be consistent with Planck's law (24.13), only if

$$\frac{A_{21}}{B_{12}} = \frac{8\pi h v^3 \mu^3}{c^3} \quad (24.14)$$

and

$$\frac{B_{21}}{B_{12}} = 1 \quad \text{or} \quad B_{12} = B_{21} \quad (24.15)$$

The above equations (24.14) and (24.15) are known as the **Einstein relations**. It follows that the coefficients are related through

$$B_{12} = B_{21} = \frac{c^3}{8\pi h v^3 \mu^3} A_{21} \quad (24.16)$$

(i) The relation (24.15) shows that the coefficients for both absorption and stimulated emission are numerically equal. The equality implies that when an atom with two energy levels is placed in the radiation field, the probability for an upward (absorption) transition is equal to the probability for a downward (stimulated) transition.

(ii) The relation (24.14) shows that the ratio of coefficients of spontaneous versus stimulated emission is proportional to the third power of frequency of the radiation. This is why it is difficult to achieve laser action in higher frequency ranges such as X-rays.

24.4 LIGHT AMPLIFICATION

Light amplification requires that stimulated emission occur almost exclusively. In practice, absorption and spontaneous emission always occur together with stimulated emission. The laser operation is achieved when stimulated emission exceeds in a large way the other two processes. Let us now look at the conditions under which the number of stimulated transitions can be made larger than the other two transitions.

24.4.1 Condition for Stimulated Emission to Dominate Spontaneous Emission

The ratio of equ.(24.6) to equ. (24.4) gives

$$\frac{\text{Stimulated transitions}}{\text{Spontaneous transitions}} = \frac{B_{21} N_2 \rho(v)}{A_{21} N_2} = \frac{B_{21}}{A_{21}} \rho(v) \quad (24.17)$$

Equ.(24.17) indicates that stimulated transitions will dominate the spontaneous transitions if the radiation density $\rho(v)$ is very large. Thus, the presence of a large number of photons in

the active medium is required. However, it will lead to more absorption transitions. Hence, large photon density alone will not guarantee more stimulated emissions.

24.4.2 Requirement of States of Larger Lifetimes

Equ.(24.17) further indicates that stimulated transitions will dominate the spontaneous transitions if the value of the ratio B_{21}/A_{21} is also large. To increase the probability of stimulated emissions, the lifetime of atoms at the excited state should be larger. In other words, it is necessary that the excited state has a longer lifetime (remember that $1/A_{21}$ represents the lifetime of the excited state).

24.4.3 Condition for Stimulated Emission to Dominate Absorption Transitions

The ratio of equ.(24.6) to equ. (24.2) yields

$$\frac{\text{Stimulated transition}}{\text{Absorption transition}} = \frac{B_{21}N_2\rho(v)}{B_{12}N_1\rho(v)} = \frac{N_2}{N_1} \quad (24.18)$$

We used here the fact that $B_{12} = B_{21}$.

The above condition indicates that the stimulated transitions will overwhelm the absorption process if N_2 is greater than N_1 . It means that there should be more atoms present in the higher energy level than in the lower energy level for stimulated emissions to dominate over the spontaneous emissions.

A medium amplifies light only when the above **three conditions** are fulfilled. Therefore to achieve high percentage of stimulated emissions, an artificial situation known as *population inversion* is to be created in the medium.

24.5 MEETING THE THREE REQUIREMENTS

24.5.1 Population Inversion

When the material is in thermal equilibrium condition, the population ratio is governed by the Boltzmann factor according to the following equation:

$$\frac{N_2}{N_1} = e^{-[E_2 - E_1]/kT} \quad (24.19)$$

It means that the population N_2 at the excited level E_2 will be far smaller than the population N_1 at the level E_1 . For example, if we take typical values for E_1 and E_2 , the population N_1 would be 10^{30} times of N_2 . The condition in which there are more atoms in the lower energy level and relatively lesser number of atoms in the higher energy level is called *normal state or equilibrium state* (Fig. 24.6 a). Thus, under thermal equilibrium condition, $N_1 \gg N_2$.

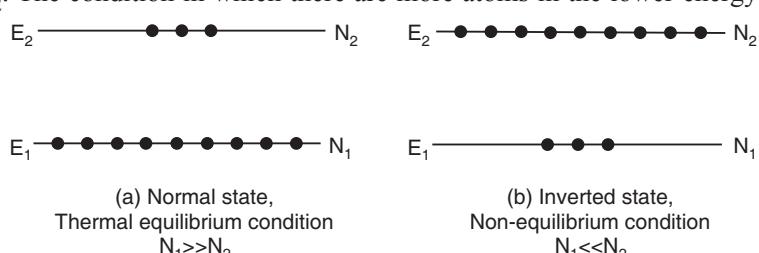


Fig. 24.6

Population inversion is the condition of the material in which population of the upper energy level N_2 far exceeds the population of the lower energy level, N_1 (Fig. 24.6 b). That is,

$$N_2 \gg N_1 \quad (24.20)$$

In this condition the population distribution between the levels E_1 and E_2 is inverted and hence it is known as the *inverted state*. This is a *non-equilibrium state* and exists only for a short time. Population inversion is obtained by employing *pumping techniques*, which transfer large number of atoms from lower energy level to higher energy level.

Example 24.1. A 10 mW He-Ne laser has efficiency of 1%. Assume that all input energy is utilized in pumping the atoms from the ground state to the excited state, which is 20 eV above the ground state. Find how many atoms are promoted to the excited state in one second.

Solution. Efficiency of laser = 1% = 0.01

$$\text{Power input} = \frac{\text{Power output}}{\text{Efficiency}} = \frac{10 \text{ mw}}{0.01} = 1 \text{ W.}$$

Therefore, energy input in one second = 1 J.

$$\text{Number of atoms excited in one second} = \frac{1 \text{ J}}{20 \text{ eV}} = \frac{1 \text{ J}}{20 \times 1.602 \times 10^{-19} \text{ J}} = 3.12 \times 10^{17}.$$

Example 24.2. Find the ratio of populations of the two states in a He-Ne laser that produces light of wavelength 6328 Å at 27°C.

Solution: The ratio of population is given by $\frac{N_2}{N_1} = e^{-(E_2 - E_1)/kT}$

$$E_2 - E_1 = \frac{12400}{6328} \text{ eV} = 1.96 \text{ eV}$$

$$\therefore \frac{N_2}{N_1} = \exp\left[\frac{-1.96 \text{ eV}}{(8.61 \times 10^{-5} \text{ eV})(300 \text{ K})}\right] = e^{-75.88} = 1.1 \times 10^{-33}$$

24.5.2 Metastable States

An atom can be excited to a higher level by supplying energy to it. Normally, excited atoms have short lifetimes and release their energy in a matter of nanoseconds (10^{-9} s) through spontaneous emission. It means that atoms do not stay long enough at the excited state to be stimulated. As a result, even though the pumping agent continuously raises the atoms to the excited level, they undergo spontaneous transition and rapidly return to the lower energy level. Population inversion cannot be established under such circumstances. In order to establish the condition of population inversion, the excited atoms are required to ‘wait’ at the upper energy level till a large number of atoms accumulate at that level. Such an opportunity would be provided by metastable states. Atoms excited to a **metastable state** remain excited for an appreciable time, which is of the order of 10^{-6} to 10^{-3} s. This is 10^3 to 10^6 times the lifetimes of the ordinary excited energy levels. Therefore, the metastable state allows accumulation of a large number of excited atoms at that level. The metastable state population can exceed the population at a lower level and establish the condition of population inversion in the lasing medium. It would be impossible to create the state of population inversion without a metastable state. Metastable state can be readily obtained in a crystal system containing impurity atoms. These levels lie in the forbidden band gap of the host crystal. Population inversion readily takes place as the lifetimes of these levels are large, and secondly, there is no competition in filling these levels, as they are localized levels. For example, **phosphorescent** materials are made up of atoms with metastable states.

There could be no population inversion and hence no laser action, if metastable states do not exist.

24.5.3 Confining Radiation within the Medium

According to equ.(24.17) a high radiation density $\rho(v)$ is required to be present in the active medium so that stimulated emission dominates spontaneous emission. If laser medium is enclosed in between a pair of optically plane parallel mirrors, photon density builds up to a very high value through repeated reflections of photons which remain within the medium. Such an arrangement is known as an **optical resonant cavity** or **optical resonator**.

24.6 COMPONENTS OF LASER

The essential components of a laser are (i) an active medium, (ii) a pumping agent and (iii) an optical resonator (see Fig. 24.7).

24.6.1 Active Medium

The **active medium** is the material in which the laser action takes place. The most important requirement for the laser medium is that we should be able to obtain population inversion in it. Atoms are in general characterized by a large number of energy levels. However, all types of atoms are not suitable for laser operation. Even in a medium consisting of different species of atoms, only a small fraction of atoms of a particular type have energy level system suitable for achieving population inversion. Such atoms can produce more stimulated emission than spontaneous emission and cause amplification of light. Those atoms, which cause laser action, are called **active centers**. The rest of the medium acts as host and supports active centers. The medium hosting the active centers is called the **active medium**. An active medium is a medium which when excited reaches the state of population inversion and promotes stimulated emissions leading to light amplification.

24.6.2 The Pump

For achieving and maintaining the condition of population inversion, we have to raise continuously the atoms in the lower energy level to the upper energy level. It requires energy to be supplied to the system. **Pumping** is the process of supplying energy. The **pump** is an external source that supplies energy needed to transfer the laser medium into the state of population inversion.

There are a number of techniques for pumping a collection of atoms to an inverted state. Optical pumping, electrical discharge and direct conversion are some of the methods of pumping. In optical pumping, a light source such as a flash discharge tube is used. This method is adopted in solid-state lasers. In electrical discharge method, the electric field causes ionization of the medium and raises it to the excited state. In semiconductor diode lasers, a direct conversion of electrical energy into light energy takes place.

24.6.3 Optical Resonator

Laser is a light source and it is analogous to an electronic oscillator. An electronic oscillator is essentially an amplifier supplied with a positive feed back. A part of the output of the amplifier is taken and fed back at its input. When the amplifier is switched on, electrical noise signal of appropriate frequency present at the input will be amplified; the output is fed back to the input and amplified again and so on. A stable output is quickly reached when the oscillator acts as a source of a particular frequency.

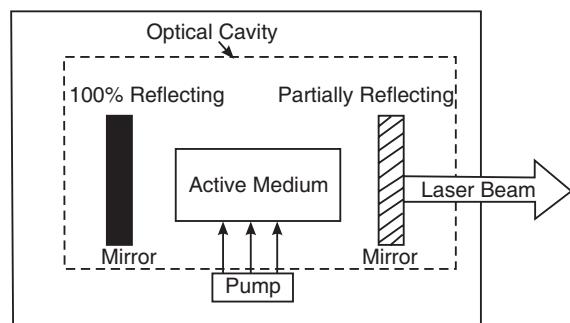


Fig. 24.7. Components of a Laser

In laser the active medium is the amplifier, which is converted into an oscillator through the feed back mechanism established by an optical resonator. A pair of optically plane parallel mirrors, enclosing laser medium in between them (Fig. 24.8), is known as an **optical resonant cavity**. One of these mirrors is partially reflecting and the other is made fully reflecting.

In laser, the role of noise is played by chance photons emitted spontaneously. The photons emitted along the optic axis of the resonant cavity travel through the medium and trigger stimulated emissions. They are reflected by the end mirror and reverse their path. The photons are thus fed back into the medium and travel toward the opposite end mirror causing more stimulated emissions. The photons are once more

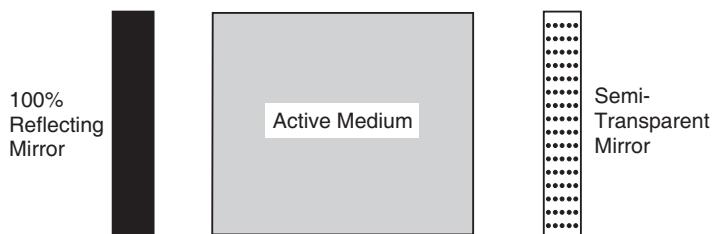


Fig. 24.8. Fabry-Perot Optical resonator

reflected at the mirror and travel toward the opposite mirror. Substantial light amplification takes place because the light beam is reflected several times at the mirrors and gains strength in each passage. Ultimately, when the amplification balances the losses in the cavity, the laser beam emerges out from the front – end mirror. *In the absence of resonator cavity, there would be no amplification of light.*

Role of the Optical Resonator

- The primary function of the optical resonator is to provide positive feed back of photons into the medium so that stimulated emission is sustained and the laser acts as a generator of light.
- The laser oscillation is initiated by photons spontaneously emitted by some of the excited atoms. Each spontaneous photon can trigger many stimulated transitions along the path of its travel. As the initial spontaneous photons are emitted in different directions, the stimulated photons would travel in different directions. The optical resonator selects the direction in which the light is to be amplified; the direction being the optical axis of the pair of mirrors. Thus, optical cavity makes the laser beam *directional*.
- In order to make the stimulated emission dominate spontaneous emission, a high radiation density $\rho(v)$ is required to be present in the active medium. The optical cavity builds up the photon density to a very high value through repeated reflections of photons and confines them within the medium.
- Optical cavity selects and amplifies only certain frequencies causing the laser output to be highly monochromatic.

24.7 LASING ACTION

Fig. 24.8 shows the active medium enclosed in optical resonator and being excited by a pumping agent. The resulting laser action, which is shown in Fig. 24.9, consists of the following steps:

Step-1: Pumping

The atoms (active centers) in the medium are in the ground state initially, as shown in Fig. 24.9 (a). By supplying energy from an external source, the atoms are excited from the ground level to an excited state.

Step-2: Population inversion

The lifetime of atoms at the excited state is extremely small, of the order of 10^{-8} sec. Therefore, the atoms drop spontaneously from the excited state to the metastable state. As the lifetime of atoms at the metastable state is comparatively longer (10^{-3} sec), the atoms go on accumulating at the metastable state. As soon as the number of atoms at the metastable state exceeds that of the ground state, the medium goes into the state of population inversion (Fig. 24.9 b).

Step-3: Spontaneous emissions

Some of the excited atoms at the metastable state may emit photons spontaneously in various directions (Fig. 24.9c). Each spontaneous photon can trigger many stimulated transitions along the direction of its propagation. As the initial spontaneous photons are moving in different directions, the photons stimulated by them also travel in different directions. Many of such photons leave the medium without reinforcing their strength. The photons emitted in a direction other than the axial direction will pass through the sides of the medium and are lost forever.

Step-4: Amplification

A majority of photons traveling along the axis cause stimulated emission and are reflected back on reaching the end mirror. They travel towards the opposite mirror and on their way stimulate more and more atoms and build up the photon strength, as shown in Fig. 24.9 (d). The photons that strike the opposite mirror are reflected once more into the medium, as shown in Fig. 24.9 (e). The photons travel once more through the medium generating more photons and more amplification. The photons are then reflected again at the mirror and travel through the medium. As the photons are reflected back and forth between the mirrors, stimulated emission sharply increases and the amplification of light takes place. The mirrors thus provide *positive feed back* of light into the medium so that stimulated emission acts are sustained and the medium operates as an oscillator.

Step-5: Oscillations

At each reflection at the front-end mirror, light is partially transmitted through it. The transmitted component constitutes a *loss of energy* from the resonator. When the losses at the mirrors and within the medium balance the gain, a steady and strong laser beam will emerge from the front-end mirror, as shown in Fig. 24.9 (f).

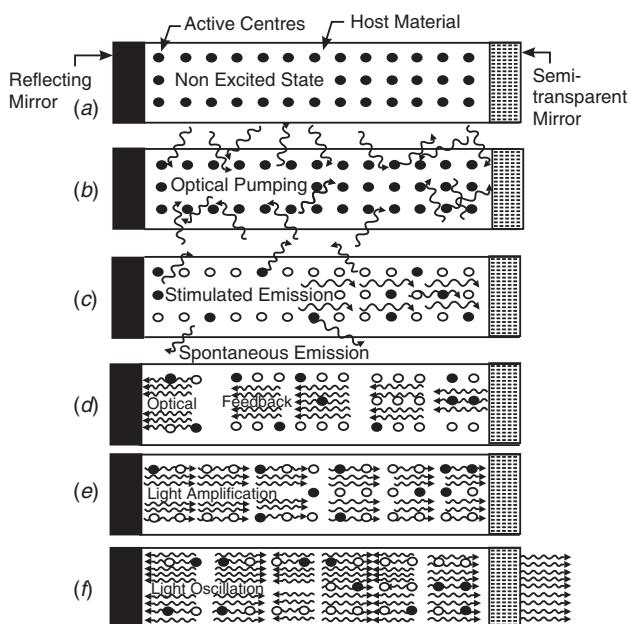


Fig. 24.9. Light amplification and oscillations due to the action of optical resonator

24.8 PUMPING METHODS

In order to create the state of population inversion in an active medium, the atoms in the material have to be pumped (excited) to particular energy levels. The most common methods of pumping are optical pumping and electrical pumping.

(a) **Optical pumping:** Optical pumping uses photons to excite the atoms. A light source such as a flash discharge tube is used to illuminate the laser medium and the photons of appropriate frequency excite the atoms to an upper energy level. From there, they drop to a metastable level to create the state of population inversion. The pump photon must have higher frequency than the emitted photon. This is because the atoms are to be excited to a level above the metastable level from the ground level or a lower energy level.

The pumping level of the atom must be a broader level. It should span a range of energies. If the level is narrow, one can use a pump photon of only one specific frequency. Such a situation severely restricts the choice of sources and also a large portion of the source power would go wasted. Fortunately, in a majority of cases the pump levels are wide bands. Therefore, light sources like flash lamps emitting a broad range of frequencies can be used for pumping. Optical pumping is suitable for any laser medium that is transparent to pump light. Optical pumping is used in solid state lasers.

(b) **Electrical pumping:** Electrical pumping can be used only in case of laser materials that can conduct electricity without destroying lasing activity. This method is limited to gases. In case of a gas laser, a high voltage pulse initially ionizes the gas so that it conducts electricity. An electric current flowing through the gas excites atoms to the excited level from where they drop to the metastable level leading to population inversion.

24.8.1 Principal Pumping Schemes

Atoms in general are characterized by a large number of energy levels. Among them only three or four levels will be pertinent to the pumping process. Therefore, only those levels are depicted in the pumping scheme diagrams. Two important pumping schemes are widely employed. They are known as *three-level* and *four-level pumping* schemes.

Two level scheme is not generally feasible for laser action. The main reason is that the energy being used to pump the atoms into the upper laser state has an equal probability of stimulating them back down. Therefore, it is not possible in general to pump more than half of atoms into the excited state.

1. Three-Level Pumping Scheme

: The three level scheme first excites the atoms to an excited state higher in energy than the upper laser state. The atoms then quickly decay down into the upper laser state. It is important for the pumped state to have a short lifetime for spontaneous emission compared to the upper laser state. The upper laser state should have as long a lifetime (for spontaneous emission) as possible, so that the atoms stay there long enough to be stimulated.

A typical three-level pumping scheme is shown in Fig. 24.10. The state E_1 is the ground state; E_3 is the pump state and E_2 is upper lasing level, which is a metastable state. When the medium is exposed to pump frequency radiation, a large number of atoms will be excited to E_3 level. However, they do not stay at that level but rapidly undergo downward transitions to the metastable state E_2 through non-radiative transitions. The atoms are trapped at this state as spontaneous transition from the level E_2 to the level E_1 is forbidden. The pumping continues

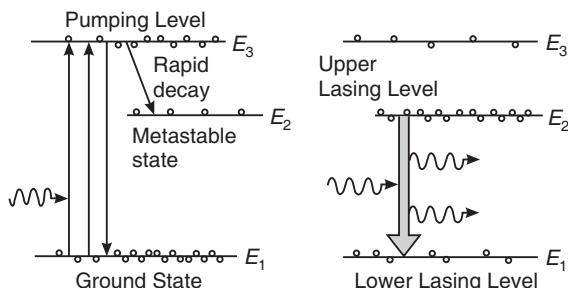


Fig. 24.10. A typical three level pumping scheme
(a) Pumping (b) Lasing action.

and after a short time there will be a large accumulation of atoms at the level E_2 . When more than half of the ground state atoms accumulate at E_2 , the population inversion condition is achieved between the two states E_1 and E_2 . Now a chance photon can trigger stimulated emission.

- In this scheme, the terminal state of the laser transition is simultaneously the ground state. Therefore, population inversion is achieved only when more than half of the ground state atoms are pumped to the upper state. Thus, the schme requires very high pump power.
- The three level scheme produces light only in pulses. Once stimulated emission commences, the metastable state is quickly emptied and the population of the ground state increases rapidly. As a result, the population inversion ends. One has to wait till the population inversion is re-established. Thus, the three-level lasers operates in pulsed mode.

2. Four-Level Pumping Scheme

A typical four level pumping scheme is shown in Fig. 24.11. The state E_1 is the ground state, E_4 the pumping level, E_3 the metastable upper lasing level and E_2 the lower lasing level. E_2 , E_3 and E_4 are the excited states. When light of pump frequency v_p is incident on the lasing medium, the active centers are readily excited from the ground state to the pumping level E_4 . The atoms stay at the E_4 level for only about 10^{-8} s, and quickly drop down to the metastable state E_3 . As spontaneous transitions from the level E_3 to level E_2 cannot

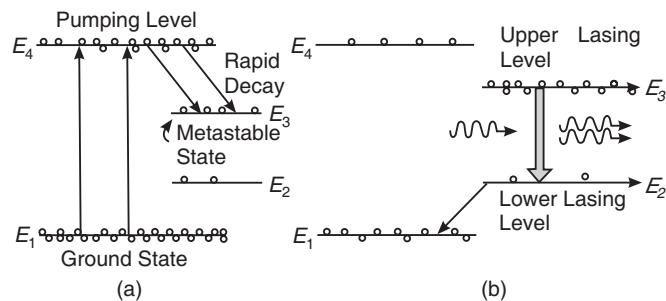


Fig. 24.11. A typical four level pumping scheme (a) Pumping
(b) Lasing Action.

take place, the atoms get trapped in the state E_3 . The population at the state E_3 grows rapidly. The level E_2 is well above the ground state such that $(E_2 - E_1) > kT$. Therefore, at normal temperature atoms cannot jump to level E_2 on the strength of thermal energy. As a result, the level E_2 is virtually empty. Therefore, population inversion is attained between the states E_3 and E_2 . A chance photon of energy $h\nu = (E_3 - E_2)$ emitted spontaneously can start a chain of stimulated emissions, bringing the atoms to the lower laser level E_2 . From state E_2 , the atoms subsequently under go non-radiative transitions to the ground state E_1 and will be once again available for excitation, making it possible for light to be emitted continuously.

- The lower laser transition level in this scheme is nearly vacant. Therefore, less pump power is sufficient to achieve population inversion.
- Four level lasers operate in continuous wave (cw) mode.

24.9 THRESHOLD CONDITION FOR LASING

As the light bounces back and forth in the optical resonator, it undergoes amplification as well as it suffers various losses. The losses occur mainly due to transmission at the output mirror and due to the scattering and diffraction of light within the active medium. For the proper build up of oscillations, it is essential that the amplification between two consecutive reflections of light from rear end mirror can balance the losses. We can determine the threshold

gain by considering the change in intensity of a beam of light undergoing a round trip within the resonator.

Let us assume that the laser medium fills the space between the mirrors M_1 and M_2 (see Fig. 24.12), which have reflectivity r_1 and r_2 respectively. Let the mirrors be separated by a distance L . Further, let the intensity of the light beam be I_0 at M_1 . Then, in travelling from mirror M_1 to mirror M_2 , the beam intensity increases from I_0 to $I(L)$, which is given by

$$I(L) = I_0 e^{(\gamma - \alpha_s)L} \quad (24.21)$$

where γ is the gain coefficient and α the loss coefficient of the active medium.

After reflection at M_2 , the beam intensity will be $r_2 I_0 e^{(\gamma - \alpha_s)L}$ and after a complete round trip the final intensity will be

$$I(2L) = r_1 r_2 I_0 e^{(\gamma - \alpha_s)2L} \quad (24.22)$$

The amplification obtained during the round trip is

$$G = \frac{I(2L)}{I_0} = r_1 r_2 e^{(\gamma - \alpha_s)2L} \quad (24.23)$$

The product $r_1 r_2$ represents the losses at the mirrors whereas α_s includes all the distributed losses such as scattering, diffraction and absorption occurring in the medium. The losses are balanced by gain, when $G \geq 1$ or $I(2L) = I_0$. It requires that

$$r_1 r_2 e^{2(\gamma - \alpha_s)L} \geq 1 \quad (24.24)$$

or $e^{2(\gamma - \alpha_s)L} \geq \frac{1}{r_1 r_2}$

Taking logarithms on both sides, we get

$$\begin{aligned} 2L(\gamma - \alpha_s) &\geq -\ln r_1 r_2 \\ \gamma - \alpha_s &\geq -\frac{1}{2L} \ln r_1 r_2 \\ \therefore \gamma &\geq \alpha_s - \frac{1}{2L} \ln r_1 r_2 \end{aligned} \quad (24.25)$$

or $\gamma \geq \alpha_s + \frac{1}{2L} \ln \frac{1}{r_1 r_2}$ (24.26)

Equ.(24.26) is known as the **condition for lasing**. It shows that the initial gain must exceed the sum of the losses in the cavity. This condition is used to determine the threshold value of pumping energy for lasing action.

γ will be dependent on how hard the laser medium is pumped. As the pump power is slowly increased, a value of γ_{th} called **threshold value** is reached and the laser starts oscillating. The threshold value γ_{th} is given by

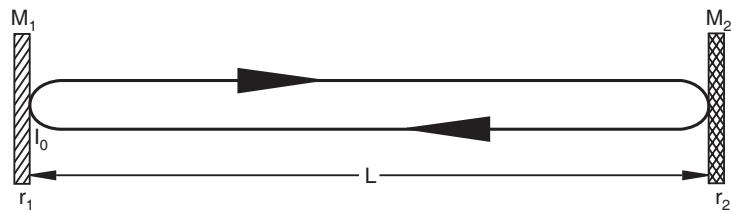


Fig. 24.12. Round trip path of the radiation through the laser cavity

$$\gamma_{\text{th}} = \alpha_s + \frac{1}{2L} \ln \frac{1}{r_1 r_2} \quad (24.27)$$

Equ.(24.27) states the condition when the net gain would be able to counteract the effect of losses in the cavity and is known as the **threshold condition for lasing**. The value of γ must be atleast γ_{th} for laser oscillations to commence.

24.10 MODES OF THE LASER BEAM

The light waves within an optical resonant cavity are characterized by their resonant modes, which are discrete resonant conditions determined by the dimensions of the cavity. The laser beam radiated from the laser cavity is thus not arbitrary. Only the waves oscillating at modes that match the oscillation modes of the laser cavity can be produced. The laser modes governed by the axial dimensions of the resonant cavity are called the **longitudinal modes**, and the modes determined by the cross-sectional dimensions of the laser cavity are called **transverse modes**.

(a) Longitudinal Modes

A light wave which moves inside the laser cavity from right to left is reflected by the left mirror, and move to the right until it is reflected from the right mirror, and so on. Thus, two waves of the same frequency and amplitude are moving in opposite directions, which is the condition for creating a standing wave. In order to create a *standing wave*, the wave must start with the same phase at the mirror (Fig. 24.13). Therefore, the optical path from one mirror to the other and back must be an integer multiplication of the wavelength.

Thus,

$$2\mu L = m\lambda_m \quad (m = 1, 2, 3, \dots)$$

or

$$\lambda_m = \frac{2\mu L}{m} \quad (24.28)$$

Light waves are amplified strongly if, and only if, they satisfy the above condition. Only those wavelengths that satisfy equation (24.28) can exist inside the cavity. Waves of other wavelengths interfere destructively with each other as they pass back and forth between the mirrors. Thus, they attenuate very quickly. Because of its relatively longer length as compared to the wavelength of light, the resonator may support simultaneously several standing waves. These standing waves are known as the **longitudinal modes**. Each mode has a distinct frequency given by

$$v_m = \frac{mc}{2\mu L} \quad (24.29)$$

Thus, frequency difference between to adjacent modes (m and $m \pm 1$) is

$$\Delta v = \frac{c}{2\mu L} \quad (24.30)$$

The frequencies given by the relation (24.29) are the *allowed frequencies* inside a laser cavity of length L . However, all these allowed frequencies will not be emitted from

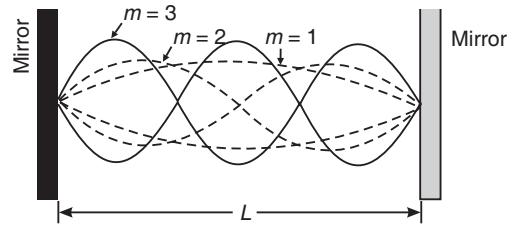


Fig. 24.13

the laser, since there are other limiting conditions. Only those frequencies (modes) that have amplification above the lasing threshold, to overcome absorption will be emitted out of the laser (see Fig. 24.14).

Fig. 24.14 shows the gain curve of a particular active medium, as a function of frequency and is marked with the **lasing threshold** and

possible longitudinal modes of the laser. The region marked under the curve and above the lasing threshold include the range where lasing can occur. For example, in Fig. 24.14 only 5 frequencies from those allowed inside the cavity, are above the lasing threshold. Thus, only these 5 frequencies can exist at the output of this laser.

Example 24.3. Because of the interaction of chromium ion with ruby lattice, the transitions responsible for ruby-laser-emission are spread over an energy, resulting in wavelength spread of 0.53 mm around 694.3 nm. If the length of ruby rod is 2 cms (refractive index 1.75) how many longitudinal cavity modes would the ruby-laser emission contain?

$$\text{Solution. Mode separation} = \frac{c/\mu}{l} = \frac{3 \times 10^8 / 1.75 \text{ m/s}}{2 \times 10^{-2} \text{ m}} = 8.6 \times 10^9 \text{ Hz}$$

Frequency spread of laser emission

$$\Delta v = (c/\lambda^2)\Delta\lambda = \frac{3 \times 10^8 (\text{m/s}) \times 0.53 \times 10^{-9} (\text{m})}{(694.3 \times 10^{-9} \text{ m})^2} = 330 \times 10^9 \text{ Hz.}$$

$$\text{No. of cavity modes} = \Delta v / \text{mode separation} = 330 \times 10^9 / 8.6 \times 10^9 = 38.5$$

$$\text{i.e.,} \quad = 38 \text{ modes}$$

Example 24.4. If the half-width of the He-Ne laser operating at wavelength 6328\AA is 1500 MHz, what must be the length of the laser cavity to ensure that only one longitudinal mode oscillates?

Solution:

$$\text{The length of cavity is given by } L = \frac{mc}{2\Delta v} = \frac{1 \times 3 \times 10^8 \text{ m/s}}{2 \times 1.5 \times 10^9 / \text{s}} = 0.1 \text{ m.}$$

(b) Transverse Modes

The configuration of the optical cavity determines the transverse modes of the laser output, which characterizes the intensity distribution across the cross-section of the laser beam. This is simply a natural consequence of electromagnetic waves being confined within the cavity and restricted by the boundary conditions. In general, the allowed modes in an optical cavity are designated as TEM_{mn} , where T, E, and M stand for transverse, electric, and magnetic modes, respectively, and m and n are integers. The simplest or the lowest-order transverse mode is TEM_{00} , which has a smooth cross-section profile with a Gaussian peak in the middle, as shown in Fig. 24.15.

The integers m and n associated with the TEM_{mn} mode represent the numbers of zeros or minima between the edges of the beam in two orthogonal directions. As shown in the upper

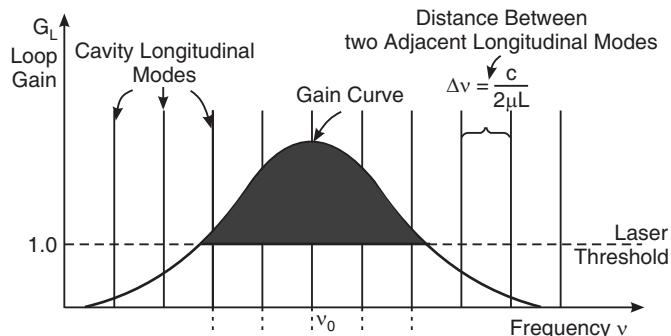


Fig. 24.14. Gain curve of a laser

picture, a TEM_{01} beam has a single minimum dividing the beam into two bright spots. A TEM_{11} beam has two perpendicular minima (one in each direction), dividing the beam into four quadrants. The larger the values of m and n , the more bright spots are contained in the laser beam. To achieve a single transverse mode operation, one can place a small circular aperture that is slightly larger than the spot size of the TEM_{00} mode in the middle of the laser cavity. Most laser cavities are designed to produce only the TEM_{00} mode. However, some lasers do operate in higher order modes, especially when they are designed to maximize the output power.

24.11 TYPES OF LASERS

Lasers are divided into different types basing on different considerations. We divide them here on the basis of the material used. Some of the important types of lasers are

1. Solid-state lasers Examples: Ruby laser, Nd:YAG laser etc
2. Gas lasers Examples: Helium-Neon laser, CO_2 laser etc
3. Semiconductor diode lasers Examples: GaAs laser, InP laser etc

Most lasers emit light in the red or IR regions. Lasers work in continuous mode or in a pulsed mode.

24.11.1 Ruby Laser

Ruby laser belongs to the class of solid-state lasers. Ruby is basically Al_2O_3 crystal containing about 0.05% of chromium atoms. Cr^{3+} ions are the actual active centers while aluminum and oxygen atoms are inert. Chromium ions have absorption bands in the blue and green regions.

Ruby rod is taken in the form of a cylindrical rod of about 4 cm in length and 1 cm in diameter. Its ends are grounded and polished such that the end faces are exactly parallel and are also perpendicular to the axis of the rod.

The end faces of the ruby rod are silvered so that they form the optical resonator. The rear face is made totally reflecting while the front face is made partially reflecting.

The laser rod is surrounded by a helical photographic flash lamp filled with xenon (Fig. 24.16 a). Whenever activated by the power supply the lamp produces flashes of white light.

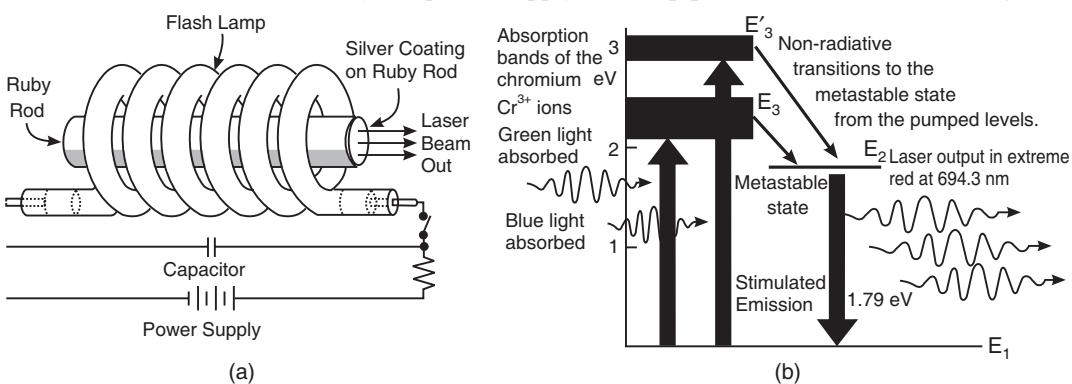


Fig. 24.16. (a) Schematic of a ruby laser (b) Energy levels and transitions in a ruby laser

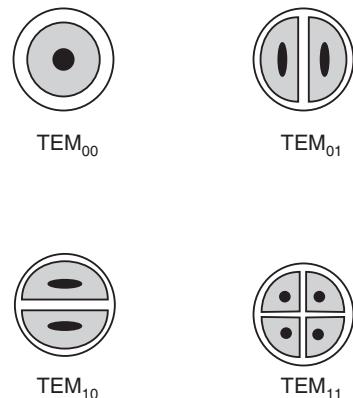


Fig. 24.15. Beam profiles of transverse modes.

Working: The energy levels of Cr^{3+} ions in the crystal lattice are shown in Fig. 24.16 (b). The energy level structure of the Cr^{3+} ions is characterized by two absorption bands and a metastable state.

The Pumping Mechanism

- When the xenon flash lamp is switched on, the discharge generates an intense burst of white light lasting for a few milliseconds.
- The Cr^{3+} ions are excited to the energy bands E_3 and E_3' by the green and blue components of white light.
- The Cr^{3+} ions undergo non – radiative transitions from these energy levels to level E_2 . E_2 is a metastable state.
- The metastable state E_2 has a lifetime of approximately 1000 times more than the lifetime of E_3 and E_3' levels. Therefore, Cr^{3+} ions accumulate at E_2 level.
- The metastable level E_2 is the upper laser level, while E_1 is the ground level and constitutes the lower laser level.

Population Inversion

- The upper laser level E_2 will be rapidly populated, as the excited Cr^{3+} ions quickly make downward transitions from the upper energy bands.
- When more than half of the Cr^{3+} ion population accumulates at E_2 level, the state of population inversion is established between E_2 and E_1 levels.

Lasing

- A chance photon is produced when a Cr^{3+} ion makes a spontaneous transition from E_2 level to E_1 level.
- This spontaneous photon stimulates another excited ion to make a downward transition.
- This stimulated photon and the initial photon trigger many excited ions to emit photons.
- Red photons of wavelength 6943 Å travelling along the axis of the ruby rod are repeatedly reflected at the end mirrors and light amplification takes place.
- On attaining sufficient energy, the laser beam emerges out through the partially reflecting mirror.
- The laser emission occurs in the visible region at a wavelength of 6943 Å (694.3 nm).
- Once stimulated transitions commence, the metastable state gets depopulated very rapidly and the state of population inversion disappears and lasing action ceases.
- The laser becomes active once again when population inversion state is reestablished.
- Therefore, the output of the laser is not a continuous wave but occurs in the form of pulses of microsecond duration.

Salient Features

- Uses three-level pumping scheme
- The active centers are Cr^{3+} ions
- Light from a xenon flash lamp is the pumping agent
- Poor efficiency
- Operates in pulsed mode

24.11.2 Nd: YAG Laser

Nd: YAG laser is one of the most popular types of solid state laser. It is a four-level laser. Yttrium aluminium garnet, $\text{Y}_3\text{Al}_5\text{O}_{12}$, commonly called YAG is an optically isotropic crystal. Some of the Y^{3+} ions in the crystal are replaced by neodymium ions, Nd^{3+} . Doping concentrations are typically of the order of 0.725% by weight. The crystal atoms do not participate in the lasing action but serve as a host lattice in which the active centres, namely Nd^{3+} ions reside.

Construction: Fig. 24.17 illustrates a typical design of Nd: YAG laser. The system consists of an elliptically cylindrical reflector housing the laser rod along one of its focus line and a flash lamp along the other focus line. The light leaving one focus of the ellipse will pass through the other focus after reflection from the silvered surface of the reflector. Thus the entire flash lamp radiation gets focused on the laser rod. The YAG crystal rods are typically of 10 cm in length and 12 mm in diameter. The two ends of the laser rod are polished and silvered and constitute the optical resonator.

Working: A simplified energy level diagram for the neodymium ion in YAG crystal is shown in Fig. 24.18. The energy

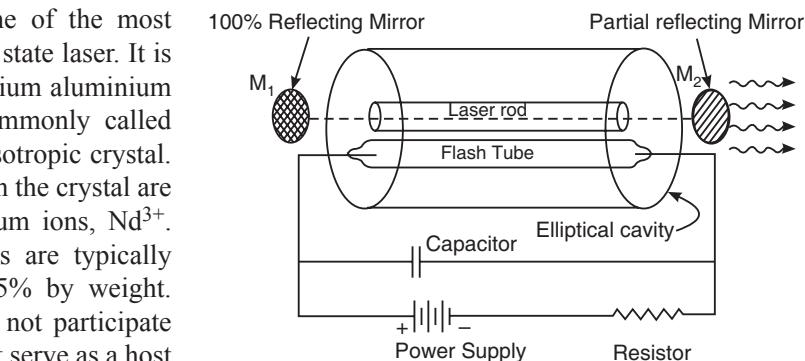


Fig. 24.17. Schematic of a Nd:YAG Laser

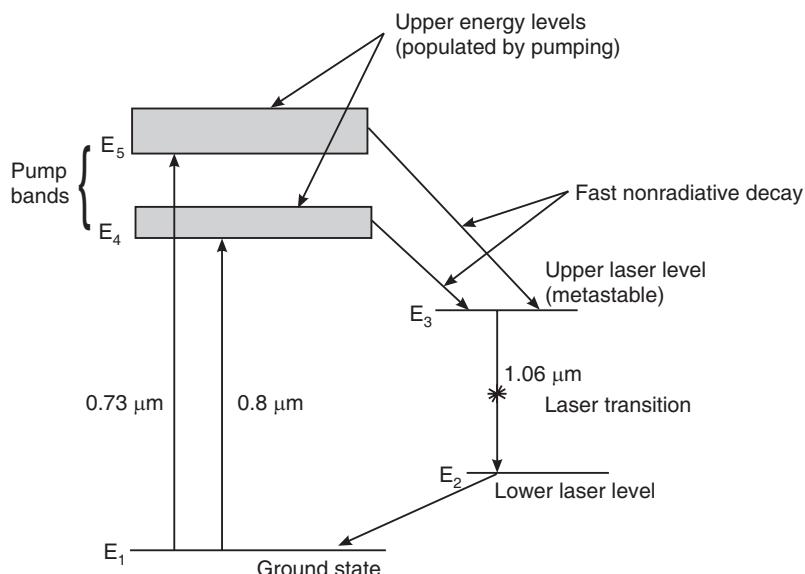


Fig. 24.18. Energy levels and transitions in a Nd : YAG laser.

level structure of the free neodymium atom is preserved to a certain extent because of its relatively low concentration. However, the energy levels are split and the structure is complex.

The Pumping Mechanism

- When the krypton flash lamp is switched on, the Nd^{3+} ions are excited to the upper energy bands E_4 and E_5 .
- The Nd^{3+} ions make a transition from these energy levels to level E_3 by non-radiative transition. E_3 is a metastable state.
- The metastable level E_3 is the upper laser level, while E_2 forms the lower laser level.

Population Inversion

- The upper laser level E_3 will be rapidly populated, as the excited Nd³⁺ ions quickly make downward transitions from the upper energy bands.
- The lower laser level E_2 is far above the ground level and hence it cannot be populated by Nd³⁺ ions through thermal transitions from the ground level.
- Therefore, the population inversion is readily achieved between the E_3 level and E_2 level.

Lasing

- A chance photon is produced when an Nd³⁺ ion makes a spontaneous transition from E_3 level to E_2 level.
- This spontaneous photon stimulates another excited atom to make a downward transition.
- This stimulated photon and the initial photon trigger many excited atoms to emit photons.
- Photons thus generated travel back and forth between the two end mirrors and gain in strength very rapidly.
- On attaining sufficient energy, the laser beam emerges out through the partially reflecting mirror.
- The laser emission occurs in infrared (IR) region at a wavelength of about 10,600 Å (1.06 μm).
- The Nd³⁺ ions return to the ground state E_1 from the lower lasing level E_2 , on their own through non-radiative transitions.

Salient Features

- Uses four-level pumping scheme
- The active centers are Nd³⁺ ions
- Light from a xenon or krypton flash lamp is the pumping agent
- Low efficiency (1%) and moderate power output (watts)
- Operates in CW/pulsed mode

24.11.3 Helium-Neon Laser

Gas lasers are the most widely used lasers. They range from the low power helium-neon laser used in college laboratories to very high power carbon dioxide laser used in industrial applications. These lasers operate with rarefied gases as the active media and are excited by an electric discharge. In gases, the energy levels of atoms involved in the lasing process are narrow and as such require sources with sharp wavelength to excite atoms. Finding an appropriate optical source for pumping poses a problem. Therefore optical pumping is not used in gas lasers. The most common method of exciting gas laser medium is by passing an electric discharge through the gas. Electrons present in the discharge transfer energy to atoms in the laser gas by collisions.

The first gas laser was He-Ne laser, which was invented in 1961 by Ali Javan, William R. Bennett, Jr. and Donald R. Herriott.

Construction:

The schematic of a He-Ne laser is shown in Fig. 24.19. Helium –

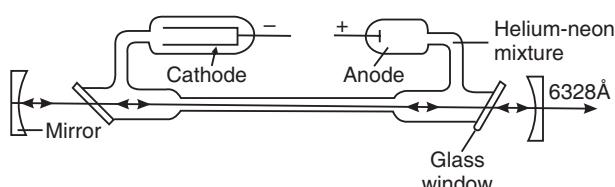


Fig. 24.19

Neon laser consists of a long discharge tube filled with a mixture of helium and neon gases in the ratio 10:1. Neon atoms are the active centers and have energy levels suitable for laser transitions while helium atoms help in exciting neon atoms. Electrodes are provided in the discharge tube to produce discharge in the gas. They are connected to a high voltage power supply. The tube is hermetically sealed by inclined windows arranged at its two ends. On the axis of the tube, two mirrors are arranged externally, which form the Fabry-Perot optical resonator. The distance between the mirrors is adjusted to be $m \lambda/2$ such that the resonator supports standing wave pattern.

Working:

The energy levels of helium and neon are shown in Fig. 24.20.

The Pumping Mechanism

- When the power is switched on, a high voltage of about 10 kV is applied across the gas mixture. It ionizes the gas.
- The electrons and ions produced in the process of discharge are accelerated towards the anode and cathode respectively. They collide with helium and neon atoms on the way.
- The energetic electrons excite helium atoms more readily, as they are lighter.
- One of the excited levels of helium $F_2(2s)$ is at 20.61 eV above the ground level. It is a metastable level and the excited helium atom cannot return to the ground level through spontaneous emission.
- However, the excited helium atom can return to the ground level by transferring its excess energy to a neon atom through collision. Such an energy transfer can take place when the two colliding atoms have identical energy levels. Such an energy transfer is known as **resonant energy transfer**.
- The neon energy level $E_5(5s)$ is at 20.66 eV, which is close to the excited energy level F_2 of helium atom. Therefore, resonant transfer of energy occurs between the excited helium atom and ground level neon atom. The kinetic energy of helium atoms provides the additional 0.05 eV required for excitation of the neon atoms.
- Helium atoms drop to the ground state after exciting neon atoms. This is the pumping mechanism in He-Ne laser.

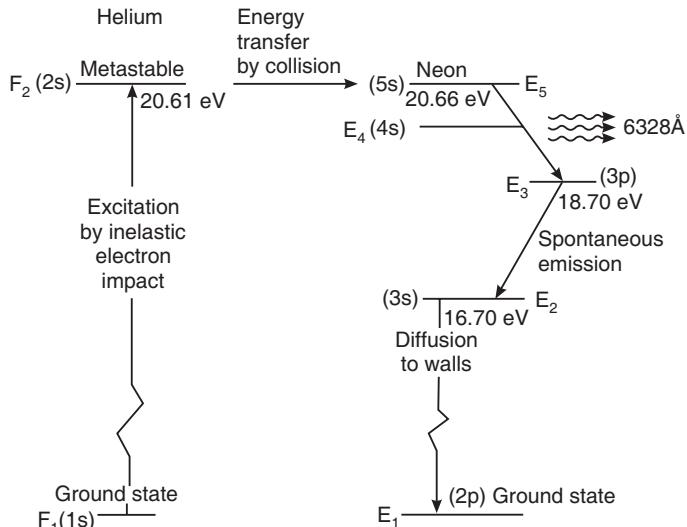


Fig. 24.20

Population Inversion

- The upper state of neon atom E_5 is a metastable state. Therefore, neon atoms accumulate in this upper state.

- The E_3 (3p) level is sparsely populated at ordinary temperatures.
- As the population at the higher energy level E_5 is greater than the population at the lower level E_3 , a state of population inversion is established between E_5 and E_3 levels.

Lasing

- Random photons of red colour of wavelength 6328 Å are emitted spontaneously by a few of the atoms at the energy level E_5 .
- The spontaneous photons traveling through the gas mixture prompt stimulated emission of photons of red colour of wavelength 6328 Å.
- The photons bounce back and forth between the end mirrors, causing more and more stimulated emission during each passage. The strength of the stimulated photons traveling along the axis of the optical cavity (discharge tube) builds up rapidly while the photons traveling at angles to the axis are lost.
- Thus, the transition $E_5 \rightarrow E_3$ generates a laser beam of wavelength 6328 Å.
- From the level E_3 the neon atoms drop to E_2 (3s) level spontaneously.
- E_2 level is a metastable state. Consequently, neon atoms tend to accumulate at E_2 level.
- Neon atoms return to the ground state E_1 through frequent collisions with the walls of the glass tube holding the helium-neon gas mixture.
- The neon atoms are once again available for excitation to higher state and participate in lasing action.
- The neon atoms are excited to the upper lasing level continuously through collisions. As the population inversion can be maintained in the face of continuous laser emission, the laser operates in continuous wave mode.

Role of helium atoms

- The role of helium atoms in the laser is to excite neon atoms and to cause population inversion. The probability of energy transfer from helium atoms to neon atoms is more, as there are 10 helium atoms per 1 neon atom in the gas mixture. The probability of reverse transfer of energy from neon to helium atom is negligible.

Necessity of narrow glass tube

- During the operation of the laser, it is necessary that the atoms accumulating at the metastable level E_2 are brought to the ground state E_1 (2p) quickly; otherwise the number of atoms at the ground state will go on diminishing and the laser ceases to function. The only way of bringing the atoms to the ground state is through collisions. Therefore, to increase the probability of atomic collisions with the tube walls, the discharge tube is made narrow.

Salient Features

- Uses four-level pumping scheme
- The active centers are neon atoms
- Electrical discharge is the pumping agent
- Low efficiency and low power output
- Operates in CW mode

24.11.4 Carbon Dioxide Laser

The carbon gas laser is a very useful and efficient laser. It is a four-level molecular laser and operates at $10.6 \mu m$ in far IR region.

Construction: The schematic of typical CO₂ laser is shown in Fig. 24.21. It is basically a discharge tube having a bore of cross section of about 1.5 mm² and a length of about 260 mm. The discharge tube is filled with a mixture of carbon dioxide, nitrogen and helium gases in 1:4:5 proportions respectively. Other additives such as water vapour are also added. The active centres are CO₂ molecules lasing on the transitions between the vibrational levels of the electronic ground state.

Energy levels of CO₂ molecule

Fig. 24.22 shows the vibrational modes and rotations of CO₂ molecule.

- The electron energy levels of an isolated atom are discrete and narrow. However, in case of molecules the energy spectrum is complicated due to many additional features.
- Each electron energy level is associated with nearly equally spaced vibrational levels and each vibrational level in turn has a number of rotational levels.
- CO₂ molecule is a linear molecule consisting of a central carbon atom with two oxygen atoms attached one on either side.
- It undergoes three independent vibrational oscillations known as the **vibrational modes** (Fig. 24.22). These vibrational degrees of freedom are quantized. At any one time, a CO₂ molecule can vibrate in a linear combination of three fundamental modes.
- The energy states of the molecule are represented by three quantum numbers (m n q).
- These numbers represent the amount of energy associated with each mode. For example, the number (020) indicates that the molecule in this energy state is in the pure bending mode with two units of energy.
- Each vibrational state is associated with rotational states corresponding to the rotation of CO₂ molecule about its centre of mass. The separations between vibrational – rotational states are much smaller on the energy scale compared to the separations between electron energy levels.

The nitrogen molecule N₂ is also characterized by similar vibrational levels.

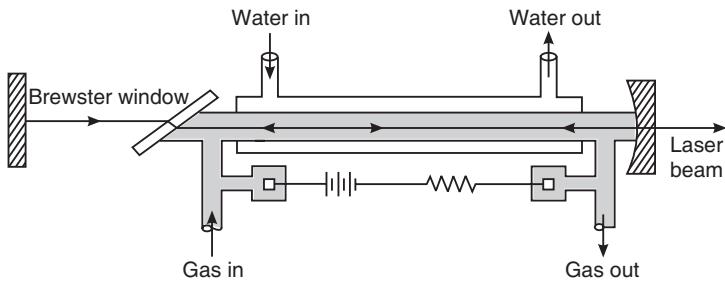


Fig. 24.21. Schematic of a carbon dioxide laser

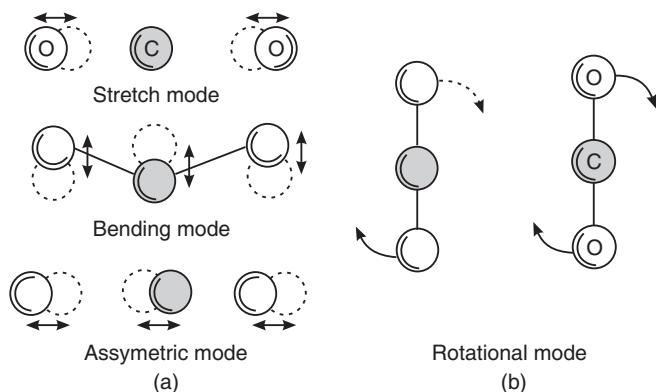


Fig. 24.22. Vibrational modes of a CO₂ molecule

Working: Fig. 24.23 shows the lowest vibrational levels of the ground electron energy state of CO_2 molecule and an N_2 molecule. The excited state of an N_2 molecule is metastable and it is identical in energy to (001) vibrational level of CO_2 molecule, indicated as E_5 in Fig. 24.23.

The Pumping Mechanism

- When current passes through the mixture of gases, the N_2 molecules get excited to the metastable state.
- The excited N_2 molecules cannot spontaneously lose their energy and consequently, the number of N_2 molecules at the metastable level builds up.
- The N_2 molecules undergo inelastic collisions with ground state CO_2 molecules and excite them to E_5 level. Some of the CO_2 molecules are also excited to the upper level E_5 through collisions with electrons.

Population Inversion

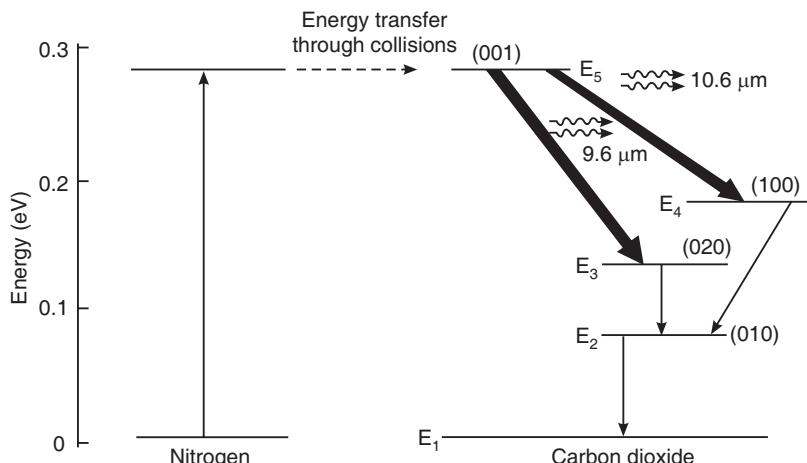


Fig. 24.23. Energy levels of nitrogen and carbon dioxide molecules and transitions between the levels

- The E_5 level is the upper lasing level while the (020) and (100) states marked as E_3 and E_4 levels act as the lower lasing levels.
- As the population of CO_2 molecules builds up at E_5 levels population inversion is achieved between E_5 level and the levels at E_4 and E_3 .

Lasing

- Random photons are emitted spontaneously by a few of the atoms at the energy level E_5 .
- The spontaneous photons traveling through the gas mixture prompt stimulated emission of photons.
- The photons bounce back and forth between the end mirrors, causing more and more stimulated emission during each passage. The strength of the stimulated photons traveling along the axis of the optical cavity (discharge tube) builds up rapidly while the photons traveling at angles to the axis are lost.
- The laser transition between $E_5 \rightarrow E_4$ levels produces far IR radiation at the wavelength $10.6 \mu\text{m}$ ($1,06,000\text{\AA}$).
- The lasing transition between $E_5 \rightarrow E_3$ levels produces far IR radiation at $9.6 \mu\text{m}$ ($96,000\text{\AA}$) wavelength.

- E_3 and E_4 levels are also metastable states and the CO_2 molecules at these levels fall to the lower level E_2 through inelastic collisions with normal (unexcited) CO_2 molecules.
- This process leads to accumulation of population at E_2 level. As the gaseous mixture heats up, the E_2 level, which is close to the ground state, tends to be populated through thermal excitations. Thus, the de-excitation of CO_2 molecules at the lower lasing level poses a problem and inhibits the laser action.
- The helium atoms de-excite CO_2 molecules through inelastic collisions and decrease the population density of CO_2 at E_2 level. It also aids cooling the gaseous mixture through heat conduction.
- The CO_2 molecules are once again available for excitation to higher state and participate in lasing action.
- CO_2 molecules are excited to the upper lasing level continuously through collisions. As the population inversion can be maintained in the face of continuous laser emission, the laser operates in continuous wave mode.

Salient Features

- Uses four-level pumping scheme
- The active centers are CO_2 molecules
- Electrical discharge is the pumping agent
- High efficiency (40%) and high power output (several kilowatts)
- Operates in CW mode

24.11.5 Semiconductor Diode Laser

A **semiconductor diode laser** is a specially fabricated p-n junction device, which emits coherent light when it is forward biased. R. N. Hall and his coworkers made the first semiconductor laser in 1962. It is made from Gallium arsenide (GaAs) which operated at low temperatures and emitted light in the near IR region. Now, p-n junction lasers are made to emit light almost anywhere in the spectrum from UV to IR.

Diode lasers are remarkably small in size (0.1mm long). They have high efficiency of the order of 40%. Modulating the biasing current easily modulates the laser output. They operate at low powers. In spite of their small size and low power requirement, they produce power outputs equivalent to that of He-Ne lasers. The chief advantage of a diode laser is that it is portable. Because of the rapid advances in semiconductor technology, diode lasers are mass produced for use in optical fibre communications, in CD players, CD-ROM drives, optical reading, high speed laser printing etc wide variety of applications.

Semiconductor Materials

- Among the semiconductors, there are two different groups. They are *direct band gap semiconductors* and *indirect band gap semiconductors*.
- Direct band gap semiconductor is the one in which a conduction band electron can recombine directly with a hole in the valence band.
- The recombination process leads to emission of light.
- Most of the compound semiconductors belong to this group.
- Direct recombination of conduction band electron with a hole in the valence band is not possible in indirect band gap semiconductors. Silicon and germanium belong to this group. The recombination of an electron and a hole produces heat in these materials.

- Direct band gap semiconductors are formed by group III-V elements and group IV-VI elements.
- Lasers are made using direct band gap semiconductors. Gallium Arsenide (GaAs) diode is an example of semiconductor diode laser.

Principle

- The energy band structure of a semiconductor consists of a valence band and a conduction band separated by an energy gap, E_g .
- The conduction band contains electrons and the valence band contains holes and electrons.
- When an electron from the conduction band jumps into a hole in the valence band, the excess energy, E_g is given out in the form of a photon.
- Thus, the electron-hole recombination is the basic mechanism responsible for emission of light.
- The wavelength of the light is given by the relation $\lambda = hc/E_g$.
- Semiconductors having a suitable value of E_g emit light in the optical region.

Types of semiconductor diode lasers

Broadly there are two types of semiconductor diode lasers. They are known as *homojunction semiconductor lasers* and *heterojunction semiconductor lasers*.

Homojunction Semiconductor Laser

A simple diode laser which makes use of the same semiconductor material on both sides of the junction is known as a homojunction diode laser.

Example: Gallium arsenide (GaAs) laser.

Heterojunction Semiconductor Laser

A diode laser which makes use of different semiconductor materials on the two sides of the junction is known as a heterojunction diode laser. These are further classified as single heterojunction diode lasers and double heterojunction diode lasers.

Example: A junction laser having GaAs on one side and GaAlAs on the other side.

24.11.5.1 Homojunction semiconductor laser

Construction: Fig. 24.24 shows the schematic of a homojunction diode laser. Starting with a heavily doped n-type GaAs material, a p-region is formed on its top by diffusing zinc atoms into it. A heavily zinc doped layer constitutes the heavily doped p-region. The diode is extremely small in size. Typical diode chips are 500 μm long and about 100 μm wide and thick. The top and bottom faces are metallized and metal contacts are provided to pass current through the diode. The front and rear faces are polished parallel to each other and perpendicular to the plane of the junction. The polished faces constitute the Fabry-Perot resonator. In practice there is no necessity to polish the faces. A pair of parallel planes cleaved

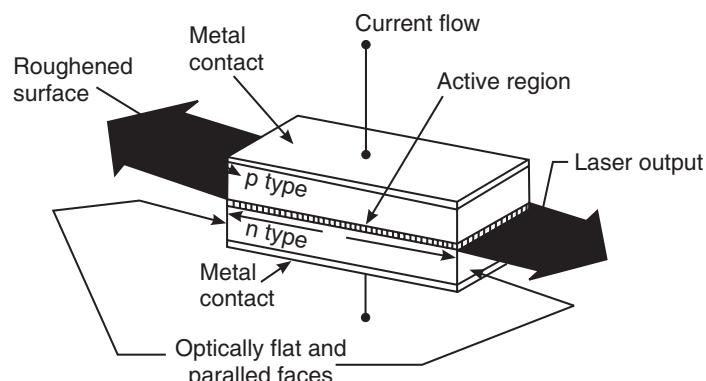


Fig. 24.24. Schematic of homojunction diode laser

at the two ends of the pn junction provides the required reflection to form the cavity. The two remaining sides of the diode are roughened to eliminate lasing action in that direction. The entire structure is packaged in small case which looks like the metal case used for discrete transistors.

Working: The energy band diagram of a heavily doped p-n junction is shown in Fig. 24.25.

- Heavily doped p- and n- regions are used in making a laser diode.
- Because of very high doping on n-side, the donor levels are broadened and extend into the conduction band. The Fermi level also is pushed into the conduction band.
- Electrons occupy the portion of the conduction band lying below the Fermi level.
- Similarly, on the heavily doped p-side the Fermi level lies within the valence band and holes occupy the portion of the valence band that lies above the Fermi level.
- At thermal equilibrium, the Fermi level is uniform across the junction.

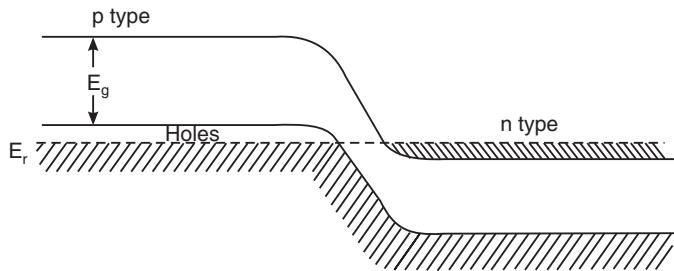


Fig. 24.25. Energy band diagram of a heavily doped p-n junction without bias

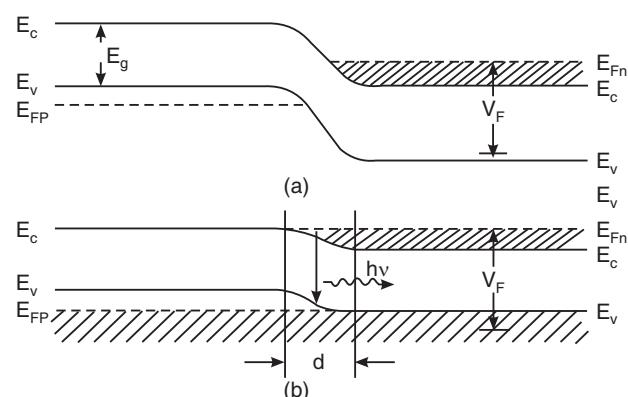


Fig. 24.26. Laser diode under forward bias

The Pumping Mechanism

- When the junction is forward-biased, electrons and holes are injected into the junction region in high concentrations (Fig. 24.26a).
- In other words, charge carriers are *pumped* by the dc voltage source.
- When the diode current reaches a threshold value (see Fig. 24.26b), the carrier concentrations in the junction region will rise to a very high value.

Population Inversion

- As a result, the region (region 'd' in Fig. 24.26b) contains a large concentration of electrons within the conduction band and *simultaneously* a large number of holes within the valence band.
- Holes represent absence of electrons.
- Thus, the upper energy levels in the narrow region are having a high electron population while the lower energy levels in the same region are vacant.

- Therefore, the condition of population inversion is attained in the narrow junction region. This narrow zone in which population inversion occurs is called an **inversion region or active region**.

Lasing

- Chance recombination acts of electron and hole pairs lead to emission of spontaneous photons.
- The spontaneous photons propagating in the junction plane stimulate the conduction electrons to jump into the vacant states of valence band.
- This stimulated electron-hole recombination produces coherent radiation.
- GaAs laser emits light at a wavelength of 9000 Å in IR region.

Drawbacks of homojunction lasers:

1. In homojunction lasers, the active region is not well defined due to the diffusion length of the carriers.
2. The semiconductor has nearly uniform refractive index throughout. Therefore, light can diffuse from active layer into the surrounding medium. As a result the cavity losses increase.
3. High threshold currents are required and the laser cannot be operated continuously at room temperature.

24.11.5.2 Heterojunction laser

Heterojunction lasers are multilayer-structures designed such that the carriers are confined in a narrow region and population is built up at lower current levels. We study here the structure and working of a double heterojunction (DH) laser.

Construction:

In a double heterojunction laser, a GaAs layer is sandwiched between two layers of GaAlAs (Fig. 24.27). The material GaAlAs has a wider energy gap and a lower refractive index than GaAs. The top and bottom faces are metallized and metal contacts are provided to pass current through the diode. The front and rear faces are polished parallel to each other and perpendicular to the plane of the junction. The polished faces constitute the Fabry-Perot resonator.

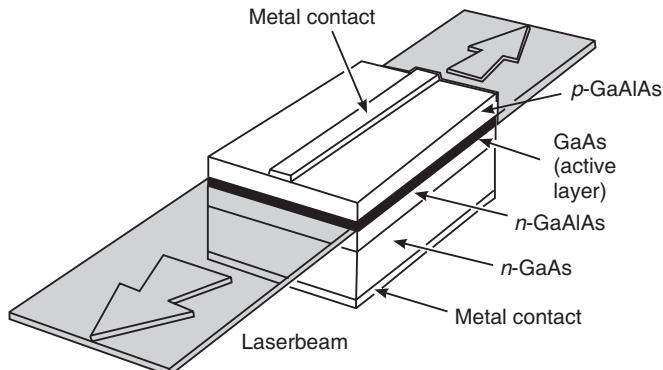


Fig. 24.27

Working:

- The basic principle of working of heterojunction diode is similar to that of a homojunction diode.
- p-type narrow band gap GaAs layer at the centre constitutes the active layer in which the lasing occurs.
- This layer is flanked by an n-type wide band gap GaAlAs layer on one side and by a p-type wide band gap GaAlAs layer on the other side (Fig. 24.28 a).

- The refractive indices of GaAlAs layers are smaller than that of GaAs layer (Fig. 24.28c).

The Pumping Mechanism

- When the junction is forward-biased, electrons and holes are injected into the active region in high concentrations.
- When the diode current reaches a threshold value, the carrier concentrations in the active region will rise to a very high value.
- The electrons injected from the n-type GaAlAs layer into the p-GaAs layer confront energy barrier at the junction where p-type GaAs and p-type GaAlAs meet and are reflected back into the active region. Similarly, the holes are reflected by the potential barrier provided by the higher band gap n-GaAlAs layer. Thus, the layers in heterostructure confine the charge carriers to the GaAs layer.

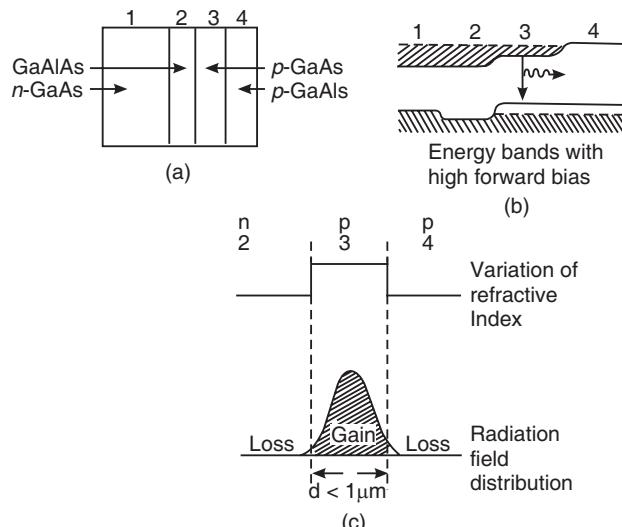


Fig. 24.28

Population Inversion

- Due to the forward bias, the active region contains a large concentration of electrons within the conduction band and *simultaneously* a large number of holes within the valence band.
- Thus, the upper energy levels in the active region are having a high electron population while the lower energy levels in the same region are vacant.
- Therefore, the condition of population inversion is attained in the narrow active region.
- Now the active region thickness is the thickness of p-GaAs layer. If its thickness is made small, a smaller drive current can lead to population inversion.

Lasing

- Chance recombination acts of electron-hole pairs lead to spontaneous emission of photons.
- The spontaneous photons propagating in the active layer stimulate the conduction electrons to jump into the vacant states of valence band and produce stimulated photons.
- The reflection at the GaAs – air interface provides sufficient feed back for laser oscillation.
- When the diode losses are off-set by the laser gain, the laser oscillations will start.
- When the radiation attains appropriate strength, laser beam emerges from the diode.
- As the refractive index of GaAs is higher than the refractive index of GaAlAs layers, the light is trapped within the active region and travels in one direction only.

- The wavelength of the light emitted by the GaAs layer is 800 nm when its band gap is 1.55 eV.

Advantages

- Heterojunction lasers have high efficiency even at room temperature.
- As a result of reduction in the threshold current density, continuous operation is possible.
- With operating currents of less than 50 mA, output powers of about 10 mW can be produced.

24.12 LASER BEAM CHARACTERISTICS

The important characteristics of a laser beam are:

- (i) directionality
- (ii) negligible divergence
- (iii) high intensity
- (iv) high degree of coherence and
- (v) high monochromaticity.

(i) Directionality: The conventional light sources emit light uniformly in all directions. When we need a narrow beam in a specific direction, we obtain it by placing a slit in front of the source of light.

In case of laser, the active material is in a cylindrical resonant cavity. Any light that is travelling in a direction other than parallel to the cavity axis is eliminated and only the light that is travelling parallel to the axis is selected and reinforced. Light propagating along the axial direction emerges from the cavity and becomes the laser beam. Thus, a laser emits light only in one direction.

(ii) Divergence: Light from conventional sources spreads out in the form of spherical wave fronts and hence it is highly divergent.

On the other hand, light from a laser propagates in the form of plane waves. The light beam remains essentially as a bundle of parallel rays. The small divergence that exists is due to the diffraction of the beam at the exit mirror. A typical value of divergence of a He-Ne laser is 10^{-3} radians. It means that the diameter of the laser beam increases by about 1 mm for every meter it travels. The extent of divergence can be estimated in a simple way as follows:

If the diameters of spot produced by the laser on a screen which is held at two different distances from the laser are measured, then the angle of divergence is given by

$$\phi = \frac{d_2 - d_1}{l_2 - l_1} \quad (24.31)$$

where d_1 is the spot diameter at the distance l_1 and d_2 is the spot diameter at the distance l_2 .

Example 24.5: Calculate the divergence of light beam issuing out of He-Ne laser, which produces spot diameters of 4 mm and 6 mm at 1m and 2m distances respectively.

Solution: Beam divergence $\theta = \frac{d_2 - d_1}{2(l_2 - l_1)} = \frac{(6 - 4) \times 10^{-3} \text{ m}}{2(2 - 1) \text{ m}} = 10^{-3} \text{ radians}$

(iii) Intensity: The intensity of light from a conventional source decreases rapidly with distance as it spreads out in space. Laser emits light in the form of a narrow beam with its

energy concentrated in a small region of space. Therefore, the beam intensity would be tremendously large and stays constant with distance. The intensity of a laser beam is approximately given by

$$I = \left[\frac{10}{\lambda} \right]^2 P \quad (24.32)$$

where P is the power radiated by the laser.

To obtain light of same intensity from a tungsten bulb, it would have to be raised to a temperature of 4.6×10^6 K.

Example 24.6: A 10 mw laser has a beam diameter of 1.6 mm. What is the intensity of the light assuming that it is uniform across the beam?

Solution: Intensity of light is given by

$$\begin{aligned} I &= \frac{\text{Power of the laser}}{\text{Area of cross section of the beam}} = \frac{P}{A} \\ &= \frac{10 \times 10^{-3} \text{ W}}{3.143(0.8 \times 10^{-3} \text{ m})^2} = 4.97 \text{ kW/m}^2 \end{aligned}$$

Example 24.7. A laser beam has a power of 50mW. It has an aperture of $5 \times 10^{-3} \text{ m}$ and wavelength 7000\AA . The beam is focused with a lens of focal length of 0.2m. Calculate the areal spread and intensity of the image.

Solution: Angular spread of the beam is $d\theta = \frac{\lambda}{d} = \frac{7000 \times 10^{-10} \text{ m}}{5 \times 10^{-3} \text{ m}} = 1.4 \times 10^{-4} \text{ rad}$.

$$\text{Areal spread} = (d\theta \times f)^2 = (1.4 \times 10^{-4} \text{ rad} \times 0.2 \text{ m})^2 = 4 \times 10^{-9} \text{ m}^2.$$

$$\text{Intensity} = \frac{\text{Power}}{\text{Area}} = \frac{50 \times 10^{-3} \text{ watt}}{4 \times 10^{-9} \text{ m}^2} = 12.5 \text{ MW/m}^2.$$

(iv) Coherence: The light that emerges from a conventional light source is a jumble of short wave trains which combine with each other in a random manner. The resultant light is incoherent. Coherence length is one of the parameters used as a measure of coherence. In case of a laser a large number of identical photons are emitted through stimulated emissions and therefore they will be in phase with each other. The resultant light exhibits a high degree of coherence.

The coherence length of light from a sodium lamp, which is a traditional monochromatic source, is of the order of 0.3 mm. On the other hand the coherence length of light emitted by an ordinary helium-neon laser is about 100 m.

(v) Monochromaticity: If light coming from a source has only one frequency (single wavelength) of oscillation, the light is said to be *monochromatic* and the source a *monochromatic source*. Light from traditional monochromatic sources spreads over a wavelength range of 100 Å to 1000 Å. On the other hand, the light from lasers is highly monochromatic and contains a very narrow range of a few angstroms (< 10 Å).

24.13 APPLICATIONS

Lasers find application in almost every field. They are used in mechanical working, industrial electronics, entertainment electronics, communications, information processing, and even

in wars to guide missiles to the target. Lasers are used in CD players, laser printers, laser copiers, optical floppy discs, optical memory cards etc. We discuss here a few of the applications in industry.

The large intensity that is possible in the focused output of a laser beam and its directionality makes laser an extremely useful tool for a variety of industrial applications.

Welding: Welding is the joining of two or more pieces into a single unit. If we consider welding of two metal plates, the metal plates are held in contact at their edges and a laser beam is made to move along the line of contact of the plates. The laser beam heats the edges of the two plates to their melting points and causes them to fuse together where they are in contact. The main advantage of the laser welding is that it is a contact-less process and hence there is no possibility of introduction of impurities into the joint. In the process, the work-pieces do not get distorted, as the total amount of input is very small compared to conventional welding processes. The heat-effected zone is relatively small because of rapid cooling. Laser welding can be done even at difficult to reach place. CO₂ lasers are used in welding thin sheets and foils.

Drilling: The principle underlying drilling is the vaporization of the material at the focus of the beam. With lasers, one can drill holes as small as 10 μm in diameter. For drilling, the energy must be supplied in such a way that rapid evaporation of material takes place without significant radial diffusion of heat into the work piece. The vaporized material is removed with the help of a gas jet. Pulsed ruby and neodymium lasers are commonly used for drilling holes of small l/D ratio, where l is the thickness of the work and D is the hole diameter.

Hardening: Heat treatment is the process, which is done for sometime to harden metals and certain other materials. Heat treatment is common in the tooling and automotive industry. Heat-treating converts the surface layer to a crystalline state that is harder and more resistant to wear. In general CO₂ lasers of about 1 kW output power operating in cw mode are used for heat treatment. As metals are more reflecting at IR frequencies, a heat absorbing coating such as graphite or zinc phosphate is applied on the surface of the work piece to help it absorb laser energy more efficiently. Laser heat treatment requires a low amount of energy input to the work piece. Laser processing is advantageous as it can provide selective treatment of the desirable areas. Heat treatment is used to strengthen cylinder blocks, gears, camshafts etc in the automobile industry. As the method is a non-contact method, stress is not induced in the work-pieces.

Electronics Industry: Electronics industry uses lasers in the manufacture of electronic components and integrated circuits. Lasers have been used to perforate and divide silicon slices having several hundred circuits. They are also used for the isolation of faulty components in a large integrated circuit by disconnecting the conducting paths by evaporation. Trimming of thick and thin film resistors using lasers is a very common application.

Measurement of atmospheric pollutants: Laser is a very useful tool for the measurement of the concentrations of various atmospheric pollutants such as N₂, CO, SO₂ etc gases and particulate matter such as dust, smoke and flyash. Conventional methods of pollution measurements require that samples of pollutants are to be collected for chemical analysis. Therefore, these methods cannot give real-time data. In contrast, laser methods permit measurements by remotely sensing the composition of atmosphere without the necessity of sample collection or chemical processing.

In one of the laser techniques, the light scattered by pollutants is studied. A pulsed laser is used as the source of light and the light scattered back is detected by a photodetector. The

distance to particulate matter and the concentrations of particulate matter is obtained in this method. The distance is inferred from the time that light takes to travel up to the pollutant region and to return back. This technique is known as LIDAR which stands for light detection and ranging. The principle is very much similar to that of RADAR. The method helps in determining the concentration of particulate matter as a function of distance. However, this method cannot provide any information regarding the nature of the scattering particles. It is mainly useful in knowing the distribution of atmospheric pollutants in different vertical sections and in monitoring their variations. Environmental agencies measure concentrations of harmful gases such as SO_2 and NO_2 using this method.

Another technique uses study of absorption of light beam by pollutants. The existence of specific gases in the atmosphere is detected using absorption spectroscopy techniques. A laser beam is transmitted through polluted sample and the attenuation of intensity of light due to absorption in the sample is detected and recorded. Each chemical absorbs light of characteristic wavelengths and from the absorption spectrum, its existence can be inferred.

A third method uses Raman effect to detect the pollutants. The Raman effect involves scattering of light by gas molecules accompanied by a shift in the wavelength of light. Raman shifts are characteristic of each molecular species. Hence, analysis of backscattered laser light reveals the constituents of the gas sample. The ozone concentration high in the atmosphere is determined using this technique.

QUESTIONS

1. What is LASER? How does a laser beam differ from a point source of light? Mention any three engineering applications of laser.
2. How will you differentiate laser light from ordinary light? Explain in brief.
(Amaravati Univ.,2005, 2006)
3. How is laser light different from an ordinary light? **(R.T.M.N.U., 2007), (C.S.V.T.U.,2007)**
4. Explain with neat diagram, the processes of absorption of light, spontaneous emission and stimulated emission of light. What are the necessary conditions for their occurrence? Why does spontaneous emission dominate over stimulated emission at normal temperature?
5. With the help of neat sketches, explain the three quantum processes that may occur when light radiation interacts with matter. **(R.T.M.N.U., 2007)**
6. Explain with neat diagram, the processes of absorption of light, spontaneous emission and stimulated emission of radiation. **(Amaravati Univ.,2004)**
7. With the help of neat sketches, explain the three quantum processes that may occur when light radiation interacts with matter.
8. Explain spontaneous and stimulated emission of radiation.
(Amaravati Univ.,2005), (Calicut Univ.,2005)
9. With the help of a well labeled diagram explain the interaction of matter and radiation in the processes of (i) absorption (ii) spontaneous emission and (iii) stimulated emission. Which of these processes is maximized for laser operation?
10. Explain the terms: Spontaneous and stimulated emission. Which of the two emission processes is maximized in LASER operation? How?
11. Explain the terms: stimulated emission, population inversion, pumping and metastable states. Highlight their role in working of a laser. **(R.T.M.N.U., 2007)**
12. Explain in laser:
 - (i) Why the active media should have preferably broad absorption band?
 - (ii) What is the role of metastable state?

13. Explain the terms:
 (a) Stimulated mission (b) Population inversion
 (c) Metastable state. (d) Optical pumping (R.T.M.N.U., 2006), (C.S.V.T.U.,2009)
14. What are Einstein's coefficients? Explain them.
15. State the necessary condition for stimulated emission and explain the Einstein's A and B coefficients. (RGPV, 2008)
16. Derive the relation between Einstein's "A" and "B" coefficients. (G.T.U.,2009)
17. Explain the conditions for light amplification.
18. What is popular inversion? Explain the necessity of population inversion for lasing. (V.T.U.,2008)
19. What is popular inversion? Give the applications of laser. (Amaravati Univ.,2003)
20. What is population inversion and how can it be achieved? (Anna Univ., 2006)
21. Why is popular inversion between two atomic levels necessary for laser action to occur?
 (Amaravati Univ.,2005)
22. Explain the concept of population inversion. What is meant by pumping? Discuss in brief optical pumping.
 (Amaravati Univ.,2006)
23. Explain with sketches the basic principle of operation of lasers. (V.T.U.,2007)
24. Explain the basic principles of laser action. (Calicut Univ.,2007)
25. Explain the working of a resonant cavity and its role in laser operation. (R.T.M.N.U., 2007)
26. Distinguish between the following:
 (i) Spontaneous and stimulated emission
 (ii) Three level and four level lasers. (R.T.M.N.U., 2006)
27. What is resonant cavity? Highlight its importance in the production of laser radiation.
 (R.T.M.N.U., 2005)
28. Why is the optical resonator required in lasers? Illustrate your answer with neat sketches.
29. Explain the role of end mirrors in a laser. (R.T.M.N.U., 2006)
30. What is the lifetime of charge carrier in metastable state? (G.T.U.,2009)
31. What is metastable state? What role do such states play in the operation of laser?
 (Amaravati Univ.,2006)
32. What is meant by active material in laser? (Anna Univ., 2004)
33. Explain briefly pumping schemes used in Laser. Explain why two level laser system is not possible. (RGPV, 2008)
34. Explain in brief three and four level pumping scheme. Why four level scheme is preferred over three level scheme?
 (R.T.M.N.U., 2007)
35. Explain three level and four level laser systems. What are the advantages of four level laser system over three level laser system?
 (V.T.U.,2008)
36. What do you understand by solid-state laser?
37. Explain the production of lasers by Ruby crystal. (Calicut Univ.,2005)
38. Explain the working of ruby laser with the help of neat energy level diagram.
 (R.T.M.N.U., 2005, 2007)
39. Explain the lasing action with the help of three level energy diagram. Give any two applications of lasers.
 (Amaravati Univ.,2004)
40. Describe the construction and working of Nd:YAG laser.
 (G.T.U., 2009), (RGPV, 2007), (Anna Univ., 2007)
41. Explain laser action in three level system. (Amaravati Univ., 2002)
42. Explain why two level pumping is not suitable for obtaining population inversion.
 (C.S.V.T.U.,2005), (RGPV, 2007)

43. What is the main drawback of ruby laser? Explain the operation of a gas laser with the essential components. How stimulated emission takes place with the exchange of energy between helium and neon atoms?
(C.S.V.T.U.,2008)
44. Explain the construction and working of He-Ne laser with the help of energy level diagram.
(V.T.U.,2007, 2008), (C.S.V.T.U.,2006)
45. Explain the working of He-Ne laser with the help of a neat energy level diagram.
(R.T.M.N.U., 2006)
46. Explain with a neat diagram the principle, construction and working of He-Ne laser. What are its merits and demerits?
(Calicut Univ.,2005)
47. What is the active material in a He-Ne laser? How population inversion is achieved in a He-Ne laser?
48. With the help of neat energy level diagram, explain how the population inversion is achieved in He-Ne laser. Explain why increase in diameter of He-Ne tube may reduce lasing efficiency.
(R.T.M.N.U., 2006)
49. In He-Ne laser, what is the function of He atoms? Why is it necessary to use a tube of narrow diameters?
50. Describe the construction and working of He-Ne laser.
(RGPV, 2007), (R.T.M.N.U., 2007), (Calicut Univ.,2007)
51. Explain the action of a Helium-neon laser. How is it superior to a Ruby laser?**(UPTU, Lucknow)**
52. (a) Explain the terms:
(i) Spontaneous emission and (ii) Stimulated emission
(b) Distinguish between Ruby laser and He-Ne laser
(c) Write the applications of lasers in medical field.
(JNTU, 2010)
53. (a) Explain the terms :
(i) Population inversion (ii) Stimulated emission (iii) Temporal coherence
(b) Describe the working of solid state Ruby Laser with the help of neat energy level diagram.
(RTMNU, 2010)
54. Differentiate between ‘spontaneous emission’ and ‘stimulated emission’ of radiation. Obtain a relation between transition probabilities of the two. Explain the applications of lasers.
(RGPV, 2010)
55. What are laser characteristics? Describe the principle and working of He-Ne laser. Why a narrow discharge tube is used in He-Ne laser?
(C.S.V.T.U.,2006, 2007)
56. Describe the construction and working of CO₂ laser.
(Anna Univ., 2007)
57. Explain the modes of vibrations of CO₂ molecule. Describe the construction and working of CO₂ laser with necessary diagrams.
58. Explain with a neat diagram the principle, construction and working of a semiconductor laser. What are its merits and demerits?
(Calicut Univ.,2005)
59. What are semiconductor diode lasers? Describe with energy band diagram the construction and working of semiconductor diode laser. Mention the uses of diode lasers.
(V.T.U.,2007), (Anna Univ., 2005)
60. What is the role of length of resonant cavity in supporting different frequency modes?
61. Explain longitudinal modes and transverse modes in a laser beam.
62. List out the properties of Laser.
(Calicut Univ.,2007), (G.T.U.,2009)
63. State important characteristics of laser beam. Explain construction and working of any one laser.
(RGPV, 2008)

PROBLEMS

1. A three level laser emits laser light at a wavelength of 5500\AA .

 - In the absence of optical pumping, what will be the equilibrium ratio of the population of the upper level to that of the lower level? Assume $T = 300$ K.
 - At what temperature for the conditions of (i) above would the ratio be $\frac{1}{2}$?
 - What conclusion can be drawn on the basis of the results obtained by you in relation to choice of pumping mechanism?

2. A typical He-Ne laser emits radiation of $\lambda = 6328\text{\AA}$. How many photons per second would be emitted by a one milli-watt He-Ne laser? **(Ans: 3×10^{15})**

3. Find the relative populations of the two states in a ruby laser that produces a light beam of wavelength 6943\AA at 300 K and 500 K. **(Ans: 8.75×10^{-19})**

4. Calculate the ratio of spontaneous emission to stimulated emission if the wavelength of the radiation is 5500\AA at 2000K . **(Ans: 4.6×10^5)**

5. For the He-Ne laser ($\lambda = 6328\text{\AA}$), estimate the broadening of the wavelength due to Heisenberg uncertainty principle, assuming that the metastable state has a lifetime of 1 ms. **(Ans: $1.06 \times 10^{-19}\text{ m}$)**

6. If the pulse width of a laser ($\lambda = 10640\text{\AA}$) is 25 ms and average output power per pulse is 0.8W , how many photons does each pulse contain? **(Ans: 107×10^{15})**

7. In an experiment on the divergence of the He-Ne laser beam, the diameters of the beam spot were measured to be 4mm and 6mm at distances of 1m and 2m respectively. Determine the divergence of the beam. **(Ans: 1 milliradian)**

8. A ruby laser produces a beam of spot diameter 5mm and divergence of 1 milliradian. It is directed towards the moon. The earth-to-moon distance is 376284 km . Compute the area on the moon that would be illuminated by the laser beam. **(Ans: 445000 km^2)**

9. Estimate the angular spread of a laser beam of wavelength 6930\AA due to diffraction, if the beam emerges through a 3mm diameter mirror. How large would be the diameter of this beam when it is incident on a satellite 300 km above the earth? **(Ans: 84.6m)**

10. A laser beam $\lambda = 6000\text{\AA}$ on earth is focused by a lens of diameter 2m on to a crater on the moon. The distance of the moon is $4 \times 10^8\text{ m}$. How big is the spot on the moon? **(Ans: $3 \times 10^{-7}\text{ rad}, 1.4 \times 10^4\text{m}^2$)**

11. A ruby laser emits light of wavelength 694.4 nm . If a laser pulse is emitted for $1.2 \times 10^{-11}\text{ s}$ and the energy release per pulse is 0.15J ,
 - What is the length of the pulse and
 - How many photons are there in each pulse?

12. Compute the coherence length of yellow light with 5893\AA in 10^{-12} seconds pulse duration. Find also the bandwidth. **(RTMNU, 2010)**

CHAPTER

25

Holography



Hologram

25.1 INTRODUCTION

Images of objects are generally obtained using photographic method. In this method a lens focuses the light reflected from a three-dimensional object onto a photographic film where a two-dimensional image of the object is formed. A negative is first obtained by developing the film and then a positive is obtained through printing. The positive print is a two-dimensional record of light intensity received from the object. It, thus, contains information about the square of the amplitude of the light wave that produced the image but information about the phase of the wave is not recorded and is lost.

In 1948 Dennis Gabor outlined a two-step lensless imaging process. It is radically a new technique of photographing the objects and is known as **wave front reconstruction**. This technique is also called **holography**. The word ‘holography’ is formed by combining parts of two Greek words: ‘holos’, meaning “whole”, and ‘graphein’ meaning “to write”. Thus holography means writing the complete image. Holography is actually a recording of interference pattern formed between two beams of coherent light coming from the same source. In this process both the amplitude and phase components of light wave are recorded on a light sensitive medium such as a photographic plate. The recording is known as a

hologram. Holography required an intense coherent light source. Laser was not available when Gabor formulated the idea of holography. Holographic technique became a practical proposition only after the invention of lasers. Leith and Upatnick prepared laser holograms for the first time. In this chapter we discuss the fundamental concept of holography.

25.2 PRINCIPLE OF HOLOGRAPHY

Holography is a two-step process. First step is the **recording** of hologram where the object is transformed into a photographic record and the second step is the **reconstruction** in which the hologram is transformed into the image. Unlike in the conventional photography, lens is not required in either of the steps. A hologram is the result of interference occurring between two waves, an object beam which is the light scattered off the object and a coherent background, the reference beam, which is the light reaching the photographic plate directly. In Gabor's original experiments, the reference beam and object beams were *coaxial*. Further advance was made by Leith and Upatnick, who used the reference beam at an *offset angle*. That made possible the recording of holograms of three-dimensional objects.

The off-axis arrangement for generating and viewing holograms is described here.

25.2.1 Recording of the Hologram

In the off-axis arrangement a broad laser beam is divided into two beams, namely a **reference beam** and an **object beam** by a beam splitter (Fig. 25.1). The reference beam goes directly to the photographic plate. The second beam of light is directed onto the object to be photographed.

Each point of the object scatters the incident light and acts as the source of spherical waves. Part of the light, scattered by the object, travels towards the photographic plate. At the photographic plate the innumerable spherical waves from the object combine with the plane light wave from the reference beam. The sets of light waves are coherent because they are from the same laser. They interfere and form interference fringes on the plane of the photographic plate. These interference fringes are a series of zone-plate like rings,

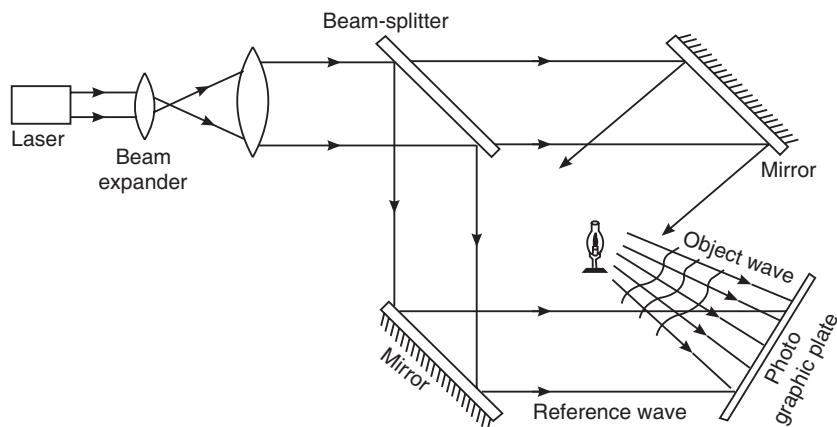


Fig. 25.1



Fig. 25.2. The image on the film for a transmission hologram

but these rings are also superimposed, making a complex pattern of lines and swirls. The developed negative of these interference fringe-patterns is a hologram. Thus, the hologram does not contain a distinct image of the object but carries a record of both the intensity and the relative phase of the light waves at each point (Fig. 25.2).

25.2.2 Reconstruction of the Image

Whenever required, the object can be viewed. For **reconstruction** of the image, the hologram is illuminated by a parallel beam of light from the laser (Fig. 25.3). Most of the light passes straight through, but the complex of fine fringes acts as an elaborate diffraction grating. Light is diffracted at a fairly wide angle. The diffracted rays form two

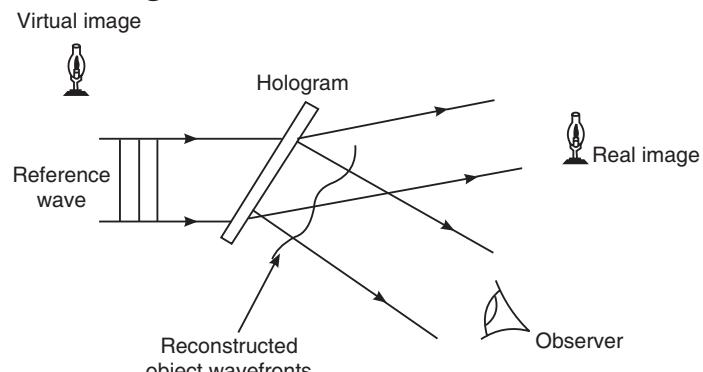


Fig. 25.3

images: a *virtual image* and a *real image*. The virtual image appears at the location formerly occupied by the object and is sometimes called as the true image. The real image is formed in front of the hologram. Since the light rays pass through the point where the real image is, it can be photographed. The virtual image of the hologram is only for viewing. Observer can move to different positions and look around the image to the same extent that he would be able to, were he looking directly at the real object (see Fig. 25.7). This type of hologram is known as a *transmission hologram* since the image is seen by looking through it. The three dimensional image is seen suspended in midair at a point which corresponds to the position of the real object which was photographed.

25.3 COAXIAL HOLOGRAPHY

The original technique adopted by Gabor for recording hologram was a **coaxial arrangement** (Fig. 25.4), where he made use of a mercury discharge lamp and *collinear* object and reference beams. In this arrangement both virtual and real images are on the same axis. The real image is located in front of the virtual image. Thus an observer focusing on one image, always sees it accompanied by the out-of-focus twin image. As such it is inconvenient for viewing or photographing the image.

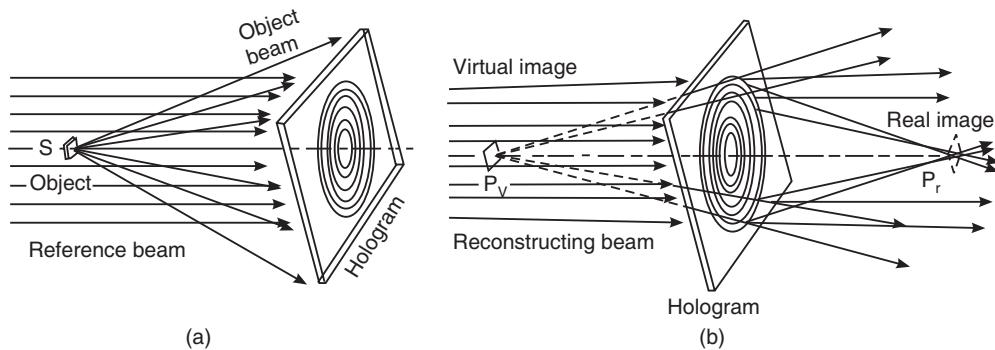


Fig. 25.4: In-line holography (a) Generation (b) viewing

25.4 OFF-AXIS HOLOGRAPHY

In 1962, Leith and Upatnieks demonstrated a technique which made it possible to separate the twin images. In this technique a separate coherent reference wave is allowed to fall on the hologram plate during the recording process, at *an offset angle* to the beam from the object (see Fig. 25.1). The exposed plate is developed by normal photographic procedures so that the amplitude transmittance of the plate after development is proportional to the exposure.

The advantage of the off-axis configuration is that it generates virtual and real images angularly separated from each other and from the direct beam also (see Fig. 25.3).

25.5 THEORY

The general theory of holography is much involved and cumbersome. We illustrate here it by taking the simple example of a point object in a coaxial configuration.

Let the light beam from a coherent source illuminate a point object P (see Fig. 25.5). The beam consists of plane waves. Most of the plane waves reach the photographic plate directly. Part of the light is scattered by the point object and spherical waves are produced. They also reach the photographic plate. The plane waves of the reference beam and spherical waves of the object beam superpose at the plane of the photographic plate and produce interference.

We may write the optical field arriving at point O on the photographic plate as

$$E = E_R + E_S \quad (25.1)$$

where E_R is the field due to the reference beam and E_S is the field scattered from the object. The scattered field E_S is not simple, both amplitude and phase vary greatly with position. The scattered wave fronts are spherical and concentric around the point of origin. We represent the field of the scattered wave front by

$$E_S = \frac{E_0}{r_0} \exp[i(k r_0 - \omega t)] \quad (25.2)$$

and the field E_R by the plane wave

$$E_R = E_r \exp[i(k z_0 - \omega t)] \quad (25.3)$$

where $r_0 = PO$ and z_0 is the distance from P to the plate. The intensity at O is

$$\begin{aligned} I &= |E_R + E_S|^2 \\ &= |E_R|^2 + \frac{|E_S|^2}{r_0^2} + \frac{E_S E_r^*}{r_0} \exp[ik(r_0 - z_0)] + \frac{E_S^* E_r}{r_0} \exp[ik(z_0 - r_0)] \end{aligned} \quad (25.4)$$

We combine the last two terms of the above equation and write it as

$$I = |E_R|^2 + \frac{|E_S|^2}{r_0^2} + K \frac{\cos[k(r_0 - z_0) + \phi]}{r_0} \quad (25.5)$$

The total intensity I is a function of cosine term and shows a series of maxima and minima. Thus, the interference of the spherical wave E_S and plane wave E_R produces a set of circular interference fringes. If we assume that the plate response is proportional to the intensity I , the power transmission of the plate, T^2 is given by

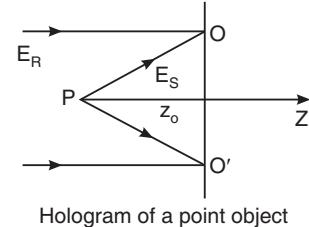


Fig. 25.5

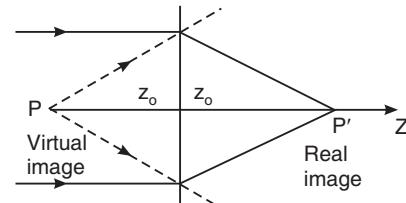


Fig. 25.6

$$T^2 = 1 - \alpha I \quad (25.6)$$

where α is a constant. Equ.(25.6) can be approximated as

$$T \cong 1 - \frac{1}{2} \alpha I \quad (25.7)$$

When the hologram is illuminated by the reference beam, the field of the transmitted wave may be written as

$$\begin{aligned} E = TE_2 &= \left[1 - \frac{\alpha}{2} I \right] E_r \exp[i(kz_0 - \omega t)] \\ &= \left[1 - \frac{\alpha}{2} |E_r|^2 - \frac{\alpha |E_s|^2}{2 r_0^2} \right] E_r \exp[i(kz_0 - \omega t)] \\ &\quad - \frac{\alpha E_s^* E_r}{2 r_0} \exp[ik(r_0 - z_0)] E_r \exp[i(kz_0 - \omega t)] \\ &= \left[1 - \alpha |E_r|^2 - \frac{\alpha |E_s|^2}{2 r_0^2} \right] E_r \exp[i(kz_0 - \omega t)] \\ &\quad - \frac{\alpha E_s |E_r|^2}{2 r_0} \exp[i(kr_0 - \omega t)] - \frac{\alpha E_s^* E_r^2}{2} \exp[i(2kz_0 - kr_0 - \omega t)] \end{aligned} \quad (25.8)$$

The first term in equ.(25.8) represents the incident plane wave with some attenuation.

The second term represents a spherical wave identical with that emitted by the object except for a constant factor. The wave surface when projected back appears to have come from an apparent object located at the place where the original object was located. This is the *virtual image* of the object.

The third term represents also a spherical wave, which is identical to the original wave but converges at a point P' . A *real image* is produced at P' which can be photographed without a lens. The hologram thus produces both a real image P and a virtual image P' (Fig. 25.6).

25.6 HOLOGRAMS

Holograms are true three-dimensional images. This is evidenced by the fact that one can move his head while viewing the image and see it in a different perspective. It reveals part of the image which was hidden at another viewing angle.

For example, three images are shown below (Fig. 25.7). They are from the *same* hologram but are obtained by looking through the hologram at *different angles*. Note that the pawn appears in different perspective in front of the king behind it.



Fig. 25.7

Recording a hologram is different from taking a photograph.

- (i) In hologram, it is necessary that a coherent light source like a laser be used to illuminate the object.
- (ii) There is a second beam of coherent light which strikes the film on which the hologram is to be recorded. This is called the reference beam. The reference beam and the object beam overlap at the surface of the film and they form an interference pattern.
- (iii) A high-resolution photographic film is used for recording the fine patterns.
- (iv) Because the lines that make up the hologram are usually less than a micron, holographic recording is more sensitive to movement and vibrations compared to photographic recording. This is actually a very serious restriction. Therefore, holography technique is restricted to the laboratory where vibration isolation is created.

The fundamental difference between a **hologram** and an **ordinary photograph** is like this.

- (i) In a photograph the information is stored in an orderly fashion: each point in the object relates to a conjugate point in the image. In a hologram there is no such relationship; light from every object point goes to the entire hologram. This has two main advantages. As the observer moves sideways in viewing the hologram, the image is seen in three dimensions.
- (ii) If the hologram were shattered or cut into small pieces, each fragment would still reconstruct the whole scene, not just part of the scene.

25.6.1 Orthoscopic and Pseudoscopic Images

A hologram reconstructs two images, one real and the other virtual which are exact replicas of the object. However, the two images differ in appearance to the observer. The virtual image is produced at the same position as the object and has the same appearance of depth and the parallax as the original three dimensional object. The virtual image appears as if the observer is viewing the original object through a window defined by the size of the hologram. This image is known as **orthoscopic image**.

The real image is also formed at the same distance from the hologram, but in front of it. In the real image, however the scene depth is inverted. This is due to the fact that the corresponding points on the two images (virtual and real) are located at the same distances from the hologram. The real image is known as **pseudoscopic image** and does not give a pleasing sensation as we do not come across objects with inverted depths in normal life.

25.7 IMPORTANT PROPERTIES OF A HOLOGRAM

1. In an ordinary photograph each region contains separate and individual part of the original object. Therefore, destruction of a portion of a photographic image leads to an irreparable loss of information corresponding to the destroyed part. On the other hand, in a hologram each part contains information about the entire object. From even a small part of the hologram the entire image can be reconstructed if only with a reduced clarity and definition of the image. Therefore, a hologram is a reliable medium for data storage.
2. It is not useful to record several images on a single photographic film. Such a record cannot give information about any of the individual images. On the other hand, several images can be recorded on a hologram. Therefore the information holding capacity of a hologram is extremely high. While a 6×9 mm photograph can hold one printed page, a hologram of the same size can store up to 300 such pages.

3. On a hologram information is recorded in the form of interference pattern. The type of the pattern obtained depends on the reference beam used to record the hologram. The information can be decoded only by a coherent wave identical to that of the reference wave. The reference wave can be chosen appropriately. Consequently without the knowledge of the shape of the reference wave front the information encoded in the form of interference pattern on the hologram cannot be deciphered.
4. The reconstruction of the image of the hologram can be done with reference beam of any wavelength if it is coherent and identical to the original reference beam. If the wavelength λ of the reconstructing beam is greater than that λ_0 of the reference beam, the reconstructed image will be a magnified image. The magnification will be proportional to the ratio of the two wavelengths.

25.8 CLASSIFICATION OF HOLOGRAMS

Holograms may be classified in a number of different ways depending on their thickness, method of recording, method of reconstruction etc.

(i) Thin Holograms or Plane Holograms

Holograms may be thin (plane) or thick (volume) (see Fig. 25.8). A hologram may be regarded as *thin* if its emulsion thickness is much less than the fringe spacings. Otherwise it is called a *thick hologram*. Thick hologram is also known as *volume hologram*.

Thin holograms produce several orders (*i*) zero order which is the directly transmitted reference beam, (*ii*) the first order diffraction producing virtual image, (*iii*) the minus first order diffraction equal in intensity to the first order producing the conjugate image and (*iv*) higher orders of decreasing intensity.

(ii) Volume hologram

In 1962 Yuri Denisyuk used a process similar to Lippmann colour photography. In this method the object wave is reflected from the object and propagates backward and overlaps the incoming reference wave. The two waves form standing wave pattern. The fringes are recorded by the photo-emulsion throughout its entire thickness to form a **volume hologram**. The hologram may be regarded as a three-dimensional grating. In the volume holograms there is an interdependence of the wavelength and the scattering angle, because the scattering follows Bragg's law, $2d \sin \theta = m\lambda$. Therefore, by successively changing the incident angle or wavelength, a number of holograms can be stored in the medium. Different and mutually incoherent laser beams may be used to produce different component holograms of the object and when they are illuminated, a multicolored image is seen. A consequence of Bragg condition is that the volume hologram reconstructs the virtual image at the original position of the object if the reconstruction beam exactly coincides with the reference beam. However, the conjugate image and higher order diffractions are absent.

(iii) Amplitude and Phase Holograms

Holograms recorded in photographic emulsions change both the amplitude and the phase of the illuminating wave. The shape of the recorded fringe planes depend on the relative phase

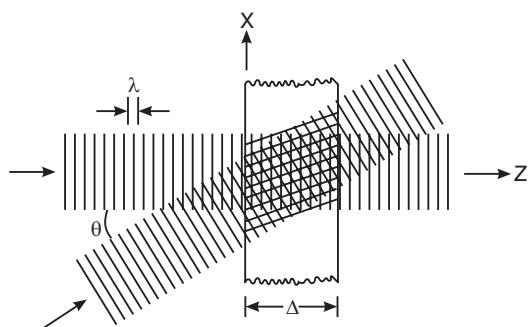


Fig. 25.8

of the interfering beams. Consequently, the reconstructed wave is reflected from the hologram according to the density of the silver deposited with the amplitude variation proportional to the amplitude of the object wave. Similarly the phase of the reconstruction wave is modulated in proportional to the phase of the object wave. Thus both amplitude and phase of the object wave are reproduced. An **absorption type** hologram produces a change in the amplitude of the reconstruction beam. The **phase type hologram** produces phase changes in the reconstruction beam due to a variation in the refractive index or thickness of the medium. Phase holograms have the advantage over amplitude holograms that no energy is dissipated within the hologram medium and have higher diffraction efficiency.

(iv) Transmission Holograms

For preparing a transmission hologram, both the reference beam and object beam are made incident from the same side of the recording medium, as shown in Fig. 25.9 (a). These two beams form interference pattern and the spacing between the maxima is given by

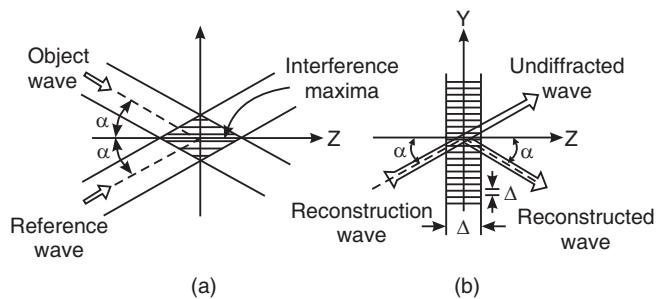


Fig. 25.9

$$\Lambda = \frac{\lambda}{2n \sin \alpha} \quad (25.9)$$

where n is the index of refraction of the recording medium and α is the angle shown in Fig. 25.9 (a). In the reconstruction process, the interference fringes act like reflective layers to the incident reference wave. These reflecting layers are perpendicular to the plane of the hologram. The incident reconstruction wave is reflected toward the other side of the hologram. The reconstructed beam appears as if the beam is transmitted through the hologram.

(v) Reflection Holograms

The distinction between transmitting and reflection holograms is due to the angle between the reference and object beams. In case of reflection hologram, the object beam is introduced at almost 180° with respect to the reference beam. Further, reflection holograms are often thicker than transmission holograms. There is more physical space for recording interference fringes. The interference pattern consists of vertical strips at a spacing given by

$$\Lambda = \frac{\lambda}{2n \cos \alpha} \quad (25.10)$$

The reflecting layers formed in the recording medium are parallel to the surface of the hologram. One can think of holograms that are made this way as having multiple **layers** that are only about half a wavelength deep. In the reconstruction process, some of the incident light reflects back toward the light source, and some continues to the next layer, where the process repeats. The light from each layer interferes with the light in the layers above it. This is known as the **Bragg effect**, and it is a necessary part of the reconstruction of the object beam in reflection holograms. The reconstructed wave is on the same side of the hologram and thus appears as if light is reflected from the hologram.

The Bragg effect can also change the way the hologram reflects light, especially in holograms that one can view in white light. At different viewing angles, the Bragg effect can be different for different wavelengths of light. This means that you might see the hologram as one color from one angle and another color from another angle. The Bragg effect is also one of the reasons why most novelty holograms appear green even though they were created with a red laser.

Reflection holograms are more useful since a laser is not required for reconstruction. They can be viewed in white light.

(vi) White-Light Reflection Holography

This method is developed by Stroke and Labeyrie. In this scheme, the hologram is generated using coherent light but in the reconstruction process an ordinary white-light beam having a wave front similar to the original coherent waves is used. Using coherent sources at different wavelengths, several holograms are stored in a single film. When the hologram is illuminated by ordinary white light, a multicolored image is seen in reflection.

(vii) Colour Holography

Colour holograms are basically multiplexed holograms which produce multicolour images. They can be recorded with three wavelengths. When reconstructed with the recording wavelengths the hologram produces overlapping images in three colours producing a multicolour image. The behaviour of the reconstructed image depends on whether the hologram has been recorded in a thin medium or in a thick medium. Colour holograms recorded in a thin recording medium suffer from cross-talk. Volume holograms effectively eliminate cross-talk images utilizing Bragg effect. Both transmission and reflection volume colour holograms can be recorded in thick media. The transmission volume holograms are reconstructed with the laser beams used to record it, while the volume colour reflection holograms are reconstructed with white light due to their inherent wavelength discrimination ability.

(viii) Rainbow Holograms

Rainbow hologram is a new type of transmission hologram capable of reconstructing a bright, sharp and monochromatic image when viewed in white light. They are made by a double holographic process where an ordinary hologram such as a transmission hologram is used as the object and a second hologram is made through a slit. A horizontal slit limits the vertical perspective of the first image so that there is no vertical parallax in the resultant rainbow hologram. This slit process removes the coherence requirement on the viewing light so that full advantage can be taken of the image brightness obtained from ordinary room light, while maintaining the three-dimensional character of the image. A true-color hologram image can be observed when the hologram is viewed in the correct plane. If the viewer moves off this plane, different shades of color can still be seen, but the color will be different from that of the original object. The steps involved in making a rainbow hologram are shown in Fig. 25.11.

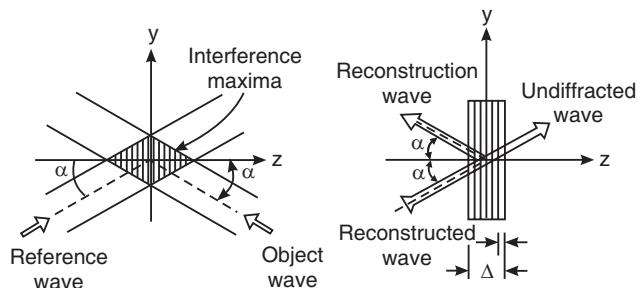


Fig. 25.10

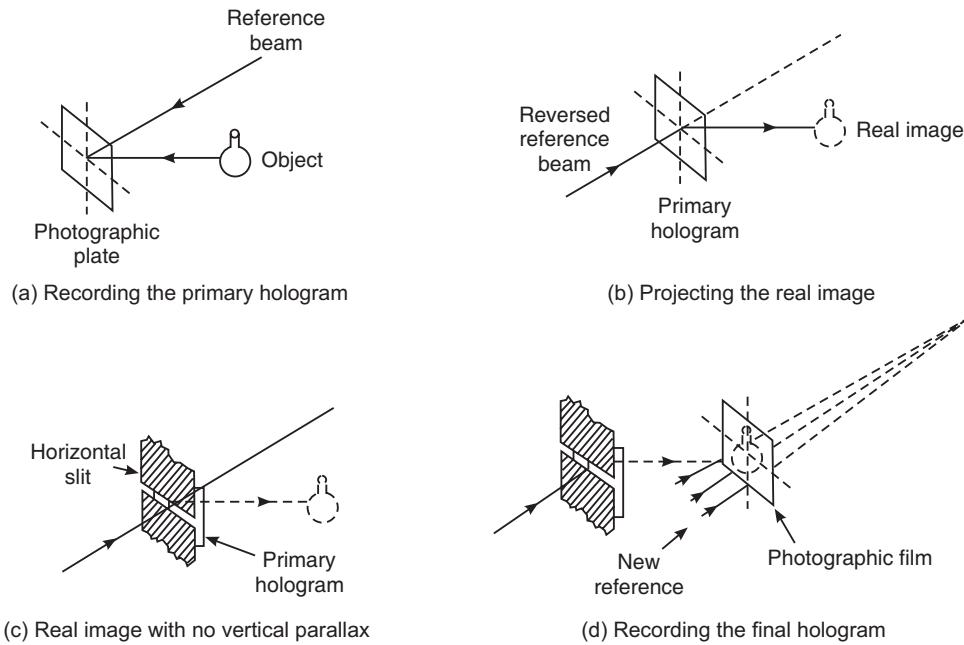


Fig. 25.11

- (a) First a conventional transmission hologram of the object is made.
- (b) The hologram is then illuminated by the conjugate of the original reference beam. It generates a diffracted wave, which is the real image of the object.
- (c) A horizontal slit is placed over the primary hologram to eliminate vertical parallax.
- (d) A second and final hologram is recorded which is used for reconstructing the object.

(ix) 360° Holograms

This type of hologram is made on a 360° circular film. A photographic film is mounted on a cylindrical surface surrounding the object to be holographed. A divergent laser beam is made to be incident on the object from the top (Fig. 25.12). A convex mirror at the bottom illuminates the object.

When the cylindrical hologram is illuminated, the virtual image is observed at the centre of the cylinder and can be viewed from all sides.

(x) Copying Holograms

Many of the commercial products bear a holographic logo and trade mark as a mark of the identity and authenticity of the manufacturer. It requires bulk production of holograms. Large number of copies of the original master hologram can be made using different ways.

Light reflected or transmitted from the master hologram and a direct beam are made to interfere on a copy plate (Fig. 25.13). This results in a copy hologram. Excellent copies can be made in this way.

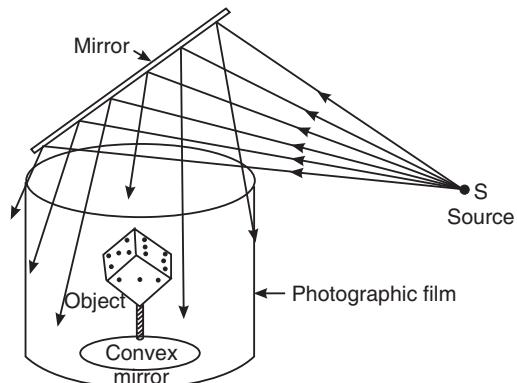


Fig. 25.12. 360° hologram

Alternately, an embossing method can be used for making copies. A phase hologram where the thickness of the emulsion varies is used in this method. This master hologram has a surface relief structure that is metallized. The

metallized hologram is used for impressing the pattern onto thin sheets of plastic.

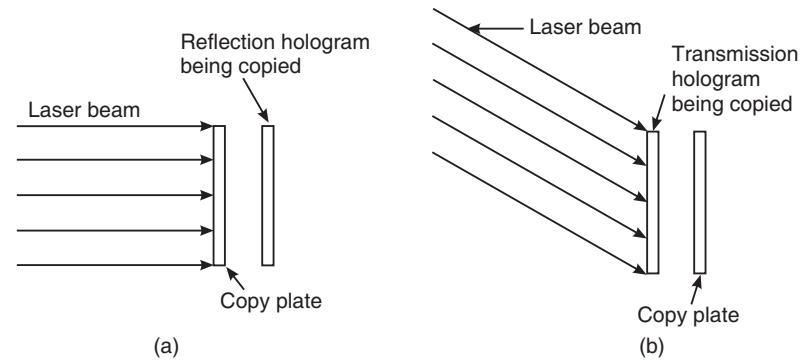


Fig. 25.13. Copy Hologram

25.9 APPLICATIONS

Holography can be used for a broad range of applications in different fields. It is not possible to describe all of them here. Only some typical applications are discussed here.

- 1. Security:** One of the most popular areas for the use of holograms is the security and product authentication. The presence of holograms indicates the authenticity of these items. They provide a powerful obstacle to counterfeiting. The security holograms have proven to be unsurpassed when added to documents, anti-counterfeiting, tamper-proofing, customizing ticket protection, identification documents including credit and phone cards, drivers licenses etc.
- 2. Three-dimensional photography:** One of the most obvious applications of holography is the production of a three-dimensional photograph, with the distance and orientation of each point of the object recorded in the image.
- 3. Microscopy:** Holography can be used in techniques of microscopy. It is possible to obtain a magnified image of an object if recording is done with light of smaller wavelength and reconstruction with light of longer wavelength. Smaller areas in an object can be examined in greater detail. This has great potential in observing micro-objects such as blood cells, amoebas, cancer affected tissues etc.
- 4. Character recognition:** Holography can also be used for character recognition. The complicated wave front from an object is generated from a hologram by the simple wave front of the reference beam. The process is reversible so that reference wave can be generated by the object wave. This principle forms the basis of holographic pattern recognition. This could be used to identify fingerprints etc.
- 5. Data storage:** Holograms can also be used for data storage devices and hence are of much use in computer technology. A large amount of information such as 10^{12} letters/digits can be stored in a cubic cm of a volume hologram. These memories have long lifetime because a small mechanical damage to the portion of a hologram will not erase the stored information.
- 6. Photolithography:** Holography is used in the production of photographic masks used to produce microelectronic circuits.
- 7. Holographic projection:** is used to display flight information at the pilot's eye level in an airplane cockpit.

8. Holographic interferometry: One of the most important applications of holography has been in interferometry. **Holographic interferometry** is an optical technique to visualize in a dark environment small deformations (200 nm to 100 μm) of objects. It is applied to objects, which are placed in a vibration-free set-up.

Holographic interferometry is used in vibrational analysis, structural analysis, stress and strain evaluation etc. There are three basic methods of holographic interferometry. They are known as real time, multiexposure and time-average holography.

Real time holography allows one to observe instantaneously the effects of minute changes in an object as some stress affects it. In this method, first the hologram of the undisturbed object is obtained. This hologram of the object is superimposed over the object subjected to some small stress (Fig. 25.14). The distortions that appear as a fringe pattern are analysed.

Multiexposure holography creates a hologram by using two or sometimes more exposures. The first exposure shows the object in an undisturbed state. Subsequent exposures are made on the same image while the object is subjected to some stress. The resulting image depicts the difference between the two states.

The technique where two exposures are made is known as **double exposure holographic interferometry**. In the technique, a hologram of the undisturbed object is first recorded on the photographic plate with an exposure to a reference wave. Then, before the hologram is removed or developed, the object is stressed and is recorded on the same photographic plate through a second exposure along with the same reference wave. After this double exposure, the hologram is developed. If the hologram is now illuminated by a reconstruction wave, there would emerge from the hologram two object waves - one corresponding to the unstressed object and the other corresponding to the stressed object. These two object waves interfere to produce interference pattern. Thus on viewing through the hologram, the object as well as the

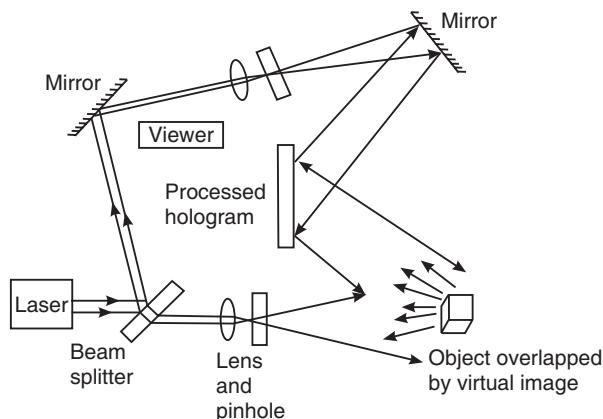
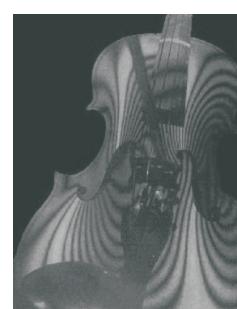


Fig. 25.14. Real time holographic interferometry



(a)



(b)

Fig. 25.15. The Zebra-like pattern of fringes on (a) a vibrating coffee cup (b) on a violin shows the nodal pattern of vibrations.

surface covered by the interference fringes caused by surface irregularities can be seen. These fringes reveal the distribution of strain in the object (Fig. 25.15).

The third technique is known as the **time –average holographic interferometry** which is used in studying vibrating objects. In this technique, a hologram of the vibrating object is prepared by exposing the photographic plate to a reference wave, for a relatively long period of time such that the vibrating object has undergone a number of oscillations during that time. The resulting hologram consists of a standing wave pattern caused by a superposition of the number of images corresponding to the successive states of vibration of the object. The bright areas in the hologram correspond to undeflected regions whereas the contour lines indicate the regions of constant amplitude of vibration. This technique is very much useful in the vibrational analysis of any vibrating system or vibrating component of a machine.

Nowadays, the above techniques are widely used in nondestructive testing. Although holography can solve many problems, it still is a relatively expensive procedure.

9. Acoustical holography

It is easy to produce coherent sound waves. Therefore, holograms can be made using ultrasonic waves initially and then visible light can be used for reconstruction of the visual image. Light waves cannot propagate considerable distances in dense liquids and solids whereas sound waves can propagate through them. Therefore, a three dimensional acoustical hologram of an opaque object can be made. By viewing such hologram in visible light the internal structure of the object can be observed. Such techniques will be highly useful in the fields of medicine and technology. In one of the techniques, two submerged coherent sound wave generators emit the reference and the signal waves scattered by an object (Fig. 25.16). On a calm surface of water, these two contributions produce ripples. The ripple pattern is the hologram. The pattern is photographed and reconstructed as and when required. As sound waves can propagate through dense liquids and solids, acoustical holography has an advantage in locating underwater submarines etc and study of internal body organs.

10. Holographic optical elements

Traditional optical elements operate on the principle of refraction. A **diffractive optical element (DOE)** operates on the principle of diffraction. DOEs can function as gratings, lenses or any other type of optical element. Large optical apertures, lightweight and lower cost are the main features of DOEs. They can offer unique optical properties that are not possible with conventional optical elements. They can be fabricated in a wide range of materials namely, aluminum, silicon, silica, plastics, etc providing the user greater flexibility in selecting the material for a particular application. DOEs can be used to perform more specialized functions, like making the panel instruments of a car visible in the windshield for increased safety. Some of the important features of DOEs are as follows.

- Several different optical elements can share the same substrate without interfering with one another. Thus, a single DOE can be used as a lens, beam splitter and spectral filter simultaneously.

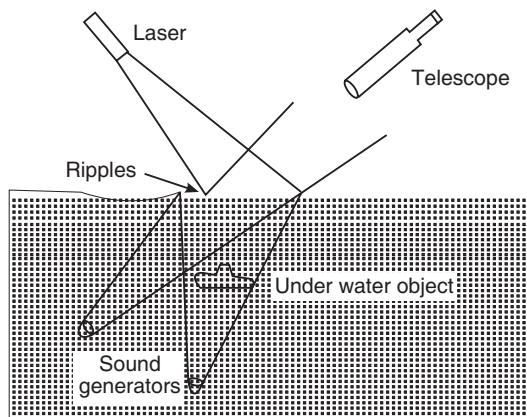


Fig. 25.16

- Diffractive elements are very light, as they are formed in thin films of a few μm thickness only.
- Because DOEs can generate unique optical functions that are not possible by conventional reflective or refractive optical elements, they provide greater flexibility in system configuration.
- At least one surface of a conventional glass lens is curved, whereas for a diffractive lens there is no such requirement. A diffractive element can be fabricated on any arbitrary shape of the substrate.
- They can be made to operate over a narrow wavelength band.
- The fabrication and replication of DOEs are relatively easy and cheap because no precision shaping of a surface is required.

25.10 MEDICAL APPLICATIONS OF HOLOGRAPHY

Holographic technique is also used in various medical applications like ophthalmology, endoscopy, otology, orthopedics and many more. Recent improvements in hologram recording techniques and the availability of tools for the interpretation of holographic interferograms and the success of holographic techniques in imaging through tissues, ophthalmology, dentistry, urology, otology, pathology, and orthopedics shows that holography may emerge as a powerful tool for medical applications. We discuss here some of the applications.

1. Holographic Endoscope: Endoscopic holography provides a powerful tool for non-contact, high resolution, 3D imaging and nondestructive measurements inside natural cavities of human body. It combines the features of holography and endoscopy. The holographic endoscopy is of two types. In one of the forms, the hologram is recorded inside the endoscope, while the other form uses an external recording device.

(a) Internal Hologram Recording Endoscope: The endoscope accommodates a miniaturized holographic setup inside the instrument and records a reflection hologram. It mainly consists of three parts; a film cartridge, a diaphragm and a single mode optical fibre (core diameter $4 \mu\text{m}$) cable. The three parts are assembled in three adjustable stainless steel tubes. The film is placed normal to the endoscope. The holograms are viewed under a powerful microscope allowing for the observation of individual cells. Due to large hologram aperture, the image with a low speckle noise and high lateral resolution is obtained. A lateral resolution of $7 \mu\text{m}$ has been obtained in the reconstructed image that shows that the technique can be used for cellular structure analysis and may even substitute biopsy in tumour diagnosis. Specific dyes can be used to enhance the contrast of the tissue before recording the holograms as has been used extensively in gynecology and gastrointestinal tract.

(b) External Hologram recording Endoscope: In the external hologram recording endoscope, a conventional endoscope is used. The system records the hologram outside the endoscope using an external reference beam. An endoscope with extremely small outer diameter can be used. In order to obtain a high signal-to-noise ratio, the holographic endoscope must use gradient-index (GRIN) rod lenses. The speckle noise is reduced by illuminating and imaging the object by the same GRIN lens. An electro-optic crystal can be used as the photographic storage device in the holographic endoscope to provide in-situ recording, reconstruction, and erasure.

Holographic endoscope has been used with success for early recognition of cancerous indurations in the wall of urinary bladder.

2. Holography in Ophthalmology: Recording of a three dimensional image of the eye was one of the earliest applications of holography in the field of ophthalmology. Any retinal detachment or intraocular foreign body can be detected. Holography can also be applied for

the measurement of corneal topography and crystalline lens changes and for the study of surface characteristics of both the nerve head and the cornea. Current methods of determining the shape of the central surface miss the central part and its periphery. The major advantage of holographic technique is the ultra high precision (sub- μm range) with which such measurements are possible. The elastic expansion of the cornea can also be measured by holographic interferometry. This information is vital for corneal surgery. The studies made so far show that holography has potential to investigate changes on the cornea, crystalline lens changes, and surface characteristics of both the nerve head and the retina.

3. Diffractive bifocal Intraocular Lens: A very useful application of diffractive optics is in the correction of refractive errors for old persons who have been operated for cataract by the use of a bifocal intraocular lens. Such persons have difficulty in changing the focus of their eyes for near distant and far distant objects. Bifocal lenses are implanted in place of the natural eye lenses. The bifocal lens is a combination of a conventional refractive lens and a diffractive lens, the former focussed to infinity and the later for near distance vision. The efficiency of the diffractive lens is set at 50%, thus both the near and the far foci are accommodated over the whole visual field. The diffractive lens is fabricated on the rear of the conventional lens. When the eyes are focussed for a distant object, a blurred image is superimposed due to the presence of diffractive lens and vice versa, which obviously reduces the image quality. In most cases, the blurred image is discarded by the human visual perception and retinal processing system.

4. Holography in Orthopedics: Holography offers an excellent tool for the contactless study of orthopedic structures, specifically external fixtures to measure strains on fixation pins and rods. Such studies are important in osteosynthesis with external fixture used for long bone fractures, to prevent dislocations of both fractured ends that are mainly caused by decrease in strength of the fixation pins.

QUESTIONS

1. What is meant by holography? Why is it called wave front reconstruction?
2. Describe how a hologram is generated and image is reconstructed using off-axis configuration?
3. What is a hologram? How does it differ from an ordinary photograph? Describe in short how a hologram is generated and viewed? **(C.S.V.T.U.,2005)**
4. Describe the recording and reconstruction processes in Holography with the help of suitable diagrams. **(V.T.U.,2008)**
5. Write the principle of holography. With neat sketches explain in brief recording of a hologram and reconstruction of images. **(V.T.U.,2008)**
6. What is holography? How does it differ from ordinary photographic technique? How do interference and diffraction phenomena related to construction reconstruction of the hologram? Explain in detail with suitable diagram. **(Bombay Univ.)**
7. What is the advantage of off-axis configuration over the coaxial configuration?
8. Explain the principle of holography.
9. Explain some of the important properties of hologram.
10. How are the holograms classified? Explain.
11. Write the differences between holography and photography. **(C.S.V.T.U.,2009)**
12. Discuss some of the important applications of holography.
13. Explain holographic interferometry.
14. Explain acoustic holography.
15. Describe some of the important medical applications of holography.
16. What are DOEs (diffractive optical elements)? Describe their salient features and applications.

CHAPTER

26

Crystal Structures

26.1 INTRODUCTION

A solid consists of atoms or clusters of atoms arranged in close proximity. The physical structure of a solid and its properties are closely related to the scheme of arrangement of atoms within the solid. In amorphous solids the arrangement of atoms is random while in crystals there is a regular arrangement of atoms. Simple geometrical concepts of a lattice and unit cell are used to describe the atomic arrangement in crystals. The advantage of such a description is that all possible crystal structures are represented by a limited number of basic unit cell geometries. Majority of metals are found to belong to three simple structures. The crystal structures are analyzed using x-ray diffraction technique invented by Max von Laue and extensively employed by Bragg and Bragg. The study of crystal geometry helps us understand the diverse behaviour of solids in their mechanical, metallurgical, electrical, magnetic and optical properties.

26.2 CLASSIFICATION OF SOLIDS

Solids are classified into the following **three categories** basing on the atomic arrangement within the solid.

- (i) **single crystals**
- (ii) **polycrystalline solids** and
- (iii) **amorphous solids.**

(i) Single crystals: Single crystals are polyhedrons that have a distinctive shape for each material and are bounded by smooth shiny faces and straight edges. When a crystal is broken, it cleaves along certain preferred directions. The same substance may crystallize under different conditions of crystal growth to form different geometrical shapes but the angles between the faces are always constant for different shapes. This is known as *law of constancy of angles*. X-ray diffraction studies have shown that the atoms in crystalline solids are arranged in a regular periodic pattern in three dimensions, as shown in Fig. 26.1(a). The arrangement of atoms in specific relation to each other is called **order**. In crystals the order exists in the immediate neighbourhood of a given atom as well as over large distances corresponding to several layers of atoms. Therefore, crystals possess both **short-range order** and **long-range order**.

In single crystals, since the periodic arrangement of atoms differs in the three directions, the physical properties vary with direction and therefore, they are called **anisotropic** substances.

Quartz, alum, diamond and rock salt are examples of solids that occur as large size single crystals.

(ii) Polycrystalline solids: Polycrystalline solids consist of fine grains, having a size of 10^3 to 10^4 Å, separated by well-defined boundaries but oriented in different directions (Fig. 26.1b). Each such grain is a single crystal of an irregular shape. Since the grains are oriented randomly, a polycrystalline material is **isotropic** and the physical properties do not vary with direction. Majority of the natural solids have polycrystalline structure. Metals are examples of polycrystalline solids.

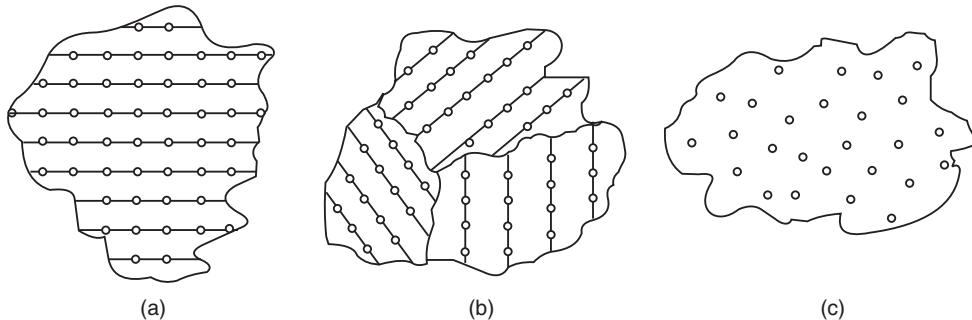


Fig. 26.1. Atomic arrangement in (a) Single crystal (b) Polycrystalline solid and (c) Amorphous substance

Types of crystals

The atoms are bound to each other by electrostatic forces. Depending on the nature of the bond between the neighbouring atoms, the crystals can be classified as metallic crystals, covalent crystals, and ionic crystals. In ionic crystals, the electron is transferred from the cation to anion and the atoms exist in the crystal as ions and hence the name **ionic crystal**. The interatomic forces in such crystals are non-directional. These forces are directional in **covalent crystals** where electron is shared by the neighbouring atoms. In **metallic crystals**, the electrons are very loosely bound to the corresponding nuclei.

(iii) Amorphous solids: Solid materials possessing only short-range order (Fig. 26.1c) are known as *amorphous* solids. These materials maintain a fixed volume and shape and resemble solids in their external features, but internally they do not have the ordered crystalline arrangement. The atomic arrangement is random in these materials. They are in fact considered as **supercooled liquids** having a very high viscosity. Glass, rubber and many polymers are amorphous solids. The amorphous materials are generally referred to as ‘glass’. Note that the word ‘glass’ here does not refer to the transparent glass used for windowpanes. The amorphous materials are called **glass** because they have features similar to glass. The physical properties of amorphous solids are not dependent on the direction of measurement and therefore they are **isotropic** substances.

Though the word ‘solid’ has been traditionally used for crystalline, polycrystalline and amorphous materials, nowadays the word ‘solid’ is meant to convey only crystalline materials.

26.2.1 Phase

When the crystals are grown there are two important requirements to be satisfied. (i) The atoms must occupy minimum space, and (ii) the bond angle requirements must be satisfied. For metallic crystals, the second condition does not exist and therefore, the atomic arrangement in metallic crystals is usually much simpler than the atomic arrangement in nonmetallic crystals.

Since the atomic arrangement everywhere in the crystal is the same, the crystal growth occurs at a fixed temperature. The atoms in the crystal arrange and disperse like soldiers in the army, each soldier going through identical movement. A given atomic arrangement in the material is called a **phase**. Thus, a given chemical like water may exist in vapour, water or ice phase. If the atomic arrangement in the crystal changes, we say that a new crystal has formed and a **phase change** has occurred.

26.2.2 Periodicity

The periodicity of atomic arrangement can be described as

$$\mathbf{r} = a \mathbf{x} + b \mathbf{y} + c \mathbf{z}$$

where $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are unit vectors in x, y, z directions respectively. Note that *periodicity is the only condition for crystallinity*.

26.3 SPACE LATTICE

A regular and periodic arrangement of atoms is the most important feature of crystals. The actual arrangement of atoms is called the **structure**. Suppose the atoms or clusters of atoms in a crystal are represented by points that correspond to their mean positions. Then we obtain a regular distribution of points in space. These points are called **lattice points** or **lattice sites**. Although lattice points represent atom locations, the lattice points and atom centres need not be coincident. The three-dimensional network of regularly arranged points is known as a **space lattice**. A point is a dimensionless and shapeless entity; therefore, a lattice is merely an imaginary geometrical framework.

A space lattice is defined as an array of points in three dimensions in which every point has surroundings identical to that of every other point. A space

lattice can be generated by successive translations of an initial point. A very simple operation of repetition consists of repeating the unit without change after translating it a distance ' t_1 '. Repeated application of a translation of given length and direction, t_1 , to an initial point generates a sequence of periodically spaced points or a row of points.

The repeated application of some other translation which is not along the same line (i.e. non-collinear), t_2 , to the above row generates a planar array of points.

A third translation not in the same plane (i.e. non-coplanar), t_3 , applied to a two-dimensional lattice generates a three-dimensional array of points called a **space lattice** or simply **lattice**.

Thus, lattice points satisfy the condition of periodicity. Hence, the crystal structure, i.e., the arrangement of atoms in the crystal can be understood with reference to a lattice. It is also observed that a given point in the lattice (Fig. 26.2c) is surrounded by a specific number of lattice points located at equal distances. The arrangement of points around a given lattice point is called its **environment**. The environment continues to change as we traverse in a particular direction from one lattice point to the next lattice point, but at the next lattice point it is

Fig. 26.2 a. A sequence of periodically spaced points

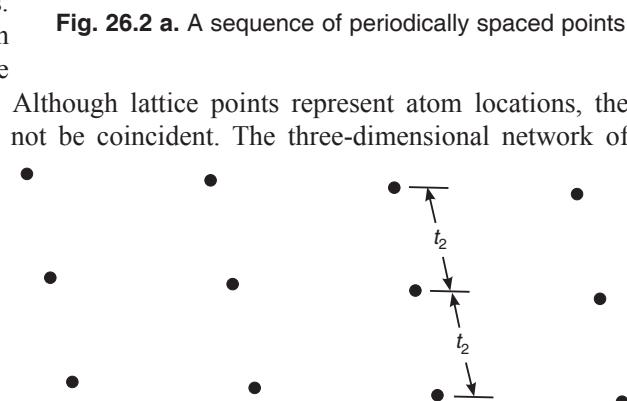
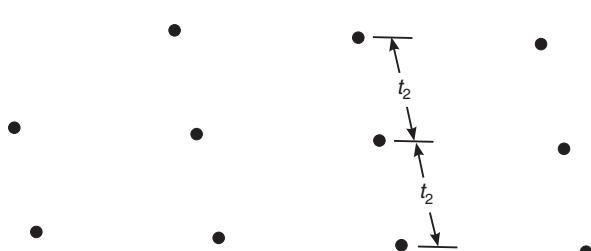


Fig. 26.2 b. A planar array of points



exactly the same as that at the initial point. Thus, the same environment would be found after equal intervals of distance in the same direction. The environment would be different in different directions. The space lattice is the skeleton upon which crystal structure is built by placing atoms on or near the lattice points.

The study of crystal structure becomes simpler when it is represented by a space lattice.

26.4 CRYSTAL STRUCTURE

A space lattice is a mathematical abstraction. The crystal structure is formed only when a group of atoms is identically attached to each lattice point, as shown in Fig. 26.3. The group of atoms that is associated with every lattice point is called a **basis**. The basis must be identical in composition, arrangement and orientation such that the crystal appears exactly the same at one point as it does at other equivalent point.

A crystal structure is thus the result of two quantities; namely a lattice and a basis. Thus,

$$\text{Lattice} + \text{basis} \rightarrow \text{crystal structure}$$

In the simplest crystals such as copper, silver, gold, iron and the alkali metals, the basis is a single atom. Often the basis contains a few atoms.

26.5 UNIT CELL

Because of the inherent periodicity, the space lattice can be represented by a unit cell. *The unit cell is the smallest geometrical unit, which when repeated in space indefinitely, generates the space*

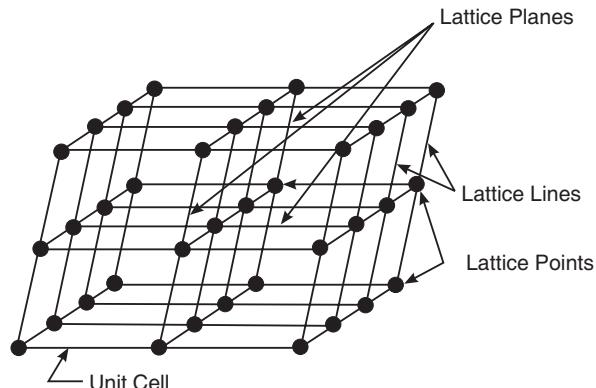


Fig. 26.2 c. Three dimensional lattice or space lattice

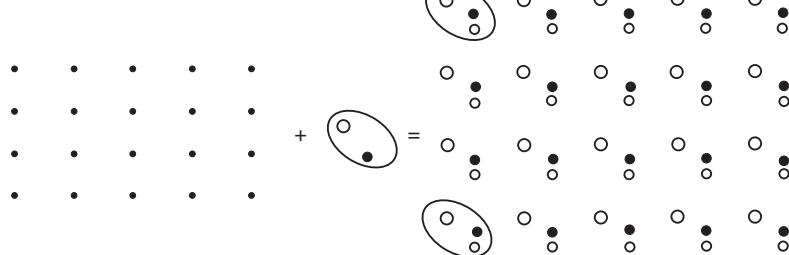


Fig. 26.3

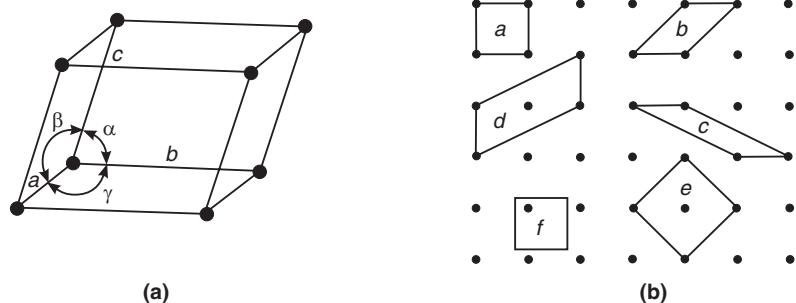


Fig. 26.4. (a) A primitive unit cell. (b) Primitive and non-primitive unit cells in a two-dimensional lattice

lattice. Hence, it is the smallest volume that carries a full description of the entire lattice. The lines drawn parallel to the lines of intersection of any three faces of the unit cell which do not lie in the same plane are called **crystallographic axes**. The three translational vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} lie along the crystallographic axes. The intercepts a , b , and c (Fig. 26.4 a) define the dimensions of the unit cell and are known as **primitives**. The angle γ represents the angle between a and b axes, the angle α represents the angle between b and c axes and the angle β represents the angle between the c and a axes. The axial lengths a , b , c and the three inter-axial angles α , β , γ are known as the basic **lattice parameters**. The volume of the unit cell is $(a \times b \times c)$. If this volume in space contains only one lattice point, we call it a **primitive unit cell**. It is not necessary that every unit cell should be a primitive cell (Fig. 26.4 b). Depending on the requirement and symmetry of the lattice we can choose larger cell containing more than one lattice point as a unit cell. If there are two or more lattice points per unit cell, then it is called a **non-primitive unit cell**. Most of the unit cells of various crystal lattices contain two or more lattice points and are non-primitive cells.

Example 26.1. The unit cell of copper is a cube. The side of the cube is 3.6 \AA . If the unit cell in copper solid are lined up side by side, how many unit cells will be there along 10 mm length of the solid?

Solution. The number of unit cells in a length l of solid is given by

$$N = \frac{\text{Length of the solid}}{\text{side of the unit cell}} = \frac{l}{a} = \frac{10^{-2} \text{ m}}{3.6 \times 10^{-10} \text{ m}} = 28 \times 10^6$$

26.6 BRAVAIS LATTICES

Bravais introduced the concept of space lattice in the study of crystal structures. A three-dimensional space lattice is generated by repeated translations of three non-coplanar vectors. One would expect that many lattices can be generated in three dimensions with different primitive and non-primitive cells. However, Bravais showed that there are only fourteen different ways of arranging identical points in three-dimensional space, satisfying the condition of periodicity, so that they are in every way equivalent in their surroundings. These fourteen types of arrangements are called the **space lattices** or **Bravais lattices**. There are seven primitive cells and seven non-primitive cells. With the 14 types of lattices and on the basis of primitive cell, crystals are grouped into 7 systems. They are cubic, tetragonal, orthorhombic, monoclinic, triclinic, hexagonal and rhombohedral (trigonal). The seven types of crystals and the corresponding Bravais lattices with their features are described below.

1. Cubic system

In cubic crystals, the crystal axes are perpendicular to one another. Thus, $\alpha = \beta = \gamma = 90^\circ$. The length of the primitives (edges of unit cell) is the same along the three axes. Thus, $a = b = c$. Cubic lattice has three possible types of arrangements, simple (SC) or cubic-P, body centered (BCC) or cubic-I and face centered (FCC) or cubic-F.

- (i) **Simple cubic lattice:** It has lattice points at all 8 corners of the unit cell.

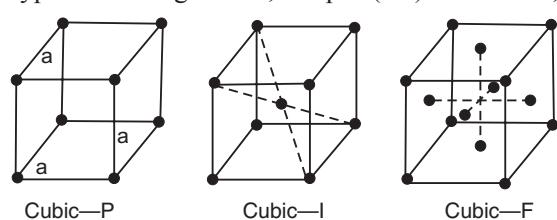


Fig. 26.5 (A): The three Bravais lattices of a cubic system

- (ii) **Body-centered cubic lattice:** It has lattice points at all 8 corners of the unit cell and one lattice point at the centre of the body.
- (iii) **Face-centered cubic lattice:** It has lattice points at all 8 corners of the unit cell and one lattice point each at the centre of six faces of the cube.

2. Tetragonal system

In tetragonal crystals, the crystal axes are perpendicular to one another. Thus, $\alpha = \beta = \gamma = 90^\circ$. The lengths of the edges of unit cell along two axes are the same but the edge along the third axis is different. That is, $a = b \neq c$. Tetragonal lattice has two possible types of arrangements, simple (P) and body centered (I).

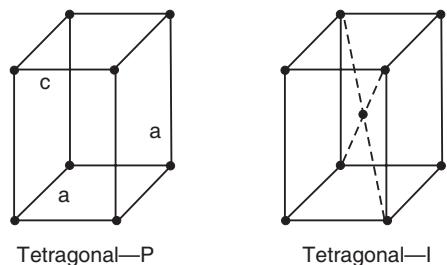


Fig. 26.5 (B): The two Bravais lattices of a tetragonal system

- (i) **Simple tetragonal lattice:** It has lattice points at all 8 corners of the unit cell.
- (ii) **Body-centered tetragonal lattice:** It has lattice points at all 8 corners of the unit cell and one lattice point at the centre of the body.

3. Orthorhombic system

In orthorhombic crystals, the crystal axes are perpendicular to one another. Thus, $\alpha = \beta = \gamma = 90^\circ$. But the lengths of the edges of unit cell along the three axes are different. That is, $a \neq b \neq c$. Orthorhombic lattice has four possible types of arrangements, simple (P), base centered (C), body centered (I) and face centered (F).

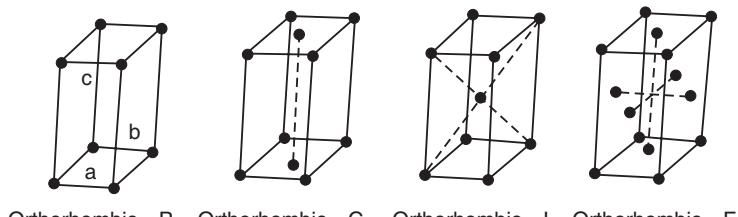


Fig. 26.5 (C): The four Bravais lattices of an orthorhombic system

- (i) **Simple orthorhombic lattice:** It has lattice points at all 8 corners of the unit cell.
- (ii) **Base-centered orthorhombic lattice:** It has lattice points at all 8 corners of the unit cell and two lattice points, one each at the base and top face of the body.
- (iii) **Body-centered orthorhombic lattice:** It has lattice points at all 8 corners of the unit cell and one lattice point at the centre of the body.
- (iv) **Face-centered orthorhombic lattice:** It has lattice points at all 8 corners of the unit cell and six lattice points, one each at the centre of six faces of the unit cell.

4. Monoclinic system

In monoclinic crystals, out of the three crystal axes, two are not perpendicular to each other but the third is perpendicular to both of them. Thus, $\alpha = \gamma = 90^\circ \neq \beta$. The lengths of the edges of unit cell along the three axes are different. That is, $a \neq b \neq c$. Monoclinic lattice has two possible types of arrangements, simple (P) and base centered (C).

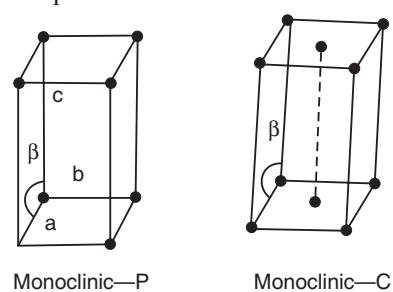


Fig. 26.5 (D): The two Bravais lattices of a monoclinic system

- (i) **Simple monoclinic lattice:** It has lattice points at all 8 corners of the unit cell.
- (ii) **Base-centered monoclinic lattice:** It has lattice points at all 8 corners of the unit cell and two lattice points, one each at the base and top face of the unit cell.

5. Triclinic system

In triclinic crystals, none of the three crystal axes is perpendicular to any of the other. Thus, $\alpha \neq \beta \neq \gamma$. The lengths of the edges of unit cell along the three axes are different. That is, $a \neq b \neq c$. Triclinic lattice has only one possible type of arrangement, namely, simple (P).

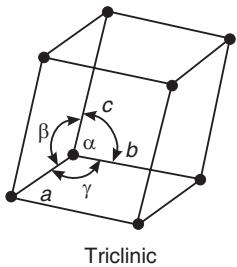


Fig. 26.5(E): The Bravais lattice of a triclinic system

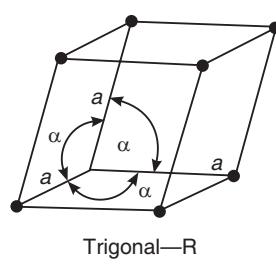


Fig. 26.5 (F): The Bravais lattice of a trigonal system

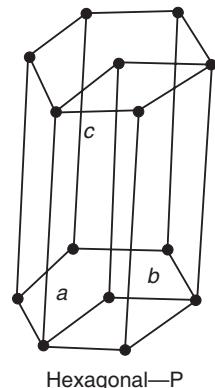


Fig. 26.5 (G): The Bravais lattice of a hexagonal system

- (i) **Simple triclinic lattice:** It has lattice points at all 8 corners of the unit cell.

6. Trigonal or Rhombohedral system

In trigonal crystals, the angle between each pair of crystal axes is the same but is not equal to 90° . Thus, $\alpha = \beta = \gamma \neq 90^\circ < 120^\circ$. The length of the edges of unit cell is the same along all the three axes. That is, $a = b = c$. Trigonal lattice has only one possible type of arrangement, namely, simple (R).

- (i) **Simple trigonal lattice:** It has lattice points at all 8 corners of the unit cell.

7. Hexagonal system

In hexagonal crystals, two of the crystal axes are 120° apart, while the third axis is perpendicular to both of them. Thus, $\alpha = \beta = 90^\circ$ and $\gamma = 120^\circ$. The length of the edges of unit cell is the same along all the axes that are 120° apart but the edge along the third axis is different. That is, $a = b \neq c$. Hexagonal lattice has only one possible type of arrangement, namely, simple (P).

- (ii) **Simple hexagonal lattice:** It has lattice points at all 12 corners of the hexagonal unit cell and two lattice points, one each at the base and top face of the hexagonal prism.

The seven crystal systems and the parameters of the corresponding unit cells are summarized in Table-1.

Table 1: The seven crystal systems

Sr. No.	Crystal System	Type and number of Bravais lattices	Unit cell axes	Angle between axes
1.	Triclinic	P,1	$a \neq b \neq c$	$\alpha \neq \beta \neq \gamma \neq 90^\circ$
2.	Monoclinic	P } 2 C }	$a \neq b \neq c$	$\alpha = \gamma = 90^\circ \neq \beta$
3.	Orthorhombic	P } C } 4 I } F }	$a \neq b \neq c$	$\alpha = \beta = \gamma = 90^\circ$
4.	Tetragonal	P } 2 I }	$a = b \neq c$	$\alpha = \beta = \gamma = 90^\circ$
5.	Cubic	P } I } 3 F }	$a = b = c$	$\alpha = \beta = \gamma = 90^\circ$
6.	Hexagonal	P, I	$a = b \neq c$	$\alpha = \beta = 90^\circ, \gamma = 120^\circ$
7.	Trigonal (Rhombohedral)	R, I	$a = b = c$	$\alpha = \beta = \gamma \neq 90^\circ < 120^\circ$

26.7 SYMMETRIES IN CRYSTALS

Some similarities of atomic arrangements may occur in crystals as a result of periodic arrangement. This is called **symmetry** of atomic arrangement. Indeed, the crystal can be characterized by the symmetry it possesses. Crystals possess different external symmetries, which are described by certain mental operations. A **symmetry operation** is one that takes the crystal into a configuration identical to the initial configuration. The crystal is said to possess a **symmetry element** corresponding to an operation, if after performing the particular operation the crystal goes into a position indistinguishable from the initial position. The most important elements of symmetry are a center, an axis and a plane.

1. A crystal is said to have a **center of symmetry**, if a point exists within the crystal such that any line drawn through it will have similar situation at both of its ends. This means that if there is an atom at some distance at one end, there is also an atom at the other end at the same distance. It is also called the **center of inversion**.
2. A crystal is said to possess an **axis of symmetry** if when the crystal is rotated about the axis, the atomic arrangement looks exactly the same more than once during one complete revolution. If the atomic arrangement looks the same twice in one revolution, the axis is said to be a *diad* axis. Similarly, the axis may be a triad axis, tetrad axis, and hexad axis. Note that only these types of axes are possible. The other types of axes are not consistent with the condition of periodicity. The diad axis is represented by 2, triad axis by 3, tetrad axis by 4, and the hexad axis by 6, corresponding to $180^\circ, 120^\circ, 90^\circ, 60^\circ$, rotations respectively.
3. A crystal is said to have a **plane of symmetry** when it can be divided by an imaginary plane into two parts, such that one is the exact mirror image of the other.

The mirror plane is represented by the letter m . Usually, the mirror plane is parallel to rotation axis or perpendicular to it.

In addition to the above point operations, there exist the following four hybrid operations.

Hybrid operations

- (i) **Rotoreflection:** This is the combination of an n -fold rotation followed by a reflection in a plane perpendicular to the rotation axis.
- (ii) **Rotoinversion:** This is the combination of an n -fold rotation followed by an inversion.
- (iii) **Screw translation:** In this the n -fold rotation axis is coupled with the translation parallel to rotation axis.
- (iv) **Glide reflection:** In this a mirror plane is coupled with a translation parallel to the reflecting plane.

Each type of symmetry is called a **symmetry element**. The crystal may have more than one symmetry element. All the symmetry elements possessed by a crystal, grouped together, are called a **symmetry group** or a **point group**. Because of the restriction of the condition of periodicity, the number of possible point groups is not large. It was shown that all possible combinations of symmetry elements lead to 230 space groups divided into 32 symmetry classes. These are grouped into the seven crystal systems.

26.7.1 Symmetry Elements in a Cubic Crystal

Basing on the internal structure, only certain elements of external symmetry can occur in a crystal. We now take the example of a cubic crystal and illustrate the elements of symmetry it exhibits.

- (i) Let us consider the cube shown in Fig. 26.6. If the body center point is I, and body diagonals are drawn through it, each diagonal connects identical lattice points located at equal distances and in opposite directions from I. The point

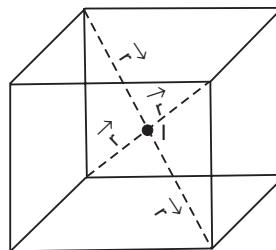


Fig. 26.6

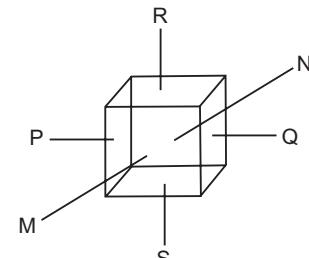


Fig. 26.7

I acts as a point mirror which generates the second lattice point at an equal distance in opposite direction. Therefore, the point I is the center of symmetry or inversion point. Thus, for a cubic crystal inversion point I is located at the body center.

- (ii) Let us consider a normal MN through mid-points of the pair of opposite parallel faces of a cube. If the cube is rotated through 90° rotation, it goes into an indistinguishable configuration. In one complete rotation of 360° , the cube becomes indistinguishable four times. Therefore, the face-normal is a 4-fold axis of symmetry. There are three such 4-fold axes of symmetry, MN, PQ and RS, one normal to each of the three pairs of parallel faces (see Fig. 26.7).

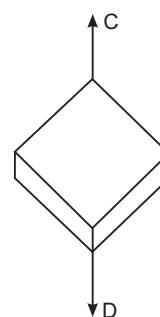


Fig. 26.8

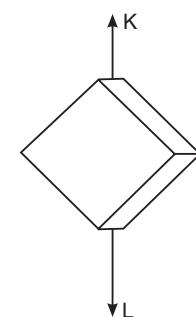


Fig. 26.9

Now consider the axis CD passing through the body diagonal of the cube (Fig. 26.8). If the cube is rotated around the body diagonal through 120° , it comes into a congruent position. In one full rotation, the cube becomes indistinguishable three times. Therefore, the body diagonal is a 3-fold axis of symmetry. A cube has four body diagonals and therefore possesses four 3-fold axes of symmetry.

Next, consider a normal KL at mid points of parallel edges of Fig. 26.9. If the cube is rotated through 180° , it becomes congruent. In one rotation of 360° , the cube assumes indistinguishable configuration twice. Therefore, KL is a 2-fold axis of symmetry. As there are 12 edges in a cube, the number of 2-fold axes of symmetry are six.

The axes of symmetry in a cube are thus

2 - fold axes (diad axes)	6
3 - fold axes (triad axes)	4
4 - fold axes (tetrad axes)	3

	13

There are in total 13 axes of symmetry in a cube.

(iii) Let us consider a plane such as PQRS (Fig. 26.10) in the middle of the cube and parallel to one pair of the faces. If it is a plane mirror, and one half of the crystal is cut and removed, the plane PQRS forms the image of that half of the crystal in it. That means, if we reflect one half of the crystal in PQRS, the image will coincide with the other half. Therefore, PQRS is called a plane of symmetry or a *mirror plane*. There are three such planes of symmetry parallel to the faces of the cube.

Further, consider the diagonal plane KLMN (Fig. 26.11) in the cube. If KLMN is imagined to be a mirror, it is readily seen that the prism behind it is the reflection of the prism in contact with it in front. Thus it is also a mirror plane. There are six such diagonal mirror planes in a cube.

The total elements of symmetry of a cubic crystal are 23 comprising of one center of inversion, 13 axes of rotation and 9 mirror planes.

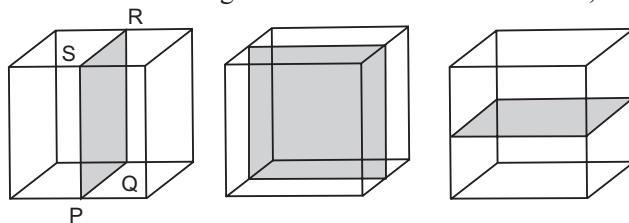


Fig. 26.10

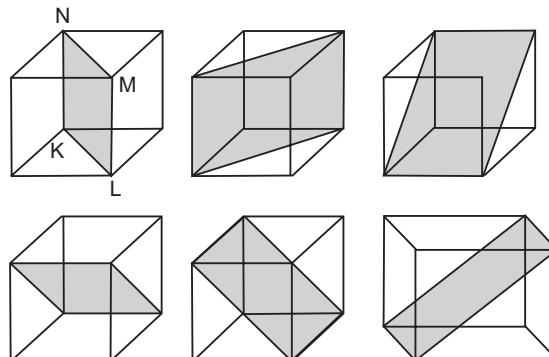


Fig. 26.11

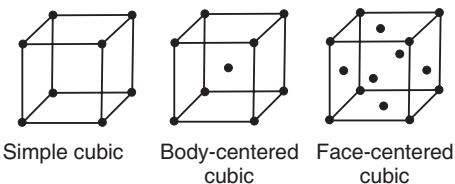


Fig. 26.12. The three space lattices of a cubic system

26.8 CALCULATION OF PARAMETERS OF A CUBIC LATTICE

A unit cell is characterized by a set of characteristics, namely, volume, effective number of atoms, coordination number, atomic packing fraction, and density. These parameters can be computed with relative ease in case of cubic unit cells. There are three varieties of the cubic

lattice as shown in Fig. 26.12. They are simple cubic (SC), body centered cubic (BCC), and face centered cubic (FCC) unit cells.

26.8.1 Simple cubic (SC) cell

A unit cell is said to be *primitive* when the cell has lattice points only at its corners. The primitive cubic unit cell is also called a simple cubic (SC) unit cell. Fig. 26.13 shows a SC cell.

(i) **Unit cell volume, V :** In case of cubic cell all the edges of the cube are of equal length, ' a '. Therefore, the volume is given by

$$V = a^3 \quad (26.1)$$

(ii) **Effective number of atoms per unit cell, Z :** A unit cell is a part of an infinite scheme, and is not an isolated entity. Therefore, several adjacent cells share each lattice point. As a result, the basis attached to a lattice site contributes only a fraction of its mass and volume to one unit cell. Let only one atom, having a radius R , be attached to one lattice point. The effective number of atoms per unit cell is given by

$$Z = Z_B + \frac{Z_F}{2} + \frac{Z_C}{8} \quad (26.2)$$

where Z_B = Number of body centered atoms,

Z_F = Number of face centered atoms and

Z_C = Number of corner atoms

In the three-dimensional array, each corner atom is linked to eight surrounding cells, as shown in Fig. 26.14 (a). Hence, in effect, the atom contributes $1/8^{\text{th}}$ of its content to a unit cell, as shown in Fig. 26.14 (b). SC cell being a primitive cell does not contain lattice points within the body volume or in the centre of faces. Therefore, the total contribution to Z comes from the corner atoms of the unit cell and it is given by

$$Z = Z_B + \frac{Z_F}{2} + \frac{Z_C}{8} = 0 + 0 + \frac{8}{8} = 0 + 0 + 1 \text{ atom / cell}$$

$$\therefore Z = 1 \text{ atom / cell} \quad (26.3)$$

(iii) **Coordination Number, CN:** The coordination number of an atom in a crystal is the number of nearest neighbour atoms. It signifies the tightness of packing of atoms in the crystal.

Around an atom in a SC cell, there would be six equally spaced nearest neighbour atoms each at a distance ' a ' from that atom, as shown in Fig. 26.15. Four atoms lie in the plane

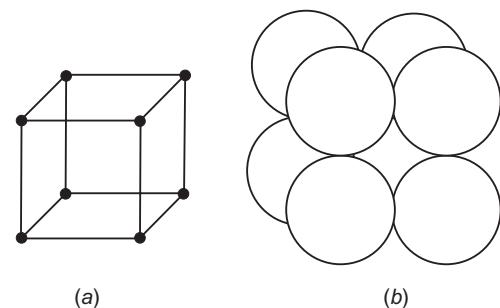


Fig. 26.13. Simple cubic cell (a) atomic site model (b) hard sphere model

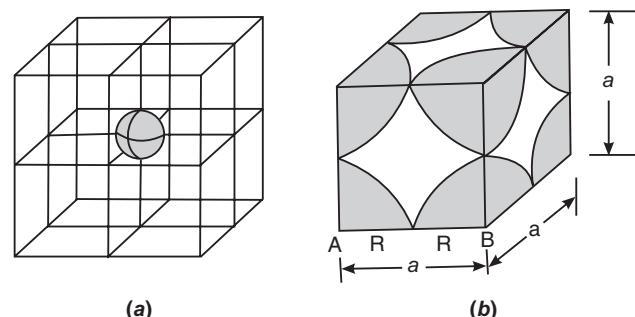


Fig. 26.14. (a) An atom contributes one eighth to a unit cell when it is located at the corner of the cell. (b) Isolated SC cell

of the atom while one is vertically above it and one vertically below. Therefore, the coordination number CN = 6.

(iv) Atomic radius, R: The relationship between the apparent size of the atom and the edge of the unit cell can be determined where one atom is attached to a lattice site. The specific direction along which atoms are in contact is identified and by applying simple geometry, the relation between the atomic size and the unit cell edge can be computed.

In a SC cell the atoms would be in contact along the edges of the cube, as seen from Fig. 26.14 (b). If 'a' is the edge of the cubic cell and R is the radius of the atom.

$$a = 2R \quad \text{or} \quad R = a / 2 \quad (26.4)$$

(v) Packing Fraction, APF: The fraction of space occupied by atoms in a unit cell is known as atomic packing fraction. It is defined as *the ratio of volume of effective number of atoms in the unit cell to the total volume of the unit cell*. Thus,

$$APF = \frac{\text{(Number of atoms/unit cell)} (\text{Volume of each atom})}{\text{Volume of the unit cell}} \quad (26.5)$$

$$\therefore APF = \frac{Zv}{V}$$

In case of SC cell, $Z = 1$

Volume of unit cell, $V = a^3 = (2R)^3 = 8R^3$.

$$\begin{aligned} \text{Volume of spherical atom, } v &= \frac{4}{3}\pi R^3 \\ \therefore APF &= \frac{1 \times \frac{4}{3}\pi R^3}{8R^3} = \frac{\pi}{6} = 0.52 \end{aligned} \quad (26.6)$$

(vi) Percentage void space: *The void space in the unit cell is the vacant space left unutilised in the cell.* It is often expressed as percentage.

$$\begin{aligned} \% \text{ void space} &= (1 - APF) \times 100 \\ &= (1 - 0.52) \times 100 \\ &= 48\% \end{aligned} \quad (26.7)$$

(vii) Density of SC crystal, ρ : As a unit cell possesses all the structural properties of a bulk crystal, the density of a unit cell must be the same as that of the bulk crystal. Thus,

$$\text{Density, } \rho = \frac{\text{Mass}}{\text{Volume}} = \frac{ZW}{V} \quad (26.8)$$

where W is the mass of each atom, which is given by

$$W = \frac{M}{N_A} \quad (26.9)$$

M is the molecular weight of the material and N_A is the Avogadro number.

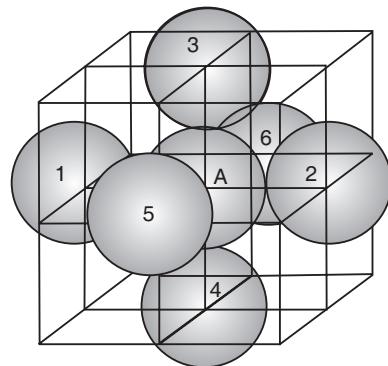


Fig. 26.15. Determination of nearest neighbours. There are six close neighbours for any selected atom in a SC cell.

Therefore, the density of a cubic crystal is given by

$$\rho = \frac{ZM}{N_A V} = \frac{ZM}{N_A a^3} \quad (26.10)$$

In case of SC crystal $Z = 1$.

Hence,

$$\rho = \frac{M}{N_A a^3} \quad (26.11)$$

26.9 BODY CENTRED CUBIC (BCC) CELL

The body centered cubic cell has a lattice point within the cell in addition to the eight corner lattice points, as shown in Fig. 26.16 (a).

(i) Unit cell volume, V : In case of cubic cell all the edges of the cube are of equal length ' a '. Therefore, the volume is given by

$$V = a^3$$

(ii) Effective number of atoms per unit cell, Z :

A BCC cell has eight lattice points at the eight corners of the cube and one lattice point at centre within the cell, as shown in Fig. 26.16 (a). There are no points in the face. An atom at the lattice point within the cell belongs completely to the cell and the atoms at the corners of the cell contribute $1/8^{\text{th}}$ each, as shown in Fig. 26.16 (b). The effective number of atoms per BCC cell is then

$$Z = Z_B + \frac{Z_F}{2} + \frac{Z_C}{8}$$

$$Z = 1 + 0 + \frac{8}{8}$$

$$\therefore Z = 2 \text{ atoms/cell} \quad (26.12)$$

(iii) Coordination Number, CN:

In the BCC cell, atoms occupying the corner lattice points do not touch each other. However, each corner atom is in contact with the atom at the body center (Fig. 26.17). As there are eight unit cells around each corner of the cell, the atom located at a corner would be simultaneously touched by the eight body centered atoms of the eight surrounding cells. Thus, the coordination number

$$CN = 8$$

(iv) Atomic radius, R:

In a BCC cell atoms would be in contact along the body diagonal. From the Fig. 26.18,

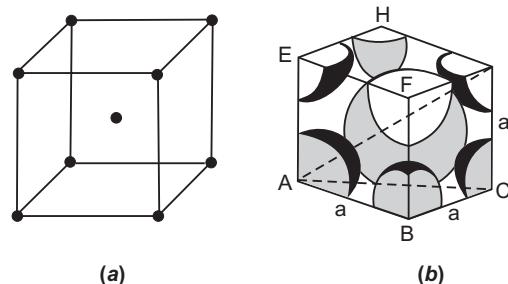


Fig. 26.16. (a) BCC cell (b) Isolated BCC unit cell

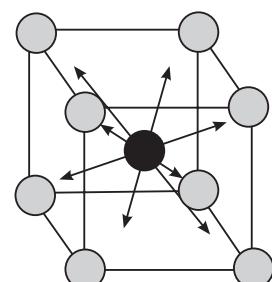


Fig. 26.17. A given atom in a BCC cell is surrounded by 8 closed neighbours

$$(AG)^2 = (AC)^2 + (CG)^2 = (AB)^2 + (BC)^2 + (CG)^2 \quad (26.13)$$

$$AG = 4R \quad \text{and} \quad AB = BC = CG = a$$

$$\therefore (4R)^2 = 3(a)^2$$

$$\therefore 4R = a\sqrt{3}$$

$$\therefore R = \frac{a\sqrt{3}}{4} \quad (26.14)$$

(v) Atomic packing fraction, APF:

$$\text{For BCC cell, } Z=2 \quad \text{and} \quad a = \frac{4R}{\sqrt{3}}$$

$$\text{Volume of the spherical atom } v = \frac{4}{3}\pi R^3$$

$$\text{Volume of the unit cell } V = a^3 = \frac{64R^3}{3\sqrt{3}}$$

$$APF = \frac{Zv}{V} = \frac{\frac{2}{3} \times \frac{4}{3}\pi R^3}{\frac{64}{3\sqrt{3}}R^3} = \frac{\pi\sqrt{3}}{8} = 0.68 \quad (26.15)$$

(vi) Percentage void space:

$$\begin{aligned} \text{Percentage void space} &= (1 - APF) \times 100 \\ &= (1 - 0.68) \times 100 = 32\% \end{aligned} \quad (26.16)$$

(vii) Density of BCC crystal, ρ :

In case of BCC crystal $Z = 2$. Hence

$$\rho = \frac{2M}{N_A a^3} \quad (26.17)$$

Example 26.2. Molybdenum has a BCC structure. Its density is $10.2 \times 10^3 \text{ kg/m}^3$ and its atomic weight is 95.94. Determine the radius of molybdenum atom.

Solution: Density of a cubic unit cell is given by $\rho = \frac{ZM}{N_A a^3}$

$$\therefore a^3 = \frac{ZM}{\rho N_A} = \frac{2 \times 95.94 \text{ kg / k.mol}}{10.2 \times 10^3 \text{ kg / m}^3 \times 6.02 \times 10^{26} / \text{k.mol}} = 31.25 \times 10^{-30} \text{ m}^3$$

$$\therefore a = 3.15 \text{ \AA.}$$

$$\text{The atomic radius in BCC structure is given by } R = \frac{a\sqrt{3}}{4} = \frac{3.15 \text{ \AA} \times 1.732}{4} = 1.364 \text{ \AA.}$$

Example 26.3. Sodium crystallizes in a cubic lattice. The edge of the unit cell is 4.3 \AA . The density of sodium is 963 kg/m^3 and its atomic weight is 23. What type of unit cell does sodium form?

Solution:

$$\text{Density of a cubic unit cell is given by } \rho = \frac{ZM}{N_A a^3}$$

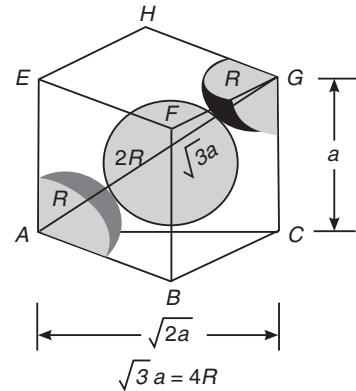


Fig. 26.18. Determination of the relation between atomic radius and lattice edge in a BCC cell

The effective number atoms per unit cell is given by $Z = \frac{\rho N_A a^3}{M}$

$$\therefore Z = \frac{963 \text{ kg/m}^3 \times 6.02 \times 10^{26} / \text{k.mol} \times (4.3 \times 10^{-10} \text{ m})^3}{23 \text{ kg/k.mol}} = \frac{46.09}{23} = 2 \text{ atoms/unit cell}$$

\therefore Sodium forms BCC structure.

26.10 FACE CENTRED CUBIC (FCC) CELL

The face centred cubic cell is a non-primitive cell having six lattice points at the centres of its six faces and eight atoms at the eight corners of the cube, as shown in Fig. 26.19.

(i) **Unit cell volume, V:** In case of cubic cell all the edges of the cube are of equal length ' a '. Therefore, the volume is given by

$$V = a^3$$

(ii) **Effective number of atoms per unit cell, Z:**

In a three dimensional array, a unit cell is surrounded by other unit cells, while each corner of the cell is shared by eight adjacent cells, and each face is shared by two adjacent cells, as shown in Fig. 26.20. Therefore, each face-centred atom contributes half of its mass and volume to one cell, while each corner atom contributes $1/8^{\text{th}}$, as shown in Fig. 26.21.

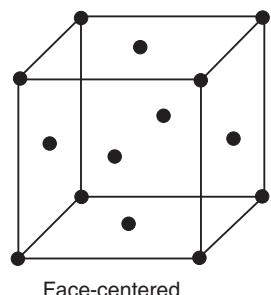


Fig. 26.19. Face Centered Cubic cell

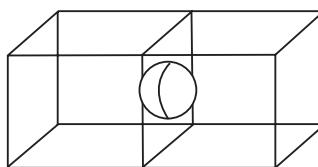


Fig. 26.20. An atom contributes half of its volume to the unit cell when it is located in the face

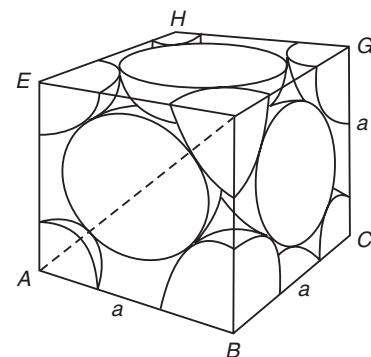


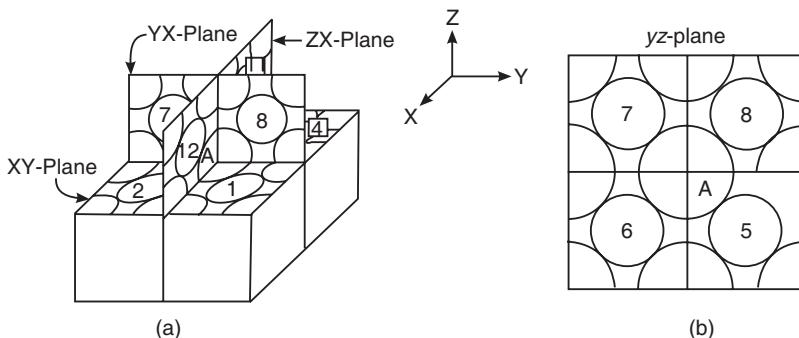
Fig. 26.21. Determination of effective number of atoms per unit cell.

There are 6 faces and 8 corners in a cubic cell. Thus, the number of atoms effectively contributing to the FCC cell is

$$\begin{aligned} Z &= Z_B + \frac{Z_F}{2} + \frac{Z_C}{8} \\ Z &= 0 + \frac{6}{2} + \frac{8}{8} \quad (26.18) \\ \therefore Z &= 4 \text{ atoms/cell.} \end{aligned}$$

(iii) **Coordination Number, CN:**

In a FCC cell, each corner atom is in contact with the face-centered atom. It would therefore be in contact with 4 atoms in the xy plane, 4 atoms in the yz -plane and 4 atoms in

**Fig. 26.22.** Determination of nearest neighbours in a FCC cell

the zx -plane, as shown in Fig. 26.22. Therefore,

$$\text{Coordination number} = 4 + 4 + 4 = 12.$$

(iv) Atomic radius, R :

In a FCC cell, atoms are in contact along the face diagonal of the cube, as shown in Fig. 26.23. It is seen from the figure that

$$AF^2 = AB^2 + BF^2 \quad (26.19)$$

$$\text{But } AF = 4R \text{ and } AB = a = BF$$

$$\therefore (4R)^2 = 2a^2$$

$$\therefore 4R = a\sqrt{2}$$

$$\therefore R = \frac{a\sqrt{2}}{4} = \frac{a}{2\sqrt{2}} \quad (26.20)$$

(v) Atomic packing fraction, APF:

$$APF = \frac{Z v}{V} = \frac{4v}{a^3}$$

$$a^3 = [2\sqrt{2}R]^3 = 16\sqrt{2}R^3$$

$$\therefore APF = \frac{4 \times \frac{4}{3}\pi R^3}{16\sqrt{2}R^3} = \frac{\pi}{3\sqrt{2}} = 0.74 \quad (26.21)$$

(vi) Percentage void space:

$$\text{Percentage void space} = (1 - APF) \times 100 = (1 - 0.74) \times 100 = 26 \% \quad (26.22)$$

(vii) Density of FCC crystal, ρ :

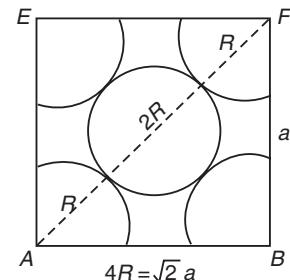
The density of a cubic crystal is given by

$$\rho = \frac{ZM}{N_A V} = \frac{ZM}{N_A a^3}$$

In case of FCC crystal $Z = 4$.

$$\text{Hence } \rho = \frac{4M}{N_A a^3} \quad (26.23)$$

Example 26.4. Lead exhibits FCC structure. Each side of the unit cell is of 4.95 \AA . Calculate the radius of a lead atom.

**Fig. 26.23.** Determination of atomic radius in a FCC cell

Solution: The radius of the lead atom is given by $R = \frac{a}{2\sqrt{2}} = \frac{4.95 \text{ \AA}}{2(1.414)} = 1.75 \text{ \AA}$.

26.11 HCP STRUCTURE

Hexagonal close packed (HCP) structure is one of the most common metallic structures. About 25 metals exhibit this structure. Metals do not crystallize into the simple hexagonal crystal structure. The atoms attain a lower energy and a more stable condition only by forming the HCP structure.

(i) Effective number of atoms per unit cell

The isolated HCP unit cell is shown in Fig. 26.24. Each corner atom of the hexagonal face is shared by six unit cells. Consequently it contributes $1/6^{\text{th}}$ of its volume and mass to one unit cell. There are six corner atoms in the base. Further, there is an atom at the center of each hexagonal face, which is shared by two adjacent unit cells. Three atoms forming a triangle in the middle layer are within the body of the cell and cannot be shared by adjacent cells. The total contribution to the effective number of atoms in the cell is thus

$$Z = \left(2 \frac{\text{base}}{\text{cell}} \right) \times \left(\frac{1}{6} \frac{\text{atom}}{\text{corner}} \right) \times \left(6 \frac{\text{corners}}{\text{base}} \right) + \left(2 \frac{\text{bases}}{\text{cell}} \right) \times \left(\frac{1}{2} \frac{\text{atoms}}{\text{base}} \right) + \left(3 \frac{\text{atoms}}{\text{cell}} \right)$$

or $Z = 6$ atoms/unit cell.

(ii) Atomic radius, R: The atoms are in contact along the edges of the hexagon as seen from Fig. 26.24. Therefore,

$$2R = a \quad \text{or} \quad R = a/2$$

(iii) Coordination number, CN: Each atom in the structure is positioned in valleys formed by three adjacent atoms of the top layer and by three adjacent atoms in the bottom layer; and is surrounded by six neighbour atoms. Thus, twelve atoms are in contact with the given atom under consideration. Therefore,

$$\text{CN} = 12$$

(iv) Volume of the unit cell, V:

The volume of the unit cell may be determined by computing the area of the base of the unit cell and then by multiplying it by the cell height. The area of the base of the unit cell is the area of the hexagon ABDEFG, shown in Fig. 26.25 (a). It is equal to the sum of the areas of the six equal-sized triangles (Fig. 26.25 b).

$$\text{Area of the hexagon ABDEFG} = 6 \text{ (area of } \Delta \text{ ABC}) = 6(1/2a)(a \sin 60^\circ)$$

$$= 3a^2 \sin 60^\circ = 3a^2 \left(\frac{\sqrt{3}}{2} \right) = \frac{3\sqrt{3}}{2} a^2$$

$$\therefore \text{Volume of the unit cell} = (\text{Area of the hexagon}) \times (\text{height of the unit cell})$$

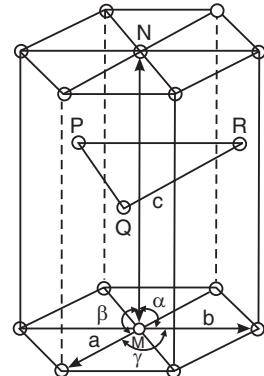


Fig. 26.24

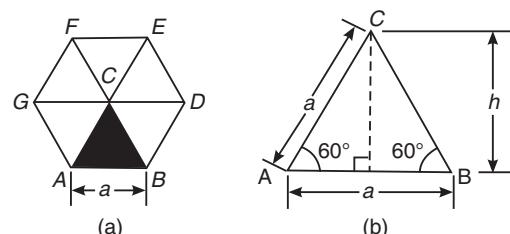


Fig. 26.25

$$= \left(\frac{3\sqrt{3}}{2} a^2 \right) c$$

or $V = \frac{3\sqrt{3}}{2} a^2 c$ (21.24)

Calculation of c/a ratio:

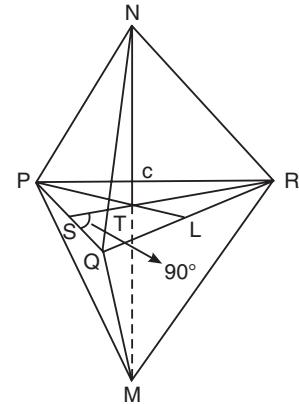
Referring to the HCP unit cell shown in Fig. 26.26, it is seen that $a = b$ and the angle $\theta = 120^\circ$. The c axis is normal to the plane containing a and b . Therefore $\alpha = \beta = 90^\circ$. Let P , Q and R be the centers of the adjacent atoms in the middle plane and N and M be the centers of the adjacent atoms in the plane immediately above and below the plane PQR , as shown in Fig. 26.26. Now let us join M to N and M and N to P , Q and R . It forms two tetrahedrons $MPQR$ and $NPQR$ with PQR as the common base. The line MN passes through the point T , which is the intersection of the three medians of the triangle PQR . RS is one of the medians and therefore,

$$TR = 2 TS$$

Further,

$$MN = c \quad \text{and} \quad PQ = QR = a$$

Fig. 26.26



As the median of an equilateral triangle is also perpendicular to the opposite side, $\angle QSR = 90^\circ$.

$$\therefore RS = \sqrt{a^2 - \left(\frac{a}{2}\right)^2} = \frac{\sqrt{3}}{2}a$$

$$\text{Also, } RT = 2TS = \frac{2}{3}(TS + TR) = \frac{2}{3}RS = \frac{2}{3} \cdot \frac{\sqrt{3}}{2}a = \frac{a}{\sqrt{3}}$$

$$\begin{aligned} \text{And } NT &= \sqrt{NR^2 - RT^2} = \sqrt{a^2 - \frac{a^2}{3}} = \sqrt{\frac{2}{3}}a && (\because NR = NQ = NP = a) \\ c &= 2NT = 2\sqrt{\frac{2}{3}}a = \sqrt{\frac{8}{3}}a \end{aligned}$$

$$\therefore \frac{c}{a} = \sqrt{\frac{8}{3}} \quad (26.25)$$

$$\therefore \text{The volume of the hexagonal cell, } V = \frac{3\sqrt{3}}{2}a^2 c = \frac{3\sqrt{3}}{2}a^2 \cdot \sqrt{\frac{8}{3}}a = 3\sqrt{2}a^3 \quad (26.26)$$

(v) Atomic Packing Fraction (APF):

$$\text{APF} = \frac{Zv}{V} = \frac{\frac{6}{3} \times \frac{4}{3} \pi r^3}{3\sqrt{2}(2r)^3} = \frac{\pi}{3\sqrt{2}} = 0.74$$

(vi) Void space

$$\text{Percentage void space} = (1 - \text{APF}) \times 100 = (1 - 0.74) \times 100 = 26\%$$

(vii) Density, ρ :

The theoretical density of HCP cell is given by

$$\rho = \frac{ZM}{N_A V} = \frac{6M}{N_A (3\sqrt{2}a^3)} = \frac{\sqrt{2}M}{N_A a^3} \quad (26.27)$$

The characteristics of the three types of cubic unit cells and the HCP cell are summarized in Table-2.

Table - 2

Sr.No.	Characteristics	Unit Cell			
		SC	BCC	FCC	HCP
1.	Unit cell volume, V	a^3	a^3	a^3	$3\sqrt{2}a^3$
2.	Atoms per unit cell, Z	1	2	4	6
3.	Atomic radius, r	$a/2$	$a\sqrt{3}/4$	$a/2\sqrt{2}$	$a/2$
4.	Coordination number, CN	6	8	12	12
5.	Atomic packing fraction, APF	$\pi/6$ = 0.52	$\pi\sqrt{3}/8$ = 0.68	$\pi/3\sqrt{2}$ = 0.74	$\pi/3\sqrt{2}$ = 0.74
6.	Void space	48%	32%	26%	26%
7.	Density, ρ	$\frac{M}{N_A a^3}$	$\frac{2M}{N_A a^3}$	$\frac{4M}{N_A a^3}$	$\frac{\sqrt{2}M}{N_A a^3}$

It is seen from the above table that the SC cell is loosely packed and the FCC and HCP cells are the close packed cells among the four types of unit cells.

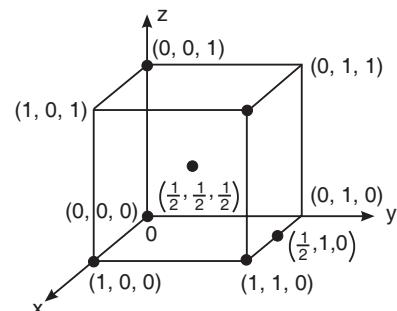
26.12 ATOM POSITIONS IN CUBIC UNIT CELLS

Let a rectangular right handed coordinate system be attached to the cubic lattice, as shown in Fig. 26.27. The position of a lattice site is described by three coordinates (x, y, z) which are expressed as

$$x = pa, \quad y = qb, \quad z = rc \quad (26.28)$$

where a , b and c are lattice constants in x , y and z directions respectively and p , q and r are integers. If the lattice constants a , b and c are taken as **unit** axial lengths, $a = 1$, $b = 1$, and $c = 1$, and the co-ordinates of the lattice site will be (p, q, r) . These are the indices of the lattice site and are written with commas separating the numbers and enclosed in parenthesis.

As an example, the indices of atomic sites in the BCC cell are shown in Fig. 26.27. The atom positions for the eight corner atoms are $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, $(1,1,1)$, $(1,1,0)$, $(1,0,1)$, and $(0,1,1)$. The body cnetre atom in the cell has the coordinates $\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$. By convention, the set of coordinates $(0,0,0)$ stands for the locations of all eight corners of cubic unit cell. Therefore, it is sufficient if the atom positions $(0,0,0)$ and $\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$ are specified for the BCC cell.

**Fig. 26.27**

The minimum number of coordinate sets necessary to specify the locations of all atoms in a unit cell are equal to the value of Z , the effective number atoms/cell.

26.13 INDICES OF CRYSTALLOGRAPHIC DIRECTION

Many crystals exhibit properties that are dependent on the direction in which they are measured and on the direction of the external stimuli, such as an electric field, magnetic field or mechanical stress, acting on the crystal. Therefore, it is essential that we know reference directions in crystals. To describe a direction in a crystal lattice, a straight line passing through the origin is chosen. The co-ordinates of the first lattice point lying on this line are utilized to denote the direction of the line. Thus, the indices of direction are the vector components of direction resolved along each of the axes and reduced to the smallest integers. The vector components are multiples of lattice constants. If the direction passes through the origin, then the indices of the lattice will also be the indices of direction.

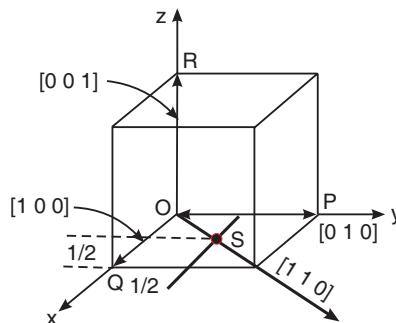


Fig. 26.28. Indices of direction

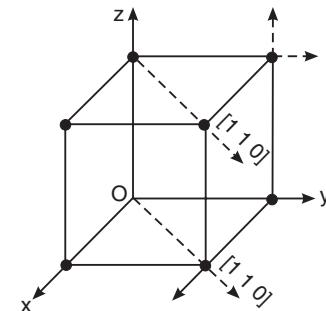


Fig. 26.29. Indices of parallel directions

The indices of direction are calculated by the following procedure.

- First, find the coordinates of the lattice site nearest to the origin in a given direction.
- The coordinates are then divided by appropriate unit translations.
- If fractions are obtained, each of the fractions is multiplied by smallest common divisor.
- The integers obtained are the indices of direction written in square brackets, as $[p \ q \ r]$.

Thus, *the indices of direction in a crystal are the set of smallest integers, which have the same ratios as the components of a vector in the desired direction, referred to the axes.*

As an example let us consider a cubic cell shown in Fig. 26.28. Let the point 'O' be the origin. The point P is at the position $(0, b, 0)$. 'b' represents one unit distance. Therefore, the indices of the point P are $(0, 1, 0)$. The direction OP is given as $[0 \ 1 \ 0]$. On the other hand, the direction of PO is specified as $[0 \bar{1} \ 0]$. A bar on the number indicates a negative direction.

Similarly, the directions OQ and OR are indicated as $[1 \ 0 \ 0]$ and $[0 \ 0 \ 1]$ respectively.

The direction of OS may be determined as follows. The position indices of the point S are $\left(\frac{1}{2}, \frac{1}{2}, 0\right)$. Then, using the instruction (iii), the direction of the line OS is obtained as

$$\left[2 \times \frac{1}{2}, 2 \times \frac{1}{2}, 0\right] = [110].$$

All parallel directions have the same direction indices (Fig. 26.29) and the parallel directions are equivalent. $\langle pqr \rangle$ represents the family of directions $[0\ 1\ 0]$, $[0\ 0\ 1]$, $[1\ 0\ 0]$, $[0\ \bar{1}\ 0]$, $[0\ 0\ \bar{1}]$, $[\bar{1}\ 0\ 0]$. All of them are grouped as $\langle 1\ 0\ 0 \rangle$.

26.14 LATTICE PLANES AND MILLER INDICES

A crystal lattice may be regarded as an aggregate of a set of parallel, equally spaced planes passing through the lattice points. The planes are called **lattice planes** and the perpendicular distance between adjacent planes is called **interplanar spacing**. A given space lattice may have infinite sets of lattice planes, each having its characteristic interplanar spacing. In a crystal the geometrical location of a plane is not important. All the planes that are parallel to each other play a similar role.

The position and orientation of a lattice plane in a crystal are determined by three smallest whole numbers which have the same ratios with one another as the reciprocals of the intercepts of the plane on the three crystal axes. These numbers are known as **Miller indices** of that plane.

We find the Miller indices of a given plane as follows. We choose a set of coordinate axes parallel to unit cell edges at a convenient point. Let OX, OY and OZ be positive directions of the axes along the three edges of the unit cell, as shown in Fig. 26.30. We consider a plane ABC oriented with respect to the coordinate axes and which intercepts the x -, y -, and z -coordinates at distances OA, OB and OC respectively. It will be seen that Miller indices for all the planes parallel to the plane ABC and containing similar type of atomic array are the same. The intercepts are not measured in metrical distances of cm or Å, but are measured in terms of respective unit lengths assigned to each edge of the cell, regardless the actual dimension of the edge. Thus, the intercepts OA, OB, and OC made by a plane such as ABC are expressed as multiples of the axial lengths.

For the plane of interest, we determine the intercepts x , y , and z on the coordinate axes.

- We express the intercepts in terms of the base vectors of the unit cell, pa , qb , and rc ; p , q and r are not necessarily integers *but they do have rational ratios*.
- We form the reciprocals $1/p$, $1/q$, and $1/r$.
- If the reciprocals are fractions, each fraction is multiplied by the least common denominator, D , and reduce them to the smallest set of integers. Thus, we obtain $D/p = h$, $D/q = k$ and $D/r = l$ respectively.
- We put the integers in parenthesis (hkl) . These are the **Miller indices** of the given plane.

All the planes, having the same structure of atoms, which are parallel to each other and therefore have the same orientation with respect to the three edges of the unit cell belong to the same family of planes and are represented by the same Miller indices. Consider a plane, of a family of similar planes, lying within the unit cell and nearest to the origin. This plane cuts off intercepts equal to $(1/h)a$, $(1/k)b$, $(1/l)c$. This implies that the (hkl) plane divides “ a ” into h parts, “ b ” into k parts and “ c ” into l parts. Other members of the family cut off intercepts $2a/h$, $2b/k$, $2c/l$, ..., $n a/h$, $n b/k$, $n c/l$ so that the ratio between the intercepts is the same.

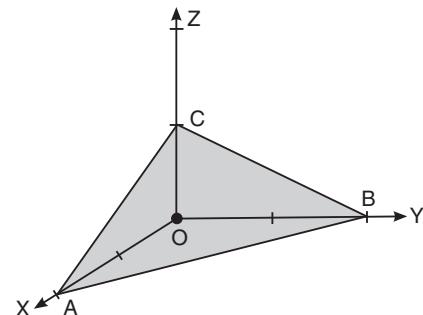


Fig. 26.30. A lattice plane intercepting the three crystallographic axes.

Some important features of Miller indices

1. Parallel planes spaced equally have the same Miller indices. Thus, all the planes in a set parallel to a particular plane (hkl) are denoted by (hkl).
2. A plane parallel to one coordinate axis has Miller index 0 for that direction.
3. A plane passing through origin is denoted by Miller indices of a parallel plane having non-zero intercepts.
4. When the intercept of a plane is negative on an axis, a bar is placed on the corresponding Miller index.

26.14.1 Miller Indices of Principal Planes in a Cubic Cell

Any plane in the crystal containing some atoms is a crystal plane. However, the effect of the planes with small atomic density on the physical properties of the crystal is negligible. Hence, the planes with appreciable atomic density are called principal planes.

Fig. 26.31 depicts the principal crystallographic planes of cubic crystal structure. One of the faces of the cube is shown shaded in Fig. 26.31 (a). The shaded plane intercepts the axes x , y , z at 1 , ∞ , ∞ respectively. The reciprocals of the intercepts are $1, 0, 0$. The Miller indices for this plane are (100) .

The plane in Fig. 26.31 (b) has the intercepts at $1, 1, \infty$. Therefore, the Miller indices are (110) and the plane is called one-one-zero plane. Finally, the plane in Fig. 26.31 (c) has intercepts at $1, 1, 1$ and hence the Miller indices are (111) . This plane is called one-one-one plane. These three planes are known as the **principal planes of cubic crystal system**.

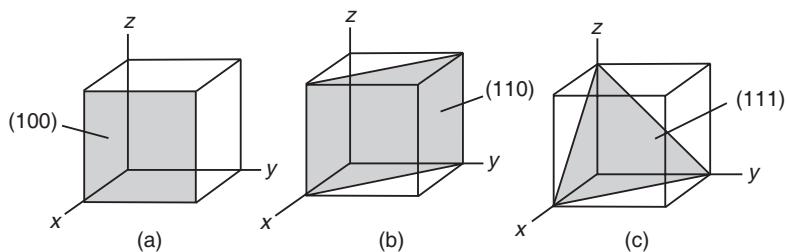


Fig. 26.31. The three principal Miller index planes in a cubic lattice.

26.14.2 Sketching a Lattice Plane (hkl)

The intercepts made by a plane can be found if the Miller indices of the plane are known.

- Let the Miller indices of a plane be (hkl) .
- A unit cell (for e.g. a cube) is drawn with three coordinate axes.
- From the Miller indices, the reciprocals are obtained $\rightarrow \frac{1}{h}, \frac{1}{k}, \frac{1}{l}$
- The intercepts are marked on the coordinate axes.
- The three non-planar points are joined and the plane is obtained.

Example: Draw a plane (321)

Step (i): A cube having *unit axial lengths* ($a = 1$) along the three coordinate axes is drawn as shown in Fig. 26.32.

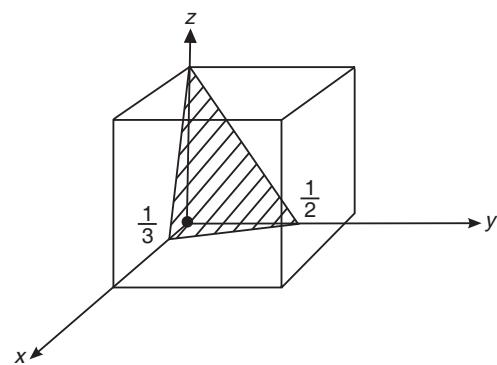


Fig. 26.32. (321) plane

Step (ii): The reciprocals of Miller indices are $p = \frac{1}{3}$, $q = \frac{1}{2}$ and $r = 1$.

Step (iii): The intercepts, $\frac{1}{3}, \frac{1}{2}$ and 1, are marked within the cube on x , y and z axes respectively.

Step (iv): A plane is drawn through the points and is shaded.

26.15 INTERPLANAR SPACING IN A CUBIC LATTICE

The distance ‘ d ’ between successive lattice planes is known as the *interplanar distance*. The interplanar distance ‘ d ’ involves the axial lengths of the unit cell and the Miller indices of the planes. We shall derive here an expression for d in the case of a cubic system only. We know that the three axes of a cubic crystal are mutually perpendicular.

Let ABC be one of the family of parallel lattice planes in the crystal (Fig. 26.33). Let its Miller indices be (hkl) and the intercepts on the crystallographic axes be $OA = a/h$, $OB = b/k$ and $OC = c/l$. The next plane of the set parallel to ABC passes through the origin of the coordinates O (which is not shown in the Fig. 26.33). Therefore, ON, the length of the normal from the origin to the plane is equal to ‘ d ’. Let α , β , and γ be the angles ON makes with the three crystallographic axes respectively. Then the direction cosines of ON are

$$\begin{aligned}\cos \alpha &= \frac{ON}{OA} = \frac{d}{a/h} \\ \cos \beta &= \frac{ON}{OB} = \frac{d}{b/k} \\ \cos \gamma &= \frac{ON}{OC} = \frac{d}{c/l}\end{aligned}\quad (26.29)$$

The sum of the squares of the direction cosines of a line equals unity. Thus,

$$\cos^2 \alpha + \cos^2 \beta + \cos^2 \gamma = 1 \quad (26.30)$$

Substituting the values of $\cos \alpha$, $\cos \beta$, and $\cos \gamma$ in the above equation, we get

$$\begin{aligned}\frac{d^2}{a^2/h^2} + \frac{d^2}{b^2/k^2} + \frac{d^2}{c^2/l^2} &= 1 \\ d^2 \left[\frac{h^2}{a^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2} \right] &= 1 \\ \therefore d &= \left[\frac{h^2}{a^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2} \right]^{-1/2}\end{aligned}\quad (26.31)$$

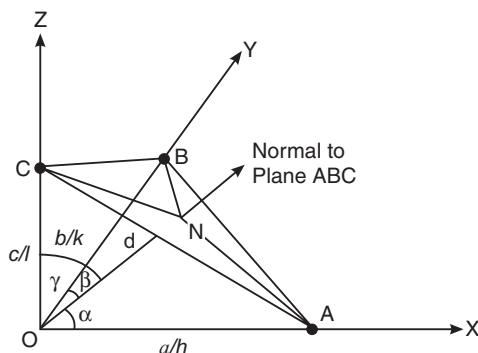


Fig. 26.33. Determination of interplanar distance in a cubic crystal

In case of cubic system, $a = b = c$. The above equation then reduces to

$$d_{hkl} = \frac{a}{\sqrt{(h^2 + k^2 + l^2)}} \quad (26.32)$$

Example 26.5. Determine lattice constant for FCC lead crystal of radius 1.746 \AA . Also find the spacing of (i) (111) planes, (ii) (200) planes and (iii) (220) planes.

Solution: The lattice constant is related to the atomic radius in FCC structure through

$$a = 2\sqrt{2}r = 2 \times 1.414 \times 1.746 \text{ \AA} = 4.938 \text{ \AA}$$

The interplanar spacing is given by $d = \frac{a}{\sqrt{h^2 + k^2 + l^2}}$

$$(i) \quad d_{111} = \frac{4.938}{\sqrt{1^2 + 1^2 + 1^2}} \text{ \AA} = 2.851 \text{ \AA}$$

$$(ii) \quad d_{200} = \frac{4.938}{\sqrt{2^2 + 0^2 + 0^2}} \text{ \AA} = 2.469 \text{ \AA}$$

$$(iii) \quad d_{220} = \frac{4.938}{\sqrt{2^2 + 2^2 + 0^2}} \text{ \AA} = 1.746 \text{ \AA}$$

Example 26.6. The density of copper is 8980 kg/m^3 and unit cell dimension is 3.61 \AA . Atomic wt. of copper is 63.54. Determine crystal structure. Calculate atomic radius and interplanar spacing of (110) plane.

Solution: (i) The effective number atoms per unit cell is given by $Z = \frac{\rho N_A a^3}{M}$

$$\therefore Z = \frac{8980 \text{ kg/m}^3 \times 6.02 \times 10^{26} / \text{k.mol} \times (3.61 \times 10^{-10} \text{ m})^3}{63.54 \text{ kg/k.mol}} = \frac{254.32}{63.54} = 4 \text{ atoms/unit cell.}$$

∴ Copper exhibits FCC structure.

(ii) The lattice constant is related to the atomic radius in FCC structure through $a = 2\sqrt{2}r$

$$r = \frac{a}{2\sqrt{2}} = \frac{3.61 \text{ \AA}}{2 \times 1.4142} = 1.276 \text{ \AA}$$

(iii) The interplanar spacing is given by $d = \frac{a}{\sqrt{h^2 + k^2 + l^2}}$

$$\therefore d = \frac{3.61 \text{ \AA}}{\sqrt{1^2 + 1^2 + 0^2}} = 2.553 \text{ \AA.}$$

Example 26.7. A crystal with primitives 1.2 \AA , 1.8 \AA and 2 \AA has a plane (231) which cuts an intercept 1.2 \AA along x-axis. Calculate the intercepts along y- and z-axes.

Solution: Intercept on X-axis, $pa = 1.2 \text{ \AA}$ $\therefore p = \frac{1.2 \text{ \AA}}{1.2 \text{ \AA}} = 1$.

$$\text{L.C.M, } D = hp = 2 \times 1 = 2.$$

$$\therefore q = \frac{D}{k} = \frac{2}{3} \text{ and } r = \frac{D}{l} = \frac{2}{1}.$$

$$\therefore \text{Y-intercept, } qb = \frac{2}{3} \times 1.8\text{\AA} = 1.2\text{\AA}$$

and Z-intercept, $rc = 2 \times 2\text{\AA} = 4\text{\AA}$.

26.16 ATOMIC PACKING

The unit cells are geometric models connecting points which are purported to be the atomic centres. Atoms are space-filling entities. Therefore, the crystal structure may be viewed, alternatively, as resulting from the packing of hard spheres in three dimensions. The most efficient packing of spheres is called **closest packing**, or simply **close packing**.

Let us consider the packing of hard spheres of identical size. They can be close packed in a row by arranging them in contact with each other. A two dimensional layer of spheres can be built in two ways as shown in Fig. 26.34. Fig. 26.34 (a) shows square packing and triangular packing of two rows. A comparison of the two types suggests that the spheres fit more compactly in triangular packing rather than in square packing. When more rows are added to build the layer, a close packing such as the one shown in Fig. 26.34 (b) is obtained.

In a layer of close packed spheres, each sphere is in contact with six neighbours. The lines joining the centres of the six spheres form a hexagon. Therefore, such a packing is commonly called **hexagonal packing**. Hexagonal packing is the most compact packing possible within a layer. Such a layer is known as a hexagonal close packed layer. Three dimensional close packed structures are built by keeping hexagonal close packed layers on top of one another in a regular sequence.

In practice, identical spheres can be packed in different ways to produce three dimensional

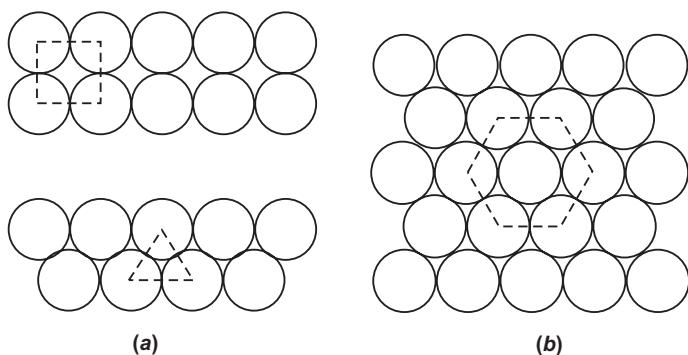


Fig. 26.34

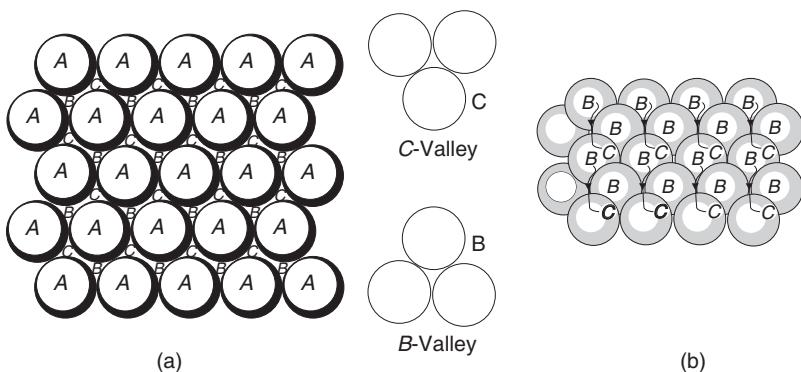


Fig. 26.35

structures. Different kinds of close packing are theoretically possible but all are combinations of two different basic structures, namely HCP structure and CCP structure. They are built starting with a hexagonal packed layer.

26.16.1 HCP Structure

Fig. 26.35 (a) shows the top view of a hexagonal close packed layer of identical spheres. Let us designate the spheres as A and hence the layer as A-layer.

Successive layers of spheres are added on top of the A-layer to form three dimensional structures. As spheres have curved surfaces, a regular array of valleys are formed between the spheres in the A layer. Each valley is formed where three spheres are in contact with each other. Each sphere in the layer is surrounded by six neighbour spheres and six valleys (Fig. 26.35 a). The valleys may be divided into two sets of differing shapes and are labeled B and C. Thus, there are three B-type valleys and three type C-valleys. If a sphere is placed in a B valley, none of the three adjacent C-valleys can be filled with a sphere. Similarly, if a sphere is placed in a C valley, none of the three adjacent B-valleys can be filled with a sphere. It means that the B and C valleys are so close that they cannot be simultaneously occupied. Therefore, when spheres are placed on top of the A-layer, all of them must roll down into one kind of valley, either B or C. Thus, the second layer spheres occupy only one kind of valley leaving the other kind vacant. It means that there is only one way of arranging the second layer. Fig. 26.36 shows the close packing of identical spheres in two layers. The centres of the spheres in the upper layer lie above the B-valleys. Let us designate this layer as B-layer.

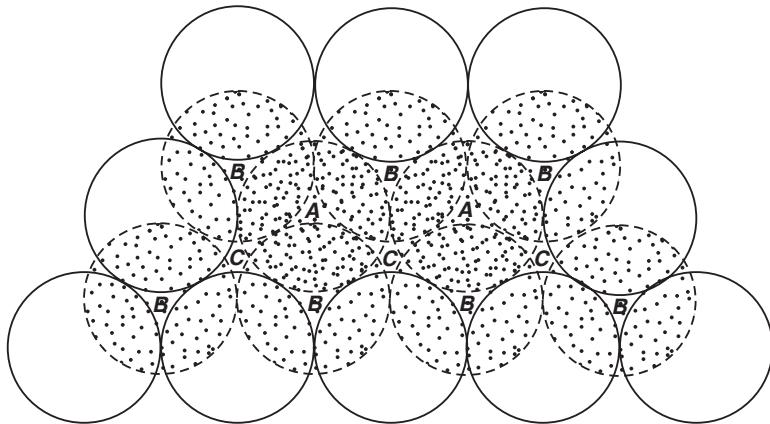


Fig. 26.36

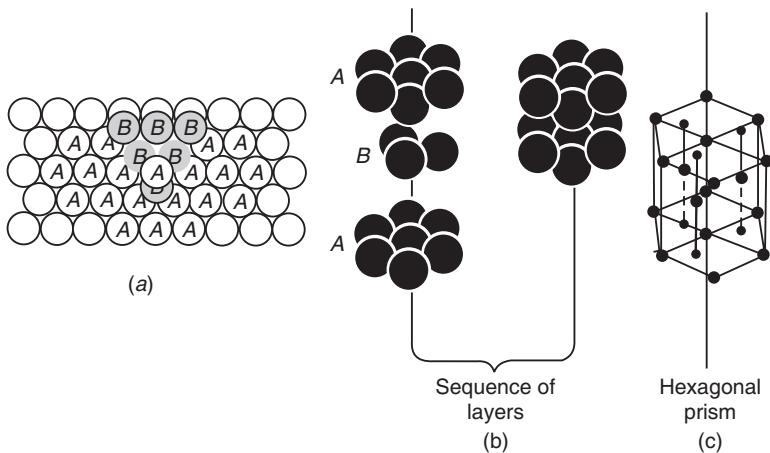


Fig. 26.37

Again in the B-layer, each sphere is surrounded by two sets of valleys marked CCC and AAA. Now there are two options. The spheres can be positioned in A valleys or C valleys to form the third layer. If spheres are placed in A valleys, the third layer becomes equivalent to the first layer at the bottom. It means that it is again an A-layer. Similarly, the fourth layer may be arranged such that its spheres sit in B valleys of third layer, forming a B-layer. Spheres may be arranged to form the fifth layer as layer, sixth layer as B-layer and so on. This kind of

stacking of spheres is essentially a repeating two-layer-close-packing. The packing is known as AB AB ABlayer sequence. In this structure, the C valleys are left vacant from the bottom layer to the topmost layer. Fig. 26.37 (a) shows the schematic of AB AB AB....layer sequence.

If the centres of six spheres surrounding a particular sphere in the first layer spheres and those of the third layer spheres are joined, a right prism having a regular hexagon for its base is obtained (Fig. 26.37 b). Therefore, the unit cell of this structure is a hexagon. In view of this, the AB AB AB....layer packing is called hexagonal close packing and the resulting three dimensional structure is known as **hexagonal close packed structure or HCP structure**.

26.16.2 CCP Structure

If the spheres of the third layer are positioned in C valleys instead of A valleys of second layer, a different sequence of layers is produced. The third layer built on C valleys is denoted as C-layer. In the C-layer, Each sphere is surrounded by AAA and BBB valleys. If the spheres of fourth layer are stacked on A-valleys,

it becomes a repetition of the first layer. Similarly, the fifth layer may be arranged as a repetition of B-layer, the sixth that of C-layer , the seventh that of A-layer and so on. The structure is symbolically represented as ABC ABC ABC.....layer sequence. Fig. 26.38 shows the sequence of layers.

The fundamental block of ABC ABC ABC...structure is a cube (Fig. 26.38 a). Therefore, the packing is known as **cubic close packed structure or CCP structure**. The atoms are located at the corners of the cube and at the centres of the faces of the cube. The CCP structure is in fact FCC structure only.

Atoms which are attracted to each other by nondirectional forces are likely to form close packed structures. Thus, the noble gases crystallize in CCP structure, with the exception of helium which crystallizes in HCP structure. Metallic bond is nondirectional and non saturable. Therefore, metals also crystallize in close packed structures.

Thus, Cd, Co, Mg, Zn, Ti etc crystallize in HCP structure whereas Ag, Al, Cu, Ni, Pt etc crystallize in CCP structure.

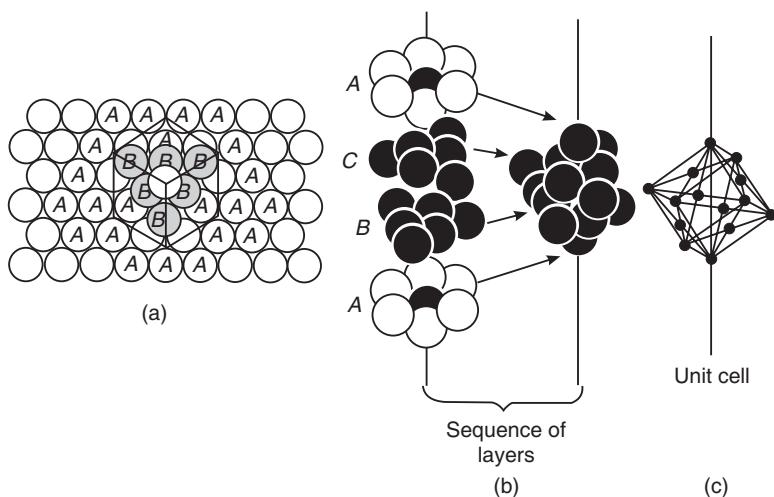


Fig. 26.38

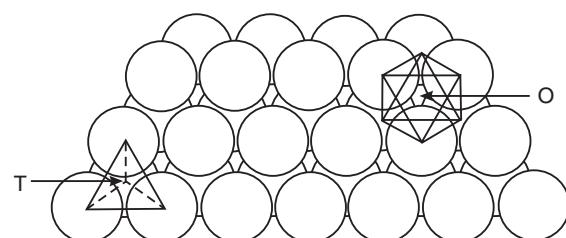


Fig. 26.39. Formation of tetrahedral and octahedral voids in a three dimensional close packed structure

26.17 VOIDS

All the available space in a crystal is not filled with atoms. Even in the close packed structures, it is seen that about 26% of the volume is left vacant. These vacant spaces between the atoms in the crystal are called **voids**. They are also known as **interstices**. Two kinds of voids form in close packed structures, such as FCC close packed structures. They are **tetrahedral voids** and **octahedral voids** (see Fig. 26.39). A **tetrahedral void** is formed by a sphere fitting into the valley formed between three adjacent spheres of a close packed layer, as shown in Fig. 26.40 (a). If the centres of the four spheres are joined, a regular tetrahedron is produced. The vacant site enclosed by the tetrahedron is known as the tetrahedral void. If the fourth sphere belongs to the upper layer and is on top of three spheres, an upright tetrahedral void forms. On the other hand, if the fourth sphere belongs to the bottom layer an inverted tetrahedral void forms. Thus, in a three-dimensional structure two tetrahedral voids form for every sphere.

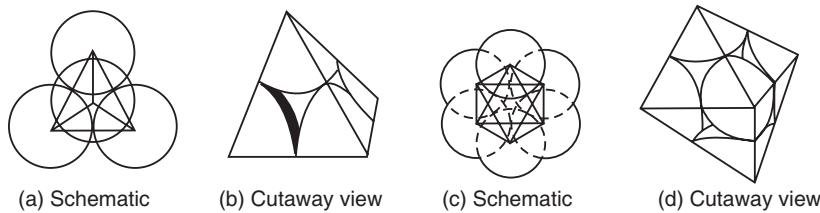


Fig. 26.40. Tetrahedral Void (a) Schematic (b) cutaway view; Octahedral void (c) Schematic (d) cutaway view

An **octahedral void** is produced when a void formed by three spheres in one layer comes on the top of the void formed by three spheres in the top layer, as depicted in Fig. 26.40 (c). When the centres of the six spheres are joined, an octahedron is formed. The space enclosed is an octahedral void. An octahedral void is surrounded by six spheres. There is no sphere in the central valley. Octahedral voids are bigger in size and smaller in number compared to tetrahedral voids. In a three-dimensional close packed structure, one octahedral void forms per one sphere.

26.17.1 Sizes of the Spheres that Fit into Voids

(i) The tetrahedron in Fig. 26.40 (b) may be used to determine the radius of the sphere that can fit into the tetrahedral void. A cube ABCDEFGH is formed using the tetrahedron CAEH (Fig. 26.41a). The point C is the apex of the tetrahedron and is one of the corners of the cube. The edges AE, AH and EH of the tetrahedron are the face diagonals of the faces ABEG, ADFH, and EHG of the cube. The body diagonal AEGD of the cube is shown separately in Fig. 26.41 (b).

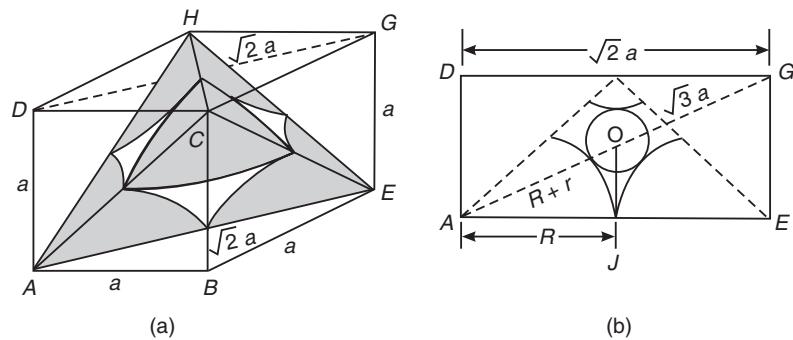


Fig. 26.41

From the Fig. 26.41 (b) it is seen that the Δ^{les} AJO and AEG are similar.

$$\therefore \frac{AO}{AJ} = \frac{AG}{AE} \quad \therefore \frac{R+r}{R} = \frac{\sqrt{3}a}{\sqrt{2}a}$$

$$\therefore 1 + \frac{r}{R} = \frac{\sqrt{3}}{\sqrt{2}} \quad \therefore r = \left[\frac{\sqrt{3} - \sqrt{2}}{\sqrt{2}} \right] R = 0.225 R \quad (26.33)$$

(ii) For finding the size of the sphere that can fit tightly into an octahedral void, let us consider the octahedron. In Fig. 26.41(d). The plane ABCD passing through four spheres surrounding the central valley is shown in Fig. 26.42.

The triangles ABC and ADC are right isosceles triangles. If each side of the square is a , in the Δ^{le} ABC

$$\frac{AC}{BC} = \frac{\sqrt{2}a}{a}$$

$$\therefore \frac{2R+2r}{2R} = \sqrt{2}$$

or $R+r = \sqrt{2}R$

$$\therefore r = (\sqrt{2}-1)R \quad \text{or} \quad r = 0.414$$

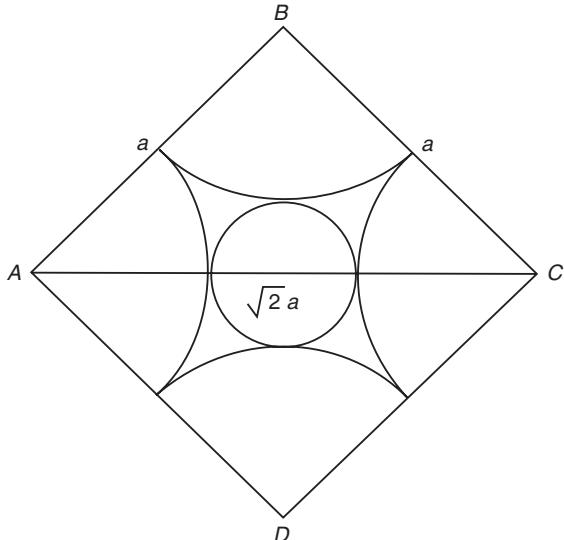


Fig. 26.42

R (26.34)

26.18 IONIC SOLIDS

Ionic solids are made up of ions of different sizes. The cations are ordinarily smaller than anions as cations have given up their valence electrons to the anions. The anions that are relatively larger in size than the cations, form close packed structures and the cations are accommodated in the voids. The packing depends upon the ionic radii of cations, r_C and anions, r_A . Each cation prefers to have as many nearest neighbours as possible. So does each anion. The coordination number is therefore determined by cation-anion radius ratio r_C/r_A . This is called **radius ratio effect**. This determines the efficiency of packing and forms the basis of many ceramic (inorganic materials) structures.

Some of the simple ionic solids are those in which there are equal number of cations and anions. These are often referred to as AX compounds where A denotes the cation and X the anion. The most common AX crystal structure is the sodium chloride type.

NaCl CRYSTAL STRUCTURE

The NaCl crystal is an ionic crystal. During the formation of the crystal, Na atom loses its outer electron and becomes a positive ion. On the other hand, the chlorine ion acquires the electron lost by the Na atom and becomes a negative ion. When the two ions approach each other, at a certain distance the forces of attraction are balanced by forces of repulsion and a stable system is produced. Na^+ and Cl^- ions arranged alternately in a cubic pattern in space so that the electrostatic attraction between the nearest neighbours is maximum. The resultant structure has

a coordination number 6. Six anions surround each cation and each anion is surrounded by six cations. Thus, the cations occupy the octahedral voids formed by the anions (Fig. 26.43 a). Since the ratio of anions to cations is 1:1 and there is one octahedral void per anion, all octahedral voids in the structure are occupied.

NaCl is an example of *face centered cubic* (fcc) Bravais lattice. Each cell has 8 corners and 8 cells meet at each corner. Thus, an ion at a corner of the cell is shared by 8 cells, i.e. only $1/8$ ion belongs to any one cell. Similarly, an ion at the center of a face of the cell is shared by 2 cells, i.e. only $1/2$ ion belongs to any one cell. Since a cell has 8 corners and, 6 faces, it has $[(8 \times 1/8) + (6 \times 1/2)] = 4$ ions of one kind, and similarly 4 ions of other kind. Thus, *there are 4 $\text{Na}^+ - \text{Cl}^-$ ion pairs (molecules) per unit cell*. The position coordinates of Na^+ and Cl^- are as follows.

Na	0 0 0	$\frac{1}{2} \frac{1}{2} 0$	$\frac{1}{2} 0 \frac{1}{2}$	$0 \frac{1}{2} \frac{1}{2}$
Cl	$\frac{1}{2} \frac{1}{2} \frac{1}{2}$	$0 0 \frac{1}{2}$	$0 \frac{1}{2} 0$	$\frac{1}{2} 0 0$

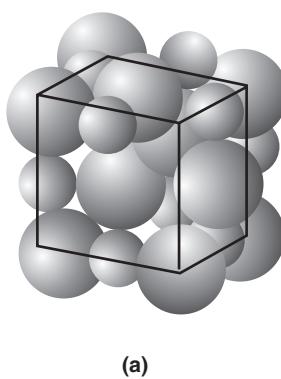
Fig. 26.43 (b) shows a unit cell of NaCl lattice. NaCl lattice can be regarded to be made up of two fcc sub-lattices, one of Na^+ ions having origin at (0,0,0) and the other of Cl^- ions having its origin midway along the cube edge.

Each Na^+ ion has 6 Cl^- ions as nearest neighbours and similarly, each Cl^- ion has 6 Na^+ ions. Hence, the coordination number of NaCl is 6, the same as that for, simple cubic lattice.

26.19 DIAMOND CUBIC STRUCTURE

In a diamond crystal the carbon atoms are linked by directional covalent bonds. Each carbon atom forms covalent bonds with four other carbon atoms that occupy four corners of a cube in a tetrahedral structure (Fig. 26.44 a). The length each bond is 1.53 Å and the angle between the bonds is 109.5° . The entire diamond lattice is constructed of such tetrahedral units.

Diamond exhibits both cubic and hexagonal type structure. The



(a)

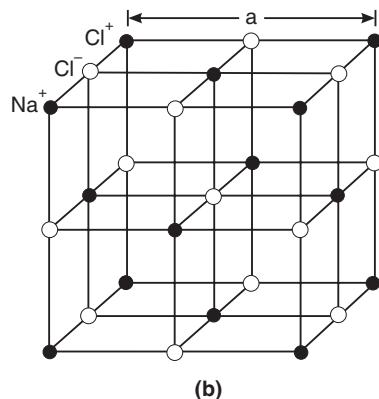
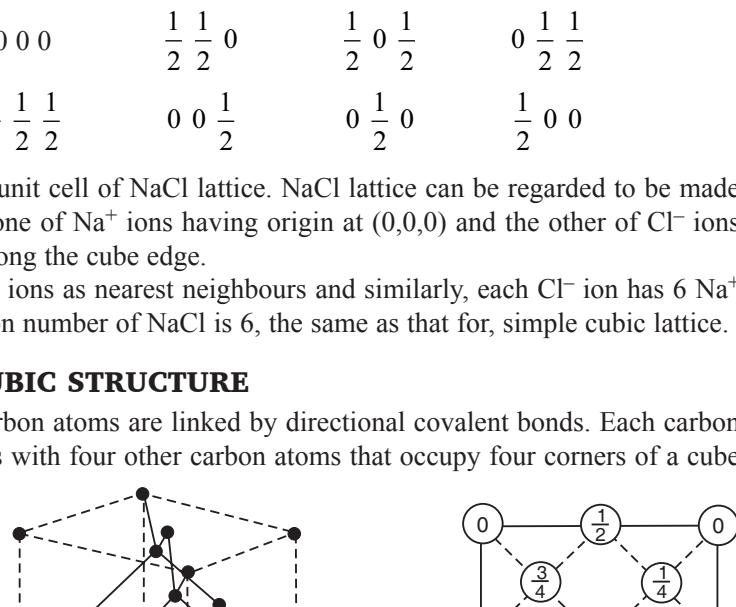


Fig. 26.43: NaCl crystal



(a)

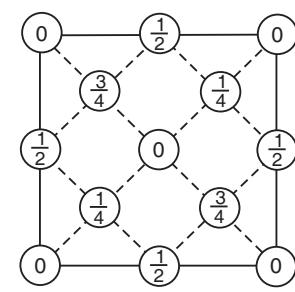


Fig. 26.44: Diamond crystal

diamond cubic (dc) structure is more common. The space lattice of diamond is face centered (fcc) with a basis of two carbon atoms associated with each lattice point. Fig. 26.44 (b) shows the position of atoms in the cubic cell of the diamond structure projected on a cubic face. The fractions denote the height above the base in units of a cube edge. The points at 0 and $1/2$ are on the fcc lattice, those at $1/4$ and $3/4$ are on a similar lattice displaced along the body diagonal by one fourth of its length. Thus, *the diamond lattice is composed of two interpenetrating fcc sublattices, one of which is shifted relative to the other by one fourth of a body diagonal.*

In the diamond lattice, each atom has four nearest neighbours with which it forms covalent bonds. Thus, the coordination number of diamond crystal is 4. The number of atoms per unit cell is 8.

Silicon, germanium and gray tin crystallize in the diamond structure.

26.20 ZnS STRUCTURE

The structure of zinc sulphide is identical to the diamond cubic structure. It consists of two interpenetrating FCC sub lattices which are occupied by two different elements and are displaced from each other by one quarter of the body diagonal. Zinc sulphide structure results when Zn atoms are placed on one FCC lattice and S atoms on the other FCC lattice, as shown in Fig. 26.45. The conventional pattern of this structure is a cube. There are four molecules per conventional cell. For each atom there are four equidistant atoms of other kind arranged at the vertices of a regular tetrahedron. Some of the important compounds which possess this structure are semiconductors such as InSb, GaAs, and CdS.

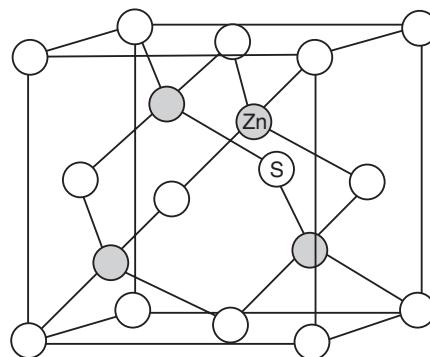


Fig. 26.45: Zinc sulphide crystal

26.21 POLYMORPHISM AND ALLOTROPY

The phenomenon by which a single substance crystallizes in two or more different forms under different conditions is called **polymorphism** and the different crystalline forms are said to be polymorphous with each other. Polymorphism is found among pure elements as well as among chemical compounds. Polymorphism of an element is frequently called *allotropy*.

An element that exists in more than one crystalline form is said to be **allotropic** and the forms are referred to as **allotropes** and the phenomenon is called **allotropy**. The prevailing crystal structure depends on the temperature and the external pressure. Pure iron has a BCC crystal structure at room temperature, which changes to FCC structure at 912°C . Another example is carbon. Graphite is the stable allotrope at ambient conditions, whereas diamond is formed at extremely high pressures.

26.22 GRAPHITE STRUCTURE

The crystalline allotropes of carbon are diamond and graphite. The atoms of carbon can form two types of space lattices, namely diamond and graphite structures. Diamond structure is discussed in Art. 26.19.

Graphite structure

The second type of packing of carbon atoms occurs in graphite. In graphite each atom is bound to four neighbours but the forces and the directions of the bonds differ (Fig. 26.46).

An atom has a strong bond with three other atoms that lie together with it in the same layer at angles of 120° and a weaker bond with a fourth atom lying in the adjacent layer. This last bond makes an angle of 90° with the plane of the layer. The unit cells are right hexagonal prisms with edges 1.42 \AA and 3.35 \AA in length. The distance between the layers is 2.36 times more than between the atoms lying in one layer and therefore the bonds between layers are weak.

In each layer, each carbon atom is covalently bonded to three others involving sp^2 hybrid orbital instead of four as in diamond. Thus, all atoms in a single plane are linked to give flat hexagons. The hexagons are held together in sheet-like structure, parallel to one another. The C-C covalent bond distance is 1.42 \AA . The distance between the sheets of layers, however is comparatively larger being about 3.35 \AA . This rules out the possibility of covalent bonding between the layers. Thus graphite has a **layer structure**.

The difference in structure of the diamond and graphite lattices is the cause of the striking differences in their physical properties. Diamond is one of the hardest substances in nature; it can cut glass, as well as any hard rock like granite. Graphite is very soft; it writes on the paper because its layers slide easily with respect to each other. The diamond is an insulator as it has no free electrons; but graphite is a conductor. Diamond is transparent whereas graphite is opaque. The density of diamond is 3500 kg/m^3 ; that of graphite is 2100 kg/m^3 .

26.23 CRYSTAL STRUCTURE ANALYSIS

Much of our knowledge of the internal structure of crystals has been obtained from x-ray diffraction experiments. Diffraction of waves occurs when the waves are scattered by a periodic array of scattering centres separated by distances of the order of a wavelength. A plane transmission grating is a familiar device used to study diffraction of light. It is made by ruling thousands of parallel lines on a piece of glass. The lines are opaque and the spaces between the lines form openings for light to pass through. A grating normally used in laboratories has 15000 lines per inch. The slit width or grating period in the grating comes to about $17,000\text{ \AA}$, which is about three times the wavelength of light. When light passes through these narrow slits, the waves spread in all directions. The diffracted waves interfere constructively in a few specific directions and destructively in all other directions, producing a spatial pattern of alternate regions of brightness and darkness. The dimension of atoms and the interatomic spacing in a crystal are of the order of 2 to 5 \AA which is of the order of the wavelength of X-rays. In view of this Max von Laue, the German physicist predicted in 1912 that a crystal acts as a natural three dimensional grating for x-rays, where the regular periodic lines of atoms serve the role of parallel ruled lines. His associates, W.Friedrich and P.Knipping duly tested and confirmed the prediction. Laue was awarded the Nobel prize in Physics in 1914.

26.23.1 Laue Method

X-rays produced by an x-ray tube are defined into a narrow beam by a set of lead screens S_1 and S_2 having pin holes at their centres. A thin crystal C is mounted in the path of the X-ray

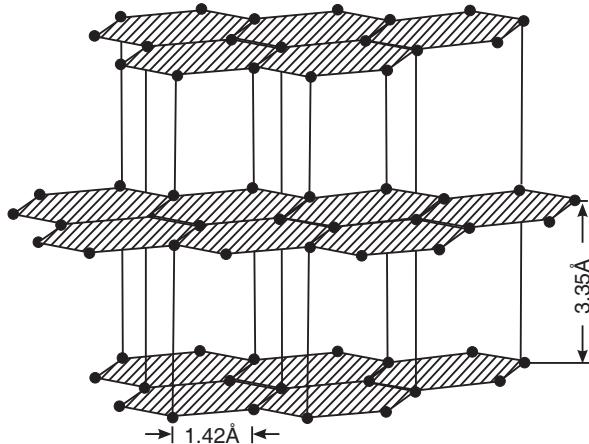


Fig. 26.46: Graphite

beam and a photographic film is positioned beyond it, as shown in Fig. 26.47 (a).

As the X-ray beam penetrates the crystal C, some of the rays are scattered

by atoms from their initial direction. The scattered x-rays emerge from the crystal in specific directions as highly narrow beams and they are intercepted by the photographic film. On developing the exposed film, a pattern of bright spots corresponding to maximum intensity are observed (Fig. 26.47 b). They are more commonly referred to as **Laue spots**. The pattern of Laue spots is uniquely characteristic of the crystal C. The central bright spot on the film corresponds to the main beam. A hole is often cut in the film so that the central spot is not recorded. In the experiment conducted by Friedrich and Knipping, the wavelength of X-rays was determined from the measurement of angular positions of Laue spots and from the knowledge of the separation of atoms in the crystal. This method has been subsequently used to determine the crystal structures using X-rays of known wavelength. However, the analysis of Laue pattern is mathematically complicated. At present the Laue method is only of historical interest.

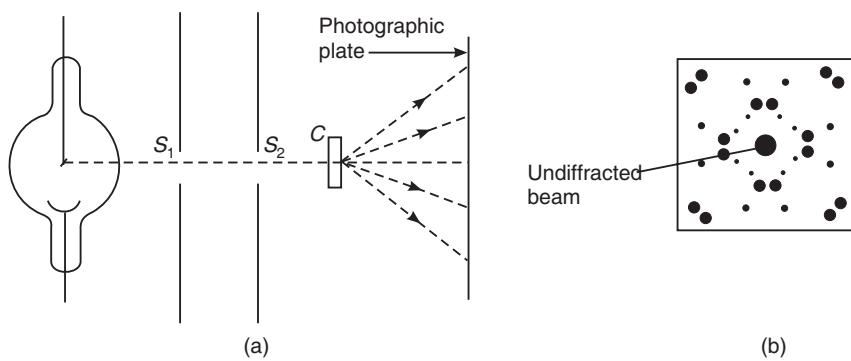
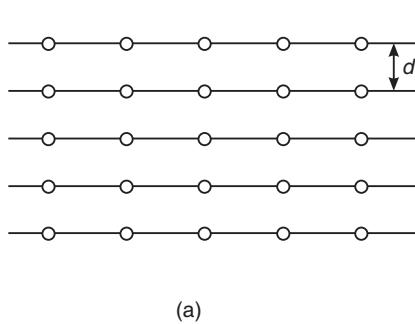


Fig. 26.47

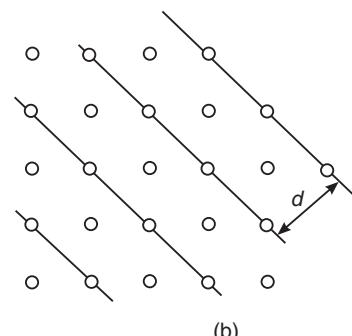
26.24 BRAGGS' LAW

In 1912 W.H.Bragg and W.L.Bragg discovered that X-rays can be regularly reflected by the cleavage planes of the crystals. The cleavage planes are successive atomic planes in the crystal (Fig. 26.48).

They are also known as **Braggs' planes**. Thus, a crystal may be regarded as consisting of a stack of several parallel planes of



(a)



(b)

Fig. 26.48. Two different sets of Bragg planes

atoms. Each plane in a given set has the same distribution of atoms. When an X-ray beam is incident on the crystal, each atom in the crystal scatters a portion of the incident beam. The scattered waves travel in all directions. It is convenient to consider the net scattering by atoms in terms of the diffraction (scattering) by the crystal planes. Each family of planes gives rise to scattering but only a certain family scatters constructively in a given direction. W.H.Bragg and W.L Bragg derived a relation between the wavelength of X-rays and the angular positions of the scattered beam and the separation of atomic planes in the crystal.

Braggs made the following simplifications regarding the diffraction of X-rays from a crystal:

1. Any crystal may be viewed as a regular stack of atomic planes;
2. The atomic planes act like semitransparent mirrors for X-rays;
3. X-rays reflected from the successive atomic planes would interfere constructively or destructively depending upon the path difference between the rays;
4. Whenever the path difference between the rays is an integer multiple of wavelength (λ), the rays interfere constructively and produce a bright spot in that direction.

Thus, the complex phenomenon of diffraction of X-rays by the atoms was converted into the problem of reflection of X-rays by the parallel atomic planes. Hence, the words ‘diffraction’ and ‘reflection’ are mutually interchangeable in Braggs’ treatment. Based on these considerations, Braggs derived a simple mathematical relationship that serves as a condition for the Bragg reflection to occur. This condition is known as Braggs’ law.

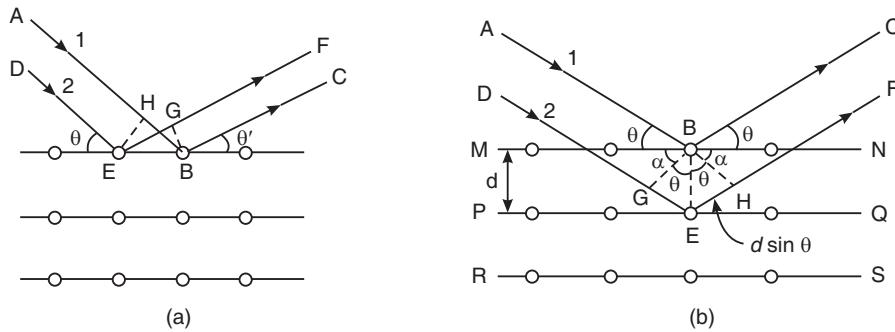


Fig. 26.49. Determination of path difference between X-rays diffracted by consecutive lattice planes

Let us consider a set of parallel atomic planes with interplanar spacing d . We represent the row of atoms as a single horizontal line, as shown in Fig. 26.49. Let a parallel beam of monochromatic X-rays of wavelength λ , represented by the parallel lines AB and DE, be incident on these planes at a glancing angle θ (Fig. 26.49 a). The scattered beam emerges along BC and EF. The contributions of two adjacent atoms in the same plane are considered in the Fig. 26.49 (a).

The rays BC and EF are coherent and reinforce each other, if they are in phase. It requires that the path lengths BK and EL are equal. They will be equal when $\theta = \theta'$. This is the *condition of reflection*.

We consider next the contributions of two adjacent atoms to the reflection from **successive planes** MN and PQ are considered (see Fig. 26.49 b). The path difference Δ between the reflected rays BC and EF is

$$\Delta = GE + EH \quad (26.35)$$

$$\angle ABG = \theta + \alpha = 90^\circ \quad (26.36)$$

BE is normal to the plane MN. Therefore,

$$\angle MBE = \alpha + \angle GBE = 90^\circ \quad (26.37)$$

From equ. (26.36) and (26.37) we get

$$\angle GBE = \theta$$

Similarly, $\angle EBH = \theta$

In the Δ^{le} BGE, $BE = d$

$$\therefore GE = d \sin \theta$$

Similarly, in the Δ^{le} EBH, $EH = d \sin \theta$

$$\therefore \text{The path difference } \Delta = GE + EH = d \sin \theta + d \sin \theta = 2d \sin \theta \quad (26.38)$$

The rays BC and EF will constructively interfere only when $\Delta = m\lambda$, where $m = 1, 2, 3, \dots$

\therefore The condition for reinforcement of scattered waves is

$$2d \sin \theta = m\lambda \quad (26.39)$$

The above equation is called **Bragg's equation or Braggs' law**.

The meaning of the above equation is that the reinforcement of reflected waves will take place only at certain values of θ , corresponding to specific values of λ and d . In these directions the resultant wave has the maximum amplitude and produces bright spots on the photographic plate placed in their path. On the other hand, at other angles, the scattered waves may not be in phase with each other and hence their amplitude will be zero, leading to dark spots in those directions.

Order of reflection:

The reflection angle θ is given by

$$\theta = \sin^{-1} \left(\frac{m\lambda}{2d} \right) \quad (26.40)$$

For a fixed wavelength, λ , the angle θ depends d_{hkl} , the interplanar distance.

When the lattice planes of indices (hkl) give rise to x-ray reflection, the path difference between the rays reflected from successive planes is one wavelength. For example, if the reflection from (100) planes of the lattice occurs, say, at an angle θ_1 , then the path difference between the reflections will be λ and the order of reflection $m = 1$. When the path difference between the reflected rays is 2λ , the reflection will occur at an angle θ_2 and the Bragg equation is satisfied with $m = 2$. Thus, as θ is increased gradually, a number of positions will be found at which the reflections are intense. These positions correspond to $m = 1, 2, 3, \dots$ values. The diffraction lines appearing for $m = 1, 2$ and 3 are called first, second, third order diffraction lines respectively. Thus, m denotes the **order of reflection**. The highest possible order is determined by the condition that $\sin \theta$ cannot exceed unity.

If the value of θ is determined experimentally, and knowing the wavelength λ , the interplanar spacing ' d ' can be determined with the help of Bragg law.

Example 26.8. X-rays of unknown wavelength give first order Bragg reflection at glancing angle 20° with (212) planes of copper having FCC structure. Find the wavelength of X-rays, if the lattice constant for copper is 3.615 \AA .

Solution: The interplanar spacing is given by $d = \frac{a}{\sqrt{h^2 + k^2 + l^2}}$

$$\therefore d = \frac{3.615 \text{ \AA}}{\sqrt{2^2 + 1^2 + 2^2}} = 1.205 \text{ \AA.}$$

According to Braggs' law

$$2d \sin \theta = m\lambda.$$

$$\therefore \lambda = \frac{2d \sin \theta}{m} = \frac{2 \times 1.205 \text{ \AA} \times \sin 20^\circ}{1} = 0.824 \text{ \AA}.$$

26.25 BRAGGS' SPECTROMETER

W.H.Bragg and W.L.Bragg devised an X-ray spectrometer in which a crystal is used as a reflection grating. It is used to measure glancing angle θ .

Construction: Braggs' spectrometer is very much similar in construction to the optical spectrometer. The schematic diagram of the Braggs' spectrometer is shown in Fig. 26.50. X-rays from an X-ray tube are collimated into a narrow beam by two slits S_1 and S_2 cut in lead plates. The beam is then allowed to be incident at a glancing angle on the face of the crystal D, which is mounted in wax on the turntable of the spectrometer. The turntable is capable of rotation about a vertical axis passing through its center and the position of the crystal can be read from the circular scale, C. Most of the incident beam passes straight through the crystal. Some of the X-rays are however scattered by the regularly arranged atoms in different crystal planes. The scattered X-rays can be regarded as having been reflected from the crystal planes rich in atoms. The reflected X-ray beam enters an ionization chamber carried by the spectrometer arm, which is capable of rotation about the same axis as the turntable. The turntable and the arm are so linked together that when the turntable rotates through an angle θ , the arm turns through an angle 2θ . In this way, the X-ray beam is always reflected into the ionization chamber whatever its incidence angle is at the crystal face. The ionization current produced by the reflected beam is measured by a sensitive electrometer E or recorded on a photographic plate.

Working: Initially, the single crystal under investigation is mounted on the turntable such that the glancing angle $\theta = 0^\circ$ and the ionization chamber is adjusted to receive X-rays. Then the crystal is moved in small angles and the corresponding deflection obtained in the electrometer is noted down. A graph is plotted for glancing angle versus intensity of diffracted X-rays.

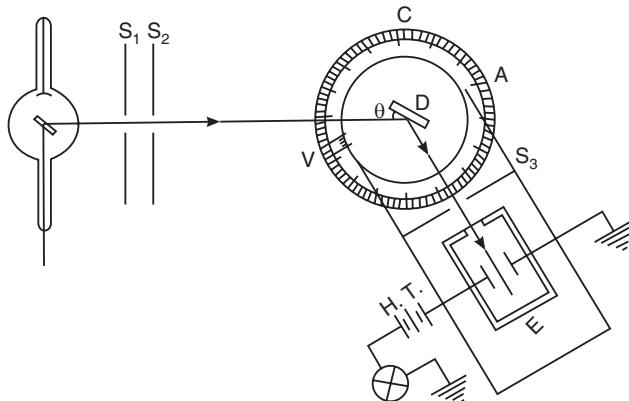


Fig. 26.50. The schematic diagram of the Bragg's spectrometer

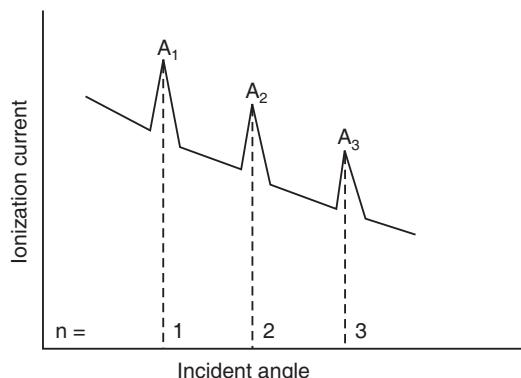


Fig. 26.51. A typical diffraction spectrum produced by a monochromatic X-rays scattered from a crystal.

X-ray spectrum: Fig. 26.51 shows a typical diffraction spectrum produced by a monochromatic X-rays scattered from a crystal. The diffraction spectrum is a graph plotted between the intensity of ionization and the glancing angle. It is seen from the graph that the intensity of ionization current increases abruptly for certain values of glancing angle. The peaks A_1, A_2, A_3 etc. in intensity occur whenever the Braggs' equation is satisfied. If θ_1, θ_2 and θ_3 are the angles of incidence where intense peaks occur, then they correspond to the different orders of reflection with $n = 1, 2$ and 3 respectively for a given wavelength. Using Braggs' law

$$2d \sin \theta_1 = \lambda, \quad 2d \sin \theta_2 = 2\lambda \quad \text{and} \quad 2d \sin \theta_3 = 3\lambda$$

That is, $2d \sin \theta_1 : 2d \sin \theta_2 : 2d \sin \theta_3 = \lambda : 2\lambda : 3\lambda$

$$\therefore \sin \theta_1 : \sin \theta_2 : \sin \theta_3 = 1 : 2 : 3 \quad (26.41)$$

The above ratio indicates that the peaks correspond to first, second and third order reflections respectively.

It may be seen from Fig. 26.51 that

- the intensity of the reflected beam decreases as the order of the spectrum increases.
- The intensity of the reflected beam never falls to zero. It reaches down only to a minimum value. This is due to the presence of continuous spectrum.

At a given time more than one set of planes give rise to reflections. Usually, the reflections are produced due to principal planes in the crystals.

26.25.1 Determination of Lattice Constant:

Bragg's law may be used to find the interplanar spacing, if the wavelength of X-rays is known. If we substitute the value of d in terms of the Miller indices of the planes in the relation connecting the interplanar spacing, d and the lattice constant, a , we can determine the lattice constant of the crystal. Thus, in case of a cubic crystal, the relation between d and a is given by

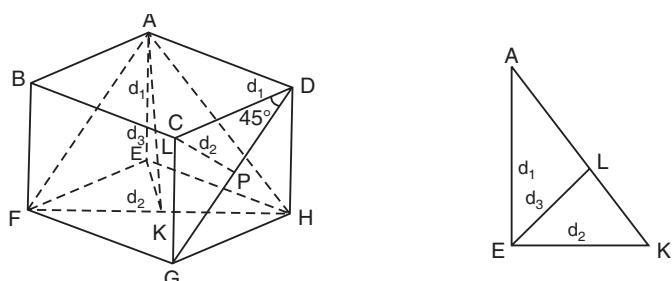
$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}} \quad (26.42)$$

Knowing the value of d , the lattice constant a can be calculated from the above equation.

26.25.2 Determination of Crystal Structure

The Braggs' method of determining crystal structure consists in finding out the interplanar spacing, d_{hkl} of the various sets of planes rich in atoms. From the knowledge of d_{hkl} and from the relative intensities of diffracted beams of different orders, the crystal structure can be deduced.

For example, let us consider a crystal of a simple cubic system. In the unit cell the atoms are located at the corners of the cell. It may be seen that three sets of planes are rich in atom.



(i) Referring to Fig. 26.52, we see that the set of planes

ABFE, CDHG, ADHE, BCGF, ABCD and EFGH are all alike. Let the d_2 distance between the consecutive planes be d_1 . These planes are known as (100) planes.

Fig. 26.52

(ii) The second set of planes consists of parallel planes like ADGF, inclined at an angle of 45° to the planes mentioned in (i). Let d_2 be the spacing between the second set of planes, as shown in Fig. 26.52 by CP. Note that a plane parallel to ADGF contains the line BC. Then, from the ΔCDP , we find that

$$\frac{d_2}{d_1} = \sin 45^\circ = \frac{1}{\sqrt{2}}$$

or
$$\frac{d_2}{d_1} = \frac{1}{\sqrt{2}} \quad (26.43)$$

These planes are known as (110) planes.

(iii) The third set of planes consists of planes like AFH. Let us draw EK perpendicular to FH and join AK to complete the triangle AEK. The perpendicular EL denoted by d_3 represents the distance between the plane AFH and a parallel plane passing through E. In ΔAEK , $AE = d_1$, $EK = d_2$ and $AK = \sqrt{d_1^2 + d_2^2}$. Further,

$$\begin{aligned} \sin A &= \frac{d_3}{d_1} \quad \text{and also} \quad \sin A = \frac{d_2}{\sqrt{d_1^2 + d_2^2}} \\ \therefore d_3 &= \frac{d_1 d_2}{\sqrt{d_1^2 + d_2^2}} = \frac{d_1 d_2}{\sqrt{2d_2^2 + d_1^2}} = \frac{d_1}{\sqrt{3}} \end{aligned} \quad (26.44)$$

These planes are known as the (111) planes. The interplanar distances are in the ratio

$$d_1 : d_2 : d_3 = 1 : 1/\sqrt{2} : 1/\sqrt{3} \quad (26.45)$$

Using a KCl crystal, Bragg found the values of θ for reflections from the faces ABCD, ADGF and AHF respectively as

$$\theta_1 = 5^\circ 22', \theta_2 = 7^\circ 36' \quad \text{and} \quad \theta_3 = 9^\circ 25'.$$

Hence for the first order spectrum,

$$\begin{aligned} d_1 : d_2 : d_3 &= 1/\sin 5^\circ 22' : 1/\sin 7^\circ 36' : 1/\sin 9^\circ 25' \\ &= 1/0.0936 : 1/0.1323 : 1/0.1636 \\ &= 1/1 : 1/1.410 : 1/1.748 \end{aligned}$$

These ratios, within the limits of experimental error, are

$$d_1 : d_2 : d_3 = 1 : 1/\sqrt{2} : 1/\sqrt{3}$$

Hence it is concluded that KCl crystal has a simple cubic structure. On the other hand, NaCl crystal which is very much similar to KCl is found to have the ratio of interplanar distances as

$$d_1 : d_2 : d_3 = 1 : 1/\sqrt{2} : 2/\sqrt{3}$$

which agree with the theoretical values for a face centred cubic system. The observed dissimilarity between the conclusions on these similar crystals can be resolved once we determine the relative intensities of the diffracted beams. From such a study, it has been confirmed that both the KCl and NaCl crystals have face centred cubic crystal structure.

Example 26.9. A beam of X-rays $\lambda = 0.842 \text{ \AA}$ is incident on a crystal at a glancing angle of $8^\circ 35'$ when the first order Braggs' diffraction occurs. Calculate the glancing angle for 3rd order diffraction.

Solution: According to Braggs' law, $2d \sin \theta = m\lambda$.

$$d = \frac{m\lambda}{2 \sin \theta} = \frac{1 \times 0.842 \text{ \AA}}{2 \times \sin 8^\circ 35'} = 2.826 \text{ \AA}$$

$$\theta = \sin^{-1}\left(\frac{m\lambda}{2d}\right) = \sin^{-1}\left(\frac{3 \times 0.842}{2 \times 2.826}\right) = 26.55^\circ.$$

Example 26.10. X-rays of wavelength 1.5 \AA are incident on Na Cl crystal having a grating spacing of 2.8 \AA . What is the highest order that the crystal can diffract?

Solution: According to Braggs' law, $2d \sin \theta = m\lambda$. m will be maximum if $\sin \theta = 1$

$$\therefore m = \frac{2d \sin \theta}{\lambda} = \frac{2 \times 2.8 \text{ \AA} \times 1}{1.5 \text{ \AA}} = 3.73.$$

It means that the highest order of diffraction is 3.

26.26 POWDER CRYSTAL METHOD

There are many materials for which it is impossible to obtain single crystals of required size. For such materials powder photography is highly suitable. One form of the powder photography is known as Debye-Scherrer method invented by P. Debye and Scherrer.

Fig. 26.53 shows the experimental arrangement. It consists of a cylindrical camera called Debye-Scherrer camera, whose width is smaller as compared to its diameter. The material

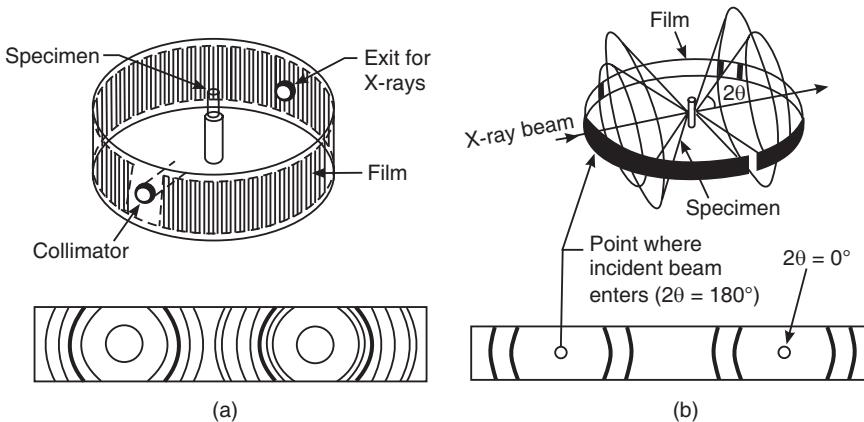


Fig. 26.53

under investigation is crushed into fine grain powder and compressed into a thin rod or packed into a small capillary tube. A strip of photographic film wrapped in opaque paper is mounted round the inside of the cylindrical drum of the camera. The specimen is positioned vertically at the centre of the drum. A narrow beam of monochromatic x-rays enter and leave the drum through the apertures on the opposite sides of the drum. Each crystallite has the same system of atomic planes. Some of the crystallites are bound to lie with their planes at glancing angle θ to the incident ray such that Braggs' condition is satisfied. Each such crystal produces a spot on the photographic plate. As the specimen contains a large number of crystallites oriented in all directions, almost all the possible values of θ and d are available. Also, for a particular

value of the angle of incidence θ , numerous orientations of a particular set of planes are possible. The diffracted rays corresponding to fixed values of θ and d lie on the surface of a cone with its apex at the sample and the semi-vertical angle 2θ . Different cones are observed for different sets of θ and d for a particular value of m and also for different combinations of θ and m for a particular value of d . Each cone of the diffracted beam leaves two impressions on the film in the form of arcs on either side of the exit aperture and their centres coinciding with the aperture.

If x is the distance at which a diffracted beam strikes the film from the centre O, then

$$2\theta = \frac{180^\circ x}{\pi R}$$

where R is the radius of the drum.

$$\therefore \theta = \frac{90^\circ x}{\pi R} \quad (26.46)$$

Let $x_1, x_2, x_3 \dots$ be the distances between symmetrical arcs on the stretched photographic film. Then,

$$\theta_1 = \frac{90^\circ x_1}{\pi R}, \quad \theta_2 = \frac{90^\circ x_2}{\pi R}, \quad \theta_3 = \frac{90^\circ x_3}{\pi R} \quad \text{and so on.}$$

Using the value of θ into Braggs' equation, the interplanar spacing d can be calculated.

The diffraction pattern obtained in this method helps us distinguish amorphous materials from crystalline materials. Amorphous materials do not have reflecting planes. Therefore, diffraction rings are not produced on the film. However, they may produce a smeared ring as there is some sort of short range order in the arrangement of its molecules.

26.27 ROTATING CRYSTAL METHOD

The rotating crystal method is used when single crystals of moderate size are available. The experimental set up of the method is shown in Fig. 26.54 (a). The crystal is mounted on a rotating spindle which is rotated by a rotator. The crystal is mounted in such a way that one of the crystal axes is along the axis of the spindle. A photographic film is mounted inside the drum which is concentric with the rotating spindle.

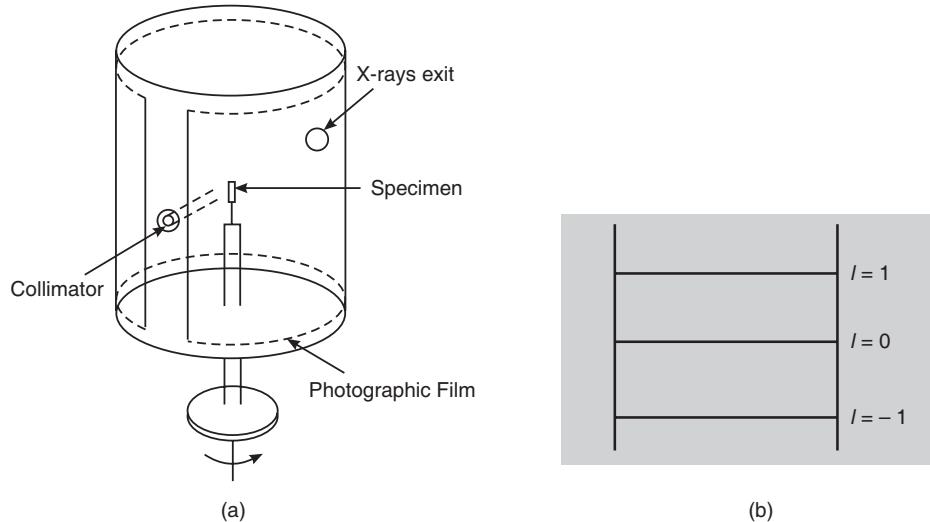


Fig. 26.54

A monochromatic beam of x -rays is allowed through the aperture in the drum to impinge on the crystal at right angles. As the crystal rotates, various planes come successively into positions for which Braggs condition is satisfied. The diffracted beams produce spots on the photographic film. The planes parallel to the axis of rotation diffract the incident rays in a horizontal plane. The planes inclined to the rotation axis produce reflections above or below the horizontal plane depending on the angle of inclination. The horizontal lines produced by diffraction spots on the film are called *layer lines*. If the crystal is placed such that its c -axis coincides with the axis of rotation, all the planes with Miller indices $(h, k, 0)$ will produce the central layer line (Fig. 26.54 b). The planes having Miller indices (h, k, l) and (h, k, \bar{l}) will produce layer lines above and below the central line respectively, and so on. The vertical spacing between the layer lines depends on the distance between the lattice points along the c -axis. Similarly, we can determine the translational vectors a and b on mounting the crystal along a and b axes, respectively. Thus, the dimensions of the unit cell of the crystal are determined.

QUESTIONS

15. Explain simple cubic, body centered and face centered cubic structures. (M.G.Univ.,2005)
16. What is atomic packing factor? Work out atomic packing factors for SC, FCC, and BCC structures. (V.T.U.,2007, 2008)
17. The edge of the unit cell of cubic lattice is 'a'. The radius of the atoms that occupy the lattice site is 'r'. Compute: (i) Number of atoms per unit cell, (ii) atomic radius, (iii) the packing fraction for SC, BCC, and FCC crystal structure. (R.T.M.N.U., 2006)
18. Consider a body centered cubic (BCC) lattice of identical atoms having radius R. Compute (i) the number of atoms per unit cell (ii) the coordination number and (iii) the packing fraction.
19. Define packing factor. Calculate the packing factor for BCC structure. (M.G.Univ.,2006)
20. Explain in detail how atomic radius is determined in different types of cubic lattices. (Univ. of Pune, 2007)
21. What is coordination number? Calculate the coordination number for simple cubic and body central cubic lattices. (M.G.Univ.,2005)
22. Which type of the cubic crystal structure has closest packing of atoms? How many nearest neighbours does an atom in this type of crystal have? Derive the relation between the atomic radius and the unit cell dimension of this crystal.
23. Obtain the relationship between lattice parameter and atomic radius for simple cubic, body centered cubic and face centered cubic lattices. Also obtain the values of atomic packing factor in each of these cases. (R.T.M.N.U., 2005)
24. Show that the FCC structure possesses maximum packing density and minimum percentage of void space among the three crystal structures SC, BCC and FCC.
25. Show that the FCC structure possesses least percentage of void space among SC, BCC and FCC cubic structures.
26. Explain how to find number of atoms per unit cell for simple cubic, B.C.C. and F.C.C. structures in a crystal. Also find the relationship between atomic radius and inter atomic distance in each of these cases.
27. What is meant by atomic packing factor? Calculate atomic packing factor for SC,BCC & FCC structures. (Univ. of Pune, 2007), (M.G.Univ.,2005)
28. Define atomic packing factor. Show that atomic packing factor for BC structure is 0.68 and for FCC structure it is 0.74. (Univ. of Pune, 2007)
29. Show that the SC structure possesses minimum packing density and maximum percentage of void space among the three crystal structures.
30. Verify that simple cubic structure possesses maximum void space amongst all cubic structures.
31. Tabulate the characteristics of SC, BCC and FCC unit cells.
32. Show that FCC structure possesses least percentage void among SC, BCC and FCC cubic structure.
33. For SC, BCC and FCC crystal structures, calculate-
 (a) Number of atoms per unit cell, (b) Coordination number
 (c) Atomic radius. (C.S.V.T.U., 2007)
34. What are Miller indices? Explain with proper example how to determine Miller indices. (G.T.U.,2009)
35. What are Miller indices? Find the Miller indices for a crystal plane. Derive the expression for interplanar distance between consecutive planes described by Miller indices (hkl). (V.T.U.,2008)
36. Explain the term Miller indices. What is their role in crystal structure? Give the important features of Miller indices. (M.G.Univ.,2005)
37. How do you proceed to index a given plane in a cubic crystal – explain. (Andhra Univ.)

38. Explain with an example, the different steps to be followed, to identify the Miller indices of a crystallographic plane. **(R.T.M.N.U., 2007)**
39. Draw separately the principal planes (100), (110), and (111) in a simple cubic crystal.
40. Deduce a relation between an interlinear distance ‘ d ’ and the Miller indices of the planes for cubic crystal. **(R.T.M.N.U., 2007)**
41. Deduce the relation between interplanar distance and Miller indices of the planes for a cubic system. **(R.T.M.N.U., 2010)**
42. Show that the interplanar spacing between (hkl) planes in a cubic lattice is given by

$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}}$$
 where, a is the lattice constant. **(R.T.M.N.U., 2005), (C.S.V.T.U., 2006)**

43. What do you understand by Miller indices of a crystal plane? Show that the spacing between consecutive parallel planes defined by Miller indices (hkl) is given by

$$d_{hkl} = \left[\frac{h^2}{a^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2} \right]^{\frac{1}{2}}$$

where a , b and c are primitive vectors along three mutually perpendicular axes.

(C.S.V.T.U., 2005)

44. What are tetrahedral and octahedral voids? **(R.T.M.N.U., 2006)**
45. Describe crystal structure of NaCl. Derive the expression for lattice constant. **(V.T.U., 2008)**
46. Describe the structure of zinc sulphide.
47. What is meant by polymorphism and allotropy?
48. Discuss the structures of diamond and graphite.
49. Explain with neat sketch the diamond crystal. **(V.T.U., 2007)**
50. Define unit cell. Describe the diamond structure giving details about its coordination number, atomic radius and number of atoms per unit cell. Calculate the packing fraction of diamond structure. **(V.T.U., 2008)**
51. Deduce Braggs’ law of X-ray diffraction in crystals. Describe and explain how Braggs’ spectrometer can be used in the study of crystal structure analysis. **(M.G.Univ., 2005)**
52. Explain crystal state of matter and obtain the expression for Braggs’ law of X-ray diffraction in crystals. **(V.T.U., 2008)**
53. Deduce Braggs’ law of X-ray diffraction. **(Univ. of Pune, 2007)**
54. State and derive Braggs’ law for diffraction in crystals. **(Calicut Univ., 2007)**
55. Derive and explain Bragg’s relation. How is it useful in crystal structure determination? Why is this law not useful for amorphous solids? Why X-rays are preferred to visible light in crystal structure determination?
56. Explain how Braggs’ X-ray spectrometer can be used for verification of Braggs’ relation. **(Univ. of Pune, 2007)**
57. Describe how Braggs’ spectrometer is used for determination of crystal structure. **(V.T.U., 2008)**
58. Explain how Braggs’ X-ray spectrometer can be used to determine the interplanar spacing. **(V.T.U., 2008)**
59. Give the theory of Braggs’ X-ray diffraction. Describe Braggs’ X-ray spectrometer. How is it used to verify Braggs’ law? **(M.G.Univ., 2006)**
60. Explain and deduce Braggs’ law in X-ray diffraction. Describe a Braggs’ spectrometer. Explain how it is used to determine the wavelength of X-rays. **(C.S.V.T.U., 2007)**
61. Explain how the structure of KCl and NaCl is determined using Braggs’ X-ray technique.
62. Describe with suitable diagram the powder method for determination of crystal structure.

63. Describe the rotating crystal method for crystal structure analysis. What is the advantage of this method?
64. Write a note on rotating crystal method of X-ray diffraction.
65. (a) Derive Bragg's law of crystal diffraction
 (b) Describe in detail, powder method to determine the crystal structure. (JNTU, 2010)

PROBLEMS

1. The density of copper is 8980 Kg/m^3 and unit cell dimension is 3.6\AA . Atomic wt. of copper is 63.54. Determine crystal structure.
 Calculate atomic radius and interplanar spacing of (110) plane.
2. The lattice constant for BCC iron at 20° C is 2.87 \AA . The density of iron is 7870 kg/m^3 . Determine its atomic mass. (Avogadro's number = $6.02 \times 10^{26} \text{ atoms/kmol}$). [Ans: 56]
3. Using the following data find the type of unit cell GaAs forms:
 Density of GaAs = 5.324 gm/cm^3 , atomic weight of Ga = 69.7, atomic weight of As = 74.9, lattice constant of GaAs = 5.65 \AA .
4. Draw crystal planes having Miller indices (111), (110) and (211).
5. A certain crystal has axial units $x : y : z$ of $0.424:1:0.367$. Find the Miller indices of crystal faces whose intercepts are $0.212:1:0.183$. [Ans: (212)]
6. Silver has fcc structure and its atomic radius is 1.441 \AA . Find the spacing of (220) planes. [Ans: 1.441 \AA]
7. Calculate the interplanar spacing for a (321) plane in a simple cubic lattice whose lattice constant is 4.2 \AA . [Ans: 1.12 \AA]
8. The lattice constant of BCC iron at 20° C is 2.87\AA . The density of iron is 7870 kg/m^3 . Determine its atomic mass?
9. Nickel is characterized by FCC lattice. The edge of the unit cell is 3.52\AA . The atomic weight of nickel is 56.71 kg/k-mol . Determine the density of the metal.
10. Draw crystal planes having Miller indices (111), (110) and (211).
11. Find the Miller indices of a set of planes with intercepts a , $2a$ and $3a$ on X, Y and Z axes respectively for a cubic crystal.
12. Draw separately the planes (100), (110) and (111) in a simple cubic crystal.
13. What are Miller indices? Draw crystal planes having Miller indices; (2, 1, 0), (1, 0, 1) and (0, 1, 0) for simple cubic structure.
14. Draw separately the principle planes (100), (110) and (111) in a simple cubic crystal.
15. Find the Miller indices of a set of planes with intercepts a , $2a$ and $3a$ on X, Y and Z axes respectively for a cubic crystal.
16. Draw crystal planes having Miller indices (111), (110) and (211).
17. Lead is face centered cubic with an atomic radius $r = 1.746 \text{ \AA}$. Find the spacing of: (200), (220) and (111) planes.
18. Calculate the longest wavelength that can be analyzed by rock salt crystal of spacing 2.82 \AA in the first order.
19. Calculate the glancing angle at which X-rays with wavelength of 0.549 nm are reflected in second order from a crystal with interplanar separation of 0.423 nm . (JNTU, 2010)
20. A beam of X-rays of $\lambda = 0.842 \text{ \AA}$ is incident on a crystal at a glancing angle of 8.35° , where the first order Bragg's reflection occurs. Calculate the glancing angle for second order reflection.

(RTMNU, 2010)

CHAPTER

27

Crystal Defects

27.1 INTRODUCTION

Solids are characterized by regular arrangement of atoms over the entire volume of the material. This is the description of an *ideal* solid. Some of the properties of solids such as specific heat, elasticity, magnetic properties etc can be explained on the basis of the above consideration. However, we cannot explain other properties such as colour of crystals, luminescence etc. In the real solids, on account of thermodynamic considerations, and also on account of conditions of crystal growth, some atoms are not found at places where they should be. The atoms may occupy sites where they are not expected. These are known as *defects* in crystals. Almost all the crystals contain defects, and these defects change the physical properties of crystals to some extent.

27.2 CRYSTAL DEFECTS

The ideal crystal has an infinite 3D repetition of identical units, which may be atoms or molecules. Therefore, it does not contain lattice imperfections. Therefore, in an ideally perfect crystal long range order would be observed over unlimited distances. However, ideal crystals neither occur in nature nor can be produced by artificial methods. The real crystals are limited in size and repeated violations of long range order would be found in them. **A defect or an imperfection is any deviation from the perfect periodic arrangement of atoms observed in real crystals.**

Classification of defects

The defects found in real crystals are classified into four main categories. They are:

1. Point defects (Zero dimensional defects)
2. Line defects or dislocations (One dimensional defects)
3. Planar defects (Two dimensional defects) and
4. Volume defects (Three dimensional defects).

27.3 POINT DEFECTS

A **point defect** is a localized interruption in the regularity of the crystal lattice. It produces strain in a small volume of the crystal surrounding the defect, but does not affect the perfection of distant parts of the crystal. The point defects can be divided into four types, namely (*i*) Vacancies or Schottky defects, (*ii*) Interstitials, or Frenkel defects, and (*iii*) Impurities and (*iv*) Electronic defects.

27.4 VACANCIES

A **vacancy** is the absence of an atom from a site normally occupied in the lattice. Vacancies are produced during solidification as a result of local disturbances or from thermal vibrations of atoms at high temperatures. At any temperature, some atoms which have higher energy than their neighbours will be able to move considerable distances from one equilibrium position to another and ultimately go over to the surface. It appears as if the atoms evaporate from their sites and condense elsewhere.

The site left by the atom is a vacancy in the lattice, as shown in Fig. 27.1. Such vacancies arise in close packed structures, that is, in metallic structures. The number of vacancies per unit volume of the crystal depends upon the temperature.

- A vacancy is an atomic site from which the atom is missing.
- Vacancies form in close packed metallic structures.
- The concentration of vacancies in a crystal depends on the temperature.

27.5 ENERGY OF FORMATION OF VACANCY IN A METALLIC CRYSTAL

Point defects are produced in a crystal due to thermal vibrations. At any given temperature, some of the lattice points normally occupied by metal ions are vacant, giving rise to vacancies. The energy of formation of a vacancy may be defined as the energy difference between the energy needed to remove an atom from the interior of the crystal and take it to infinity, and the energy gained when the atom is placed on the surface of the crystal.

When a vacancy is produced, the internal energy of the crystal increases. At the same time, it increases the entropy of the crystal also. The *free energy* of the system, denoted by G is defined as

$$G = H - TS$$

where H is the *enthalpy* of the system, and S is *entropy* of the system.

At thermal equilibrium at any temperature, the free energy of the crystal is a minimum. That is, $\partial G = 0$.

Let N be the total atomic sites in the crystal. Let n be the number of vacancies at temperature T . The remaining $(N-n)$ sites are occupied by atoms. The vacancies are indistinguishable particles and can be arranged among themselves in $n!$ ways. Similarly, the atoms are indistinguishable and can be arranged among themselves in $(N-n)!$ ways. The number of ways in which we can arrange n vacancies and $(N-n)$ atoms among N sites will be equal to ${}^N C_n$. The probability of a distribution in which n vacancies can be created is given by

$$W = {}^N C_n = \frac{N!}{(N-n)!n!} \quad (27.1)$$

The change in entropy as a result of creation of vacancies, is given by Boltzmann equation as

$$\Delta S = k \log_e W \quad (27.2)$$

where k is the Boltzmann constant.

$$\therefore \Delta S = k \log_e \frac{N!}{(N-n)!n!}$$

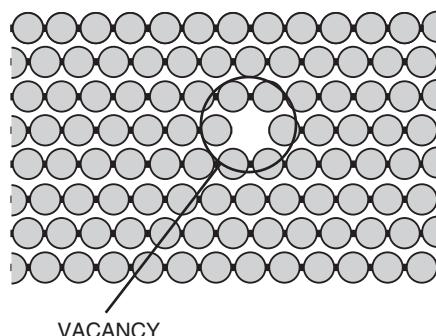


Fig. 27.1

$$= k[\log_e N! - \log_e(N-n)! - \log_e n!] \quad (27.3)$$

Using Stirling's approximation $\log_e x! \approx x \log_e x - x$ for $x \gg 1$ into the above equation, we get

$$\begin{aligned} \Delta S &= k[(N \log_e N - N) - (N-n) \log_e(N-n) + (N-n) - n \log_e n + n] \\ \text{or} \quad \Delta S &= k[N \log_e N - (N-n) \log_e(N-n) - n \log_e n] \end{aligned} \quad (27.4)$$

In formation of vacancies, a change in free energy occurs due to a change in enthalpy and a change in entropy. It is expressed as

$$\Delta G = \Delta H - T\Delta S$$

where ΔH is *enthalpy of vacancy formation*. It is the change in enthalpy resulting from the addition of vacancies and arises from the increase in internal energy caused by breaking interatomic bonds (i.e. removing atoms).

As the process takes place at constant volume,

$$\Delta H = \Delta E,$$

where ΔE is the change in the internal energy of the crystal. Hence,

$$\Delta G = \Delta E - T\Delta S \quad (27.5)$$

If E_v is the average energy required to create a vacancy, then ΔE is the energy change due to creation of n vacancies. Thus,

$$\Delta E = n E_v \quad (27.6)$$

Substituting for ΔE and ΔS into the eq.(27.5), we get

$$\Delta G = n E_v - kT[N \log_e N - (N-n) \log_e(N-n) - n \log_e n] \quad (27.7)$$

The equilibrium state of the crystal at any temperature requires that its free energy is a minimum. As the process takes place at constant volume, the change of free energy occurs due to a change in the number of vacancies.

$$\begin{aligned} \therefore \frac{\partial \Delta G}{\partial n} &= \frac{\partial}{\partial n} \{n E_v - kT[N \log_e N - (N-n) \log_e(N-n) - n \log_e n]\} \\ &= E_v - kT \frac{\partial}{\partial n} [-(N-n) \log_e(N-n) - n \log_e n] \quad (\because N \text{ is a constant}) \\ &= E_v - kT \left\{ - \left[\left(\frac{N-n}{N-n} \right) (-1) + \log_e(N-n)(-1) + \frac{n}{n} + \log_e n \right] \right\} \\ &= E_v - kT \{-(-1 - \log_e(N-n) + 1 + \log_e n)\} \\ &= E_v - kT \log_e \left(\frac{N-n}{n} \right) \end{aligned} \quad (27.8)$$

Therefore, the equilibrium condition is that $\frac{\partial \Delta G}{\partial n} = 0$

$$\begin{aligned} \therefore E_v - kT \log_e \left(\frac{N-n}{n} \right) &= 0 \\ \text{or} \quad E_v &= kT \log_e \left(\frac{N-n}{n} \right) \end{aligned} \quad (27.10)$$

Eq.(27.10) indicates the average energy of formation of a vacancy in a crystal.

27.5.1 Equilibrium Concentration of Vacancies in a Metallic Solid

The energy of formation of a vacancy in a crystal is given by

$$E_v = kT \log_e \left(\frac{N-n}{n} \right)$$

$$\therefore \left(\frac{N-n}{n} \right) = \exp \left(\frac{E_v}{kT} \right)$$

The number of vacancies is much smaller than the total number of atoms. That is, $n \ll N$ and therefore we can approximate that $(N-n) \approx N$.

$$\therefore \left(\frac{N}{n} \right) = \exp \left(\frac{E_v}{kT} \right) \quad \text{or} \quad \left(\frac{n}{N} \right) = \exp \left(\frac{-E_v}{kT} \right)$$

$$\therefore n = N \exp \left(\frac{-E_v}{kT} \right) \quad (27.11)$$

Equ.(27.11) shows that the concentration of vacancies in a metal depends on temperature. It increases rapidly as the temperature of the metal is increased. The formation energy of vacancy is characteristic of the metal; it depends on the nature of the metal.

Example 27.1. In the germanium crystal the equilibrium vacancy concentration decreased by six orders of magnitude when the temperature was reduced from 600 to 300°C. Calculate the energy of formation of the vacancy.

Solution. The equilibrium vacancy concentration is given by $n = N \exp \left(\frac{-E_v}{kT} \right)$.

$$\therefore \text{Equilibrium vacancy concentration at } 873 \text{ K is } n_{873} = N \exp \left(\frac{-E_v}{873k} \right)$$

and the equilibrium vacancy concentration at 573 K is $n_{573} = N \exp \left(\frac{-E_v}{573k} \right)$

$$\therefore \frac{n_{873}}{n_{573}} = \exp \left[-\frac{E_v}{k} \left(\frac{1}{873} - \frac{1}{573} \right) \right] = \exp \left[\frac{E_v}{k} \left(\frac{1}{573} - \frac{1}{873} \right) \right]$$

$$= \exp \left[\frac{E_v \times 6 \times 10^{-4}}{8.625 \times 10^{-5} \text{ eV}} \right]$$

$$\therefore 10^6 = e^{0.696 E_v}$$

$$\ln 10^6 = 0.696 E_v$$

$$\therefore E_v = 1.99 \text{ eV}$$

27.6 SCHOTTKY DEFECT

In case of ionic crystals, the formation of a vacancy requires a local readjustment of charge such that charge neutrality is maintained in the crystal as a whole. Therefore, a pair of vacancies arises due to missing of one cation and one anion from the structure. Such a pair of vacant sites is called a **Schottky defect**.

- Schottky defects are ion vacancies.
- A Schottky defect is the combination of one cation vacancy and one anion vacancy.

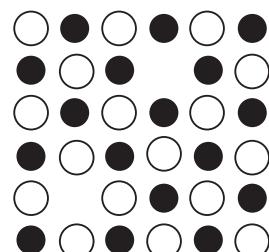


Fig. 27.2. A missing cation (vacancy) and also a missing anion nearby is called a Schottky defect

- This type of point defect forms in ionic crystals.
- The concentration of Schottky defects decreases the density of the crystal.

A Schottky defect is shown in Fig.27.2.

The vacancies disturb the arrangement of atoms around them. The atoms surrounding a vacancy collapse towards the vacancy producing tensile stress.

27.7 INTERSTITIALS

An **interstitial** is an atom on a non-lattice site, as shown in Fig. 27.3.

An interstitial defect arises when an atom occupies a position in the lattice that is not normally occupied in the perfect crystal.

- Interstitial atom produces lattice strain because it tends to push the surrounding atoms further apart.

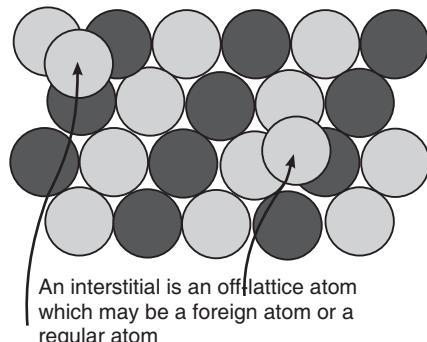


Fig. 27.3

27.8 EQUILIBRIUM CONCENTRATION OF SCHOTTKY DEFECTS IN AN IONIC CRYSTAL

In case of ionic crystals, point defects occur in pairs. A Schottky defect occurs when a cation-anion pair goes missing from the respective lattice positions.

Let us consider that the crystal contains equal number of cations and anions and let N be the total number of ions. Let n be the number of Schottky defects, i.e., n cation vacancies and n anion vacancies in the crystal. The number of different ways in which cation defects can be produced is given by ${}^N C_n$. Similarly, the number of different ways in which anion defects can be produced is given by ${}^N C_n$. The probability of a distribution in which n Schottky defects (defect pairs) can be produced is given by the compound probability, which is the product of the two probabilities. Thus,

$$W = \left[\frac{N!}{(N-n)!n!} \right]^2 \quad (27.12)$$

The change in entropy as a result of creation of vacancies, is given by Boltzmann equation as

$$\Delta S = k \log_e W \quad (27.13)$$

where k is the Boltzmann constant.

$$\begin{aligned} \therefore \Delta S &= 2k \log_e \frac{N!}{(N-n)!n!} \\ &= 2k[\log_e N! - \log_e(N-n)! - \log_e n!] \end{aligned} \quad (27.14)$$

Using Stirling's approximation $\log_e x! \approx x \log_e x - x$ for $x \gg 1$ into the above equation, we get

$$\begin{aligned} \Delta S &= 2k [(N \log_e N - N) - (N-n) \log_e(N-n) + (N-n) - n \log_e n + n] \\ \text{or } \Delta S &= 2k [N \log_e N - (N-n) \log_e(N-n) - n \log_e n] \end{aligned} \quad (27.15)$$

In formation of defects, a change in free energy occurs due to a change in internal energy and a change in entropy. It is expressed as

$$\Delta G = \Delta H - T\Delta S$$

As the process takes place at constant volume,

$$\Delta H = \Delta E,$$

where ΔE is the change in the internal energy of the crystal. Hence,

$$\Delta G = \Delta E - T\Delta S \quad (27.16)$$

If E_S is the average energy required to create a Schottky defect, then ΔE is the energy change due to creation of n defects. Thus,

$$\Delta E = nE_S \quad (27.17)$$

Substituting for ΔE and ΔS into the equ.(27.16), we get

$$\Delta G = nE_S - 2kT[N \log_e N - (N-n) \log_e(N-n) - n \log_e n] \quad (27.18)$$

The equilibrium state of the crystal at any temperature requires that its free energy is a minimum. As the process takes place at constant volume, the change of free energy occurs due to a change in the number of vacancies.

$$\begin{aligned} \therefore \frac{\partial \Delta G}{\partial n} &= \frac{\partial}{\partial n} \{nE_S - 2kT[N \log_e N - (N-n) \log_e(N-n) - n \log_e n]\} \\ &= E_S - 2kT \frac{\partial}{\partial n} [-(N-n) \log_e(N-n) - n \log_e n] \quad (\because N \text{ is a constant}) \\ &= E_S - 2kT \left\{ - \left[\left(\frac{N-n}{N-n} \right) (-1) + \log_e(N-n)(-1) + \frac{n}{n} + \log_e n \right] \right\} \\ &= E_S - 2kT \{-(-1 - \log_e(N-n)) + 1 + \log_e n\} \\ &= E_S - 2kT \log_e \left(\frac{N-n}{n} \right) \end{aligned} \quad (27.19)$$

Therefore, the equilibrium condition is that $\frac{\partial \Delta G}{\partial n} = 0$ (27.20)

$$\therefore E_S - 2kT \log_e \left(\frac{N-n}{n} \right) = 0$$

$$\text{or } E_S = 2kT \log_e \left(\frac{N-n}{n} \right) \quad (27.21)$$

$$\therefore \left(\frac{N-n}{n} \right) = \exp \left(\frac{E_S}{2kT} \right)$$

$$\begin{aligned} \therefore \left(\frac{N}{n} \right) &= \exp \left(\frac{E_S}{2kT} \right) \quad \text{or} \quad \left(\frac{n}{N} \right) = \exp \left(-\frac{E_S}{2kT} \right) \\ \therefore n &= N \exp \left(-\frac{E_S}{2kT} \right) \end{aligned} \quad (27.22)$$

Eq.(27.22) shows that the concentration of defects in an ionic crystal depends on temperature. It increases exponentially as the temperature of the crystal is increased. The average energy required to produce Schottky defects is characteristic of the solid; it depends on the nature of the solid.

Example 27.2. If the average energy required for producing a Schottky defect is 1.97 eV in the ionic crystal NaCl, calculate the density of Schottky defects at 27°C. Given that the interionic distance is 2.82 Å.

Solution. A unit cell of NaCl crystal contains 4 molecules of NaCl.

$$\text{Volume of unit cell} = (2 \times 2.82 \times 10^{-3} \text{ m}) = 1.794 \times 10^{-28} \text{ m}^3.$$

$$\text{Concentration of molecules} = \frac{4}{1.794 \times 10^{-28}} = 2.23 \times 10^{28}.$$

$$E_S = 1.97 \times 1.602 \times 10^{-19} \text{ J}$$

$$\text{The concentration of Schottky defects is given by } n = N \exp\left(-\frac{E_S}{2kT}\right).$$

$$\therefore n = 2.23 \times 10^{28} \times \exp\left[-\frac{1.97 \times 1.602 \times 10^{-19} \text{ J}}{2 \times 1.38 \times 10^{-23} \text{ J/K} \times 300 \text{ K}}\right]$$

$$= 2.23 \times 10^{28} \times e^{-38} = 2.23 \times 10^{28} \times 3.14 \times 10^{-17} = 7 \times 10^{11}.$$

27.9 FRENKEL DEFECT

In ionic crystals, a cation goes into an interstitial position and the vacancy-interstitial pair is called a **Frenkel defect**. Frenkel defect is illustrated in Fig. 27.4.

- In case of ionic crystals, an ion displaced from the lattice site into an interstitial site is called a Frenkel defect.
- As cations are generally smaller in size, they may get displaced into the void space present between in the crystal. Anions are bigger in size and they do not go into void spaces.
- A Frenkel defect is the combination of one cation vacancy and one cation interstitial defect.
- A Frenkel defect does not change the overall electrical neutrality of the crystal.
- The concentration of Frenkel defects does not change the density of the crystal.

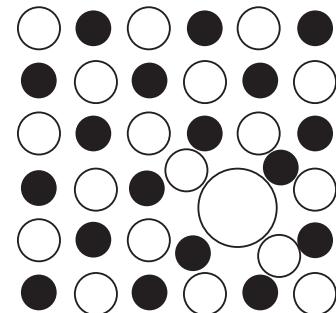


Fig. 27.4. A cation displaced from its structurally correct site (to create a vacancy) becomes interstitial and is called a Frenkel defect

27.10 EQUILIBRIUM CONCENTRATION OF FRENKEL DEFECTS IN AN IONIC CRYSTAL

Frenkel defects are produced when the crystal is heated or a high energy radiation is used to irradiate the crystal. Some of the atoms are knocked out of their normal positions in the lattice and enter into the interstices of the structure. Thus, vacancies and interstitials are simultaneously produced.

Let us consider that the crystal contains equal number of cations and anions and let N be the total number of ions. Let n be the number of Frenkel defects, i.e., n cation or anion vacancies and n interstitial ions in the crystal. The number of different ways in which vacancy defects can be produced is given by ${}^N C_n$. If N_i is the number of interstitial positions, the number of different ways in which interstitial defects can be produced is given by ${}^{N_i} C_n$. The

probability of a distribution in which n Frenkel defects can be produced is given by the compound probability, which is the product of the two probabilities. Thus,

$$W = \frac{N!}{(N-n)!n!} \cdot \frac{N_i!}{(N_i-n)!n!} \quad (27.23)$$

The change in entropy as a result of creation of vacancies, is given by Boltzmann equation as

$$\Delta S = k \log_e W \quad (27.24)$$

where k is the Boltzmann constant.

$$\begin{aligned} \therefore \Delta S &= k \log_e \left[\frac{N!}{(N-n)!n!} \cdot \frac{N_i!}{(N_i-n)!n!} \right] \\ &= k[\log_e N! - \log_e(N-n)! - \log_e n! + \log_e N_i! - \log_e(N_i-n)! - \log_e n!] \end{aligned} \quad (27.25)$$

Using Stirling's approximation $\log_e x! \approx x \log_e x - x$ for $x \gg 1$ into the above equation, we get

$$\Delta S = k \left[(N \log_e N - N) - (N-n) \log_e(N-n) + (N-n) - n \log_e n + n \right. \\ \left. + (N_i \log_e N_i - N_i) - (N_i-n) \log_e(N_i-n) + (N_i-n) - n \log_e n + n \right]$$

$$\text{Or } \Delta S = k[N \log_e N + N_i \log_e N_i - (N-n) \log_e(N-n) - (N_i-n) \log_e(N_i-n) - 2n \log_e n] \quad (27.26)$$

In formation of defects, a change in free energy occurs due to a change in internal energy and a change in entropy. It is expressed as

$$\Delta G = \Delta H - T\Delta S$$

As the process takes place at constant volume,

$$\Delta H = \Delta E,$$

where ΔE is the change in the internal energy of the crystal. Hence,

$$\Delta G = \Delta E - T\Delta S \quad (27.27)$$

If E_f is the average energy required to create a Frenkel defect, then ΔE is the energy change due to creation of n defects. Thus,

$$\Delta E = nE_f \quad (27.28)$$

Substituting for ΔE and ΔS into the equ.(27.27), we get

$$\begin{aligned} \Delta G &= nE_f - kT[N \log_e N + N_i \log_e N_i - (N-n) \log_e(N-n) \\ &\quad - (N_i-n) \log_e(N_i-n) - 2n \log_e n] \end{aligned} \quad (27.29)$$

The equilibrium state of the crystal at any temperature requires that its free energy is a minimum. As the process takes place at constant volume, the change of free energy occurs due to a change in the number of vacancies.

$$\begin{aligned} \therefore \frac{\partial \Delta G}{\partial n} &= \frac{\partial}{\partial n} \{nE_f - kT[N \log_e N + N_i \log_e N_i - (N-n) \log_e(N-n) \\ &\quad - (N_i-n) \log_e(N_i-n) - 2n \log_e n]\} \\ &= E_f - kT \frac{\partial}{\partial n} [-(N-n) \log_e(N-n) - (N_i-n) \log_e(N_i-n) - 2n \log_e n] \\ &\quad (\because N \text{ is a constant}) \end{aligned}$$

$$= E_f - kT \left\{ - \left[\left(\frac{N-n}{N-n} \right) (-1) + \log_e(N-n)(-1) + \left(\frac{N_i-n}{N_i-n} \right) (-1) \right] \right\}$$

$$\begin{aligned}
& + \log_e(N_i - n)(-1) + \frac{2n}{n} + \log_e n \Big] \Big\} \\
& = E_f - kT \{- (-1 - \log_e(N - n) - 1 - \log_e(N_i - n) + 2 + \log_e n)\} \\
& = E_f - kT \log_e \left(\frac{N-n}{n} \right) \left(\frac{N_i-n}{n} \right) \\
& = E_f - kT \log_e \frac{(N-n)(N_i-n)}{n^2} \tag{27.30}
\end{aligned}$$

Therefore, the equilibrium condition is that $\frac{\partial \Delta G}{\partial n} = 0$ (27.31)

$$\begin{aligned}
\therefore E_f - kT \log_e \frac{(N-n)(N_i-n)}{n^2} &= 0 \\
\text{or } E_f &= kT \log_e \frac{(N-n)(N_i-n)}{n^2} \tag{27.32} \\
\therefore \frac{(N-n)(N_i-n)}{n^2} &= \exp \left(\frac{E_f}{kT} \right)
\end{aligned}$$

The number of defects is much smaller than the total number of ion pairs. That is, $n \ll N$ and $n \ll N_i$, therefore we can approximate that $(N-n) \approx N$ and $(N_i-n) \approx N_i$.

$$\begin{aligned}
\therefore \left(\frac{NN_i}{n^2} \right) &= \exp \left(\frac{E_f}{kT} \right) \quad \text{or} \quad \left(\frac{n^2}{NN_i} \right) = \exp \left(-\frac{E_f}{kT} \right) \\
\therefore n^2 &= NN_i \exp \left(-\frac{E_f}{kT} \right) \\
\text{or } n &= \sqrt{NN_i} \exp \left(-\frac{E_f}{2kT} \right) \tag{27.33}
\end{aligned}$$

Equ.(27.33) shows that the concentration of defects in an ionic crystal depends on temperature. It increases exponentially as the temperature of the crystal is increased. The average energy required to produce Frenkel defects is characteristic of the solid and depends on the nature of the solid.

Example 27.3. If the average energy required to produce a Frenkel defect in an ionic crystal is 1.4 eV, find out the ratio of the number of Frenkel defects at 20°C and 300°C in one gram of the crystal.

Solution. The concentration of Frenkel defects is given by $n = \sqrt{NN_i} \exp \left(-\frac{E_f}{2kT} \right)$.

$$\therefore n_{293} = (NN_i)^{1/2} \exp \left(-\frac{1.4}{586k} \right)$$

and $n_{573} = (NN_i)^{1/2} \exp \left(-\frac{1.4}{1146k} \right)$

$$\therefore \frac{n_{293}}{n_{573}} = \exp\left[\left(-\frac{1.4}{k}\right)\left(\frac{1}{586} - \frac{1}{1146}\right)\right] = \exp\left(-\frac{1.4}{8.625 \times 10^{-5}}\right) \times 8.3 \times 10^{-4} = e^{-13.47}$$

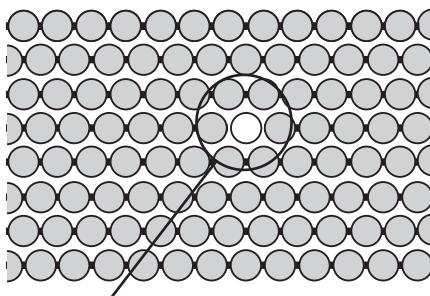
or $\frac{n_{293}}{n_{573}} = 1.4 \times 10^{-6}$

27.11 IMPURITIES

An *impurity* is the substitution of a regular lattice atom with an atom that does not normally occupy that site. The impurities can occupy two different sites in the lattice. An impurity atom may replace an atom of the parent lattice site. Such impurities are known as *substitutional impurity* or *substitutional defect*. Or an impurity may occupy an interstitial position which is a non-atomic site. These impurities are known as *interstitial impurities*.

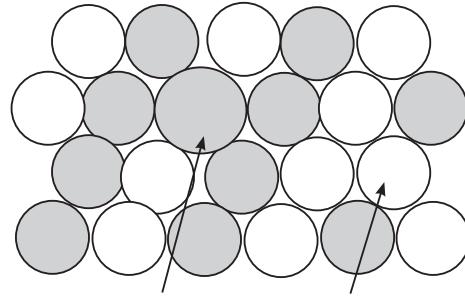
Substitutional impurity

A substitutional impurity arises when a host atom in the lattice is replaced by a foreign atom. The substitutional atom remains in at the regular lattice site. The substitutional defects are illustrated in Fig. 27.5.



Substitutional Impurity atom

(a)



A foreign atom, or a regular atom out of place, is an impurity.

(b)

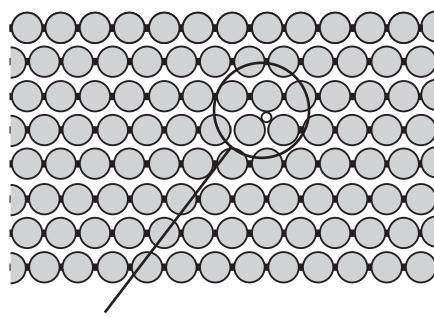
Fig. 27.5

A small substitutional atom causes a tensile stress in the lattice (Fig. 27.5a), while a large substitutional atom produces a compressive stress (Fig. 27.5b).

A controlled addition of impurity to an intrinsic semiconductor crystal is the basis of manufacture of many semiconductor devices. Here, impurity atoms of nearly the same size substitute in place of some of the host semiconductor atoms. During the production of brass alloy, zinc atoms are doped in copper lattice. Such alloys are called substitutional solid solutions.

Interstitial impurity

An interstitial impurity is a small sized atom occupying the void space in the parent crystal, without dislodging any of the host atoms from its site (Fig. 27.6). A foreign atom can occupy an interstitial space when it is substantially smaller than the host atom. Carbon steel is made by introducing carbon atoms in the void spaces formed in iron lattice. Such addition of carbon to iron increases the mechanical strength of iron.



Interstitial Impurity Atom

Fig. 27.6

26.12 ELECTRONIC DEFECTS

Errors in charge distribution in solids are called **electronic defects**. These defects are produced when the composition of an ionic crystal does not correspond to the exact stoichiometric formula. **Stoichiometry** is a state for ionic compounds where there is the exact ratio of cations to anions as predicted by the chemical formula. For example, NaCl is stoichiometric if the ratio of Na^+ ions to Cl^- ions is exactly 1:1. A compound is nonstoichiometric if there is any deviation from the exact ratio.

The nonstoichiometric composition is caused by an excess of metal ions. When a stoichiometric compound ZnO is heated in the presence of zinc vapour, a nonstoichiometric compound Zn_yO is produced. Some of ionized zinc atoms enter the lattice of ZnO and stay around as interstitial cations, as shown in Fig. 27.7. Since the crystal must be electrically neutral, the excess of cations must be balanced by an equal number of electrons. These defects, electrons, are free to move in the crystal under the influence of an electric field and contribute to the electrical conductivity of the material.

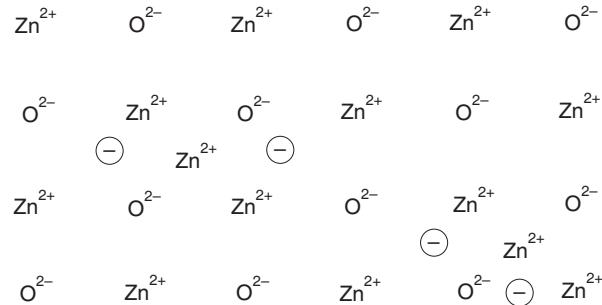


Fig. 27.7

27.13 EFFECT OF POINT DEFECTS

Point defects are sources of distortion and cause residual strains in the lattice. But in some cases, they give rise to beneficial effects by improving mechanical and electrical properties of the materials.

- The addition of carbon atoms in iron increases its tensile strength.
- By changing impurities (carbon, manganese, chromium etc) or concentration of the impurities, various grades of steels are manufactured.
- Addition of copper atoms in gold increases its ductility so that it can be drawn into wires.
- Tin atoms substituted in copper lattice increase the bearing properties of copper.
- Pure germanium and silicon semiconductor materials are doped with pentavalent and trivalent impurities and as a result their electrical conductivity increases considerably.
- Copper atoms in silver increases the electrical resistivity of silver.
- Vacancies help the transport of atoms through the lattice for annealing etc purposes.

27.14 LINE DEFECTS

Line defects are one-dimensional defects and are also called *dislocations*. Dislocations are areas where the atoms are out of position in the crystal structure. A **dislocation** is a one-dimensional defect around which some of the atoms are misaligned. They may be defined as disturbed region between two perfect parts of a crystal. It is a defect in a crystal structure whereby a part-plane of atoms is displaced from its symmetrically stable position in the array. Dislocations appear in crystals due to growth accidents, thermal stresses, phase transformations etc. They can be observed in crystalline materials with the help of electron microscope.

Dislocations are divided into two basic types:

- (i) Edge dislocation, and
- (ii) Screw dislocation.

Most of the dislocations found in crystals are neither pure edge nor pure screw dislocations, but contain components of both these types. They are called **mixed dislocations**.

(i) Edge dislocation

In a perfect crystal, atoms are arranged in both vertical and horizontal planes. The atoms are in perfect equilibrium in their positions and all bond lengths are in equilibrium state. If one of the vertical planes does not extend to the full length, but abruptly terminates in between, within the crystal, then it may be viewed as an extra half-plane inserted between a set of parallel planes (Fig. 27.8). The edge of such a plane constitutes a linear defect. It is a linear defect centered on the line along the end of the extra half-plane of atoms. Within the region around the edge, there is some localized distortion. It may be seen that the atoms above the edge are squeezed and are in a state of compression while the atoms below the edge are pulled apart and are in a state of tension. The region of distortion spreads a few lattice distances around the edge. This distortion is known as **edge dislocation**. The edge is in the direction perpendicular to the plane of the Fig. 27.8 and constitutes the dislocation line.

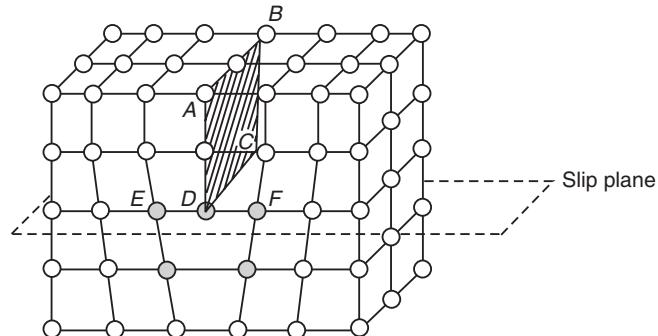


Fig. 27.8

The dislocation is called a line defect because the locus of defective points produced in the lattice by the dislocation lie along a line. This line runs along the bottom of the extra half-plane. The inter-atomic bonds are significantly distorted only in the immediate vicinity of the dislocation line. The plane in which the dislocation line is present is called a **slip plane**. The dislocation line forms a boundary between slipped and unslipped portions.

An edge dislocation is said to be **positive** when the half-plane appears to have been inserted from the top of the crystal, as in Fig. 27.8. It is denoted by \perp . An edge dislocation is said to be **negative** if the extra half-plane lies in the lower part of the crystal. It is denoted by T .

(ii) Screw dislocation

The screw dislocation is also called **Burger's dislocation**. The geometry of screw dislocation is shown in Fig. 27.9. Let us visualize a perfect crystal cut by a plane P. Then let the crystal on the right side of the cut be shifted (sheared) relative to that on the left side by an amount \mathbf{b} . The resulting arrangement is shown in Fig. 27.9. From this figure it is seen that the shift has occurred for the length of one-atomic spacing in the predominant part of the crystal and reached the line OO'. Away from this line, the crystal structure is undisturbed and shows an ordered atomic arrangement. Near the line OO', atoms of atomic planes lying to the left and right of the shear plane P have undergone displacement with

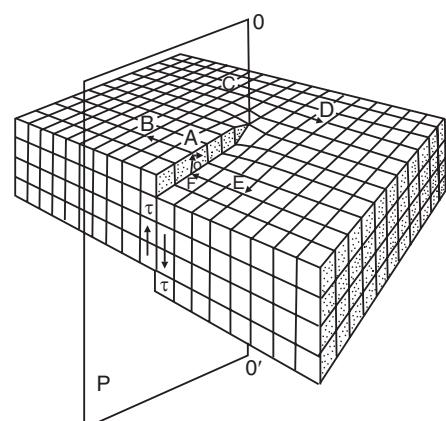


Fig. 27.9

respect to each other. This displacement causes a local distortion called a **screw dislocation**. The line OO' is the *line of dislocation*. It may be viewed now that the atoms in the dislocation region lie on a surface which spirals from one end of the crystal to the other with the line of dislocation as the axis of the spiral. The displacement of the atoms from their original positions in the perfect crystal is described by the following equation.

$$r = \frac{b}{2\pi} \theta \quad (27.34)$$

where r is the displacement along the dislocation line and the angle θ is measured from a line perpendicular to the dislocation line.

27.15 BURGERS VECTOR

Dislocations are quantitatively described by the Burgers vector, and is denoted by \mathbf{b} . It tells us the direction and the magnitude of the lattice distortion associated with a dislocation in a crystal. It describes the slip which one part of the crystal undergoes relative to the rest of the crystal. The nature of the dislocation is defined by the relative orientations of dislocation line and Burgers vector. The Burger vector of an edge dislocation is perpendicular to the dislocation line and that of a screw dislocation is parallel to the dislocation line. In case of dislocations of mixed nature, the vector \mathbf{b} will be neither perpendicular nor parallel to the dislocation line.

The magnitude of Burgers vector is found by drawing a closed circuit, called the *Burgers circuit* surrounding the dislocation line. First, let us consider the perfect crystal as given in Fig. 27.10 (a). In this perfect crystal structure, a rectangle whose lengths and widths are integer multiples of lattice vector is drawn encompassing a centre. The number of lattice vectors travelled along each side of the rectangle is noted. As an example, starting from the point P, let us go to right by n steps ($n = 4$), then take m steps ($m = 5$) down, then n steps to the left and finally m steps up. We reach the point where we started, having described a close path. This movement around a centre is called a **Burgers circuit**. In a perfect crystal we reach the starting point after completion of the circuit.

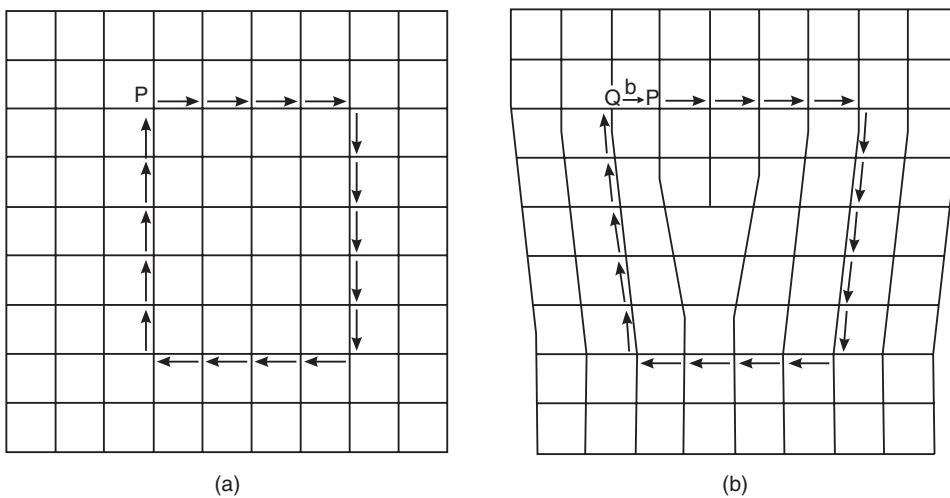


Fig. 27.10

Edge dislocation

Now we consider the Burgers circuit in a crystal that contains edge location (Fig. 27.10 b). Burgers circuit in this case is drawn by proceeding through the undistorted

region surrounding the dislocation, moving the same number of lattice vectors along each direction as in Fig. 27.10 (a).

Starting from P, if we trace the Burgers circuit in a plane normal to the dislocation line, the circuit would not be completed since the end point Q and the starting point P do not coincide. From Q, we need an extra step "b" to arrive at the starting point P. This extra step $\mathbf{QP} = \mathbf{b}$ is called *Burgers vector* (BV). The Burgers vector gives both the magnitude and direction of the displacement.

$$BV = \mathbf{QP} = \mathbf{b}$$

It is seen from Fig. 27.10 (b) that the Burgers vector of an edge dislocation is perpendicular to the dislocation.

Screw dislocation

Now we consider the Burgers circuit in a crystal that contains screw dislocation. In the perfect crystal, starting from P, if we trace the Burgers circuit, the circuit is a closed path PMNOP (Fig. 27.11 a). In case of a crystal with a screw dislocation, the circuit would not be completed and requires an extra step $\mathbf{b} = \mathbf{FA}$ (Fig. 27.11 b), parallel to the dislocation axis, to close the circuit (also see Fig. 27.9). This additional vector \mathbf{b} is *Burgers vector*. The Burgers vector gives both the magnitude and direction of the displacement.

The Burgers vector of a screw dislocation is parallel to the dislocation.

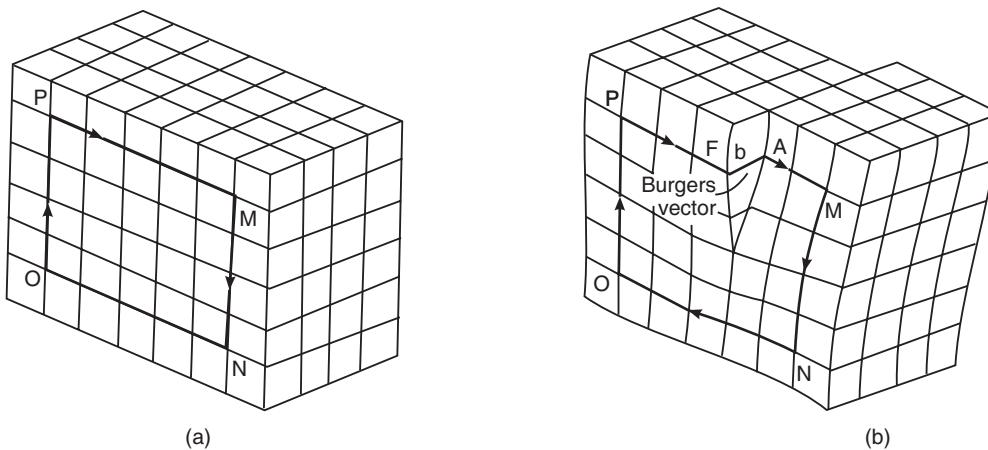


Fig. 27.11

27.16 PLANAR DEFECTS OR SURFACE DEFECTS

Planar defects are two dimensional defects that separate two regions of the crystal. These defects are: the external surfaces, grain boundaries, twin boundaries, and stacking faults.

External surfaces

The surface of a crystal itself is a defect. The surface atoms have formed bonds with neighbours only on one side, whereas the atoms inside the crystal formed bonds with neighbours on either side of them. Since the surface atoms are not bonded to the maximum number of nearest neighbours, they are in a higher energy state than the atoms in the interior of the crystal. This gives rise to a surface energy. The crystal tends to minimize the total surface area in order to reduce the surface energy, but it is not possible because solids are mechanically rigid.

Grain boundaries

The most important defect in crystalline samples is the grain boundary. A **grain Boundary** is a general planar defect that separates regions of different crystalline orientation

(i.e. *grains*) within a polycrystalline solid. Solids are formed when the melt of materials are cooled. Depending upon the rate of cooling, either single crystals or polycrystalline solids form. During solidification, infinitesimally small crystallites are formed first at some places in the melt. They are called nuclei. As the solidification proceeds, these nuclei grow in size. They are randomly oriented. Crystallites with particular orientation tend to join those with the same orientation. All the crystallites that have the same orientation form a bigger block called grain. When the solidification is completed, the solid consists of a number of grains attached to each other. The result is a polycrystalline sample. In between the grains, there exists a boundary called grain boundary. The boundary region is of a few atomic diameters wide. There occurs some atomic mismatch in the region of transition from one grain to its adjacent grain. The atoms are bonded less regularly along a grain boundary and there is grain boundary energy. The magnitude of this energy is a function of the degree of misalignment. Grain boundaries are therefore more chemically reactive than the grains themselves.

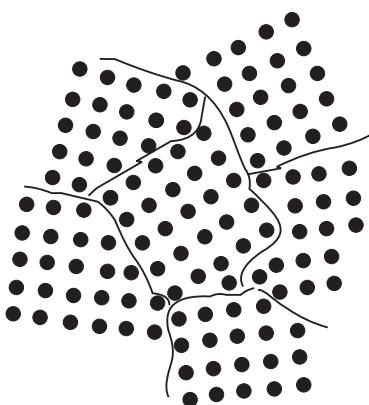


Fig. 27.12

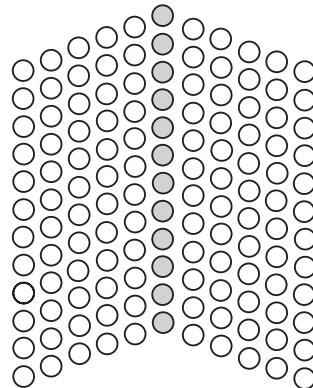


Fig. 27.13

Twin boundaries

A twin boundary is a special type of grain boundary. Atoms on one side of a twin boundary are located in mirror-image positions of the atoms on the other side. The boundary between the twinned crystals will be a single plane of atoms. The crystals on either side of the plane are mirror images of each other. There is no region of disorder and the boundary atoms can be viewed as belonging to the crystal structures of both twins.

Stacking faults

An FCC (face centered cubic) crystal is regarded as formed by the stacking of hexagonal close packed atomic layers, A,B and C in the following sequence:

ABC ABC ABC

If one of the layers go missing, then stacking fault occurs in the crystal. For example, if the A layer in the third set of sequence fails to appear, the sequence changes to ABC ABC BCA , then the sequence is interrupted and a stacking fault is produced. The stacking faults are usually produced during the growth of the crystals. Stacking faults are observed in FCC metals.

27.17 VOLUME DEFECTS

Volume defects such as cracks may arise in crystals during the process of crystal growth. A crack may result during the growth due to a possible small electrostatic dissimilarity between

the stacking layers. A large vacancy may arise due to missing of clusters of atoms which is a volume defect. Inclusion of foreign particles or non-crystalline regions of the size of 20 Å also belong to the category of volume defects.

QUESTIONS

1. What are crystal imperfections and how they affect properties of crystals? Explain point defect. **(Univ. of Pune, 2007)**
2. What are crystal defects? Mention the different kinds of crystal imperfections.
3. What is crystal defect? How does it arise? Explain effect of point defect and line defect on properties of crystals. **(Univ. of Pune, 2008)**
4. Explain the various point defects in a crystal.
5. How are vacancies created in a lattice?
6. Derive the expression for the energy required to produce a vacancy in a crystal.
7. What are point defects in crystals? Obtain an expression for the equilibrium concentration of vacancies in a metal at a given temperature.
8. Derive an expression for the concentration of Schottky defects in a crystal.
9. Distinguish between Schottky and Frenkel defects in ionic crystals. **(Univ. of Pune, 2007)**
10. Discuss the Schottky defect in ionic crystals.
11. Obtain an expression for the equilibrium concentration of Frenkel defects at a given temperature in an ionic crystal.
12. Explain interstitial defects.
13. (a) Explain edge and screw dislocations with neat diagrams.
(b) What is Burger circuit? Draw Burger's circuit for an edge dislocation and screw dislocation. What is the significance of the Burger's vector? **(Andhra Univ.)**
14. What are grain boundaries? Explain. **(Andhra Univ.)**
15. Write notes on: Stacking faults and twin boundary.

PROBLEMS

1. If the average energy required to produce a vacancy in a metal is 1eV, calculate the ratio of vacancies in a metal at 1000K and 500K. **[Ans: 1.08×10^5]**
2. The fraction of vacancy sites in a metal is 1×10^{-10} at 500°C. What will be the fraction of vacancy sites at 1000°C? **[Ans: 8.4×10^{-7}]**
3. Calculate the equilibrium number of vacancies per unit volume at a temperature of 1000°C. The energy of formation of vacancy in copper is 0.90 eV. What is the vacancy fraction at 500°C? Given that density of copper metal is 8960 kg/m³ and the atomic weight is 63.5. **[Ans: 1.6×10^{-6}]**
4. Calculate the equilibrium number of vacancies at 300K in aluminium to that produced by rapid quenching at 800K. Enthalpy of formation of vacancies in aluminium is 68 kJ/mol. **[Ans: 3.75×10^{-8}]**
5. The density of Schottky defects in a certain sample of NaCl is $5 \times 10^{11}/\text{m}^3$ at 25°C. If the interionic distance is 2.82 Å, what is the energy required to create one Schottky defect? **[Ans: 1.97 eV]**
6. The concentration of Schottky defects in an ionic crystal is 1 in 10^{10} at a temperature of 300K. Estimate the average separation in terms of the lattice spacings between the defects at 300K and calculate the value of concentration to be expected at 1000K.

CHAPTER

28

Conductors

28.1 INTRODUCTION

Materials having low electrical resistivity are known as conductors. Metals and their alloys belong to this group of materials. In metals the valence electrons are loosely bound to their individual atoms. They become free and are responsible for the conduction of electricity and heat in metals. Free electron theory was proposed by Drude by assuming that the valence electrons become free in metals and move about randomly within the metal much the same way as molecules in a gas confined to a container. H.A. Lorentz later refined Drude's theory by assuming that the velocity distribution of the electrons obeyed the classical Maxwell-Boltzmann law. The Drude-Lorentz theory could successfully explain the Ohm's law and the high electrical conductivity of metals. However, the theory suffered from a number of setbacks, because of the various assumptions made. Sommerfeld later extended the free electron model by incorporating Fermi-Dirac statistics. It too could not succeed to explain all the experimental observations found on solids. Ultimately, the band theory of solids successfully explained the electrical behaviour of solids.

28.2 ELECTRICAL CONDUCTION

If a potential difference V is applied across a solid, as shown in Fig. 28.1, it establishes an electric field E in the solid. E is given by

$$E = \frac{V}{L} \quad (28.1)$$

where L is the length of the solid along which charge carriers move. The electric field accelerates the charge carriers and causes a flow of electric current through the solid. The current I passing across an area A is defined as the net charge Q transported through the area per unit time. Thus,

$$I = \frac{Q}{t} \quad (28.2)$$

When electric current flows through a material, it is said to be **conducting** electricity. Any material can conduct electricity if it contains mobile charge carriers. A free electron is an example of charge carrier. Other carriers include mobile positive or negative ions, holes etc.

The magnitude of the electrical current, I , passing through a solid (Fig. 28.1) at a constant temperature is directly proportional to the potential difference V applied across the solid.

$$I = \frac{V}{R} \quad \text{Ohm's law} \quad (28.3)$$

where R is the electrical **resistance** offered by the solid to the current flow.

When electrons travel in a vacuum, they are not impeded in their motion. They travel along straight line paths and acquire kinetic energy which is equal to the work done by the accelerating electric field. On the contrary, electrons encounter opposition while moving through a solid. The opposition to the electron motion in materials is manifest as electrical resistance. The electrical resistance offered by a solid is found to depend on the physical dimensions of the solid.

$$R \propto \frac{L}{A}$$

where L is the length and A the area of cross-section of the solid. Thus,

$$R = \rho \frac{L}{A} \quad (28.4)$$

ρ is called **electrical resistivity**. It is a material constant and does not depend on the dimensions of the solid.

$$\rho = \frac{RA}{L} \quad \text{Ohm.m} \quad (28.5)$$

The reciprocal of the electrical resistivity is known as **electrical conductivity**, σ . Thus,

$$\sigma = \frac{1}{\rho} = \frac{L}{RA}$$

Using the relation (28.3) into the above equation, we get

$$\sigma = \frac{IL}{VA} \quad \text{S/m} \quad (28.6)$$

Electrical conductivity, σ , characterizes the ability of a material to conduct electricity.

28.3 CLASSIFICATION OF MATERIALS

The resistivity of a solid can be determined using the circuit shown in Fig. 28.1. The electrical resistivity and conductivity of some materials are given in Table.1.

Table 1: Resistivities and conductivities of some solids

Material	Resistivity Ohm.m	Conductivity S/m
Silver	1.47×10^{-8}	68×10^6
Copper	1.78×10^{-8}	58×10^6
Aluminium	2.63×10^{-8}	38×10^6
Steel	20.00×10^{-8}	5×10^6
Lead	22.00×10^{-8}	4×10^6
Carbon	3500×10^{-8}	0.03×10^6
Germanium	6.00×10^{-1}	1.67

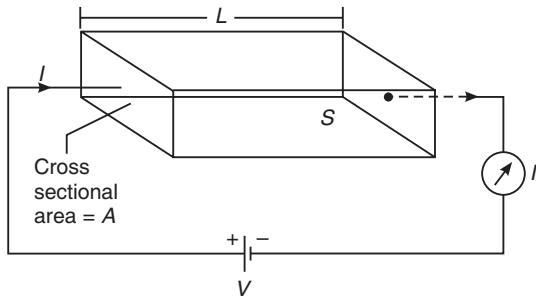


Fig. 28.1. Schematic of the circuit for determining the conductivity of a solid

Silicon	2300.00	4.35×10^{-4}
Aluminium glass	10^{10} to 10^{12}	10^{-10} to 10^{-12}
Borosilicate glass	10^{13}	10^{-13}
Polyethylene	10^{13} to 10^{15}	10^{-13} to 10^{-15}

It is seen from Table-1 that materials differ greatly in their conductivity. However, they may be broadly classified into three groups.

- Metals and alloys exhibit large conductivity of the order of 10^8 S/m and are therefore, called **conductors**.
- Materials such as metal oxides, glasses, plastics are found to possess very low conductivity of the order less than 10^{-12} S/m. They are called **insulators**.
- Materials such as silicon and germanium have values of conductivity, of the order of 10^4 to 10^{-4} S/m, intermediate to those of conductors and insulators. They are hence called **semiconductors**.

28.4 FREE ELECTRON MODEL OF SOLIDS

Electrical conduction is one of the important properties of solids. It was generally accepted that the valence electrons are involved in electrical conduction in metals and alloys. The evidence was provided at a later stage by the experiments of Stewart and Tolman. Much earlier to the experimental confirmation, the free electron model of solids was proposed by Paul Drude. He assumed that the valence electrons become *free* in solids and move about randomly within the solids much the same way as molecules in a gas confined to a container. This is the **free electron model** which is applicable to all the three categories of solids. This model is used in explaining not only the electrical properties but also the thermal, optical and magnetic properties of solids.

The free electron theory underwent successive modifications in an attempt to explain the electrical behaviour and the distinction between the three types of solids.

1. **Classical free electron theory:** This theory was proposed by Paul Drude in 1900 and later was extended by Lorentz. Therefore this theory is also known as the **Drude-Lorentz theory**. In this theory it was assumed that valence electrons become free in metals and move about randomly within the metal. In this theory it was assumed that the free electrons move in a region of constant potential. Just as the velocities of molecules in a container, the velocities of electrons in a solid obey the classical Maxwell-Boltzmann distribution. This theory successfully explained the Ohm's law and the high electrical conductivity of metals. However, the theory failed to explain other features and the distinction between conductors, insulators and semiconductors.
2. **Quantum free electron theory:** This theory was developed by Sommerfeld in 1928. Unlike neutral molecules, electrons are charged particles and obey Pauli Exclusion Principle. An assembly of free electrons obeys Fermi-Dirac statistics. Sommerfeld took into account these facts and modified the Drude's classical free electron theory. However, in this theory also it was assumed that the free electrons move in a region of constant potential. This theory is based on the particle character of electron and did not take into account of its wave character. The theory failed to explain other features and the distinction between conductors, insulators and semiconductors.

3. **Band theory of solids:** This theory was formulated by Felix Bloch in 1928. This theory takes into account that electrons exhibit wave character as they move between atoms in a solid. It further assumed that the potential varies in a periodic manner in the solid. This theory successfully explained the classification of solids into three groups, namely conductors, insulators and semiconductors.

28.5 CLASSICAL FREE ELECTRON THEORY OF METALS

Free electron gas

After the discovery of electron by J.J.Thomson in 1897, Drude developed the free electron theory of metals. According to this theory, metals consist of positive ion cores and valence electrons. The ions cores are immobile and consist of positive nucleus and the bound electrons. The valence electrons get detached from the parent atoms during the process of formation of the metal and move randomly among these cores. Hence they are known as **free electrons**. The potential field of the ion cores is assumed to be constant throughout the metal and the mutual repulsion among the electrons is neglected. The behaviour of free electrons moving within the metal is considered to be similar to that of atoms in perfect gas. These free electrons are therefore referred to as **free electron gas**. As the potential energy of a stationary electron inside the metal is less than the potential energy of an identical electron just outside it, the movement of free electrons is restricted to the boundaries of the metal. This energy difference serves as a potential barrier and stops the free electrons from leaving the surface of the metal. Thus, the free electron gas is confined to a potential energy box. The free electrons are called **conduction electrons** as they are responsible for conduction of electricity in the metals.

Thermal motion of free electrons

The free electrons keep moving randomly in all directions through the lattice structure of the metal due to thermal energy. The average speed is very high and is of the order of 10^6 m/s. As free electrons in a metal move through a maze of positive ion cores, they suffer repeated collisions with the ion cores or other electrons and get deflected. The direction of motion of each free electron changes on every collision and the electron moves along a zigzag path. On the average for every electron moving in a particular direction, there is another electron moving in the opposite direction. Therefore, the thermal motion of free electrons does not cause flow of current through the metal. The electron motion between two collisions is linear and uniform.

Drift motion of free electrons

When the ends of a piece of metal are connected to the terminals of a battery, an electric field is impressed across the metal and the equilibrium condition is disturbed. The electric field accelerates the electrons. The electrons acquire velocity and move in a direction opposite to that of electric field. The directional motion of electrons due to the action of electric field is called **drift**. The drift velocity gained by an electron due to acceleration is lost completely whenever a collision occurs. After that, the electron gets accelerated once again and loses its velocity at the next collision. The process goes on repeating and the electron moves on an average with a **mean drift velocity** v_d . The magnitude of the drift velocity is limited by the decelerations caused by collisions. Thus, the collisions play the role of a frictional force.

The drift speed is typically of the order of 10^{-2} m/s, which is very small compared to the thermal speed, 10^6 m/s. Thus, the motion of a free electron in the presence of electric field consists of a much slower directional drift motion superposed over random zigzag motion.

The drift motion is directional and causes current flow in a conductor called **drift current** or **conduction current**.

Mean Collision time, τ

An electron moving in a particular direction inside a metal encounters another electron or fixed positive ion and deviates from its direction through some angle with the initial direction. This is the process of **scattering** of electrons. For simplicity we say that electron underwent a collision during its motion. After some time another collision occurs and the electron deviates in another direction and this goes on. The time elapsed between two successive collisions is not a constant but varies. The number of collisions per second that an electron makes with the ion cores is proportional to its speed. The average duration of time that elapses between two successive collisions is called **mean collision time**, τ of the electron. It means that the electron on the average travels for a time τ before its next collision, and has traveled for a time τ since its last collision. The collision time is given by

$$\tau = \frac{\lambda}{\bar{v}} \quad (28.7)$$

where λ is known as mean free path and \bar{v} the *rms* (root mean square) velocity.

Mean free path, λ

The average distance traveled by an electron between any two consecutive collisions is known as **mean free path**, λ . The mean free path is given by the product of rms velocity of electrons and mean collision time.

$$\therefore \lambda = \bar{v}\tau \quad (28.8)$$

28.6 DRIFT VELOCITY

When an electric field E is impressed on a metal the electrons in it experience a force $-eE$.

If v is the velocity of free electron and τ is the average time between two consecutive collisions, the frictional force opposing the continuous acceleration of the electron may be written as $-mv/\tau$. Using Newton's second law, the equation of motion of free electron may be written as

$$m \frac{dv}{dt} = -eE - m \frac{v}{\tau} \quad (28.9)$$

where m is the mass of the electron and E is the intensity of the applied electric field.

Under the steady state condition, $\frac{dv}{dt} = 0$, and the electron attains a steady value of velocity v_d in dynamic equilibrium. Therefore, under steady state condition, equ.(28.9) reduces to

$$0 = -eE - m \frac{v_d}{\tau}$$

$$\therefore v_d = -\frac{eE\tau}{m} \quad (28.10)$$

where v_d is the steady state velocity of the electron and is known as **drift velocity**. The drift motion of electrons causes current flow in a conductor.

28.7 ELECTRICAL CONDUCTIVITY

Let n be the number of free electrons per unit volume of the conductor, S (Fig. 28.1). n is called the **free electron density** or **free electron concentration** in the solid.

The total number of electrons in the metal specimen is given by

$$N = (\text{electrons per unit Volume}) (\text{Total volume})$$

\therefore

$$N = nAL \quad (28.11)$$

The total charge present in the solid block may be written as

$$Q = Ne = nAle \quad (28.12)$$

Current flowing through the solid is given by

$$I = \frac{Q}{t} = \frac{nALE}{t} \quad (28.13)$$

The term L/t represents velocity and gives the average *drift velocity*, v_d of electrons in the solid.

$$I = neAv_d \quad (28.14)$$

$$\text{The current density is defined as } J = \frac{I}{A}. \quad (28.15)$$

\therefore

$$J = nev_d \quad (28.16)$$

$$\text{Also, } J = ne\left(\frac{eE\tau}{m}\right) = \frac{ne^2\tau}{m}E \quad (28.17)$$

$$\text{According to equ. (28.6), } \sigma = \frac{IL}{VA}$$

$$\text{As } \frac{V}{L} = E, \text{ the above equation may be rewritten as } \sigma = \frac{J}{E}$$

$$\text{or } J = \sigma E \quad \text{Point form of Ohm's law} \quad (28.18)$$

Equating R.H.S. of equ. (28.17) and (28.18), we obtain

$$\sigma = \frac{ne^2\tau}{m} \quad (28.19)$$

Similarly, equating R.H.S. of equ. (28.16) and (28.18), we obtain

$$\sigma = \frac{nev_d}{E} \quad (28.20)$$

or

$$\sigma = ne\mu \quad (28.21)$$

where μ is called electron **mobility**.

The equation (28.19) or (28.21) indicates that electrical conductivity of a material depends mainly on the free electron concentration in it. The large electrical conductivity of metals is explained as due to the presence of large number of free electrons in them. Similarly, one may explain the very low conductivity of insulators as due to non-availability of free electrons in them and the moderate conductivity in semiconductors as due to the presence of modest number of free electrons in them. However, the equation cannot explain the variation in conductivity due to external influences such as the temperature, radiation and impurities in case of semiconductors and insulators.

Example 28.1. Find the drift velocity of free electrons in a copper wire of cross-sectional area 10 mm^2 when the wire carries a current of 100 A . Assume that each copper atom contributes one electron to the free electron gas. Density of copper is 8969 kg/m^3 and its atomic weight is 63.54 .

Solution: The electron concentration in copper, $n = \text{Atomic density, } N \times \text{contribution from each atom, } x$

$$n = \frac{\text{Density} \times N_A}{M} \times x = \frac{8969 \text{ kg/m}^3 \times 6.02 \times 10^{26} / \text{k.mol}}{63.54 \text{ kg/kmol}} \times 1 = 8.49 \times 10^{28} / \text{m}^3.$$

Drift velocity of free electrons in a metal is given by

$$v_d = \frac{I}{nAe} = \frac{100 \text{ A}}{8.49 \times 10^{28} / \text{m}^3 \times (10^{-5} \text{ m}^2) (1.602 \times 10^{-19} \text{ C})} = 7.4 \times 10^{-4} \text{ m/s.}$$

28.8 MOBILITY

When an electric field is applied across a solid, it accelerates the electrons in the direction of electric field. As electrons moving through a solid undergo repeated collisions with the atoms in the solid and therefore, move with a steady velocity known as drift velocity, v_d . The drift velocity is proportional to the electric field applied. Thus,

$$v_d \propto E$$

or

$$v_d = \mu E \quad (28.22)$$

where μ is the proportionality constant and is called electron **mobility**.

We define *electron mobility* as the drift velocity of electrons per unit electric field.

$$\therefore \mu = \frac{v_d}{E} \quad (28.23)$$

Mobility indicates the ease with which electrons move in a solid. We find that electron mobility in metals is of the order of $10^{-3} \text{ m}^2/\text{V.s}$ and in semiconductors it is of the order of $10^{-1} \text{ m}^2/\text{V.s}$.

Example 28.2. Find the mobility of electrons in copper if there are 9×10^{28} valence electrons/ m^3 and the conductivity of copper is $6 \times 10^7 \text{ mho/m}$.

Solution. Electrical conductivity is given by $\sigma = n e \mu$.

$$\therefore \mu = \frac{\sigma}{ne} = \frac{6 \times 10^7 \Omega^{-1}/\text{m}}{9 \times 10^{28} / \text{m}^3 \times 1.602 \times 10^{-19} \text{ C}} = 4.16 \times 10^{-3} \text{ m}^2/\text{V.s}$$

28.9 RELAXATION TIME

When electric field is applied to a metal, the free electrons in it move with an average velocity v_d in a direction opposite to that of the applied field. If the external field is switched off, the velocity of the free electrons decays exponentially as a result of collisions with positive ions. The decay is given by the expression

$$v_d(t) = v_d(0)e^{-t/\tau}$$

where v_0 is the initial velocity of the electrons before application of electric field.

$$\text{If } t = \tau, \quad v_d(t) = v_d(0)e^{-\tau/\tau} = \frac{v_d(0)}{e} \quad (28.24)$$

The duration of time in which the drift velocity of an electron decays to $\frac{1}{e}$ times of its initial velocity is known as **relaxation time**. It gives the time taken by electrons in a conductor to return from non-equilibrium condition to the equilibrium condition, after the applied electric field is turned off. Its value is of the order of 10^{-14}s .

Example 28.3. Find the relaxation time of conduction electrons in a metal if its resistivity is $1.54 \times 10^{-8} \Omega\text{m}$ and it has 5.8×10^{28} conduction electrons/ m^3 .

Solution.

$$\sigma = \frac{ne^2\tau}{m}$$

$$\therefore \tau = \frac{m}{\rho ne^2} = \frac{9.11 \times 10^{-31} \text{ kg}}{1.54 \times 10^{-8} \Omega \text{ m} \times 5.8 \times 10^{28} / \text{m}^3 \times (1.602 \times 10^{-19} \text{ C})^2} = 3.9 \times 10^{-14} \text{ s.}$$

28.10 THERMAL CONDUCTIVITY

Heat conduction is the transfer of thermal energy from the hotter to the colder part of a material. All solids conduct heat. The rate at which heat is conducted depends on the temperature gradient and the material of the solid. Let us consider a metal bar of length x whose ends are held at different temperatures. Heat that flows through a cross-section of the bar divided by time and area is proportional to the temperature gradient $\frac{dT}{dx}$. Thus,

$$\frac{\Delta Q}{A \cdot \Delta t} \propto \frac{dT}{dx}$$

where ΔQ is the thermal energy conducted through a cross-sectional area A in time Δt between two planes with a temperature gradient of $\frac{dT}{dx}$. The above relation may be written in equation form as

$$J = -K \frac{dT}{dx} \quad (28.25)$$

where J is the thermal flux and K is the proportionality constant, called the **thermal conductivity**. The negative sign in the above equation indicates that the heat flows from the hot end to the cold end of the bar. Using the above expressions, we can define the thermal conductivity K of a solid. It is defined as *the quantity of heat crossing per unit time through a unit area and maintaining a unit temperature difference across the body*. Thus,

$$K = Q \left[\frac{dT}{dx} \right]^{-1} \quad (28.26)$$

28.10.1 Expression for Thermal Conductivity

Since good electrical conductors are also good conductors of thermal energy, it is assumed that the highly mobile free electron gas is responsible for transporting thermal energy through the metal via random collision processes. Treating free electron gas as a perfect gas, we obtain an expression for thermal conductivity making use of the kinetic theory of gases.

Let us now consider a uniform metal bar, which is heated at one of its ends (Fig. 28.2). Let the left side of the metal bar be hot and the right side be cold. Thus, a temperature gradient $\frac{dT}{dx}$ exists in the x -direction. Let us select a

volume at the center of the bar whose faces have the size of a unit area and whose length is 2λ , where λ is the mean free path of electrons. Let us assume that at the distance λ from the centre x_0 the average electron had its last

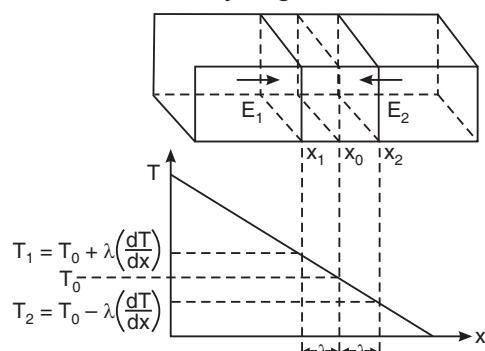


Fig. 28.2

collision and gained energy of that place. Now we calculate the energy E_1 carried by electrons that drift from the left into the selected volume.

$$E_1 = \text{Number of electrons} \times \text{average energy of an electron} = z \times \frac{3}{2} kT_1.$$

Let n be the electron concentration in the material and v be the velocity of electrons. From the kinetic theory of gases, z is given by

$$\begin{aligned} z &= \frac{1}{6} n v \\ \therefore E_1 &= \frac{m v}{6} \cdot \frac{3}{2} k \left[T_0 + \lambda \left(-\frac{dT}{dx} \right) \right] = \frac{m v}{6} \cdot \frac{3}{2} k \left[T_0 - \lambda \frac{dT}{dx} \right] \end{aligned} \quad (28.27)$$

The same number of electrons drifts from right to left through the volume under consideration. These electrons carry a lower energy E_2 because of the lower temperature of the particles at the site of interaction. Thus,

$$E_2 = \frac{m v}{6} \cdot \frac{3}{2} k \left[T_0 + \lambda \frac{dT}{dx} \right] \quad (28.28)$$

The excess thermal energy transferred per unit time into the unit volume is

$$J = E_1 - E_2 = \frac{m v}{6} \cdot \frac{3}{2} k \left[2\lambda \frac{dT}{dx} \right] \quad (28.29)$$

Comparing equ. (28.29) with equ. (28.25), we obtain the expression for thermal conductivity as

$$K = \frac{m v k \lambda}{2} \quad (28.30)$$

As $\lambda = v\tau$, the above equation may be expressed as

$$K = \frac{m v^2 k \tau}{2} \quad (28.31)$$

28.11 WIEDEMANN-FRANZ LAW

If free electrons are responsible for both electrical and thermal conduction in metals, then the ratio K/σ should be a universal constant, the same for all metals. In 1853, Wiedemann-Franz observed experimentally such a simple connection between K and σ . Wiedemann-Franz law states that **the ratio of thermal to electrical conductivity of a metal is proportional to the absolute temperature and the ratio is constant for all metals at a given temperature.**

That is,

$$\frac{K}{\sigma} \propto T \quad (28.32)$$

or

$$\frac{K}{\sigma T} = L \quad (28.33)$$

where L is a constant called the **Lorentz number**.

Derivation

Let us consider a metal. The thermal conductivity of the metal is given by

$$K = \frac{m v^2 k \tau}{2}$$

And its electrical conductivity is given by

$$\sigma = \frac{ne^2\tau}{m}$$

The ratio of thermal conductivity to electrical conductivity is given by

$$\frac{K}{\sigma} = \frac{mv^2k}{2e^2} \quad (28.34)$$

But $\frac{1}{2}mv^2$ = kinetic energy = $\frac{3}{2}kT$.

$$\therefore \frac{K}{\sigma} = \frac{3}{2}kT \cdot \frac{k}{e^2}$$

$$\text{or } \frac{K}{\sigma} = \frac{3}{2}\left(\frac{k}{e}\right)^2 T$$

$$\text{or } \frac{K}{\sigma} = LT \quad (28.35)$$

$$\text{where } L = \frac{3}{2}\left(\frac{k}{e}\right)^2 \quad (28.36)$$

Thus, it is seen from eq. (28.35) that the ratio of thermal to electrical conductivity of a metal is proportional to the absolute temperature and the ratio is a constant.

Wiedemann-Franz law confirmed that the motion of the free electrons is mainly responsible for both electrical and thermal conductivity.

28.12 LORENTZ NUMBER

The Lorentz number is the ratio of thermal conductivity of a metal to the product of its electrical conductivity and its absolute temperature.

$$\text{Thus, } L = \frac{K}{\sigma T}$$

Using the classical theory, we find that

$$L = \frac{3}{2}\left(\frac{k}{e}\right)^2$$

Now substituting the values of k and e in the above expression, we get the value of L .

$$L = \frac{3}{2}\left(\frac{1.38 \times 10^{-23} J/K}{1.602 \times 10^{-19} C}\right)^2 = 1.12 \times 10^{-8} \text{ watt.ohm/deg}^2$$

This classical value is only half of the experimental value. Using the expressions derived from quantum theory, we find that L is given by

$$L = \frac{\pi^2}{3}\left(\frac{k}{e}\right)^2 \quad (28.37)$$

Substituting the values of k and e into the above expression, we get

$$L = 2.45 \times 10^{-8} \text{ watt.ohm/deg}^2$$

The above value agrees well with the experimental value.

Example 28.4. The electrical resistivity of copper at $27^\circ C$ is $1.72 \times 10^{-8} \Omega m$. Compute its thermal conductivity if the Lorentz number is $2.26 \times 10^{-8} W\Omega K^{-2}$.

Solution. $\frac{K}{\sigma T} = L \quad \therefore \quad K = \frac{LT}{\rho} = \frac{2.26 \times 10^{-8} \text{ W}\Omega\text{m}^{-2} \times 300\text{ K}}{1.72 \times 10^{-8} \Omega\text{m}} = 394 \text{ Wm}^{-1}\text{K}^{-1}$.

Example 28.5. The thermal and electrical conductivities of copper at 200°C are $390 \text{ Wm}^{-1}\text{K}^{-1}$ and $5.87 \times 10^7 \Omega^{-1} \text{ m}^{-1}$. Calculate Lorentz number.

Solution. $L = \frac{K}{\sigma T} = \frac{390 \text{ Wm}^{-1}\text{K}^{-1}}{5.87 \times 10^7 \Omega^{-1}\text{m}^{-1} \times 293\text{ K}} = 2.267 \times 10^{-8} \text{ W}\Omega\text{K}^{-2}$.

28.13 RESISTANCE

Origin of Resistance

The origin of electrical resistance is assumed to be due to repeated collisions of electrons with obstacles in the material. In the classical free electron model, these obstacles were thought to be the stationary positive ions in the lattice. If it were true, the value of electron mean free path should be around a few angstroms, which is the order of distance between the ions. Experimentally determined values for mean free path are of the order of a few hundred angstroms. Secondly, the mean free path of electrons should not depend on the temperature of the material and should depend only on the concentration of the ions in it. However, resistivity of a metal is inversely proportional to the electron mean free path and a decrease in resistivity with temperature indicates an increase in the value of the mean free path. At extremely low temperatures, it is then expected that the mean free path values reach macroscopic values. It is obvious that the classical picture was incorrect because we have not taken into account the wave nature of electrons.

According to new interpretation, electrical resistance arises due to the scattering of electrons at the imperfections in the lattice. The imperfections are the deviations from perfect periodicity of the lattice. Vibrations of the ions in the lattice, vacant lattice sites, different lattice defects etc are the imperfections that exist in a real solid and all these cause electron scattering and contribute to the resistivity of the material.

Temperature dependence of Resistivity

Electrical resistivity of metals has been observed to increase linearly with increase of temperature. However, at very low temperatures, the resistivity varies as T^5 and remains practically constant and independent of temperature below 10 K down to nearly 0 K, as shown in Fig. 28.3(a).

1. The constant value of resistivity is attributed to the presence of impurity atoms in small concentrations and also to geometrical imperfections like grain boundaries, point defects etc in the metal. This resistance is called **residual resistivity**. The residual resistivity, ρ_{res} , arises due to the electron scattering at the imperfections in the metal. A relaxation time, τ_i , can be assigned to the impurity and imperfection scattering.

2. At temperatures well above 10 K, the positive ions, which are arranged in a periodic array in a metal, starts vibrating like simple harmonic oscillators. Since, the ions are bound to each other through elastic and electric forces, they vibrate with different frequencies. These vibrations are similar to standing waves with fixed energies. Therefore, they are considered to be packets of energy called **phonons**. The phonon

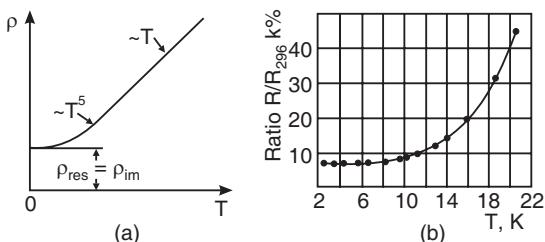


Fig. 28.3

spectrum and phonon concentration are strongly dependent on temperature. At low temperatures, phonon number reduces considerably. At low temperatures, the specific heat of metallic lattice is proportional to T^3 and therefore, the density of phonons and the probability of scattering are proportional to T^3 . In addition to this, the relaxation time is dependent on the scattering angle that would give rise to resistivity variation with temperature as T^2 . Consequently, $\rho_{ph} \propto T^3 \cdot T^2 = T^5$. At much higher temperatures, the electrical resistivity increases linearly with temperature. The relaxation time, τ_{ph} can be assigned to phonon scattering.

The total resistivity of a metal at any given temperature is a sum of resistivity due to impurity scattering, ρ_i , and phonon scattering, ρ_{ph} . Thus,

$$\rho(T) = \rho_i + \rho_{ph} \quad (28.38)$$

The above equation is known as "**Matthiessen's rule**".

As $\rho = \frac{m}{ne^2\tau}$, we can write equ. (28.38) as

$$\begin{aligned} \frac{m}{ne^2\tau} &= \frac{m}{ne^2\tau_i} + \frac{m}{ne^2\tau_{ph}} \\ \text{or } \frac{1}{\tau} &= \frac{1}{\tau_i} + \frac{1}{\tau_{ph}} \end{aligned} \quad (28.39)$$

28.14 DRAWBACKS OF CLASSICAL FREE ELECTRON THEORY

The free electron model is highly successful in explaining many physical properties of metals such as their high electrical and thermal conductivities, their high luster etc. However, it failed to account for some of the other properties. We cite some important failures of the model here.

1. Monovalent metals (copper, silver etc) have been found to have higher electrical conductivity than divalent (cadmium, zinc etc) and trivalent (aluminum, indium etc) metals. If the conductivity is proportional to electron concentration, then monovalent metals should have lesser electrical conductivity compared to the divalent and trivalent metals.
2. The model cannot explain the classification of materials into conductors, semiconductors and insulators.
3. Metals are expected to exhibit negative Hall coefficient since the current carriers in them are electrons. However, some of the metals such as zinc have positive values for Hall coefficient. The free electron model cannot explain why zinc and other metals have positive Hall coefficient.
4. According to the classical theory, the conductivity of a metal is given by the expression (28.19). Thus, the resistivity is given by

$$\rho = \frac{m}{ne^2\tau} \quad (28.40)$$

When the above expression is used to calculate the mean free path of electron in metals, we find it to be the order of about 3 Å and is consistent with the assumption of classical theory that the origin of resistivity is due to frequent collisions of electrons with the lattice ions. However, the experimentally measured values are very high and are of the order of 50 Å. The experimental results suggest that electrons pass a long distance through the lattice without collisions.

5. According to classical free electron theory, all the valence electrons can absorb thermal energy. According to the law of equipartition energy, each free electron possesses an average kinetic energy of $(3/2)kT$. If we consider a monovalent crystal,

each atom contributes one valence electron to the electron gas and there will be N free electrons per unit volume of the crystal. Then the total energy of electrons is given by

$$E = \frac{3}{2} N k T$$

When the metal is heated, the free electrons also absorb part of the heat energy and the electronic specific heat is given by

$$[C_v]_{\text{el}} = \left(\frac{dE}{dT} \right) = \frac{3}{2} N k \quad (28.41)$$

Substituting the values of N and k , we get $[C_v]_{\text{el}} = 12.5 \text{ kJ/kmol/K}$. This value is about hundred times greater than the experimentally measured value. This result implies that free electrons do not contribute significantly to the heat capacity of the metal. We conclude that the law of equipartition and hence Maxwell-Boltzmann statistics is not applicable to the free electrons in a metal.

28.15 QUANTUM FREE ELECTRON THEORY

The quantum theory for the assembly of free electrons in a metal was first advanced by Sommerfeld in 1928. He had retained the vital features of classical free electron theory and included (i) the Pauli Exclusion Principle and (ii) the Fermi-Dirac statistics to formulate his quantum free electron theory.

The main assumptions of this theory are:

- (i) The eigen values of conduction electrons are quantized and are realized in terms of a set of energy levels.
- (ii) The distribution of electrons in various allowed energy levels takes place according to Pauli's exclusion principle.
- (iii) The electrons move in a constant potential inside the metal and are confined within defined boundaries.
- (iv) Mutual attraction between electrons and lattice ions and the repulsion between individual electrons may be ignored.

28.16 DENSITY OF ENERGY STATES

Number of Energy States

Let us consider a specimen of a metal. For the sake of simplicity, we consider it to have the shape of a cube with the side L . We assume that the free electrons travel absolutely freely within the volume of the specimen. The sea of electrons obeying the Pauli Exclusion Principle is called **Fermi gas**. Since the electrons are confined inside the specimen, their wave properties limit the energy values that they may take. The application of Schrödinger equation to the electron motion in the three directions reveals that the electron energy is quantized. The quantized value of energy is given by

$$E = \frac{\hbar^2}{8mL^2} (n_x^2 + n_y^2 + n_z^2) \quad (28.42)$$

The state of a free electron is determined by the four quantum numbers n_x , n_y , n_z and by the spin quantum number $m_s = \pm \frac{1}{2}$.

Equ. (28.42) indicates that the energy of an electron is determined by the sum of the squares of the quantum numbers n_x , n_y , and n_z . We get the same value for energy for several different combinations of these three quantum numbers. Corresponding to each set we can

find a specific energy state (or energy level) E . We can therefore represent a *quantum state* by a point in quantum number space (Fig. 28.4 a). Quantum number space is an imaginary space where the values of quantum numbers are denoted along the three axes respectively.

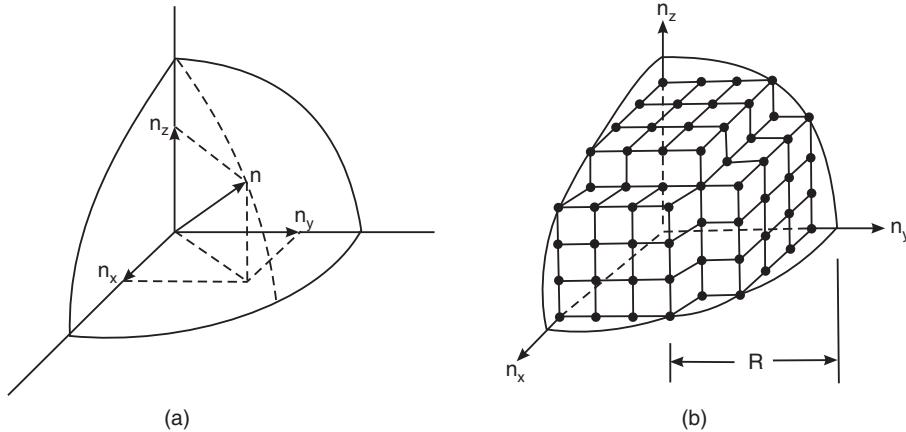


Fig. 28.4

In this space, a radius vector n may be drawn from the origin of the coordinate system to a point (n_x, n_y, n_z) where

$$n^2 = n_x^2 + n_y^2 + n_z^2 \quad (28.43)$$

It is easy to see that all points on the surface of a sphere of radius n will correspond to the same energy. All points within the sphere represent quantum states with energies smaller than E .

$$\text{Number of energy states within a sphere of radius } n = \text{Volume of the sphere} = \frac{4\pi}{3} n^3.$$

Since the quantum numbers can have only positive integer values, the n -values can only be defined in the positive octant of the sphere (Fig. 28.4 b).

Therefore, the number of quantum states with energy equal to or smaller than E is proportional to the first octant of the sphere.

$$\text{Number of energy states within one octant of the sphere of radius } n = \frac{1}{8} \times \frac{4\pi}{3} n^3 \quad (28.44)$$

Similarly, the number of energy states within one octant of a sphere of radius $(n + dn)$ corresponding to energy $(E + dE)$

$$= \frac{1}{8} \left[\frac{4\pi}{3} (n + dn)^3 \right]$$

The number of energy states having energy values between E and $(E + dE)$ is given by

$$N(E) dE = \frac{1}{8} \left[\frac{4\pi}{3} (n + dn)^3 \right] - \frac{1}{8} \left(\frac{4\pi}{3} \right) n^3 \cong \frac{\pi}{6} (3n^2 dn)$$

Terms corresponding to higher powers of dn are negligibly small and are hence neglected.

$$\therefore N(E)dE = \frac{\pi}{2} n^2 dn \quad (28.45)$$

We know that

$$E = \frac{h^2}{8mL^2} (n_x^2 + n_y^2 + n_z^2) = \frac{n^2 h^2}{8mL^2}$$

$$\therefore n^2 = \frac{8mL^2}{h^2} E \quad (28.46)$$

or $n = \left[\frac{8mL^2}{h^2} \right]^{1/2} E^{1/2}$

Differentiating the equation (28.46), we get

$$\begin{aligned} n \, dn &= \frac{4mL^2}{h^2} dE \\ \therefore N(E)dE &= \frac{\pi}{2} n^2 dn = \frac{\pi}{2} n(n \, dn) \\ &= \frac{\pi}{2} \left[\frac{8mL^2}{h^2} \right]^{1/2} E^{1/2} \times \frac{4mL^2}{h^2} dE \\ &= \frac{\pi}{4} \left[\frac{8mL^2}{h^2} \right]^{3/2} E^{1/2} dE \end{aligned} \quad (28.47)$$

There are two spin states $m_s = \pm \frac{1}{2}$ for an electron. According to Pauli Exclusion Principle, two electrons of opposite spin can occupy each state. Hence, the number of *energy states* available for electron occupancy is double the above value (28.47) and equals to

$$\begin{aligned} N(E)dE &= \frac{\pi}{2} \left[\frac{8mL^2}{h^2} \right]^{3/2} E^{1/2} dE \\ \text{or } N(E)dE &= \frac{4\pi}{h^3} (2m)^{3/2} L^3 E^{1/2} dE \end{aligned} \quad (28.48)$$

Density of Energy States

The *density of states* is given by the number of available electron states per unit volume per unit energy range at a certain energy level, E .

$$\therefore Z(E)dE = \frac{N(E)dE}{L^3} = \frac{4\pi}{h^3} (2m)^{3/2} E^{1/2} dE \quad (28.49)$$

$Z(E)$ is called the **density of states function**. It may be noted that $Z(E)$ is independent of the dimensions (L) of the potential box and hence is applicable for any case. We define $Z(E)$ as *the number of available states per unit energy interval centered around E* .

28.16.1 Energy Distribution of Electrons

Density of energy states in the energy interval E and $E + dE$ is given by

$$Z(E) dE = \frac{4\pi}{h^3} (2m)^{3/2} E^{1/2} dE$$

The density of states plotted against the energy gives a parabola (Fig. 28.5). The area under the curve represents the number of electrons in the metal. Note that *the number of available energy levels at the lower end of the parabola is considerably less than at higher energies*.

The plot of $Z(E)$ versus E at $T = 0$ drops abruptly to zero at $E = E_F$. This distribution of electron energies is understandable. The electrons cannot crowd in the lower energy levels

since they obey Pauli exclusion principle. They start with the lowest energy level and go on occupying higher and higher levels in pairs until all of them are accommodated. The highest energy occupied is E_F .

The plot of $Z(E)$ versus E at a much higher temperature is also shown in the Fig. 28.5. Since at any temperature the area under the corresponding curve gives the number of electrons in the metal, the areas under the two curves, shown in Fig. 28.5, must be equal. It is seen that even for a large increase in temperature, the distribution curve changes only very slightly.

On heating the conductor, electrons are excited to higher energy levels. The curve shows that for most of the electrons lying deep in the conduction band, the thermal energy is not sufficient to cause transitions to upper levels and electrons in lower levels are left undisturbed. Only those electrons occupying the energy levels near the Fermi level are thermally excited. These levels make up a narrow band of width kT directly adjacent to the Fermi level. Therefore, electrons having energy a little below E_F jump into levels with energy somewhat above E_F and a new energy distribution of electrons is obtained.

28.17 CARRIER CONCENTRATION IN METALS

The density of states represents the number of states that could be occupied by charge carriers. However, all the available energy states are not filled in a metal. A particular energy level E is occupied or not, is determined by the probability $f(E)$ that a charge carrier can have the energy E . Hence, the number of carriers per unit volume within a given energy range depends both on the number of available states lying in that range and on the probability that carriers acquire sufficient energy to occupy the states. Hence, the carrier concentration in energy range dE is obtained by multiplying the density of states in that range with the probability of their being occupied. Thus, at thermal equilibrium, the concentration of electrons, dn , having energy between E and $E + dE$ is given by

$$dn = f(E) \cdot Z(E)dE \quad (28.50)$$

The probability that an electron occupies an energy level E at thermal equilibrium is given by

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]} \quad (28.51)$$

where E_F is known as **Fermi level**.

By substituting for $f(E)$ and for $Z(E)$ in eq. (28.50), we get

$$dn = \frac{4\pi}{h^3} (2m)^{3/2} \cdot \frac{E^{1/2}dE}{e^{(E-E_F)/kT} + 1} \quad (28.52)$$

Therefore, the electron concentration, i.e., number of electrons per unit volume of the conductor is given by

$$n_C = \frac{4\pi}{h^3} (2m)^{3/2} \int_0^\infty \frac{E^{1/2}dE}{e^{(E-E_F)/kT} + 1} \quad (28.53)$$

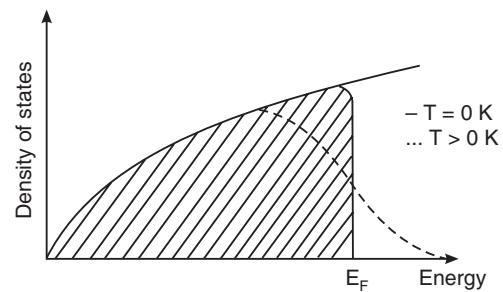


Fig. 28.5

28.18 FERMI ENERGY, E_F

Let the specimen of metal contain N free electrons. In a conductor at absolute zero temperature, the electrons fill the available states starting from the lowest energy level. Therefore, all the levels with an energy E less than a certain value $E_F(0)$ will be filled with electrons, whereas the levels with $E > E_F(0)$ will remain vacant. The energy $E_F(0)$ is known as the **Fermi energy** and the corresponding energy level is known as **Fermi level**. The total number of free electrons is equal to the number of quantum states up to the energy E_F . Note the distinction between energy level and energy state as used here; a quantum state accommodates only one electron whereas an energy level accommodates two electrons. If we set the number of states equal to the number of electrons in eq. (28.48), then we find that

$$N = \int_0^{E_F} N(E) dE = \frac{4\pi}{h^3} (2m)^{3/2} L^3 \int_0^{E_F} E^{1/2} dE$$

or

$$N = \frac{8\pi V}{3} \left(\frac{2m}{h^2} \right)^{3/2} [E_F(0)]^{3/2} \quad (28.54)$$

where we have denoted $L^3 = V$ and E_F at absolute zero temperature as $E_F(0)$.

$$\therefore E_F(0) = \frac{h^2}{2m} \left[\frac{3N}{8\pi V} \right]^{2/3} \quad (28.55)$$

In terms of electron concentration, the above equation may be written as

$$E_F(0) = \frac{h^2}{2m} \left[\frac{3n_C}{8\pi} \right]^{2/3} \quad (28.56)$$

where $n_C = \frac{N}{V}$. Eq. (28.56) shows that Fermi energy of a metal depends only on the electron concentration in the metal.

Let us estimate the value of $E_F(0)$. The concentration of conduction electrons, n_C in metals is of the order of $5 \times 10^{28}/\text{m}^3$. Therefore, $E_F(0)$ is of the order of 5 eV.

28.18.1 Variation of Fermi Energy with Temperature

Fermi energy decreases when the temperature of the metal is increased. It can be shown that when $k_T \ll E_F$, we obtain the following expression for the Fermi level E_F .

$$E_F \approx E_F(0) \left[1 - \frac{\pi^2}{12} \left(\frac{kT}{E_F(0)} \right)^2 \right] \quad (28.57)$$

From the above expression it follows that the temperature dependence of the Fermi level is very slight and for all practical purposes we assume that $E_F = E_F(0)$.

Sommerfeld theory was successful to a good extent but failed to give any basis for the classification of solids into conductors, insulators and semiconductors.

Example 28.6. Calculate the Fermi energy of sodium at 0K assuming that it has one free electron per atom and density of sodium is 970 kg/m^3 and atomic weight 23.

$$\text{Solution. } n_C = \frac{N\rho}{M} = \frac{6.02 \times 10^{26} / \text{k.mol} \times 970 \text{ kg/m}^3}{23} = 2.5 \times 10^{28}/\text{m}^3$$

$$\therefore E_F(0) = \frac{h^2}{2m} \left[\frac{3n_C}{8\pi} \right]^{2/3} = \frac{(6.625 \times 10^{-34} \text{ Js})^2}{2 \times 9.11 \times 10^{-31} \text{ kg}} \left[\frac{3 \times 2.5 \times 10^{28} / \text{m}^3}{8 \times 3.142} \right]^{2/3} = 5.11 \times 10^{-19} \text{ J}$$

$$= \frac{5.11 \times 10^{-19}}{1.602} eV = 3.1 \text{ eV.}$$

28.19 FERMI-DIRAC DISTRIBUTION FUNCTION

We are interested in knowing how electrons are distributed among the various energy levels in a conductor at a given temperature. We cannot apply Maxwell-Boltzmann distribution to electrons because (i) they obey *exclusion principle* and (ii) they are *indistinguishable* particles. The statistical distribution function applicable to quantum particles is the *Fermi-Dirac distribution* function.

The probability that an electron occupies an energy level E at thermal equilibrium is given by eq. (28.51).

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]}$$

In general E_F may or may not correspond to an energy level but it provides a reference with which other energies can be compared. The function $f(E)$ is known as *Fermi factor*.

The above equation is known as *Fermi-Dirac equation* or *Fermi-Dirac distribution function*. Note that the probability of the electron to occupy the energy level E increases with temperature.

28.20 QUANTUM FREE ELECTRON THEORY OF ELECTRICAL CONDUCTION

It is assumed in the classical free electron theory that electrons follow Maxwell-Boltzmann distribution of velocities and energies. According to Maxwell-Boltzmann distribution many electrons can simultaneously possess the same energy (or velocity). According to quantum theory, electrons obey Pauli Exclusion Principle and hence follow Fermi-Dirac distribution. According to quantum theory, only two electrons can occupy the same energy level. Hence even at 0K, conduction (or free) electrons occupy different discrete energy levels. Electrons occupying the higher energy levels would possess higher energies. The highest energy level is called the *Fermi energy level*, E_F . Therefore, the largest energy which the electrons can have in a metal at 0K is the Fermi energy, E_F . A large number of electrons possess this energy since the density of states is highest around E_F .

Each conduction electron occupies a particular energy state and possess a particular momentum, $p = \hbar k$.

If the momentum of the electron is plotted in k -space, we obtain a point (Fig. 28.6). Similarly, if we plot the momenta of conduction electrons in k -space, we obtain points. The points correspond to the tips of the k -vectors. In the absence of the electric

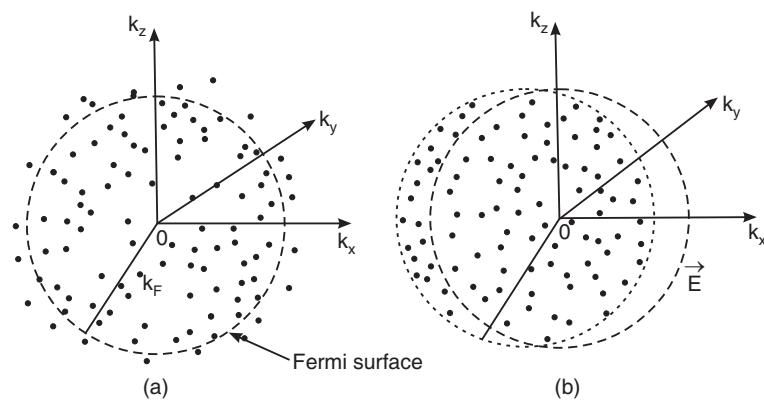


Fig. 28.6

field, the electron momenta are random and the momentum surface will have a spherical symmetry at any given temperature. At 0K, all the levels from the lowest level to the highest level would be occupied, with the highest level being the Fermi level, E_F . The outermost boundary containing all possible momenta will be a sphere, of radius, k_F (Fig. 28.6). Such a sphere is called a **Fermi surface**. In two dimensions, the Fermi surface is a circle (Fig. 28.6). All points inside the Fermi sphere are occupied. Thus, Fermi surface separates occupied states from unoccupied states. For every occupied state k there is an occupied state $(-k)$. As a consequence, the net momentum of the electrons is zero. Therefore, there is no current flow across the material.

When an electric field E is applied, electrons experience a force F , which is given by

$$F = \frac{dp}{dt} = -eE$$

or

$$\hbar \frac{dk}{dt} = -eE \quad (\because p = \hbar k) \quad (28.58)$$

Equ. (28.58) represents the equation of motion of electron in the presence of applied field. Under the action of the electric field, electrons near the Fermi surface move into vacant levels adjacent to them. However, the electrons deep in the lower levels cannot move as all the energy levels lying above them are occupied. Since the electric field acts on all of the electrons, the equilibrium is disturbed for a moment. Some of the electrons undergo collisions while others are accelerated by the field. Soon, the system returns to equilibrium. In this process, the Fermi sphere gets displaced slightly opposite to the field direction. The majority of the electron velocities cancel each other pair wise. However, some electrons remain uncompensated and cause the observed current.

Integrating eq. (28.58), we get

$$k(t) - k(0) = -\frac{eEt}{\hbar}$$

It means that the center of the Fermi sphere moves in a time $t = \tau_F$ to a center at

$$\Delta k = -\frac{eE\tau_F}{\hbar}$$

If n is the electron density per unit volume near the Fermi surface, the steady state current density would be given by

$$J = n(-e)v = -\frac{n e \hbar \Delta k}{m} = \sigma E \quad \left[\because v = \frac{p}{m} \right]$$

$$\frac{n e \hbar}{m} \left[\frac{e E \tau_F}{\hbar} \right] = \sigma E$$

$$\frac{n e^2 \tau_F}{m} E = \sigma E$$

$$\therefore \text{Conductivity } \sigma = \frac{n e^2 \tau_F}{m} \quad (28.59)$$

According to eq. (28.59), the electrical conductivity of a metal depends largely on the population density of electrons near the Fermi surface. A more rigorous expression for conductivity is given by

$$\sigma = \frac{n e^2 \tau_F}{m^*} \quad (28.60)$$

where m^* is known as the effective mass of electron.

28.21 FAILURE OF QUANTUM FREE ELECTRON THEORY

The quantum free electron model of solids could explain the properties of conductors such as electrical conductivity, thermal conductivity, heat capacity etc better than the classical theory. However, it failed to explain the distinction between conductors, insulators and semiconductors. Occurrence of positive Hall coefficients in case of some metals like zinc could not be accounted on the basis of this model.

Quantum mechanical calculations are to be applied to determine the behaviour of an electron in solid. The solution of the Schrodinger equation in a periodic lattice shows that an electron moving in a crystal have a large number of possible energy values, and these values are distributed into *energy bands*, each band consisting of a large number of closely spaced energy levels. The band theory successfully explained the classification of solids into conductors, insulators and semiconductors.

QUESTIONS

1. A rectangular block of a solid is connected to a dc voltage source. Obtain the expression:
 - (a) For the current density flowing through the block and
 - (b) For the conductivity of the material in terms of the concentration of carriers in it.
2. Explain the terms: (i) Drift velocity and (ii) Carrier mobility.
3. What are bound and free electrons? (Anna Univ., 2006)
4. Define mean free path of electrons. (Anna Univ., 2006)
5. What are the basic assumptions of classical free electron theory?
6. Discuss the important postulates of free electron theory of metals. (G.T.U., 2009)
7. Define the terms mobility and relaxation time of free electrons in a metal. (Anna Univ., 2005)
8. Using the free electron model derive an expression for electrical conductivity in metals. (VTU, 2007), (Anna Univ., 2006)
9. How does classical free electron theory lead to Ohm's law?
10. Based on free electron theory, derive an expression for electrical conductivity of metals. How does electrical resistance change with impurity and temperature? (VTU, 2008)
11. State Mathiessen's rule and give an account of the nature of total resistivity both at high and low temperatures. (VTU, 2007)
12. (i) What are the special features of classical free electron theory of metals?
 (ii) Derive an expression for the electrical conductivity of a metal. (Anna Univ., 2006)
13. Based on Drude-Lorentz theory, derive the expression for electrical conductivity and assuming the classical expression for thermal conductivity derive Wiedemann-Franz law. (Anna Univ., 2006)
14. Define Wiedemann-Franz law. (Anna Univ., 2007)
15. State and derive Wiedemann-Franz law. (Anna Univ., 2005, 2006)
16. What is meant by Lorentz number? (Anna Univ., 2005)
17. Explain Sommerfeld theory of free electron gas.
18. Write any two drawbacks of the classical free electron theory of metals. (Anna Univ., 2006)
19. Explain density of states. (VTU, 2007)
20. Derive an expression for the density of energy states and carrier concentration in a solid material (metal) by using Fermi distribution function. (Anna Univ., 2007)

21. Derive an expression for density of energy states. Obtain an expression for Fermi energy in metals at $T = 0$ K.
22. State and explain the Fermi-Dirac distribution function.
23. Explain Fermi energy and Fermi factor. Discuss the variation of Fermi factor with temperature and energy. **(VTU, 2007)**
24. Describe Fermi-Dirac distribution and discuss the same for different temperature conditions. **(VTU, 2008)**
25. Why is that only the electrons near the Fermi level contribute to electrical conductivity?
26. Derive an expression for electrical conductivity of a conducting material basing on quantum mechanical treatment.
27. Elucidate the difference between classical free electron theory and quantum free electron theory. **(VTU, 2007)**

PROBLEMS

1. The density of silver is 10.5×10^3 kg/m³. The atomic weight of silver is 107.9. Assuming that each silver atom provides one conduction electron (i) calculate the density of free electrons. The conductivity of silver at 20°C is 6.8×10^7 ohm⁻¹m⁻¹. (ii) Calculate the electron mobility in silver.
[Ans: 5.86×10^{28} ; 7.25×10^{-3} m²/V.s]

CHAPTER

29

Band Theory of Solids

29.1 INTRODUCTION

X-ray diffraction studies show that a solid is an ordered structure. In the solid atoms occupy the lattice sites and the spacing between the atoms is of the same order as that of the linear dimensions of atoms. Therefore, atoms in a solid interact strongly and set up an internal electric field, which is *periodic* in nature. The periodic electric field affects the motion of free electrons. Electron is not an ordinary particle but possesses a wave character also. The application of quantum mechanics to the motion of electrons in solid shows that the allowed values of electron energy are distributed into bands, each band consisting of a sequence of closely spaced energy levels arranged in a manner akin to the steps of a ladder. In 1928 Felix Bloch developed **zone theory** for the electrons moving in a periodic field provided by a crystal lattice. This theory is popularly known as the **band theory of solids**. Knowledge of the formation of energy bands and the consequent restrictions imposed on electron motion in a solid are obtained from the band theory. Such considerations led to the invention of a gamut of solid state devices, which has revolutionized the field of electronics leading to miniaturization, micro-miniaturization and mass production of devices and systems.

29.2 THE BAND THEORY OF SOLIDS-A QUALITATIVE EXPLANATION

A solid may be imagined as formed by allowing initially free atoms to gradually approach one another. As long as the atoms are widely separated their interactions are negligible. Every atom has the *same* energy-level diagram. The energy-level diagram for the entire system of N atoms resembles the energy-level diagram for a single atom; now each state of the *system* can be occupied by N electrons instead of just one. As the atoms come together to create a close packed periodic structure, they interact strongly due to their proximity to each other. By **interaction** we mean that the positive nucleus of one atom attracts the electrons and repels the nucleus of the adjacent atom and vice versa. As a result, instead of one energy level the same for all N isolated atoms, there arise N closely spaced separate levels, which fall into groups. The energy levels are so closely spaced in the group that they form a virtual continuum, which is called an **energy band**. The formation of energy bands can be qualitatively understood using molecular orbital concept.

Let us consider an imaginary situation where N hydrogen atoms approach each other to form solid hydrogen. Each hydrogen atom is characterized by one electron residing at $1s$ energy level corresponding to an atomic orbital ψ . As long as the separation r of two atoms A and B is much larger than the size d of the atoms ($r \gg d$), the atoms do not interact and the atomic orbitals ψ_A and ψ_B are not affected (see Fig. 29.1 a). The two atoms have identical

energy levels marked E in Fig. 29.2. When the two atoms come into contact ($r = d$), the orbitals ψ_A and ψ_B overlap and form molecular orbitals, where each electron may be said to orbit around both nuclei. A linear combination of these two atomic orbitals yields two types molecular orbitals, namely

$$\Psi_+ = C_1 \psi_A + C_2 \psi_B \quad \text{and} \quad \Psi_- = C_3 \psi_A - C_4 \psi_B$$

where Ψ_+ and Ψ_- denote molecular orbitals and C_1, C_2, C_3 and C_4 are constants.

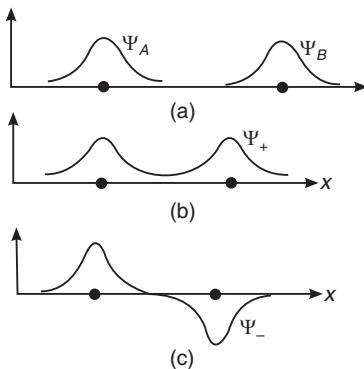


Fig. 29.1

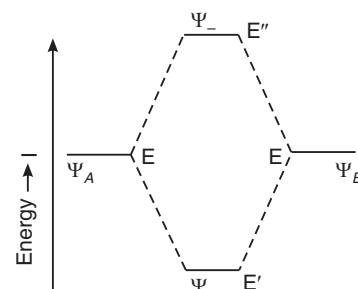


Fig. 29.2

The molecular orbital Ψ_+ is called the **bonding orbital** and has a lower energy E' ($E' < E$). The molecular orbital Ψ_- is called the **antibonding orbital** and has a higher energy E'' ($E'' > E$). Thus, the combination of two atomic orbitals results in two molecular orbitals, which extend over both the atoms. The first important consequence of the molecular orbital formation is that the individual valence electrons are not localized to their original atomic positions and they belong to both the atoms. The second important result is that one of the molecular orbitals (bonding orbital) is of lower energy than either of the individual atomic orbitals and the other molecular orbital (antibonding orbital) is of higher energy. It means that the original energy level E of each electron is split into two energy levels (see Fig. 29.2). *The transformation of a single energy level into two or more separate energy levels is known as energy level splitting.*

When two atoms come close, one energy level splits into two energy levels (Fig. 29.3 b). When three atoms approach each other closely, the original level splits into three levels; four atoms produce four levels and so on (Fig. 29.3 c). In general, N interacting atoms cause a particular energy level to split into N levels. The group of energy levels resulting from splitting is so closely spaced that it is called an *energy band* (see Fig. 29.3 d).

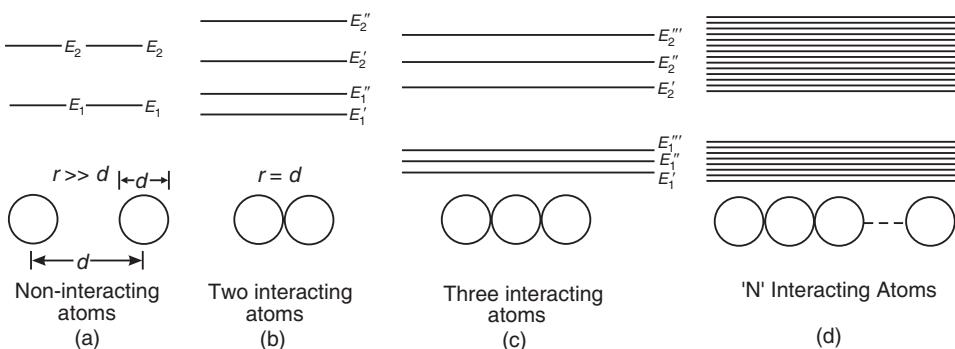


Fig. 29.3. Energy level splitting and band formation.

The individual valence electrons no longer belong to individual atoms; but they now belong to all nuclei in the solid.

29.3 THE BAND THEORY OF SOLIDS—QUANTUM MECHANICAL EXPLANATION

In order to find the allowed energies of electrons in solids, we have to apply Schrödinger wave equation for an electron in a crystal lattice. Fig. 29.4(a) shows the actual potential as seen by an electron in the crystal lattice in one dimension.

29.3.1 The Kronig-Penny Model

Kronig and Penny suggested a simplified model consisting of an infinite row of rectangular potential wells separated by barriers of width b . This one-dimensional representation of periodic lattice is known as *Kronig-Penny model* and is shown in Fig. 29.4 (b). Each well has a width b and a depth V_0 .

The period of the potential is $(a + b)$. In regions where $0 < x < a$, the potential energy is assumed to be zero and in regions such as $-b < x < 0$, it is V_0 . Through this model, Schrodinger equation can be solved explicitly in terms of elementary functions.

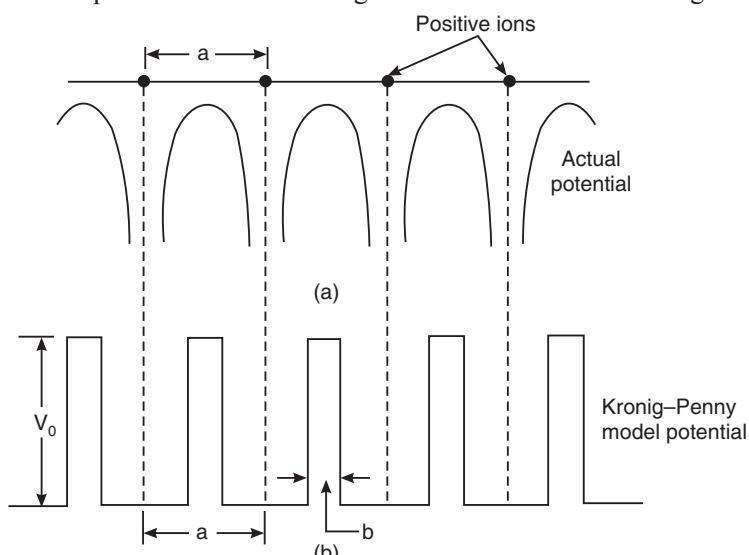


Fig. 29.4. Kronig-Penny Model

29.3.2 Bloch Theorem

The Schrodinger equation for the two regions can be written as

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2m}{h^2} E\psi = 0 \quad \text{for } 0 < x < a \quad (29.1)$$

and $\frac{d^2\psi}{dx^2} + \frac{8\pi^2m}{h^2} (E - V_0)\psi = 0 \quad \text{for } -b < x < 0 \quad (29.2)$

We rewrite the above equation as

$$\frac{d^2\psi}{dx^2} + \alpha^2 \psi = 0 \quad \text{for } 0 < x < a \quad (29.3)$$

and $\frac{d^2\psi}{dx^2} + -\beta^2 \psi = 0 \quad \text{for } -b < x < 0 \quad (29.4)$

There is an important theorem known as the **Bloch theorem**. According to this theorem, the solution of the Schrodinger equation for a periodic potential would be of the form of a plane wave modulated with the periodicity of the lattice. It means that the solution can

be represented as the product of two functions: a free particle wave function and a periodic function $u(x)$ that has the same period as the lattice. Thus

$$\psi(x) = u(x)e^{ikx} \quad (29.5)$$

with

$$u(x) = u(x + a) \quad (29.6)$$

The wave functions of the above type are known as **Bloch functions** which change periodically with increasing x .

29.3.3 Energy Bands

We substitute the above wave functions into the Schrodinger equation and solve it in the usual way. When we apply the periodic boundary condition, we get the following expression.

$$\frac{maV_0b}{\hbar^2} \cdot \frac{\sin \alpha a}{a} + \cos \alpha a = \cos k a \quad (29.7)$$

where

$$\alpha = \frac{\sqrt{2mE}}{\hbar}$$

Equ. (29.7) provides the allowed solutions to the Schrodinger equation. As the relation involves trigonometric functions, only certain values of α are possible. The right hand side of equation (29.7) is cosine function and can take values only between -1 and +1. Therefore, the left-hand side of the equation is restricted to vary between those two limits. Hence, only certain values of α are allowed. It means that energy E is restricted to lie within certain ranges.

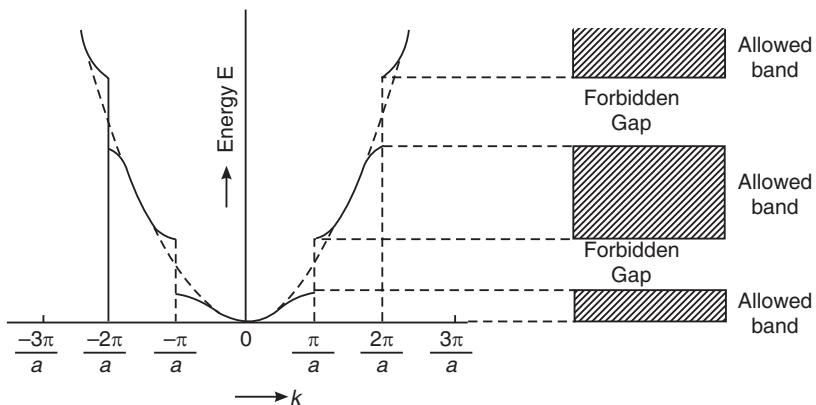


Fig. 29.5. Electron energy, E versus wave number, k plot for a solid

This concept is best understood by drawing the plot of energy E as a function of the wave number, k . The plot is shown in Fig. 29.5. The parabolic relation between E and k obtained in case of free electron is interrupted at certain values of k , as shown by the broken curve.

Fig. 29.5 shows discontinuities in E . The discontinuities occur at $ka = \pm n\pi$ i.e., at

$$k = \pm \frac{\pi}{a}, \pm \frac{2\pi}{a}, \pm \frac{3\pi}{a}, \dots \quad (29.8)$$

The origin of the allowed energy bands and forbidden gaps is clear from Fig. 29.5.

29.4 ENERGY BAND STRUCTURE OF A SOLID

A crystal (i.e., solid) consists of an enormous number of atoms arranged in a regular periodic structure. The extent of energy level splitting in the solid depends on the nearness of atoms in it. Let us assume that N identical atoms form the crystal. The energy levels of the isolated atoms are shown in Fig. 29.6(c). All the N atoms have identical sets of energy levels. The electrons fill the energy levels in each atom independently. Fig. 29.6 (b) shows an atom sitting

at the origin of the coordinate system. Now let us imagine that other atoms approach this atom along the three axial directions and halt at the distance a_0 , which is the lattice constant of the crystal. As the atoms approach, a continuously increasing interaction occurs between the atoms. Each of the energy levels splits into many distinct levels and form energy bands, as shown in Fig. 29.6 (b). Fig. 29.6 (a) depicts the effect of slicing of Fig. 29.6 (b) at a_0 and it represents the *energy band structure* of the crystal. It is seen that corresponding to each allowed energy level of an isolated atom, there forms an allowed energy band; and that the allowed energy bands are separated by forbidden bands of energy.

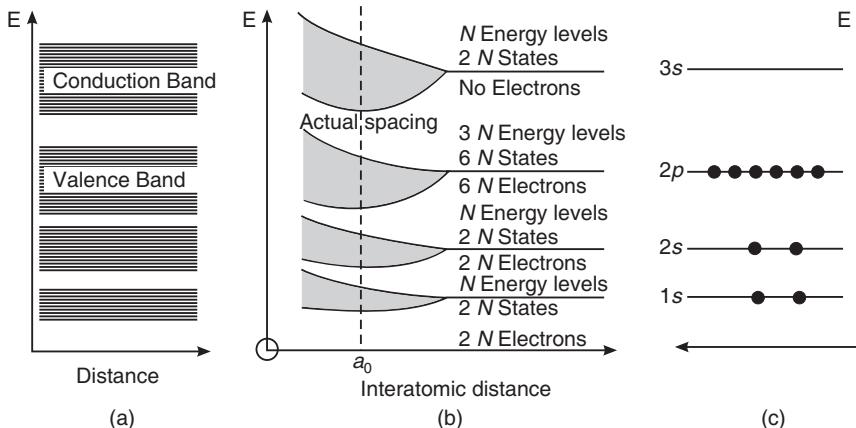


Fig. 29.6. Energy level splitting in a solid as a function of interatomic distance (a) Energy band structure of the solid corresponding to the actual spacing of atoms in the solid. (b) Energy level splitting as a function of distance. (c) Discrete energy levels in an isolated atom.

The degree of splitting of energy levels depends on their depth in the atom. The electrons in outer shells screen the core electrons belonging to inner shells. Consequently, the energy levels of inner shell electrons are split to a lesser degree. They form narrow core bands. They are always completely filled and do not participate in electrical conduction. In contrast the energy levels of valence electrons are split more and form wider bands.

In general, N interacting atoms cause an energy level to split into $(2l + 1)N$ levels. Thus, s-level ($l = 0$) splits into N levels whereas the p-level, consisting of three sublevels p_x , p_y and p_z , splits into $3N$ levels. Thus, in a solid each level of an individual atom splits into $(2l + 1)N$ number of levels where N is the number of atoms in the system. Consequently, the maximum electron capacity of an s-band is $2N$ electrons whereas the capacity of a p-band is $6N$ electrons.

While occupying an energy band, electrons start from the lowest energy level in the band and fill the levels one after the other in the ascending order of energy. When $2N$ electrons occupy the N levels available in the band, the band is said to be **completely filled**. In case of non-availability of $2N$ electrons, the energy band gets **partially filled**. When there are no electrons to occupy the levels, the energy band remains **vacant**.

The width of an allowed or forbidden energy band is generally of the order of a few electron-volts. As N is very large, the energy separation between successive energy levels in an allowed band is very small and is of the order of 10^{-27} eV. At room temperature, the kinetic energy of the electrons of the order of kT (≈ 0.026 eV) which is very large compared to the energy level separation in an allowed band. Consequently, electrons can easily move into higher vacant levels within the allowed energy band either due to thermal energy or due to a small externally applied electric field. On the other hand, electrons cannot jump across a

forbidden band under normal thermal energy possessed by them or due to an applied electric field. High temperatures are required to cause inter-band electron transitions.

29.5 ELECTRICAL CONDUCTION FROM THE VIEW POINT OF BAND THEORY

In a solid, the allowed values of electron energy are distributed into bands (Fig. 29.6), each band consisting of a sequence of closely spaced discrete energy levels arranged in a manner akin to the rungs of a ladder. The electrons are distributed in the energy levels according to the Pauli exclusion principle. The motion of an electron corresponds to its transition from a lower energy level to an upper vacant energy level. This implies that the following two conditions are to be fulfilled for electrical conduction to take place in a solid:

- (i) There should be free electrons available in the solid.
- (ii) Vacant energy levels should be available immediately above the levels occupied by free electrons.

If a band has vacant energy levels but is devoid of electrons, there would be no carriers to move through the vacant levels when energy is supplied to the solid from a source such as a battery. Hence current does not flow through the solid.

On the other hand, if all the energy levels within a band were completely occupied by electrons, there would be no energy level to which an electron can jump. Therefore, even though the energy is supplied to the solid from a source such as a battery, the electrons cannot acquire energy and electrical conduction cannot occur in the solid.

If an energy band is partially filled, then the electrons will have vacant upper energy levels into which they can jump. On acquiring energy from the electric field applied across the solid, the electrons move into successive upper energy levels and cause electrical conduction. Thus, *partially filled energy band is required for electrical conduction* in a solid.

29.6 ENERGY BAND DIAGRAM

An *energy band diagram* is a graphic representation of the energy levels associated with top energy band and the next lower energy band in a solid. The energy band diagram shows two bands with a gap in-between (see Fig. 29.7). The upper band is called the **conduction band** and the lower energy band is called the **valence band**. These two bands are separated by a *forbidden gap*. This energy gap is more popularly called **band gap** and is denoted by the symbol E_g . The conduction band corresponds to the energy values of *free electrons* that have broken their valence bonds, and hence have become free to move in the crystal. The bottom of the conduction band represents the smallest energy that the electron must possess to become free. Only the free electrons can move in the crystal under the influence of the externally applied electric field. Hence, these electrons are called **conduction electrons** and the energies of such electrons constitute the **conduction band**. The band showing the energy values of **valence electrons** that are engaged in covalent bonding is called the **valence band**.

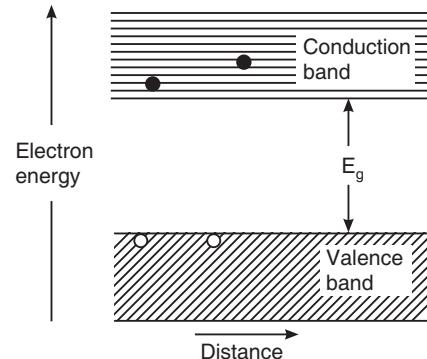


Fig. 29.7. Energy band diagram

29.7 CLASSIFICATION OF SOLIDS

The concept of energy bands helps us in understanding the division of solids into three groups. The nature of the energy bands determines whether the solid is an electrical conductor

or insulator. According to the band theory, the electrical conductivity a solid is characterized by the energy gap E_g separating the outermost energy bands namely, the valence band and the conduction band. The ability of electrical conduction is decided by the order of magnitude of the energy gap E_g .

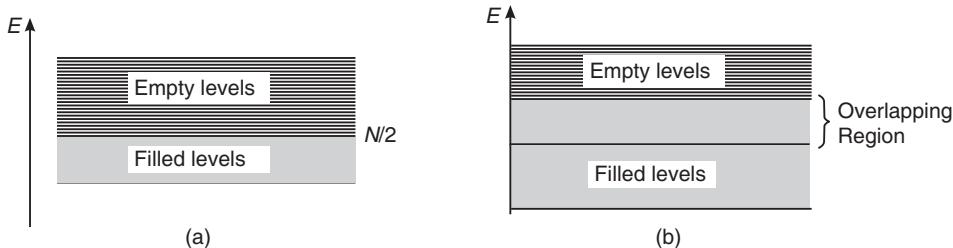


Fig. 29.8. Energy band formation in a conductor (a) Half filled conduction band.
(b) Empty upper band overlaps on a totally filled lower band.

In some solids, an upper vacant band overlaps the valence band or the valence band itself is half-filled, as shown in Fig. 29.8. It means that electrons in the valence band have easy access to levels in the upper vacant band. For this reason, very large numbers of electrons are available for conduction, even at extremely low temperatures. When electric field is impressed across the solid, electrons readily jump into upper unoccupied energy levels of the vacant band and current flows in a large measure in the solid. Therefore, these solids exhibit good electrical conductivity and are called **conductors**.

In some solids the band gap is narrow and of the order of 2 eV or less, as shown in Fig. 29.9. Acquisition of small amounts of energy from the vibrations of atom can raise electrons from the valence band to the conduction band. The conduction band is then partially filled. If a potential is applied across the material, it causes the electrons in the conduction band to move to upper levels. As a result, current flows in a modest measure in the solid. Such solids are called **semiconductors**.

Some solids (Fig. 29.10) have band gaps that are very wide ($E_g > 3$ eV). It would require the acquisition of very large amounts of energy to cause an electron to jump from the valence band to the conduction band. Very few electrons can get this large amount of energy to jump from valence band to conduction band at ambient temperature. Hence, there are very few electrons in the conduction band. When a voltage is applied across

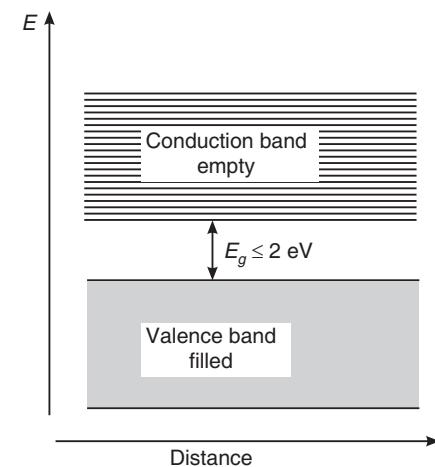


Fig. 29.9. Energy band structure of a semiconductor

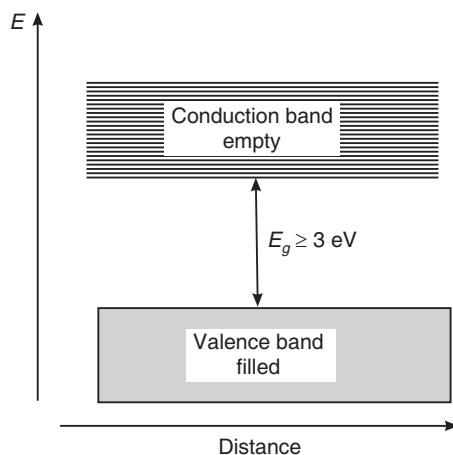


Fig. 29.10. Energy band structure of an insulator

the solid, negligible current flows and the solid exhibits very low electrical conductivity. These solids are called **insulators**.

29.8 ENERGY BAND DIAGRAMS FOR SOME TYPICAL SOLIDS

29.8.1 Lithium

Let us consider the element lithium belonging to Group I in the periodic table. The electron configuration of lithium atom is $1s^2 2s^1$. The $1s$ shell is closed and there is only one electron at the $2s$ level. In solid lithium, $1s$ and $2s$ bands form corresponding to the $1s$ and $2s$ levels, as illustrated in Fig. 29.11.

Both $1s$ and $2s$ bands have N levels each. The $1s$ band is completely filled as $2N$ electrons occupy N energy levels whereas the $2s$ band is half-filled because the N available electrons fill $N/2$ lower levels in the band leaving the upper $N/2$ levels vacant. In general, the solids of Group-I elements form such half-filled energy bands at the top and therefore belong to the group of conductors.

29.8.2 Beryllium

Let us next consider the case of alkaline earth elements of Group-II. The first element in this group is beryllium. Its electron configuration is $1s^2 2s^2$. From the electron configuration, it is expected that beryllium solid would be an insulator. However, it is known to be a conductor. The reason is that the upper vacant $2p$ band overlaps the lower completely filled $2s$ band leading to the formation of a partially filled hybrid band, as shown in Fig. 29.12.

In general, solids of Group-II elements exhibit such partially filled bands, and therefore belong to the conductors group.

29.8.3 Energy Band Diagrams for Silicon and Diamond

Silicon belongs to Group IV elements in the periodic table. The electron configuration of silicon atom is $1s^2 2s^2 2p^6 3s^2 3p^2$. It is seen that the inner K and L shells are closed and the corresponding bands would be completely filled. In the outer subshells $3s$ and $3p$, $3s$ -subshell is closed. The $3p$ sub-shell is partially filled. Hence it is expected to behave as a good conductor. But because of formation of a hybrid band, which later branches out, the Si solid behaves as a semiconductor.

In the crystal formation process, when the atoms are very far apart, as at position 'd' in Fig. 29.13, the electrostatic interaction among them is negligible. Consequently, the electronic

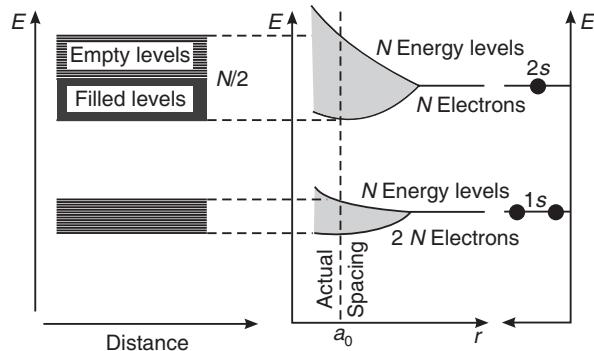


Fig. 29.11. Energy level splitting and energy band configuration in lithium solid showing half-filled $2s$ band.

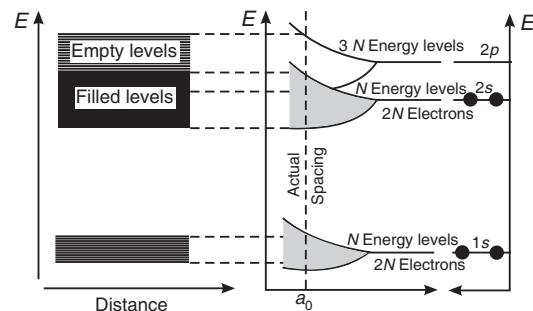


Fig. 29.12. Energy level splitting and energy band configuration in beryllium solid showing overlapping of completely filled $2s$ band and vacant $2p$ band.

energy levels of the crystal will be the same as those of isolated atoms. As the separation between atoms decreases, the $3s$ and $3p$ levels split and two bands are formed, as shown at position 'c' in Fig. 29.13. The band corresponding to $3s$ level has N energy levels and the band corresponding to $3p$ level has $3N$ levels. $2N$ electrons occupy N levels in $3s$ -band and $2N$ electrons occupy N levels in $3p$ -band. It may now be noted that there is an energy gap between the two bands. The energy gap is seen to decrease with the decrease in atomic spacing. At

position 'b' in Fig. 29.13 the two bands merge and form a composite band. The $3N$ upper levels merge with N lower levels, giving rise to a total of $4N$ levels. These levels have to be occupied by the $4N$ electrons available in total, and so the lowermost $2N$ levels are filled. When the atomic distance in our imaginary crystal is further reduced, the interaction among the atoms becomes very strong. Beyond the lattice spacing 'b' in Fig. 29.13, we find that the composite band branches out and once again two bands are formed, separated by a forbidden gap, E_g . The significant point is that the $4N$ energy levels are equally divided between the two branches. There is an equal distribution of levels, $2N$ in each, in the two bands. The $4N$ electrons available in total at $3s$ and $3p$ levels, now occupy the lower energy band and leave the upper band vacant. The lower band constitutes the valence band and the upper band the conduction band. This is the situation at the actual spacing ' a_o ' in the silicon crystal. At position a_o the two bands are not widely separated from each other. The value of E_g at 0 K is 1.12 eV . At normal temperatures, a significant number of electrons will be thermally excited from valence band to conduction band. The electrons excited to conduction band respond to the external voltage and produce a modest flow of current. Thus, Si behaves as a semiconductor.

Diamond

It is evident from Fig. 29.13 that the energy gap between the two branches goes on increasing with decreasing atomic distance. At the interatomic distance corresponding to line at 'a' in Fig. 29.13, the distance between the two bands becomes considerably large. In case of diamond the two bands are separated by 5.47 eV . Even at high temperatures, the thermal energy would be insufficient to excite enough number of electrons to the conduction band. Because of the non-availability of electrons in the conduction band electrical conduction cannot take place in the material and hence diamond behaves as insulator.

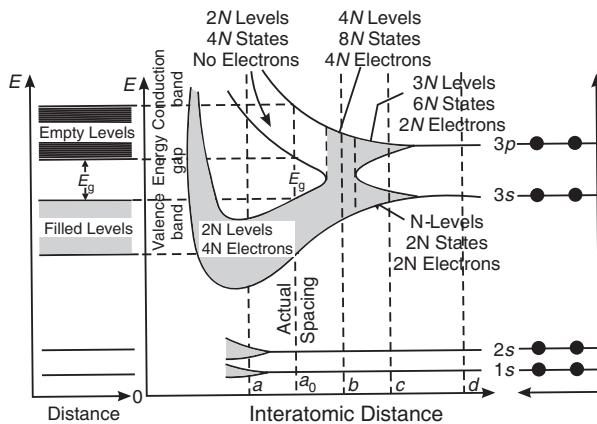


Fig. 29.13. Energy level splitting and band formation in crystals of Group IV A elements

29.9 ENERGY BAND STRUCTURE OF A CONDUCTOR

Conductors are characterized by a single energy band, namely conduction band which contains more energy levels than free electrons. At the temperature 0K , electrons occupy the lower energy levels in the conduction band up to a certain energy level called the Fermi level E_F .

29.9.1 Fermi-Dirac Distribution Function

We are next interested in knowing how electrons are distributed among the various energy levels in the conduction band at a given temperature. We cannot apply Maxwell-Boltzmann distribution to electrons because (i) they obey *exclusion principle* and (ii) they are *indistinguishable* particles. The statistical distribution function applicable to quantum particles is the *Fermi-Dirac distribution* function.

The probability that an electron occupies an energy level E at thermal equilibrium is given by

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]}$$

where E_F is known as **Fermi level**. In general E_F may or may not correspond to an energy level but it provides a reference with which other energies can be compared. The function $f(E)$ is known as **Fermi factor**.

The above equation is known as *Fermi-Dirac equation* or *Fermi-Dirac distribution function*. Note that the probability of the electron to occupy the energy level E increases with temperature. We first discuss about the distribution function and the related topics with reference to conductors. We shall find later that these concepts are equally applicable to other cases.

Example 29.1. Evaluate the Fermi function for energy kT above the Fermi energy.

Solution. The Fermi function is given by $f(E) = \frac{1}{1 + e^{(E-E_F)/kT}}$

$$\text{If } (E - E_F) = kT, \text{ then } f(E) = \frac{1}{1 + e^{(E-E_F)/kT}} = \frac{1}{1 + e^1} = \frac{1}{1 + 2.78} = \frac{1}{3.78} = 0.269.$$

Example 29.2. In a solid, consider the energy level lying 0.01 eV below Fermi level. What is the probability of this level not being occupied by an electron?

Solution. $(E_F - E) = [E_F - (E_F - 0.01)] = 0.01 \text{ eV}$ and $kT = 0.026 \text{ eV}$ at $T = 300 \text{ K}$

The probability of an energy level E not being occupied by an electron is given by $[1 - f(E)]$.

$$\begin{aligned} [1 - f(E)] &= 1 - \frac{1}{1 + e^{(E-E_F)/kT}} = \frac{1}{1 + e^{(E_F-E)/kT}} = \frac{1}{1 + e^{0.01\text{eV}/0.026\text{eV}}} = \frac{1}{1 + e^{0.385}} \\ &= \frac{1}{1 + 1.47} = 0.405 \end{aligned}$$

29.9.2 Fermi Level

The occupancy of the energy levels by electrons in conductors is described by the Fermi-Dirac distribution function.

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]} \quad (29.9)$$

We distinguish two situations - one at absolute zero and the other at higher temperatures.

Case 1: $T = 0 \text{ K}$

Fig. 29.14 (a) depicts the conduction band of a conductor at 0K. At absolute zero, electrons occupy energy levels in pairs starting from the bottom of the band up to an upper level designated as E_F , leaving the upper levels vacant. **Fermi level** can be, therefore, defined as *the uppermost filled energy level in a conductor at 0K*. Correspondingly, **Fermi energy**

is defined as *maximum energy that a free electron can have in a conductor at 0K*. To use an analogy, the electron distribution in the conduction band can be likened to water at rest in a container. The Fermi level corresponds to the top surface of water. The Fermi function at 0K is shown in Fig. 29.14 (b).

Let us now apply equ. (29.9) to the solid taking the value of T as 0 K.

(i) For energy levels E lying below E_F , $E < E_F$, $(E - E_F)$ is a negative quantity.

$$\therefore f(E) = \frac{1}{1 + e^{-(E-E_F)/0}} = \frac{1}{1 + e^{-\infty}} = \frac{1}{1 + 0} = 1$$

$f(E) = 1$ indicates that all the energy levels lying below the level E_F are occupied.

(ii) For energy levels located above E_F , $E > E_F$, $(E - E_F)$ is a positive quantity.

$$\therefore f(E) = \frac{1}{1 + e^{(E-E_F)/0}} = \frac{1}{1 + e^{\infty}} = \frac{1}{1 + \infty} = \frac{1}{\infty} = 0$$

The result $f(E) = 0$ implies that all the levels above E_F are vacant at $T = 0K$.

(iii) For $E = E_F$, the quantity $(E - E_F) = 0$.

$$\therefore f(E) = \frac{1}{1 + e^{0/0}} = \text{indeterminate}$$

The above result implies that the occupancy of Fermi level at 0K ranges from zero to one.

Case 2: $T > 0K$

On heating the conductor, electrons are excited to higher energy levels. In general, $E_F \gg kT$. Therefore, for most of the electrons lying

deep in the conduction band, the thermal energy is not sufficient to cause a transition to an upper unoccupied level. At normal temperatures, only those electrons occupying the energy levels near the Fermi level can be thermally excited. These levels make up a narrow band of width kT directly adjacent to the Fermi level. Therefore, upon heating the solid, electrons having energy a little below E_F , jump into levels with energy somewhat above E_F and a new energy distribution of electrons is obtained.

Thus, as a result of thermal excitation, the probability of finding electrons in the levels immediately below E_F will decrease. On the same hand, the probability of finding electrons in

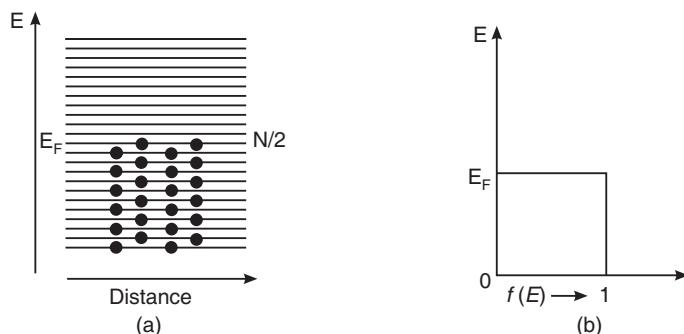


Fig. 29.14

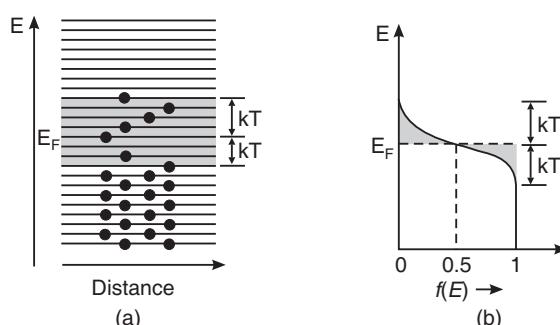


Fig. 29.15

the levels immediately above E_F increases. This fact is reflected in the graph (Fig. 29.15 b) as a blurring of the step plot.

At $T > 0\text{K}$, if we consider an electron at Fermi level, then $E = E_F$.

$$\therefore f(E) = \frac{1}{1 + e^{0/kT}} = \frac{1}{1+1} = \frac{1}{2}$$

This implies that the probability of occupancy of Fermi level at any temperature above 0K is 0.5 or 50%. Now we can define Fermi level as the energy level, which has a probability of occupancy of 0.5. An operational definition of Fermi energy can be given now. *Fermi energy is the average energy possessed by electrons participating in conduction in metals at temperatures above 0K.*

29.9.3 Effect of Temperature on Fermi Function

The Fermi-Dirac distribution curves for different temperatures are shown in Fig. 29.16. At $T = 0\text{ K}$, there is an abrupt jump in the value of $f(E)$ from 1 to zero at E_F . At $T > 0\text{K}$ the change is gradual. The higher the temperature, more gradual is the change.

It is seen from the curves for different temperatures in Fig. 29.16 that they all pass through a **crossover point C**, at which the probability of occupancy is 0.5. This is due to the fact that $f(E)$ has a value of 0.5 for any temperature greater than 0K.

We may deduce from the curves that *Fermi energy E_F is the average energy possessed by electrons that participate in conduction process in a conductor at temperatures above absolute zero.*

Example 29.3. The Fermi level for potassium is 2.1 eV. Calculate the velocity of the electrons at the Fermi level.

Solution: $E_F = \frac{1}{2}mv_F^2$

$$\therefore v_F^2 = \frac{2E_F}{m} = \frac{2 \times 2.1 \times 1.602 \times 10^{-19} \text{ C} \cdot \text{V}}{9.10 \times 10^{-31} \text{ kg}} = 0.74 \times 10^{12} \text{ m}^2/\text{s}^2$$

$$v_F = 8.6 \times 10^5 \text{ m/s.}$$

Example 29.4. The Fermi level of silver is 5.5 eV. Calculate the fraction of free electrons at room temperature located up to a width of kT on either side of the Fermi level.

Solution. The number of electrons that occupy levels above E_F at a temperature T is proportional to kT . Therefore, the fraction of electrons that occupies levels higher than E_F is given by

$$\frac{kT}{E_F} = \frac{0.026 \text{ eV}}{5.5 \text{ eV}} = 0.0047$$

Similarly, the fraction of electrons that are deprived of partners ≈ 0.0047

$$\therefore \text{The fraction of free electrons that is located up to a width } kT \text{ on either side of } E_F \\ = 2 \times 0.0047 = 0.0094 \approx 0.01$$

Example 29.5. At what temperature we can expect a 10% probability that electrons in silver have an energy which is 1% above the Fermi energy? The Fermi energy of silver is 5.5 eV.

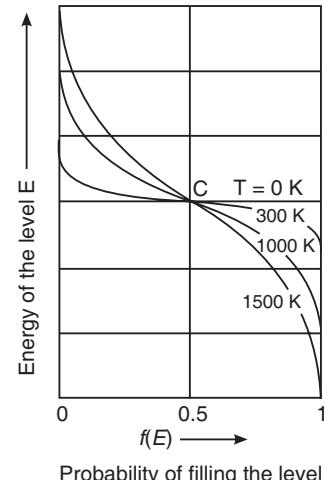


Fig. 29.16

Solution. Given that the electron energy is $E = E_F + 1\%E_F$.

$$\therefore E - E_F = 1\%E_F = \frac{5.5 \text{ eV}}{100} = 0.055 \text{ eV}$$

Also

$$f(E) = \frac{1}{1 + e^{(E-E_F)/kT}}$$

$$\text{Putting } \frac{E - E_F}{kT} = x, \quad f(E) = \frac{1}{1 + e^x}$$

$$\text{As } f(E) = 10\% = 0.1, \text{ we get } \frac{1}{1 + e^x} = 0.1 \quad \text{or} \quad x = 2.197$$

$$\therefore \frac{E - E_F}{kT} = 2.197$$

$$\text{or } T = \frac{E - E_F}{2.197 \times k} = \frac{0.055 \text{ eV}}{2.197 \times 8.61 \times 10^{-5} \text{ eV/K}} = 290 \text{ K.}$$

Example 29.6. Find the temperature at which there is 1% probability that a state with energy 2 eV is occupied. Given that Fermi energy is 1.5 eV.

Solution. The probability of an energy state E being occupied by an electron is given by

$$f(E) = \frac{1}{1 + e^{(E-E_F)/kT}}$$

$$E - E_F = 2 \text{ eV} - 1.5 \text{ eV} = 0.5 \text{ eV} \quad \text{and} \quad f(E) = 1\%$$

$$\therefore \frac{1}{100} = \frac{1}{1 + e^{0.5/kT}} \quad \text{or} \quad e^{0.5/kT} = \frac{0.99}{0.01} = 99$$

Taking logarithm on both the sides, we get

$$\frac{0.5 \text{ eV}}{kT} = 2.303 \log 99$$

$$\text{or } \frac{0.5 \text{ eV}}{kT} = T = \frac{0.5 \text{ eV}}{2.303 \log 99 \times 8.61 \times 10^{-5} \text{ eV}} = 1262 \text{ K.}$$

Example 29.7. Show that the probability of finding an electron of energy ΔE above the Fermi level is same as probability of not finding an electron at energy ΔE below the Fermi level.

OR

Show that the probability that a state ΔE above the Fermi level E_F is filled equals the probability that a state ΔE below E_F is empty.

Solution. Let us consider an energy level E_2 that is above the Fermi level by an amount of energy ΔE . The probability that the energy level E_2 is occupied is given by

$$\begin{aligned} f(E_2) &= f(E_F + \Delta E) = \frac{1}{1 + \exp[(E_2 - E_F)/kT]} \\ &= \frac{1}{1 + \exp[(E_F + \Delta E - E_F)/kT]} \\ \therefore f(E_2) &= \frac{1}{1 + \exp(\Delta E / kT)} \end{aligned} \tag{1}$$

Next, let us consider the energy level E_1 that is below the Fermi level by energy ΔE . $[1 - f(E_1)]$ gives the probability that the level E_1 is not occupied.

$$\begin{aligned}
 [1 - f(E_1)] &= [1 - f(E_F - \Delta E)] \\
 &= 1 - \frac{1}{1 + \exp[(E_F - \Delta E - E_F) / kT]} \\
 &= 1 - \frac{1}{1 + \exp[-\Delta E / kT]} \\
 &= \frac{\exp[-\Delta E / kT]}{1 + \exp[-\Delta E / kT]}
 \end{aligned}$$

or

$$[1 - f(E_1)] = \frac{1}{1 + \exp[(\Delta E) / kT]} \quad (2)$$

The R.H.S of equation (1) and (2) are the same.

$$\therefore f(E_2) = [1 - f(E_1)]$$

It means that the probability of an energy level $[E_F + \Delta E]$ (ΔE above the Fermi level) being occupied is the same as the probability of an energy level $[E_F - \Delta E]$ (ΔE below E_F), being vacant.

Example 29.8. Show that the occupancy probabilities of two states whose energies are equally spaced above and below the Fermi energy add up to one.

Solution. Let us consider two energy levels E_2 and E_1 , which are equally spaced above and below the Fermi energy E_F .

Let

$$E_2 = E_F + \Delta E \quad \text{and}$$

$$E_1 = E_F - \Delta E$$

The probability of occupancy of the level E_2 is given by

$$\begin{aligned}
 F(E_2) &= F(E_F + \Delta E) = \frac{1}{1 + \exp[(E_2 - E_F) / kT]} \\
 &= \frac{1}{1 + \exp[(E_F + \Delta E - E_F) / kT]} \\
 \therefore F(E_2) &= \frac{1}{1 + \exp(\Delta E / kT)}
 \end{aligned}$$

The probability of occupancy of the level E_1 is given by

$$f(E_1) = f(E_F - \Delta E) = \frac{1}{1 + \exp[(E_F - \Delta E - E_F) / kT]}$$

or

$$f(E_1) = \frac{1}{1 + \exp[-\Delta E / kT]}$$

\therefore

$$\begin{aligned}
 f(E_1) + f(E_2) &= \frac{1}{1 + \exp(\Delta E / kT)} + \frac{1}{1 + \exp(-\Delta E / kT)} \\
 &= \frac{1}{1 + \exp(\Delta E / kT)} + \frac{\exp(\Delta E / kT)}{1 + \exp(\Delta E / kT)} \\
 &= \frac{1 + \exp(\Delta E / kT)}{1 + \exp(\Delta E / kT)} = 1
 \end{aligned}$$

$$\therefore f(E_1) + f(E_2) = 1$$

Thus, the occupancy probabilities of two states whose energies are equally spaced above and below the Fermi energy add up to one.

29.10 ENERGY BAND STRUCTURE OF AN INSULATOR

Insulators are characterized by two energy bands – conduction band and valence band, separated by a large energy gap. At 0K all valence electrons are engaged in covalent bonds, the valence band is full. The absence of mobile charge carriers keeps the conduction band vacant. The situation is same even at higher temperatures (300K), as the valence band and conduction bands are separated by a large gap (> 3 eV) and it is not possible to excite electrons from valence band to conduction band by thermal energy ($kT \approx 0.026$ eV). Consequently, insulators do not allow flow of current even at temperatures higher than room temperature.

The concept of Fermi level can be extended to insulators also. As the energy levels in valence band are filled, $f(E)$ is equal to unity throughout the valence band. As there are no electrons in the conduction band, $f(E)$ is equal to zero throughout the conduction band. Since the Fermi function is symmetrical about E_F , the Fermi level may be expected to be situated in the middle of the energy gap. Even though there are no energy levels and no electrons in the band gap, the meaning of Fermi level remains the same. It is a *reference energy position*. The energy band diagram for an insulator is shown in Fig. 29.17 along with probability function.

29.11 ENERGY BAND STRUCTURE OF A SEMICONDUCTOR

A semiconductor is characterized by two energy bands – conduction band and valence band separated by a smaller energy gap. At normal temperatures, a significant number of electrons are thermally excited from valence band to conduction band. An equal number of vacancies are produced in the valence band. These vacancies are treated as particles having a mass equal to that of electron and carry positive charge. They are called **holes**.

The Fermi-Dirac distribution function is applicable to a semiconductor. Fig. 29.18 depicts the probability function plotted alongside the band diagram for a semiconductor. Because the probability of electron occupancy of the conduction

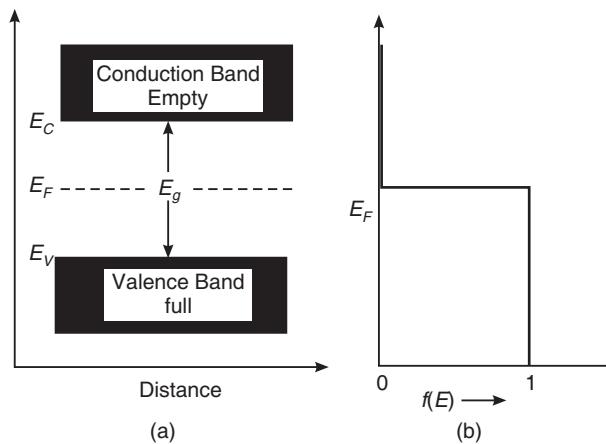


Fig. 29.17

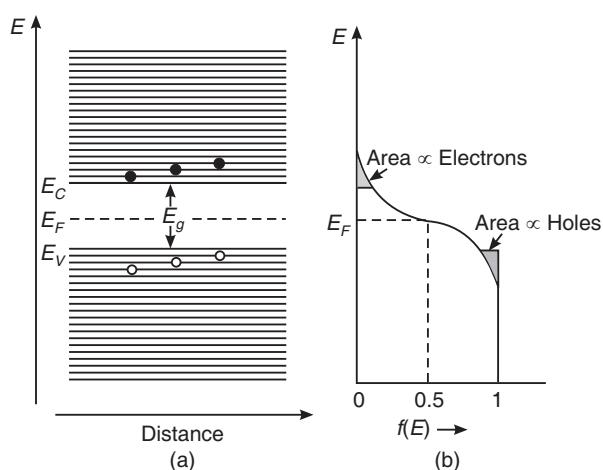


Fig. 29.18

band increases at temperatures greater than 0K, the probability function is blurred and tapers off towards higher energy in the conduction band. Similarly, the probability of hole occupancy of the valence band increases, and the probability curve is blurred near the top edge of the valence band. The extent of blurring of blurring of probability curve in both the bands is equal indicating that the concentration of electrons in the conduction band and that of holes in the valence band are equal. Secondly, the probability function $f(E)$ rapidly approaches zero value with increasing E . It suggests that the electrons in the conduction band are clustered very close to the bottom edge of the band. In a similar way, the holes are grouped very close to the top edge of the valence band. Therefore, it may be approximated that electrons are located right at the bottom edge of the conduction band whereas the holes are at the top edge of the valence band.

The Fermi level represents the average energy of charge carriers participating in conduction. Both electrons and holes participate in conduction in semiconductor and they lie in two different bands separated by a forbidden gap. Therefore, it is expected that the Fermi level lies in the middle of the forbidden gap. If the Fermi level is located elsewhere, it would mean that the number of electrons in the conduction band would be different from the number of holes in the valence band. It, in turn, would imply that the material does not exhibit overall neutrality which is not at all true.

29.12 EFFECTIVE MASS

We generally assume that the mass of an electron in a solid is the same as the mass of a free electron. However, experimentally measured values indicate that in some solids the electron mass is larger while for others it is slightly smaller than the free electron mass. The experimentally determined electron mass is called the **effective mass m^*** . The cause

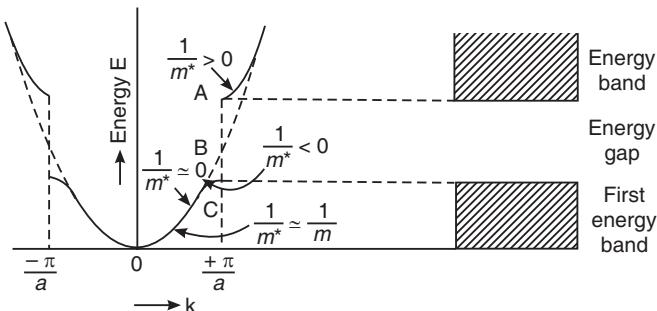


Fig. 29.19. The effective mass of electron in a solid depends on electron's location in the band.

for the deviation of the effective mass from the free electron mass is due to the interactions between the drifting electrons and the atoms in a solid. It has been found that the effective mass is inversely proportional to the curvature of an allowed energy band. It means that the effective mass depends on the location of an electron in the allowed energy band (Fig. 29.19). Considering the electron as a wave packet having a group velocity, v_g , an expression for the effective mass is derived as

$$m^* = \frac{\hbar^2}{d^2 E / dk^2} \quad (29.10)$$

- (i) Near the bottom of the band, the form of the $E - k$ curve does not differ much from the curve for free electron. Therefore, in these regions $m^* \approx m$.
- (ii) At the point of inflection B, the derivative $d^2E/dk^2 = 0$. Therefore, in these regions $m^* \approx \infty$. It means that an external field cannot exert any action on the motion of the electron in the region.

(iii) Near the top of the allowed band, C the derivative $d^2E/dk^2 < 0$. Therefore, the effective mass m^* of the electrons occupying levels near the top of the band is negative.

The concept of effective mass provides a satisfactory description of the charge carriers in crystals. In crystals such as alkali metals, which have partially filled energy band, the conduction takes place mainly through electrons. However, in crystals for which the energy band is nearly full, the negative charge and negative mass vacancies may be considered as positive charge and positive mass particles called *holes*. It explains the origin of positive Hall coefficient in certain metals such as zinc.

QUESTIONS

1. Define energy level and energy band. Explain with proper diagrams, how on the basis of band theory, solids are classified as conductors, insulators and semiconductors.
(C.S.V.T.U., 2005, 2007, 2009)
2. Describe the formation of energy bands in a crystalline solid.
Define valence band, conduction band and forbidden gap in the energy band structure.
Hence classify solids into conductors, semiconductors and insulators.
(Bombay Univ.)
3. Explain formation of energy bands in solids on the basis of band theory of solids.
(R.T.M.N.U., 2007)
4. Explain the 'Kronig-Penny' model of solids and show that it leads to energy band structure of solids.
(RGPV, 2010)
5. (a) Discuss with suitable mathematical expressions, the motion of an electron in a periodic potential.
(b) Explain how the above theory leads to the concept of band structure of solids.
(c) What is effective mass of electron ?
(JNTU, 2010)
6. Explain quantitatively band theory of solids. Explain energy band diagram and distinguish metal, semiconductor and insulator on the basis of above theory.
(RGPV, 2008)
7. Explain how solids are classified on the basis of energy band gap.
(Calicut Univ., 2005)
8. Describe in short the formation of energy bands in solids and hence explain how it helps to classify the materials into conductors, semiconductors and insulators (with an example in each).
(C.S.V.T.U., 2008)
9. Explain how the materials are classified into conductors, semiconductors and insulators with the help of energy band diagrams.
(G.T.U., 2009)
10. Explain the formation of energy bands in solids and briefly explain how solids are classified on the basis of energy band gap.
(Calicut Univ., 2006)
11. According to band theory, a completely filled or empty band is not associated with electrical conduction. Only partially filled band is responsible for electrical conduction. Explain, why?
(R.T.M.N.U., 2007)
12. How does the band theory differ from the free electron model in explaining the properties of metals?
13. Explain energy band diagram of silicon showing a graph of variation of potential energy with distance. Explain semi-conducting nature of silicon. With similar band structure why is diamond insulator?
14. Draw a graph showing variation of electron energy in germanium crystals as a function of interatomic distance and explain why it shows semiconducting behaviour?

15. Draw a graph showing variation of electron energy levels of germanium as a function of its interatomic distance. Explain from it why germanium is an insulator at 0°K and semiconductor at 7°K. **(R.T.M.N.U., 2007)**
16. What is Fermi level and Fermi energy?
17. Explain Fermi energy function. How does it vary with temperature? **(RGPV, 2008)**
18. Write and explain Fermi function. Explain with the help of a diagram how it varies with change of temperature.
19. Write down the Fermi-Dirac equation for the probability of occupation of an energy level E by an electron. Show that the probability of its occupancy by an electron is zero if $E > E_F$ and unity if $E < E_F$ at temperature 0K.
20. Define Fermi distribution function. Show that at all temperatures ($T > 0$ K) probability of occupancy of Fermi level is 50%.
21. What is Fermi function? Draw a graph showing its variation with energy at different temperatures and discuss it.
22. Write down Fermi distribution function $f(E)$. Show graphically and analytically that $f(E)$ as function of E always passes through a point $\left(E_F, \frac{1}{2}\right)$ at different temperatures.
23. Why is that only the electrons near the Fermi level contribute to electrical conductivity?
24. Explain the concept of hole. **(R.T.M.N.U., 2006)**
25. What is meant by effective mass of electron?
26. Explain the concept of negative mass on the basis of band theory.
27. Write down an expression for the probability of occupancy of a particular energy state of an electron in an intrinsic semiconductor. Represent it graphically at 0° K and at room temperature.
28. Explain in brief the concept of Fermi level. Show diagrammatically the Fermi level in metals, intrinsic semiconductors and insulators at 0° K and at higher temperature. **(R.T.M.N.U., 2006)**
29. How are the band structures of insulators and semiconductors similar? How are they different? **(R.T.M.N.U., 2006)**

PROBLEMS

1. In a solid, consider the energy level lying 0.01 eV above Fermi level. What is the probability of this level being *occupied* by an electron at 200 K? **[Ans: 0.359]**
2. In a solid, consider the energy level lying 0.01 eV below Fermi level. What is the probability of this level being occupied by an electron at 300 K? **[Ans: 0.595]**
3. In a solid, consider the energy level lying 0.01 eV above Fermi level. What is the probability of this level being *occupied* by an electron at 300 K? **[Ans: 0.405]**
4. In a solid, consider the energy level lying 0.01 eV above Fermi level. What is the probability of this level being *not occupied* by an electron at 300 K? **[Ans: 0.595]**

CHAPTER

30

Semiconductors

30.1 INTRODUCTION

Semiconductors are materials having electrical conductivity considerably greater than that of an insulator but significantly lower than that of a conductor. Of all the elements in the periodic table, eleven elements are semiconductors. Germanium and silicon are the most widely used semiconductors in device manufacturing applications. They are known as **elemental semiconductors**. Besides these, there are certain **compound semiconductors** such as Gallium Arsenide, Indium Phosphide etc which are formed from the combinations of the elements of groups III and V or groups II and VI. The unique and interesting feature of semiconductors is that they are bipolar and two charge carriers, namely electrons and holes, transport current in these materials. The electrical conductivity of a *pure semiconductor*, known as intrinsic conductivity, is significantly low and is drastically influenced by temperature. As such pure semiconductors cannot be used in device fabrication. Through the technique of doping, the conductivity of a semiconductor can be increased in magnitude to a desired value and can be made independent of temperature in a certain temperature interval. Doped semiconductors are known as *extrinsic semiconductors*. The remarkable feature of extrinsic semiconductors is that current is transported in them by two different charge carriers, *electrons* and *holes*; and through two different processes, *drift* and *diffusion*. Extrinsic semiconductors are widely used in fabrication of solid-state devices. An understanding of the mechanism of conduction in intrinsic and extrinsic semiconductors helps us understand the working of solid-state devices.

30.2 CRYSTAL STRUCTURE

Silicon and germanium are elemental semiconductors. They belong to Group IVA in the periodic table. Silicon atom has 14 electrons and germanium atom 32 electrons. Each of them has four valence electrons which are distributed among the outermost *s* and *p* orbitals. In case of silicon atom, there are four energy levels, $3s$, $3p_x$, $3p_y$, and $3p_z$, out of which the lower two levels are filled by four valence electrons. In an actual piece of silicon crystal, the $3s$ and $3p$ orbitals of the atom combine to yield $3sp^3$ hybrid molecular orbitals. The $3sp^3$ orbitals form four covalent bonds of equal angular separation, leading to a tetrahedral arrangement of atoms in space (Fig. 30.1 *a*). Atoms occupy the corners of a regular tetrahedron and each atom in turn bonds with four neighbours and so on. The resulting crystal structure is known as **diamond cubic (DC) crystal structure**, shown in Fig. 30.1 *b*.

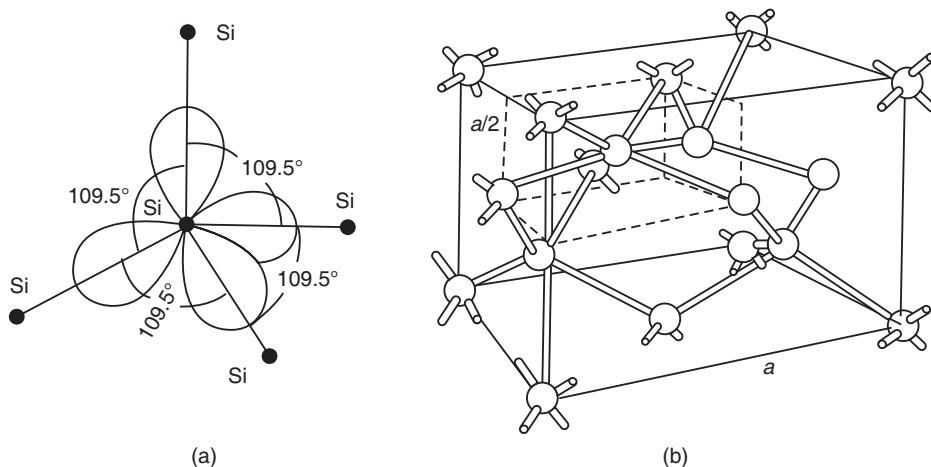


Fig. 30.1

30.3 INTRINSIC SEMICONDUCTOR

Chemically *pure* semiconductors are known as **intrinsic semiconductors**. A semiconductor is considered to be pure when there is less than one impurity atom in a billion host atoms.

A two-dimensional representation of silicon crystal is shown in Fig. 30.2a. Each silicon atom forms covalent bonds with four surrounding atoms. The shaded circles in Fig. 30.2a, represent the cores of the silicon atoms. The four valence electrons are shown by the small black dots surrounding each hatched circle. The probability of valence electrons being in any place between the bonding atoms is indicated by the dashed curves. In terms of energy band diagram, a conduction band and a valence band separated by a smaller energy gap characterize a semiconductor (Fig. 30.2b).

In a real crystal, the concentration of atoms N is given by

$$N = \frac{N_A \rho}{M} \quad (30.1)$$

where N_A is the Avogadro number, ρ the density and M the atomic weight of the material.

Using the data for silicon into equ. (30.1), we obtain

$$N = \frac{(6.02 \times 10^{26} \text{ atoms/k.mol})(2330 \text{ kg/m}^3)}{28.09 \text{ kg / k.mol}} \\ N = 5 \times 10^{28} \text{ atoms/m}^3 \quad (30.2)$$

The valence and conduction bands of silicon crystal contain $2N$ energy levels each. Therefore, the number of energy levels in each band is 10^{29} levels/ m^3 . In other words, there are 2×10^{29} states/ m^3 . The number of valence electrons available in the silicon crystal is $4N = 2 \times 10^{29}$ electrons/ m^3 . These electrons occupy the valence band and leave the conduction band vacant.

(a) At 0K an Intrinsic Semiconductor Behaves as a Perfect Insulator

At 0K and temperatures close to 0K, all valence electrons are locked in covalent bonds (Fig. 30.2 a) and spend most of the time between neighbouring atoms. Since all the valence electrons are engaged in covalent bonds, the bonds are complete. The energy available at 0K is not sufficient to break the covalent bonds.

Therefore, there are no free electrons within the material at absolute zero. Consequently, the semiconductor at 0K cannot conduct electricity and acts as a perfect insulator.

In terms of energy band diagram, the situation is as follows. There are $2N = 10^{29}$ energy levels/m³ in the valence band and $2N = 10^{29}$ energy levels/m³ in the conduction band. The total number of valence band electrons available in the crystal is $4N = 2 \times 10^{29}$ electrons/m³. At 0K these 2×10^{29} electrons/m³ completely occupy the 10^{29} energy levels/m³ available in the valence band. There are no electrons left to go into the conduction band (Fig. 30.2b). At 0K, electrons in the valence band do not possess enough energy to jump into the conduction band. As free electrons do not exist in the conduction band, an externally applied electric field cannot cause flow of current through the crystal. Hence, the intrinsic semiconductor behaves as a *perfect insulator* at 0K.

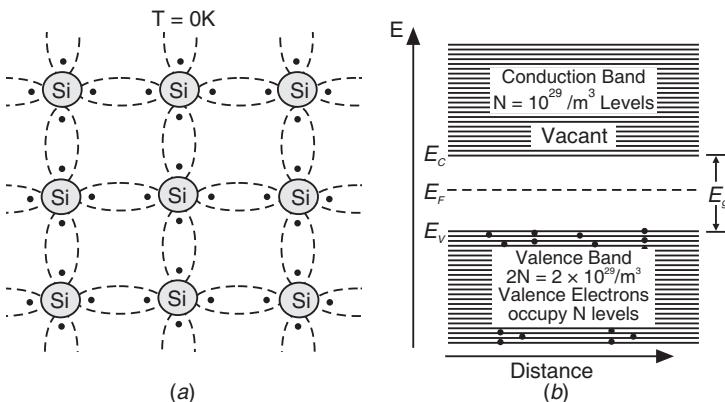


Fig. 30.2: Silicon Crystal at 0K

(b) Mechanism of Conduction in an Intrinsic Semiconductor

At temperatures above absolute zero, the finite thermal energy causes each atom in the crystal to vibrate about its mean position. When the vibrations become violent, some of the electrons acquire sufficient energy and break away from covalent bonds (Fig. 30.3a). Whenever a covalent bond is ruptured by thermal energy, a valence electron becomes free. The higher the temperature, the more covalent bonds are broken. The electrons liberated from bonds move randomly in the void spaces between the atoms in the crystal. If an electric field is applied, these free electrons cause electrical conduction.

From the energy band point of view, it means that some of the electrons in the valence band convert part of their thermal energy into potential energy. Those electrons which acquire energy equal to or in excess of the band gap

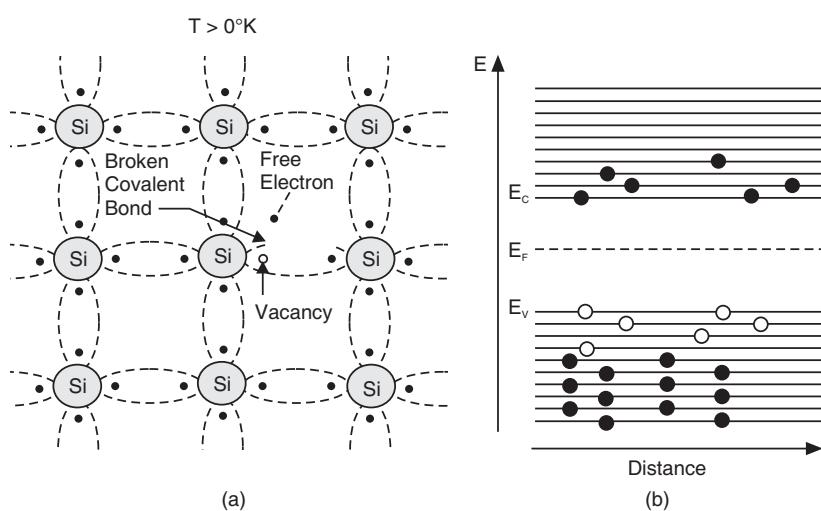


Fig. 30.3: Silicon Crystal at $T > 0\text{K}$

energy E_g are excited to the conduction band (Fig. 30.3b). Thus, the band gap energy E_g is the minimum amount of energy required to excite an electron from valence band to conduction band. E_g is characteristic of the material. The number of electrons excited to the conduction

band depends on the amount of thermal energy received by the crystal. For example, the concentration of broken bonds in a silicon crystal at 300K is $1.5 \times 10^{16} /m^3$. Thus, there are 1.5×10^{16} electrons/m³ in the conduction band at 300K. In fact, the conduction band can accommodate 2×10^{29} electrons/m³ and hence it is partially filled.

When an electron from the valence band jumps to the conduction band, an **empty state** (quantum vacancy) arises in the valence band. In silicon crystal at 300K, 1.5×10^{16} vacant states/m³ appear in the valence band, which are very small in number compared to the number of electrons remained in the band. Thus, now both the bands are partially filled. The electrons in the conduction band and the electrons in the valence band can be excited to upper vacant levels within the respective bands. Therefore, if an electric field is applied, these electrons can move into higher vacant levels and current flows in the crystal at ordinary temperatures. The motion of valence electrons in the valence band is customarily described in terms of a fictitious particle called **hole** which is bequeathed with a positive charge $+e$ and a mass m_h equal to that of an electron.

In pure semiconductors all available charge carriers, electrons and holes, arise due to thermally ruptured bonds and these thermally generated electron-hole pairs cause electrical conduction. Thermal generation is an *intrinsic process*. Therefore, we may define an intrinsic semiconductor as follows.

Definition: An *intrinsic semiconductor* is a semiconductor crystal in which electrical conduction arises due to **thermally excited electrons and holes**.

Significance of Band Gap E_g

The band gap energy E_g is the minimum amount of energy required for breaking a covalent bond. It is the minimum amount of energy required to excite an electron from valence band to conduction band. Also, it is the minimum amount of energy required to convert a bound electron into a free electron. The energy required to break a covalent bond in a germanium crystal is about 0.72 eV at 300K and that in silicon is 1.12 eV.

30.4 CORRELATION BETWEEN CRYSTAL LATTICE AND ENERGY BAND DESCRIPTIONS

The correlation between the crystal lattice description and the energy band description of a semiconductor can be summed up as follows:

The statement that an electron is in the valence band means that the electron is participating in a covalent band. The excitation of the electron from the valence band to the conduction band implies rupturing of a covalent bond. Accordingly, the band gap E_g may be regarded as representing the strength of the covalent bond. When it

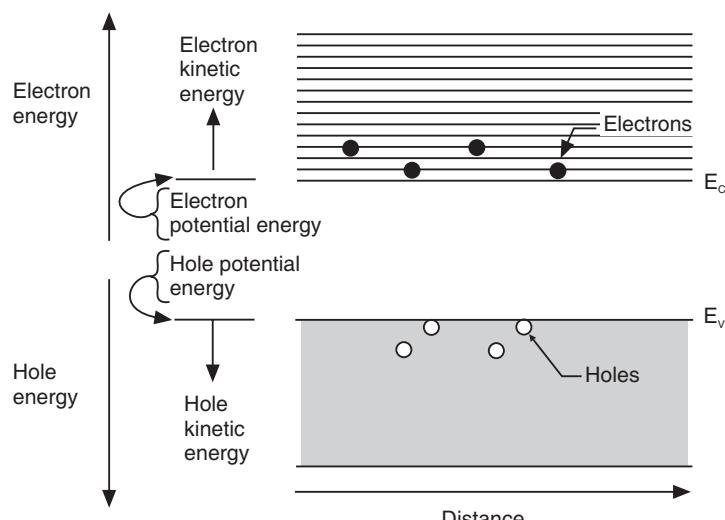


Fig. 30.4

is said that the electron is in the conduction band, it means that the electron is wandering haphazardly in the interstitial spaces of the crystal. The electrons in the conduction band are equally attracted to all nuclei and hence they experience a constant potential in the crystal. The magnitude of the potential energy may be taken as E_g . It is the energy corresponding to the lowest energy level in the conduction band.

Out of the thermal energy E supplied to a valence electron, an amount E_g is translated into the potential energy and the electron reaches the bottom edge E_C of the conduction band (Fig. 30.4). Any additional energy goes into the form of kinetic energy. Thus,

$$E = E_g + \frac{1}{2}mv^2 \quad (30.3)$$

Consequently, the electrons in the conduction band are characterized by their velocity only and act as **free electrons**. It is seen from the relation (30.3) that an electron with an energy E_g just reaches the bottom edge of the conduction band and will be at rest since $v = 0$. With more energy, it is positioned higher and higher above the conduction band edge E_C . The situation is very much akin to the billiard balls on a billiard table. The balls are confined to the table but can have wide ranging velocities.

The situation is similar for holes. The top of the valence band, E_V represents the potential energy of the hole (Fig. 30.4). The energies below E_V indicate the increasing kinetic energy of a hole. The directions of increasing energy for electrons and holes are opposite to each other because they have charges of opposite sign.

30.5 HOLES

In a silicon crystal, when a covalent bond is broken and an electron is set free, there arises a **quantum vacancy** at the site of the broken bond. The removal of a negative charge from an otherwise neutral atmosphere accords an effective positive charge $+e$ to the resulting vacancy. Obviously, the actual seat of the positive charge is the nucleus of the silicon atom which is deprived of its valence electron. The positively charged vacancy, say, created at the site A can attract an electron from the adjacent bonds. An electron from site, say, B jumps into the vacancy at A and bridges the bond but leaves a vacancy at B and so on. In this way, a new type of electron motion is created in the valence band. In quantum mechanical terms, the electron at a lower energy level makes an upward transition to a vacant level. As the electron moved up, the vacancy moves down. Note that the vacancy and the electron move in opposite directions in the valence band. As both motions are related, the hopping of a large number of valence electrons can be equivalently described in terms of normal motion of a small number of holes. The concept of hole helps in differentiating the two different kinds of electron motion, namely the motion of high energy electrons in the conduction band and the motion of low energy electrons in the valence band.

The concept of holes simplifies the understanding of electrical conduction in semiconductors. It is seen as follows:

The instantaneous current produced by one electron moving at a speed v_k is given by

$$i = -ev_k$$

The current set up by the motion of many electrons in a valence band is therefore

$$I = -e \sum_i v_i$$

For a completely filled band, the resultant current is zero. That is,

$$I = -e \sum_i v_i = 0 \quad (30.4)$$

The above equation may be rewritten as

$$I = -e \left[\sum_{i \neq k} v_i + v_k \right] = 0$$

$$\therefore -e \sum_{i \neq k} v_i = -e(-v_k) = ev_k \quad (30.5)$$

$$\therefore \sum_{i \neq k} v_i = -v_k \quad (30.6)$$

The above relation (30.6) implies that if the k^{th} electron is absent in the valence band, then the sum of the velocities of the remaining electrons is $-v_k$. The relation (30.5) shows that these electrons will set up a current equal to $+ev_k$. It means that the current set up by all the electrons in the valence band having one vacancy is equivalent to the current resulting from the motion of a single particle with a positive charge $+e$ that occupies the vacancy.

If a vacancy is to be given the status of a particle, a mass is also to be ascribed to it. We can arrive at the mass from the band theory. In solids, electrons experience forces due to atoms and other electrons. The motion of electron in a solid subjected to an electric field may be written in the form

$$F = m_e^* a$$

where m_e^* is the effective mass of the electron. The effective mass of an electron heavily depends on its location in the energy band, as discussed in Art. 29.12 of Chapter 29. Electrons near the bottom of the conduction band have an effective mass which is nearly identical to the mass of a free electron. Electrons near the top of valence band have negative effective mass. The concept of negative effective mass is not easy to grasp. The removal of an electron with a negative effective mass is identical to creating in its place a particle of positive effective mass.

Thus, a hole is assigned an effective mass, m_h^* .

A hole in the sea of valence electrons in the valence band is analogous to a bubble in a liquid. A bubble has shape, size and also a velocity. Its behaviour is the same as that of a particle. In reality, it is the absence of liquid. One can in principle describe the motion of a bubble by describing the motion of the surrounding liquid; but it will be a tedious task. In a similar way, the motion of the valence electrons is difficult to be described. On the other hand, the description of their motion in terms of holes is a far simpler task.

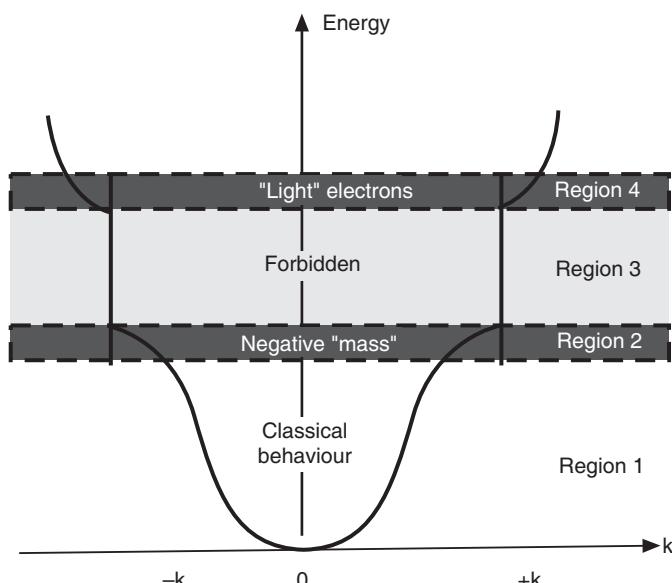
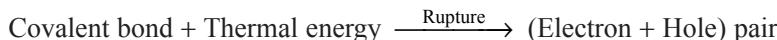


Fig. 30.5

Direct evidence for the existence of holes furnished by the results of Hall effect measurements justifies the use of parlance of holes.

30.6 GENERATION AND RECOMBINATION

In semiconductors a single event of covalent bond breaking leads to the generation of two charge carriers, an electron in the conduction band and a hole in the valence band (Fig. 30.3). The electron and hole are produced simultaneously as a pair and the process is called **electron-hole pair generation**. The process may be represented as



In the process of generation, a covalent bond is broken and a bound electron is transformed into a free electron. Thermal energy is one of the agents which causes pair generation. Another agent is optical illumination. At any temperature T , the number of electrons generated would be equal to the number of holes produced. If n denotes the concentration of electrons in the conduction band and p is the concentration of holes in the valence band, then

$$n = p \quad (30.7)$$

After generation the charge carriers move independently. The electrons move in the conduction band and the holes move in the valence band. Their motion is at random in the respective bands, as long as external electric field is not applied.

It is likely that the electron in conduction band may lose its energy due to collision with other particles in the lattice and fall into the valence band (Fig. 30.6). When a free electron falls into valence band, it merges with a hole. This process is called *recombination*. When a recombination event occurs, the free electron enters a ruptured covalent bond and re-bridges it.



Therefore, recombination means that a free electron transforms into a valence electron and that a ruptured covalent bond is re-bridged. In the process the electron-hole pair disappears and energy is released. The released energy is mainly in the form of thermal energy.

At a steady temperature a dynamic equilibrium exists which balances the two processes of electron-hole pair generation and electron-hole recombination (Fig. 30.6).

Just as thermal energy generates electron-hole pairs, light radiation can produce electron-hole pairs in a semiconductor. When a semiconductor material is irradiated with optical radiation, electron-hole pairs are generated if the frequency v of the radiation satisfies the condition

$$hv \geq E_g \quad (30.8)$$

When the radiation is switched off the excess electron-hole pairs recombine and the material returns to thermal equilibrium. Optically generated electron-hole pairs are responsible for the working of LDRs, photodiodes, etc optical detectors.

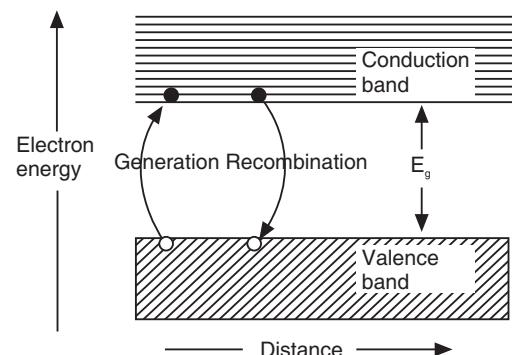


Fig. 30.6: Generation and recombination of electron-hole pairs

30.7 INTRINSIC CONDUCTIVITY

A single event of bond breaking in a pure semiconductor leads to generation of an **electron-hole** pair. At any temperature T , the number of electrons generated will be equal to the number of holes generated per unit volume. As the two charge carrier concentrations are equal, they are denoted by a common symbol n_i , which is called *intrinsic density* or *intrinsic concentration*. Thus,

$$n = p = n_i \quad (30.9)$$

In a semiconductor under thermal equilibrium condition, free electrons move in the conduction band and holes in the valence band, which are in a state of random motion. At a temperature T , they possess an average kinetic energy given by

$$\frac{1}{2}mv_{th}^2 = \frac{3}{2}kT$$

where v_{th} is the mean **thermal velocity**. When a potential difference is applied across the solid, the equilibrium condition is disturbed. The electric field accelerates the electrons and holes but their motion is hindered due to interactions with the lattice vibrations. In the steady state condition there arises a net movement of electrons in a direction opposite to that of the electric field and movement of holes in the direction of the electric field. This net movement of electrons and holes is called **drift**, and the corresponding mean velocity is known as **drift velocity**, v_d . The drift motion is superposed on the random thermal motion of the charge carriers. The drift motion is directional and causes **drift current** flow, which is more often called **conduction current**.

The drift velocity is given by

$$v_d = \mu E \quad (30.10)$$

Since the electrons move in relatively less populated conduction band the properties such as mobility, conductivity etc of electrons are larger compared to those of holes as the latter move in nearly full valence band. We designate the drift velocity of electron with v_{de} and that of hole with v_{dh} . Similarly, we denote the mobility of electron and hole with μ_e and μ_h respectively. Then, the current density due to electrons is given by

$$J_e = nev_{de} = ne\mu_e E \quad (30.11)$$

and the current density due to holes is

$$J_h = pev_{dh} = pe\mu_h E \quad (30.12)$$

Comparing the above expressions with Ohm's law $J = \sigma E$, we obtain the expressions for electronic and hole conductivities as follows.

$$\sigma_e = ne\mu_e \quad (30.13)$$

and

$$\sigma_h = pe\mu_h \quad (30.14)$$

Let us now consider a sample of semiconductor across which a potential difference V is applied. The potential difference V establishes an electric field E in the semiconductor. It causes a current I_e due to electrons drifting in the conduction band and a current I_h due to holes drifting in the valence band (see Fig. 30.7). The total current through the semiconductor is

$$I = I_e + I_h \quad (30.15)$$

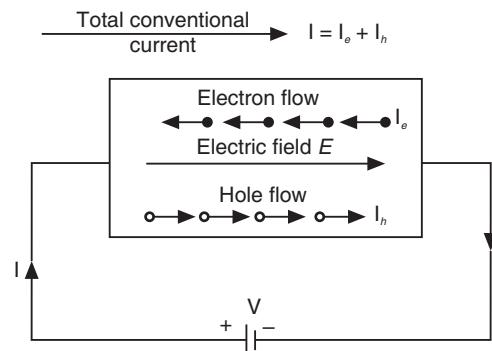


Fig. 30.7: Components of drift current due to electron and hole motion in an intrinsic semiconductor

\therefore Current density $J = \frac{I_e}{A} + \frac{I_h}{A}$ where A is cross-sectional area of the crystal.

or

$$\begin{aligned} J &= J_e + J_h \\ &= (ne\mu_e + pe\mu_h)E \end{aligned} \quad (30.16)$$

Therefore, the intrinsic conductivity is given by

$$\sigma = (ne\mu_e + pe\mu_h) \quad (30.17)$$

As $n = p = n_i$ in an intrinsic semiconductor, equ. (30.17) may be rewritten as

$$\sigma = en_i(\mu_e + \mu_h) \quad (30.18)$$

The equation (30.18) does not explain the temperature dependence of electrical conductivity in semiconductors. In general the variation of mobility with temperature is too small and the large variation in electrical conductivity in semiconductors is linked to the variation of electron concentration with temperature. Thus, as a first approximation, taking e , μ_e and μ_h as constants, we find that

$$\sigma(T) \propto n_i(T) \quad (30.19)$$

The expression for the intrinsic carrier concentration is derived in the band theory. Here, we take the final results from the band theory and use them to obtain an expression for $n(T)$, the variation of intrinsic carrier concentration with temperature.

30.8 CARRIER CONCENTRATIONS

With an increase in temperature covalent bonds are broken in an intrinsic semiconductor and electron-hole pairs are generated. We expect that a large number of electrons can be found in the conduction band and similarly, a large number of holes in the valence band. As electrons and holes are charged particles, they are together called **charge carriers**. **Carrier concentration** is the number of electrons in the conduction band per unit volume (n) and the number of holes in the valence band per unit volume (p) of the material. Carrier concentration is also known as the **density of charge carriers**. We would like to calculate the electron concentration, n , in the conduction band and the hole concentration, p , in the valence band.

30.8.1 Calculation of Electron Density

Let dn be the number of electrons whose energy lies in the energy interval E and $E + dE$ in the conduction band. Then,

$$dn = Z(E)f(E)dE \quad (30.20)$$

where $Z(E)$ dE is the density of states in the energy interval E and $E + dE$ and $f(E)$ probability that a state of energy is occupied by an electron.

The electron density in the conduction band is given by integrating the above equation between the limits E_C and ∞ . E_C is the energy corresponding to the bottom edge of the conduction band and ∞ the energy corresponding to the top edge of the conduction band. As the probability of electrons occupying upper levels of conduction band $f(E)$ readily approaches zero for higher energies, the upper limit, namely the top of conduction band is taken as ∞ . Thus,

$$n = \int_{E_C}^{\infty} Z(E)f(E)dE \quad (30.21)$$

The density of states in the conduction band is given by

$$Z(E)dE = \frac{4\pi}{h^3} (2m_e^*)^{3/2} E^{1/2} dE \quad \text{for } E > E_C \quad (30.22)$$

The bottom edge of the conduction band E_C corresponds to the potential energy of an electron at rest. Therefore, $(E - E_C)$ will be the kinetic energy of the conduction electron at higher energy levels. Hence, equ. (30.22) is to be modified as follows.

$$Z(E)dE = \frac{4\pi}{h^3} (2m_e^*)^{3/2} (E - E_C)^{1/2} dE \quad (30.23)$$

The probability of an electron occupying an energy level is given by

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]}$$

When the number of particles is very small compared to the available energy levels, the probability of an energy state being occupied by more than one electron is very small. Such a situation is valid when $(E - E_F) \gg 3kT$. Under this circumstance, the number of available states in the conduction band is far larger than the number of electrons in the band. Then, Fermi-Dirac function can be approximated to Boltzmann function,

$$f(E) = \exp[-(E - E_F)/kT]. \quad (30.24)$$

Using the equations (30.23) and (30.24) into (30.21), we obtain

$$\therefore n = \frac{4\pi}{h^3} (2m_e^*)^{3/2} \int_{E_C}^{\infty} (E - E_C)^{1/2} e^{-(E-E_F)/kT} dE \quad (30.25)$$

$$\text{or } n = \frac{4\pi}{h^3} (2m_e^*)^{3/2} e^{(E_F-E_C)/kT} \int_{E_C}^{\infty} (E - E_C)^{1/2} e^{-(E-E_C)/kT} dE \quad (30.26)$$

The integral in eq. (30.26) is of the standard form which has a solution of the following form.

$$\int_0^{\infty} x^{1/2} e^{-ax} dx = \frac{\sqrt{\pi}}{2a\sqrt{a}}$$

where $a = 1/kT$ and $x = (E - E_C)$.

$$\therefore n = \frac{4\pi}{h^3} (2m_e^*)^{3/2} e^{(E_F-E_C)/kT} \left[\frac{\sqrt{\pi}}{2} (kT)^{3/2} \right] \quad (30.27)$$

Rearranging the terms, we get

$$n = 2 \left[\frac{2\pi m_e^* k T}{h^2} \right]^{3/2} e^{-(E_C - E_F)/kT} \quad (30.28)$$

The above equation is the expression for the **electron concentration** in the conduction band of an intrinsic semiconductor.

$$\text{Designating } N_C = 2 \left[\frac{2\pi m_e^* k T}{h^2} \right]^{3/2} \quad (30.29)$$

in the above equation, we obtain

$$n = N_C e^{-(E_C - E_F)/kT} \quad (30.30)$$

N_C is a temperature dependent material constant known as the **effective density of states** in the conduction band. In silicon at 300 K, $N_C = 2.8 \times 10^{25}/\text{m}^3$.

The importance of the relation (30.30) is that it relates the equilibrium electron concentration to a single variable, namely the Fermi level E_F . Therefore, electron concentration is specified, if E_F is specified.

30.8.2 Calculation of Hole Density

Let dp be the number of holes whose energy lies in the energy interval E and $E + dE$ in the valence band. Then,

$$dp = Z(E)[1 - f(E)]dE \quad (30.31)$$

where $Z(E)dE$ is the density of states in the energy interval E and $E + dE$ and $[1 - f(E)]$ the probability that a state of energy is vacant and not occupied by an electron. If $f(E)$ is the probability for occupancy of an energy state at E by an electron, then the probability that the energy state is vacant is given by $[1 - f(E)]$. Since a hole represents a vacant state in valence band, the probability for occupancy of a state at E by a hole is equal to the probability of absence of electron at that state.

We can write

$$[1 - f(E)] = 1 - \frac{1}{1 + e^{(E-E_F)/kT}} = \frac{1}{1 + e^{(E_F-E)/kT}} \approx e^{-(E_F-E)/kT} \quad (30.32)$$

The density of states in the valence band is given by

$$Z(E)dE = \frac{4\pi}{h^3} (2m_h^*)^{3/2} E^{1/2} dE \quad (30.33)$$

The top edge of the valence band E_V corresponds to the potential energy of a hole at rest. Therefore, $(E_V - E)$ will be the kinetic energy of the hole at lower energy levels. Hence, equ. (30.23) is to be modified as follows.

$$Z(E)dE = \frac{4\pi}{h^3} (2m_h^*)^{3/2} (E_V - E)^{1/2} dE \quad (30.34)$$

The number of holes in the energy interval E and $E + dE$ is

$$dp = \frac{4\pi}{h^3} (2m_h^*)^{3/2} (E_V - E)^{1/2} e^{-(E_F-E)/kT} dE \quad (30.35)$$

In order to calculate the number of holes in the valence band equ. (30.35) is to be integrated between the limits $-\infty$ and E_V . The hole density in the valence band is therefore given by

$$p = \frac{4\pi}{h^3} (2m_h^*)^{3/2} \int_{-\infty}^{E_V} (E_V - E)^{1/2} e^{-(E_F-E)/kT} dE \quad (30.36)$$

$$= \frac{4\pi}{h^3} (2m_h^*)^{3/2} e^{-(E_F-E_V)/kT} \int_{-\infty}^{E_V} (E_V - E)^{1/2} e^{-(E_V-E)/kT} dE \quad (30.37)$$

The integral in equ. (30.37) is of the standard form which has a solution of the following form.

$$\int_0^\infty x^{1/2} e^{-ax} dx = \frac{\sqrt{\pi}}{2a\sqrt{a}}$$

where $a = 1/kT$ and $x = (E_V - E)$.

$$\therefore p = \frac{4\pi}{h^3} (2m_h^*)^{3/2} e^{-(E_F-E_V)/kT} \left[\frac{\sqrt{\pi}}{2} (kT)^{3/2} \right] \quad (30.38)$$

Rearranging the terms we get

$$\text{or } p = 2 \left[\frac{2\pi m_h^* k T}{h^2} \right]^{3/2} e^{-(E_F - E_V)/kT} \quad (30.39)$$

The above equation is the **expression for the hole concentration** in the valence band of an intrinsic semiconductor.

$$\text{Denoting } N_V = 2 \left[\frac{2\pi m_h^* k T}{h^2} \right]^{3/2} \quad (30.40)$$

$$\therefore p = N_V e^{-(E_F - E_V)/kT} \quad (30.41)$$

N_V is called the **effective density of states** in the valence band. For silicon at 300 K, $N_V = 10^{25}/m^3$. It is seen that $N_C/N_V = 2.8$. To a first approximation, N_C is taken to be equal to N_V .

The electron and hole concentrations in an intrinsic semiconductor are shown in Fig. 30.8.

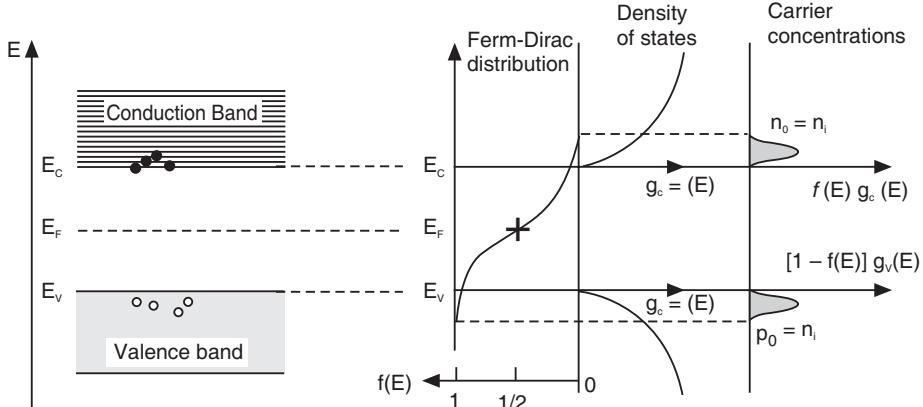


Fig. 30.8

30.9 INTRINSIC CARRIER CONCENTRATION

A single event of bond breaking in a pure semiconductor leads to generation of an **electron-hole** pair. At any temperature T , the number of electrons generated will be equal to the number of holes generated. As the two charge carrier concentrations are equal, they are denoted by a common symbol n_i , which is called *intrinsic density* or *intrinsic concentration*. Thus,

$$n = p = n_i$$

We can write that

$$n_i^2 = np \quad (30.42)$$

$$= (N_C e^{-(E_C - E_F)/kT})(N_V e^{-(E_F - E_V)/kT})$$

$$= (N_C N_V)^{-1} e^{-(E_C - E_V)/kT} \quad (30.43)$$

The term $(E_C - E_V)$ stands for the difference in energy between the top level of valence band and the bottom level of conduction band. Thus, it represents the separation between the valence and conduction bands, which is the band gap E_g .

$$(E_C - E_V) = E_g \quad (30.44)$$

$$\therefore n_i^2 = (N_C N_V) e^{-E_g/kT}$$

Substituting the values of N_C and N_V into the above equation, we obtain

$$\begin{aligned} &= 4 \left[\frac{2\pi kT}{h^2} \right]^3 \left(m_e^* m_h^* \right)^{3/2} e^{-E_g/kT} \\ \therefore n_i &= 2 \left[\frac{2\pi kT}{h^2} \right]^{3/2} \left(m_e^* m_h^* \right)^{3/4} e^{-E_g/2kT} \end{aligned} \quad (30.45)$$

This is the expression for intrinsic carrier concentration.

30.9.1 Variation of Intrinsic Carrier Concentration with Temperature

Eqn. (30.45) may be written as

$$n_i = 2 \left[\frac{2\pi k}{h^2} \right]^{3/2} \left(m_e^* m_h^* \right)^{3/4} T^{3/2} e^{-E_g/2kT} \quad (30.46)$$

The above relation shows that the free charge carrier concentration varies with temperature.

Eqn. (30.46) may be approximated to

$$n_i = 10^{21.7} T^{3/2} \times 10^{-2500 E_g/T} \quad (30.46a)$$

The following important points may be inferred from the relation (30.46):

- The intrinsic concentration is independent of Fermi level.
- The intrinsic concentration has an exponential dependence on the band gap value E_g .
- It strongly depends on the temperature.
- The factor 2 in the exponent indicates that two charge carriers are produced for one covalent bond broken.

The experimental value of n_i in silicon at room temperature is 1.5×10^{16} carriers/m³ and in germanium 2.5×10^{19} carriers/m³.

Example 30.1: The forbidden gap in pure silicon is 1.1 eV. Compare the number of conduction electrons at temperatures 37°C and 27°C.

Solution: Let n_1 be the number of conduction electrons at 27°C and n_2 at 37°C. Using eqn. (30.46 a), we get

$$n_1 = 10^{21.7} (300 \text{ K})^{3/2} (10^{-2500 \times 1.1 \text{ eV}/300 \text{ K}})$$

and

$$n_2 = 10^{21.7} (300 \text{ K})^{3/2} (10^{-2500 \times 1.1 \text{ eV}/310 \text{ K}})$$

$$\frac{n_2}{n_1} = \frac{(310 \text{ K})^{3/2} \times 10^{-8.87}}{(300 \text{ K})^{3/2} \times 10^{-9.2}} = 2.96$$

Thus the number of electrons in the conduction band increased nearly three-fold as the temperature of the material increased from 27°C to 37°C.

Example. 30.2: Compute the concentration of intrinsic charge carriers in a germanium crystal at 300 K. Given that $E_g = 0.72$ eV and assume $m_e^* = m_e$

Solution: Intrinsic charge carrier concentration, $n_i = 2 \left[\frac{2\pi m_e^* k T}{h^2} \right]^{3/2} \exp \left(-\frac{E_g}{2kT} \right)$

$$2 \left[\frac{2\pi m_e^* k}{h^2} \right]^{3/2} = 2 \left[\frac{2 \times 3.143 \times 9.11 \times 10^{-31} \text{ kg} \times 1.38 \times 10^{-23} \text{ J/K}}{\left(6.626 \times 10^{-34} \text{ J.s} \right)^2} \right]^{3/2}$$

$$= 4.83 \times 10^{21}, T^{3/2} = 300^{3/2} = 5196$$

and

$$\exp\left(-\frac{E_g}{2kT}\right) = \exp\left(-\frac{0.72 \text{ eV}}{2 \times 8.61 \times 10^{-5} \text{ eV/K} \times 300 \text{ K}}\right)$$

$$= \exp(-13.846) = 9.7 \times 10^{-7}$$

$$\therefore n_i = 4.83 \times 10^{21} \times 5196 \times 9.7 \times 10^{-7} = 3.4 \times 10^{19} \text{ m}^3.$$

30.10 THE FRACTION OF ELECTRONS IN THE CONDUCTION BAND

The Fermi-Dirac probability function gives the fractional occupancy of the energy states.

Thus, $f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]}$ gives the probability that an electron occupies the energy state E . The probability that an electron occupies the energy state E_C may then be found from

$$f(E_C) = \frac{1}{1 + \exp[(E_C - E_F)/kT]} \quad (30.47)$$

$$\text{But, according to the definition of probability, } f(E_C) = \frac{n}{N}$$

where n is the number of electrons excited to conduction band levels and n is the total number of electrons available in the valence band initially.

$$\therefore \frac{n}{N} = \frac{1}{1 + \exp[(E_C - E_F)/kT]} \quad (30.48)$$

As $(E_C - E_F) = E_g/2$, we write

$$\frac{n}{N} = \frac{1}{1 + \exp[E_g/2kT]}$$

Since $E_g > 2kT$, the factor unity may be neglected in comparison to the exponential term.

$$\therefore \frac{n}{N} = \frac{1}{\exp[E_g/2kT]} = e^{-E_g/2kT} \quad (30.49)$$

The above equation gives the fraction of electrons in the conduction band.

Example 30.3. What is the probability of an electron being thermally promoted to the conduction band in diamond at 27°C, if the bandgap is 5.6 eV wide?

Solution. The probability that an electron being thermally promoted to the conduction band is given by

$$\begin{aligned} f(E_C) &= \frac{1}{1 + \exp(E_g/2kT)} = \frac{1}{1 + \exp\left[\frac{5.6 \text{ eV}}{2(0.026) \text{ eV}}\right]} \\ &= \frac{1}{1 + e^{107.69}} = 1.7 \times 10^{-47} \end{aligned}$$

Example 30.4. Estimate the fraction of electrons in the conduction band at 300°K of

(i) Germanium ($E_g = 0.72 \text{ eV}$) (ii) Silicon ($E_g = 1.1 \text{ eV}$) and (iii) Diamond ($E_g = 5.6 \text{ eV}$)
What is the significance of these results?

Solution.

$$(i) \text{ Germanium: } f(E_C) = e^{-E_g/2kT} = e^{\frac{0.72 \text{ eV}}{2 \times 0.026 \text{ eV}}} = e^{-13.85} = 9.66 \times 10^{-7}$$

$$(ii) \text{ Silicon: } f(E_C) = e^{-E_g/2kT} = e^{\frac{1.1 \text{ eV}}{2 \times 0.026 \text{ eV}}} = 6.5 \times 10^{-10}$$

$$(iii) \text{ Diamond: } f(E_C) = e^{-E_g/2kT} = e^{\frac{5.6 \text{ eV}}{5 \times 0.026 \text{ eV}}} = e^{-107.7} = 1.7 \times 10^{-47}$$

The above results show that the larger the band gap the smaller the electrons that can go into the conduction band, at a given temperature.

Example 30.5: Assuming that the number of electrons near the top of the valence band available for thermal excitation is $5 \times 10^{25}/\text{m}^3$ and the intrinsic carrier density is $2.5 \times 10^{19}/\text{m}^3$, calculate the energy gap of germanium at room temperature.

Solution: Fraction of electrons in conduction band $\frac{n}{N} = \exp\left[-\frac{E_g}{2kT}\right]$

$$\therefore E_g = -2kT \ln\left(\frac{n}{N}\right)$$

$$= -2(8.61 \times 10^{-5} \text{ eV/K})(300 \text{ K}) \ln\left[\frac{2.5 \times 10^{19}/\text{m}^3}{5 \times 10^{25}/\text{m}^3}\right]$$

$$= 0.052 \times 14.509 \text{ eV} = 0.75 \text{ eV}.$$

Example 30.6. Estimate the fraction of electrons in conduction band at room temperature in Ge with $E_g = 0.72 \text{ eV}$ and in diamond with $E_g = 5.6 \text{ eV}$.

Solution. Fraction of electrons in conduction band $\frac{n}{N} = \exp\left[-\frac{E_g}{2kT}\right]$

(i) Fraction of electrons in conduction band in Ge,

$$\frac{n}{N} = \exp\left[-\frac{0.72 \text{ eV}}{2 \times 0.026 \text{ eV}}\right] = \exp(-13.846) = 9.7 \times 10^{-7}.$$

(ii) Fraction of electrons in conduction band in diamond,

$$\begin{aligned} \frac{n}{N} &= \exp\left[-\frac{5.6 \text{ eV}}{2 \times 0.026 \text{ eV}}\right] \\ &= \exp(-107.692) = 1.7 \times 10^{-47}. \end{aligned}$$

30.11 FERMI LEVEL IN INTRINSIC SEMICONDUCTOR

In a pure semiconductor, the electrons in the conduction band cluster very close to the bottom edge of the band, and we assume that electrons are located right at the bottom edge of the conduction band, as shown in Fig. 30.9. Similarly, we assume that the holes are at the top edge of the valence band. The electron concentration in the conduction band is given by

$$n = N_C e^{-(E_C - E_F)/kT}$$

The hole concentration in the valence band is given by

$$p = N_V e^{-(E_F - E_V)/kT}$$

In an intrinsic semiconductor, the electron and hole concentrations are equal. Thus, $n = p$

$$N_C e^{-(E_C - E_F)/kT} = N_V e^{-(E_F - E_V)/kT} \quad (30.50)$$

Taking logarithm on both sides, we get

$$\begin{aligned} \frac{(E_C - E_F)}{kT} &= \ln \frac{N_V}{N_C} - \frac{(E_F - E_V)}{kT} \\ -E_C + E_F &= kT \ln \frac{N_V}{N_C} - E_F + E_V \\ 2E_F &= (E_C + E_V) + kT \ln \frac{N_V}{N_C} \\ \therefore E_F &= \frac{E_C + E_V}{2} + \frac{1}{2} kT \ln \frac{N_V}{N_C} \end{aligned} \quad (30.51)$$

But $N_C = 2 \left[\frac{2\pi m_e^* kT}{h^2} \right]^{3/2}$ and $N_V = 2 \left[\frac{2\pi m_h^* kT}{h^2} \right]^{3/2}$

$$\begin{aligned} \therefore \frac{N_V}{N_C} &= \left(\frac{m_h^*}{m_e^*} \right)^{3/2} \\ \therefore \ln \left(\frac{N_V}{N_C} \right) &= \frac{3}{2} \ln \left(\frac{m_h^*}{m_e^*} \right) \\ \therefore E_F &= \frac{E_C + E_V}{2} + \frac{3}{4} kT \ln \left(\frac{m_h^*}{m_e^*} \right) \end{aligned} \quad (30.52)$$

We can also write the above equation as

$$E_F = \frac{E_C + E_V}{2} - \frac{3}{4} kT \ln \left(\frac{m_e^*}{m_h^*} \right) \quad (30.53)$$

If the effective mass of a free electron is assumed to be equal to the effective mass of a hole, i.e.,

$$\begin{aligned} m_h^* &= m_e^* \\ \ln \left(\frac{m_h^*}{m_e^*} \right) &= 0 \\ \therefore E_F &= \frac{E_C + E_V}{2} \end{aligned} \quad (30.54)$$

To make the meaning of the above equation more explicit, we write

$$E_F = \frac{E_C - E_V}{2} + E_V$$

$$E_F = \frac{E_g}{2} + E_V$$

If we denote the top of the valence band E_V as zero level, $E_V = 0$.

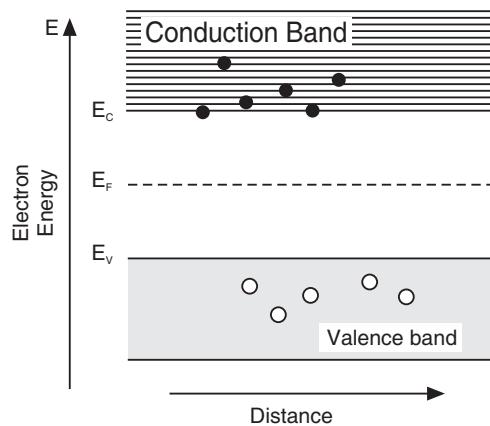


Fig. 30.9: Fermi level in intrinsic semiconductor

$$\therefore E_F = \frac{E_g}{2} \quad (30.55)$$

The above result shows that in an intrinsic semiconductor the Fermi level lies in the middle of the forbidden gap. We made the following assumptions in obtaining the relation (30.55).

- It was assumed that the electrons undergo transitions from the top edge level, E_V , of the valence band to the bottom edge level, E_C , of the conduction band. In reality, transitions are possible between the other levels also. However, the above result does not appreciably differ if other transitions are also taken into account.
- It was assumed that the effective mass of electrons in the conduction band, m_e^* , is exactly equal to the effective mass of the holes in the valence band, m_h^* . In practice, the effective masses differ from each other. However, the difference does not alter the above result significantly.

An important point to be noted here is that the **Fermi level is not an allowed energy level** in semiconductors. It only serves as a reference energy with reference to which we specify the energies of electrons and holes in a semiconductor.

30.11.1 Variation of Fermi Level with Temperature in an Intrinsic Semiconductor

With an increase in temperature, the Fermi level gets displaced upward to the bottom edge of the conduction band if $m_h^* > m_e^*$ or downward to the top edge of the valence band if $m_h^* < m_e^*$, as indicated in Fig. 30.10.

In most of the materials, the shift of Fermi level on account of $m_h^* \neq m_e^*$ is insignificant. The Fermi level in an intrinsic semiconductor may be considered as independent of temperature and as staying in the middle of the band gap.

Example 30.7: Determine the position of Fermi level in silicon semiconductor at 300 K. Given that the band gap is 1.12 eV, and $m_e^* = 0.12 \text{ m}$ and $m_h^* = 0.28 \text{ m}$.

$$\text{Solution: } E_F = \frac{E_g}{2} + \frac{3kT}{4} \ln\left(\frac{m_h^*}{m_e^*}\right)$$

$$= \frac{1.12 \text{ eV}}{2} + \frac{3 \times 8.61 \times 10^{-5} \text{ eV/K} \times 300 \text{ K}}{4} \ln\left(\frac{0.28 \text{ m}}{0.12 \text{ m}}\right)$$

$$= 0.56 \text{ eV} + (0.0194) \ln 2.333 \text{ eV}$$

$$= 0.56 \text{ eV} + (0.0194) (0.8473) \text{ eV}$$

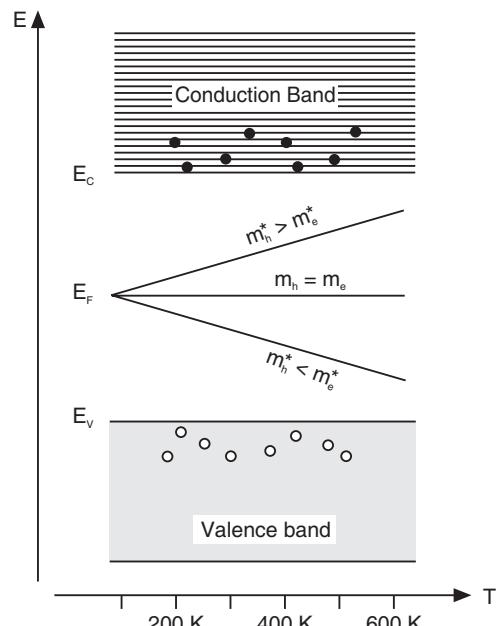


Fig. 30.10: Variation of Fermi level with temperature in an intrinsic semiconductor

$$= 0.56 \text{ eV} + 0.016 \text{ eV} = \mathbf{0.576 \text{ eV.}}$$

The Fermi level is 0.016 eV above the centre of the forbidden gap. In other words, it is at 0.576 eV from the top of the valence band.

Example 30.8: If the effective mass of an electron is equal to twice the effective mass of hole, determine the position of the Fermi level in an intrinsic semiconductor from the center of forbidden gap at room temperature.

Solution:

$$E_F = \frac{E_g}{2} + \frac{3}{4}kT \ln\left(\frac{m_h^*}{m_e^*}\right)$$

$$\frac{E_g}{2} - \frac{3}{4}kT \ln 2 = \frac{E_g}{2} - \frac{3}{4}(0.026 \text{ eV})(0.69)$$

or

$$E_F = \frac{E_g}{2} - 0.0135 \text{ eV}$$

The Fermi level is below the centre of the forbidden gap by 0.014 eV.

30.12 VARIATION OF INTRINSIC CONDUCTIVITY WITH TEMPERATURE

Incorporating the expression for n_i from equ. (30.43) into the equation (30.18), we get

$$\sigma = \sqrt{N_C N_V} e^{(\mu_e + \mu_h)} e^{-E_g/2kT}$$

$$\text{or } \sigma = \sigma_o e^{-E_g/2kT} \quad (30.56)$$

where σ_o is a constant.

The relation (30.56) gives the temperature dependence of conductivity of an intrinsic semiconductor, which is dominated by the exponential term $e^{-E_g/2kT}$.

Taking logarithm on both sides of eq. (30.56), we get

$$\ln \sigma_i = \ln \sigma_o - \frac{E_g}{2kT} \quad (30.57)$$

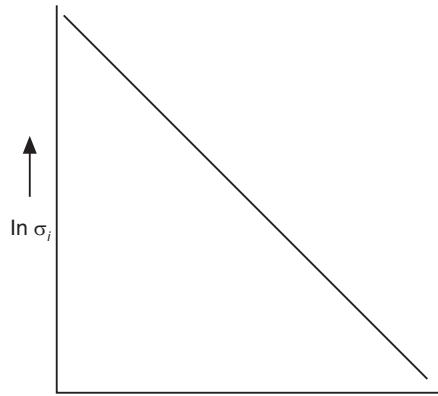


Fig. 30.11: Variation of intrinsic conductivity with temperature

Fig. 30.11 shows a plot of $\ln \sigma_i$ versus $1/T$. It shows that the conductivity increases with temperature.

Example 30.9: Find the resistivity of intrinsic germanium at 300°K. Given that the intrinsic density carriers is $2.5 \times 10^{19}/\text{m}^3$.

Solution:

$$\begin{aligned} \sigma_i &= en_i(\mu_e + \mu_h) \\ &= 1.602 \times 10^{-19} \text{ C} \times 2.5 \times 10^{19} \times (0.39 + 0.19) \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1} \\ &= 2.32 \text{ mho/m} \end{aligned}$$

$$\therefore \rho_i = \frac{1}{\sigma_i} = \mathbf{0.43 \text{ ohm-m.}}$$

30.13 DETERMINATION OF BAND GAP

Equ. (30.56) may be rewritten in terms of resistivity of the intrinsic semiconductor as follows:

$$\rho_i = \rho_o e^{E_g/2kT} \quad (30.58)$$

or

$$\frac{R_i A}{L} = \rho_o e^{E_g/2kT}$$

$$\therefore R_i = C e^{E_g/2kT} \quad (30.59)$$

Taking logarithms on both sides, we obtain

$$\ln R_i = \ln C + \frac{E_g}{2kT} \quad (30.60)$$

A plot of $\ln R_i$ versus $1/T$ gives a straight line, as shown in Fig. 30.12. The slope of the straight line gives the value of E_g . In practice we measure the resistance of intrinsic semiconductor at different temperatures with the help of four-point probe technique and draw a plot between $\ln R_i$ and $10^3/T$. The slope of the straight line thus obtained gives the value of $E_g/(2k)$. Hence,

$$E_g = m(2k) = \left(\frac{dy}{dx} \right) 2k \quad (30.61)$$

where $m \left(= \frac{dy}{dx} \right)$ is the slope of the

straight line.

Example 30.10: The resistivity of an intrinsic semiconductor is $4.5 \Omega \cdot m$ at $20^\circ C$ and $2.0 \Omega \cdot m$ at $32^\circ C$. Find the energy gap.

Solution:

$$\begin{aligned} E_g &= 2k \left(\frac{dy}{dx} \right) = 2k \left(\frac{\log \rho_2 - \log \rho_1}{(1/T_1 - 1/T_2)} \right) \\ &= 2 \times 8.61 \times 10^{-5} \text{ eV/K} \left(\frac{\log 2 - \log 4.5}{(1/293) - (1/305)} \right) \text{ K} \\ &= 2 \times 8.61 \times 10^{-5} \text{ eV} \left(\frac{0.6532 - 0.3010}{0.134 \times 10^{-3}} \right) \\ &= 2 \times 8.61 \times 10^{-5} \text{ eV} \times 2.63 \times 10^3 = \mathbf{0.45 \text{ eV}} \end{aligned}$$

Example 30.11. The resistivity of intrinsic silicon is $2.3 \times 10^3 \Omega \cdot m$ at 300K . Calculate its resistivity at $100^\circ C$. Assume $E_g = 1 \text{ eV}$ and $k = 1.38 \times 10^{-23} \text{ J/K}$.

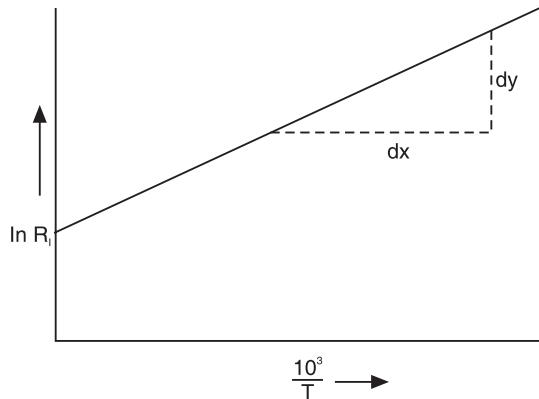


Fig. 30.12: Variation of resistance with temperature in intrinsic semiconductor

Solution. The resistivity of an intrinsic semiconductor is given by $\rho_i = \rho_o e^{E_g/2kT}$.
 \therefore The resistivity of the intrinsic semiconductor at 300 K is given by $\rho_1 = \rho_o e^{E_g/2k(300\text{ K})}$
and the resistivity of the semiconductor at 373 K is given by $\rho_2 = \rho_o e^{E_g/2k(373\text{ K})}$.

$$\begin{aligned}\therefore \frac{\rho_2}{\rho_1} &= \frac{\exp\left[\frac{1\text{eV}}{2 \times 8.61 \times 10^{-5}\text{eV} \times 373\text{K}}\right]}{\exp\left[\frac{1\text{eV}}{2 \times 8.61 \times 10^{-5}\text{eV} \times 300\text{K}}\right]} \\ \therefore \rho_2 &= \rho_1 \frac{\exp\left[\frac{1\text{eV}}{2 \times 8.61 \times 10^{-5}\text{eV} \times 373\text{K}}\right]}{\exp\left[\frac{1\text{eV}}{2 \times 8.61 \times 10^{-5}\text{eV} \times 300\text{K}}\right]} \\ &= 2.3 \times 10^3 \frac{\exp\left[\frac{1\text{eV}}{2 \times 8.61 \times 10^{-5}\text{eV} \times 373\text{K}}\right]}{\exp\left[\frac{1\text{eV}}{2 \times 8.61 \times 10^{-5}\text{eV} \times 300\text{K}}\right]} \Omega\text{.m} \\ &= 52 \Omega\text{.m.}\end{aligned}$$

30.14 LIMITATIONS OF INTRINSIC SEMICONDUCTOR

Intrinsic semiconductors are not useful for device manufacture because of low conductivity and the strong dependence of conductivity on temperature. If we take a crystal of pure silicon or germanium and connect it in the circuit, we will find that the current in the circuit will gradually increase as the temperature of the crystal is increased. We would also expect the current to increase if the voltage is increased and it does. But *increasing the voltage increases the current only proportionally, obeying Ohm's law, while increasing the temperature increases the current at an exponential rate*. Thus the temperature over which we have no control exerts more influence upon the current than the voltage, which we customarily do control.

We summarize the limitations as follows:

- Conductivity is low. Germanium has a conductivity of 1.67 S/m, which is nearly 10^7 times smaller than that of copper.
- Conductivity is a function of temperature and increases exponentially as the temperature increases.
- Conductivity cannot be controlled from outside.

30.15 EXTRINSIC SEMICONDUCTORS

A judicious introduction of impurity atoms in an otherwise perfect semiconductor crystal produces useful modifications of its electrical conductivity. It makes the current more voltage dependent than temperature dependent. An intentional introduction of controlled amount of impurity into an intrinsic semiconductor is called **doping**. The impurity added is called a **dopant**. A semiconductor doped with impurity atoms is called an **extrinsic semiconductor**. *The impurity-produced electrons are not temperature-dependent but are voltage-dependent and they will be under our control.*

One of the important methods of doping is to add precisely determined quantities of impurity to the *melt* from which the semiconductor crystal is grown. This way, a crystal

having a constant impurity density is obtained. Typical doping levels range from 10^{20} to 10^{27} impurity atoms/m³. Pentavalent elements from Group V or trivalent elements from Group III are used as dopants. The atoms belonging to these two groups are nearly of the same size as silicon or germanium atoms and easily substitute themselves in place of some of the host atoms in the semiconductor crystal. Thus, they are substitutional impurities and do not cause any distortion in the original crystal structure. Depending on the two different types of doping, two types of extrinsic semiconductors are possible. They are **n-type** and **p-type** semiconductors.

Common Dopant Elements for Silicon and Germanium

n-type	p-type
Phosphorous	Aluminum
Arsenic	Boron
Antimony	Gallium
	Indium

Advantages of Extrinsic Semiconductors

- Conductivity is high.
- Conductivity can be tailored to the desired value through the control of doping concentration.
- Conductivity is not a function of temperature.

30.16 n-TYPE SEMICONDUCTOR

An *n*-type semiconductor is produced when a pure semiconductor is doped with a pentavalent impurity such as phosphorous. A phosphorous atom has five valence electrons. Out of the five electrons, only four participate in bonding with four host silicon atoms while the fifth electron remains loosely bound. The host silicon lattice is a dielectric medium having a dielectric constant of 12. As a result, the Coulomb force between the phosphorous nucleus and the fifth electron is smaller than that it would be in free space. Therefore, the ionization energy of the fifth electron is very small. It is found to be 0.045 eV. The ionization energy is so small that the thermal energy can easily liberate the fifth electron from the nucleus. It means that the energy levels corresponding to phosphorous atoms are nearer to the bottom edge of the conduction band. At normal temperatures, the fifth electron becomes free to move about in the crystal and acts as a charge carrier. That is, the electron jumps into the conduction band leaving behind the **positive phosphorous ion** that is fixed in the crystal lattice. As the phosphorous atom is donating an electron for the purpose of electrical conduction, it is called a **donor atom**.

Energy Band Diagram

The energy band diagram of *n*-type semiconductor is shown in Fig. 30.13. If the donor atom density is low, the donor atoms are distantly spaced from one another, approximately, by 100

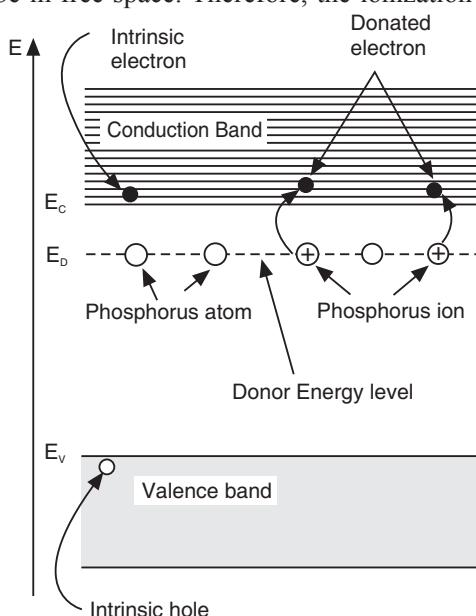


Fig. 30.13: Energy band diagram of n-type semiconductor

atom spacings. In such a situation, the donor atoms cannot interact with each other and their energy levels are discrete levels, E_D . They are called **donor levels** and represent the ground state of the fifth electron of impurity atom. As even small amount of thermal energy can readily liberate the fifth electron from the atom and send it into the conduction band, the donor levels are expected to be located very near to the bottom edge of the conduction band.

30.16.1 Temperature Variation of Carrier Concentration

The general dependence of electron concentration on temperature is shown in Fig. 30.14. At 0 K, the donor atoms are not ionized which means that all the donor electrons are bound to the donor atoms. The conduction band is empty, while the valence band is full. The material behaves essentially as an insulator.

At slightly elevated temperatures, the donor atoms are ionized and the donor electrons go into the conduction band (Fig. 30.14, ionization region) by getting energy from lattice vibrations. In this process holes are not produced in the valence band. At about 100K, the donor levels are all ionized. Once all the donor atoms are ionized, further increase in temperature does not produce electrons and the curve levels off. The plateau region is called the *depletion region*. In the depletion region, the electron concentration in the conduction band is nearly identical to the concentration of the dopant atoms.

If N_D is the concentration of donor atoms, then

$$\ln n \cong N_D \quad (30.62)$$

where n is the electron concentration in the conduction band of n-type material.

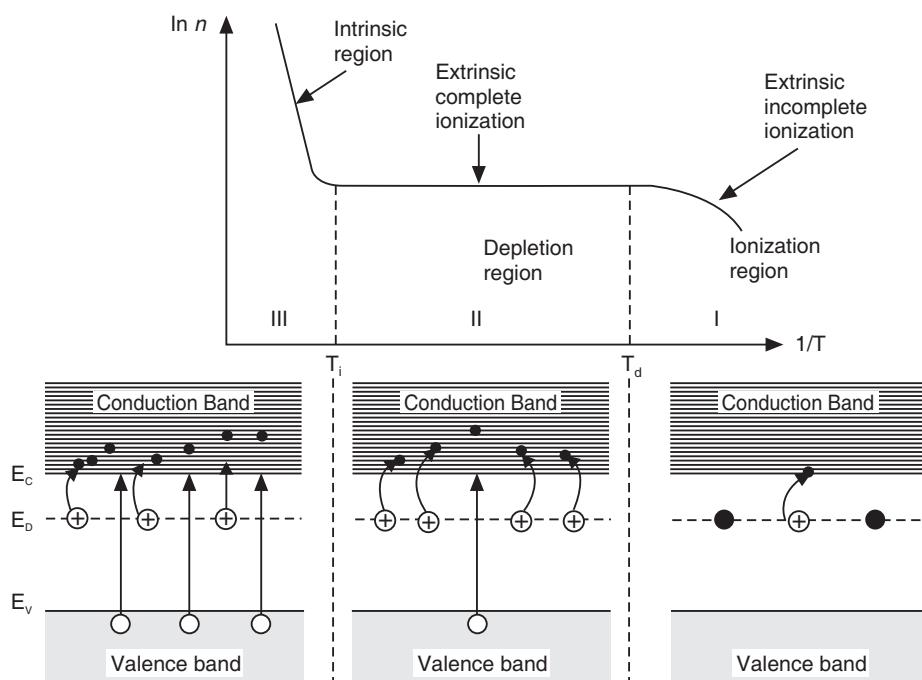


Fig. 30.14. Variation of electron concentration in an n-type semiconductor as a function of the inverse of temperature.

As temperature increases further, some electrons from the valence band are also excited into the conduction band (Fig. 30.14, depletion region). The conduction band, therefore,

contains electrons that have come through two different processes-namely (i) donor atom ionization and (ii) intrinsic process. The intrinsic process produces holes in the valence band.

At high enough temperatures the intrinsic behaviour takes over since large number of electrons in the valence band gets thermally excited to conduction band and their number far exceeds the number of donor electrons (Fig. 30.14, intrinsic region). Therefore, in intrinsic region

$$n_n = n_i \quad (30.63)$$

In an n type semiconductor at moderate temperatures, there are some 10^{10} electrons and 10^{10} holes which have been generated due to heat-ruptured bonds in addition to the 10^{16} free electrons received from the donors. Donor electrons are in the majority and their number is essentially constant at all usual temperatures; this is because they are not part of covalent bonds and only 0.05 eV is required to free them. Thus, by doping we produce a less temperature-sensitive semiconductor with an abundance of electrons.

In n -type material the electrons outnumber the holes and constitute the *majority carriers* (in region II). Holes are *minority carriers*. The number of carriers is independent of temperature in the depletion region. The current in this type of crystal is mainly due to the negatively charged electrons and hence the material is called **n -type semiconductor**.

30.16.2 Carrier Concentration in n -type Semiconductor at Low Temperatures: (In the Ionization Region)

The energy band diagram of a n -type semiconductor is shown in Fig. 30.13. Let N_D be the concentration of donors in the material. At 0K, the donor atoms are not ionized and are at the level E_D which is very near to E_C . When the temperature is raised above 0°K, the donor atoms get ionized and free electrons appear in the conduction band. With increase in temperature more and more donor atoms get ionized and the electron concentration in the conduction band increases. Electrons require an energy E_D , for their transition to the conduction band from the donor levels. Therefore, we may assume that the electron concentration, n , in the conduction band is

$$n = N_D^+$$

or

$$n = N_D - N_D^0$$

where N_D^+ is the number of donor atoms that are ionized and N_D^0 is the number of atoms left unionized at the energy level E_D .

$$\text{The concentration of ionized donors } N_D^+ = (N_D - N_D^0) = N_D [1 - f(E_D)] = \frac{N_D}{1 + e^{-(E_D - E_F)/kT}}$$

$$\therefore n = \frac{N_D}{1 + e^{-(E_D - E_F)/kT}}$$

From the operational definition of Fermi level it is expected that the Fermi level in n -type semiconductor lies a few kT above E_D . Therefore, the above equation may be simplified as

$$n = N_D e^{(E_D - E_F)/kT} \quad (30.64)$$

But the electron concentration, n , in the conduction band is given by

$$n = N_C e^{-(E_C - E_F)/kT} \quad (30.65)$$

$$\therefore N_D e^{(E_D - E_F)/kT} = N_C e^{-(E_C - E_F)/kT}$$

Taking logarithm and rearranging the terms we get

$$\left(\frac{E_D - E_F}{kT}\right) + \left(\frac{E_C - E_F}{kT}\right) = \ln \frac{N_C}{N_D} \quad (30.66)$$

$$(E_D + E_C) - 2E_F = (kT) \ln \frac{N_C}{N_D}$$

or $E_F = \frac{E_D + E_C}{2} - \left(\frac{kT}{2}\right) \ln \frac{N_C}{N_D}$

or $E_F = \frac{E_D + E_C}{2} + \left(\frac{kT}{2}\right) \ln \frac{N_D}{N_C}$

or $E_F = \frac{E_D + E_C}{2} + \left(\frac{kT}{2}\right) \ln \frac{N_D}{2(2\pi m_e^* kT/h^2)^{3/2}} \quad (30.67)$

It is seen from equ. (30.67) that at T = 0K,

$$E_F = \frac{E_D + E_C}{2} \quad (30.68)$$

That is, the equilibrium Fermi level lies midway between the bottom of the conduction band and donor levels. Now,

$$\begin{aligned} \exp\left[\frac{E_F - E_C}{kT}\right] &= \exp\left[\frac{E_D + E_C}{2kT} + \left(\frac{1}{2}\right) \ln \frac{N_D}{2(2\pi m_e^* kT/h^2)^{3/2}} - \frac{E_C}{kT}\right] \\ &= \exp\left[\frac{E_D - E_C}{2kT} + \left(\frac{1}{2}\right) \ln \frac{N_D}{2(2\pi m_e^* kT/h^2)^{3/2}}\right] \\ &= \exp\left[\frac{E_D - E_C}{2kT} + \ln \sqrt{\frac{N_D}{2(2\pi m_e^* kT/h^2)^{3/2}}}\right] \quad \left[\because \frac{1}{2} \ln x = \ln \sqrt{x}\right] \\ &= \exp\left[\left(\frac{E_D - E_C}{2kT}\right)\right] \cdot \exp\left[\ln \sqrt{\frac{N_D}{2(2\pi m_e^* kT/h^2)^{3/2}}}\right] \quad \left[\because \exp(a+b) = \exp(a) + \exp(b)\right] \\ &= \exp\left[\left(\frac{E_D - E_C}{2kT}\right)\right] \cdot \left[\sqrt{\frac{N_D}{2(2\pi m_e^* kT/h^2)^{3/2}}}\right] \quad [\exp(\ln x) = x] \\ \therefore n &= N_C \exp\left[\frac{E_F - E_C}{kT}\right] = 2 \left[\frac{2\pi m_e^* kT}{h^2} \right]^{3/2} \exp\left[\left(\frac{E_D - E_C}{2kT}\right)\right] \cdot \left[\sqrt{\frac{N_D}{2(2\pi m_e^* kT/h^2)^{3/2}}}\right] \end{aligned}$$

or $n = (2N_D)^{\frac{1}{2}} \left[\frac{2\pi m_e^* kT}{h^2} \right]^{3/4} \exp\left[\left(\frac{E_D - E_C}{2kT}\right)\right] \quad (30.69)$

Thus, the electron concentration in the conduction band of an *n*-type semiconductor is proportional to the square root of the donor concentration at moderately low temperatures. Electrons are the majority carriers in an *n*-type semiconductor and the above expression therefore gives the majority carrier concentration.

30.17 p-TYPE SEMICONDUCTOR

A *p*-type semiconductor is produced when a pure semiconductor is doped with a trivalent impurity such as boron. Boron atom has three valence electrons. Therefore, it falls short of one electron for completing the four covalent bonds with its neighbours. When an electron from a neighbouring atom acquires energy and jumps into the vacancy to form the fourth bond, it leaves behind a hole. The boron atom having acquired an additional electron becomes a negative ion. The hole can move freely in the valence band whereas the impurity ion is fixed in position by the covalent bonds. As the boron atom accepted an electron from the valence band, it is called an *acceptor atom*. The acceptor impurity atoms produce holes without the simultaneous generation of the electrons in the conduction band.

Energy Band Diagram

The energy band diagram of *p*-type semiconductor is shown in Fig. 30.15. If the acceptor atom density is low, the acceptor atoms are distantly spaced from one another. As such, the acceptor atoms cannot interact with each other and their energy levels are discrete levels, E_A . They are called **acceptor levels** and represent the ground state of the hole. As even small amount of thermal energy can make an electron in the valence band jump into the acceptor level, the acceptor levels are expected to be located very near to the top edge of the valence band. They are at about 0.01 eV above the valence band. A hole may be said to have moved from the acceptor atom to the valence band.

At 0 K the acceptor levels are vacant and the valence band is full. The conduction band is also vacant. The material behaves essentially as an insulator.

At slightly elevated temperature, electrons from the valence band jump into the acceptor levels and holes are generated in the valence band (Fig. 30.15). In this process holes are generated without the simultaneous generation of electrons. At normal temperatures the acceptor levels are saturated and a few electrons are excited to the conduction band also. At about 100K, the acceptor atoms are all ionized. Once all the acceptor atoms are ionized, further increase in temperature does not produce holes and we say the acceptor levels are saturated. The region is called the *saturation region*. In the saturation region, the hole concentration in the valence band is nearly identical to the concentration of the acceptor atoms.

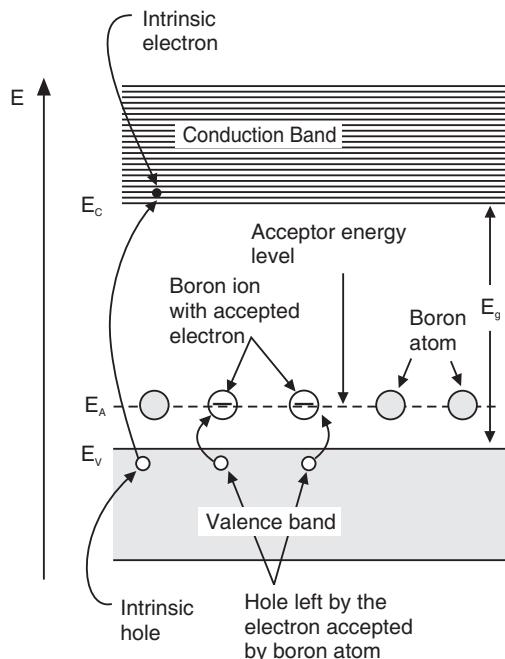


Fig. 30.15. Energy band diagram for a *p*-type semiconductor

If N_A is the concentration of acceptor atoms, then

$$p_p \equiv N_A \quad (30.70)$$

where p_p is the hole concentration in the valence band of p -type material.

As temperature is increased further, some electrons from the valence band are excited into the conduction band. The valence band now contains holes that have been generated by two different processes-namely, (i) acceptor atom ionization and (ii) intrinsic process. The intrinsic process causes electrons to appear in the conduction band.

At high enough temperatures, a large number of electron-hole pairs are generated and the number of holes generated thermally far exceeds the number of holes due to acceptor impurity. The material behaves as an intrinsic semiconductor. In the intrinsic region

$$p_p = n_i \quad (30.71)$$

In p -type material the holes outnumber the electrons and constitute the *majority carriers*. Electrons are *minority carriers*. The number of majority carriers is independent of temperature in the depletion region. The current in this type of crystal is mainly due to the positively charged holes and hence the material is called **p-type semiconductor**.

30.17.1 Carrier Concentration in p-type Semiconductor at Low Temperatures : (In the Ionization Region)

The energy band diagram of a p -type semiconductor is shown in Fig. 30.15. Let N_A be the concentration of acceptors in the material. At 0K, the acceptor atoms are not ionized and are at the level E_A which is very near to E_V . When the temperature is raised above 0°K, the acceptor atoms get ionized and holes appear in the valence band. With increase in temperature more and more acceptor atoms get ionized and the hole concentration in the valence band increases. Since transition of electrons to the acceptor levels from the valence band requires an energy E_A , they can go to acceptor levels and ionize acceptor atoms. Therefore, we may assume that the hole concentration, p , in the conduction band is

$$p = N_A^-$$

where N_A^- is the number of acceptor atoms that are ionized.

$$\begin{aligned} \text{The concentration of ionized acceptors } N_A^- &= N_A f(E_A) = N_A \exp\left(\frac{E_F - E_A}{kT}\right) \\ \therefore p &= N_A \exp\left(\frac{E_F - E_A}{kT}\right) \end{aligned} \quad (30.72)$$

But the hole concentration, p , in the valence band is given by

$$p = N_V e^{-(E_F - E_V)/kT} = N_V \exp\left(\frac{E_V - E_F}{kT}\right) \quad (30.73)$$

$$\therefore N_A \exp\left(\frac{E_F - E_A}{kT}\right) = N_V \exp\left(\frac{E_V - E_F}{kT}\right)$$

Taking logarithm and rearranging the terms we get

$$\begin{aligned} \left(\frac{E_F - E_A}{kT}\right) - \left(\frac{E_V - E_F}{kT}\right) &= \ln \frac{N_V}{N_A} \\ -(E_V + E_A) + 2E_F &= (kT) \ln \frac{N_V}{N_A} \\ \text{or } E_F &= \frac{E_V + E_A}{2} + \left(\frac{kT}{2}\right) \ln \frac{N_V}{N_A} \end{aligned}$$

or

$$E_F = \frac{E_V + E_A}{2} - \left(\frac{kT}{2} \right) \ln \frac{N_A}{N_V}$$

or

$$E_F = \frac{E_V + E_A}{2} - \left(\frac{kT}{2} \right) \ln \frac{N_A}{2(2\pi m_h^* kT/h^2)^{3/2}} \quad (30.74)$$

It is seen from equ. (30.74) that at T = 0K,

$$E_F = \frac{E_V + E_A}{2} \quad (30.75)$$

That is, the equilibrium Fermi level lies midway between the bottom of the valence band and acceptor levels. Now,

$$\begin{aligned} \exp \left[\frac{E_V - E_F}{kT} \right] &= \exp \left[\frac{E_V}{kT} - \frac{E_V + E_A}{2kT} + \left(\frac{1}{2} \right) \ln \frac{N_A}{2(2\pi m_h^* kT/h^2)^{3/2}} \right] \\ &= \exp \left[\frac{E_V - E_A}{2kT} + \left(\frac{1}{2} \right) \ln \frac{N_A}{2(2\pi m_h^* kT/h^2)^{3/2}} \right] \\ &= \exp \left[\frac{E_V - E_A}{2kT} + \ln \sqrt{\frac{N_A}{2(2\pi m_h^* kT/h^2)^{3/2}}} \right] \quad \left[\because \frac{1}{2} \ln x = \ln \sqrt{x} \right] \\ &= \exp \left[\left(\frac{E_V - E_A}{2kT} \right) \right] \cdot \exp \left[\ln \sqrt{\frac{N_A}{2(2\pi m_h^* kT/h^2)^{3/2}}} \right] \quad \left[\because \exp(a+b) = \exp(a) + \exp(b) \right] \\ &= \exp \left[\left(\frac{E_V - E_A}{2kT} \right) \right] \cdot \left[\sqrt{\frac{N_A}{2(2\pi m_h^* kT/h^2)^{3/2}}} \right] \quad \left[\because \exp(\ln x) = x \right] \\ \therefore p &= N_V \exp \left[\frac{E_V - E_A}{kT} \right] = 2 \left[\frac{2\pi m_h^* kT}{h^2} \right]^{3/2} \exp \left[\left(\frac{E_V - E_A}{2kT} \right) \right] \cdot \left[\sqrt{\frac{N_A}{2(2\pi m_h^* kT/h^2)^{3/2}}} \right] \\ \text{or } p &= (2N_A)^{\frac{1}{2}} \left[\frac{2\pi m_h^* kT}{h^2} \right]^{3/4} \exp \left[\left(\frac{E_V - E_A}{2kT} \right) \right] \quad (30.76) \end{aligned}$$

Thus, the hole concentration in the valence band of a *p*-type semiconductor is proportional to the square root of the acceptor concentration at moderately low temperatures. Holes are the majority carriers in a *p*-type semiconductor. Therefore, the above expression represents the majority carrier concentration in a *p*-type semiconductor.

30.18 BAND DIAGRAMS OF EXTRINSIC SEMICONDUCTORS AT 0K AND 300K

The energy band diagrams of extrinsic semiconductors at 0K and 300K showing the position of Fermi level in each case is shown in Fig. 30.16 and Fig. 30.17 respectively.

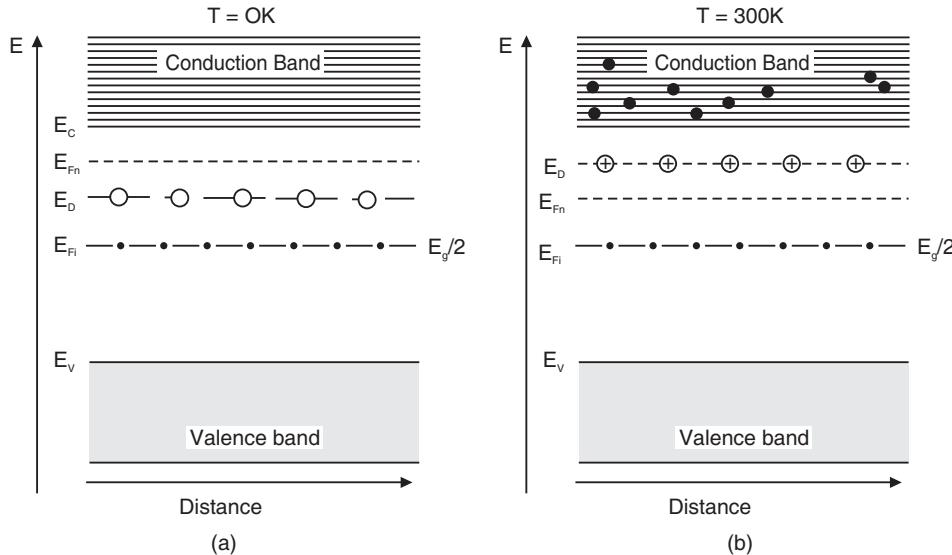


Fig. 30.16 Energy band diagram of an n-type semiconductor (a) at 0K (b) at 300K

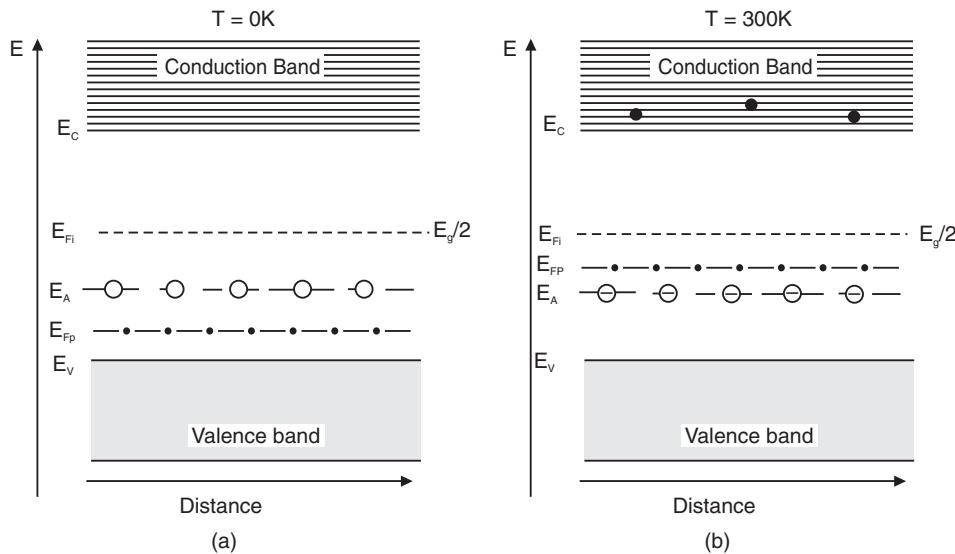


Fig. 30.17 Energy band diagram of a p-type semiconductor (a) at 0K (b) at 300 K

30.19 EXTRINSIC CONDUCTIVITY

The temperature range in which a doped semiconductor used is the extrinsic region where all the impurity atoms are ionized (Region II, Fig. 30.14). This region is of practical interest since the carrier concentration is essentially independent of temperature in this region and any desired conductivity can be achieved by controlling the amount of impurities added from

outside. Note that whenever extrinsic semiconductor is discussed, we invariably imply its characteristics in the extrinsic region. The conductivity of n-type semiconductor is given by

$$\sigma_n = n_n e \mu_e + p_n e \mu_h$$

As $p_n \ll n_n$, the second term is negligible and it is the electrons that contribute to the conductivity. Therefore,

$$\sigma_n = n_n e \mu_e$$

As $n_n = N_D$, we can write the above relation as

$$\sigma_n = N_D e \mu_e \quad (30.77)$$

Similarly, the conductivity of p-type semiconductor is given by

$$\sigma_p = p_p e \mu_h$$

or

$$\sigma_p = N_A e \mu_h \quad (30.78)$$

A general dependence of conductivity on temperature in extrinsic semiconductors is shown in Fig. 30.18.

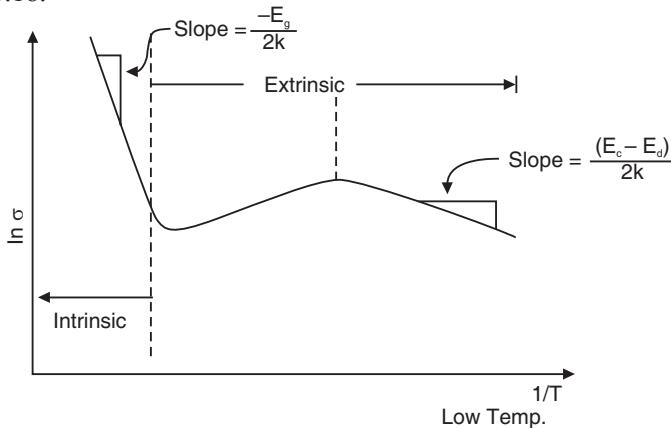


Fig. 30.18

Example 30.12. In a doped semiconductor, there are 4.52×10^{24} holes and 1.25×10^{14} electrons per cubic metre. What will be the carrier density in undoped specimen? Electron and hole mobilities are $0.38 \text{ m}^2/\text{V.s}$ and $0.18 \text{ m}^2/\text{V.s}$ respectively. Calculate the conductivity of intrinsic and the doped semiconductors.

Solution. Intrinsic carrier density, $n_i = \sqrt{p_p n_p}$

$$\therefore n_i = \sqrt{(4.52 \times 10^{24} / \text{m}^3)(1.25 \times 10^{14} / \text{m}^3)} \\ = (5.65 \times 10^{38} / \text{m}^6)^{1/2} = 2.38 \times 10^{19} / \text{m}^3$$

Conductivity of the intrinsic semiconductor

$$\begin{aligned} \sigma_i &= e n_i (\mu_e + \mu_h) \\ &= (1.602 \times 10^{-19} \text{ C})(2.38 \times 10^{19} / \text{m}^3)(0.38 + 0.18) \frac{\text{m}^2}{\text{V.s}} \\ &= 2.14 \frac{\text{C}}{\text{m.V.s}} = 2.14 \text{ S/m} \end{aligned}$$

Conductivity of the doped semiconductor

$$\sigma_p = e p_p \mu_h$$

$$= (1.602 \times 10^{-19} \text{ C}) (4.52 \times 10^{24} / \text{m}^3) (0.18) \frac{\text{m}^2}{\text{V.s}}$$

$$= 1.30 \times 10^5 \frac{\text{C}}{\text{m.V.s}} = 130 \text{ kS/m}$$

Unit: $1 \frac{\text{C}}{\text{m.V.s}} = 1 \frac{\text{A}}{\text{mV}} = 1 \Omega^{-1} \text{ m}^{-1} = 1 \text{ S/m}$

Example 30.13: Silicon has a conductivity of only $5 \times 10^{-4} \Omega^{-1}\text{m}^{-1}$ in its pure form. An engineer wanted it to have conductivity of $200 \Omega^{-1}\text{m}^{-1}$ and doped it with aluminium to produce p-type semiconductor. Calculate the impurity concentration. Assume $\mu_h = 0.05 \text{ m}^2/\text{Vs}$.

Solution.

$$\sigma = pe\mu_h = N_A e\mu_h$$

$$\therefore N_A = \frac{\sigma}{e\mu_h} = \frac{200 \Omega^{-1}\text{m}^{-1}}{1.602 \times 10^{-19} \text{ C} \times 0.05 \text{ m}^2/\text{Vs}} \\ = 2.5 \times 10^{22} \text{ atoms/m}^3.$$

30.20 LAW OF MASS ACTION

In case of intrinsic semiconductors the product of n and p is a constant for a certain semiconductor at a certain temperature and is given by

$$np = n_i^2 = N_C N_V e^{-E_g/kT}$$

There is no condition in the expression that restricts it to intrinsic semiconductors because E_g does not change with impurity concentration and N_C and N_V are constants. This product is therefore a constant equally valid for intrinsic as well as for extrinsic semiconductors.

The electron and hole concentrations in extrinsic semiconductors may be given by expressions similar to equ. (30.30) and (30.41). Denoting the electron concentration in n-type semiconductor by n_n and the hole concentration by p_n , we can write

$$n_n = N_C e^{-(E_C - E_F)/kT} \quad (30.79)$$

and

$$p_n = N_V e^{-(E_F - E_V)/kT} \quad (30.80)$$

$$\therefore n_n p_n = N_C N_V e^{-E_g/kT} \quad (30.81)$$

$$\therefore n_n p_n = n_i^2 \quad (30.82)$$

Similarly, if we denote the hole concentration in a p-type semiconductor by p_p and electron concentration by n_p , we can express

$$p_p = N_V e^{-(E_F - E_V)/kT} \quad (30.83)$$

and

$$n_p = N_C e^{-(E_C - E_F)/kT} \quad (30.84)$$

$$\therefore p_p n_p = N_C N_V e^{-E_g/kT} \quad (30.85)$$

$$\therefore p_p n_p = n_i^2 \quad (30.86)$$

The relations (30.82) and (30.86) show that the product of majority and minority carrier concentrations in an extrinsic semiconductor at a particular temperature is a constant and is equal to the square of intrinsic carrier concentration at that temperature.

The law of mass action is very important relation because it in conjunction with charge neutrality condition enables us calculate minority carrier concentration. The law suggests that the addition of impurities to an intrinsic semiconductor increases the concentration of one type of carrier, which consequently becomes majority carrier and simultaneously decreases the concentration of the other carrier, which as a result becomes minority carrier. The minority carriers decrease in number below the intrinsic value because the majority carriers increase the rate of recombinations. *The law of mass action states that the product of majority and minority carriers remains constant in an extrinsic semiconductor and it is independent of the amount of donor and acceptor impurity concentrations.* Note that when the doping is heavy, the minority carrier concentration will be low and if doping is lighter, the minority carrier concentration will be larger.

30.21 CHARGE NEUTRALITY CONDITION

An extrinsic semiconductor is an electrically neutral body in its equilibrium condition. In n-type semiconductor, the total number of electrons in the conduction band must be equal to the sum of electrons originated from the donor atoms and electrons excited from the valence band. Electrons coming from donor levels leave behind positive donor ions while electrons excited from the valence band leave behind holes. These processes have not created any additional charges, and the equality between positive and negative charges remain undisturbed. The **charge neutrality condition** applied to the n-type semiconductor implies that the total negative charge of mobile electrons is equal to the total positive charge created in the crystal. It means that

$$n_n = N_D + p_n$$

where N_D is the donor impurity concentration and all the donor atoms are assumed to have got ionized.

But $n_n > p_n$
 $\therefore n_n = N_D$ (30.87)

The above relation indicates that *the majority carrier concentration, n_n , in an n-type semiconductor is equal to the donor impurity concentration, N_D .*

The charge neutrality for a p-type semiconductor requires that

As $p_p = N_A + n_p$
 $p_p > n_p$
 $p_p = N_A$ (30.88)

30.22 CALCULATION OF MINORITY CARRIER CONCENTRATION

In the case of an n-type semiconductor, the majority carrier concentration n_n is given by

$$n_n = N_D$$

where N_D is the donor impurity concentration and all the donor atoms are assumed to have got ionized. The above relation indicates that *the majority carrier concentration, n_n , in an n-type semiconductor is equal to the donor impurity concentration, N_D .* The **minority carrier concentration**, p_n , is then given by eq. (30.82), as

$$p_n = \frac{n_i^2}{n_n}$$

or $p_n = \frac{n_i^2}{N_D}$ (30.89)

Similarly, in case of a *p*-type semiconductor, the minority carrier concentration, n_p is given by

$$n_p = \frac{n_i^2}{p_p}$$

or

$$n_p = \frac{n_i^2}{N_A} \quad (30.90)$$

Example 30.14. A sample of intrinsic germanium at room temperature has a carrier concentration of $2.4 \times 10^{19} / m^3$. It is doped with antimony at a rate of one antimony atom per million atoms of germanium. If the concentration of the germanium atoms is $4 \times 10^{28} / m^3$, determine the hole concentration.

Solution.

$$N_D = \frac{4 \times 10^{28} \text{ atoms}/m^3}{10^6 \text{ atoms}/m^3} = 4 \times 10^{22} \text{ donors}/m^3$$

$$n_n = N_D = 4 \times 10^{22} \text{ electrons}/m^3$$

$$\therefore p_n = \frac{n_i^2}{n_n} = \frac{(2.4 \times 10^{19} \text{ carriers}/m^3)^2}{(4 \times 10^{22} \text{ electrons}/m^3)}$$

or

$$p_n = 1.4 \times 10^{16} \text{ holes}/m^3.$$

Example 30.15: A sample of intrinsic silicon at room temperature has a carrier concentration of $1.5 \times 10^{16} / m^3$. A donor impurity is added to the extent of 1 donor atom per 10^8 atoms of silicon. If the concentration of silicon atoms is $5 \times 10^{28} \text{ atoms}/m^3$, determine the resistivity of the material. (Given $\mu_e = 0.135 \text{ m}^2/\text{V.s}$ and $\mu_h = 0.048 \text{ m}^2/\text{V.s}$)

Solution:

$$\text{Donor atom density is given as } N_D = \frac{N}{10^8} = \frac{5 \times 10^{28} / m^3}{10^8} = 5 \times 10^{20} / m^3$$

$$\text{Free electron concentration is given by } n_n = N_D = 5 \times 10^{20} / m^3$$

$$\text{Hole concentration is given by } p_n = \frac{n_i^2}{n_n} = \frac{(1.5 \times 10^{16} / m^3)^2}{5 \times 10^{20} / m^3} = 4.5 \times 10^{11} / m^3.$$

$$\begin{aligned} \text{Resistivity of the material } \rho &= \frac{1}{\sigma} = \frac{1}{e(n_n \mu_e + p_n \mu_h)} \\ &= \frac{1}{(1.602 \times 10^{-19} \text{ C})(5 \times 10^{20} \times 0.135 \text{ m}^2/\text{V.s} + 4.5 \times 10^{11} \times 0.048 \text{ m}^2/\text{V.s})} = 0.092 \Omega \cdot \text{m}. \end{aligned}$$

Example 30.16: A sample of intrinsic germanium at room temperature has a carrier concentration of $2.4 \times 10^{19} / m^3$. It is doped with antimony at a rate of one antimony atom per million atoms of germanium. If the concentration of the germanium atoms is $4 \times 10^{28} / m^3$, determine the hole concentration.

Solution.

$$\text{Donor atom density is given as } N_D = \frac{N}{10^6} = \frac{4 \times 10^{28} / m^3}{10^6} = 4 \times 10^{22} / m^3$$

$$\text{Free electron concentration is given by } n_n = N_D = 4 \times 10^{22} / m^3$$

$$\text{Hole concentration is given by } p_n = \frac{n_i^2}{n_n} = \frac{(2.4 \times 10^{19}/\text{m}^3)^2}{4 \times 10^{22}/\text{m}^3} = 1.4 \times 10^{16}/\text{m}^3.$$

30.23 FERMI LEVEL IN EXTRINSIC SEMICONDUCTORS

In intrinsic semiconductors, the Fermi level, E_F lies in the middle of the band gap. However, the situation is different with extrinsic semiconductors. In an n-type semiconductor, the Fermi level lies in the upper half of the gap, as the majority carriers reside in the conduction band and their average energy is more than E_{Fn} . In a p-type semiconductor, the Fermi level lies in the lower half of the gap, as the majority carriers reside in the valence band and their average energy is less than E_{Fp} . The carrier concentrations in extrinsic semiconductors vary with temperature and impurity concentration. It means that the probability of occupancy of respective bands varies and consequently the position of Fermi level changes with temperature and impurity concentration.

(a) VARIATION OF FERMI LEVEL WITH TEMPERATURE IN AN n-TYPE SEMICONDUCTOR

In the *n*-type semiconductor at low temperatures, some donor atoms are ionized and provide electrons to the conduction band while others remain neutral. As electrons in the conduction band are only due to the transitions from the donor levels, the Fermi level must lie between the impurity donor levels and the bottom of the conduction band. When $T = 0\text{K}$, E_{Fn} lies midway between the donor levels and the bottom of the conduction band. It is thus,

$$E_{Fn} = \frac{E_C + E_D}{2} \text{ at } T = 0\text{ K} \quad (30.91)$$

As the temperature increases the donor levels gradually get depleted and the Fermi level moves downward. At the temperature of complete depletion of donor levels, T_d , the Fermi level coincides with the donor level E_D . Thus

$$E_{Fn} = E_D \text{ at } T = T_d \quad (30.92)$$

As the temperature grows further above T_d , the Fermi level shifts downward in an approximately linear fashion. At a temperature T_i , the intrinsic process contributes to electron concentration significantly. At higher temperatures, the *n*-type semiconductor loses its extrinsic character and behaves as an intrinsic semiconductor. In the intrinsic region, the electron concentration in conduction band increases exponentially and the Fermi level approaches the intrinsic value. Thus,

$$E_{Fn} = E_{Fi} = \frac{E_g}{2} \quad \text{at } T \geq T_i \quad (30.93)$$

The variation of Fermi level E_{Fn} in an *n*-type semiconductor with temperature is illustrated in Fig. 30.19.

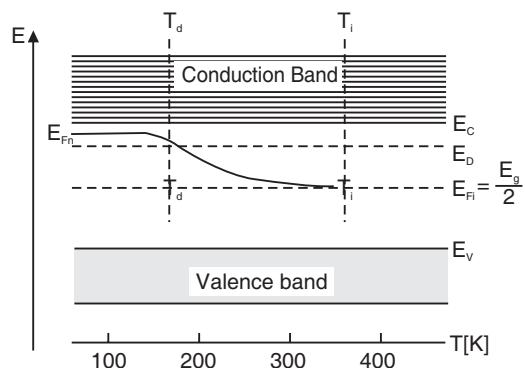


Fig. 30.19: Qualitative dependence of Fermi level on temperature in an *n*-type semiconductor

Variation of Fermi Level with Temperature in a p-type Semiconductor

In case of *p*-type semiconductor, in the low temperature region, holes in the valence band are only due to the transitions of electrons from the valence band to the acceptor levels. As the valence band is the source of electrons and the acceptor levels are the recipients for them, the Fermi level must lie between the top of the valence band and the impurity acceptor levels. When $T = 0$, Fermi level lies midway between the acceptor levels and the top of the valence band. Thus,

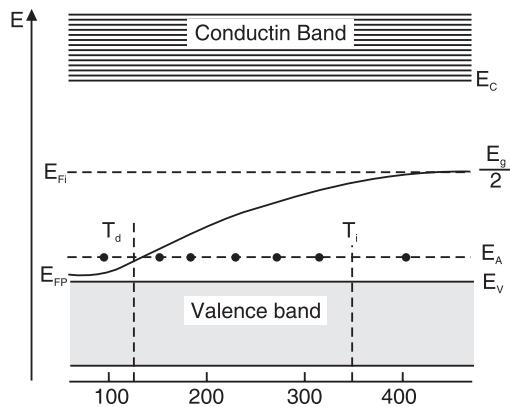


Fig. 30.20: Qualitative dependence of Fermi level on temperature in a *p*-type semiconductor

$$E_{FP} = \frac{E_V + E_A}{2} \quad (30.94)$$

As the temperature increases the acceptor levels gradually get filled and the Fermi level moves upward. At the temperature of saturation T_s , the Fermi level coincides with the acceptor level E_A . Thus,

$$E_{FP} = E_A \quad \text{at } T = T_s \quad (30.95)$$

As the temperature grows above T_s , the Fermi level shifts upward in an approximately linear fashion.

At a temperature T_i intrinsic behaviour sets in. At higher temperatures, the *p*-type semiconductor loses its extrinsic character and behaves as an intrinsic semiconductor. In the intrinsic region, the hole concentration in the valence band increases exponentially and the Fermi level approaches the intrinsic value. Thus

$$E_{FP} = E_i = \frac{E_g}{2} \quad \text{at } T = T_i \quad (30.96)$$

The variation of Fermi level E_{Fn} in an *p*-type semiconductor with temperature is illustrated in Fig. 30.20.

30.24 VARIATION OF FERMI LEVEL WITH IMPURITY CONCENTRATION

n-type Semiconductor

The addition of donor impurity to an intrinsic semiconductor leads to the formation of discrete donor levels below the bottom edge of the conduction band. At low impurity concentrations, the impurity atoms are distantly spaced from one another, approximately, by 100 atom spacings. Therefore, they do not interact with each other. With an increase in the impurity concentration, the separation between impurity atoms decreases and they tend to interact. As a result, the donor levels undergo splitting and form an energy band below the conduction band, as shown in Fig. 30.21.

The larger the doping concentration, the broader is the impurity band; and at one stage the impurity band overlaps on the conduction band. Then the upper vacant levels in the conduction band are accessible to the donor electrons. The broadening of donor levels into a band is accompanied by a decrease in the width of the forbidden gap and also by the upward displacement of Fermi level. The Fermi level shifts closer and closer to the conduction band

with increasing impurity concentration and finally moves into the conduction band when the donor band overlaps on the conduction band.

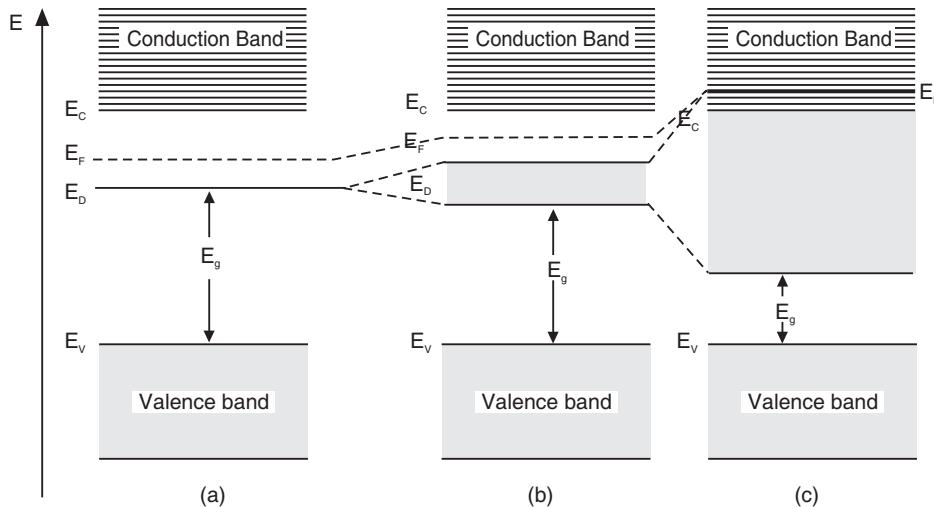


Fig. 30.21: Energy band diagrams of an *n*-type semiconductor at three different levels of doping; (a) low level doping; (b) medium doping; (c) heavy doping.

p-type Semiconductor

In *p*-type semiconductor, the acceptor levels broaden and form into a band with increasing impurity concentration. The acceptor band ultimately overlaps on the valence band. The Fermi level moves down closer to the valence band and finally at very high impurity concentration it will shift into the valence band.

30.25 DRIFT AND DIFFUSION CURRENTS

Under the condition of thermal equilibrium, the electrons and holes are uniformly distributed in the crystal and in the absence of an external stimulus their average velocity is zero and no current flows through the crystal. This is equally true for an intrinsic or an extrinsic semiconductor.

The thermal equilibrium may be disturbed by an external agent and the chaotic motion of charge carriers acquire a directional movement leading to a flow of current in the material. Electric field and concentration gradients are examples of such disturbing agents.

1. Drift Current: When an electric field E is applied across a semiconductor, the charge carriers acquire a directional motion over and above their thermal motion and produce drift current.

The electrons drifting in the conduction band produce a current component J_e given by

$$J_e(\text{drift}) = ne\mu_e E$$

The holes drifting in the valence band cause a current component J_h given by

$$J_h(\text{drift}) = pe\mu_h E$$

Therefore, the total drift current density is,

$$J_{dr} = J_e(\text{drift}) + J_h(\text{drift})$$

Drift current occurs only when external electric field is present across the solid. Although electrons and holes move in opposite directions, the direction of conventional current flow due to both the carriers is in the same direction.

$$\therefore J_{dr} = e(n\mu_e + p\mu_h)E \quad (30.97)$$

2. Diffusion Current: In case of semiconductors, current can also flow without the application of an external electric field. If a spatial variation of carrier density is created in the semiconductor, current flows in it. If we consider an arbitrary surface in the volume of the solid and if there are more charge carriers on its one side than on the other side, we say there is a **concentration gradient**. This concentration gradient causes a directional movement of charge carriers, which continues until all the carriers are evenly distributed throughout the material. Any movement of charge carriers constitutes an electric current, and this type of movement produces a current component known as **diffusion current**.

Concentration gradient may be produced in an extrinsic semiconductor by applying heat or light locally at one region. Suppose an external agent such as light or heat acts momentarily at one end of a *p*-type semiconductor, as shown in Fig. 30.22. The external agent generates additional electron-hole pairs leading to a sudden increase in the concentration of charge carriers at that end. In the rest of the volume, the concentration of carriers is at equilibrium value. The difference in the concentration of charge carriers initiates the carriers to diffuse from the region of higher concentration to the region of lower concentration in order to restore the equilibrium condition. As the carriers are charged particles, their migration produces a current flow, which is the diffusion current. The diffusion current strength is proportional to the concentration gradient, i.e. the rate of change of carrier concentration per unit length. In case of electrons moving left to right (see Fig. 30.22), current flows from right to left in the negative *x*-direction.

The current component due to electron diffusion is given by

$$J_e(\text{diff}) = eD_e \frac{dn}{dx} \quad (30.98)$$

The current component due to hole diffusion is given by

$$J_h(\text{diff}) = -eD_h \frac{dp}{dx} \quad (30.99)$$

D_e and D_h are *diffusion coefficients* for electrons and holes respectively.

Drift and diffusion currents coexist in semiconductors. The total current density due to drift and diffusion of electrons may be written as

$$J_e = e \left(n\mu_e E + D_e \frac{dn}{dx} \right) \quad (30.100)$$

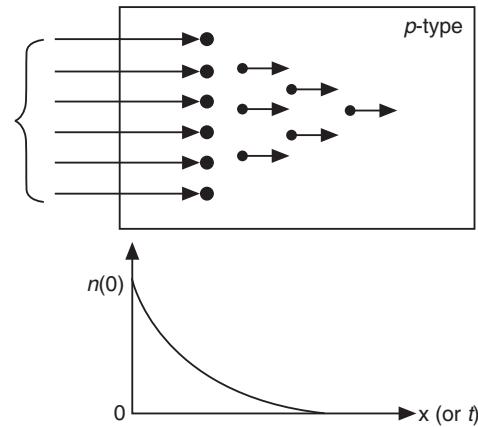


Fig. 30.22

Similarly, for holes we can write

$$J_h = e \left(p \mu_h E - D_h \frac{dp}{dx} \right) \quad (30.101)$$

30.26 MINORITY CARRIER DIFFUSION

In extrinsic semiconductors the external agent disturbs the equilibrium concentration of minority carriers only. For instance, let us again consider *p*-type semiconductor with an acceptor impurity concentration of 2.25×10^{22} atoms / m³. Under the equilibrium condition, the majority carrier concentration is 2.25×10^{22} holes / m³ whereas the minority carrier concentration is

$$n_p = \frac{n_i^2}{N_A} = \frac{(1.5 \times 10^{16} \text{ carriers/m}^3)^2}{2.25 \times 10^{22} \text{ atoms/m}^3} = 10^{10} \text{ electrons/m}^3$$

Let us assume that one end of the semiconductor is heated momentarily and 10^{10} electron-hole pairs/m³ are generated due to the sudden local rise in temperature. It is seen that due to the action of temperature, the number of electrons is doubled but there is no change in the number of holes ($10^{22} \gg 10^{10}$). Consequently, electrons experience a large concentration gradient and diffuse towards the other end of the semiconductor. As they diffuse into the bulk of the material, recombinations take place along the length of their travel. Electrons continue to diffuse till their concentration reaches the equilibrium value. On the other hand, holes do not experience concentration gradient and therefore do not diffuse. It all means that the minority carriers control the electrical behaviour of a semiconductor device.

30.26.1 Diffusion Length and Mean Lifetime

Once again let us take the case of a *p*-type semiconductor (Fig. 30.22) where an excess of carriers is generated at the end $x = 0$. The concentration of electrons rises above equilibrium value at $x = 0$. The electrons tend to diffuse to the right and as they diffuse, they undergo rapid recombinations. The excess electron concentration gradually falls off with distance in an exponential manner and the decrease with distance is governed by the relation

$$\Delta n(x) = n(0) e^{-x/L_n} \quad (30.102)$$

where $n(0)$ is the electron concentration at $x = 0$ and L_n is the *diffusion length* for electrons. The **diffusion length** is the average distance covered by an excess carrier during its lifetime τ . It may also be defined as the distance covered by an excess carrier between its generation and recombination.

The recombination of carriers occurs with distance and in time. The decrease in the excess of electrons with time is also governed by an exponential relation.

$$\Delta n(t) = n(0) e^{-t/\tau_n} \quad (30.103)$$

where τ_n is called the **mean lifetime** of electrons.

The diffusion length L_n and the mean lifetime τ_n are intimately related to each other.

$$L_n = \sqrt{D_n \tau_n} \quad (30.104)$$

Similar equations are valid for excess holes generated in an *n*-type semiconductor.

$$\Delta p(x) = p(0) e^{-x/L_h} \quad (30.105)$$

$$\Delta p(t) = p(0) e^{-t/\tau_h} \quad (30.106)$$

$$L_h = \sqrt{D_h \tau_h} \quad (30.107)$$

30.26.2 Einstein Relations

Although drift and diffusion are two seemingly different processes, the parameters μ , the mobility and D , the diffusion length are not independent. There exists a close relationship between them, since both these parameters are determined by the thermal motion and scattering of the free carriers. They are related as follows:

$$\frac{D_h}{\mu_h} = \frac{kT}{e} \quad (30.108)$$

Similar relation holds good for electrons also.

$$\frac{D_n}{\mu_e} = \frac{kT}{e} \quad (30.109)$$

The equations (30.108) and (30.109) are known as **Einstein relations**. From these relations we get

$$\frac{D_n}{D_h} = \frac{\mu_e}{\mu_h} \quad (30.110)$$

Example 30.17: Find the diffusion coefficient of electrons in silicon at 300K if μ_e is $0.19 \text{ m}^2/\text{V.s}$.

Solution:

$$\frac{D_n}{\mu_e} = \frac{kT}{e} .$$

Therefore,

$$D_n = \frac{kT}{e} \mu_e = \frac{1.38 \times 10^{-23} \text{ J/K} \times 300\text{K}}{1.602 \times 10^{-19} \text{ C}} \times 0.19 \text{ m}^2/\text{V.s}$$

$$= 0.0045 \text{ m}^2/\text{V.s}$$

30.27 COMPOUND SEMICONDUCTORS

Silicon and germanium are the commonly used semiconductors. They are elements belonging to the IV group. They are used either in pure form or doped form. These intrinsic and extrinsic semiconductors are known as **elemental semiconductors**.

Compound semiconductors are semiconductor compounds composed of elements from two or more different groups of the periodic table, such as III-IV, II-VI, IV-VI semiconductors. They are formed by combining equal atomic fractions of the above group elements. III-V semiconductors are composed of elements from group III (B, Al, Ga, In) and Group V (N, P, As, Sb, Bi). Thus, GaP, GaAs, InP, InAs, and AlSb are examples of the III-IV compound semiconductors. Similarly, ZnO, ZnS, ZnSe, CdS, and HgS are examples of II-VI compounds. The range of possible compounds is very broad. Binary compounds are formed by combining two elements as in GaAs. Ternary compounds are obtained when three elements are combined as in InGaAs. Quaternary compounds are formed when four elements are combined, as in AlInGaP. It is possible to adjust the bandgap and change the optical properties by forming their ternary and quaternary derivatives by doping with impurities. GaAs is of great technical interest because it has large band gap and electron mobility is large in the material.

The elemental semiconductors Si, and Ge are indirect band gap semiconductors whereas III-IV compound semiconductors are direct gap semiconductors. The elemental semiconductors are not suitable for producing light, while III-IV compound semiconductors have high optoelectronic conversion efficiency and are widely used in fabrication of optoelectronic devices, such as lasers, LEDs etc.

30.28 HALL EFFECT

If a metal or a semiconductor carrying a current I is placed in a transverse magnetic field B , a potential difference V_H is produced in a direction normal to both the magnetic field and current directions. This is known as *Hall effect*. This effect was discovered by E.H. Hall in 1879 and showed that it is negatively charged particles that carry current in metals.

30.28.1 Importance of Hall Effect

The importance of Hall effect in the field of semiconductors is that it helps to determine

- (i) the type of semiconductor,
- (ii) the sign of majority charge carriers,
- (iii) the majority charge carrier concentration,
- (iv) the mobility of majority charge carriers, and
- (v) the mean drift velocity of majority charge carriers.

30.28.2 Experimental Arrangement

The experimental set up for the measurement of Hall voltage and determination of Hall coefficient is shown in Fig. 30.23 (a). A thin rectangular semiconductor wafer is mounted on an insulating strip and two pairs of electrical contacts are provided on opposite sides of the wafer. One pair of contacts is connected to a constant current source. And the other pair is connected to a sensitive voltmeter. This arrangement is mounted in between two pole pieces of an electromagnet such that the magnetic field acts perpendicular to the lateral faces of the semiconductor wafer.

30.28.3 Hall Voltage

Let us assume that the semiconductor is a *p*-type semiconductor. Let a potential difference V be applied across its ends. A current of strength I flows through it along the x -direction (Fig. 30.23 b). Holes are the majority charge carriers in the *p*-type semiconductor. The current through the wafer is given by

$$I = peAv_d \quad (30.111)$$

where p is the hole concentration

A is the area of cross-section of the end face of semiconductor wafer,

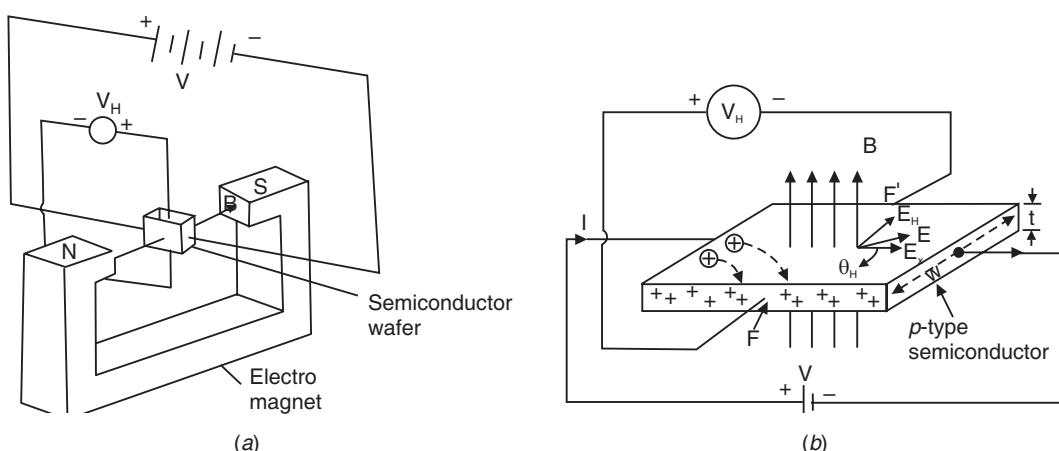


Fig. 30.23: (a) Basic experimental arrangement to study Hall effect (b) Generation of Hall voltage.

e is the electrical charge associated with a hole, and v_d is the average drift velocity of holes.

$$\text{The current density} \quad J_x = \frac{I}{A} = pev_d \quad (30.112)$$

Any plane perpendicular to the current flow direction is an equipotential surface. Therefore, the potential difference between the front and rear faces F and F' is zero (see Fig. 30.24).

Now, if a magnetic field B is applied normal to the wafer surface and hence to the direction of current flow in it, then a transverse potential difference is produced between faces F and F' . It is known as **Hall Voltage** V_H .

Before the application of magnetic field, holes move parallel to faces F and F' . On application of magnetic field B , the holes experience a sideways deflection due to the magnetic force F_L , which is given by

$$F_L = eBv_d \quad (30.113)$$

Holes are deflected toward the front face F and pile up there. Due to this, a corresponding equivalent negative charge is left on the rear face F' . These opposite charges produce a transverse electric field, E_H , whose direction is from the front to the rear face. Due to the action of E_H , holes experience an electric force in addition to the Lorentz force. When the force F_E due to this transverse electric field balances the magnetic force F_L , equilibrium condition is attained and the holes once again flow along x -direction parallel to the faces F and F' .

In the equilibrium condition

$$\therefore eE_H = ev_d B$$

$$\text{If 'w' is the width of the semiconductor wafer, } E_H = \frac{V_H}{w}$$

$$\therefore \left(\frac{V_H}{w} \right) = Bv_d \quad (30.114)$$

From equ. (30.112), we have $v_d = \frac{J_x}{pe}$

Therefore, we can write equ. (30.114) as

$$\therefore \frac{V_H}{w} = \frac{BJ_x}{pe} \quad (30.115)$$

$$V_H = \frac{wBJ_x}{pe} = \frac{wBI}{peA}$$

If 't' is the thickness of the semiconductor plate, $A = w t$.

$$\therefore V_H = \frac{BI}{pet} \quad (30.116)$$

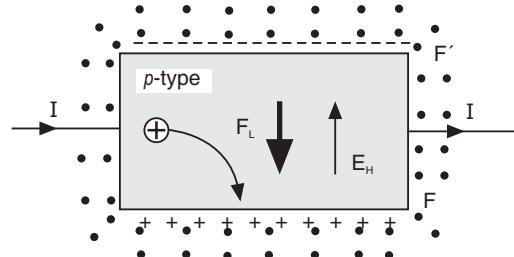


Fig. 30.24: Top view of the wafer—The directions of magnetic force and Hall field in *p*-type semiconductor

30.28.4 Hall Coefficient

Hall coefficient, R_H is defined as *Hall field per unit current density per unit magnetic induction*. Thus,

$$R_H = \frac{E_H}{J_x B} = \frac{V_H / w}{J_x B}$$

Using equ. (30.114), we get

$$R_H = \frac{BJ_x}{peJ_x B}$$

or

$$R_H = \frac{1}{pe} \quad (30.117)$$

Hall voltage, V_H can now be written as

$$V_H = R_H \frac{BI}{t} \quad (30.118)$$

$$\therefore R_H = \frac{V_H t}{BI} \quad (30.119)$$

The above equation is derived assuming that the p -semiconductor contains only holes. When the two types of charge carriers are taken into account, it is shown that the Hall coefficient is given by

$$R_H = \frac{(p\mu_h^2 - n\mu_e^2)}{e(p\mu_h + n\mu_e)^2} \quad (30.120)$$

From the above equation, we see that the Hall coefficient and Hall voltage are smaller for intrinsic materials than for extrinsic materials.

30.28.5 Drift Velocity

According to the equilibrium condition, namely $F_E = F_L$, we have

$$e\left(\frac{V_H}{w}\right) = eBv_d$$

$$v_d = \frac{V_H}{Bw} \quad (30.121)$$

As it is possible to experimentally measure the Hall voltage, magnetic field B and width of the wafer w , the mean drift velocity of the charge carriers can be obtained from the above equation.

30.28.6 Carrier Concentration

From the measurement of Hall voltage and current through the wafer and knowing the strength of magnetic induction and thickness of the wafer, R_H can be computed using equ. (30.119). Once R_H is known, the carrier concentration can be determined with the help of equ. (30.108), which may be rewritten as follows:

$$p = \frac{1}{R_H e}$$

In case of conductors and n -type semiconductors, the expression for Hall coefficient is

$$R_H = -\frac{1}{ne} \quad (30.122)$$

where n represents the concentration of electrons in the conductor or n -type semiconductor.

$$\therefore n = -\frac{1}{R_H e} \quad (30.123)$$

30.28.7 Doping Level

It is possible to estimate the doping concentration, N_A , from the value of p using the relation $p_p \equiv N_A$.

30.28.8 Carrier Sign and Type of Semiconductor

With the direction of the magnetic field and the current as depicted in Fig. 30.23 (b), the sign of the Hall voltage is positive. For an n -type semiconductor, the Hall voltage will be negative when the directions of B and I are kept same. Therefore, knowing the sign of Hall voltage, the type of the semiconductor and hence the sign of the majority carrier can be determined.

30.28.9 Hall Mobility

Mobility is defined as the drift velocity acquired in unit electric field. We know that the current density J is given by the following expressions.

$$\begin{aligned} J &= pev_d \text{ and } J = \sigma E \\ \therefore pev_d &= \sigma E \\ \text{or } \frac{v_d}{E} &= \frac{\sigma}{pe} \\ \therefore \mu_h &= \sigma R_H \end{aligned} \quad (30.124)$$

30.28.10 Hall Angle

The net electric field E in the semiconductor is a vector sum of E_x and E_H . It acts at an angle θ_H to the x -axis, where θ_H is called the **Hall angle**. From the Fig. 30.23 (b), we have

$$\tan \theta_H = \frac{E_H}{E_x} \quad (30.125)$$

$$\text{But } E_H = \frac{V_H}{w} = \frac{BJ_x}{pe}$$

$$\text{Also } E_x = \frac{J_x}{\sigma} = \rho J_x$$

$$\text{Thus, } \tan \theta_H = \frac{BJ_x}{pe\rho J_x}$$

$$\begin{aligned} \therefore \tan \theta_H &= \frac{B}{pe\rho} \\ \frac{E_H}{E_x} &= \frac{B}{pe\rho} \end{aligned} \quad (30.126)$$

Using $\frac{1}{pe} = R_H$ and $\frac{1}{\rho} = \sigma$, equ. (30.125) can be rewritten as

$$\tan \theta_H = \sigma R_H B$$

$$\text{or } \tan \theta_H = \mu_h B \quad (30.127)$$

30.28.11 Factors Affecting Hall Voltage

Hall voltage is given the following expression.

$$V_H = \frac{BI}{pet}$$

It is seen from the above equation that Hall voltage is directly proportional to the magnetic field strength and the current passing through the wafer. Therefore, the Hall voltage will be larger, the larger the magnetic field or the current. Further, the Hall voltage is larger, the smaller is the carrier concentration, and the thinner is the wafer. It is easy to select thin wafers of the material, so making t small. The fact that the concentration of carriers must be small implies that large Hall voltage is obtained with poor conductors, i.e. those with relatively few free electrons or holes per unit volume. Thus, the best voltage is obtained with semiconductors like silicon and germanium that have been lightly doped with impurities to make them conducting slightly.

For the semiconductors the number of charge carriers per unit volume is about $10^{23}/\text{m}^3$ while in case of metals it is about $10^{28}/\text{m}^3$. Therefore, the Hall voltage is about 10^5 greater in semiconductors than in metals.

30.28.12 Variation of Hall Coefficient with Temperature

In metals it was found that the Hall coefficient does not depend on temperature. Hence, it may be concluded that the concentration of free electrons does not vary with temperature in metals.

In semiconductors, the Hall coefficient decreases sharply with increase in temperature. It indicates that the concentration of charge carriers in semiconductors increases with increasing temperature.

30.28.13 Applications

Hall effect is widely used in various fields for a variety of applications. In almost all cases a Hall effect sensor is employed. A Hall effect sensor is a transducer that produces its output voltage in response to changes in magnetic fields.

1. **Determination of semiconductor type:** The Hall coefficient is negative for a p -type semiconductor and positive for a n -type semiconductor. Therefore, the sign of the Hall coefficient can be used to determine whether a given semiconductor is n - or p -type.
2. **Determination of carrier concentration:** By measuring the Hall coefficient, the carrier concentration in a semiconductor can be determined making use of the relations

$$n = \left| \frac{1}{R_H e} \right| \quad \text{or} \quad p = \frac{1}{R_H e}.$$

3. **Determination of carrier mobility:** By measuring the Hall coefficient and conductivity of the semiconductor, the carrier mobility can be determined using the relation

$$\mu_h = \sigma |R_H|$$

4. **Measurement of magnetic fields:** Hall voltage is proportional to the magnetic field intensity, for a given current through the sample. Therefore, one of the important applications of Hall effect consists in measuring magnetic fields. Knowing the parameters of the Hall probe, and applied current, we can determine the intensity of the magnetic field. Hall probes can be used for static as well as high-frequency magnetic fields. Hall probes measure variable magnetic fields up to a frequency of 10^{12} Hz.

5. Measurement of power in an electromagnetic wave: In an electromagnetic wave in free space, the electric and magnetic fields are at right angles. Now, if a semiconductor is kept parallel to E , it will produce a current I in the semiconductor. Since the semiconductor is simultaneously subjected to a transverse magnetic field, Hall voltage is produced across the sample. The Hall voltage is proportional to the product EH which represents power of the wave. Thus, Hall effect can be used to determine the flow of power of an electromagnetic wave.

6. Miscellaneous applications: Hall sensors are used for proximity switching, positioning, speed detection, and current sensing applications.

Current flowing through a conductor produces a magnetic field that varies with current, and a Hall sensor can be used to measure the current without interrupting the circuit. They are especially used in measuring extremely heavy currents, where conventional ammeters cannot be used.

A Hall sensor is combined with circuitry to act as a digital switch. Such switches are used in consumer equipment; for example in computer printers to detect missing paper and open covers.

Example 30.18: An *n*-type germanium sample has a donor density of $10^{21}/\text{m}^3$. It is arranged in a Hall experiment having magnetic field of 0.5 T and the current density is 500 A/m^2 . Find the Hall voltage if the sample is 3 mm wide.

Solution: Hall voltage

$$V_H = \frac{BI}{net} = \frac{BJA}{net} = \frac{BJwt}{net}$$

As

$$n = N_D$$

∴

$$\begin{aligned} V_H &= \frac{BJw}{N_D e} \\ &= \frac{(0.5 \text{ T})(500 \text{ A/m}^2)(3 \times 10^{-3} \text{ m})}{(10^{21}/\text{m}^3)(1.602 \times 10^{-19} \text{ C})} \\ &= 4.7 \times 10^{-3} \text{ V} = 4.7 \text{ mV} \end{aligned}$$

Units:

$$1 \frac{TAm^2}{C} = 1 \cdot \frac{N}{A \cdot m} \cdot \frac{A \cdot m^2}{C} = 1 \cdot \frac{N \cdot m}{C} = 1 \frac{J}{C} = 1 \text{ V}$$

Example 30.19. A copper strip 2.0cm wide and 1.0mm thick is placed in a magnetic field with $B=1.5 \text{ wb/m}^2$. If a current of 200 A is set up in the strip, calculate Hall voltage that appears across the strip. Assume $R_H = 6 \times 10^{-7} \text{ m}^3/\text{C}$.

Solution. Hall voltage

$$V_H = R_H \frac{IB}{t}$$

$$\begin{aligned} &= 6 \times 10^{-7} \text{ m}^3/\text{C} \cdot \frac{200 \text{ A} \times 1.5 \text{ wb/m}^2}{10^{-3} \text{ m}} \\ &= 0.18 \text{ V} \end{aligned}$$

Example 30.20. An electric field of 100 V/m is applied to a sample of *n*-type semiconductor whose Hall coefficient is $-0.0125 \text{ m}^3/\text{C}$. Determine the current density in the sample, assuming $\mu_e = 0.6 \text{ m}^2 / \text{Vs}$.

Solution. Current density

$$J = \frac{\mu_e E}{R_H}$$

$$= \frac{(0.36 \text{ m}^2/\text{V.s})(100 \text{ V/m})}{-0.0125 \text{ m}^3/\text{C}}$$

$$= -2880 \text{ C/s.m}^2 = -\mathbf{2880 \text{ A/m}^2}.$$

Example 30.21. In a Hall coefficient experiment, a current of 0.25 A is sent through a metal strip having thickness 0.2 mm and width 5 mm. The Hall voltage is found to be 0.15 mV when a magnetic field of 2000 gauss is used.

(a) What is the carrier concentration?

(b) What is the drift velocity of the carriers?

Solution. (a) The carrier concentration,

$$n = \frac{IB}{V_H e t}$$

$$= \frac{(0.25 \text{ A})(0.2 \text{ T})}{(0.15 \times 10^{-3} \text{ V})(1.602 \times 10^{-19} \text{ C})(0.2 \times 10^{-3} \text{ m})}$$

$$= \mathbf{1.04 \times 10^{25} \text{ carriers/m}^3}.$$

(b) Drift velocity of the carriers,

$$v_d = \frac{V_H}{wB}$$

$$= \frac{0.15 \times 10^{-3} \text{ V}}{5 \times 10^{-3} \text{ m} \times 0.2 \text{ T}}$$

$$= 0.15 \text{ V/m.T} = \mathbf{0.15 \text{ m/s}.}$$

QUESTIONS

1. Show that the intrinsic concentration n_i for a semiconductor is given by

$$n_i = (N_C N_V)^{1/2} e^{-E_g/2kT}$$

where the symbols have their usual meaning.

2. Write the expressions for electron and hole concentrations in an intrinsic semiconductors and hence derive the expression:

$$E_F = \frac{E_C + E_V}{2} + \frac{3}{4} kT \ln\left(\frac{m_h^*}{m_e^*}\right)$$

for Fermi level in the intrinsic semiconductor. Assume the symbols to have their usual meanings.

3. Draw a neat sketch of a band diagram of intrinsic semiconductor at room temperature and show that the Fermi level in an intrinsic semiconductor lies in the middle of the energy gap.

(R.T.M.N.U., 2007)

4. Derive an expression for Fermi energy in intrinsic semiconductor. What is the effect of temperature on Fermi level in an intrinsic semiconductor?
5. Describe a method of determining the band gap of a semiconductor. How does electrical conductivity vary with temperature for an intrinsic semiconductor? (Anna Univ., 2004)

6. Using the expressions of electron concentration and hole concentration for an intrinsic semiconductor show that the intrinsic carrier density is independent of Fermi level.
7. Obtain an equation for the conductivity of an intrinsic semiconductor in terms of carrier concentration and carrier mobility.
8. Why are holes not generated in metals? Can the electrical conduction of semiconductors be improved to the extent possible in metals? Discuss. **(R.T.M.N.U., 2007)**
9. Differentiate the *n*-type and *p*-type semiconductors with their Fermi level diagram. **(Bombay Univ.)**
10. P- and N-type semiconductors have more holes and electrons respectively. Explain why they are electrically neutral. **(Bombay Univ.)**
11. What is the effect of temperature on the working of *n*- and *p*-type semiconductors? Draw diagrams to show the variation in the position of Fermi level with rise in temperature. **(Bombay Univ.)**
12. Draw energy band diagrams for *n*-type semiconductor at 0 K and at room temperature. Show that at 0K Fermi level lies midway between valence band and conduction band. **(C.S.V.T.U., 2006)**
13. Distinguish between extrinsic and intrinsic semiconductors. **(Calicut Univ., 2005)**
14. Discuss the effect of increasing amounts of dopants on the Fermi level in extrinsic semiconductors.
15. Explain the effect of impurity concentration on the Fermi level in an extrinsic semiconductor. **(R.T.M.N.U., 2006)**
16. Define Hall effect. **(C.S.V.T.U., 2009)**
17. Explain Hall effect and obtain an expression for Hall coefficient for an extrinsic semiconductor. **(R.T.M.N.U., 2006)**
18. What is Hall effect? Obtain an expression for Hall coefficient. Does R_H depend on the doping concentration? **(R.T.M.N.U., 2005)**
19. What is Hall effect? Derive expressions for Hall voltage and Hall coefficient. Mention important applications of Hall effect. **(C.S.V.T.U., 2005, 2006)**
20. Explain in brief the concept of Fermi level. Show diagrammatically the Fermi level in metals, intrinsic semiconductors and insulators at 0°K and at higher temperature.
21. Explain Hall effect and its significance. Give its applications. **(R.T.M.N.U., 2006)**
22. How does the Fermi level change with temperature in extrinsic semiconductors? Discuss the effect of increasing amounts of dopants in extrinsic semiconductors.
23. What do you understand by intrinsic and extrinsic semi-conductors?
24. Write short notes on:
 - (i) Elemental and compound semiconductors.
 - (ii) Intrinsic and extrinsic semiconductors. **(Calicut Univ., 2006)**
25. Explain in brief the concept of Fermi level. Derive an expression for Fermi energy in intrinsic semiconductor. What is the effect of temperature on Fermi level in an intrinsic semiconductor? Draw a neat energy band diagram of intrinsic semiconductor at 0 K and at room temperature.
26. Differentiate between an intrinsic and an *n*-type semiconductor on the basis of (i) Crystal representation and (ii) band representation and (iii) relation between *n* and *p*.
27. Draw energy band diagrams for
 - (a) Intrinsic semiconductor (b) *n*-type semiconductor
 - (c) *p*-type semiconductor
 at 0 K and room temperature. **(R.T.M.N.U., 2005)**
28. An *n*-type extrinsic semiconductor is in equilibrium at room temperature. What is the net charge on the impurity atom? Discuss.
29. How does the Fermi level change with temperature in extrinsic semiconductors? Discuss the effect of increasing amounts of dopants in extrinsic semiconductors.

30. Derive the expression: $E_F = \frac{E_g}{2} + \frac{3}{4}kT \ln\left(\frac{m_h^*}{m_e^*}\right)$, for Fermi level in an intrinsic semiconductor.

31. For a n-type semiconductor material with high doping concentration and low temperature show that: $n = (N_c N_d)^{\frac{1}{2}} \exp\left(\frac{E_D - E_g}{2kT}\right)$ Where the symbols have their usual meaning. (Anna Univ., 2004)

32. Obtain an expression for density of electrons in the conduction band of an n-type extrinsic semiconductor by assuming Fermi-Dirac distribution. (Anna Univ., 2004)

33. Obtain an expression for density of holes in the valence band of p-type extrinsic semiconductor. (Anna Univ., 2004)

34. What is law of mass action? How does this lead to the relation $np = n_p^2$, where symbols have their usual meaning? (R.T.M.N.U., 2005)

35. Derive the expression for the Fermi energy at any temperature for an n-type semiconductor. (Calicut Univ., 2006)

36. Write short note on: Drift and Diffusion currents. (C.S.V.T.U., 2006)

37. Define Hall effect and Hall coefficient. Obtain an expression for the Hall coefficient. (Calicut Univ., 2005)

38. What is Hall effect? Derive a formula for density of charge carriers in a p-type semiconductor give two of its applications.

39. What is Hall Effect? How does this effect show whether holes or electrons predominate in a semiconductor?

40. Explain the term Hall effect. Dervie the relation between Hall voltage and Hall coefficient. (G.T.U., 2009)

41. Explain Hall effect and its importance. Show that the ratio of Hall electric field E_H to the electric field E which is responsible for the current in a n-type semiconductor wafer kept in a uniform magnetic field B is given by

$$\frac{E_H}{E} = \frac{B}{nep} \quad \text{(C.S.V.T.U., 2008)}$$

42. What is Hall effect? Give expressions for each of the following:

 - (a) Hall coefficient
 - (b) Hall Voltage
 - (c) Hall angle
 - (d) Hall mobility

Mention some applications of Hall effect. (C.S.V.T.U., 2007)

43. Show that the minimum conductivity of a semiconductor sample occurs when $n_o = n_i \sqrt{\mu_p / \mu_n}$.

PROBLEMS

- If effective mass of an electron is equal to twice the effective mass of hole, determine the position of the Fermi level in an intrinsic semiconductor from the centre of forbidden gap at room temperature. **[Ans: The Fermi level is 0.014 eV below the centre of forbidden gap]**
 - Determine the fraction of electrons in conduction band in silicon at 27°C and 227°C. Given: $E_g = 1.1$ eV and $k = 1.38 \times 10^{-23}$ J/K. **[Ans: 5.7×10^{-10} ; 2.8×10^{-6}]**
 - A silicon wafer is doped with 10^{21} phosphorus atoms/m³. Calculate
 - The majority carrier concentration
 - The minority concentration and
 - The electrical resistivity of the doped silicon at room temperature. Assume complete ionization of the dopant atoms; $n_i = 1.5 \times 10^{16}$ /m³, $\mu_e = 0.135$ m²/V.s. and $\mu_h = 0.048$ m²/V.s.

4. The resistivity of intrinsic silicon at 270°C is 3000 Ωcm. Calculate the intrinsic carrier density.
Assume: $\mu_e = 0.17 \text{ m}^2/\text{V.s.}$ and $\mu_h = 0.035 \text{ m}^2/\text{V.s.}$ [Ans: $1.0 \times 10^{16}/\text{m}^3$]
5. Calculate the temperature at which Si ($E_g = 1.14 \text{ eV}$) will have the same fraction of electrons in the conduction band as germanium ($E_g = 0.72 \text{ eV}$) has at 300 K. [Ans: 475K]
6. Compute the current produced in a Ge plate of 1 cm² area, 0.03 mm thickness across its faces. Assume a free electron concentration of $2 \times 10^{19} \text{ m}^{-3}$, the electron and hole mobility's being 0.39 m²/V.s and 0.19 m²/V.s respectively. [Ans: 12.4 A]
7. Find the fraction of electrons in the valence band of intrinsic germanium which can be thermally excited across the forbidden energy gap of 0.7 eV into the conduction band at:
(i) 50 K
(ii) 300 K
(iii) 1000 K [Ans: $5.1 \times 10^{-36}; 1.3 \times 10^{-6}; 0.017$]
8. What fraction of the conductivity of intrinsic silicon at room temperature is due to
(a) Electrons and
(b) Holes? [Ans: 73.8% due to electrons and 26.2% due to holes]
The electrons and the hole mobility's are, $\mu_e = 0.135 \text{ m}^2/\text{V.s.}$ & $\mu_h = 0.048 \text{ m}^2/\text{V.s.}$
9. The energy gap in silicon crystal is 1.12 eV. Its electron and hole mobilities at room temperature are 0.48 m²/V.s and 0.013 m²/V.s respectively. Find its conductivity. [Ans: $1.29 \times 10^{-3} \Omega^{-1} \text{ m}^{-1}$]
10. The forbidden band gap in pure silicon is 1.1 eV. Compare the number of conduction electrons at temperature 27°C and 37°C.
11. A silicon wafer is doped with 10^{21} phosphorus atoms/m³. Calculate
(i) The majority carrier concentration
(ii) The minority concentration and
(iii) The electrical resistivity of the doped silicon at room temperature. Assume complete ionization of the dopant atoms;
 $n_i = 1.5 \times 10^{16} / \text{m}^3$
 $\mu_e = 0.135 \text{ m}^2/\text{V.s.}$
 $\mu_h = 0.048 \text{ m}^2/\text{V.s.}$ [Ans: $10^{21}/\text{m}^3; 2.25 \times 10^{11}/\text{m}^3; 0.046 \Omega \cdot \text{m}$]
12. A sample of intrinsic silicon at room temperature has a carrier concentration of $1.5 \times 10^{16} \text{ m}^{-3}$. If a donor impurity is added to the extent of 1 donor atom per 10^8 atom/m³, determine the resistivity of the material. Given: $\mu_e = 0.135 \text{ m}^2/\text{V.s.}$; $\mu_h = 0.048 \text{ m}^2/\text{V.s.}$ [Ans: 0.09Ω.m]
13. A current density of 10^3 A/m^2 flows through an n-type germanium crystal which has resistivity 0.05 ohm-m. Calculate the time taken for electrons in material to drift a $50 \mu\text{m}$ distance. The mobility of electron is $0.38 \text{ m}^2/\text{V.s.}$ [Ans: $2.6 \times 10^{-6} \text{ s}$]
14. A sample of Ge is doped to the extent of 10^{20} donor atoms/m³ and 7×10^{19} acceptor atoms/m³. At the temperature of the sample the resistivity of intrinsic germanium is 0.6 ohm-m. If the applied electric field is 200 V/m, find the total conduction current density.
Assume: $\mu_e = 0.38 \text{ m}^2/\text{V.s.}$, $\mu_h = 0.18 \text{ m}^2/\text{V.s.}$ [Ans: 431.6 A/m^2]
15. An n-type Ge sample has a donor density of $10^{21} / \text{m}^3$. It is arranged in Hall Effect experiment having magnetic field of 0.5 Tesla and current density 500 A/m^2 . Find the Hall voltage, if the sample is 3 mm wide. [Ans: 4.7 mV]

CHAPTER

31

Semiconductor Diodes

31.1 INTRODUCTION

Pure semiconductors are of very limited use. Semiconductors that are doped with impurities form the basis of the practical devices. A semiconductor that has been doped with acceptor impurities and into the surface of which donor atoms are diffused forms a *p-n* junction diode. A *p-n* junction diode is also known as a **semiconductor diode**. The most remarkable property of the *p-n* junction is that it allows current flow in one direction and opposes it in the opposite direction. This property is known as *rectifying action*. Semiconductor diodes are widely used as rectifiers, which convert input ac voltage to dc voltage. Practically, all semiconductor devices contain at least one *p-n* junction. The production techniques enable the fabrication of *p-n* junction to suit specific purposes. Thus, a varicap that acts as a variable capacitor, a tunnel diode and a Gunn diode as oscillators, a zener diode as a voltage stabilizer, a photodiode as a light detector, a solar cell as a voltage source, an LED and a laser as light sources are all *p-n* junctions. A junction transistor is fabricated with two *p-n* junctions in close proximity. Therefore, *p-n* junction constitutes the most basic component of solid state devices and a thorough understanding of its electrical behaviour is essential for appreciation of the operation of many semiconductor devices.

31.2 p-n JUNCTION DIODE

By doping an intrinsic semiconductor crystal, it can be converted into an *n*-type or *p*-type extrinsic semiconductor crystal. When taken individually, *p*-type and *n*-type semiconductors conduct with equal facility in both the directions just like a resistor (Fig. 31.1) and exhibit *linear* conduction characteristic. Such behaviour is known as **ohmic behaviour**. It is indeed possible to grow a semiconductor crystal with part of the crystal *n*-type and the other part *p*-type. ***p-n* junction** is the boundary between one region of a semiconductor with *p*-type impurities and another region containing *n*-type impurities. The technology of junction fabrication is beyond the scope of our discussion. The most remarkable property of the *p-n* junction is that it has a *non-linear* conduction characteristic.

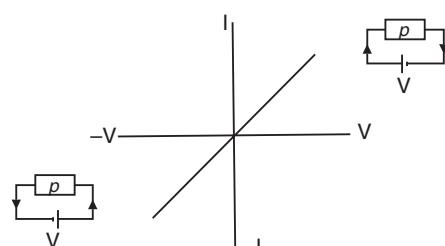


Fig. 31.1

(see Fig. 31.2) and allows current in one direction and opposes it in the opposite direction. This is known as *rectifying property*.

p-n junction diodes are widely used in rectifiers, which convert ac voltages to dc voltages. A *p-n* junction is the basic component of solid state devices and is used as a varicap, a tunnel diode, a Gunn diode, a zener diode, a photodiode, a solar cell, an LED, a laser, etc.

31.2.1 A Physical Description of the p-n Junction

A *p-n* junction is formed when a *p*-type and an *n*-type semiconductor are joined metallurgically. Within the semiconductor block there is a more or less abrupt transition from *p*-type to *n*-type material. This transition is called a **junction**. In the ideal case, the transition from one type to the other would take place abruptly. Such junctions are known as **abrupt junctions**. The distance over which the change from *p*-type to *n*-type occurs takes place in a very short length much less than $1\mu\text{m}$.

We assume here uniform *p*-doping on one side of the junction and uniform *n*-doping on the other side; and that doping concentrations at either side of the junction are identical, i.e., $N_A = N_D$.

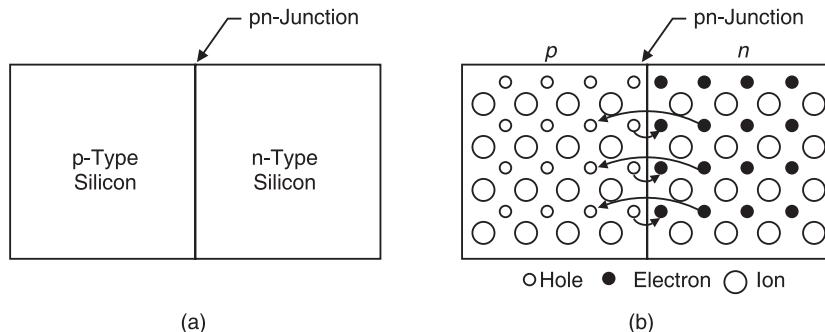


Fig. 31.3: *p-n* junction (a) a *p-n* junction is formed when *p*-type and *n*-type semiconductors are metallurgically joined (b) majority carriers diffuse across the junction at the instant of formation of the junction.

Our primary interest lies in understanding the formation of the junction and the properties of junction. The line in Fig. 31.3 (a) implies some sort of barrier between the two regions. Without such a barrier all the free electrons in the *n* type region gradually diffuse into the holes in the *p* region, leaving neither free electrons nor holes there. In the beginning, when the crystal was formed, it was without a barrier and electrons began to diffuse into holes and formed a transition zone, which had neither free electrons nor holes. The transition zone is the barrier between the regions. The question now concerns the nature of this barrier and its genesis.

31.2.2 Formation of p-n Junction

The *p*-region contains holes as the majority carriers which are produced by the acceptor atoms and very few thermally generated electrons, which are minority carriers. The *n*-side contains electrons as the majority carriers contributed mainly by the donor atoms and very few thermally generated holes, which are minority carriers. The impurity atoms are ionized and firmly fixed in the lattice through covalent bonds. Before contact, each side is electrically

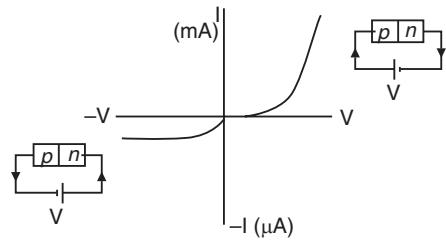


Fig. 31.2

neutral. One hole is associated with each ionized acceptor in the *p*-type material and one electron with each ionized donor on *n*-side. Thus, the charges on mobile carriers balance the fixed ion charges in each region. When the two regions are joined, the following events take place and lead to the formation of *p-n* junction.

31.2.3 Diffusion of Majority Carriers and Formation of Space Charge

When the two sections are brought together, the holes in the *p*-region are more than the holes in the *n*-region. Because of the concentration difference, holes nearer to the junction begin to diffuse from *p*-side to the *n*-side (see Fig. 31.3 *b*). Similarly, the electrons in the *n*-region are more than the electrons in the *p*-region. Due to the concentration gradient, electrons nearer to the junction begin to diffuse from *n*-side to the *p*-side. Thus, majority carriers start moving into opposite regions. As electrons and holes are charged particles, their motion produces *electron diffusion current*, J_{en} and *hole diffusion current*, J_{hp} respectively. The current components due to holes and electrons add up although the carriers are moving in opposite directions (the electron current is opposite to the direction of electron flow). The first letters ‘*h*’ and ‘*e*’ of the subscripts in the designation denote the carrier and the second letters ‘*p*’ and ‘*n*’ indicate the region of their origin.

At the junction the holes and electrons meet each other and undergo recombination. As a hole recombines with an electron, both the hole and electron disappear. This leads to the disappearance of mobile charge carriers in the junction region.

The majority holes diffusing out of the *p*-region leave behind acceptor ions (N_A) on *p*-side of the junction, thus producing negative ions in a previously neutral region (see Fig. 31.4). In the same way, the electrons diffusing from the *n*-side leave behind the uncompensated positively ionized donor atoms (N_D). The double layer of ions around the junction is known as the **space charge region**. This narrow space-charge region is depleted of mobile charges and contains only the immobile uncompensated ions. Therefore, this region is also called the **depletion region**, which is about $1 \mu\text{m}$.

The array of fixed ions produces **electric field**, E , which is directed from the donor ions on *n*-side toward the acceptor ions on *p*-side. This electric field is in a direction that opposes the diffusion of majority carriers into opposite sides.

The space charge layers tend to reduce the diffusion of holes and electrons. The holes leaving *p*-region must overcome the repulsion of the positive ion layer on *n*-side of the junction in order to continue their forward movement. Similarly, the electrons diffusing from the *n*-side must overcome the repulsion of the negative ion layer on *p*-side of the junction (Fig. 31.5a). Thus, the **internal electric field acts as a barrier** to the flow of majority charge carriers. The barrier increases

till majority carriers cannot diffuse further across the junction. As the barrier builds up, diffusion of majority carriers is halted. The process is self-arresting. Ultimately, the electric field establishes an equilibrium potential difference V_0 across the depletion region. V_0 is known as the **internal potential barrier**. Thus, the very electric field which is produced due to the diffusion of the majority carriers across the junction, inhibits the diffusion.

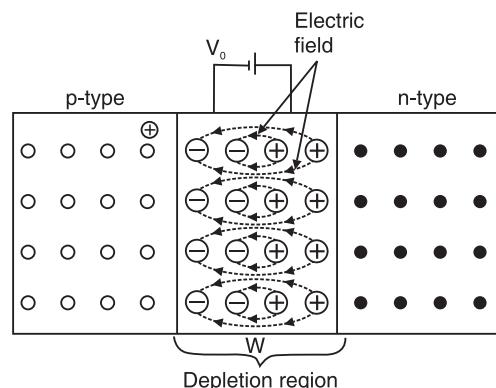


Fig. 31.4

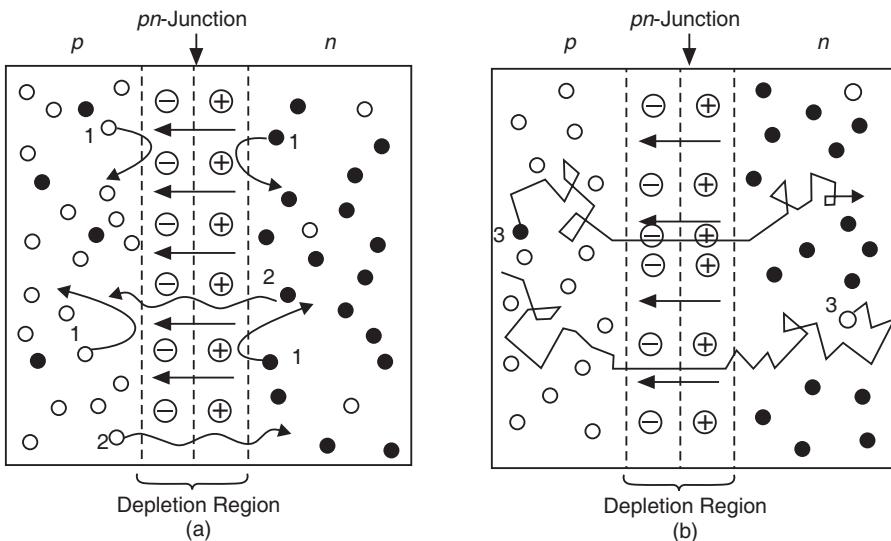


Fig. 31.5

However, occasionally a majority hole in *p*-region diffuses into *n*-region whenever it has energy equal to or greater than ' eV_o '. Similarly, a majority electron in *n*-region diffuses into the *p*-region when it has a minimum energy ' eV_o ' (see particles 2, Fig. 31.5a).

The initial diffusion of charge carriers and the creation of the resultant barrier potential occur when the junction is formed during the crystal growth process. Without such a barrier all the free electrons in the *n*-type region gradually diffuse and recombine with neither the holes in the *p*-region, leaving neither free electrons nor holes.

The diffusion of majority carriers causes *diffusion current* to flow across the junction. It is easy to see that the current components due to holes and electrons add up although the carriers are moving in opposite directions (the electron current is opposite to the direction of electron flow). The net diffusion current density flowing across the junction is given by

$$J(\text{diff}) = J_{hp} + J_{en} \quad (31.1)$$

31.2.4 Space Charge and Drift of Minority Carriers

It may however be noted that the field due to space charge has the right direction to cause the flow of *minority carriers* across the junction. Electrons reaching the edge of the junction on *p*-side are accelerated by the electric field into *n*-region and similarly, the holes reaching the edge of the junction on *n*-side are accelerated into *p*-region (see particles 3, Fig. 31.5b).

As a consequence, an electric current flows across the junction. This current, which is caused by electric field, is called *drift current*. We may readily see that the current components due to drift motion of holes and electrons are in the same direction and add to each other. If we designate the hole drift current density as J_{hn} and the electron drift density as J_{ep} , then the net drift current through the junction is

$$J(\text{drift}) = J_{hn} + J_{ep} \quad (31.2)$$

31.2.5 Thermal Equilibrium Condition

When external voltage is not applied and no net current flows across the *p-n* junction, it is said to be in **equilibrium condition**. At thermal equilibrium the net diffusion current through the junction must be equal and opposite to the net drift current so that the total current is zero, as it must be for an open-circuited device (see Fig. 31.6).

$$J(\text{diff.}) = J(\text{drift})$$

i.e.,

$$J_{hp} + J_{en} = J_{hn} + J_{ep} \quad (31.3)$$

However at any given time there are very few carriers within the transition region, since the electric field sweeps out carriers which have wandered into the junction region.

The equilibrium condition demands further that the hole and electron currents must be *separately* zero. Holes and electrons are in transit from one side of the junction to the other. Some holes diffuse from *p*-side to *n*-side, and some holes drift from *n* to *p* side under the action of the junction field. Conversely, some electrons diffuse from *n*-side to *p*-side and some drift from *p* to *n*-side. In the equilibrium condition, the neutrality of *p*- and *n*-regions is preserved.

Since there can be no net build up of holes or electrons on either side as a function of time, the drift and diffusion currents must cancel for *each* type of carrier. In other words the hole drift current must exactly cancel the hole diffusion current and the electron drift current must cancel the electron diffusion current. Thus,

$$J_{hp} - J_{hn} = 0$$

$$J_{en} - J_{ep} = 0$$

or

$$\left. \begin{array}{l} J_{hp} = J_{hn} \\ J_{en} = J_{ep} \end{array} \right\} \quad (31.4)$$

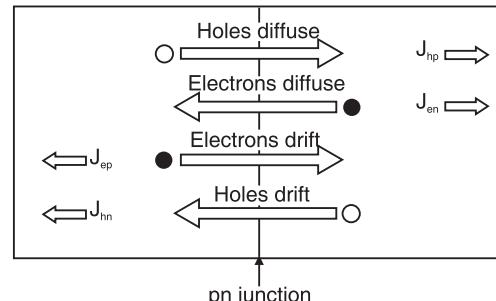


Fig. 31.6

31.2.6 Energy Band Diagram of p-n Junction at Equilibrium

Let us now understand the formation of *p-n* junction from the point of view of energy band structure. When the two semiconductors are in contact, equilibrium is attained only when there is no net current flow across the junction region. Current can be zero when the probability of occupancy of a given energy level is the same in both the semiconductor regions. It implies that the reference energy level, namely Fermi level E_F must be at the same level in the two semiconductor regions. If it is not so, the different probability of occupancy in neighbouring regions would lead to carrier migration and current flow. Fermi level is similar to the liquid level in a container, temperature of a body etc. Their equality throughout a medium ensures no-flow-condition and equilibrium state. In the same way, an energy band structure having equalized Fermi level in both the regions is characteristic of thermal equilibrium of a *p-n* junction.

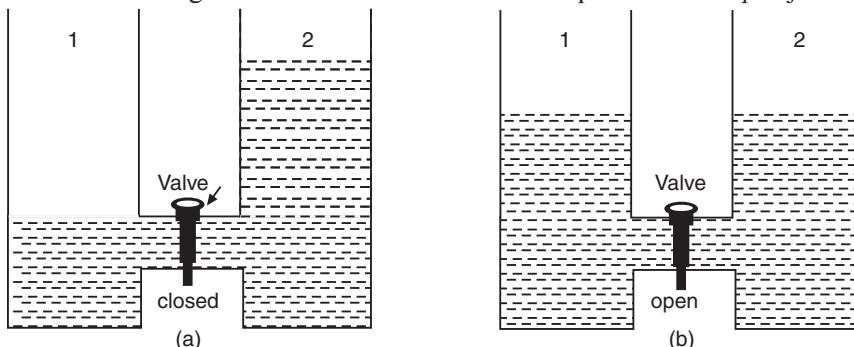


Fig. 31.7

The equalization of Fermi levels in two different regions can be understood with the help of the following analogy. Let us consider two containers (1) and (2) which are interconnected through a valve (Fig. 31.7a). Initially, the valve is closed and the containers are filled with water to different levels. When the valve is opened, water flows from container (2) into container (1). As a result, the water level in container (1) rises and water level in container (2) falls down. The process continues till the levels in both the containers are equalized as shown in Fig. 31.7 (b).

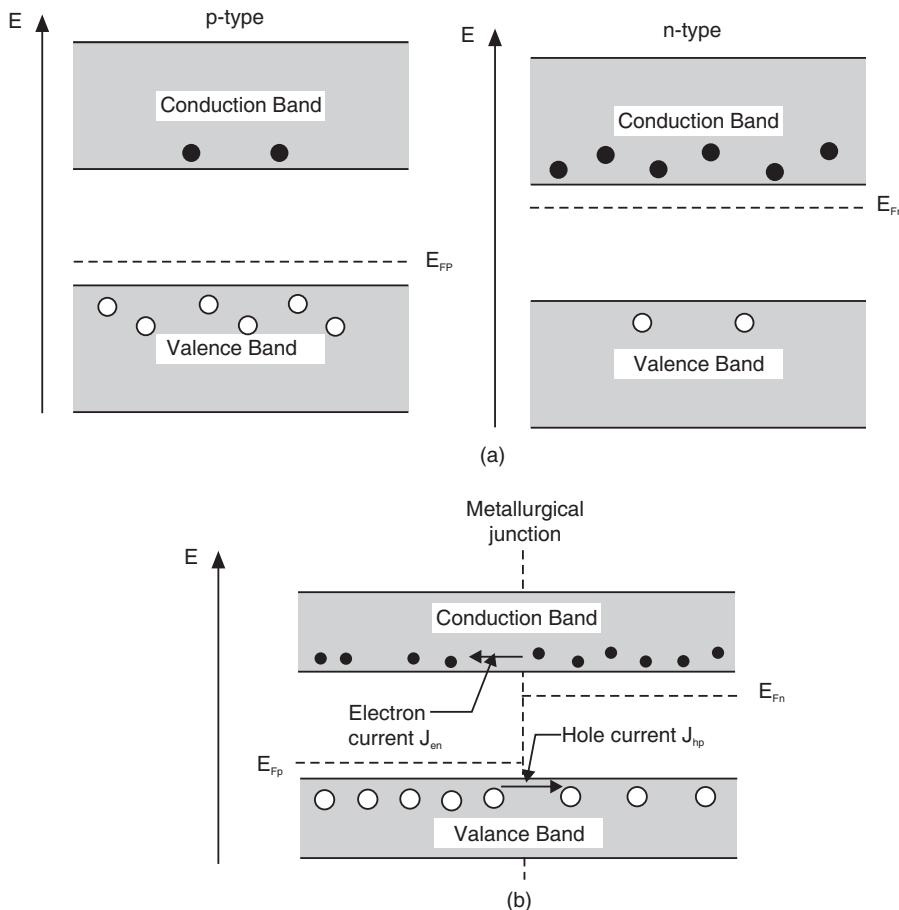


Fig. 31.8

The energy band diagrams of the individual p - and n -semiconductors are shown in Fig. 31.8 (a). Note that the Fermi levels E_{Fp} and E_{Fn} are at different levels. At the instant of joining (Fig. 31.8b) the levels in the two semiconductor regions are not aligned. The occupancy of energy levels by electrons in the conduction band on n -side is high while it is low on p -side. Therefore, the electrons occupying the energy levels in the conduction band on n -side move into the conduction band levels on p -side. Similarly, the occupancy of energy levels by holes in the valence band on p -side is high while it is low on n -side. Hence, the holes occupying the energy levels in the valence band on p -side move into the valence band levels on n -side. As high energy electrons leave n -region, the Fermi level E_{Fn} which represents the average energy of electrons moves downwards. Since the Fermi level is fixed relative to the band structure of

the region, its movement causes downward shift of the entire band structure in the *n*-region. On the *p*-side, holes having higher energy leave the valence band in that region. The direction of decrease in hole energy is upward and hence the Fermi level E_{Fp} moves upward. Along with E_{Fp} , the entire band structure in the *p*-region shifts upward. The shifting of energy bands continues till the energy levels E_{Fp} and E_{Fn} attain the same level in both the regions. When the two levels are equalized, the carrier migration comes to a halt and equilibrium condition is established.

The displacement of the energy bands in opposite directions on both the sides causes a bending of the energy bands in the junction region. Fig. 31.9 shows the energy band diagram of a *p-n* junction in equilibrium condition. Each side takes up a different electrostatic potential.

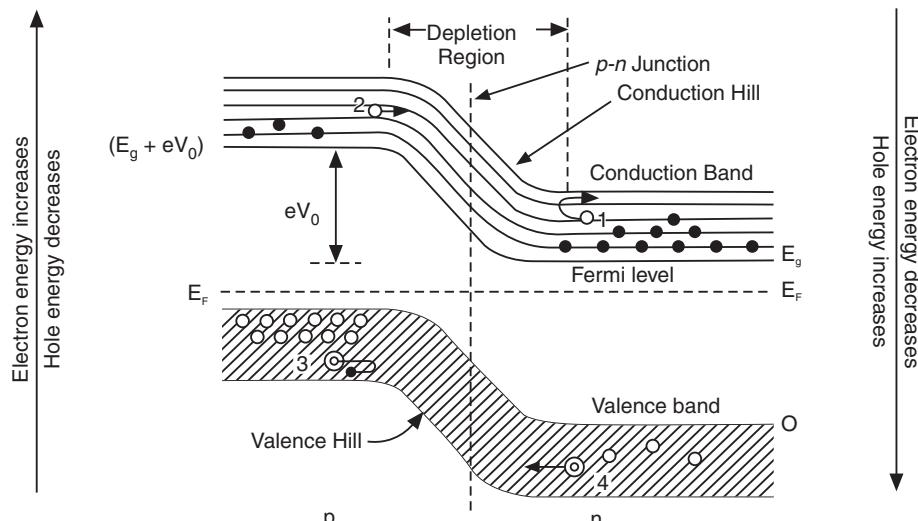


Fig. 31.9: Energy band diagram of *p-n* junction at equilibrium

It results in a potential barrier V_o or an energy hill of height eV_o . Electrons in the conduction band of *n*-region face an energy hill, namely *conduction hill*. Electrons approaching the junction region cannot surmount the conduction hill unless they have a minimum energy of eV_o . For example, electron marked (1) in Fig. 31.9 fails to climb the conduction hill. Occasionally, a few of the electrons that have kinetic energy equal to or greater than eV_o overcome the conduction hill and go into *p*-region. On the other hand, the electrons near the junction in *p*-region can roll down the conduction hill effortlessly and pass into *n*-region. For example, the electron marked (2) in Fig. 31.9 can roll down easily the conduction hill. The components of currents due to such migration of electrons are in opposite directions and balance each other.

As the direction of increasing energy is downward for holes, the holes in the valence band of *p*-region encounter an energy hill, namely *valence hill*. The holes on the *p*-side cannot go into the *n*-region unless they have a minimum energy of eV_o . The hole marked (3) in Fig. 31.9 fails to surmount the valence hill whereas a few holes having kinetic energy equal to or greater than eV_o succeed in going into *n*-region. On the other hand, holes near the junction on the *n*-side can readily float up the hill irrespective of their energy. For example, hole marked (4) in Fig. 31.9 floats up the valence hill. The two components of current due to the opposite flow of holes balance each other.

Thus, the current due to occasional diffusion of majority carriers is balanced by the occasional drift of minority carriers and the net current across the p - n junction is zero.

31.2.6.1 Calculation of the internal potential barrier, V_o

The magnitude of the potential barrier V_o can be estimated from the knowledge of the electron concentrations in p - and n -region of the diode. Referring to Fig. 31.9. E_g is the edge of the conduction band on the n -side. The electron concentration in the conduction band on the n -side can be written as

$$n_n = N_C \exp[-(E_g - E_F)/kT] \quad (31.5)$$

The edge of the conduction band on the p -side is given by $(E_g + eV_o)$. The electron concentration on p -side can be expressed as

$$n_p = N \exp[-\{(E_g + eV_o) - E_F\}/kT] \quad (31.6)$$

Dividing equ. (31.5) by equ. (31.6) we get

$$\frac{n_n}{n_p} = \exp\left[\frac{eV_o}{kT}\right] \quad (31.7)$$

The relation (31.7) shows that at thermal equilibrium the concentrations of electrons on both sides of the junction are related through the Boltzmann factor $e^{eV_o/kT}$. The concentrations of holes on both the sides are related by an equation similar to (31.7). Taking logarithms on both sides of equation (31.7), we obtain

$$V_o = \frac{kT}{e} \ln \frac{n_n}{n_p} \quad (31.8)$$

The above equation can be written as

$$V_o = \frac{kT}{e} \ln \frac{n_n p_p}{n_p p_p}$$

At room temperature, all the impurities are ionized and therefore, we can write

$$n_n = N_D \text{ and } p_p = N_A$$

Further,

$$p_p n_p = n_i^2$$

Using these relations, we can write equation (31.8) as

$$V_o = \frac{kT}{e} \ln \frac{N_D N_A}{n_i^2} \quad (31.9)$$

The factor kT/e has the dimensions of voltage and is denoted by V_T . Writing $kT/e = V_T$ in equ. (26.9), we get

$$V_o = V_T \ln \frac{N_D N_A}{n_i^2} \quad (31.10)$$

Equ. (31.10) indicates that the barrier potential in a junction diode depends on the equilibrium concentrations of the impurities in p - and n -regions and does not depend on the charge density in the depletion region.

Example 31.1. Calculate the potential barrier for a germanium pn junction at room temperature, if both the p - and n -regions are doped equally and to the extent of one atom per 10^6 germanium atoms.

Solution.

$$V_o = \frac{kT}{e} \ln \frac{N_A N_D}{n_i^2}$$

$$\begin{aligned}
 &= \frac{1.38 \times 10^{-23} J / ^\circ K \times 300 K}{1.602 \times 10^{-19} C} \ln \frac{(4.4 \times 10^{22})^2}{(2.4 \times 10^{19})^2} \\
 &= 0.0258 \times 16.12 V = \mathbf{0.42 \text{ Volts}}
 \end{aligned}$$

31.2.6.2 Circuit symbol and bias

A *p-n* junction diode is schematically represented by the symbol shown in Fig. 31.10. The arrowhead indicates the conventional direction of current flow when the diode is forward biased. The *p*-side of the diode is positive and is called *anode*. The *n*-side is the *cathode* and is the negative terminal when the diode is forward biased.

When a dc voltage is applied to a device, the device is said to be **biased**. A *p-n* junction can be biased in two ways.

If the dc voltage is connected in such a way that the positive terminal of the source is connected to the *p*-region and the negative terminal to the *n*-region, then the junction is said to be **forward biased**.

The junction is said to be **reverse biased** when the positive terminal of the external source is connected to the *n*-region and the negative terminal to the *p*-region.

31.3 p-n JUNCTION UNDER FORWARD BIAS

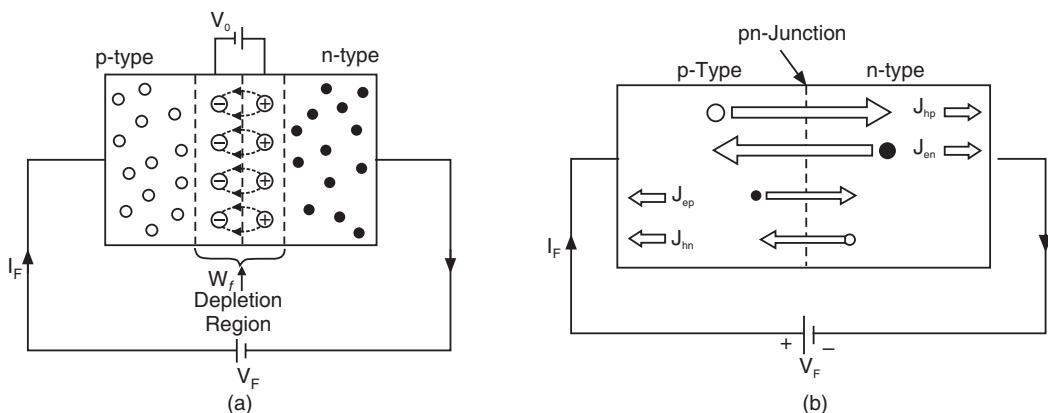


Fig. 31.11: (a) *p-n* junction under forward bias (b) Current components across forward biased junction

When a dc voltage, V_F , is connected to the diode in such a way that the positive terminal of the source is connected to the *p*-region and the negative terminal to the *n*-region, then the junction is forward biased (Fig. 31.11 a). Since the junction region is depleted of mobile charges, its resistance is very high. The resistance of the remaining parts of the semiconductor, where mobile charges are in plenty, is very low. Therefore, the forward bias voltage V_F appears across the junction and acts opposite to the internal potential V_o . The effective voltage across the junction is $(V_o - V_F)$ now. Therefore, the potential barrier is reduced from

eV_o to $e(V_o - V_F)$. The majority carriers push into the depletion region and the width of the depletion region is reduced from the equilibrium value of W to W_f . As a result, the diffusion of majority carriers across the junction increases. The diffusion current density increases from J_{diff} to a value J_{diff}^* .

However, the reduction in the potential does not influence the minority carriers. As usual, they are in a favourable position to move into opposite regions. Hence the drift current density J_{drift} remains unaffected by the forward bias. Consequently, $J_{diff}^* > J_{drift}$. Therefore, a net current density equal to $J_{diff}^* - J_{drift}$ flows through the junction (Fig. 31.11 b).

31.3.1 Energy Band Diagram

The steady state situation of the energy band structure under forward bias condition is shown in Fig. 31.12. The negative terminal of the external voltage source causes an increase in electron energy and an upward shift of all energy levels on the *n*-side. Similarly, the positive terminal connected to *p*-side causes an increase in hole energy and hence a lowering of all levels on *p*-side. As the displacements of the energy levels occur in opposite directions, the Fermi levels E_{Fn} and E_{Fp} get separated by an amount of energy eV_F added by the voltage source. As a result of the shifts, the heights of the potential barrier are reduced by an amount of energy eV_F to a value of $e(V_o - V_F)$. Hence, the energy required by the majority carriers to move into opposite regions is reduced to $e(V_o - V_F)$ from eV_o . Due to the reduction of height of the barrier, the movements of the majority carriers is promoted. As a result, the components J_{hp} and J_{en} increase.

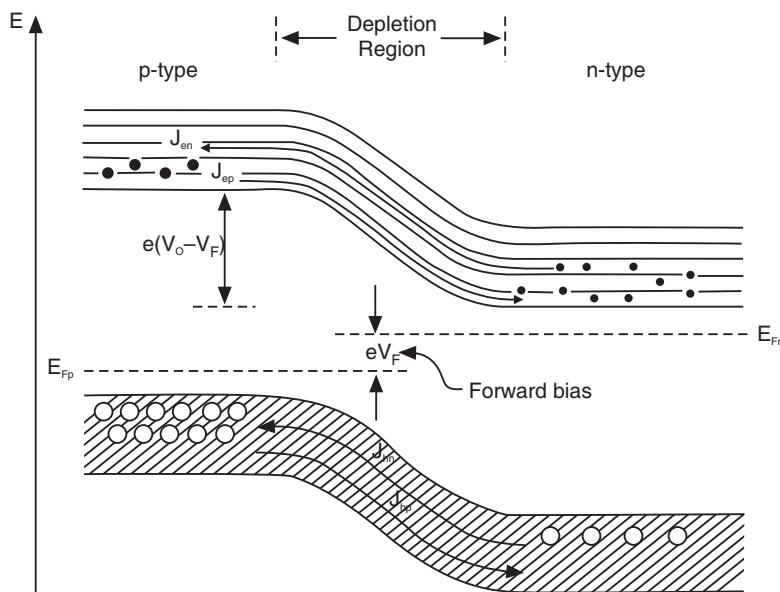


Fig. 31.12: Energy band diagram of *p-n* junction under forward bias

The reduction in the height of the potential barrier does not affect the minority carrier movement. As usual those carriers which are near the depletion region are in a favourable position to cross the junction. Electrons slide down the conduction hill from *p*-side to *n*-side, while holes float up the valence hill from *n*-side to *p*-side. The small drift current density components J_{ep} and J_{hn} due to minority carriers do not change due to forward bias.

Therefore, a net current flows through the junction, which is determined by the diffusion of majority carriers.

31.3.2 The Law of the Junction

Note that the majority carriers that cross the junction and arrive in the opposite region become minority carriers in that region. Therefore, due to forward bias the number of minority carriers in that region increases to a very large value. The increase in minority carrier concentration can be estimated as follows.

Under equilibrium condition, the electron concentration on *p*-side of the junction is given by equ. (31.7). It may be rewritten as

$$n_p = n_n \exp\left[-\frac{eV_o}{kT}\right] \quad (31.11)$$

Due to the application of forward bias, V_o reduces to $(V_o - V_F)$. It causes an increase in the value of n_p^* . Let n_p^* be the enhanced concentration of electrons at the junction on *p*-side.

Then, equ. (31.11) may be written as

$$\begin{aligned} n_p^* &= n_n \exp\left[-\frac{e(V_o - V_F)}{kT}\right] \\ &= n_n \exp\left[-\frac{eV_o}{kT}\right] \cdot \exp\left[\frac{eV_F}{kT}\right] \end{aligned} \quad (31.12)$$

or

$$n_p^* = n_p \exp\left[\frac{eV_F}{kT}\right] \quad (31.13)$$

Equ. (31.13) indicates that upon applying a small forward bias, an exponential increase in the number of electrons occurs on *p*-side of the junction. To get an idea of increase in electron concentration on *p*-side, let us assume that $V_F = 0.1$ volt. The new value of the electron concentration on *p*-side is

$$n_p^* = n_p \exp\left[\frac{0.1}{0.025}\right] \approx 50 n_p$$

31.3.3 Injection of Minority Carriers

The penetration of majority carriers through the *p-n* junction under forward bias and the consequent increase in the number of minority carriers on the opposite side of the junction is called the **injection of minority carriers**.

When the doping levels on the two sides are different, the junction is said to be asymmetric and the number of carriers injected from the region with higher doping level exceeds the number of carriers injected from the weakly doped region. The region from which larger number of carriers is injected is called the **emitter** and the region into which the injection is directed is called the **base**.

Note that the majority carrier concentrations are not affected in each region by the application of forward bias. Each region acts as a reservoir of majority carriers and does not suffer in strength when some of the carriers leave the region, since the number of carriers leaving the region is only a small fraction.

31.3.4 Diffusion of Minority Carriers

It is seen from the equ. (31.13) that under forward bias, a large concentration of minority carriers are found at the depletion layer edge of both the sides. For example, the hole

concentration is large at the depletion layer edge on *n*-side and the electron concentration suddenly rises on *p*-side edge. The concentration of minority carriers is low in the rest of the bulk of the material. Thus, the forward bias produces a concentration gradient for minority carriers in each region.

Consequently, diffusion of excess minority carriers takes place, away from the junction and into the bulk of the semiconductor on both sides of the depletion layer in an attempt to reach a state of equilibrium. The concentrations of minority carriers fall off rapidly with distance from the junction. This is due to the reason that they are surrounded by majority carriers and a swift recombination of carriers takes place. If the supply were limited, the concentration of these carriers would be that at thermal equilibrium. However, as the majority carriers diffuse across the junction, new carriers are supplied due to the forward bias. Therefore, thermal equilibrium is never reached and a continuous diffusion current is maintained. The external voltage source replenishes the carriers lost through recombinations. The processes of diffusion and recombination transport charge through the device and cause current flow. The current continues to flow in the device as long as the external voltage source is connected in the circuit.

It should be recognized at this stage that the flow of minority carriers in the two regions is due to the concentration gradient rather than due to the electric field established by the battery. There will be only an insignificantly small voltage drop across the bulk of the semiconductor due to its high conductivity, while the major portion of the voltage appears across the depletion layer which is a high resistance region. As a result, the concentration gradient for minority carriers is more effective in causing diffusion of minority carriers.

31.3.5 Carrier Movement in the Forward Bias Circuit

The movement of charge carriers through the forward biased *p-n* junction may be traced as follows:

An electron leaves the negative terminal of the battery and moves toward the *n*-region. It enters the *n*-region and becomes a majority carrier. Gradually, it drifts towards the junction. As it approaches the junction, the electron slows down under the repelling influence of electric field acting across the junction. While slowing down it gives its energy to the neighbouring electrons through collisions. The electron that absorbs the extra energy overcomes the barrier and goes into *p*-region, where it becomes a minority carrier. In the *p*-region the electron diffuses due to concentration gradient. As it diffuses, it is surrounded by a large number of holes and the electron recombines with one of the holes. Recombination means that a free electron becomes a valence electron. Thus, the moving electron suddenly stops and its kinetic energy is communicated to other valence electrons in the lattice. As a result, a covalent bond is broken elsewhere and an electron-hole pair is generated. The electron of the pair diffuses forward and it in turn undergoes recombination after some time. Another electron-hole pair is generated. This process goes on till the electron reaches the terminal end of the *p*-region. At the end, out of the electron-hole pair, the hole drifts towards the junction while the electron flows into the connecting wire and returns to the positive terminal of the battery. This completes the circuit. The current density in the circuit is composed of two components J_e and J_h , as shown in Fig. 31.13.

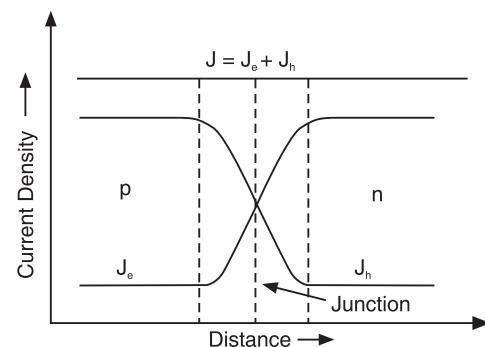


Fig. 31.13

31.4 p-n JUNCTION UNDER REVERSE BIAS

Fig. 31.14 shows a reverse biased *p-n* junction.

It is seen that the external voltage V_R adds to the barrier voltage V_o and the potential barrier increases to $e(V_o + V_R)$. As a result, the electric field in the depletion region increases and the majority carriers are pushed farther from the depletion region. This action leads to an increase in the width of the depletion region from W_f to W_{rb} (see Fig. 31.14). The increase in potential barrier inhibits the diffusion of majority carriers and the diffusion current falls to zero at higher bias values (Fig. 31.15). However, the increase in the potential barrier does not influence minority carriers. The increase in the electric field increases the acceleration of minority carriers without causing a change in their number. The drift current density remains the same as in unbiased and forward bias conditions. The current through the junction is only due to drift current caused by the minority carriers. The components of drift current J_{hn} and J_{ep} are shown in Fig. 31.15. The current is of the order of a nanoampere or less. Thus, in a reverse biased *p-n* junction

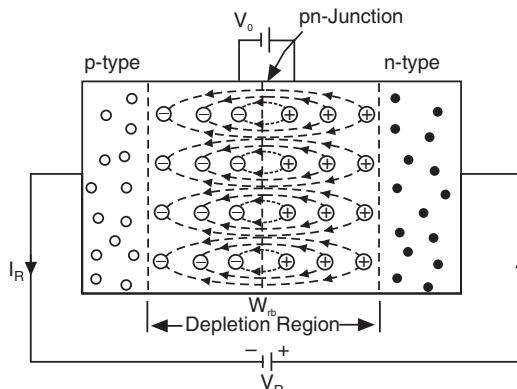


Fig. 31.14

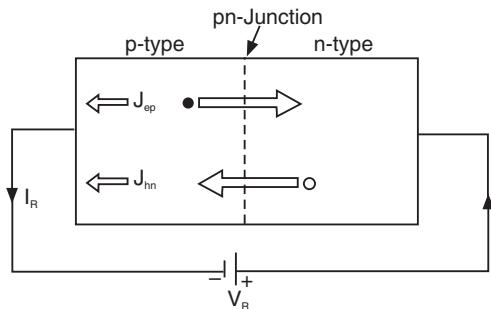


Fig. 31.15

$$\begin{aligned} J_{diff} &= 0 \\ J_{drift} &= -(J_{hn} + J_{ep}) = -J_o \end{aligned} \quad (31.14)$$

31.4.1 Energy Band Diagram

The energy band diagram for a *p-n* junction under reverse bias condition is shown in Fig. 31.16. The positive terminal of the voltage source, V_R connected to the *n*-side reduces the energy of the electrons. Therefore, the energy levels on *n*-side are displaced downwards. Similarly, the negative terminal connected to *p*-side reduces the energy of holes on *p*-side. Therefore, the energy levels on *p*-side are displaced upwards. Due to such displacement of energy levels on both sides, the Fermi levels E_{Fn} and E_{Fp} get separated by an amount of energy eV_R . Now, the barrier height increase to a value of $e(V_o + V_R)$. The majority carriers in *n*-region, electrons cannot climb the conduction hill to go into *p*-region. Similarly, the majority carriers in the *p*-region, holes cannot float up the valence hill to go into *n*-region. The diffusion of majority carriers is totally stopped. However, the barrier does not influence the drift motion of the minority carriers. The electrons generated near the depletion layer edge on *p*-side can easily slide down the conduction hill and the holes generated near the depletion layer edge on *n*-side can easily float down the valence hill. Thus, the minority carriers can cross the junction and cause the flow of drift current. This current is independent of the height of the potential

hills, as the minority carriers are only falling from a higher energy region into lower energy regions. Therefore, the diffusion current density becomes zero while the drift current density is the same as that in equilibrium condition.

31.5 THE DIODE EQUATION

When the diode is forward biased, the potential barrier is lowered by an amount of energy eV_F and the probability of a majority carrier crossing the junction is increased by a factor $e^{eV_F/kT}$ [see equ. (31.13)].

Therefore, the diffusion current density increases by a factor $e^{eV_F/kT}$. Thus, the diffusion current density components J_{hp}^* and J_{en}^* in a forward biased diode are given by

$$J_{hp}^* = J_{hp} e^{eV_F/kT} = J_{hn} e^{eV_F/kT} \quad (31.15)$$

$$J_{en}^* = J_{en} e^{eV_F/kT} = J_{ep} e^{eV_F/kT} \quad (31.16)$$

where J_{hp} and J_{en} are diffusion current densities in unbiased diode. The drift current density components have not changed and have the same magnitude as in equilibrium case.

Therefore, the net hole current density across the forward biased junction is

$$J_h = J_{hp}^* - J_{hn} = J_{hn} (e^{eV_F/kT} - 1) \quad (31.17)$$

Similarly, the net electron current density across the junction is

$$J_e = J_{en}^* - J_{ep} = J_{ep} (e^{eV_F/kT} - 1) \quad (31.18)$$

The total current density across the forward biased $p-n$ junction is a sum of electron and hole current density components, i.e.,

$$\begin{aligned} J &= J_e + J_h \\ &= (J_{hn} + J_{ep}) (e^{eV_F/kT} - 1) \\ &= J_o (e^{eV_F/kT} - 1) \end{aligned} \quad (31.19)$$

where $J_o = (J_{hn} + J_{ep})$.

If the area of cross section of the junction is A , then the current is $I = JA$. Therefore,

$$I = I_o (e^{eV_F/kT} - 1) \quad (31.20)$$

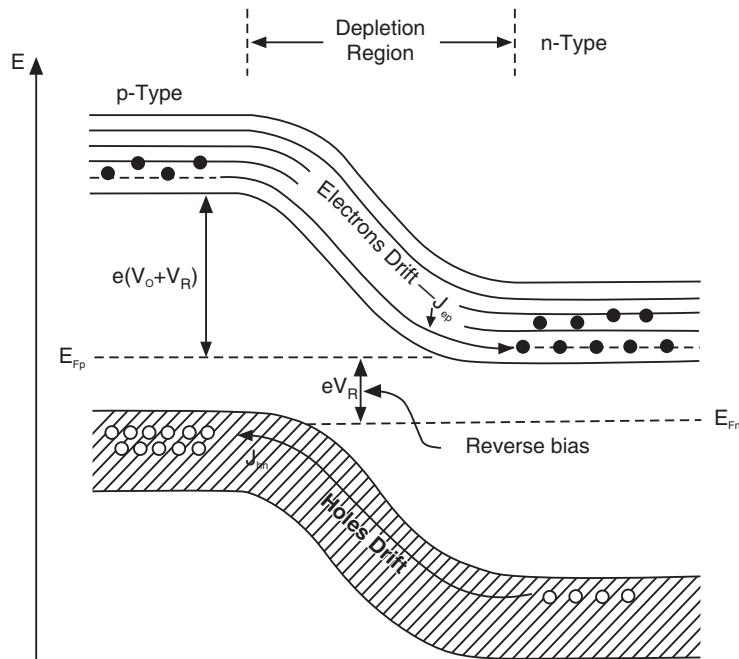


Fig. 31.16: The energy band diagram for a pn junction under reverse bias condition

where

$$I_o = J_o A.$$

When the diode is reverse-biased, the connection from the voltage source is reversed and a voltage ($-V_R$) is applied to the diode. Therefore, using ($-V_R$) in place of V_F in the equation (31.19), we get

$$J = J_o (e^{-eV_R/kT} - 1) \quad (31.21)$$

For larger values of V_R ,

$$e^{-eV_R/kT} \rightarrow 0.$$

∴

$$J = -J_o \quad (31.22)$$

∴

$$I = -I_o \quad (31.23)$$

Equations (31.19) and (31.21) can be combined into a single equation by denoting the forward and reverse voltages by a single symbol V . Thus, we write

$$J = J_o (e^{eV/kT} - 1) \quad (31.24)$$

In terms of current,

$$I = I_o (e^{eV/kT} - 1) \quad (31.25)$$

where $V = V_F$ for forward bias and $V = -V_R$ in case of reverse bias.

Eq. (31.25) is known as the **rectifier equation or diode equation**.

The term I_o is known as **reverse saturation current**.

Example 31.2: Current flowing in a *p-n* junction is $0.2 \mu\text{A}$ at room temperature when a large reverse bias voltage is applied. Calculate the current when a forward bias of 0.1 V is applied.

Solution: The current flowing through a junction is given by $I = I_o [e^{eV/kT} - 1]$

$$\therefore I = 2 \times 10^{-7} A [e^{(0.1 \text{ V})/(0.026 \text{ V})} - 1] = 9.2 \mu\text{A}.$$

31.6 VOLTAGE-AMPERE CHARACTERISTIC

A graph that shows the variation in current in a device with the variation of voltage applied across it is called **volt-ampere characteristic**. Fig. 31.17 shows the volt-ampere characteristic of a *p-n* junction diode. It is seen that the characteristic is not linear and hence a *p-n* junction is a **non-linear** device. The *p-n* junction acts as a *closed switch* in forward bias condition allowing large current to flow through it; and acts as an *open switch* in reverse bias condition causing an insignificantly low current through it. This unidirectional conduction is called **rectifying action**.

For a forward bias voltage V_F , the current through the junction is negligibly small as long as V_F is less than V_o . The voltage V_o at which the current raises sharply is called the **cut-in voltage**. It is in fact the same as barrier voltage. As V_F increases beyond V_o , the barrier disappears and an exponential increase in forward current I_F occurs. The forward current is equal to the difference between the diffusion current and drift current.

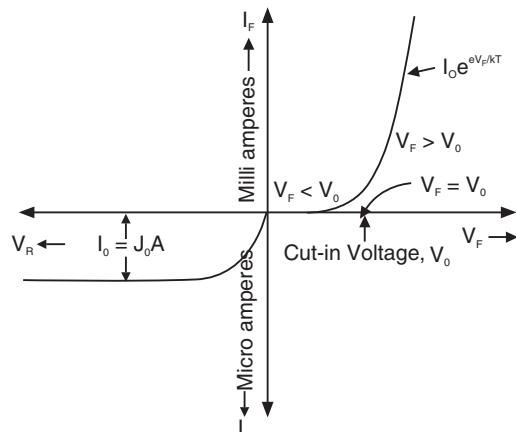


Fig. 31.17

$$I_F = I_o e^{eV_F/kT} - I_o = I_{\text{diffusion}} - I_{\text{drift}} \quad (31.26)$$

When V_F exceeds V_o , the barrier collapses and the drift current becomes nearly zero. However, the barrier does not totally vanish. The bulk resistance of the diode causes the bias voltage to be distributed over the entire length of the semiconductor regions. Therefore, at higher values of V_F the characteristic becomes practically linear (i.e., ohmic).

The reverse current through the junction is very much smaller than the forward current. Therefore, a different scale is used for plotting the reverse bias characteristic. The reverse current I_R is given by

$$I_R = I_{\text{drift}} - I_{\text{diffusion}} = (-I_o) - I_o e^{-eV_R/kT} \quad (31.27)$$

In the initial stage the reverse current increases rapidly due to an exponential decrease of diffusion current with increasing voltage. At higher V_R values, the diffusion current drops to zero and the reverse current becomes equal to the drift current due to the minority carriers. As the number of minority carriers at a given temperature is constant and does not change with the bias voltage, the reverse current stays constant with increase in the reverse bias.

31.6.1 Reverse Saturation Current

The reverse saturation current I_o is constant and is caused by the minority carriers. Minority carriers are generated in the material due to rupture of covalent bonds and have succeeded in arriving into the depletion region. The number of minority carriers that can come into the depletion region is determined by the diffusion length. Electrons generated in p -region at distances greater than the diffusion length L_e from the junction cannot reach the depletion region since they undergo recombinations before reaching the depletion layer. So is the case with holes in n -region. Holes generated in n -region at distances greater than the diffusion length L_h from the junction cannot reach the depletion region since they undergo recombinations before reaching the depletion layer. The diffusion length is of the order of 10^{-4} m in case of germanium. Therefore, only those minority carriers generated in the neighbourhood of the outer edges of depletion layer succeed in coming under the influence of internal electric field. Only those carriers are accelerated across the junction and cause the drift current.

Minority carriers are generated in the material due to rupture of covalent bonds which depends only on the temperature of the material. As long as the temperature remains constant, the rate of generation of minority carriers is constant. Therefore, the current due to their flow is the same whether the applied reverse bias voltage is small or large. Hence, the drift current due to minority carriers is known as **reverse saturation current**. It is denoted by I_o . It is very small as the number of minority carriers is very small. It is of the order of nanoamperes in silicon $p-n$ junction and microamperes in germanium $p-n$ junction.

31.6.2 Reverse Breakdown

Ordinary $p-n$ junction normally does not conduct when it is reverse biased. From Fig. 31.17 it is seen that the reverse saturation current in a semiconductor diode is negligibly small. However, if the reverse bias voltage is increased gradually, a point is reached where the junction breaks down and starts conducting heavily. This critical value of voltage is called the **breakdown voltage**. Once the breakdown occurs, even a very small increase in voltage causes large change in the reverse current. In normal operation the condition of breakdown should be avoided as it permanently damages the crystal structure of the two regions and renders the junction useless. As the breakdown phenomenon is irreversible and damages the diode permanently, ordinary diodes are never operated in this region.

The breakdown occurs in reverse biased diode mainly due to two different mechanisms.

- (i) **Avalanche breakdown:** This type of breakdown occurs in lightly doped junctions when high reverse voltage is applied across the junction. The avalanche breakdown

occurs as follows. When high reverse bias is applied to the diode, the electric field in the depletion region will be sufficiently high. Minority electrons entering the depletion region from the *p*-side are likely to acquire high kinetic energy. The high energy electrons collide with host atoms and remove valence electrons from some of the covalent bonds. The new carriers in turn produce additional charge carriers and the process multiplies and an avalanche of carriers is produced in a very short time to give large reverse current. The avalanche breakdown is self-sustaining as long as the reverse voltage is present across the depletion region. The breakdown voltage is found to increase with temperature.

- (ii) **Zener breakdown:** The Zener breakdown mechanism occurs in thin, abrupt and heavily doped *p-n* junctions and requires relatively low reverse voltage for its operation.

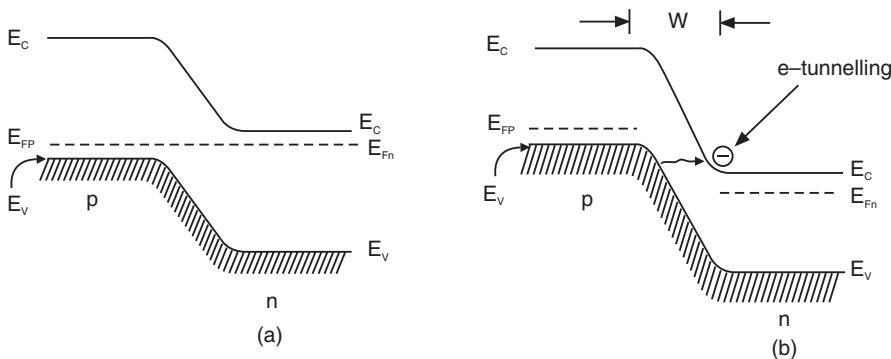


Fig. 31.18

When the *p* and *n* regions are heavily doped, the depletion region becomes very thin, of the order of 10 nm. Application of reverse bias voltage causes the conduction and valence bands to bend to the extent that the *n*-side conduction band appears opposite to the *p*-side valence band. As a result, a number of vacant states in the *n*-side conduction band are brought directly opposite to large number of filled states in the *p*-side valence band (Fig. 31.18). As the barrier separating the two bands is very narrow, the electrons tunnel through it under the effect of a small reverse bias. The electrons from the valence band on *p*-side reach the conduction band on *n*-side, thus causing a sudden large reverse current from *n* to *p* side. This is the *zener effect*. Zener breakdown takes place usually at low reverse voltages of the order of 4 volts or less in silicon diodes.

31.7 APPLICATIONS

A *p-n* junction diode is used as a rectifier, a switch, a clumper etc. We shall study here the application of diode as a rectifier.

Rectifiers

Government supplies electrical energy in the form of ac voltage because it is more economic and efficient to distribute ac power. However, dc voltage is required for operating most of the electronic equipments and gadgets. It is therefore necessary to convert the incoming mains supply to dc voltage. The process of converting ac voltage into dc voltage is called **rectification** and the device or circuit that converts ac voltage into dc voltage is called **rectifier**. The unidirectional current conduction property of diode is used in rectifiers. A single diode or more diodes can be connected into a circuit to form different types of rectifier circuits.

31.7.1 Half-Wave Rectifier

A circuit employing a single diode whereby the diode conducts during only one half-cycle of the input ac cycle is called a **half-wave rectifier**. The basic half wave rectifier circuit is shown in Fig. 31.19(a).

In the circuit, an alternating voltage is applied to a single diode connected

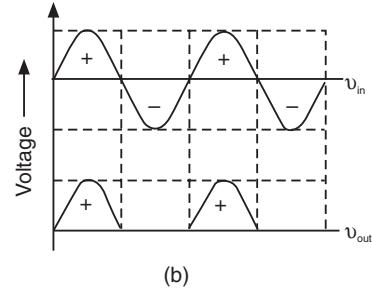
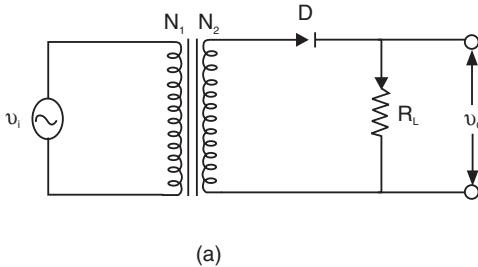


Fig. 31.19: Half-wave rectifier (a) Circuit diagram (b) input and output voltage wave forms

resistor R_L . Resistor R_L represents a device to which the circuit delivers power. As the device draws power and performs work, it is called **load**. Therefore, R_L is called **load resistor**. The input voltage v_i is a sinusoidal voltage, which changes in polarity 100 times in a second. During the positive half cycle of the input voltage the diode is forward biased and offers a very low resistance. As a result, a large current I_L flows through the load resistor. A voltage v_0 develops across R_L . The voltage drop across the diode is usually negligible. Therefore, the entire power appears across R_L . During the negative half cycle of the input waveform, the diode is reverse-biased and offers a very high resistance. Therefore, only a negligible reverse saturation current flows through the circuit. As such $I_L = 0$ and hence $v_0 = 0$. It means that the negative half-cycle of input voltage waveform is suppressed and is not utilized for delivering power to the load. In this condition all the input voltage appears across the diode itself.

The input and output waveforms are shown in Fig. 31.19 (b). It is seen that the output voltage is no longer an ac voltage. It is a **unidirectional fluctuating voltage**. It has an average dc value over which a number of ac components are superimposed. The undesired ac components are called the **ripple**. In case of the half-wave rectifier, the lowest ripple frequency is the same as the frequency of the input voltage.

31.7.2 Full Wave Rectifier

A full wave rectifier uses two diodes and rectifies both the half cycles of the input voltage. The circuit is shown in Fig. 31.20 (a).

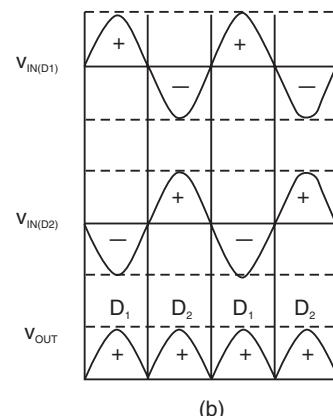
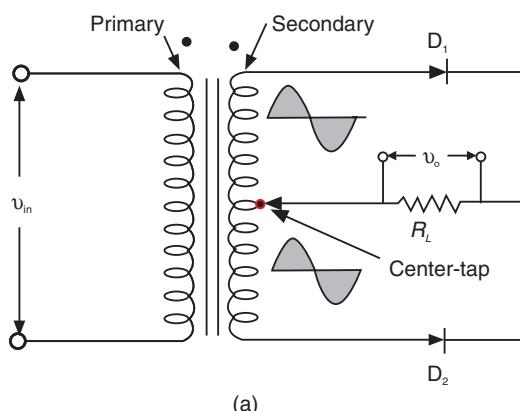


Fig. 31.20: Full-wave rectifier (a) Circuit diagram (b) input and output voltage wave forms

The input is supplied to the full wave rectifier from a transformer with a center-tapped secondary winding. During the positive half-cycles of the secondary voltage, the diode D_1 is forward biased and the diode D_2 is reverse biased. The current i_L flows through the diode D_1 , the load resistor R_L and the upper half-winding of the secondary. The output voltage v_0 develops across R_L . During the negative half cycles, the diode D_1 is reverse biased and the diode D_2 is forward biased. The current now flows through the diode D_2 , the load resistor R_L and the lower half winding of the secondary. The direction of the current flow through the resistor R_L is the same in both the cases. Hence during both the alternations, current i_L passes through R_L and produces an output voltage v_0 . Thus, in the full wave rectifier, both the half cycles are utilized to produce the output. The output consists of a continuous series of positive half cycles of alternating voltage. The input and output waveforms are shown in Fig. 31.20 (b).

The pulsating dc voltage obtained from a rectifier is smoothed out with the help of **filter circuits** and a stable dc voltage is obtained.

31.8 ZENER DIODE

Zener diode is a semiconductor diode specially designed to operate in the breakdown region of the reverse bias. Zener diodes are always operated in reverse bias condition. By varying the impurity concentration and other parameters, it is possible to design the breakdown voltage to suit specific applications. In Zener diodes, the breakdown phenomenon is reversible and harmless.

Fig. 31.21 (a) shows the symbol of a zener diode. It may be seen that it is like that of an ordinary diode except the bar is turned into Z-shape. The volt-ampere characteristic of the Zener diode is shown in Fig. 31.21 (b). Zener diode acts very much

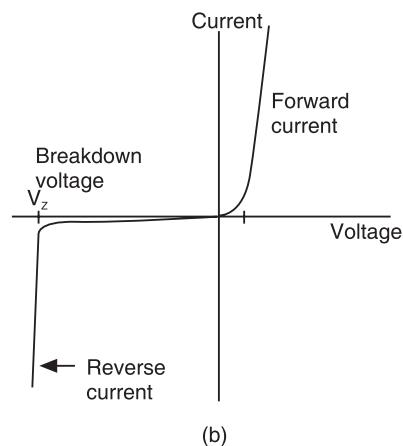
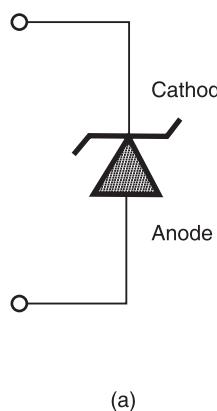


Fig. 31.21

similar to ordinary diode under forward bias condition. However, the zener diode is generally not used in the forward bias condition. The reverse bias characteristic is very much different from that of an ordinary diode. As the reverse voltage is increased, the reverse current remains constant till a certain value is reached. At that value, the reverse current increases abruptly. The voltage at which such sudden increase in reverse current occurs is called *zener breakdown voltage* or *zener voltage*, V_Z . The breakdown voltage of an ordinary diode is high, but if a reverse current above that value is allowed to pass through it, the diode is permanently damaged. **Zener diodes** are designed so that their zener voltage is much lower - for example just 2.4 Volts. When a reverse voltage above the zener voltage is applied to a zener diode, there is a *controlled breakdown* which does not damage the diode. In the zener region the voltage across zener diode remains constant but the current changes depending on the supply voltage. The voltage drop across the zener diode is equal to the zener voltage of that diode no matter how high the reverse bias voltage is above the zener voltage.

The location of zener region can be controlled by varying doping levels. An increase in doping will decrease the zener potential. Zener diodes are available in the range from 2V to 200V.

Applications

Zener diodes are widely used in electronic circuits. Some of the important uses are in

1. voltage regulators
2. reference element
3. meter protection etc.

ZENER DIODE AS A VOLTAGE REGULATOR

Voltage regulation is the ability of a circuit to maintain a constant output voltage even when either input voltage or load current varies. A zener diode can serve as a voltage regulator.

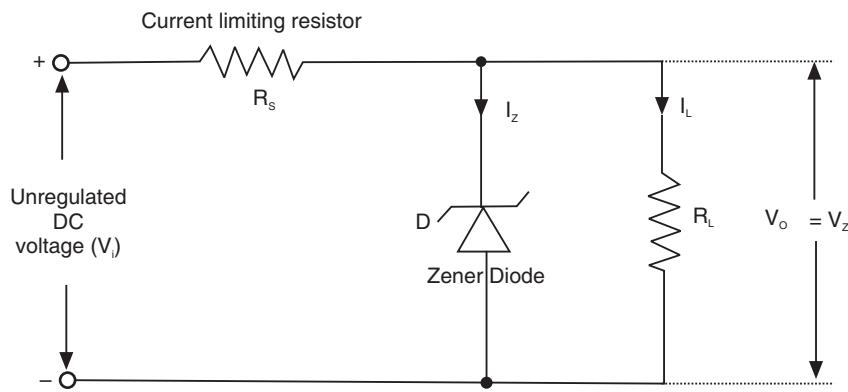


Fig. 31.22

It can be used to provide a constant voltage from a source whose voltage may vary over a considerable range. Fig. 31.22 shows the use of zener diode as a voltage regulator. An unregulated power supply provides the input voltage V_i , R_s is the current limiting resistor and R_L is the load resistor. The zener diode of zener voltage V_Z is reverse connected across V_i . The load resistance R_L across which constant output is desired is connected parallel with the diode. The series resistance R_s absorbs the output voltage fluctuations so as to maintain constant voltage across the load. Zener diode has maximum current rating equal to I_{zmax} . The zener current should not exceed this limit. Minimum current may go to zero.

The total current I passing through R_s equals the sum of diode current and load current, i.e.,

$$I = I_z + I_L$$

This circuit makes use of the fact that under reverse bias breakdown voltage, the voltage across the zener diode remains constant even if larger current is drawn. Since the load resistance, R_L is parallel to the zener diode, the voltage across the load resistance does not vary even though current through the load changes. Hence, the voltage across the load is regulated against the variations in the load current.

Operation

Case 1: We assume that the load resistance R_L is constant and V_i is varying.

As R_L is constant, I_L is also constant because $I_L = \frac{V_z}{R_L}$. But supply current keeps changing

due to change in V_i . The current through R_s is $I = \frac{V_i - V_z}{R_s}$ and is the sum of $I = I_z + I_L$.

If V_i increases, then the current I will increase. But as V_z and R_L are constant, the load current I_L will remain constant. Naturally, the increase in current will increase the zener current I_z . Thus, the increase in I will be absorbed by the zener diode without affecting I_L . The increase in V_i results in a larger voltage drop across R thereby keeping V_o as constant.

If V_i decreases, I will decrease causing I_z to decrease. The diode takes a smaller current and voltage drop across R is reduced. As a result, the output V_o remains constant. Thus, whenever V_i changes, I and IR drop in such a way as to keep V_o constant.

Case 2: We assume that V_i is constant and the load resistance R_L is varying.

If R_L decreases then I_L will increase. But as I is constant, the zener current will decrease, thereby keeping I and IR drop constant. The output voltage remains constant.

If R_L increases then I_L will decrease. With decrease in I_L , the zener current I_z will increase in order to keep I and IR drop constant. Again, the output voltage remains constant.

31.9 VARACTOR DIODE

The varactor is a semiconductor diode with the properties of a voltage-dependent capacitor. Specifically, it is a variable-capacitance, *p-n* junction diode that makes good use of the voltage dependency of the depletion-area capacitance of the diode.

The depletion region of a junction diode has a large concentration of positive charges stored on the *n*-side of the junction and a large concentration of negative charges on the *p*-side of the junction, separated by a distance ' w ', the width of the depletion region. The situation is similar to a charged capacitor where one of the parallel plates is positively charged and the other is negatively charged. Therefore, the depletion region in the junction diode leads to junction capacitance. This capacitance is called *junction capacitance* or *transition capacitance*, C_T . It is given by

$$C_T = \frac{\epsilon A}{w} \quad (31.28)$$

where A is the area of cross-section of the junction,

ϵ is the permittivity of semiconductor and

w is the width of the depletion region.

The junction capacitance of a normal diode is about 20 pf with no external bias. Junction capacitance depends on reverse bias and doping concentration.

- Reverse bias causes an increase in the depletion region width and hence the junction capacitance decreases (Fig. 31.23 a). The change in the width of the depletion region in case of an abrupt junction is given by

$$w \propto \sqrt{V_R}$$

$$\therefore C_T \propto \frac{1}{\sqrt{V_R}} \quad (31.29)$$

Thus, junction capacitance C_T decreases with an increase in reverse bias.

Since the depletion region acts as a capacitor, the diode will perform as a variable capacitor that changes with the applied bias voltage. Thus, a varactor diode is a junction diode in which the junction capacitance of the diode is tailored to vary with the reverse bias voltage. The capacitance of a typical varactor can vary from 2 to 50 picofarads for a bias variation of just 2 volts.

- If doping level is increased, the width of the depletion layer reduces. Therefore, C_T increases with increase in doping level.

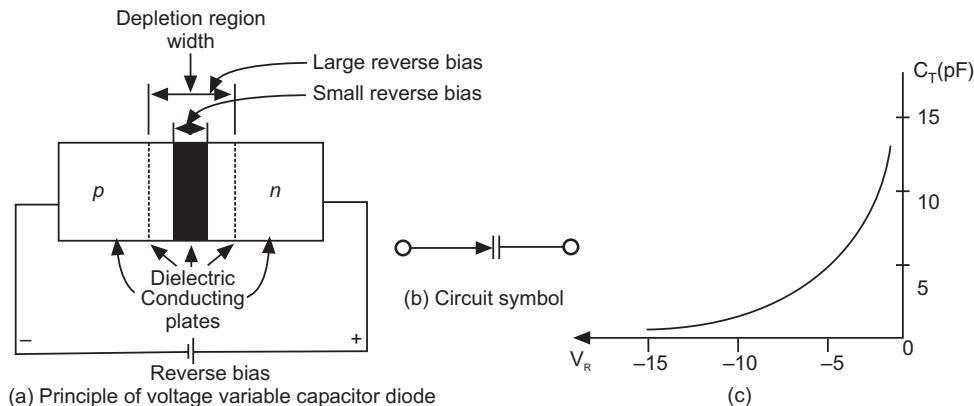


Fig. 31.23

Varactor diodes are also known as **varicaps**. The circuit symbol of the varactor diode is shown in Fig. 31.23 (b). Varactor diode shows normal volt-ampere characteristics as the junction diode in both forward and reverse bias conditions. Fig. 31.23 (c) shows the variation of transition capacitance with applied reverse bias in a typical silicon diode.

The semiconductor material used for the fabrication of varactor diodes should have high carrier mobility, low dielectric constant, large band gap, low ionization potential, and high thermal conductivity etc. The material having most of the required properties is GaAs and therefore, GaAs is used for the fabrication of varactor diodes.

Applications

- TV receivers,
- F.M. receivers,
- Tuned devices,
- Communication equipments.

If a diode is used in the reverse-biased mode as a capacitor in an LC resonant circuit, the value of C and hence the resonant frequency can be varied by changing the reverse bias voltage applied to the diode; no moving parts are required.

Fig. 31.24 shows the use of a varactor diode in an LC resonant circuit. The resonant frequency is given by

$$f = \frac{1}{2\pi\sqrt{L(C + C_T)}}$$

By changing the C_T by changing the applied voltage, the resonant frequency of the resonant circuit can be changed.

Example 31.3: Calculate the junction capacitance of a germanium diode whose area is $1 \text{ mm} \times 1 \text{ mm}$ and depletion region width is $2 \mu\text{m}$. The relative permittivity of germanium is 16.

Solution:

$$C_T = \frac{\epsilon A}{w}$$

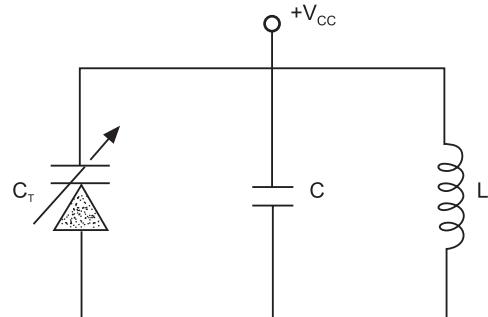


Fig. 31.24

$$\begin{aligned}
 &= \frac{8.854 \times 10^{-12} F / m \times 16 \times 10^{-3} m \times 10^{-3} m}{2 \times 10^{-6} m} \\
 &= 70.8 \text{ pF}.
 \end{aligned}$$

Example 31.4: Junction capacitance of an abrupt junction diode is 20 pF at 5 V. Compute the decrease in capacitance for 1 V increase in bias.

Solution:

$$\frac{C_{T2}}{C_{T1}} = \left(\frac{V_{R1}}{V_{R2}} \right)^{\frac{1}{2}}$$

$$\therefore C_{T2} = C_{T1} \left(\frac{V_{R1}}{V_{R2}} \right)^{\frac{1}{2}} = 20 \times 10^{-12} F \times \left(\frac{5 \text{ V}}{6 \text{ V}} \right)^{\frac{1}{2}} = 18.25 \text{ pF}.$$

Decrease in capacitance = $(20 - 18.25) \text{ pF} = 1.75 \text{ pF}$.

31.10 LIGHT EMITTING DIODE (LED)

A light emitting diode (LED) is a semiconductor diode that gives off light when it is forward biased. LEDs are generally fabricated using III-IV compound semiconductors, such as GaAs, which have a direct band gap.

Principle: When a *p-n* junction is forward biased, minority carriers flow in large numbers into regions where they can recombine with majority carriers producing light in the visible or infra red region. The wavelength of light is given by

$$\lambda = \frac{hc}{E_g} = \frac{1.24}{E_g (\text{eV})} \mu \text{ m}. \quad (31.30)$$

This effect is known as **injection electroluminescence**. A significant light output is obtained only when there is large number of electro-hole recombinations occurring per second. To ensure this, the *p* and *n* regions are heavily doped.

Theory: The energy band diagram of a heavily doped *p-n* junction is shown in Fig. 31.25a. There is a large concentration of electrons in the conduction band of *n*-region and a large concentration of holes in the valence band of *p*-region. When forward bias is applied the electrons push into the depletion region and occupy energy levels in the conduction band. Similarly, holes push forward into the depletion region and occupy energy levels in the valence band. The electrons in the conduction band are directly above the holes at the edge of the valence band (Fig. 31.25 b). The situation is highly conducive for direct recombination of electrons and holes. When an electron from the conduction band jumps into the hole in the valence band, recombination occurs and the excess energy is emitted in the form of a light photon.

The colour of the emitted light depends on the type of material used.

	Material used	Colour of the emitted light
1.	Gallium Arsenide, GaAs	Infrared
2.	Gallium Arsenide-phosphide, GaAsP	Red or Yellow
3.	Gallium Phosphide, GaP	Red or Green

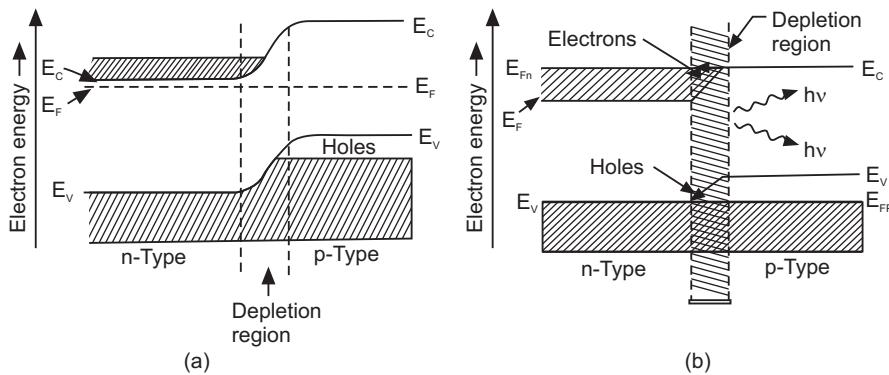


Fig. 31.25: Energy band diagram of an LED- (a) without bias (b) under forward bias

Construction

We describe here the structure of a surface emitting LEDs. These LEDs emit light in a direction perpendicular to the *p-n* junction plane. The construction of a surface emitting LED is shown in Fig. 31.26 (a). An *n*-type layer is grown on a substrate and a *p*-type layer is grown on it by the process of diffusion. The *p*-layer is made very thin to prevent loss of photons due to absorption in the layer. Metal connections are made at the edges of the *p*-layer in order to allow more central surface for the light to escape. A metal film is deposited at the bottom of the substrate for reflecting as much light as possible towards the surface of the device and also to provide electrode connection. The light generated at the junction may not emerge from the surface of the device as it is likely to suffer total internal reflection at the semiconductor-air boundary. Therefore, the device is encapsulated in a clear epoxy resin of suitable refractive index (Fig. 31.26 b).

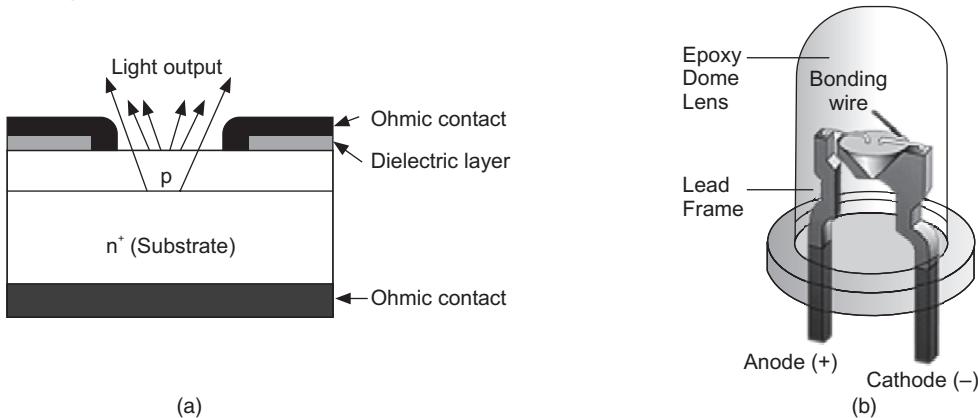


Fig. 31.26

Working

The circuit symbol of LED and a simple circuit to illustrate the working of an LED are shown in Fig. 31.27.

LED is always forward biased. The forward voltage across an LED is considerably greater than an ordinary diode. Typically the maximum forward voltage for LED is between 1.2 V and 3.2 V depending on the device. The LED emits light in response to a sufficient forward current. The amount of light emitted is directly proportional to the forward current, as

shown in Fig. 31.27 (c). The reverse breakdown voltage of LED is of the order of 3V and an LED is never reverse biased.

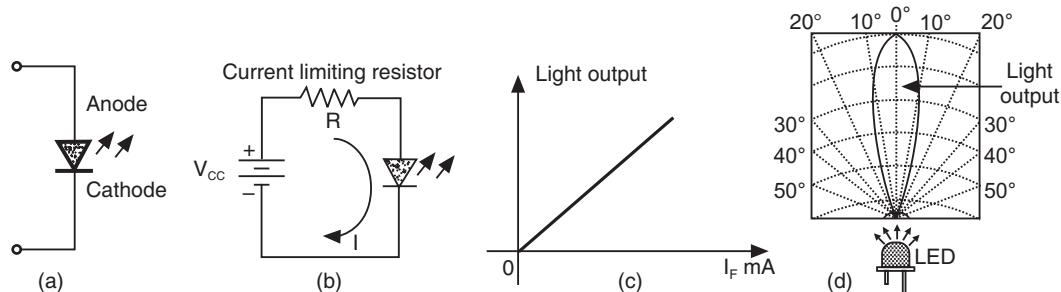


Fig. 31.27

Applications

LEDs are used in many applications. Discrete LEDs are used as indicators and as light sources in fibre-optic communications. A number of LEDs may be grouped to form a display. The LEDs may be arranged in the form of a seven-segment display where by energizing a proper combination of segments the decimal numbers 0 to 9 may be displayed. Or they may be arranged in the form of a 5×7 matrix which may be used to generate a decimal number or alphabetical character.

Example 31.5: A light emitting diode is made of GaAsP having a band gap of 1.9 eV. Determine the wavelength and colour of radiation emitted.

Solution: $\lambda = \frac{hc}{E_g} = \frac{1.24}{E_g(eV)} \mu\text{m} = \frac{1.24}{1.9} \mu\text{m} = 652.6 \text{ nm}$. \therefore Colour of the radiation is red.

31.11 PHOTODETECTORS

Photoconductive Effect

When light is incident on an intrinsic semiconductor, electrons are excited from the valence band to the conduction band. Such electrons leave behind holes in the valence band. Thus, free electrons and holes are generated in the material; but they do not leave the material. Therefore, an increase of free charge carrier concentration occurs within the semiconductor. This is known as **internal photoelectric effect**. An electron gets excited to the conduction band from the valence band by a light photon provided the photon energy, $h\nu$ is greater than the band gap energy, E_g . That is, $h\nu \geq E_g$. It means that the frequency of the photon should satisfy the following condition.

$$\nu \geq \frac{E_g}{h}$$

We can express the above condition in terms of wavelength as

$$\lambda \leq \frac{hc}{E_g} .$$

The largest wavelength that can cause the electron transition is therefore given by

$$\lambda_g = \frac{1.24}{E_g(eV)} \mu\text{m} \quad (31.31)$$

An increase in free charge carriers leads to an increase in the conductivity of the semiconductor. The light-induced increase in the electrical conductivity called **photoconductive effect** or simply **photoconductivity**. The application of an electric field to the semiconductor causes the drifting of electrons and holes through the material and as a result, an electric current flows in the circuit.

Photodetectors are devices that absorb optical energy and convert it to electrical energy. The operation of photoelectric detectors is based on the internal photoelectric effect.

There are three main types of photodetectors, namely, photodiodes, *pin* diodes and avalanche photodiodes, which are widely used in optical communication systems.

31.11.1 Photodiode

Photodiodes are essentially the same as the *p-n* junction diodes. During the fabrication of the *p-n* diode, a depletion layer forms at the junction region by immobile negatively charged acceptor atoms in the *p* type material and immobile positively charged donor ions in the *n* type material. The electric field due to these ions stops the motion of majority carriers but accelerates minority carriers across the junction. When a photon is incident on the device, an electron-hole pairs are generated. In case of electron-hole pairs generated within the depletion region, the electric field acting across the region causes the pair to separate as shown in Fig. 31.28. This charge separation can be utilized in two ways. If the diode is short-circuited externally, a current flows between *p* and *n* regions. It is known as the **photoconductive mode** of operation. The diode is reverse biased for photoconductive operation. On the other hand, if the diode is left on open-circuit, an externally measurable voltage appears between *p* and *n* regions. This is known as **photovoltaic mode** of operation. This mode of operation is used in solar cells.

A **semiconductor photodiode** is a reverse biased *p-n* junction. The structure of a photodiode is shown in Fig. 31.29 (a). When a reverse bias is applied across the junction (Fig. 31.29 b), the depletion layer widens as mobile carriers are swept to their respective majority sides. The motion of minority carriers causes the *reverse leakage current* of the diode. Thus, even when no light radiation is present (zero light), a small leakage current exists. This leakage current is called **dark current**. The amount of dark current depends on the reverse bias voltage, the series resistance and the ambient temperature.

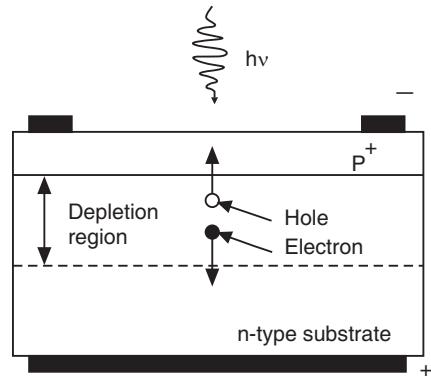


Fig. 31.28: Motion of photo-generated carriers in a *p-n* photodiode

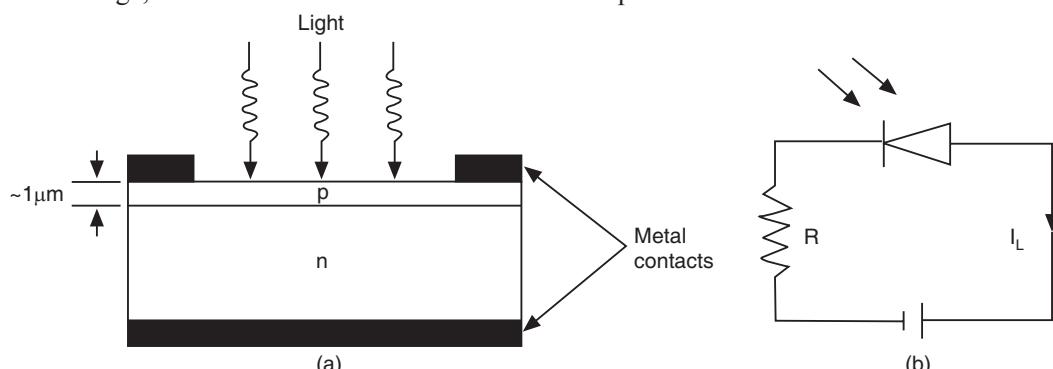


Fig. 31.29: A photo diode (a) Side view (b) A reverse biased *pn*-junction

When the diode is illuminated by light, photons are absorbed mainly in the depletion layer and also in the neutral regions. A photon of energy $h\nu \geq E_g$ incident in or near the depletion layer of the diode will excite an electron from the valence band to the conduction band. This process generates a hole in the valence band. Thus an electron-hole pair is generated by the optical photon. These are known as **photocarriers**. The electron-hole pairs generated in the depletion layer separate and drift in opposite directions under the action of the electric field. Such a transport process induces an electric current in the external circuit in excess of the already existing dark current (reverse leakage current). The photocurrent created in the external circuit is always in the reverse direction, i.e., from the *n* to the *p* region. Increasing the level of illumination increases the reverse current flowing. The light incident in the neutral region, on either side of the depletion layer, also produces electron-hole pairs. Electrons and holes generated within a diffusion length of the depletion layer will move randomly and slowly diffuse into the depletion region and are accelerated by the bias, thereby contributing to the photocurrent. Thus, optical excitation leads to an increase in the reverse-biased current. It is desirable that the depletion region be sufficiently wide so that a large fraction of incident light can be absorbed. Therefore, the diode can be used as a photodetector—using a reverse bias voltage—as the measured photocurrent is proportional to the incident light intensity.

The illumination (I-V) characteristic of a photodiode is very much similar to that of a *p-n* junction diode and is shown in Fig. 31.30. It may be noted that the I-V characteristic passes through the first, third and fourth quadrants. When the diode is used as a photodetector, it is operated in the third quadrant.

Characteristics of a Photodetector

The primary characteristics of a photodetector are its quantum efficiency, responsivity, the dark current and the bandwidth. The *quantum efficiency*, η , is the number of the electron-hole pairs generated per incident photon of energy $h\nu$ and is given by

$$\begin{aligned}\eta &= \frac{\text{number of electron-hole pairs generated}}{\text{number of incident photons}} \\ &= \frac{I_L / q}{P_o} \quad (31.32)\end{aligned}$$

The performance of a photodiode is characterized by the responsivity \mathfrak{R} , which specifies the photocurrent generated per unit optical power. This is given by

$$\mathfrak{R} = \frac{I_L}{P_o} \quad (31.33)$$

where I_L is the output photocurrent in amperes and P_o is the incident optical power in watts.

It can be shown that the responsivity \mathfrak{R} is related to quantum efficiency through the following relation.

$$\mathfrak{R} = \frac{\eta q}{h\nu} = \frac{\eta \lambda (\mu\text{m})}{1.24} \quad (31.34)$$

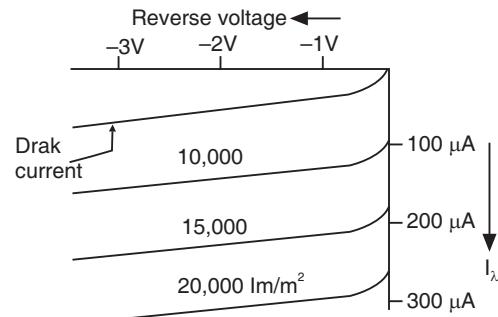


Fig. 31.30: I-V characteristic of a photodiode

It is seen that the responsivity is directly proportional to the quantum efficiency at a particular wavelength.

The maximum photocurrent in a photodiode equals

$$I_{L(max)} = \frac{q}{h\nu} P_o \quad (31.35)$$

This maximum photocurrent occurs when each incoming photon creates one electron-hole pair, which contributes to the photocurrent. The photocurrent is in fact much less than this because of various factors such as reflection at the surface of the photodiode and absorption of photons within the material and recombination of electron-hole pairs.

Disadvantages of a p-n Photodiode:

- The depletion region is a relatively small portion of the total volume of the diode. Hence, only those few photons, which are absorbed in the depletion region, cause current in the circuit whereas many of the photons absorbed in the bulk of the diode do not result in current.
- To increase the width of the depletion region, the reverse bias applied across the junction has to be increased; which may not be possible beyond a certain value.
- The electrons and holes generated by the photons in the bulk of the diode recombine before they cause current in the circuit.

31.11.2 The p-i-n Photodiode

The structure of a *p-i-n* photodiode is shown in Fig. 31.31. It is a device that consists of *p* and *n* regions separated by a very lightly doped intrinsic region (*i*). The first and most important feature of *p-i-n* photodiode is that its depletion region extends well into the intrinsic region, as it is lightly doped. Under sufficiently large reverse bias, the depletion region could extend through the intrinsic region, whereby the entire intrinsic region could be made free of charge carriers. The intrinsic layer in effect widens the depletion region and therefore increases area available for capturing light. Thus, the *p-i-n* structure of the photodiode enables us to increase the width of the depletion region to a value far greater than what it could be in a simple *pn*-junction.

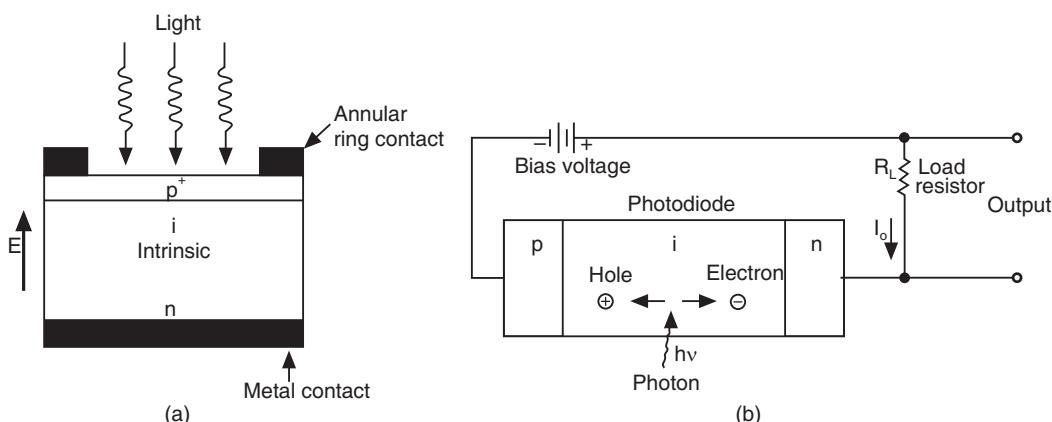


Fig. 31.31: *p-i-n* photodiode (a) a side view of the structure (b) circuit diagram

The *p-i-n* diode has a wide intrinsic semiconductor layer between *p*- and *n*-regions. The intrinsic layer has no free charges, so its resistance is high and most of the bias voltage drops across it. When an incident photon has energy greater than or equal to the band gap energy of

the semiconductor material, the photon can give up its energy and excite an electron from the valence band to the conduction band. This process generates free electron-hole pairs. These carriers are mainly generated in the depletion (depleted intrinsic) region where most of the incident light is absorbed. The high electric field present in the depletion region causes the free carriers to separate and be collected across the reverse biased junction. This gives rise to a current flow in the external circuit. As the intrinsic layer is wide enough, most of the photons are absorbed and larger photocurrent is produced. Therefore, $p-i-n$ photodiode is more sensitive than pn -photodiode.

Advantages of p-i-n-photodiode

- The depletion region is very wide and extends throughout the intrinsic region and hence the reverse bias need not be varied to widen the depletion region.
- The reverse bias applied is small, of the order of 5V.
- As the depletion area is wider, most of the incident photons are absorbed in this region and hence the efficiency of this device is high.
- The dark current in this device is smaller.

31.11.3 Avalanche Photodiode

An avalanche photodiode (APD) is more sophisticated than a $p-i-n$ diode and incorporates internal gain mechanism so that the photoelectric current is amplified within the detector. It will be very much useful when very low levels of light are to be detected. The structure of a typical APD is shown in Fig. 31.32. This configuration is known as $p^+ \pi p^- n^+$ reach-through structure.

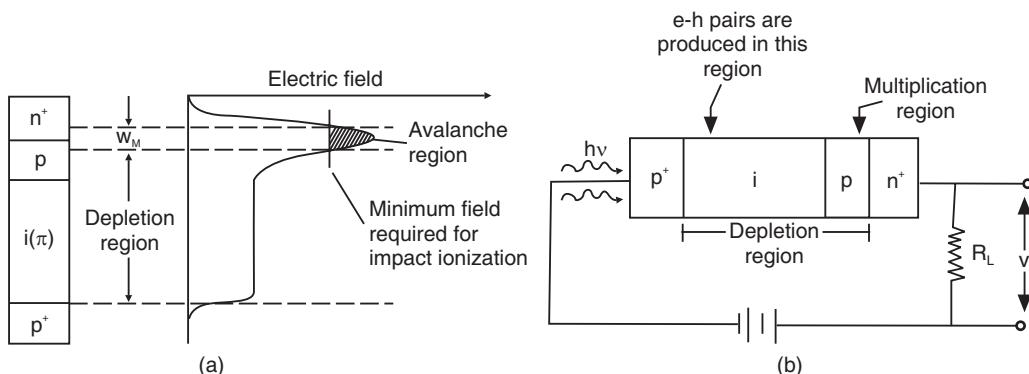


Fig. 31.32: APD photodiode (a) side view of the structure (b) circuit diagram

The device is essentially a reverse-biased $p-n$ junction. The n^+ and p^+ are heavily doped semiconductors and have very low resistance. The π region is very lightly doped and hence is nearly intrinsic. Most of the incident light passes into the intrinsic region through the thin p^+ region and electron-hole pairs are generated in the intrinsic region. Under reverse bias most of the applied voltage drops across the $p-n^+$ junction. With increase in the reverse bias voltage, the depletion region across this junction widens. Under sufficient reverse bias, the depletion region widens enough to reach through to the intrinsic layer. In this condition, the internal field intensity near the junction becomes very high and the junction approaches the breakdown condition. Therefore, the electrons and holes photogenerated in the depletion layer acquire sufficient energy from the field to liberate secondary electrons and holes within the layer by a process of *impact ionization*. The newly generated carriers are also accelerated by the high electric field, thus gaining enough energy to cause further impact ionization.

The number of carriers multiplies in geometrical progression and this phenomenon is called **avalanche effect**. In this device multiplication is initiated by electrons. Holes generated in the π region drift to p^+ electrode and hence do not take part in the multiplication process.

A photon that enters through the p^+ region is absorbed in the intrinsic region and the resulting electron-hole pair is separated by the electric field in the π region. The hole drifts towards the p^+ and do not take part in the multiplication process. The electron drifts through the π region to the pn^+ junction. There, the electric field due to high reverse bias accelerates the electron. The electron acquires enough kinetic energy to ionize neutral atoms in its path. The electrons thus produced get in turn accelerated and ionize atoms lying in their paths. The effect is cumulative and builds up into an avalanche. As a result, one electron-hole pair will on an average produce M electron-hole pairs in the process, where M is the multiplication factor. Thus there occurs a carrier multiplication and internal amplification. This internal amplification process enhances the responsivity of the detector.

Advantages of APD

- Internal current gain due to carrier multiplication
- High frequency response

31.12 SOLAR CELL

A solar cell is basically a $p-n$ junction that can generate electrical power, when illuminated. Solar cells are usually large area devices typically illuminated with sunlight and are intended to convert the solar energy into electrical energy.

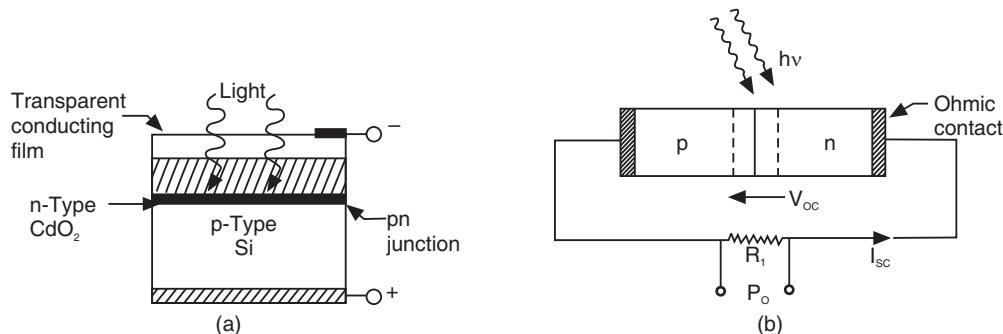


Fig. 31.33

The schematic of a solar cell is shown in Fig. 31.33 (a). It consists of a p -type chip on which a thin layer of n -type material is grown. When the solar radiation is incident on the cell, electron-hole pairs are generated in the n and p regions. The majority of them cannot recombine in the regions. They reach the depletion region at the junction where an electric field due to the space charge separates them. Electrons in the p -region are drawn into the n -region and holes in the n -region are drawn into the p -region. It results in accumulation of charge on the two sides of the junction and produces a potential difference called **photo emf**. Its magnitude is of the order of 0.5 V. The overall power-conversion efficiency of single-crystalline solar cells ranges from 10 to 30 % yielding 10 to 30 mW/cm². If a load is connected across the cell a current flows through it. The sign convention of the current and voltage is shown in Fig. 31.33 (b). It considers a current coming out of the cell to be positive as it leads to electrical power generation. The power generated depends on the solar cell itself and the load connected to it.

The $I-V$ characteristic of a solar cell is shown in Fig. 31.34. We identify the open-circuit voltage, V_{oc} , as the voltage across the illuminated cell at zero current. The short-circuit current, I_{sc} , is the current through the illuminated cell if the voltage across the cell is zero. The short-circuit current is close to the photocurrent while the open-circuit voltage is close to the turn-on voltage of the diode as measured on a current scale similar to that of the photocurrent. The power equals the product of the diode voltage and current and at first increases linearly with the diode voltage but then rapidly goes to zero around the turn-on voltage of the diode. The maximum power is obtained at a voltage labeled as V_m with I_m being the current at that voltage.

Solar cells can be connected in parallel or series into solar panels, which can deliver power output of several kilowatts. Solar panels are used in numerous applications in remote locations and in space. Solar cells of all kinds are used in different consumer products – from watches and calculators to power supplies for laptop computers.

31.12.1 Expressions for V_m and I_m

The expression for the total current in a solar cell is given by

$$I = I_o (e^{eV/kT} - 1) - I_L \quad (31.36)$$

We obtain the expression for short circuit current if we put $V = 0$ in the above equation. Thus,

$$I_{sc} = -I_L \quad (31.37)$$

The open circuit voltage is obtained if we put $I = 0$ in the equation (31.36). Then

$$I_o (e^{eV/kT} - 1) = I_L$$

$$\text{or } e^{eV/kT} = 1 + \frac{I_L}{I_o}$$

Taking logarithms on both the sides and designating $V = V_{oc}$ and $kT/e = V_T$, we get

$$V_{oc} = V_T \ln \left(1 + \frac{I_L}{I_o} \right) \approx V_T \ln \left(\frac{I_L}{I_o} \right) \quad (31.38)$$

However, no power is delivered when the solar cell operates at either of these points. It is necessary that the operating point should be selected such that the output power becomes maximum. The output power is given by

$$P = VI = V [I_o (e^{eV/kT} - 1) - I_L] \quad (31.39)$$

The condition for maximum power is given by $\frac{dP}{dV} = 0$. Therefore, by differentiating the above equation and equating it to zero, we get the expressions for V_m and I_m at the point of maximum power. Thus,

$$\frac{dP}{dV} = I_o \left[e^{eV/kT} \left(1 + \frac{eV}{kT} \right) - 1 \right] - I_L$$

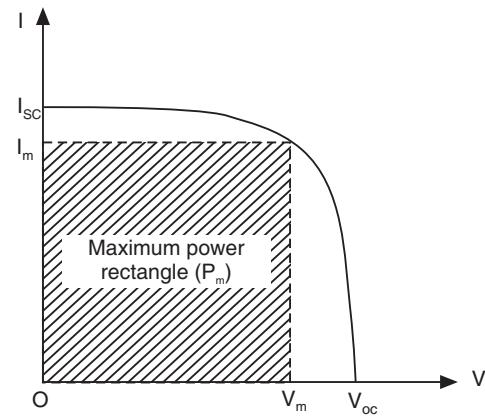


Fig. 31.34

$$\text{As } \frac{dP}{dV} = 0, \quad I_o \left[e^{eV/kT} \left(1 + \frac{eV}{kT} \right) - 1 \right] - I_L = 0 \quad (31.40)$$

When $\frac{dP}{dV} = 0$, $V = V_m$. Therefore, the above equation may be written as

$$e^{eV_m/kT} \left(1 + \frac{eV_m}{kT} \right) = 1 + \frac{I_L}{I_o}$$

$$\text{or } e^{eV_m/kT} = \frac{(1 + I_L/I_o)}{(1 + eV_m/kT)}$$

Taking logarithm on both sides of the above equation, we obtain

$$\begin{aligned} \frac{eV_m}{kT} &= \ln \left[\frac{(1 + I_L/I_o)}{(1 + eV_m/kT)} \right] \\ \text{or } V_m &= V_T \ln \frac{1 + I_L/I_o}{1 + eV_m/kT} = V_{oc} - V_T \ln \left(1 + \frac{V_m}{V_T} \right) \end{aligned} \quad (31.41)$$

where $kT/e = V_T$. Now, using the expression (31.41) into the equation (31.36) we obtain an expression for I_m .

$$\begin{aligned} I_m &= I_o \left[\exp \left\{ \frac{e}{kT} \cdot \frac{kT}{e} \ln \left[\frac{(1 + I_L/I_o)}{(1 + eV_m/kT)} \right] \right\} \right] - (I_o + I_L) \\ &= I_o \left[\frac{(1 + I_L/I_o)}{(1 + eV_m/kT)} \right] - (I_o + I_L) \\ &= \left(\frac{I_o + I_L}{1 + eV_m/kT} \right) - (I_o + I_L) \\ &= (I_o + I_L) \left[\frac{1}{1 + V_m/V_T} - 1 \right] \end{aligned}$$

$$\text{or } I_m = -(I_o + I_L) \left[\frac{V_m}{V_m + V_T} \right] \quad (31.42)$$

The maximum power is given by

$$P_m = V_m I_m \cong I_o \left[V_{oc} - V_T \ln \left(1 + \frac{V_m}{V_T} \right) - V_T \right] \quad (31.43)$$

Conversion efficiency of solar cell is defined as

$$\begin{aligned} \eta &= \frac{\text{Electrical Power delivered}}{\text{Solar Power incident}} = \frac{V_m I_m}{P_i} \\ &= \frac{I_L \left[V_{oc} - V_T \ln \left(1 + \frac{V_m}{V_T} \right) - V_T \right]}{P_i} \end{aligned}$$

$$= \frac{FF \cdot I_L V_{oc}}{P_i} \quad (31.44)$$

where FF is known as the **fill factor**.

31.13 LIGHT SOURCES FOR FIBER OPTIC SYSTEMS

Light sources are a key element in any fiber optic system. A light source converts the electrical signal into a corresponding light signal that can be injected into the fiber. It has to satisfy the following requirements in order that it can be used in transmitters.

1. Its physical dimensions must be compatible with the size of the fiber optic cable being used. This means it must emit light in a cone with cross sectional diameter 8-100 microns; otherwise it cannot be coupled into the fiber optic cable.
2. The optical source must be able to generate enough optical power so that transmission over longer distances is possible inspite of intrinsic losses in the fibre.
3. There should be high efficiency in coupling the light generated by the optical source into the fiber optic cable.
4. The optical source must be capable of easily modulated with an electrical signal.
5. Other requirements are small size, low weight, low cost and high reliability.

There are two types of light sources that satisfy the above requirements. They are

- light emitting diode (LED) and
- laser diode (LD).

LEDs are simpler and generate incoherent, lower power, light. It can be modulated at the needed speeds. LDs are more complex and generate coherent, higher power light. Both the LED and LD generate an optical beam with such dimensions that it can be coupled into a fiber optic cable. However, the LD produces an output beam with much less spatial width than an LED. This gives it greater coupling efficiency.

31.13.1 Light Emitting Diodes (LEDs)

The basic LED types used for fiber optic communication systems are the surface-emitting LED (SLED) and the edge-emitting LED (ELED).

(a) Surface-Emitting LEDs

The surface-emitting LED (shown in Fig. 31.35) is also known as the Burrus LED in honour of C. A. Burrus, its developer. In SLEDs, the size of the primary active region is limited to a small circular area of 20 μm to 50 μm in diameter. The active region is the portion of the LED where photons are emitted. The primary active region is below the surface of the semiconductor substrate perpendicular to the axis of the fiber.

A **well** is etched into the substrate to allow direct coupling of the emitted light to the optical fiber. The etched well allows the optical fiber to come into close contact with the emitting surface. In addition, the epoxy resin that binds the optical fiber to the SLED reduces the refractive index mismatch, increasing coupling efficiency.

(b) Edge-Emitting LEDs

Fig. 31.36 shows a typical ELED structure. It shows the different layers of semiconductor material used in the ELED. The primary active region of the ELED is a narrow stripe, which lies below the surface of the semiconductor substrate. The semiconductor substrate is cut or polished so that the stripe runs between the front and back of the device. The polished or cut surfaces at each end of the stripe are called facets. In an ELED the rear facet is highly reflective and the front facet is antireflection-coated. The rear facet reflects the light propagating toward

the rear end-face back toward the front facet. By coating the front facet with antireflection material, the front facet reduces optical feedback and allows light emission. ELEDs emit light only through the front facet. ELEDs emit light in a narrow emission angle allowing for better source-to-fiber coupling. They couple more power into small NA fibers than SLEDs. ELEDs can couple enough power into single mode fibers for some applications. ELEDs emit power over a narrower spectral range than SLEDs. However, ELEDs typically are more sensitive to temperature fluctuations than SLEDs.

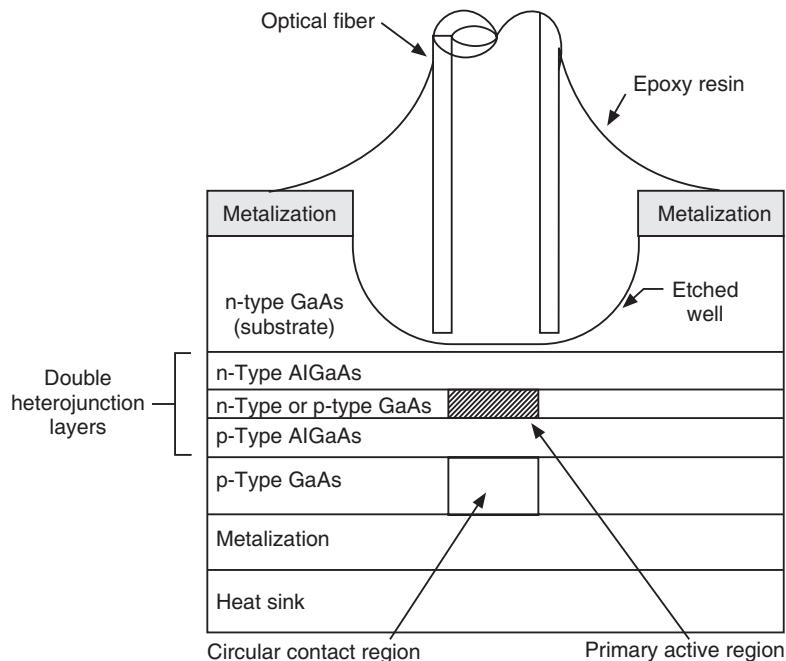


Fig. 31.35

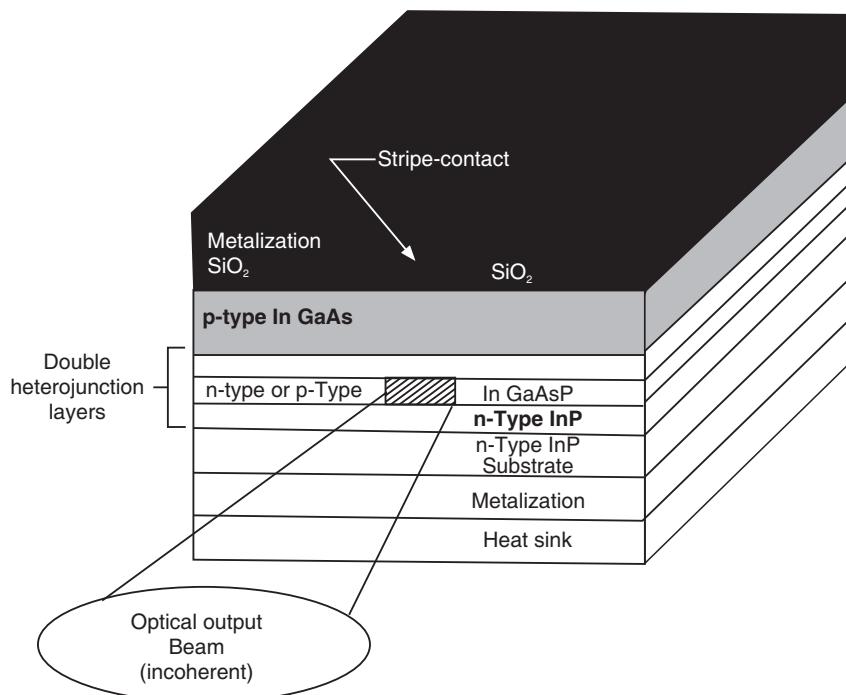


Fig. 31.36

Typically LEDs for the 850-nm region are fabricated using GaAs and AlGaAs. LEDs for the 1300-nm and 1550-nm regions are fabricated using InGaAsP and InP.

31.13.2 Laser Diodes (LD)

Ordinary lasers, having Fabry-Perot structure are multimode lasers and emit light at a number of discrete wavelengths. However, in high-speed, long-distance communications, single-mode lasers that emit light at a single frequency are required. There are two basic types of single mode laser diode structures: Distributed feedback (DFB) lasers and Vertical cavity surface-emitting lasers (VCSELs).

(a) Distributed feedback (DFB) lasers

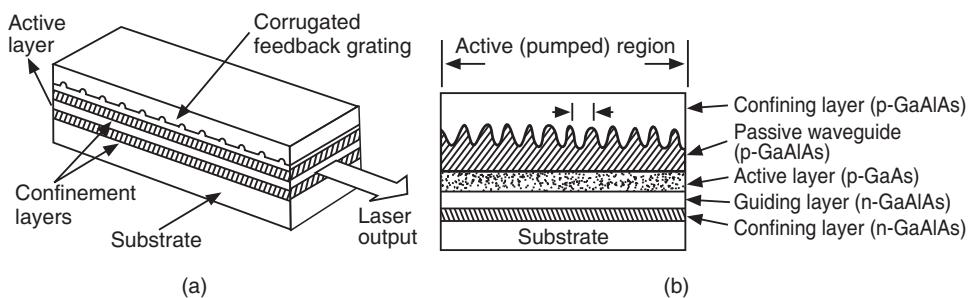


Fig. 31.37

The structure of a DFB laser is shown in Fig. 31.37. It has a corrugated layer etched internally just above the active region. The corrugation functions as an optical grating. The wave propagates parallel to the grating. At each discontinuity in the side of the wave guide, some reflection occurs. If the period of the corrugations is equal to $\frac{1}{2}$ of the internal wavelength, then the reflections reinforce. A laser having one side of the active region corrugated in this way can oscillate at a frequency defined by the corrugations. There is no need for end mirrors in this laser.

The operating wavelength is given by Braggs' law

$$\lambda_o = \frac{2n_e\Lambda}{m} \quad (31.45)$$

where Λ is the period of the corrugations, m is the order of reflection, and n_e is the effective refractive index of the mode.

(b) Vertical Cavity Surface-Emitting Lasers (VCSELs)

VCSEL emits laser light vertically from its surface and has vertical laser cavity. There are many designs of VCSEL structure. Fig. 31.38 illustrates one of the structures of VCSEL. The heart of the VCSEL is an active region, which emits light. The cavity length is very short, typically 1-3 wavelengths of the emitted light. As a result, in a single pass of the cavity, a

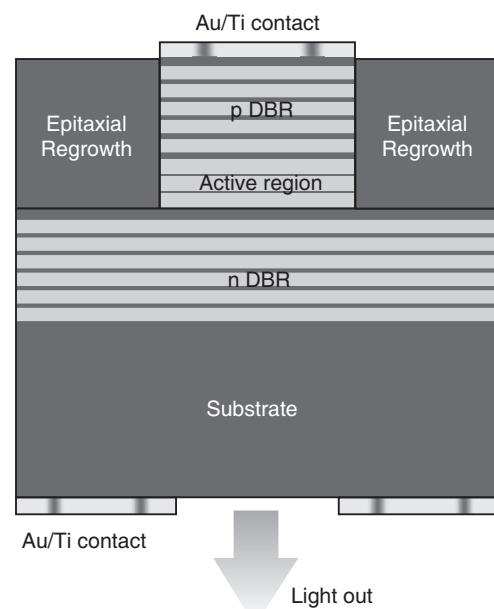


Fig. 31.38: A typical VCSEL

photon has a small chance of triggering a stimulated emission event at low carrier densities. Therefore, VCSELs require highly reflective mirrors to be efficient. Such a high reflectivity cannot be achieved by the use of metallic mirrors. In VCSELs, the reflectors above and below the active region are made from a stack of dielectric materials. The refractive index of the layers alternate between high and low values, resulting in high reflection. They are known as Distributed Bragg Reflectors (DBRs). Bragg-reflectors with as many as 120 mirror layers form the laser reflectors. Each mirror reflects a narrow range of wavelengths back into the cavity causing light emission at a single wavelength.

QUESTIONS

1. What causes majority carriers to flow at the moment when p -region and n -region are brought together? Why does this flow not continue until all the carriers have recombined?
2. Explain the formation of potential barrier across the p - n junction region. **(R.T.M.N.U., 2006)**
3. Explain the formation of depletion region in a p - n junction diode.
4. In a p - n junction, what are the diffusion and drift currents? Hence explain the existence of depletion layer.
5. Draw neat energy band diagrams for symmetrically doped p - n junction when it is
(i) Unbiased, (ii) Forward biased and (iii) Reverse biased.
6. Draw neat energy bands diagram for p - n junction when unbiased, forward biased and reverse biased. Obtain the condition of equilibrium of current when unbiased. How does biasing change this equilibrium and hence explain forward and reverse bias characteristics.
7. Discuss (draw) the energy diagram for
(i) forward biased p - n junction and
(ii) reverse biased p - n junction.
8. What is p - n -junction diode ? Explain the characteristics of p - n -junction under reverse and forward bias. **(RGPV, 2010)**
9. Derive the rectifier equation for p - n junction diode. **(RTMNU, 2010)**
10. Differentiate between drift and diffusion currents.
11. Briefly discuss application of p - n junction as half wave rectifier.
12. Draw the energy band diagram of p - n junction diode in equilibrium. Hence, obtain an expression for height of potential energy barrier. **(R.T.M.N.U., 2006)**
13. What is reverse saturation current in a diode?
14. Why is reverse saturation current independent of reverse bias?
15. What is reverse breakdown?
16. What is meant by photoconductivity?
17. What is meant by photovoltaic emf?
18. What is a zener diode? Explain the operation of a zener diode in the forward and reverse bias condition.
19. Explain how a zener diode maintains constant voltage across the load.
20. Discuss the avalanche breakdown and zener breakdown. **(Calicut Univ., 2005, 2006)**
21. Discuss the avalanche and breakdown voltage of zener diode and also discuss the zener diode as a voltage regulator. Write a short note on solar cells. **(Calicut Univ., 2007)**
22. Give a brief note on the principle, construction and working of LED. What are its advantages and disadvantages? **(Calicut Univ., 2005)**
23. What is an LED? Explain the construction and working of LED.

24. Compare LED and LCD. (Calicut Univ., 2006)
25. What is an LDR? Explain its construction and working.
26. Draw the I-V characteristic of a simple photodiode. What is meant by dark current?
27. Explain the principle of a photoconductive cell.
28. Explain construction and discuss I-V characteristics of the following semiconductor devices:
(a) Photodiode
(b) Solar cell (RGPV, 2008)
29. Sketch typical illumination characteristics for a photodiode and explain the theory of the device.
30. Explain the phenomenon of photovoltaic effect.
31. Explain the principle of photovoltaic cell.
32. Explain the working of a solar cell.
33. Draw and explain the V-I characteristics of a solar cell.
34. What are the parameters of solar cell?
35. Write short notes on:
(i) solar cell
(ii) *p-n* junction diode. (C.S.V.T.U., 2007)
36. Write short notes on:
(i) Solar cell
(ii) Phototransistor
(iii) Photoresistor (Calicut Univ., 2006)

PROBLEMS

1. In a pn-junction germanium diode the reverse saturation current is 10^{-5} amp. at 27°C . What will be the forward current in this diode for voltage 0.2 volts across it ? (RGPV, 2010)

CHAPTER

32

Bipolar Junction Transistor

32.1 INTRODUCTION

The transistor was invented in 1947 by the American physicists John Bardeen, Walter Brattain, and William Shockley at Bell Telephone Laboratories. The transistor heralded the modern era of solid-state electronics. The electronic circuits grew smaller and smaller, became lighter and inexpensive. Profound changes are brought in all fields of human endeavour. Exploration of outer space and deep sea, advancements in atomic power and arsenal, communications and computers, entertainment and medicine, automatization of production processes – all these and other fields are enriched. The inventors of transistor were honoured with Nobel Prize in physics in 1956.

The term transistor is mainly associated with the bipolar junction transistor (BJT). In BJT the action of both holes and electrons is important and therefore, it is called bipolar junction transistor. There are other types of transistors such as FET, MOSFET, UJT which operate on the action of only one type of charge carrier. We study about BJT in this chapter.

32.2 TRANSISTOR STRUCTURE

A **transistor** is a semiconductor device consisting of three regions separated by two distinct *p-n* junctions. The central region is called the **base**. It may be a *p*-type or *n*-type semiconductor. The two outer regions are called **emitter** and **collector** (Fig. 32.1). They are of the same type extrinsic semiconductor but different from that of base. Thus, if the base is *p*-type the emitter and collector are *n*-type and if the base is *n*-type the emitter and collector are *p*-type.

Thus, two types of transistors are available. They are called *npn* and *pnp* transistors. The ***npn* transistor** is constructed using *n* type material as the emitter and collector while the base is made of *p* type material. The ***pnp* transistor** is constructed using *p* type material as the emitter and collector while the base is made of *n* type.

The *n*-region contains free electrons which are negative charge carriers and *p*-region contains mobile holes which are positive carriers. Thus, two types of charge carriers, namely electrons and holes, are involved in current flow through an *npn* or *pnp* transistor. Therefore, these transistors are known as bipolar junction transistors.

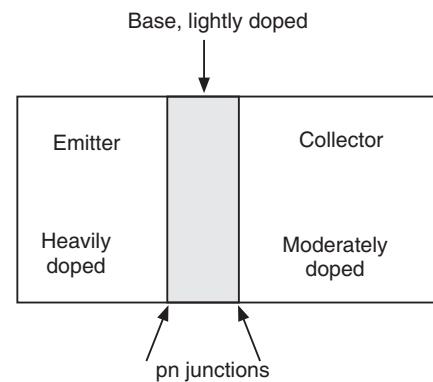


Fig. 32.1

The function of each element is as follows:

- (i) The emitter provides the majority carriers necessary to support current flow.
- (ii) The base controls the flow of the majority carriers within all elements of the transistor.
- (iii) The collector supports the majority of the current flow in the transistor. In most cases the current that flows through the collector accomplishes the work done by a transistor.

32.3 SCHEMATIC REPRESENTATION

The schematic symbols of *npn* and *pnp* transistors are shown in Fig. 32.2

In the symbols, the emitter is always indicated by an arrowhead. This arrowhead serves to tell us three things:

- (i) the location of the emitter;
- (ii) the type of the transistor that is being represented; and
- (iii) the direction in which the conventional current flows.

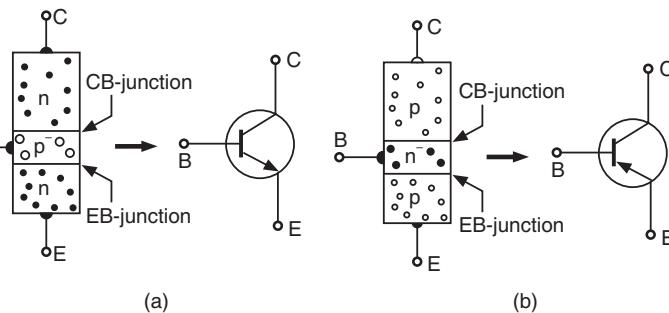


Fig.32.2: Circuit symbol of transistor (a) *n-p-n* transistor (n) *p-n-p* transistor

In the *npn* transistor, the arrow points away from the base. In this device electrons flow from the emitter into the base and hence the current flows from the base to the emitter. In the *pnp* transistor, the arrow points toward the base. The holes flow from the emitter into the base and the current flows from the emitter into the base.

A simple way of remembering the direction of arrow is as follows. In an *npn* transistor, *n*-regions are outside *p*-region and the arrow points outward and in a *pnp*-transistor, *n*-region is in between *p*-regions and the arrow points inward. Thus,

npn *n* outside, arrow outward

pnp *n* inside, arrow inward

32.4 FORMATION OF DEPLETION REGIONS

Each transistor has two *p-n* junctions. The junction that separates the base and the emitter is called the **emitter-base (EB) junction** and the one separating the base and the collector is called the **collector-base (CB) junction**. Each transistor is actually one piece of crystalline material that has been doped to create the three elements.

During the process of formation of junctions, diffusion of majority carriers takes place and depletion layers form (Fig. 32.3).

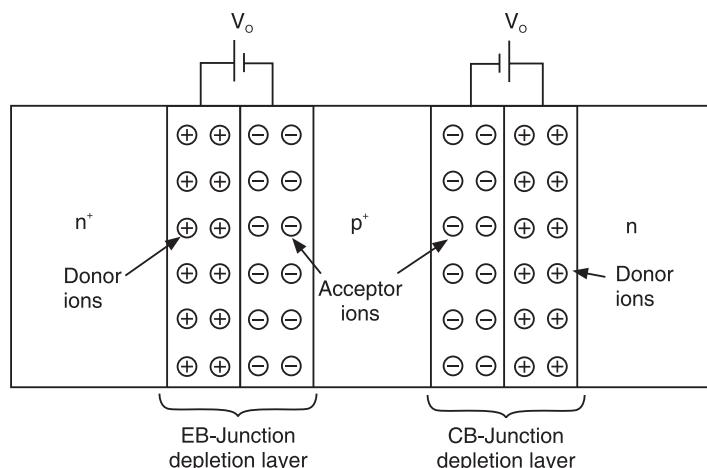


Fig.32.3

As the doping levels in the three regions are different, the two depletion layers form with different widths. Because the emitter is heavily doped and the base is lightly doped, the depletion layer at EB-junction penetrates only slightly into the emitter region and deeply into the base region. Similarly, at the CB-junction, the depletion layer extends deep into the base region while it penetrates to a lesser extent into the collector region. This results in a narrow depletion layer at EB-junction and a wide depletion layer at CB-junction. The base region becomes thinner compared to its actual physical dimension, as the two depletion layers encroach on it. The built-in barrier voltages across the two depletion layers are the same and will be of the order of 0.7 V in case of silicon transistor.

The two $p-n$ junctions can be viewed as two diodes. Therefore, a transistor may be regarded as two $p-n$ junction diodes arranged back-to-back with the base being common to both the diodes. As both the diodes have the base in common, they influence each other strongly.

32.5 ENERGY BAND DIAGRAM OF UNBIASED TRANSISTOR

(a) $n-p-n$ transistor

The energy band diagram of an unbiased $n-p-n$ transistor is shown in Fig. 32.4. At the instant of formation of the junctions, diffusion of majority carriers occurs and the energy bands in each region get shifted. The energy bands in n -region move down while those in p -region are shifted upward. The displacement of energy bands and diffusion of majority charge carriers come to a halt as soon as the Fermi levels in the three regions are equalized. The process of relative displacement of energy bands causes bending of bands in the junction regions and creation of potential barriers of height eV_o . The potential barriers inhibit the majority carrier diffusion but promote minority carrier drifting. In the equilibrium condition, the current components due to diffusion and drift of carriers balance each other. Hence, the net current through each junction is zero.

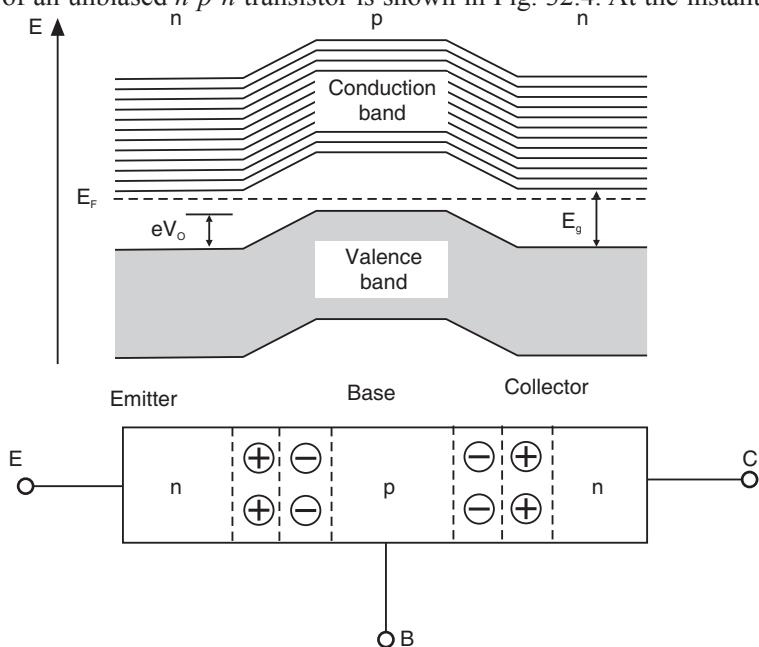


Fig.32.4: Energy band diagram of an unbiased $n-p-n$ transistor

(b) $p-n-p$ transistor

The energy band diagram of an unbiased $p-n-p$ transistor is shown in Fig. 32.5. At the instant of formation of the junctions, diffusion of majority carriers occurs and the energy bands in each region get shifted. The energy bands in n -region move down while those in p -region are shifted upward. The displacement of energy bands and diffusion of majority charge

carriers come to a halt as soon as the Fermi levels in the three regions are equalized. The process of relative displacement of energy bands causes bending of bands in the junction regions and creation of potential barriers of height eV_0 . The potential barriers inhibit the majority carrier diffusion but promote minority carrier drifting. In the equilibrium condition, the current components due to diffusion and drift of carriers balance each other. Hence, the net current through each junction is zero.

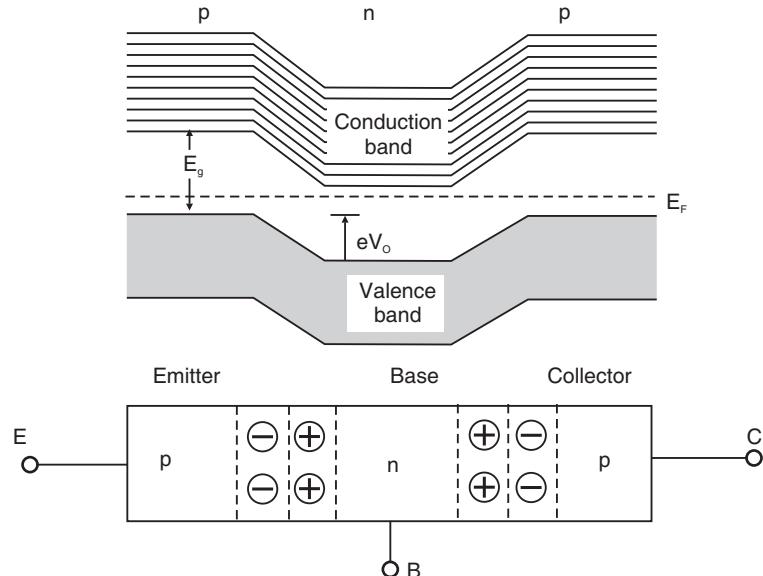


Fig.32.5: Energy band diagrams of an unbiased $p-n-p$ transistor

32.6 BIASING THE TRANSISTOR

The two junctions of a transistor can be biased in four different ways.

- (i) Both the junctions may be forward biased. It causes large currents to flow across the junctions. The currents join together in the base and flow down the common lead. Then the transistor is said to be operating in **saturation region**.
- (ii) Both the junctions may be reverse biased. Very small currents flow through the junctions. The transistor is said to be in **cut-off region**.
- (iii) EB-junction may be reverse biased and CB-junction forward biased. The transistor is said to operate in an **inverted mode**.
- (iv) EB junction may be forward biased and the CB junction reverse biased. Such biasing arrangement causes a large current to flow across the EB-junction as well as CB-junction. Further, the collector current is controlled by the emitter current or base current. With such biasing, the transistor is said to operate in **active region or in normal mode**.

We are interested in the particular biasing where the transistor operates in normal mode.

In the normal mode, when the current flows from emitter to base, it meets with a low resistance of about $15-20 \Omega$. Hence, the current is usually of the order of a few milliamperes for a voltage of a fraction of volt applied suitably between the emitter and the base. On the contrary the current encounters a large resistance of the order of $100 \text{ k}\Omega$ when it flows from the base to collector. It will be seen later that this is very advantageous for setting up amplifier circuits using transistors.

32.7 CIRCUIT CONFIGURATIONS

A transistor is a three-terminal device. There are three possible ways in which it may be connected into a circuit. They are known as circuit configurations. When the transistor is connected with its base terminal common to both the EB-junction and CB-junction, the

configuration is known as **common-base (CB) configuration**. The other configurations are known as **common emitter** and **common collector** configurations. The circuit configurations for an *npn* transistor are shown in Fig. 32.6.

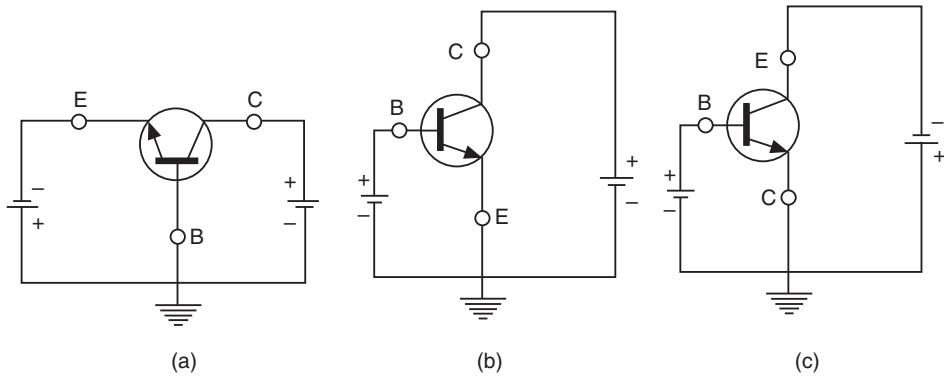


Fig.32.6: The three different ways of connecting a transistor (a) Common-base configuration.
(b) Common-emitter configuration. (c) Common-collector configuration.

32.8 ACTION OF THE BIAS

Let an *npn* transistor be connected into the circuit as shown in Fig. 32.7. The battery V_{EE} forward biases the EB-junction when the switch S_1 is closed. The source V_{CC} reverse biases CB-junction when the switch S_2 is closed.

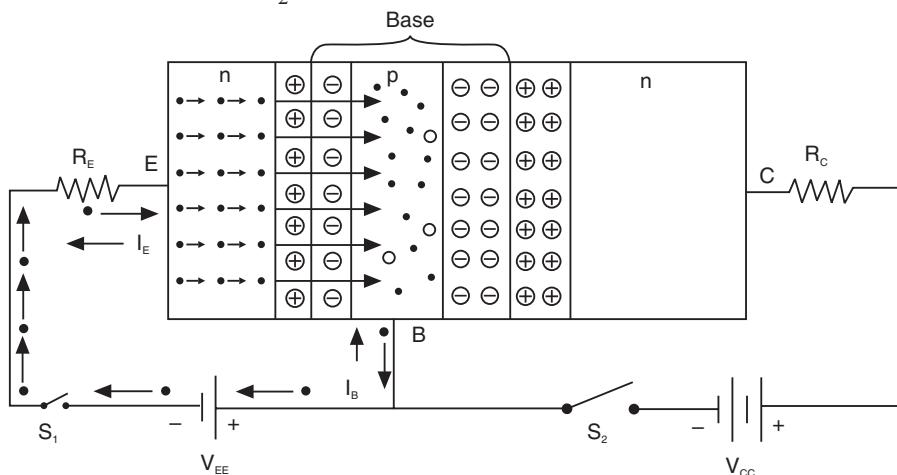


Fig.32.7

In the first step, let the switch S_1 be closed and switch S_2 be kept open. The closing of the switch S_1 forward biases the EB-junction. As a result, the potential barrier at EB-junction gets reduced and majority carriers diffuse across the junction. Electrons diffuse from emitter to base and holes from base to emitter. In a transistor the base region is intentionally doped lightly and holes are smaller in number in that region. Therefore, hole diffusion current will be very much small compared to the electron diffusion current. The total current across the junction practically consists of electron current flowing from emitter to base. It constitutes the emitter current I_E . In other words, the forward bias causes intense injection of electrons across EB-junction. As a result the electron concentration in the base region in the vicinity of EB-junction will rise to high value. The electrons entering the base become minority carriers and diffuse in the base region. They move towards the terminal end B and flow out into the

connecting wire and proceed towards the positive terminal. The emitter current I_E flowing through the base becomes the base current I_B . Hence I_B is large and is equal to I_E . As the collector is open, current does not flow into collector and the collector current I_C is zero.

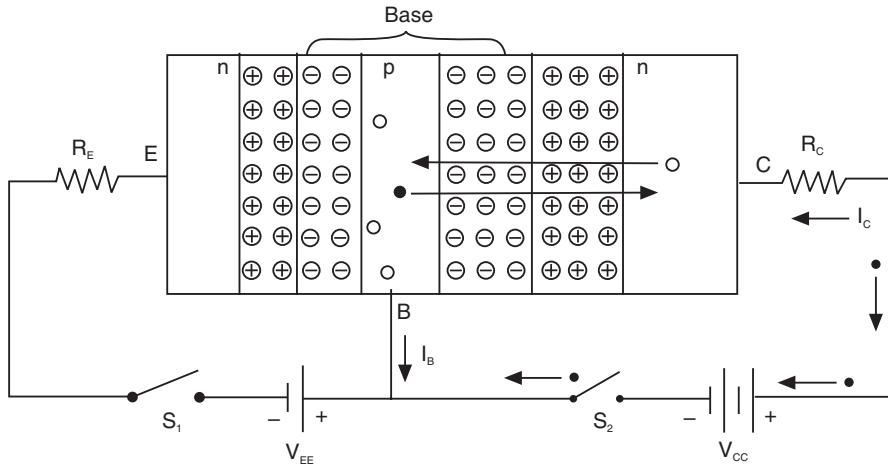


Fig. 32.8

In the next step, let us consider the situation where the switch S_1 is kept open and switch S_2 is closed (Fig. 32.8). The closing of switch S_2 reverse biases the CB-junction. The reverse bias causes drift of minority carriers. Holes from the collector region move into base region and electrons from base region flow into collector region. Because of the large difference in doping levels of base and collector regions, there are more electrons in the base region than are holes in the collector region. The electric field due to reverse bias acting across the collector depletion layer accelerates electrons and sweeps them into the collector region. Therefore, electron concentration in the base region in the vicinity of CB-junction becomes practically equal to zero. The current through the CB-junction is mainly due to electrons moving from base to collector. This current is known as the reverse saturation current. As the minority carrier concentration is small, the magnitude of the reverse saturation current is very small. This current is not at all dependent on the magnitude of the reverse bias. It is often called the **collector leakage current** I_{CBO} , where the subscript implies that it is a current through CB-junction when the emitter lead is open. Thus, in this case the collector current is I_{CBO} and as the same current flows through the base lead, the base current I_B is equal to I_{CBO} and is thus very small.

When both the switches S_1 and S_2 are closed as shown in Fig. 32.9, EB-junction is forward biased and CB-junction is reverse biased. It is expected that I_E and I_B will be large and I_C will be very small. Contrary to the expectation, I_B becomes very small and I_C will be as large as I_E which is unexpected. This unexpected result endows the transistor with a special capability and gives it an important place among devices.

32.9 TRANSISTOR ACTION

Let us now study the operation of transistor.

Referring to Fig. 32.9, we see that the emitter-base junction is forward biased. Hence, the potential barrier at the junction gets lowered and majority carriers diffuse in large number across the junction. Electron current is made much larger than the hole current by doping the base region lightly. Consequently, the emitter current is practically due to electrons flowing from emitter to base. The sum of electron and hole currents constitute the emitter current I_E . The ratio of the electron current to the total emitter current I_e/I_E is known as **emitter injection ratio**, γ . γ is typically of the order of 0.995. It means that only 0.5% of I_E consists of hole current.

Under forward bias, an intense injection of electrons into base region takes place and as a result the electron concentration in the base region nearer to EB-junction steeply rises to a value many times higher than the equilibrium value (Fig. 32.10 a). Because of reverse bias at the CB-junction, the electron concentration in the base region nearer to CB-junction is practically zero. Therefore, a large concentration gradient is established for electrons in the base region.

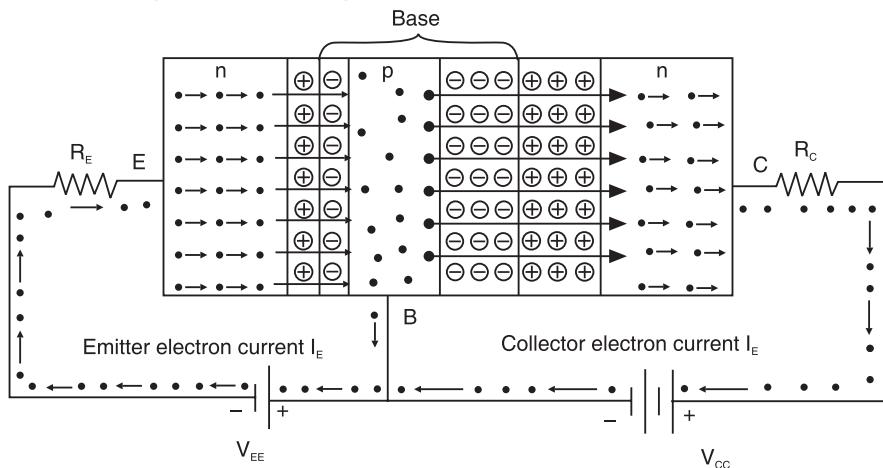


Fig.32.9

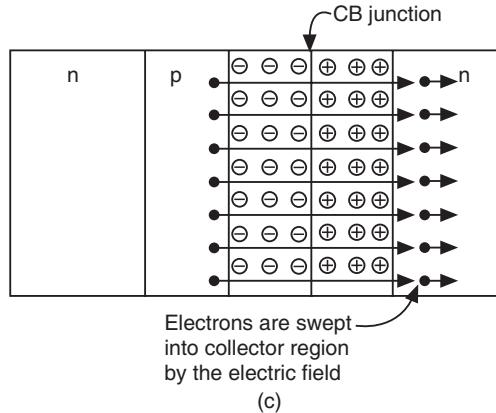
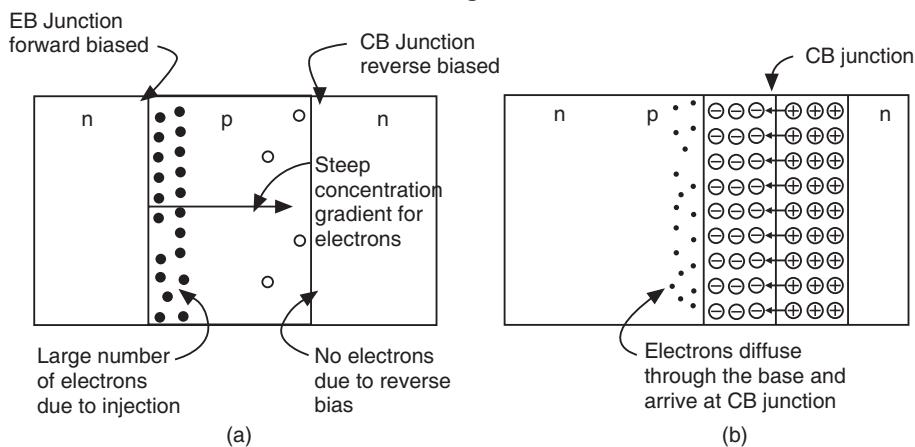


Fig.32.10

Now the electrons in the base region have two options.

1. One is that they may recombine with the holes in the base causing large base current. But due to light doping of the base region, sufficient number of holes is not available; recombination cannot take place in large way. It is necessary that recombinations should take place for the electrons to flow into the base-emitter circuit. Since recombinations are precluded in the base region, the base current I_B is very small.
2. As electron concentration is very high on the emitter side and zero on the collector side of the base region, the possibility is that electrons swiftly diffuse towards the collector-base junction under the influence of the concentration gradient across the base. The base region is narrow originally and is made further narrower due to the encroachment of depletion layers into the base and due to the action of the biases applied. Owing to this electrons quickly reach the CB-junction. Once they arrive in the vicinity of the junction they will be acted upon by the strong electric field due to reverse bias and get swept into the collector region, as shown in Fig. 32.10 (c). Consequently, a great majority of electrons emitted by the emitter flow into the collector. It causes a large reverse current I_C which is nearly equal to I_E to flow across CB-junction.

A small base current I_B is caused by the few electrons that undergo recombination in the base. The emitter-base junction is forward biased and therefore, it has a low resistance. The collector-base junction is reverse biased and has a high resistance. Almost the same current flows through the two junctions. Thus, the current is transferred from a lower resistance to higher resistance level. Hence, the device is called **trans-resistor**, which is shortened to **transistor**.

The principal particle flows in an *n-p-n* transistor biased in active mode are shown in Fig. 32.11.

The ratio of the number of electrons arriving at the collector to the number of electrons emitted by the emitter is called the **base transportation factor**, denoted by β' . Typically, it is of the order of 0.995.

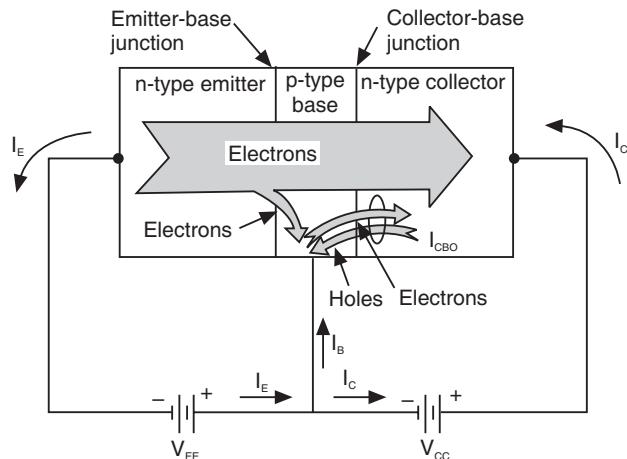


Fig.32.11: Principal particle flows when an *n-p-n* transistor is biased in the normal mode

32.10 ROLES OF Emitter, BASE AND COLLECTOR

Base

The main purpose of an electrical circuit is to drive a load. It requires that as much large power is supplied to load as possible. In the transistor circuit also, the aim is to allow almost all the current to flow through the load R_L . Any current flowing through the base will not be of much utility. The emitter is the source of current and it is required that maximum emitter current flows into the collector without getting diverted into base circuit. The base current is minimized through the following steps in an *n-p-n* transistor:

- (a) **Base region is lightly doped:** If base region is heavily doped, more holes would be present in the base and the incoming electrons would have more chance of undergoing recombination. The number of electrons flowing into collector would have thus

decreased. It leads to lesser collector current and more base current. To reduce this possibility, base region is lightly doped.

- (b) **Base region is made narrow:** It enables the electrons injected into base to quickly diffuse and come under the action of electric field due to reverse bias across junction, which sweeps them into the collector. Thus, the chance of electrons recombining with holes and causing a base current is precluded.

Emitter is Heavily Doped

In a transistor, the emitter is the source of current. It is required that a maximum of the majority carriers is injected into the base so that the emitter current I_E will be large. The function of the emitter is to provide charge carriers in large number. Hence emitter is heavily doped compared to base and collector.

Collector is Wider

Collector current is produced by minority carriers. Current by minority carriers is a drift current and requires only the presence of electric field acting in a favourable direction. Whatever may be the strength of the electric field, minority carriers are accelerated into the collector region. The minority carriers are in fact rolling down the barrier. Whether the barrier is high or low it does not matter for rolling down it. Therefore whether the reverse bias (collector voltage) is large or small it does not influence the strength of the collector current. The minority carriers rolling down the high potential barrier acquire large kinetic energy. They produce large amount of heat while transferring part of their energy to the lattice through collisions. In order to dissipate away the heat, the collector region is made larger.

32.11 RELATION BETWEEN CURRENTS IN CB CONFIGURATION

It is seen from Fig. 32.11 that in a transistor connected in CB configuration the emitter current divides itself into the base current and collector current. Thus,

$$I_E = I_B + I_C \quad (32.1)$$

Assuming that the emitter current I_E remains constant, it is seen from the relation (32.1) that the smaller the base current, the larger the collector current. In order to make the collector current as large as possible, the base current must be minimized. The collector current I_C consists of two components:

(i) a fraction of the emitter current $\alpha_{dc}I_E$ and (ii) the reverse leakage current I_{CBO} . Thus,

$$I_C = \alpha_{dc}I_E + I_{CBO} \quad (32.2)$$

where α_{dc} represents the fraction.

$$\therefore \alpha_{dc} = \frac{I_C - I_{CBO}}{I_E}$$

The reverse leakage current is negligible compared to the total collector current I_C . Therefore,

$$\alpha_{dc} = \frac{I_C}{I_E} \quad (32.3)$$

α_{dc} is called the **common base current gain factor**. As I_C is always less than I_E , α_{dc} is less than unity. α_{dc} is called **current gain** though there is no gain in current. However, as the ratio of the output current to the input current is termed gain, α_{dc} representing the ratio of I_C to I_E is also called current gain. Its value ranges from 0.95 to 0.99.

Example 32.1. In an *n-p-n* transistor circuit, the collector current is 15 mA. If 95% of the electrons emitted by the emitter reach the collector, what is the base current?

Solution. Base current is given by $I_B = I_E - I_C$.

$$\therefore I_B = (I_C/0.95) - I_C = 15 \text{ mA} (1/0.95 - 1) \\ = 0.79 \text{ mA}$$

Example 32.2. A transistor is connected in common-base configuration. If the emitter current is 2 mA and base current is 20 μA , what are the values of I_C and α ?

Solution.

$$I_C = I_E - I_B = (2 - 0.02) \text{ mA} = 1.98 \text{ mA}$$

$$\alpha = \frac{I_C}{I_E} = \frac{1.98 \text{ mA}}{2 \text{ mA}} = 0.99$$

32.12 ENERGY BAND DIAGRAM OF A TRANSISTOR BIASED IN NORMAL MODE

(a) *n-p-n* transistor

The energy band diagram for an *n-p-n* transistor biased in normal mode is shown in Fig. 32.12. The transistor is connected in common base configuration. As the EB-junction is forward biased, the energy levels in the emitter (*n*-region) and base (*p*-region) undergo relative displacement. The energy bands in the *n*-region are pushed up and those in the *p*-region are pulled down due to addition of energy. It causes a reduction in the height of the potential barrier by an amount eV_{EE} . Hence, a large number of electrons from emitter can move forward into the base region causing a current flow. Let J_{en} be the electron current density. Similarly, holes from the base region can move into the emitter region causing a hole current density, J_{hp} . As the emitter is heavily doped and larger than the base, $J_{en} \gg J_{hp}$.

The reverse bias at the collector-base junction results in an increase in the barrier height to $e(V_o + V_{CC})$ from eV_o . Majority carrier diffusion cannot occur across CB-junction as it is reverse biased. The current that flows across the CB-junction is due to minority carriers. Electrons roll down from the base region into collector region and holes from collector region float up into the base region. They produce the current density components J_{ep} and J_{hn} . The current density J_{ep} will be larger than the current density J_{hn} , as the base is lightly doped. The electrons moving into the base due to forward bias become minority carriers in the base region and are in a position to roll down into the collector region. As a result, the total current density that flows across CB junction is nearly equal to the current density flowing through EB

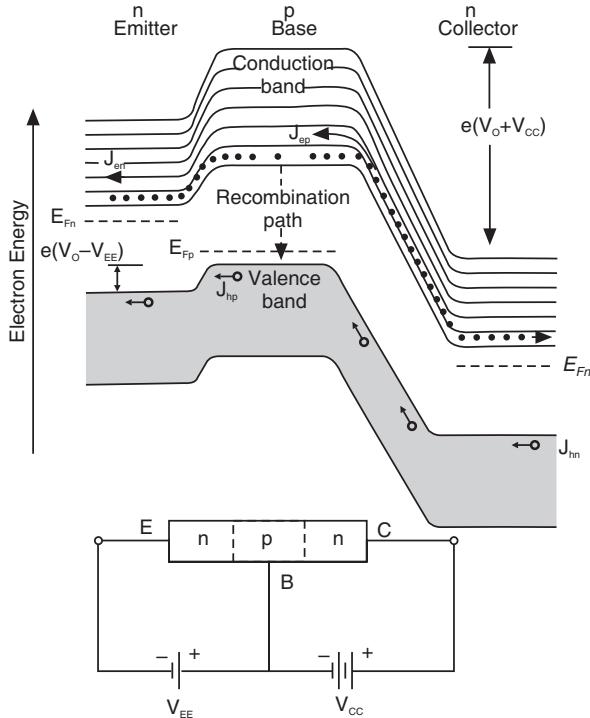


Fig. 32.12: Energy band diagram of an *n-p-n* transistor biased in normal mode

junction. It would be nearly equal to $(J_{en} + J_{ep})$. Out of the electrons injected into the base, about 2% undergo recombination by falling into the valence band from the conduction band. Such recombinations produce a small base current.

(b) *p-n-p* Transistor

The energy band diagram for a *p-n-p* transistor biased in normal mode is shown in Fig.32.13. The transistor is connected in common base configuration. As the EB-junction is forward biased, the energy levels in the emitter (*p*-region) and base (*n*-region) undergo relative displacement. The energy bands in the *p*-region are pushed down and those in the *n*-region are pulled up due to addition of energy. It causes a reduction in the height of the potential barrier by an amount eV_{EE} . Hence, a large number of holes from emitter can

move forward into the base region causing a current flow. Let J_{hp} be the hole current density. Similarly, electrons from the base region can move into the emitter region causing an electron current density, J_{en} . As the emitter is heavily doped and larger than the base, $J_{hp} \gg J_{en}$.

The reverse bias at the collector-base junction results in an increase in the barrier height to $e(V_o + V_{CC})$ from eV_o . Majority carrier diffusion cannot occur across CB-junction as it is reverse biased. The current that flows across the CB-junction is due to minority carriers. Holes float up from the base region into collector region and electrons from collector region roll down into the base region. They produce the current density components J_{ep} and J_{hn} . The current density J_{hn} will be larger than the current density J_{ep} , as the base is lightly doped. The holes moving into the base due to forward bias become minority carriers in the base region and are in a position to float up into the collector region. As a result, the total current density that flows across CB junction is nearly equal to the current density flowing through EB junction. It would be nearly equal to $(J_{hp} + J_{hn})$. Out of the holes injected into the base, about 2% undergo recombination because of electrons falling into the valence band from the conduction band. Such recombinations produce a small base current.

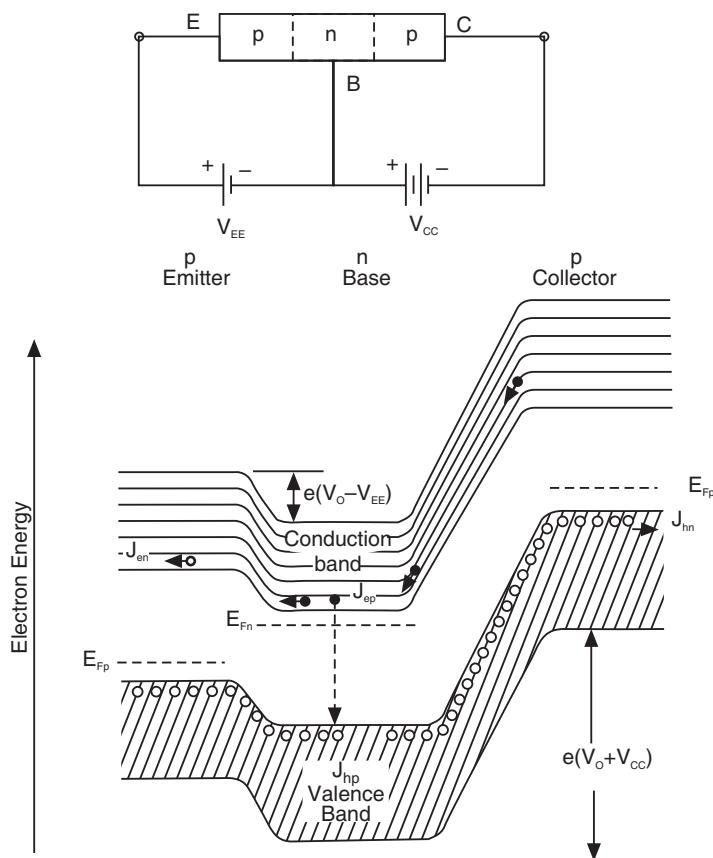


Fig.32.13: Energy band diagram of a *p-n-p* transistor biased in normal mode

32.13 COMMON EMITTER CONFIGURATION

In common emitter configuration the emitter is made common to the input and the output. An *n-p-n* transistor connected in common emitter configuration is shown in Fig. 32.14. The

transistor is biased in normal mode. The EB-junction is forward biased with the help of the battery V_{BB} . The battery V_{CC} is connected between the emitter and collector. Since the base is at $+V_{BB}$ potential with respect to the emitter and the collector is at $+V_{CC}$ potential with respect to the emitter, the net potential of the collector with respect to the base is $(V_{CC} - V_{BB})$. This potential reverse biases the CB-junction. As V_{CC} is much larger than V_{BB} , the reverse bias voltage at the CB-junction is nearly V_{CC} .

The voltage source V_{BB} forward biases the EB-junction and electrons are injected into the base from the emitter. The electrons pass through the thin weakly doped base and reach the collector. About 98% of the injected electrons are swept into the collector region and a large reverse current flows through the reverse biased CB-junction. It constitutes the current I_C . About 2% electrons recombine with holes in the base region and then onwards travel as valence electrons and flow into the base lead causing a base current I_B . A small change in forward bias of EB-junction causes a change in the value of I_B and hence of I_C .

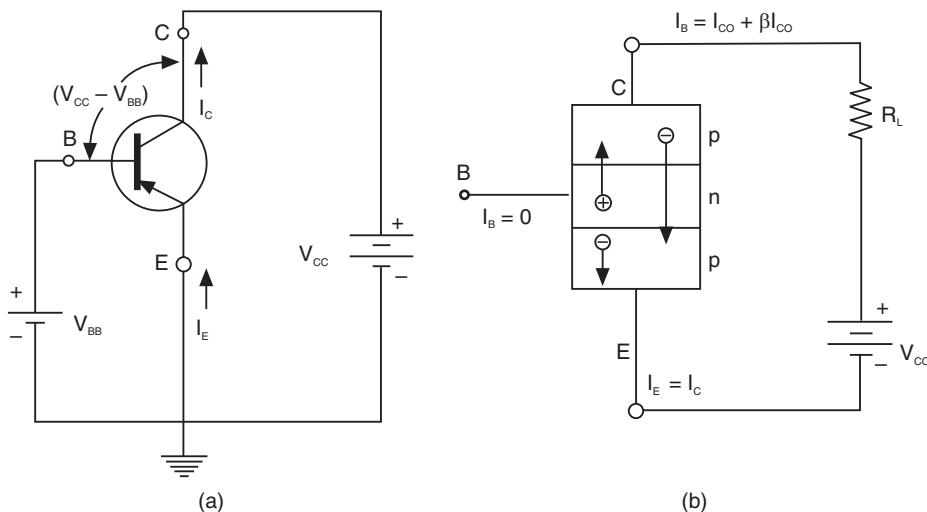


Fig.32.14

In case of common base configuration, reverse current I_{CBO} flows only through CB-junction when the emitter-base circuit is open. In contrast, current flows through both EB-junction and CB-junction in case of CE configuration (Fig. 32.14 b). The leakage current in the device is not just due to the flow of minority carriers across CB-junction. Holes from the n -type collector region have also to pass through the EB-junction in order to complete the circuit. The current due to hole flow forward biases the EB-junction, though the junction is kept open. It acts as if a base current I_{CO} is supplied by it. It causes an additional collector current βI_{CO} through the movement of electrons from n -type emitter to the base under forward bias. Thus, the current I_{CEO} through the reverse biased CB-junction will be

$$I_{CEO} = I_{CO} + \beta I_{CO}$$

or

$$I_{CEO} = (1 + \beta) I_{CO} \quad (32.4)$$

32.14 CURRENT RELATIONS IN CE CONFIGURATION

In CB configuration, I_E is the input current and I_C is the output current. They are related through the equations

$$I_E = I_C + I_B$$

and

$$I_C = \alpha_{dc} I_E + I_{CBO}$$

\therefore

$$I_C = \alpha_{dc} (I_C + I_B) + I_{CBO}$$

or

$$(1 - \alpha_{dc}) I_C = \alpha_{dc} I_B + I_{CBO}$$

or

$$I_C = \frac{\alpha_{dc}}{1 - \alpha_{dc}} I_B + \frac{1}{1 - \alpha_{dc}} I_{CBO}$$

or

$$I_C = \beta_{dc} I_B + I_{CEO} \quad (32.5)$$

where

$$\beta_{dc} = \frac{\alpha_{dc}}{1 - \alpha_{dc}} \quad (32.6)$$

and

$$I_{CEO} = \frac{I_{CBO}}{1 - \alpha_{dc}} \quad (32.7)$$

In *CE* configuration, I_B becomes the input current and I_C is the output current. As α_{dc} is less than unity, β_{dc} will be much larger than α_{dc} and I_{CEO} much larger than I_{CBO} .

The factor β_{dc} is called the **common-emitter dc current gain**. Typically, β_{dc} has values in the range from 20 to 300. From the relation (32.5) we can write

$$\beta_{dc} = \frac{I_C - I_{CEO}}{I_B}$$

$$\text{As } I_C \gg I_{CEO}, \quad \beta_{dc} = \frac{I_C}{I_B} \quad (32.8)$$

Equ. (32.8) is a linear relation. A small change in I_B causes a large variation in I_C . Thus, I_C is controlled by I_B . In view of this transistor is said to be a **current controlled device**.

32.15 TRANSISTOR AS AN AMPLIFIER

An **amplifier** is an electronic circuit that causes an increase in the voltage or power level of a given signal. Fig. 32.15 shows an *npn* transistor connected in common base configuration. The transistor is biased to operate in the active region. The battery V_{EE} forward biases the *EB* junction and the battery V_{CC} reverse biases the *CB* junction. A signal source v_i is connected in the input circuit. A load resistance R_L is connected in the output circuit. An output voltage v_o is developed across R_L .

The dc voltage V_{EE} is a fixed voltage and causes a dc current I_E to flow through *EB* junction. When the ac voltage v_i is superimposed on V_{EE} , the emitter-base voltage varies with time. For example, if V_{EE} is 10 volts and the peak voltage of the signal v_i is

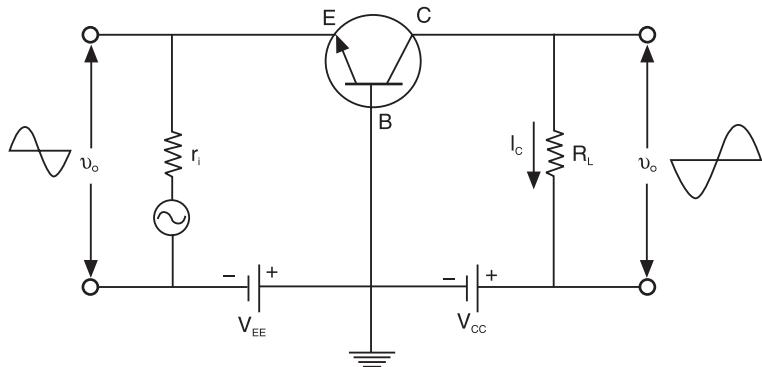


Fig. 32.15: Transistor amplifier

1 V, the emitter-base voltage swings from 9 volts to 11 volts. This variation causes corresponding variations in emitter current I_E and the collector current I_C . The varying current flows through the load resistor R_L and develops a varying voltage across it. It is the output voltage v_o .

Since the *EB* junction is forward biased, it has a low resistance r_i of the order of 100 ohms. The emitter current variation ΔI_E due to emitter-base voltage variation can be expressed as

$$\Delta I_E = v_i / r_i \quad (32.9)$$

The collector current I_C changes by ΔI_C due to the variation ΔI_E caused in I_E .

$$\Delta I_C = \alpha_{dc} \Delta I_E \quad (32.10)$$

This current ΔI_C flows through R_L causing a voltage drop $(\Delta I_C) R_L$. Therefore,

$$v_o = (\Delta I_C) R_L \quad (32.11)$$

$$= \alpha (\Delta I_E) R_L$$

$$= \alpha (v_i / r_i) R_L \quad (32.12)$$

The voltage gain of an amplifier is defined as the ratio of output signal voltage v_o to the input signal voltage v_i . Thus,

$$\text{Gain} = \frac{\text{Output Voltage, } v_o}{\text{Input Voltage, } v_i} \quad (32.13)$$

$$= \frac{\alpha R_L}{r_i}$$

$$\approx \frac{R_L}{r_i} \quad (\because \alpha \approx 1) \quad (32.14)$$

R_L is of the order of kilo ohms and is far larger than r_i . Consequently, v_o is larger than v_i and the gain of the circuit is larger than unity. It means that the transistor amplifies a small input voltage to give a larger output voltage.

QUESTIONS

1. Why is the base thin and lightly doped in transistors?
2. What is a transistor? Explain the operation of *p-n-p* transistor. (Calicut Univ., 2007)
3. Explain why in a transistor :
 - (i) Base is thin and lightly doped.
 - (ii) Collector region has larger area of cross section.
4. Draw an energy band diagram of a transistor when
 - (i) Unbiased
 - (ii) Biased in common base mode showing various currents.
5. Give reasons:
 - (i) Base region of transistor is narrow and lightly doped.
 - (ii) The collector current is large in a transistor though its collector-base junction is reverse biased.
6. Draw a neat energy band diagram (not circuit diagram) for *n-p-n* transistor when
 - (i) unbiased and (ii) biased in common base mode, showing various currents. (R.T.M.N.U., 2010)
7. Draw the energy band diagram for *n-p-n* transistor when biased in common base mode.
8. With a neat labeled diagram, explain transistor action, when it is biased to operate in the active region. Draw energy band diagram of *n-p-n* transistor in common-base mode. (R.T.M.N.U., 2006)
9. (a) What will happen to the various currents in a transistor if its base is made thick?
 (b) Why does transistor has low input resistance and high output resistance?
 (c) Why is the collector current large in a transistor though its collector-base junction is reverse biased?
10. Explain the working of transistor as an amplifier?

CHAPTER

33

Dielectrics

33.1 INTRODUCTION

Electrical circuits use insulator materials in various forms. Capacitors use dielectrics as medium. The dielectric materials, which are also known as insulators, constitute very important group of electrical (and electronic) engineering materials. They are characterized by dielectric constant, dielectric loss, dielectric strength and high resistivity. The two properties namely, dielectric constant and dielectric loss are strongly frequency dependent. The use of a dielectric in a specific application is dictated by the frequency of the applied voltage. The dielectric loss is required to be a minimum so that the performance of the dielectric does not deteriorate with time. Ferroelectric materials exhibit very high dielectric constant and low dielectric loss. They are used in making miniature capacitors. Piezoelectric materials are dielectric materials which have very interesting properties. Mechanical deformation is produced in them in response to electrical force and electrical effects are produced in response to mechanical forces. These two piezoelectric effects are vastly exploited in all areas of technology. The dielectric loss manifests in the form of heat. The dielectric heating can be employed in cases where heating by other means is difficult or is not possible. It finds application in the fields such as food processing. Material such as glass, ceramics, polymers and paper are nonconducting materials. They prevent flow of current through them. *When the main function of nonconducting materials is to provide electrical insulation, they are called insulators.*

33.2 DIELECTRICS

Dielectric materials are non-conducting materials. There are no free charge carriers in a dielectric. When dielectric materials are placed in an electric field, they modify the electric field and they themselves undergo appreciable changes because of which they act as stores of electrical charges. *When charge storage is the main function, the materials are called dielectrics.* For a material to be a good dielectric, it must be an insulator. Hence any insulator is a dielectric.

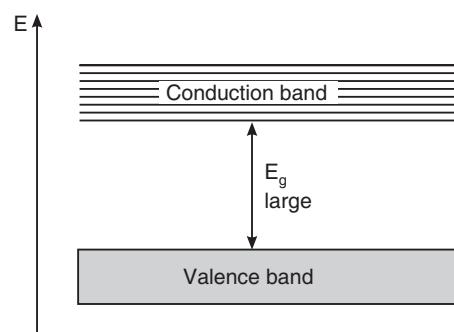


Fig. 33.1

A dielectric may be described in terms of the energy band structure. The forbidden gap E_g (Fig. 33.1) is very large in dielectrics and excitation of electrons from the normally full valence band to the empty conduction band is not possible under ordinary conditions. Therefore, conduction cannot occur in a dielectric. Even if a dielectric contains impurities, extrinsic conduction such as the one observed in doped semiconductors is not possible. The resistivity of an ideal dielectric should be therefore infinitely high. However, in practice, dielectrics conduct electric current to a negligible extent and their resistivities range from 10^{10} to 10^{20} ohm-m.

33.3 DIELECTRIC CONSTANT

A dielectric is chiefly characterized by its dielectric constant. **Dielectric constant** of a dielectric is defined and measured as *the ratio of capacitance of a capacitor containing the dielectric medium to the capacitance of the same capacitor with air as the medium*.

$$\epsilon_r = \frac{C}{C_0} \quad (33.1)$$

where C_0 is the capacitance with air as the medium between the plates and C is the capacitance with dielectric as medium.

ϵ_r is called **dielectric constant** or **relative permittivity**. It is a dimensionless quantity, which is always greater than unity in case of dielectrics, and it is independent of the size or shape of the dielectric. In fact, ϵ_r describes *the ability of the dielectric material to store electric charges*. At times another quantity known as the permittivity of the medium, ϵ , is used. It is given by

$$\epsilon = \epsilon_0 \epsilon_r \quad (33.2)$$

where $\epsilon_0 = 8.854 \times 10^{-12}$ F/m represents the permittivity of free space.

33.4 DIELECTRIC POLARIZATION

Let us consider an electrically neutral slab of an isotropic dielectric inserted between the plates of a charged parallel plate capacitor, as shown in Fig. 33.2.

Very low conductivity of the dielectric rules out the presence of free charges and their possible motion in the electric field. Hence current does not flow in the material. However, the electric field can act on the bound charges in the dielectric. These bound charges are not free to migrate through the dielectric. The action of the field E_0 on the bound charge consists in displacing the bound charges relative to one another. The negative charges (electrons) are displaced in a direction opposite to that of the electric field, while the positive charges (nuclei) are displaced in the same direction as that of the electric field. Each atom or molecule then acts as an elementary dipole and acquires an electric dipole moment in the direction of the field. The cumulative effect of formation of such dipoles is that negative charge is induced by the electric field on the dielectric surface adjacent to the positive capacitor plate while a positive charge of equal magnitude is induced on the dielectric surface adjacent to negative capacitor plate. Thus, the action of the electric field on a dielectric is to induce charges on its surfaces. When charges of opposite polarity are induced on the surfaces of a dielectric,

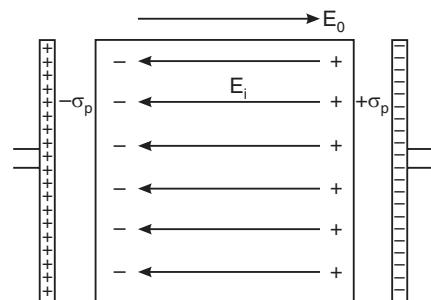


Fig. 33.2

the dielectric is said to be **polarized**. The effect is known as dielectric **polarization**. The polarized dielectric is equivalent to a big dipole consisting of polarization charges separated by a distance d , which is the thickness of the slab. *The intensity of polarization P is defined as the total dipole moment per unit volume of the material.* Thus,

$$P = \frac{\sum d\mu}{V} \quad (33.3)$$

where $d\mu$ is the dipole moment of an elemental volume and V is the total volume of the dielectric. In fact, we can consider the polarized dielectric as a big dipole consisting of induced charges separated by distance d . Thus,

$$\mu = (A\sigma_p)d = \sigma_p V \quad (33.4)$$

where A is the area of the slab and σ_p is the surface charge density due to polarization. Comparing the eqns. (33.3) and (33.4), we find that

$$P = \sigma_p \quad (33.5)$$

It follows that polarization is equal to the surface density of the induced charges in a dielectric.

The effect of polarization is to reduce the magnitude of the external field E_0 . The induced surface charges on the dielectric give rise to an induced electric field E_i which opposes the external field E_0 . Therefore, the net electric field E in the dielectric has a magnitude given by

$$E = E_0 - E_i \quad (33.6)$$

33.5 GAUSS LAW

Gauss law states that the total electric flux, φ , through a closed surface is equal to the charge enclosed by the surface. Thus,

$$\varphi = \int \mathbf{D} \cdot d\mathbf{A} = q_0$$

where \mathbf{D} is the displacement vector. It is related to the electric field through the relation

$$\mathbf{D} = \epsilon \mathbf{E}$$

Using the above relation, Gauss law can be rewritten as

$$\epsilon \int \mathbf{E} \cdot d\mathbf{A} = q_0 \quad (33.7)$$

33.5.1 Gauss Law Applied to a Dielectric

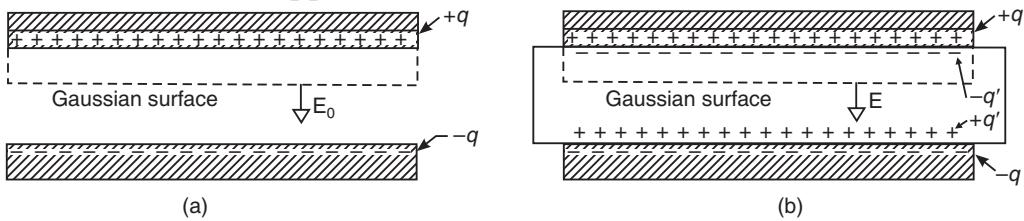


Fig. 33.3. A parallel plate capacitor (a) without and (b) with a dielectric.

Let us apply Gauss law to a parallel plate capacitor without a dielectric first and then with a dielectric. Fig. 33.3(a) shows a Gaussian surface drawn enclosing the charge q on one of the plates. q is known as the free charge on the capacitor plate. The electric field E_0 at any point on the Gaussian surface is given by

$$\int \mathbf{E}_0 \cdot d\mathbf{A} = \frac{q}{\epsilon_0}$$

or

$$E_0 A = \frac{q}{\epsilon_0}$$

$$E_0 = \frac{q}{\epsilon_0 A} \quad (33.8)$$

Now, let us consider the case of the capacitor with a dielectric and draw a Gaussian surface, as shown in Fig.33.3 (b) enclosing the free charge on the capacitor plate and induced charge on the dielectric surface. Let q' be the induced charge on the surface of the dielectric. q' is known as the bound charge. Note that q' is negative charge. Then $(q - q')$ is the net charge within the Gaussian surface. Let E be the resultant field inside the dielectric. Then, according to Gauss theorem

$$\int \mathbf{E} \cdot d\mathbf{A} = \frac{q - q'}{\epsilon_0} \quad (33.9)$$

or

$$E = \frac{q}{\epsilon_0 A} - \frac{q'}{\epsilon_0 A} \quad (33.10)$$

Equ. (33.10) indicates that the induced surface charge q' weakens the original field when the dielectric is present. The initial field and the resultant field are related through the relation

$$\begin{aligned} \mathbf{E}_0 &= \epsilon_r \mathbf{E} \\ \therefore \frac{E_0}{\epsilon_r} &= \frac{q}{\epsilon_0 A} - \frac{q'}{\epsilon_0 A} \end{aligned}$$

Using equ. (33.8), we write the above relation as

$$\begin{aligned} \frac{q}{\epsilon_0 \epsilon_r A} &= \frac{q}{\epsilon_0 A} - \frac{q'}{\epsilon_0 A} \\ \frac{q}{\epsilon_r} &= q - q' \end{aligned} \quad (33.11)$$

Using (33.11) into equ. (33.9), we get

$$\begin{aligned} \int \mathbf{E} \cdot d\mathbf{A} &= \frac{q - q'}{\epsilon_0} = \frac{q}{\epsilon_0 \epsilon_r} \\ \text{or } \epsilon_0 \epsilon_r \int \mathbf{E} \cdot d\mathbf{A} &= q \\ \text{or } \epsilon \int \mathbf{E} \cdot d\mathbf{A} &= q \end{aligned} \quad (33.12)$$

33.6 DIELECTRIC SUSCEPTIBILITY

The magnitude of polarization is directly proportional to the intensity of the electric field. Thus,

$$\mathbf{P} = \chi \epsilon_0 \mathbf{E} \quad (33.13)$$

χ (chi) is the proportionality constant and is called the **dielectric susceptibility** of the material. Dielectric susceptibility characterizes the ease with which a dielectric material can be influenced by an external electric field. It is a measure of the polarization produced in the material per unit electric field.

33.7 THE THREE FIELD VECTORS

The resultant field inside the dielectric is given by equ. (33.10) as

$$E = \frac{q}{\epsilon_0 A} - \frac{q'}{\epsilon_0 A}$$

From the Fig. 33.3 (b), it is seen that

$$\begin{aligned}\frac{q'}{A} &= \sigma_p = P \\ \therefore \frac{q}{\epsilon_0 A} &= E + \frac{P}{\epsilon_0}\end{aligned}$$

or

$$\frac{q}{A} = \epsilon_0 E + P$$

The quantity $\frac{q}{A}$ is called electric displacement, D .

Thus,

$$D = \frac{q}{A} \quad (33.14)$$

Therefore,

$$D = \epsilon_0 E + P \quad (33.15)$$

The three electric vectors, E , D and P are shown in Fig. 33.4. The expression (33.14) for D shows that this vector is related to the free charges only; they are the charges stored on the capacitor plates. The expression (33.5) for P shows that it is related to the bound charges only. The expression for the vector E shows that it is connected with both types of charges present. It may be seen from Fig. 33.4 that the lines of D begin and end on free charges; the lines of P begin and terminate on induced charge and the lines of E change from one medium to the other.

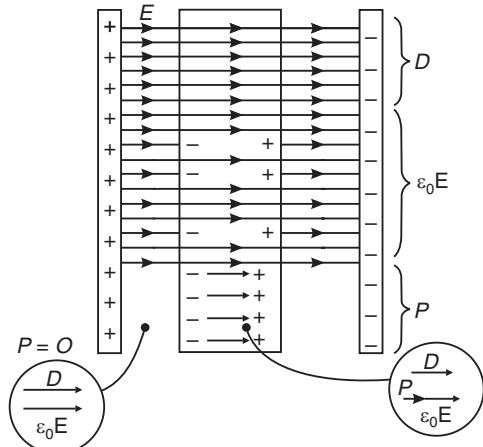


Fig. 33.4. The field vectors E , D and P . The electric field E is connected to the free charges, the vector P is related to polarization charges and the vector D to all charges.

33.8 RELATION BETWEEN ϵ_r AND χ

In order to describe the combined effects of the applied electric field E and electric polarization P , we have introduced the auxiliary vector D , called the **displacement vector**.

$$D = \epsilon_0 E + P$$

Substituting the expression for P from equ. (33.13), we get

$$D = (1 + \chi)\epsilon_0 E$$

which we can write as

$$\begin{aligned}D &= \epsilon_0 \epsilon_r E = \epsilon E \\ \epsilon_r &= 1 + \chi\end{aligned} \quad (33.16)$$

33.9 RELATION BETWEEN P AND E

From the relation (33.11), we find that

$$q' = q \left[1 - \frac{1}{\epsilon_r} \right]$$

But

$$\mathbf{P} = \sigma_p = \frac{\mathbf{q}'}{A}$$

Therefore,

$$\mathbf{P} = \frac{\mathbf{q}'}{A} = \frac{\mathbf{q}}{\epsilon_r A} (\epsilon_r - 1) = \frac{D}{\epsilon_r} (\epsilon_r - 1)t$$

or

$$\mathbf{P} = \epsilon_0 (\epsilon_r - 1) \mathbf{E} \quad (33.17)$$

Example 33.1. When NaCl crystal is subjected to an electric field of 50 V/cm, the resulting polarization is $2.215 \times 10^{-7} \text{ C/m}^2$. Calculate relative permittivity of NaCl.

Solution. $\mathbf{P} = \epsilon_0 (\epsilon_r - 1) \mathbf{E}$

$$\therefore \epsilon_r = 1 + \frac{P}{\epsilon_0 E} = 1 + \frac{2.215 \times 10^{-7} \text{ C/m}^2}{8.85 \times 10^{-12} \text{ F/m} \times 50 \text{ V/cm} \times 100 \text{ cm/m}} = 6.006$$

33.10 INDUCED DIPOLES

In order to understand the action of electric field on a dielectric, it is necessary to understand its action on an atom. In an atom the nucleus is about 10^{-15} m in diameter and it can be regarded as a point. The electron cloud is about 10^{-10} m in diameter and it may be assumed that its negative charge is concentrated at its centre. Therefore, the centres of gravity of positive and negative charges in an atom coincide (Fig. 33.5 a). Consequently, such an atom neither produces any electric field of its own nor is acted upon by an external field.

If now the atom is placed in an electric field of strength E , the electron cloud will be displaced in the direction opposite to that of E by a distance “ d ” with respect to the nucleus (Fig. 33.5 b). The centres of gravity of positive and negative charges in the atom no more coincide. The atom is now equivalent to a system of two charges of equal magnitude $q = Ze$ but opposite in sign and separated by a distance “ d ”. Such a system is called an **electric dipole** or simply a dipole. We say that a dipole is induced in the atom due to the action of the external electric field. Though a dipole is electrically neutral, the induced dipole sets up its own electric field in a direction opposite to that of the external electric field.

The product of the magnitude of the charges and the distance of their separation is called the **dipole moment**, μ of the electric dipole. Thus,

$$\mu = q d \quad (33.18)$$

The dipole moment is a vector directed along the axis of the dipole from the negative charge to the positive charge.

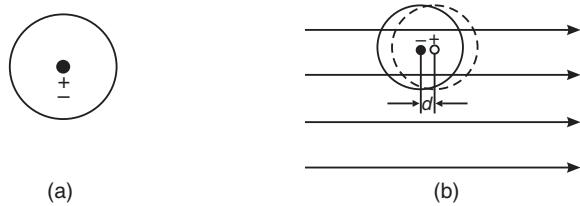


Fig. 33.5

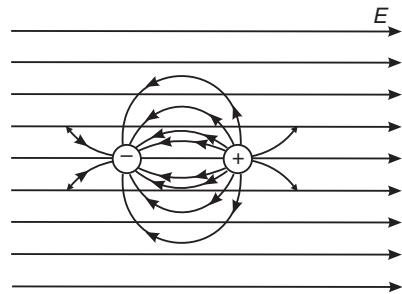


Fig. 33.6

In general any neutral system of N point charges $Q_1, Q_2, Q_3, \dots, Q_N$ occupying a volume having linear dimensions “ d ” acts as a dipole. The sum of charges $\sum Q_i$ in the volume should be equal to zero to ensure the neutrality of the system. The dipole moment of such a neutral system of point charges is given by

$$\mu = \sum_i Q_i \mathbf{r}_i \quad (33.19)$$

where \mathbf{r}_i is a vector drawn from the origin of coordinate system to the position of the charge Q_i .

Dielectric materials are made up of atoms and molecules, which are neutral systems. When a molecule is subjected to an electric field, the electric field tends to displace the equilibrium positions of bound charges, as a result of which dipole moment is induced in the molecule. The amount of **induced dipole** moment, μ , will be proportional to the field strength, E . The larger the field, the greater is the displacement of charges and hence the larger the induced dipole moment. As the charges are displaced along the field direction, the dipole moment is induced in the same direction. The molecule is then said to be **polarized** by the field. When a molecule becomes polarized, restoring forces due to coulomb attraction come into play, which tend to pull the displaced charges together. The charges separated until the restoring force balances the force due to the electric field.

Restoring forces vary in magnitude from one kind of molecule to another and therefore, the extent of dipole moment induced differs. As the amount of induced dipole moment is proportional to the field strength, we write

$$\mu_{\text{ind}} \propto E$$

or

$$\mu_{\text{ind}} = \alpha E \quad (33.20)$$

where α is the proportionality constant and is known as the **polarizability** of the molecule. *Polarizability characterizes the capacity of the electric charges in the molecule to suffer displacement in an external field.* It has the dimensions of volume. Induced dipole moment vanishes as soon as the electric field is switched off.

33.11 PERMANENT Dipoles

In some molecules, known as **polar molecules**, the centers of gravity of the charges of opposite sign are separated even in the absence of an external electric field. Such molecules are said to have **intrinsic** dipole moment and carry **permanent dipoles**.

When a molecule having intrinsic dipole moment is placed in a uniform electric field E , the field exerts a

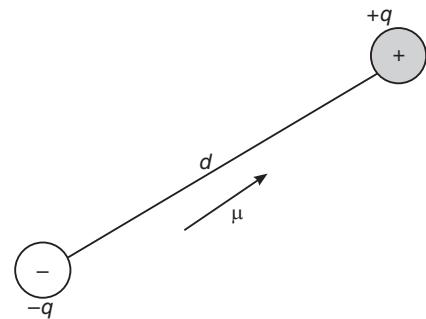


Fig. 33.7

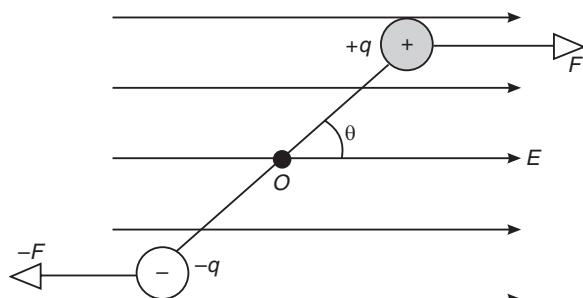


Fig. 33.8

force $+qE$ on charge $+q$ and $-qE$ on charge $-q$. The net force on the dipole is zero since the two forces acting on it are equal and opposite to each other. Therefore, there is no translational force on the dipole in a uniform electric field. However, the forces are antiparallel and constitute a couple which tends to rotate the dipole. The torque acting on the dipole is given by

$$\tau = q E d \sin \theta = \mu E \sin \theta$$

or

$$\tau = \mu \times \mathbf{E} \quad (33.21)$$

Thus, a dipole experiences a torque in a uniform electric field and rotates in an attempt to align with the field direction. In fact, a free dipole aligns its axis with the field direction.

Further, the electric field can also induce a dipole moment in the molecule. Therefore, the total dipole moment of the molecule is a sum of the induced and permanent dipole moments. Thus,

$$\mu_T = \mu_{\text{ind}} + \mu_{\text{per}} \quad (33.22)$$

However, in case of polar molecules, $\mu_{\text{ind}} \ll \mu_{\text{per}}$ and therefore,

$$\mu_T \approx \mu_{\text{per}}$$

33.12 NONPOLAR AND POLAR DIELECTRICS

Dielectrics are broadly divided into two major groups, namely, *nonpolar* and *polar* dielectrics basing on dipole moment.

A molecule is a neutral system in which the algebraic sum of all the charges is equal to zero. However, the spatial arrangement of charges in a molecule may differ from material to material. All positive charges of a molecule may be replaced by one equivalent positive charge located at the center of gravity of positive charges. Similarly, all negative charges in it may be replaced with a single equivalent negative charge located at the center of gravity of all negative charges. The two resultant charges are equal in magnitude. Their points of action in space may coincide or may not coincide. When the points of action coincide, the molecule will not possess a permanent dipole moment. Such molecules are called **nonpolar molecules** and the material is known as a **nonpolar dielectric**. Their permittivities are low and range from 1 to 2.2.

If the points of the resultant charges of a molecule do not coincide in space, the molecule possess an **intrinsic dipole moment**. Such molecules are called **polar molecules** and the materials made up polar molecules are called **polar dielectrics**. In a polar molecule consisting of several bonds, each bond may carry a permanent dipole moment. The resultant dipole moment of the molecule may then be obtained through the vector addition of the moments associated with the different bonds. The permittivities of polar dielectrics are high ranging from 3 to 8 and more.

Whether a molecule is a polar or nonpolar can be judged from its structure. It is obvious that symmetric molecules are nonpolar since the centres of gravity of positive and negative charges coincide with each other. Thus, monoatomic molecules He, Ne, Ar, and Xe are nonpolar. Molecules consisting of two identical atoms linked by a homopolar bonds such as H_2 , N_2 , Cl_2 are nonpolar.

On the other hand, asymmetric molecules are polar. The molecules of ionic compounds with a heteropolar bond, such as potassium iodide (KI) have a high dipole moment and are polar.

In estimating the intrinsic dipole moment of a molecule by its structure, it is necessary to consider the actual distribution

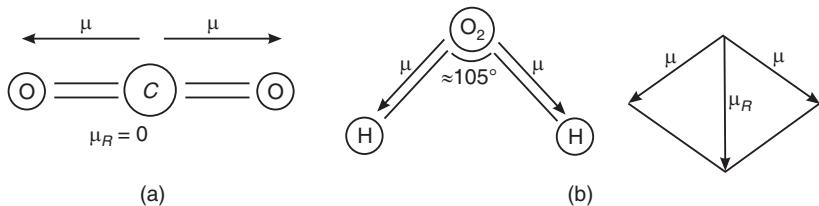


Fig. 33.9

of charges in space rather than its chemical formula taken in conventional form. For example, the chemical formula of carbon dioxide (CO_2) and water (H_2O) look identical in form. But CO_2 molecule is nonpolar whereas H_2O molecule is polar. The dipole moments of the two $\text{C}=\text{O}$ bonds in CO_2 molecule are oppositely directed and cancel each other (Fig. 33.9 a). Therefore, the resultant dipole moment of CO_2 molecule is zero. On the other hand, the water molecule has the form of an isosceles triangle with a bond angle of 104.5° . Consequently, the resultant dipole moment of water molecule comes to 6.1×10^{-30} C.m. and the molecule is polar.

All hydrocarbons are nonpolar. The intrinsic dipole moment in these molecules is either zero or very small. But hydrocarbons become polar substances when hydrogen atoms are replaced by other atoms or groups of atoms. Let us consider the example of methane. Methane is the simplest hydrocarbon and its chemical formula is CH_4 . The dipole moment of methane is zero. When the hydrogen atoms are replaced one after the other with chlorine atoms, we obtain methyl chloride CH_3Cl , methylene chloride CH_2Cl_2 , chloroform CHCl_3 and carbon tetrachloride CCl_4 their structures and dipole moments are shown below (Fig. 33.10).

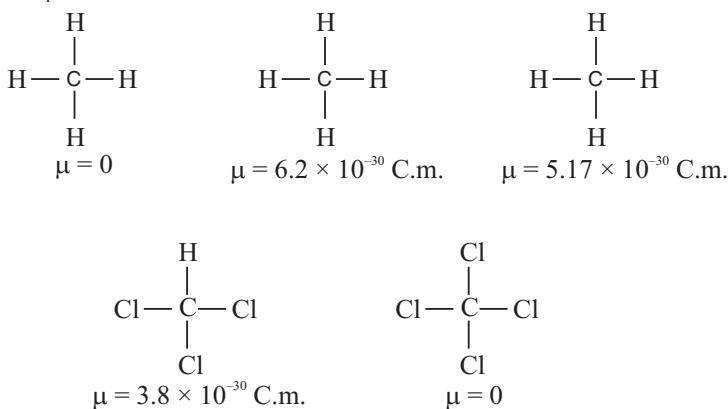


Fig. 33.10

33.13 POLARIZATION-AN ATOMIC VIEW

Let us once more consider a slab of dielectric located between the plates of a parallel plate capacitor. In the absence of an external electric field each elementary volume of the dielectric has no dipole moment. If the dielectric is a nonpolar material, the constituent molecules do not possess intrinsic dipole moments and in case the dielectric is a polar material, the individual molecular dipoles are randomly oriented so that in an elementary volume $\sum \mu = 0$. Hence, the polarization is zero. When the electric field is switched on dipoles are induced in nonpolar molecules, which form chains along the field lines, as shown in Fig. 33.11.

The polarization is given by

$$P = N \mu = N \alpha E \quad \text{Nonpolar dielectric} \quad (33.23)$$

where N is the number of molecules per unit volume.

In a polar dielectric, the molecular dipoles experience a torque that tends to align them with the field direction. Total alignment is not achieved because of the disordering effects of thermal agitation. An average alignment $\langle \mu \rangle$ is achieved in the direction of the field. The polarization is therefore given by

$$P = N \langle \mu \rangle \quad \text{Polar dielectric} \quad (33.24)$$

Thus, the action of electric field brings the dipoles into a certain ordered arrangement in space. It is seen that the ends of adjacent dipoles carrying opposite charges neutralize each other. Only the charges of the dipole ends terminating on the opposite faces of the slab remain uncompensated. Thus, the application of an electric field to a dielectric produces a displacement of charge within the material through a progressive orientation of intrinsic or induced dipoles. This is known as **dielectric polarization**.

33.14 TYPES OF POLARIZATION

Dielectric polarization is classified into four basic types.

- (i) Electronic polarization,
- (ii) Ionic polarization,
- (iii) Orientation polarization, and
- (iv) Space charge polarization.

One or two of the polarizations are always present and at a particular temperature or frequency of applied field, one or another may contribute in a large measure to the total polarization.

33.14.1 Electronic Polarization

This is the polarization that results from the displacement of the electron clouds of atoms, molecules and ions with respect to heavy 'fixed' nuclei to a distance that is less than the dimensions of the atoms, molecules or ions. It occurs in all dielectrics for any state of aggregation. The phenomenon is illustrated in Fig. 33.12. The electronic polarization sets in over a very short period of time, of the order of 10^{-14} to 10^{-15} s. It is independent of temperature.

Expression for Electronic Polarization

Let us consider a single atom with atomic number Z . The charge on its nucleus is $+Ze$ and Z electrons move around the nucleus. Let us assume that the nucleus is a point charge and the total negative charge $-Ze$ is homogeneously distributed throughout a sphere of radius, R . When this atom is subjected to an electric field E , the nucleus and the electron cloud

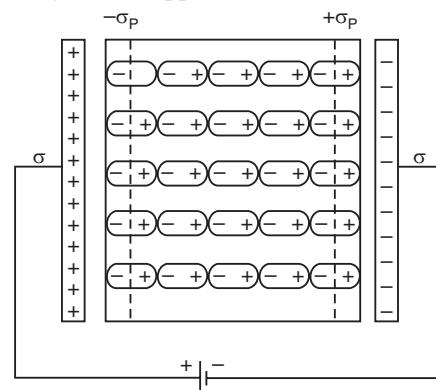


Fig. 33.11. Polarization of a dielectric according to atomic point of view. The induced charge density σ_p on the dielectric is less than the free charge density σ on the plates.

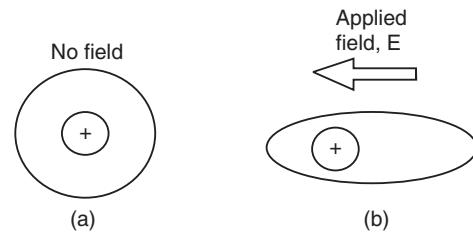


Fig. 33.12. Electronic polarization- (a) Atoms are not polarized in the absence of the electric field, (b) Electronic polarization results from the distortion of electron cloud by an applied electric field.

will move in opposite directions. The coulomb attractive force opposes the movement, which acts as the restoring force here. Equilibrium condition will be attained in which the nucleus is displaced relative to the center of the electron cloud by the amount, x . The force on the nucleus along the field direction is

$$F = ZeE \quad (33.25)$$

To determine the coulomb attraction on the nucleus, we divide the electron cloud into two regions. One region is the one that is inside the sphere of radius x and the other is the annular region lying between the two spherical surfaces of radii x and R . By applying Gauss theorem, we find that the force experienced by the nucleus arises due to the negative charge lying within the spherical region of radius x . The charge inside this region is given by $-\frac{Zex^3}{R^3}$. The force exerted by this charge on the nucleus is given by

$$F = \frac{1}{4\pi\epsilon_0} \cdot \frac{(Ze)(Zex^3 / R^3)}{x^2} \quad (33.26)$$

The equilibrium condition is that the above two forces balance each other. Thus,

$$ZeE = \frac{1}{4\pi\epsilon_0} \cdot \frac{(Ze)(Zex^3 / R^3)}{x^2}$$

Therefore, the displacement of the nucleus is

$$x = \frac{4\pi\epsilon_0 R^3}{Ze} E \quad (33.27)$$

Now the dipole moment induced in the atom due to the displacement is

$$\begin{aligned} \mu_{\text{ind}} &= (Ze)x \\ &= (Ze) \frac{4\pi\epsilon_0 R^3}{Ze} E \end{aligned}$$

or

$$\mu_{\text{ind}} = 4\pi\epsilon_0 R^3 E \quad (33.28)$$

∴

$$\mu_{\text{ind}} = \alpha_e E$$

where

$$\alpha_e = 4\pi\epsilon_0 R^3 \quad (33.29)$$

It follows that the electronic polarization in a unit volume of the dielectric is given by

$$P_e = N\alpha_e E \quad (33.30)$$

where α_e is the **electronic polarizability**. The contribution of P_e to the dielectric constant may be obtained as follows

$$\begin{aligned} \varepsilon &= 1 + \chi = 1 + \frac{P_e}{\epsilon_0 E} = 1 + \frac{N\alpha_e}{\epsilon_0 E} \\ \varepsilon_r &= 1 + \frac{N\alpha_e}{\epsilon_0} \end{aligned} \quad (33.31)$$

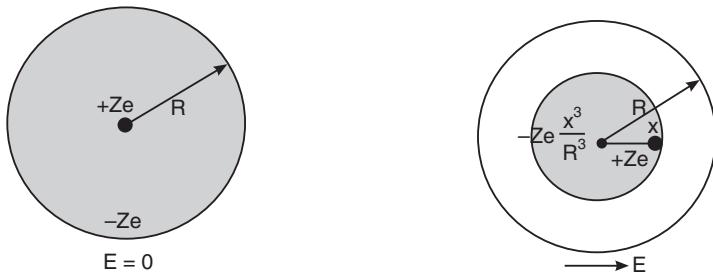


Fig. 33.13

The above expression indicates the contribution to dielectric constant due to electronic polarization alone and thus gives the dielectric constant of a nonpolar gas. Note that it depends on the polarizability of a molecule and the number of molecules in a unit volume of the dielectric. In case of monoatomic gas $\alpha_e = 4\pi\epsilon_0 R^3$.

$$\therefore \epsilon_r = 1 + 4\pi N R^3 \quad (33.32)$$

Example 33.2. Calculate the electronic polarizability of argon atom. Given $\epsilon_r = 1.0024$ at NTP and $N = 2.7 \times 10^{25}$ atoms/m³.

Solution.

$$\epsilon_r = 1 + \frac{N\alpha_e}{\epsilon_0}$$

$$\therefore \alpha_e = \frac{\epsilon_0(\epsilon_r - 1)}{N} = \frac{(8.85 \times 10^{-12} \text{ F/m})(1.0024 - 1)}{2.7 \times 10^{25} / \text{m}^3} = 7.9 \times 10^{-40} \text{ F.m}^2.$$

Example 33.3. The number of atoms in hydrogen gas is 9.8×10^{20} atoms/cc. The radius of hydrogen atom is 0.053 nm. Calculate its electronic polarizability and relative permittivity.

Solution. $\alpha_e = 4\pi\epsilon_0 R^3 = 4 \times 3.14 \times 8.85 \times 10^{-12} \text{ F.m} \times (0.053 \text{ nm})^3 = 1.657 \times 10^{-41} \text{ F.m}^2.$
 $\epsilon_r = 1 + 4\pi NR^3 = 1 + 4 \times 3.14 \times 9.8 \times 10^{26}/\text{m}^3 \times (0.053 \text{ nm})^3 = 1.0018.$

33.14.2 Ionic Polarization

Ionic polarization occurs in ionic crystals. It occurs due to the elastic displacement of positive and negative ions from their equilibrium positions. Let us take the example of sodium chloride crystal. A sodium chloride molecule consists of Na⁺ ions bound to Cl⁻ ions through ionic bond. If the interatomic distance is d , the molecule exhibits an intrinsic dipole moment equal to ' ed '. When a dc electric field is applied to the molecule, the sodium and chlorine ions are displaced in opposite directions (Fig. 33.14) until ionic bonding forces stop the process. The dipole

moment of the molecule increases consequently. When the field direction is reversed the ions move closer and again the dipole moment undergoes a change. Thus, dipoles are induced. The induced dipole moment is proportional to the applied field and is expressed as

$$\mu_i = \alpha_i E \quad (33.33)$$

where α_i is known as **ionic polarizability**.

Expression for ionic Polarization

Let us consider the sodium crystal being subjected to an external electric field E . Due to the action of electric field, the positive ions (Na⁺) displace in the direction of electric field through a distance of, say, x_1 units and the negative ions (Cl⁻) by units in a direction opposite to that of the field.

The net displacement of the ions $x = x_1 + x_2$ (33.34)

The force on the Na⁺ ion due to electric field = $+eE$

The force on the Cl⁻ ion due to electric field = $-eE$

The restoring force acting on the Na⁺ ion = $-k_1 x_1$

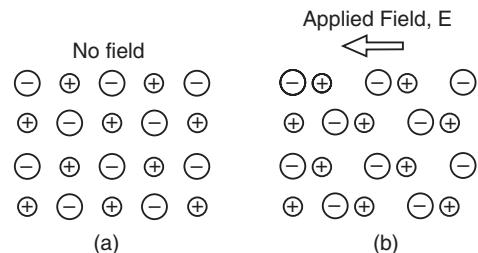


Fig. 33.14. Ionic Polarization - (a) unpolarized dielectric. (b) Ionic polarization results from the relative displacement of electrically charged ions in response to an applied electric field.

The restoring force acting on the Cl^- ion = $+k_2 x_2$
where k_1 and k_2 are the force constants. They are given by

$$k_1 = M\omega_0^2 \quad \text{and} \quad k_2 = m\omega_0^2.$$

In the above expressions, M is the mass of the positive ion (Na^+) and m the mass of the negative ion (Cl^-) and ω_0 is the natural angular frequency of the molecule.

The shifting of ions halts and equilibrium is attained when the electric force and restoring force are equal and opposite to each other. Thus, the equilibrium condition requires that

$$\begin{aligned} eE &= k_1 x_1 \quad \text{and} \quad eE = k_2 x_2 \\ x_1 &= \frac{eE}{k_1} = \frac{eE}{M\omega_0^2} \quad \text{and} \quad x_2 = \frac{eE}{k_2} = \frac{eE}{m\omega_0^2} \end{aligned}$$

Using the above expressions into equ. (33.34), we obtain the net displacement of ions as

$$\begin{aligned} x &= \frac{eE}{M\omega_0^2} + \frac{eE}{m\omega_0^2} \\ \text{or} \quad x &= \frac{eE}{\omega_0^2} \left[\frac{1}{M} + \frac{1}{m} \right] \end{aligned} \quad (33.35)$$

The induced dipole moment $\mu = e x$.

$$\mu = \frac{e^2 E}{\omega_0^2} \left[\frac{1}{M} + \frac{1}{m} \right] \quad (33.36)$$

$$\therefore \text{The electronic polarizability } \alpha_i = \frac{e^2}{\omega_0^2} \left[\frac{1}{M} + \frac{1}{m} \right] \quad (33.37)$$

It is seen from the above expression for ionic polarizability that

- It is inversely proportional to the square of the natural frequency of the molecule.
- It is directly proportional to the reduced mass of the molecule, $\left[\frac{1}{M} + \frac{1}{m} \right]$.
- It does not depend on temperature.

The ions experience electronic polarization in addition. For most materials, the ionic polarizability is less than the electronic polarizability. Typically

$$\alpha_i = \frac{1}{10} \alpha_e$$

The ionic polarization is given by

$$\begin{aligned} P_i &= N \alpha_i E \\ \text{or} \quad P_i &= \frac{Ne^2}{\omega_0^2} \left[\frac{1}{M} + \frac{1}{m} \right] E \end{aligned} \quad (33.38)$$

Ionic polarization takes 10^{-11} to 10^{-14} s to build up, and is not influenced by temperature.

33.14.3 Orientation Polarization

The orientation polarization is characteristic of polar dielectrics, which consist of molecules having permanent dipole moment. In the absence of external electric field, the orientation

of dipoles is random resulting in a complete cancellation of each other's effect, as illustrated in Fig. 33.15 (a).

When the electric field is impressed the molecular dipoles rotate about their axis of symmetry to align with the applied field. In case of electronic and ionic polarizations, the force due to the external field is balanced by a restoring force due to coulomb attraction, but for orientation polarization, restoring forces do not exist. However, the dipole alignment is counteracted by thermal agitation. The higher the temperature, the greater is the thermal agitation. The dipoles can turn only through a small angle, as illustrated in Fig. 33.15 (b). Even in case of liquids or gases, where molecules are free to rotate, a complete alignment cannot be achieved due to the randomizing effect of the temperature. However, it is estimated that it is enough if one molecular dipole in 10^5 completely aligns with the field to produce orientation polarization of the order of electronic polarization.

Thus, orientation polarization is strongly temperature dependent. This type of polarization occurs in gases, liquids and amorphous viscous substances. In case of solids, the molecules are fixed in their positions and their rotation is highly restricted by the lattice forces, leading to a great reduction in their contribution to orientation polarization. Because of this reason, while the dielectric constant of water is about 80, that of solid ice is about only 10.

As the process of orientation polarization involves rotation of molecules, it takes relatively longer time than the electronic and ionic polarizations. The build up time is of the order of 10^{-10} s or more.

Expression for Orientation Polarizability - Langevin-Debye theory

Let us consider a polar gas dielectric. If it is subjected to an electric field, E (of appropriate low frequency), the individual dipoles experience torque and tend to align themselves with the electric field. If the electric field is sufficiently strong, all the dipoles in the gas would be completely aligned into the field direction and the orientation polarization would reach the saturation value

$$P_0 = N\mu$$

where N is the number of molecular dipoles per unit volume of the gas. However, at the modest electric field strengths usually employed, the dipole alignment is not complete and is much less than the saturation value, given above. This is due to the effect of temperature, which counteracts the ordering effect of the electric field. The thermal agitation of molecules tends to randomize the orientation of dipoles. A modest orientation of dipoles is achieved in thermal equilibrium. In the equilibrium state, the molecular dipoles are distributed over all directions making angles varying from 0 to π radians with the field direction.

The potential energy of a dipole is given by

$$U = - \mu E \cos \theta \quad (33.39)$$

According to statistical mechanics, the number of dipoles, dN , having orientation between θ and $\theta + d\theta$ is proportional to

$$e^{(-U/kT)} d\Omega$$

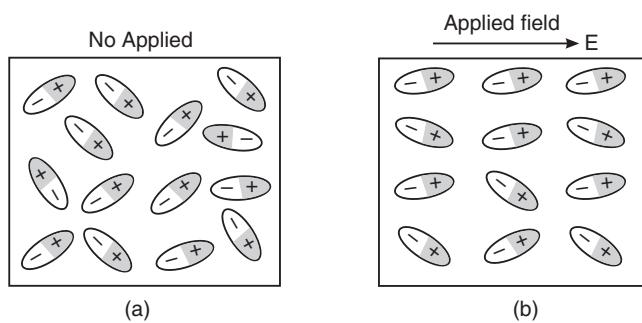


Fig. 33.15. Orientation polarization - (a) molecular dipoles are randomly oriented when $E = 0$ (b) When the field is applied the dipoles are partially aligned.

where $d\Omega$ is the solid angle. Referring to Fig. 33.16, it is given by

$$d\Omega = 2\pi \sin \theta d\theta$$

A dipole making an angle θ with the field direction contributes a component of dipole moment $\mu \cos \theta$ parallel to the field. Hence the contribution of the dN dipoles to the orientation polarization is

$$\begin{aligned} dP_o &= \mu d\Omega dN \cos \theta \\ \therefore dP_o &= 2\pi \mu e^{\mu E \cos \theta / kT} \cos \theta \sin \theta d\theta \end{aligned} \quad (33.40)$$

Therefore, the average contribution to polarization is given by

$$P_{ave} = \frac{\text{Total polarization due to all dipoles}}{\text{Total number of dipoles}} \quad (33.41)$$

$$= \frac{\mu \int_0^\pi \exp\left[\frac{\mu E \cos \theta}{kT}\right] \cos \theta \sin \theta d\theta}{\int_0^\pi \exp\left[\frac{\mu E \cos \theta}{kT}\right] \sin \theta d\theta} \quad (33.42)$$

Putting $\mu E / kT = \beta$, and $\cos \theta = y$ into the above equation, we rewrite it as follows.

$$P_{ave} = \frac{\mu \int_{-1}^1 y e^{\beta y} dy}{\int_{-1}^1 e^{\beta y} dy}$$

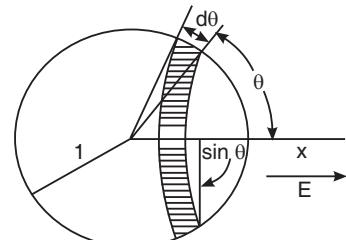


Fig. 33.16

$$\text{or } \frac{P_{ave}}{\mu} = \frac{(e^\beta + e^{-\beta})}{(e^\beta - e^{-\beta})} - \frac{1}{\beta} = \coth \beta - \frac{1}{\beta} \equiv L(\beta) \quad (33.43)$$

where $L(\beta)$ is called the Langevin function. For electric fields that are not too high and for temperatures not too low, $L(\beta)$ is given by

$$L(\beta) \approx \frac{1}{3} \beta = \frac{\mu E}{3kT} \quad (33.44)$$

$$\therefore P_{ave} = \frac{\mu^2 E}{3kT} \quad (33.45)$$

The total orientation polarization of the dielectric is

$$P_0 = NP_{ave} = \frac{N\mu^2 E}{3kT} \quad (33.46)$$

which states that the orientation polarization is directly proportional to the square of the permanent dipole moment and inversely proportional to the temperature.

$$\text{As } P_o = \frac{N\mu^2 E}{3kT} = N \alpha_0 E,$$

The orientation polarizability α_0 is given by

$$\alpha_0 = \frac{\mu^2}{3kT} \quad (33.47)$$

We have from equ. (33.17) that $\mathbf{P} = \epsilon_0 (\epsilon_r - 1)\mathbf{E}$.

Considering the contribution only from orientational polarization, we write the above expression as

$$\mathbf{P}_0 = \epsilon_0 (\epsilon_r - 1) \mathbf{E} \quad (33.48)$$

Equating the expressions (33.46) and (33.48), we obtain

$$\epsilon_0 (\epsilon_r - 1) = \frac{N\mu^2}{3kT} \quad (33.49)$$

33.14.4 Space Charge Polarization

Space charge polarization occurs in heterogeneous dielectric materials in which there is a change of electrical properties between different phases and in the homogeneous dielectrics that contain impurities, pores filled with air, inclusions of hygroscopic water etc. In particular, the properties of plastics, ceramics etc materials in their outer layers are likely to differ from those in depth because of the environmental effects and as a result they may behave as heterogeneous materials. When an electric field is applied, the electric charges those migrate within the impurity regions store up at the interfaces. The accumulation of charges takes place with opposite polarity on the interfaces, as shown in Fig. 33.17. The space charge polarization takes generally a longer time and this polarization therefore occurs at low frequencies. Space charge polarization is also known as **interfacial polarization** or **migrational polarization**.

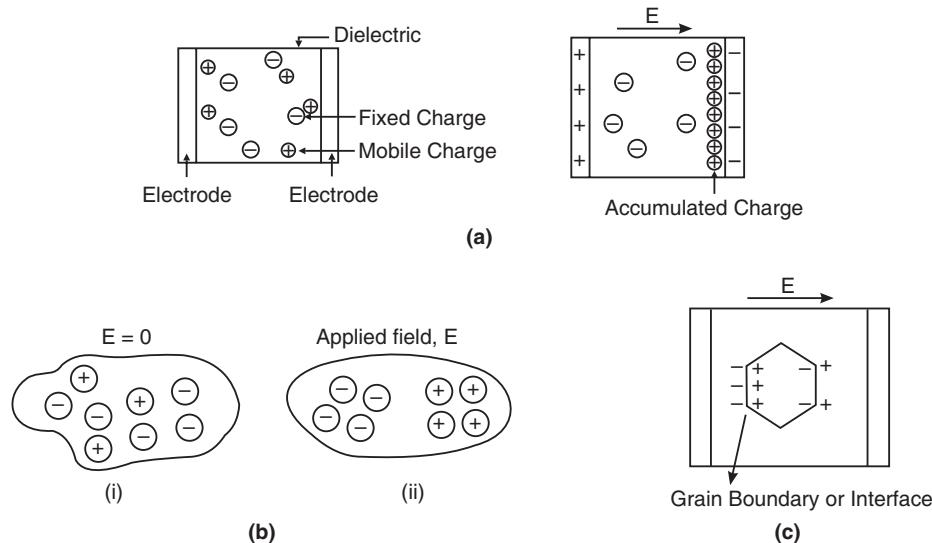


Fig. 33.17. Space charge polarization.

Interfacial polarization occurs whenever there is an accumulation of charge at an interface between two materials or between two regions within a material (Fig. 33.17 a). One of the typical examples of interfacial polarization is the grain boundaries that frequently lead to interfacial polarization as they can trap charges migrating under the influence of an applied field (Fig. 33.17 c). Interfaces also arise in heterogeneous dielectric materials for example, when there is a dispersed phase within a continuous phase (Fig. 33.17 b).

Normally, interfacial polarization exists in materials, however perfect they may be, as they contain crystal defects, impurities, and various mobile charge carriers such as electrons, holes, or ionized impurity ions. H^+ , Li^+ ions cause this type of polarization in ceramics and glasses.

While the other mechanisms are amenable to calculations, interfacial polarization defies any basic treatment. There is no general way to calculate the charges on either interfaces or their contribution to the total polarization of a dielectric.

Interfacial polarization is therefore often omitted from the discussion of dielectric properties. However, interfacial polarization is important because on the one hand many dielectrics in real capacitors rely on interface polarization while, it may harm electronic devices such as MOS transistors.

33.14.5 Total Polarization

In a material, which can experience all forms of polarization, the total polarization is equal to the sum of the electronic, ionic, orientation and migrational polarizations. The total polarization is given by

$$P_{\text{total}} = P_e + P_i + P_0 + P_m$$

In general, the migrational polarization is very small and negligible. Therefore, total polarization in a material may be taken as due to the other three contributions only. Thus,

$$P_{\text{total}} = P_e + P_i + P_0 \quad (33.50)$$

The total polarization of a polar dielectric is therefore given by

$$\begin{aligned} P &= N[\alpha_e + \alpha_i + \alpha_0]E \\ &= N \left[4\pi\epsilon_0 R^3 + \frac{e^2}{\omega_0^2} \left(\frac{1}{M} + \frac{1}{m} \right) + \frac{\mu^2}{3kT} \right] E \end{aligned} \quad (33.51)$$

The total polarizability is given by

$$\alpha = \alpha_e + \alpha_i + \alpha_0$$

$$\text{or } \alpha = 4\pi\epsilon_0 R^3 + \frac{e^2}{\omega_0^2} \left(\frac{1}{M} + \frac{1}{m} \right) + \frac{\mu^2}{3kT} \quad (33.52)$$

It is possible for one or more of the contributions to the polarization to be either absent or negligible in magnitude relative to the others. For instance, orientation polarization does not exist in non polar dielectrics. Similarly, ionic polarization will not be found in covalently bonded materials. Electronic polarization will be negligible compared to orientation polarization in polar dielectrics.

33.15 TEMPERATURE DEPENDENCE OF POLARIZATION

It is seen from equ. (33.51) that the electronic and ionic polarization do not depend on temperature and remain constant at all temperatures. However, the orientation polarization is inversely proportional to the temperature and decreases as the temperature increases. If polarization P is plotted as function of $1/T$, a straight line will be obtained, as shown in Fig. 33.18. The intercept of the line with y-axis at $1/T = 0$ gives the value of $N(\alpha_e + \alpha_i)$ from which $(\alpha_e + \alpha_i)$ can be evaluated. The dipole moment μ can be computed from the slope of the straight line, knowing the value of N , and the number of molecules per m^3 .

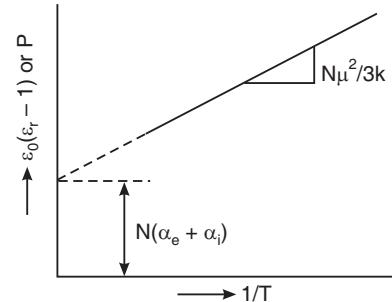


Fig. 33.18. The variation of total polarization as a function of $1/T$.

33.16 FREQUENCY DEPENDENCE OF TOTAL POLARIZATION

In many practical situations, a dielectric is subjected to an alternating electric field. An ac field changes its direction with time. With each direction reversal, the polarization components are required to follow the field reversals in order to contribute to the total polarization of the dielectric. It follows that the total polarization depends on the ability of dipoles to orient themselves in the direction of the field during each alternation of the field. The dependence of P on frequency of the electric field is sketched, in Fig. 33.19, for a polar dielectric.

In audio frequency region, all types of polarization are possible and the dielectric is characterized by a polarizability $\alpha = \alpha_e + \alpha_i + \alpha_o$ and the polarization $P = P_e + P_i + P_o$. At low frequencies, the dipoles will get sufficient time to orient themselves completely along the instantaneous direction of the field. This orientation occurs first in one direction and then in the other, following the changes in the direction of the field (Fig. 33.20). The average time taken by the dipoles to reorient in the field direction is known as the **relaxation time** τ . The reciprocal of the relaxation time is called the **relaxation frequency** v . If the frequency of the applied electric field is much higher than the relaxation frequency of the dipoles, the dipoles cannot reverse fast enough. If the dipole relaxation time τ is less than half the period of the electric field T ($\ll T/2$), the dipole can easily follow electric field alternations and contribute to orientation polarization. Consequently, the orientation polarization, which is effective at low frequencies, is damped out for higher frequencies, ($f_{\text{field}} > f_{\text{relax}}$). Usually in the radio frequency or microwave band region, the permanent dipoles fail to follow the field reversals and the polarization falls to a value corresponding to $(P_i + P_e)$. As a result, ϵ_r decreases considerably.

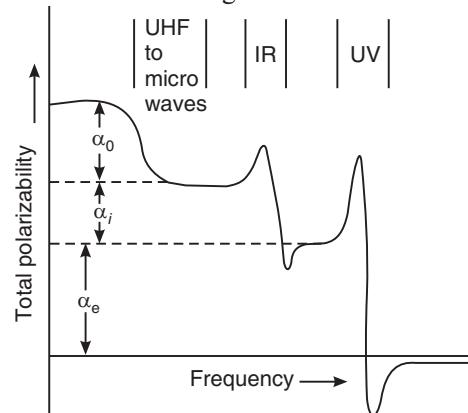


Fig. 33.19

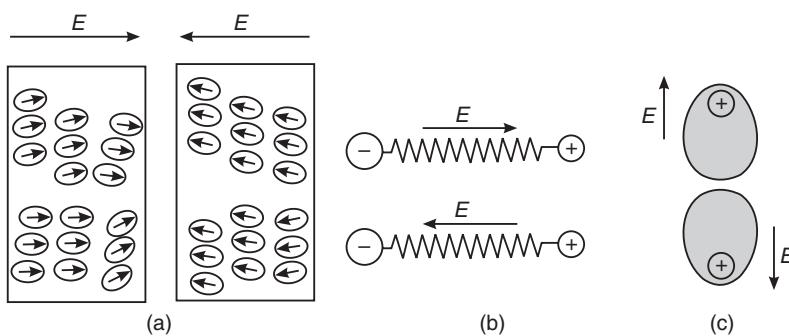


Fig. 33.20: The behaviour of (a) permanent and (b) and (c) induced dipoles in an alternating electric field.

Again, typically in the infrared region the ionic polarization fails to follow the field reversals due to the inertia of the system and the contribution of ionic polarizability ceases. In this region, only electronic polarization contributes to the total polarization. Therefore $P = P_e$. In the optical region, the electron cloud follows the field variations and the material exhibits

an electronic polarizability α_e . The relative permittivity in the optical region will be equal to the square of the refractive index 'n' of the dielectric. Thus,

$$[\epsilon_r]_{\text{optical region}} = n^2 \quad (33.53)$$

In the ultraviolet region, the electron cloud too fails to follow the field alternations and electronic contribution to the polarization ceases. Consequently, the total polarization becomes zero. It follows from equation (33.52) that the relative permittivity approaches unity at frequencies above the ultraviolet range. Thus,

$$[\epsilon_r]_{\text{X-ray}} = 1$$

To cite the example of water, the low frequency dielectric constant, generally referred to as **static dielectric constant**, at room temperature is about 80. It falls to about 1.9 in the optical region.

33.17 THE INTERNAL FIELD IN SOLIDS

In gases the atoms are in constant random motion and are separated by sufficiently large distances. As such the interaction between the atoms can be neglected. When an external field E is applied, the intensity of the electric field felt by a given atom in the gas will be equal to the applied field E . In case of solids and liquids subjected to external electric field, the atoms are surrounded on all sides by other polarized atoms, and the internal intensity of the electric field at a given point of the material is, in general, not equal to the intensity of the applied field E . The **internal field** \mathbf{E}_i , which is defined as the electric field acting at the location of a given atom, is given by the sum of the electric fields created by the neighbouring atoms and the applied field. In evaluation of the bulk polarization, the additional effects of the surrounding polarized atoms are to be taken into account. The effective field intensity E_i in the dielectric is given by

$$\mathbf{E}_i = \mathbf{E} + \mathbf{E}'$$

where E' is the field due to neighbouring atoms. The value of E' can be evaluated by the summation of all the effects of the surrounding atoms. To illustrate the method of evaluation, let us consider a one-dimensional solid consisting of a string of equidistant identical atoms, each of polarizability α_e as depicted in Fig. 33.21.

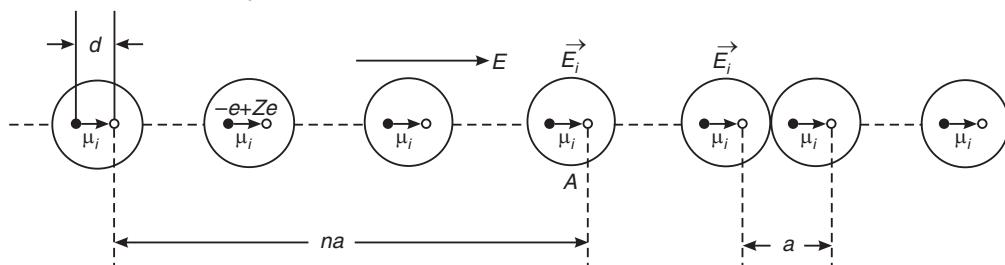


Fig. 33.21. The electric field \mathbf{E}_i seen by the atoms is different from the external field \mathbf{E} which is, say given by $\mathbf{E} = V/L$. A one-dimensional solid is considered for computation of local field.

Let us consider an external field \mathbf{E} applied in a direction parallel to the string. We shall determine the net internal field \mathbf{E}_i experienced by one of the atoms, say A. The field seen by all other atoms will also be the same, and from the consideration of symmetry \mathbf{E}_i will be parallel to \mathbf{E} . The dipole moment induced in each of the atoms of the string is therefore

$$\mu_{\text{ind}} = \alpha_e E_i$$

The field at A due to the dipole induced in an atom located at a distance ' na ' from it is given

$$\begin{aligned} E_n &= \frac{Ze}{4\pi\epsilon_0} \left[\frac{1}{(na)^2} - \frac{1}{(na+d)^2} \right] = \frac{Ze}{4\pi\epsilon_0} \left[\frac{(na+d)^2 - (na)^2}{(na)^2(na+d)^2} \right] \\ &= \frac{Ze}{4\pi\epsilon_0} \left[\frac{2nad + d^2}{(na)^2(na+d)^2} \right] \cong \frac{2Zed}{4\pi\epsilon_0(na)^3} \quad (\text{since } d \ll na) \\ &= \frac{\mu_i}{2\pi\epsilon_0(na)^3} \quad (\text{since } \mu_i = Zed) \end{aligned}$$

The total field E_i at A is given by

$$E_i = E + \frac{\mu_i}{2\pi\epsilon_0} \left[2 \sum_{n=1}^{\infty} \frac{1}{(na)^3} \right] \quad (33.54)$$

The factor 2 in the parenthesis takes into account the atoms to the left and to the right of atom A .

$$E_i = E + \frac{\mu_i}{\pi\epsilon_0 a^3} \sum_{n=1}^{\infty} \frac{1}{n^3}$$

or

$$E_i = E + \frac{1.2\mu_i}{\pi\epsilon_0 a^3} \quad (33.55)$$

Thus, the combined effect of induced dipoles of neighbouring atoms is to produce a net field at the location of a given atom, which is larger than the applied field. It is seen from the equation (33.55) that the greater the polarizability α_e or the smaller the intermolecular spacing ' a ', the larger is the internal field.

33.18 LORENTZ FIELD

The local field in a three dimensional solid is determined by the structure of the solid. An accurate calculation of the internal field in solids and liquids is in general very complicated.

Let us consider a dielectric slab kept in a uniform electric field, \mathbf{E} (Fig. 33.22). Let a molecule be at the point O and be surrounded by a spherical cavity of radius r . Let r be arbitrary but sufficiently large compared to molecular dimensions and sufficiently small compared to the dimensions of the dielectric slab. The spherical cavity contains many molecules within it. The molecule at O experiences three electric fields acting on it.

1. The external electric field \mathbf{E}
2. The field \mathbf{E}_1 due to induced charges on the surface of the spherical cavity
3. The field \mathbf{E}_2 due to the molecular dipoles present in the spherical cavity.

Therefore, the total internal field intensity, E_i is given by

$$\mathbf{E}_i = \mathbf{E} + \mathbf{E}_1 + \mathbf{E}_2 \quad (33.56)$$

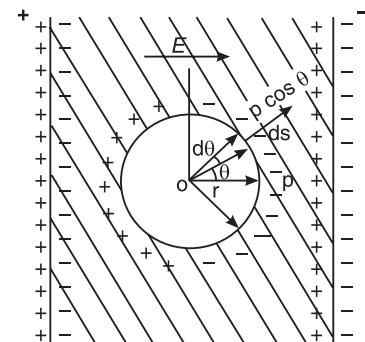


Fig. 33.22. Computation of Lorentz field.

To calculate E_1 , let us imagine that the dielectric is removed from the sphere. For the actual pattern of the electric field not to be distorted, a surface electric charge should be placed on the spherical surface. At each point of the sphere, the surface charge density is given by

$$\sigma = P \cos \theta$$

where θ is the angle between radius vector \mathbf{r} and the direction of \mathbf{E} . The charge on element dS of the surface of the sphere will be

$$dq = \sigma dS = P \cos \theta dS \quad (33.57)$$

This charge will produce electric field intensity $d\mathbf{E}_1$ at the center of the sphere.

$$d\mathbf{E}_1 = \frac{dq}{4\pi\epsilon_0 r^2} = \frac{P}{4\pi\epsilon_0 r^2} \cos \theta dS \quad (33.58)$$

This electric field can be resolved into two components: one component $dE_1 \cos \theta$ parallel to the direction of \mathbf{E} and the other $dE_1 \sin \theta$ perpendicular to the direction of \mathbf{E} .

$$dE_1 \cos \theta = \frac{P}{4\pi\epsilon_0 r^2} \cos^2 \theta dS \quad (33.59)$$

and

$$dE_1 \sin \theta = \frac{P}{4\pi\epsilon_0 r^2} \cos \theta \sin \theta dS \quad (33.60)$$

It is obvious that the perpendicular components of the upper and lower half of the sphere cancel each other and only the parallel components contribute to the total intensity E_1 . E_1 is obtained by integrating dE_1 over the whole surface area of the sphere. Thus,

$$E_1 = \int_0^\pi dE_1 \cos \theta dS = \frac{P}{4\pi\epsilon_0 r^2} \int_0^\pi \cos^2 \theta dS$$

But $dS = 2\pi r^2 \sin \theta d\theta$. Therefore,

$$E_1 = \frac{P}{2\epsilon_0} \int_0^\pi \cos^2 \theta \sin \theta d\theta$$

Let $\cos \theta = x$ and therefore, $-\sin \theta d\theta = dx$.

$$\therefore E_1 = -\frac{P}{2\epsilon_0} \int_1^{-1} x^2 dx = -\frac{P}{2\epsilon_0} \left[\frac{x^3}{3} \right]_1^{-1} = \frac{2P}{6\epsilon_0}$$

or

$$E_1 = \frac{P}{3\epsilon_0} \quad (33.61)$$

It may be deduced from Fig. 33.22 that the direction of \mathbf{E}_1 coincides with the direction of \mathbf{E} . As there exists symmetrical distribution of molecular dipoles around the molecule at O within the cavity, their contributions cancel each other.

$$\therefore E_2 = 0.$$

Hence the total internal field is given by

$$\mathbf{E}_i = \mathbf{E} + \mathbf{E}_1$$

$$\therefore \mathbf{E}_i = \mathbf{E} + \frac{\mathbf{P}}{3\epsilon_0} \quad (33.62)$$

The field given by the above equation (33.62) is called **Lorentz field or local field**.

33.19 CLAUSIUS-MOSOTTI EQUATION

Let us consider now the simple case of an elemental solid dielectric, which exhibits only electronic polarizability. Solids such as diamond, silicon and germanium crystals are made up of single type of atoms. If α_e is the electronic polarizability per atom, it is related to the bulk polarization P through the relation

$$\alpha_e = \frac{P}{NE_i} \quad (33.62)$$

where N is the number of atoms per m^3 and E_i is the local field. Following the equation (33.61) for E_i , we write

$$\alpha_e = \frac{P}{N \left[E + \frac{\gamma P}{\epsilon_0} \right]} \quad (33.63)$$

where γ is known as the internal field constant.

According to equation (33.17)

$$E = \frac{P}{\epsilon_0(\epsilon_r - 1)}$$

If the internal field is assumed to be Lorentz field, $\gamma = 1/3$ (see eqn. 33.62), and equation (33.63) becomes

$$\alpha_e = \frac{P}{N \left[E + \frac{P}{3\epsilon_0} \right]}$$

Using the relation (33.17) into the above equation, we obtain

$$\alpha_e = \frac{P}{N \left[\frac{P}{\epsilon_0(\epsilon_r - 1)} + \frac{P}{3\epsilon_0} \right]}$$

or

$$\frac{N\alpha_e}{\epsilon_0} = \frac{1}{\left[\frac{1}{\epsilon_r - 1} + \frac{1}{3} \right]} = \frac{1}{\left[\frac{\epsilon_r + 2}{3(\epsilon_r - 1)} \right]}$$

or

$$\frac{3(\epsilon_r - 1)}{\epsilon_r + 2} = \frac{N\alpha_e}{\epsilon_0}$$

∴

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{N\alpha_e}{3\epsilon_0} \quad (33.64)$$

The above equation is known as **Clausius-Mosotti equation** which is valid for nonpolar solids having cubic crystal structure.

The measured values of ϵ_r for the three elements of IV group of the periodic table are as follows:

	Diamond	Silicon	Germanium
ϵ_r	5.68 →	12 →	16 →

Example 33.4: The atomic weight and density of sulphur are 32 and 2.08 gm/cm^3 respectively. The electronic polarizability of the atom is $3.28 \times 10^{-40} \text{ F.m}^2$. If sulphur solid has cubic symmetry, what will be its relative permittivity?

Solution.

$$\begin{aligned}\frac{\epsilon_r - 1}{\epsilon_r + 2} &= \frac{N\alpha_e}{3\epsilon_0} = \frac{N_A \rho \alpha_e}{3M\epsilon_0} \\ \frac{\epsilon_r - 1}{\epsilon_r + 2} &= \frac{(6.023 \times 10^{26})(2.08 \times 10^3 \text{ kg/m}^3)(3.28 \times 10^{-40} \text{ F.m}^2)}{3 \times 32 \times 8.85 \times 10^{-12} \text{ F/m}} \\ &= 0.483 \\ \therefore \epsilon_r &= \frac{1.966}{0.517} = 3.8\end{aligned}$$

Example 33.5. A dielectric material has $\epsilon_r = 4.94$ and $n^2 = 2.69$. Calculate the ratio between electronic and ionic polarizability of this material.

Solution.

$$\begin{aligned}\frac{\epsilon_r - 1}{\epsilon_r + 2} &= \frac{N\alpha}{3\epsilon_0} = \frac{N(\alpha_e + \alpha_i)}{3\epsilon_0} \quad \because \alpha_0 \text{ is negligibly small.} \\ \therefore \frac{N(\alpha_e + \alpha_i)}{3\epsilon_0} &= \frac{4.94 - 1}{4.94 + 2} = 0.568\end{aligned} \quad (i)$$

At optical frequencies, $\epsilon_r = n^2$

$$\begin{aligned}\therefore \frac{n^2 - 1}{n^2 + 2} &= \frac{N\alpha}{3\epsilon_0} \\ \text{or} \quad \frac{N\alpha}{3\epsilon_0} &= \frac{2.69 - 1}{2.69 + 2} = 0.360\end{aligned} \quad (ii)$$

Dividing equ. (i) by (ii), we get

$$\begin{aligned}\frac{N(\alpha_e + \alpha_i)}{N\alpha_e} &= \frac{0.568}{0.360} = 1.578 \\ \text{or} \quad 1 + \frac{\alpha_i}{\alpha_e} &= 1.578 \\ \text{or} \quad \frac{\alpha_i}{\alpha_e} &= 0.578 \\ \therefore \frac{\alpha_e}{\alpha_i} &= 1.73\end{aligned}$$

33.20 DIELECTRIC LOSS

When a conductor is subjected to an a.c. or d.c. electric field, it dissipates part of the electrical energy, which gets converted to heat energy. That part of energy is lost or wasted, as no useful work is done by it. The term '**power loss**' denotes the average electrical power dissipated in a material during a certain interval of time. The power loss in conductors is also called **I²R loss or Joules heat**.

As distinct from conductors, the power (I^2R) loss in dielectrics subjected to dc voltages will be very small due to the high resistance of the dielectric materials, whereas the power loss in ac fields will be quite large. The absorption of electrical energy by a dielectric subjected to an alternating electric field is known as the *dielectric loss*. The dielectric loss caused by an ac field also results in dissipation of the electrical energy as heat in the materials. An ideal dielectric does not absorb electrical energy. However, a real dielectric always causes some loss of electrical energy.

The origin of dielectric loss may be understood as follows. An ac field changes its direction with time. With each direction reversal, the molecules are required to follow the field reversals in order to contribute to the polarization of the dielectric. When a capacitor is charged in one half-cycle, the molecules of the dielectric medium are polarized. When the capacitor is discharged in the second half-cycle, the molecules should revert to their initial condition. When it happens, the energy spent in charging the capacitor is completely returned. However, in the process of returning to their initial state, the molecules jostle with each other and lose energy due to friction. The energy lost due to friction takes the form of heat. This energy loss will increase with increase in frequency.

33.20.1 Loss Angle and Loss Tangent

Let us consider a parallel plate capacitor C , constituted by plates of area A and separated by a distance d . Let a dielectric having permittivity ϵ_r fill the space between the plates. Let a sinusoidal voltage V of angular frequency ω be applied to the capacitor. The current through the capacitor is given by

$$I = j\omega CV + \frac{V}{R} \quad (33.65)$$

or

$$I = j I_r + I_a \quad (33.66)$$

The above relations indicate that two kinds of current flow through the dielectric the conduction current $I_a = V/R$ and the displacement current I_r , given by

$$I_r = \omega CV$$

The resultant current $I = \sqrt{I_r^2 + I_a^2}$ lags behind the displacement current by an angle δ .

In case of an ideal dielectric, $R \approx \infty$ and $I_a = 0$; and it would not absorb electric energy. In such a case, the resultant current I would be ahead of the voltage V precisely by a phase angle $\phi = 90^\circ$ and the current would have been purely reactive current I_r .

$$I = I_r = \frac{\omega \epsilon_0 \epsilon_r A V}{d}$$

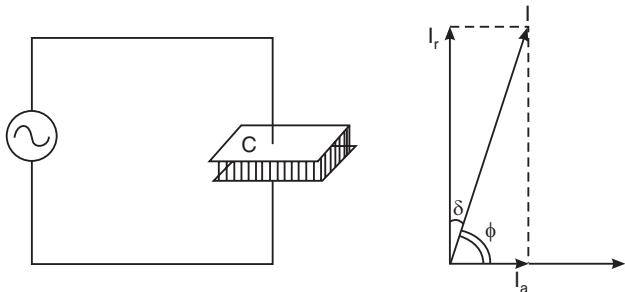


Fig. 33.23. A dielectric in an alternating field. The current through the capacitor is sum of reactive and absorption current components.

However, for a lossy dielectric, the total current is

$$I = I_a + j |I_r|$$

The phase angle ϕ between V and I is now slightly less than 90° . The angle $\delta = (90 - \phi)$ is called the **loss angle**. It is given by

$$\tan \delta = \frac{I_a}{I_r} \quad (33.67)$$

or

$$\tan \delta = \frac{V/R}{\omega CV} = \frac{1}{\omega CR} \quad (33.68)$$

$\tan \delta$ is known as the **loss tangent**. It represents the electrical power lost which is often in the form of heat. Hence it is also called **dissipation factor**.

The real power loss in the dielectric is given by

$$\begin{aligned} P_L &= VI_a = VI_r \tan \delta = \omega CV^2 \tan \delta \\ &= \frac{\omega \epsilon_0 \epsilon_r A V^2}{d} \tan \delta = \frac{2\pi f \epsilon_0 \epsilon_r (Ad)V^2}{d^2} \tan \delta \end{aligned}$$

or

$$P_L = 2\pi v \epsilon_0 E^2 f \epsilon_r \tan \delta \quad (33.69)$$

Substituting the values of 2π and ϵ_0 into the above equation, we obtain

$$P_L = 5.565 \times 10^{-11} v E^2 f \epsilon_r \tan \delta$$

where $v = Ad$ is the volume of the dielectric and $E = V/d$.

It follows from equ. (33.69) that the power loss P_L in a dielectric is related to (i) the dissipation factor (ii) the dielectric constant, (iii) the frequency of the electric field, (iv) the electric field and (v) the volume of the dielectric.

33.20.2 Complex Relative Permittivity

In most of the materials, the dielectric behaviour is more complex indicating the presence of other sources of dielectric loss. To include losses from all sources, we rewrite the equation (33.66) as

$$I = I_a + j |I_r| = j |I_r| \left[1 - j \frac{I_a}{|I_r|} \right] = \frac{j \omega A V \epsilon_0 \epsilon_r}{d} (1 - j \tan \delta) \quad (33.70)$$

Equation (33.70) suggest that the lossy dielectric can be described with the aid of a complex relative permittivity ϵ_r^* given by

$$\begin{aligned} \epsilon_r^* &= \epsilon'_r (1 - j \tan \delta) \\ \text{or } \epsilon_r^* &= \epsilon'_r - j \epsilon''_r \end{aligned} \quad (33.71)$$

$$\text{where } \tan \delta = \frac{\epsilon''_r}{\epsilon'_r} \quad (33.72)$$

Therefore the product $(\epsilon'_r \tan \delta)$ is known as the **loss factor**. A lossy dielectric is represented by a resistance parallel to the capacitor, as shown in Fig. 33.24.

Using (33.72) into (33.69), we can write

$$\text{Power dissipation per unit volume } \frac{P_L}{V} = \omega \epsilon_0 \epsilon''_r E^2 \quad (33.73)$$

At any given frequency, ϵ''_r , produces the same type of macroscopic effect as the conductivity σ and for all practical purposes the two are indistinguishable.

33.20.3 Dielectric Loss Spectrum

The typical variation of ϵ'_r and ϵ''_r of a polar dielectric with frequency are shown in Fig. 33.25.

At audio frequencies, the electric field reverses slowly so that the molecular dipoles can keep shifting their orientation directions in step with the field alternations (Fig. 33.26). There is no power loss in this region. In the r. f. region, the rotation of the dipoles in order to align with the electric field direction is opposed by the internal friction of the material and

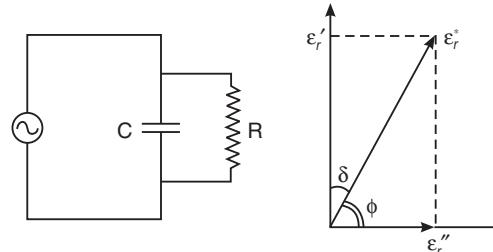


Fig. 33.24. Representation of a lossy dielectric. The complex dielectric constant is at angle to the real part of dielectric constant.

the thermal agitation of the molecules. Consequently, the dipoles lag behind the forces that cause the motion. A phase difference develops between polarization and the electric field resulting in a fall of ϵ'_r as the frequency increases. It is accompanied by heating of the dielectric and therefore by a loss of energy. Energy supplied to maintain the rotation of the dipoles accounts for power loss.

At frequencies of the field near the relaxation frequency of dipoles, the rotation becomes more rapid and the energy loss approaches a maximum at the relaxation frequency. At frequencies above the relaxation frequency, the electric field reverses so rapidly that the dipoles cannot follow the field reversals due to inertia. By the time the dipole attempts to align along the particular direction, the field direction changes. Therefore, the dipoles fail to respond and maintain a random orientation and no more become aligned with the field. Consequently, the dielectric constant is reduced and the power loss decreases after going through a maximum. At much higher frequencies, the losses become negligible. As a result the ϵ''_r vs frequency variation exhibits a bell shaped profile with a maximum at the molecular dipole relaxation frequency and reaching zero on either side. It is seen thus that the greatest loss occurs at frequencies at which the dipoles can almost but not completely be reoriented. At lower frequencies, losses are low because the dipoles have time to rotate. At higher frequencies losses are low because the dipoles do not rotate at all. The losses in this process are known as **relaxation losses**.

Up to the infrared region, ionic dipoles follow electric field variations. When one constituent of the dipole is displaced relative to the other, it again experiences a restoring force proportional to the displacement and executes simple harmonic motion. The natural frequencies of the simple harmonic motion lie in the range of 10^{13} to 10^{14} Hz. When the frequencies of the applied field approaches this region, **resonance** of ionic dipoles occurs and power is absorbed from the field. When the frequency of the field exceeds their natural frequency, the ionic dipoles do not respond and cannot absorb power. The losses encountered in this process are known as **resonance losses**.

Till the optical frequencies, the electron clouds in the molecule respond to the electric field variation of the applied voltage. The electron cloud executes simple harmonic motion, the natural frequency of this motion lying in the range of 10^{17} and 10^{18} Hz. When the frequency of the driving field reaches this value, resonance absorption takes place because of

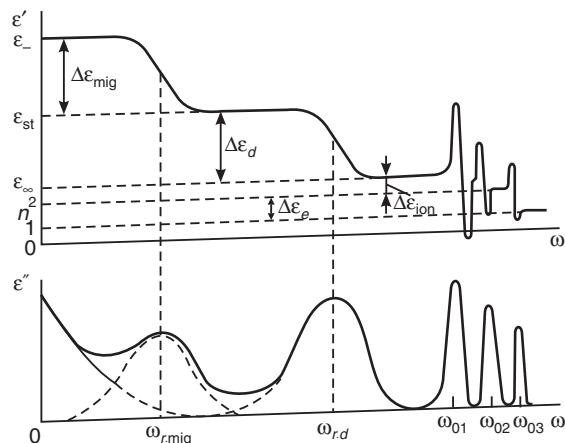


Fig. 33.25. Schematic representation of the frequency dependence of the real and imaginary parts of relative permittivity of a dielectric.

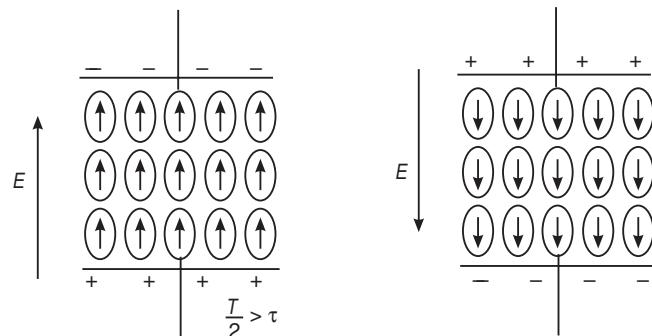


Fig. 33.26

which electrons get excited to higher energy levels. Subsequently, they reemit this energy in a random fashion, in the form of optical photons. For frequencies greater than these, electrons fail to follow the excursions of the field and cannot absorb power any more. The losses encountered in this process are also known as **resonance losses**.

33.21 DIELECTRIC BREAKDOWN

When a dielectric is subjected to very high electric fields, a considerable number of covalent bonds may be torn away and electrons may get excited to energies within the conduction band. These electrons acquire a large kinetic energy and cause localized melting, burning and vapourization of material leading to irreversible degradation and failure of the material. Conducting channels running from electrode to electrode form. It results in high electrical conductivity and total loss of the charge storage property of the dielectric. The formation of such conducting paths in a dielectric under the action of an applied electric field is termed **dielectric breakdown**.

33.21.1 Dielectric Strength

The dielectric strength of a material is a measure of the ability of that material to withstand high electric fields. It is defined as *the maximum electric field that the dielectric can withstand without suffering electrical breakdown*. Thus,

$$\text{Dielectric strength, } E_{\max} = \left(\frac{V_{\max}}{d} \right) \quad (33.74)$$

In other words, the dielectric strength is the limiting electric field intensity above, which a breakdown occurs and the charge storage property of the dielectric disappears. The dielectric strength depends on the thickness of the insulating material and on the length of time for which the dielectric is subjected to electric field. Moisture, contamination, elevated temperature, ageing and mechanical stress usually tend to decrease the dielectric strength of the material.

33.21.2 Breakdown Mechanisms in Solid Dielectrics

The physical pattern of breakdown of solid dielectrics may differ in various cases. However, the fundamental breakdown mechanisms are as follows.

- | | |
|------------------------|----------------------------------|
| 1. Intrinsic breakdown | 2. Thermal breakdown |
| 3. Discharge breakdown | 4. Electrochemical breakdown and |
| 5. Defect breakdown | |

1. Intrinsic breakdown:

When a dielectric is subjected to very high electric fields, a considerable number of covalent bonds may be torn away and electrons may get excited to energies within the conduction bond. Since the electric field is very high, the electrons acquire a large kinetic energy. They collide with other atoms and molecules and release more electrons, which in turn collide with more atoms thereby liberating more electrons. The number of electrons increases very rapidly with time. Conducting channels form running from electrode to electrode and it results in high electrical conductivity. Ultimately, dielectric breakdown occurs. Localized melting, burning and vaporization of material take place at this stage causing irreversible degradation and failure of the material. This type of breakdown is called **avalanche breakdown**.

The characteristics of intrinsic breakdown are as follows:

- It occurs at larger electric fields.

- It occurs at ordinary temperatures.
- It occurs in thin samples.
- The breakdown time is of the order of microseconds.

2. Thermal breakdown:

In dielectric materials, energy due to the dielectric loss appears as heat. This heat must be dissipated away to the surroundings. If the rate of heat generation is larger than the rate of heat dissipation, the temperature of the dielectric increases which results in local melting. Eventually the dielectric breaks down. Thus, dielectric breakdown occurs when the rate of heat generation is larger than the rate of heat dissipation.

The characteristics of thermal breakdown are as follows:

- It occurs at high temperatures.
- The breakdown time is of the order of milliseconds.
- In ac fields, the breakdown strength is lower.
- The breakdown strength depends on the size and shape of the material sample.

3. Discharge breakdown:

Breakdown of dielectrics by gas discharge is classified as external breakdown or internal breakdown.

External breakdown is caused by a glow or corona discharge and is observed at sharp edges of electrodes. The discharge causes gradual deterioration of the solid dielectric held between the electrodes. Such deterioration is accompanied by the formation of carbon. Therefore, the damaged areas become conducting. Eventually, conducting paths are formed leading to a powerful arc and total failure of the dielectric. Such breakdown is caused mainly due to the contamination of the dielectric surface by conducting impurities such as dust, moisture etc.

Internal breakdown occurs due to the presence of gas or liquid filled cavities within the solid dielectric. The inherent strength to electric stress of such cavities is low relative to the solid portion. As a result partial discharges may occur in such cavities and cause gradual deterioration of the adjacent solid dielectric.

The characteristics of thermal breakdown are as follows:

- It occurs at low electric fields.
- It depends on the frequency of applied voltage.

4. Electrochemical breakdown:

Electrochemical breakdown is very much related to thermal breakdown. Many materials have free ions which cause leakage current in the presence of electric field. When temperature increases, mobility of ions increases and also increases leakage current. Electrochemical reaction takes place in the material. Field induced chemical reactions reduces the resistance of the dielectric and finally results in breakdown.

The characteristics of electrochemical breakdown are as follows:

- It depends on the concentration of ions and magnitude of leakage current.
- It occurs at ordinary temperatures.

5. Defect breakdown:

If the surface of the dielectric material has defects such as cracks and porosity, impurities such as dust or moisture may deposit at these defects. These impurities lead to breakdown.

33.22 APPLICATIONS

Two most important applications of dielectric materials are as insulating materials and as medium in capacitors. For insulating materials application the dielectric is required to have

low dielectric constant, low dielectric loss, high resistance and high dielectric strength. Further, they should possess adequate chemical stability, high moisture resistance, and suitable mechanical properties for particular service condition.

A. Solid Insulating Materials

Polymers and ceramics are the widely used solid insulators. A variety of plastics, rubbers, waxes, paper, synthetic fibres and fabrics are applied in the form of films, sheets, slabs, tapes, sleeving, tubing, rods and moulding. Plastics such as polyethylene, polytetrafluoroethylene (PTFE) and polystyrene have low ϵ_r and practically no dielectric loss. Porcelain towers are used in high voltage power lines because of their high dielectric strength. The dielectric strength of porcelain bodies is enhanced by glazing their surfaces. Porcelain, glass, mica, alumina and asbestos are widely used ceramics.

Capacitors

A capacitor is an electronic component that stores energy in the form of electric field. Basically, it consists of two conducting plates separated by a dielectric. Capacitors are widely used in electrical and electronic equipments.

(i) Paper Capacitors: In this type of capacitors, one or more layers of extremely thin kraft or linen paper is used as the dielectric medium. The paper is kept between aluminium foils which act as the metal plates. The whole assembly is rolled into a cylindrical element. The dielectric is impregnated with mineral oil or waxes to prevent absorption of moisture.

(ii) Plastic Capacitors: Plastics can be formed in thin, uniform and non-porous films. Such thin plastic films are used as dielectric medium in these capacitors. Some of the materials used are polyester, polycarbonate, polyethylene, polystyrene, polypropylene, poly tetrafluoroethylene (PTFE) and polythene Terephthalate films.

(iii) Ceramic Capacitors: These capacitors use ceramic as the dielectric medium. Low loss low permittivity capacitors are made from steatite which formed in the form of a thin plate or foil. High permittivity capacitors use barium titanate as the dielectric material.

(iv) Mica Capacitors: Muscovite mica is a naturally occurring material and can be laminated into very thin sheets. This material has good mechanical strength and can be used up to high temperatures of the order of 500°C. Impregnants like polystyrene improve the properties of mica.

(v) Glass Capacitors: Very thin plates of glass are used as dielectric in these capacitors. The plates are interleaved with aluminium foil and fused together to form a solid block.

(vi) Electrolytic Capacitors: In electrolytic capacitors, a metallic anode has oxide film grown over it and this oxide layer acts as a dielectric. The anode is surrounded by an electrolytic solution of ammonium borate or sodium phosphate which acts as cathode.

In aluminium electrolytic capacitors, etched aluminium foil is used as anode. Aluminium oxide film is grown over it which acts as a dielectric film. The electrolyte in liquid form is held in contact with dielectric film. Another etched aluminium foil is used as the cathode. The assembly is sealed in an aluminium can.

B. Liquid Insulating materials

Liquid insulating materials are mainly mineral oils and synthetic oils, which are used for the *purpose of insulation* as well as *cooling* in transformers.

Transformers

A transformer is a device used for transmitting power from one circuit to another or from one place to another place. It consists of two windings, primary and secondary windings, linked by a common magnetic flux. During the construction of transformers, the windings are impregnated by varnishes. In case of H.V. transformers used in distribution of power where

very high voltages are present, proper provisions are to be provided to distribute away the heat produced and to provide high dielectric strength. These transformers are usually immersed in liquid dielectrics.

(i) Mineral insulating oil: Mineral oil has very high dielectric strength and is highly viscous. It transfers heat from the transformer windings and core to the outer shield and enables dissipation of the heat generated. The oil should be perfectly free from moisture to maintain its high dielectric strength. Even small traces of water significantly reduce the dielectric strength. Therefore, the oil is periodically dehydrated. Secondly, sludge formation takes place in the oil due to constant heating of the oil during its working and it also should be removed periodically to maintain its initial quality.

(ii) Synthetic insulating oil: Nowadays, synthetic oils are being used in place of mineral oils because synthetic oils are much more resistant to oxidation and fire hazards. Sovol, sovotol etc are some of the synthetic oils widely used in H.V. transformers.

(iii) Miscellaneous insulating oils: Petroleum oils, silicone oils, and vegetable oils belong to this category. They have high thermal stability. They are mainly used as filling medium for transformers, circuit breakers etc and as impregnants for high voltage cables.

33.22.1 Dielectric Heating

Insulating materials can be efficiently heated up by subjecting them to a high voltage of suitable frequency, namely the frequency at which dielectric loss is maximum. The dielectric loss manifests in the form of heat. Adequate heating may be obtained at high voltages of the order of 20 kV having a frequency of about 30 MHz. The chief advantage of this method is that the material is heated up quickly as the heat is produced in the insulating material itself.

Cooking in microwave oven is one of the popular examples of dielectric heating. Water invariably exists in all articles of food, which exhibits dielectric loss in microwave region. In an oven, microwaves produced by a source are distributed by reflection from the metal walls. They pass through the glass-cooking dish and are absorbed by water molecules. The food is cooked due to the heat produced in the absorption process.

Dielectric heating is widely employed in dehydration of food, tobacco etc. Wooden sheets are preferred to be glued by this method. The heat produced in the glue due to the dielectric absorption leads to binding of the wooden sheets. The advantage of this method is that the moisture content of wooden sheets remains unaltered.

33.23 PIEZOELECTRICITY

Dielectric materials may be divided into the two following categories:

- (i) linear dielectrics and
- (ii) nonlinear dielectrics.

(i) Linear dielectrics are those materials in which the polarization \mathbf{P} and displacement \mathbf{D} are directly proportional to the intensity of the electric field \mathbf{E} ; and relative permittivity ϵ_r and susceptibility χ do not depend on the intensity of electric field. They are also known as *passive dielectrics*.

(ii) Nonlinear dielectrics are those materials in which relative permittivity ϵ_r and susceptibility χ depend on the intensity of electric field. These materials are known as *active dielectrics*.

Piezoelectrics, pyroelectrics, ferroelectrics and some of the optical media belong to the category of nonlinear dielectrics. Piezoelectric crystals provide a coupling between electrical and mechanical forces and hence serve as transducers which produce or detect electrical or mechanical signals. Hence they are used to detect very small mechanical displacements

and small amounts of electric charge. All commercial piezoelectric materials used today are ferroelectrics.

33.23.1 Piezoelectric Effect

The French physicists Pierre Curie and Paul-Jean Curie discovered the piezoelectric effect in 1880. When one pair of opposite faces of certain asymmetric crystals such as quartz is compressed, opposite electric charges appear on the other pair of opposite faces of the crystal (Fig. 33.27). If the crystals are subjected to tension, the polarities of the charges are reversed. The development of charges as a result of the mechanical deformation is known as the **direct piezoelectric effect**. Crystals that exhibit piezoelectric effect are called *piezoelectric crystals*.

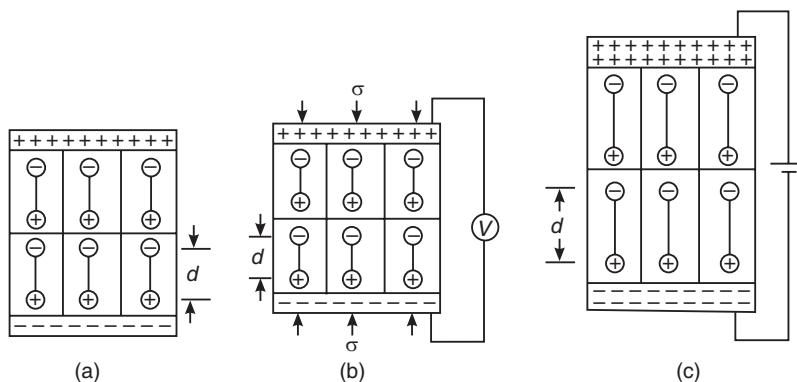


Fig. 33.27. (a) Electric dipoles in a piezoelectric crystal. (b) Mechanical forces cause appearance of polarization charges (c) An external voltage applied across ends of the crystal causes dimensional changes.

Piezoelectric effect is exhibited by a crystal only if the crystalline symmetry is non-centrosymmetric. Crystals are classified into 32 point groups according to their crystallographic symmetry. Out of them, there are 21 point groups, which do not have center of symmetry. Crystals belonging to 20 of these point groups are piezoelectric.

In some ionic crystals the center of positive charge in the unit cell of the crystal does not coincide with the center of the negative charge. Therefore there is a net dipole moment associated with the unit cell of such crystals. Ammonium phosphate, quartz, PZT (lead zirconate titanate) are examples of piezoelectric materials.

The converse effect can also occur. If an electric field is applied across one pair of faces of a piezoelectric crystal, it gets deformed along the direction of the other opposite pair of faces. If an alternating voltage is applied between the two opposite faces of the crystal, it vibrates with the frequency of the field. The mechanical deformation of piezoelectric materials caused by an external electric field is known as the **inverse piezoelectric effect**.

In ordinary solids, a stress causes a proportional strain ' s ' related by an elastic modulus. Piezoelectricity is the additional creation of the electric charge by an applied stress. The induced polarization P , in *direct piezoelectric effect*, is directly proportional to the applied mechanical stress, $σ$. Thus,

$$P = d\sigma \quad (33.75)$$

where d is the proportionality constant and is known as **piezoelectric coefficient** and is expressed in Coulombs/Newton. It may be defined as the charge developed per unit force. A change in sign of $σ$, reverses the sign of polarization. The value of d should be high for practical applications.

In the *inverse piezoelectric effect*, an electric field \mathbf{E} produces a proportional strain, s . Thus,

$$s = d\mathbf{E} \quad (33.76)$$

Thermodynamics proves that the piezoelectric coefficient d of direct and inverse piezoelectric effects are equal for the same dielectric.

In practice, **electromechanical coupling factor**, k is used to describe the piezoelectric effect in actual piezoelectric elements. Energy can be given to a piezoelectric element either mechanically by stressing it or electrically by charging it. All the energy given to it is not converted in producing the effect. Therefore, the piezoelectrics are characterized by *strength of piezoelectric effect*. This strength is measured by the electromechanical coupling factor, k . In case of direct piezoelectric effect, the external force is expended not only on the deformation of the element but also on its polarization. The square of the piezoelectric coupling factor is defined as the ratio of the electrical energy generated by the piezoelectric element to the total energy expended on the deformation. Thus,

$$k^2 = \frac{\text{Mechanical energy converted to electrical energy}}{\text{Total input mechanical energy}} \quad (33.77)$$

In case of inverse piezoelectric effect, the external voltage is expended not only on charging the element but also on its deformation. The square of the piezoelectric coupling factor is defined as the ratio of the electrical energy generated by the piezoelectric element to the total energy expended on the deformation. Thus,

$$k^2 = \frac{\text{Electrical energy converted to mechanical energy}}{\text{Total input electrical energy}} \quad (33.78)$$

The numerical values of k obtained from the direct and inverse piezoelectric effects are found to coincide.

33.23.2 Quartz Crystal

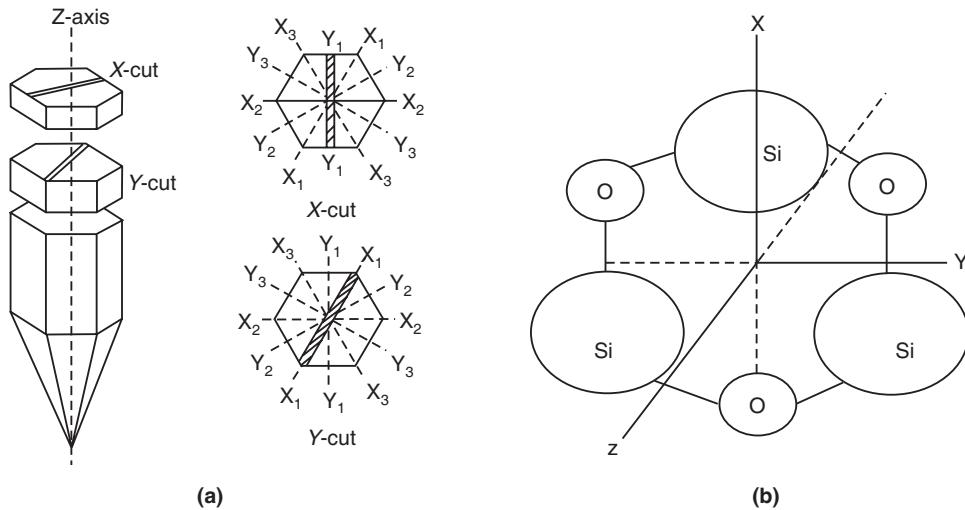


Fig. 33.28

Among piezoelectric crystals, quartz crystal is the most popular piezoelectric. It is the most common variety of silica, SiO_2 . The natural quartz crystal has the shape of a hexagonal prism with a pyramid attached to each end. The axis along the longest dimension of the natural

crystal is called **optic axis** or **z-axis** (see Fig. 33.28 a). The three lines, which pass through the opposite corners of the crystal, constitute its three **x-axes** or **electrical axes**. Similarly, the three lines, which are perpendicular to the sides of the hexagon, form the three **y-axes**, which are known as **mechanical axes**. The arrangement of atoms in the crystal is shown in Fig. 33.28 (b). In the absence of the external stress, all the charges in a unit cell are balanced and the net polarization is zero. When, an external stress is applied to the crystal the balance is disturbed and the crystal gets polarized. The charge developed per unit force is the piezoelectric coefficient d . The d -coefficient for quartz is $2.3 \times 10^{12} \text{ C/N}$ at 550°C .

33.24 FERROELECTRICITY

Ferroelectric materials constitute a very important group of dielectrics. They are anisotropic crystals that exhibit spontaneous polarization. Spontaneous polarization is the dielectric polarization, which occurs under the action of the internal processes and without the application of an electric field. In the absence of an electric field, if the centers of gravity of the positive and negative charges do not coincide, it results in a resultant dipole moment, which is the cause of spontaneous polarization. The materials, which possess special structure that permits spontaneous polarization, are called ferroelectrics and the phenomenon of spontaneous polarization is called **ferroelectricity**. Ferroelectricity was first discovered in Rochelle salt ($\text{Na KC}_4\text{H}_4\text{O}_6 \cdot 4\text{H}_2\text{O}$). It exhibits spontaneous polarization over a range of temperature -18°C to 22°C . Barium titanate, potassium phosphate, and potassium niobate are other examples of ferroelectrics.

The main characteristics of ferroelectric substances are as follows:

1. They possess very high values of permittivity ϵ_r of the order of 1000 to 10,000.
2. The static dielectric constant of ferroelectric materials change with temperature according to the following relation.

$$\epsilon = \frac{C}{T - T_C} \quad (T > T_C) \quad (33.79)$$

Eq. (33.79) is known as **Curie-Weiss law**. C is called the **Curie constant** and T_C the **Curie temperature**. The variation of ϵ_r with temperature in barium titanate is shown in Fig. 33.29.

3. They possess spontaneous electric polarization, that is, polarization without the help of an external electric field. However, the spontaneous polarization occurs only within a definite temperature range and up to the Curie temperature T_C . The variation of spontaneous polarization with temperature is shown in Fig. 33.30.

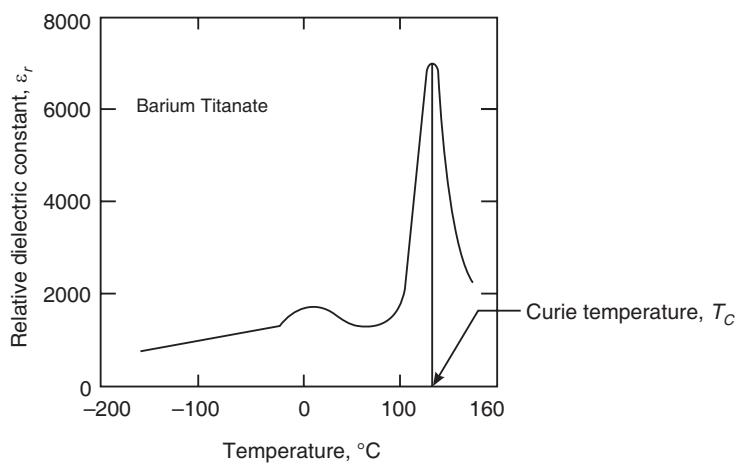


Fig. 33.29

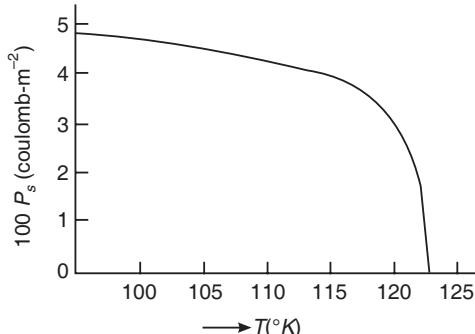


Fig. 33.30

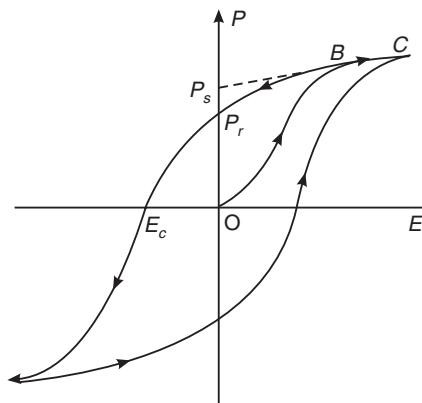


Fig. 33.31

4. In a ferroelectric, the dielectric polarization depends nonlinearly on the applied electric field. In ordinary dielectrics, the polarization varies linearly with the applied electric field. Because of this, ferroelectrics are known as **nonlinear dielectrics**. They exhibit hysteresis under the action of an alternating voltage. The polarization versus electric field curve is known as a *ferroelectric hysteresis loop* (Fig. 33.31).

33.24.1 Polarization Catastrophe

We know that

$$P = \epsilon_0 (\epsilon_r - 1)E$$

$$\therefore E = \frac{P}{\epsilon_0 (\epsilon_r - 1)}$$

We have earlier obtained expression for the internal field in non-polar dielectrics. Now let us assume that it could as well be applied to polar dielectrics. If α is the polarizability per atom, it is related to the bulk polarization P through the following relation

$$P = N\alpha E_i$$

where N is the number of atoms per m^3 and E_i is the local field. The internal field in a solid is given by

$$E_i = E + \frac{\gamma P}{\epsilon_0}$$

Then,
 \therefore

$$P = N\alpha \left[E + \frac{\gamma P}{\epsilon_0} \right]$$

$$P \left[1 - N\alpha \frac{\gamma}{\epsilon_0} \right] = N\alpha E$$

$$P = \frac{N\alpha E}{1 - N\alpha \gamma / \epsilon_0} \quad (33.80)$$

and the dielectric susceptibility $\chi = \frac{N\alpha / \epsilon_0}{1 - N\alpha \gamma / \epsilon_0} \quad (33.81)$

In common solid dielectrics of average permittivity, $N\alpha$ is approximately equal to ϵ_0 . For cubic and isotropic materials, the internal field constant $\gamma = 1/3$. Therefore, $N\alpha\gamma / \epsilon_0 \approx 1/3$. In these cases, the dipole interaction does not exert a substantial influence on the dielectric

properties. However, in cases of certain crystals $N\alpha\gamma / \epsilon_0 \rightarrow 1$, because of higher polarizability and larger internal field constant. From the expression (33.81), it is seen that χ tends to infinity and polarization occurs even without the action of an external electric field. This is **spontaneous polarization**. In ferroelectrics, spontaneous polarization occurs below a certain temperature, known as Curie point due to an increase in polarizability, α , or density, N .

33.24.2 Dielectric Behaviour of Ferroelectrics

Ferroelectricity is closely related to the ionic polarizability. The ions in ferroelectric crystals suffer an asymmetrical shift in their positions. The Clausius-Mosotti relation can be applied to ferroelectrics.

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{N\alpha}{3\epsilon_0}$$

where α is the polarizability of an ion and N the number of ions per unit volume. Now, we assume that α is independent of temperature and the temperature dependence of ϵ_r is solely due to the variation in N which changes due to thermal expansion of the crystal. We rewrite the above equation as

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \beta N \quad (33.82)$$

where β is a constant. Differentiating equ. (33.82) w.r.t. T , we get

$$\frac{1}{(\epsilon_r + 2)} \left[1 - \frac{(\epsilon_r - 1)}{(\epsilon_r + 2)} \right] \frac{d\epsilon_r}{dT} = \beta \frac{dN}{dT}$$

Dividing the above equation by equ. (33.82), we obtain

$$\frac{1}{(\epsilon_r - 1)} \left[1 - \frac{(\epsilon_r - 1)}{(\epsilon_r + 2)} \right] \frac{d\epsilon_r}{dT} = \frac{1}{N} \frac{dN}{dT}$$

or
$$\left[\frac{3}{(\epsilon_r - 1)(\epsilon_r + 2)} \right] \frac{d\epsilon_r}{dT} = \frac{1}{N} \frac{dN}{dT} = -\gamma'$$

where γ' is the volume expansion constant of the material.

As $\epsilon_r \gg 1$, we can approximate $(\epsilon_r - 1)(\epsilon_r + 2) \approx \epsilon_r^2$. Therefore,

$$\frac{3}{\epsilon_r^2} d\epsilon_r = -\gamma' dT$$

Integrating the above equation,

$$\int_{\epsilon_r}^{\infty} \frac{3}{\epsilon_r^2} d\epsilon_r = - \int_T^{T_C} \gamma' dT$$

we obtain

$$\epsilon_r = \frac{3/\gamma'}{T - T_C} \quad (33.83)$$

This relation is similar in form to the Curie-Wiess law. It implies that the temperature dependence of the dielectric constant of the ferroelectric crystal is associated with the expansion of the lattice.

33.24.3 Spontaneous Polarization in a Ferroelectric

Let us consider the case of barium titanate crystal. A unit cell of the barium titanate is shown in Fig. 33.32 (a). It exhibits tetragonal symmetry at temperatures below Curie temperature. The Ba^{2+} ions are located at the corners of the unit cell, the O^{2-} ions in the centers of faces and the Ti^{4+} ion nearly at the center of the unit cell body. In effect one ion of barium and three

ions of oxygen belong to the unit cell and correspond to the empirical formula BaTiO_3 . The titanium ion does not occupy the exact body center of the unit cell. The relative displacement of the O^{2-} ions from their symmetrical positions is shown in the side view of the unit cell (Fig. 33.32 b). It is seen that the Ti^{4+} ion is displaced upward from the center of the unit cell while the O^{2-} ions are located slightly below the centers of each of the six faces. Consequently, a permanent ionic dipole moment arises in each unit cell. Strong interactions between the adjacent permanent dipoles cause all the dipoles to mutually align in the same direction within some volume of the solid. Such regions of spontaneous polarization are known as *ferroelectric domains*. In an unpolarized ferroelectric solid, the polarization vectors of different domains orient in different directions, so that the net polarization of the solid is zero. Above the Curie temperature, the thermal energy causes transformation of tetragonal unit cell into a cubic unit cell. It leads to a change in the relative positions of Ti^{4+} and O^{2-} ions so that the center of action of negative charges is coincident with that of positive charges. Therefore, the net dipole moment becomes zero and the spontaneous polarization vanishes.

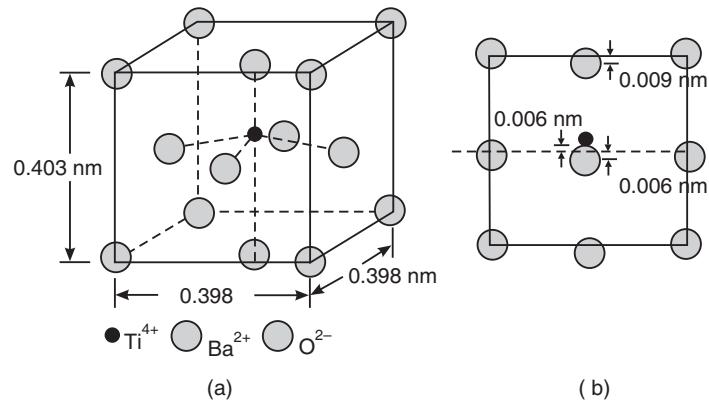


Fig. 33.32. (a) The structure of BaTiO_3 below its Curie temperature 120° C which is slightly tetragonal (b) Due to a slight shift of the central ion Ti^{4+} with respect to the surrounding O^{2-} ions of the unit cell, an electric dipole moment arises in the unit cell.

33.24.4 Ferroelectric Hysteresis Loop

When a virgin ferroelectric crystal is subjected to an alternating electric field, the polarization P versus electric field E describes a closed loop called a ferroelectric hysteresis loop (Fig. 33.31). The polarization increases nonlinearly and reaches saturation at a certain value, P_s . The polarization will not increase further even if the electric field is increased. When the electric field is switched off, the value of the polarization does not return to zero and the crystal retains a *residual polarization* P_r . In order to bring back the polarization to zero, an electric field E_c must be applied in the opposite direction. E_c is known as the *coercive field*. There is a close similarity between the electric properties of nonlinear dielectrics and the magnetic properties of ferromagnetic materials. It is for this reason that the nonlinear dielectrics are called ferroelectrics.

The hysteresis in ferroelectrics is explained on the basis of ferroelectric domains. In the absence of an external electric field, the domains are oriented randomly and the net polarization is zero. When electric field is applied, domains oriented in a favorable direction start growing in size at the expense of unfavorably oriented domains. The growth occurs initially slowly, and then more rapidly. Finally, the unfavourable domains are rotated into the favourable direction till all the domains are lined up in the direction of the applied field. If the electric field is switched off, the domains cannot rotate back to their original orientation and the sample retains remanent polarization, P_r . An electric field ($-E_c$) in the opposite direction is to be applied to disorient the domains.

33.24.5 Poling

Normally, the piezoelectric or ferroelectric materials are used in **ceramic** form i.e. polycrystalline form. They are shaped in the form of discs, rectangular blocks or rings. A polycrystalline disc consists of small crystallites or grains. Each crystallite has a polarization (P_s) oriented in a particular direction. The polarization vectors are randomly oriented in the sample as shown in the Fig. 33.33. The summation of the polarization in the entire sample comes out to be a small value or zero compared to that of a single crystal. Such a sample does not exhibit piezoelectric effect. To impart piezoelectric properties to a ceramic, it is necessary to subject the material to polarization. For this purpose, the sample is held in a strong electric field of the order of 2 to 4 MV/m at temperature of 100° to 150°C for about an hour.

After such a treatment, the polarization vectors in the grains become oriented in a direction close to the field direction and such a sample is called **electrically poled** ceramic sample. After field removal, the sample stays polarized owing to stable remnant polarization.

The ceramic sample thus transforms into an anisotropic body and has a preferred axis of polarization in the direction of the remnant polarization.

Hence, poled ceramics exhibit the piezoelectric or ferroelectrics effects nearly as single crystals even though they cannot replace a single crystal. Their physical properties related to piezoelectricity and ferroelectricity depend on the amount of poling. The electrical poling of a ceramic is analogous to the magnetizing of a permanent magnet.

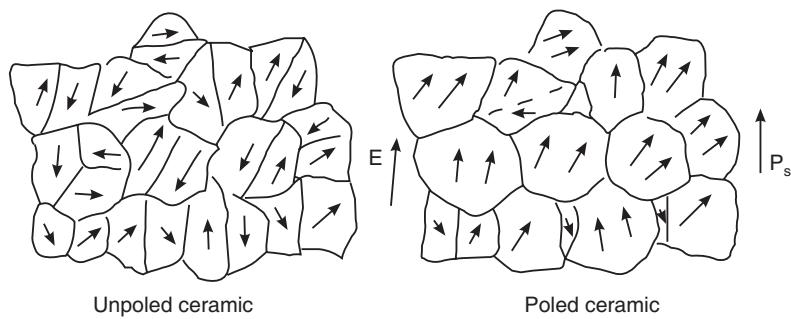


Fig. 33.33

33.25 PYROELECTRICITY

Polarization of a dielectric subjected to uniform heating or cooling is known as pyroelectricity or the **pyroelectric effect**. A Pyroelectric crystal develops an electric polarization with temperature changes. This effect can occur only in substances that display spontaneous polarization. Pyroelectric effect is the change in spontaneous polarization when the temperature of the material is changed. The Pyroelectric properties of a material are described by a Pyroelectric coefficient, p . It is defined as the change in polarization per unit temperature change of the material. Thus,

$$p = \frac{dP}{dT} \quad (33.84)$$

Change in polarization results in change in charge on the surface. With a suitable electrometer, it is now possible to detect a charge of 10^{-16} coulombs. Therefore, temperature changes as small as 10^{-6} can be measured using Pyroelectric effect.

33.26 MATERIALS

Ammonium phosphate, quartz, PZT (lead zirconate titanate) are examples of piezoelectric materials. Quartz crystals are widely used in filter and resonator applications. Rochelle salt

is used as transducer in ear phones, microphones and hearing aids. Barium titanate, lead zirconate and lead titanate are ceramic piezoelectric materials used in gas lighters, accelerometers and transducers. GaS, ZnO, CdS etc are piezoelectric semiconductors which are used in making ultrasonic wave amplifiers. Ferroelectric materials such as quartz, lithium niobate, barium titanate, calcium barium titanate and lead barium niobate are used in making pressure transducers, ultrasonic transducers etc. Pyroelectric materials such as barium titanate, lithium niobate are used in making infrared detectors. TGS, NaNO₂ and PZT ceramics are used in fabrication of pyroelectric image tubes.

33.27 APPLICATIONS

All ferroelectric materials are **pyroelectric**. Since all pyroelectric materials are piezoelectric, ferroelectric materials are inherently **piezoelectric**. Therefore, ferroelectric materials exhibit pyroelectric and piezoelectric properties. Pyroelectric effect is the change in spontaneous polarization when the temperature of the specimen is changed. The magnitude of the spontaneous polarization is greatest at temperatures well below the Curie temperature and approaches zero as the Curie temperature is neared. If in response to an applied mechanical load, the material produces an electric charge proportional to the load, then the material is said to be piezoelectric. Similarly, the material produces a mechanical deformation in response to an applied voltage. Application of the ferroelectric materials utilizes the pyroelectric, piezoelectric or ferroelectric properties of the materials.

1. Capacitors: A capacitor consists of a dielectric material sandwiched between two electrodes. The total capacitance for this device is given by

$$C = \frac{\epsilon_0 \epsilon_r A}{d} \quad (33.85)$$

where 'C' is the capacitance, ϵ_0 is the permittivity of free space, ϵ_r is the relative dielectric permittivity, 'd' is the distance between the electrodes, and 'A' is the area of the electrodes.

To get a high volumetric efficiency (capacitance per unit volume) the dielectric material between the electrodes should have a large dielectric constant, a large area and a small thickness. BaTiO₃ based ceramics show dielectric constant values as high as 15,000 as compared to 5 or 10 for common ceramic and polymer materials. The use of a high dielectric constant ceramic like BaTiO₃, allows large capacitance values to be achieved in relatively small volume capacitor devices.

2. Generation of ultrasonic waves: Piezoelectric crystals provide a coupling between electrical and mechanical forces and hence serve as transducers which produce or detect electrical or mechanical signals. Hence they are used to detect very small mechanical displacements and small amounts of electric charge. All commercial piezoelectric materials used today are ferroelectrics.

High frequency oscillations are transformed into mechanical oscillations in the production of ultrasonic waves. Inverse piezoelectric effect is used in this. Specially cut quartz crystal discs are generally used in this application. Frequencies as high as 50 MHz can be achieved.

3. Vibrators: Another important application is their use as vibrators. When an alternating voltage is applied across a piezoelectric element, it vibrates and at a particular frequency of the field, it resonates. The resonant frequency of the element depends on the thickness. The mechanical vibrations can be transferred into solids or liquids as ultrasonic waves. This frequency range lies from 1000 Hz to 10 MHz. Hard type piezoelectric ceramics are used in this application. The speakers and buzzers utilize the piezoelectric vibrations.

4. Detectors: Piezoelectric ceramics are used in the generation and reception of sound waves in water. They are used in ultrasonic cleaners, and under water detectors of sounds. The ultrasound is also useful for fault detection i.e. for finding internal cracks and other hidden defects in solid bodies like bars, rods, plates etc.

5. Pyroelectric Detectors: Pyroelectricity is the polarization produced due to a small change in temperature. Single crystals of triglycine sulfate (TGS), LiTaO_3 , and $(\text{Sr},\text{Ba})\text{Nb}_2\text{O}_6$ are widely used for heat sensing applications. In these applications ferroelectric thin films for pyroelectric devices is advantageous because of the high cost of growing single crystals and also the thin film geometry is convenient for device design. PbTiO_3 , $(\text{Pb},\text{La})\text{TiO}_3$ and PZT are widely used for thin film pyroelectric sensing applications.

6. Gas Ignitors: In the gas igniter, a very high voltage is generated spontaneously across a piezoelectric ceramic element when a strong mechanical stress pulse is given to the element. This voltage gives rise to a spark. A typical design for a voltage generator for gas ignitors is shown in Fig. 33.34.

It consists of two oppositely poled ceramic cylinders attached end to end in order to double the charge available for the spark. The compressive force has to be applied quickly to avoid the leakage of charge across the surfaces of the piezoelectric ceramic. Usually PZT ceramic disks are used for this application.

7. Accelerometers: An accelerometer is a device which gives an electrical output proportional to the acceleration. A typical accelerometer is shown in Fig. 33.35. The transducer is a piezoelectric cylinder which is poled along its axis but has its poling electrodes removed and the sensing electrodes applied to its inner and outer surfaces. The cylinder is joined to the fixed central pole on the inside and a cylindrical mass on the outside. When an axial acceleration takes place the cylinder is subjected to a shear force between the outer mass and inner pole. Any motion in the radial direction does not give any output. So the device is highly directional.

8. Piezoelectric Transformers:

Low voltage to high voltage transformation can be done by using a piezoelectric plate. Fig. 33.36 shows a flat plate having electrodes on half of it larger face and on an edge. The region between the larger face

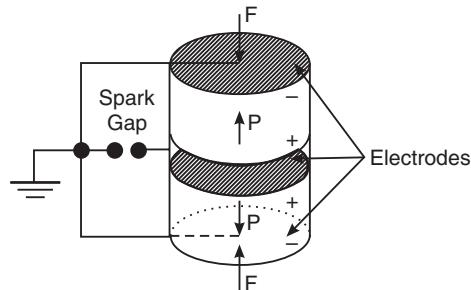


Fig. 33.34

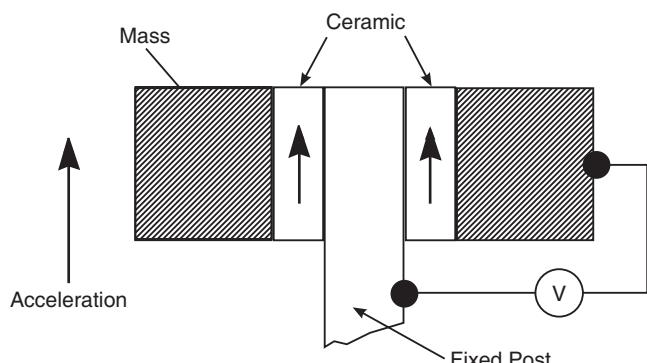


Fig. 33.35. An accelerometer.

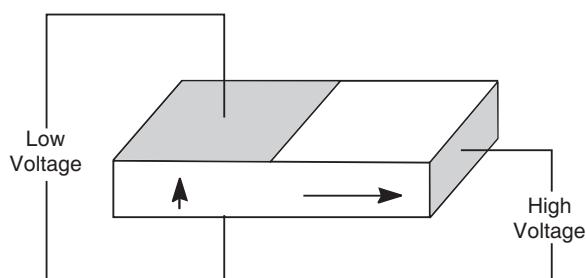


Fig. 33.36. A piezoelectric transformer with the arrows indicating the poling directions

electrodes and the edge electrode are poled separately. A length mode resonance is excited by applying a low AC voltage source between the larger face electrodes. The step up voltage ratio would be proportional to the ratio of the input to output capacitance and the efficiency of the device. This principle has been used for making EHT transformers for miniature television receivers.

9. Impact Printer Head: Dot matrix impact printer heads have piezoelectric displacement sensors. The advantages of the dot matrix printer over the conventional electromagnetic drive printers are high printing speeds, low energy consumption and noiseless printing. Inkjet printer heads are also based on displacement piezoelectric elements.

10. Ferroelectric Memories: Ferroelectric materials spontaneously polarize on cooling below the T_C . The magnitude and direction of polarization can be reversed by the application of an external electric field. The ferroelectric RAMs (FRAMs) made from ferroelectric thin films make use of this phenomenon to store data. Data is stored by localized polarization switching in the microscopic regions of ferroelectric thin films. The ferroelectrics, which exhibit a square loop in polarization versus electric field characteristics, are more useful in memory technology. Barium-strontium titanate, strontium bismuth titanates are used in this application.

Semiconductor memories such as dynamic random access memories (DRAMs) and static random access memories (SRAM's) currently dominate the market. However, the disadvantage of these memories is that they are volatile, i.e. the stored information is lost when the power fails. The non-volatile memories available include complementary metal oxide semiconductors (CMOS) with battery backup and electrically erasable read only memories (EEPROMs). These non-volatile memories are very expensive. The FRAM's are non-volatile because the polarization remains in the same state after the voltage is removed. Further advantage is that FRAMs are little affected by radiation and allow for the use of devices containing these memories in harsh environments such as outer space.

QUESTIONS

1. Explain the behaviour of dielectrics under static electric fields. Derive a relation between polarization P , the external electric field E and displacement vector D . **(C.S.V.T.U., 2008)**
2. What do you understand by dielectric constant? Define dielectric susceptibility. Derive a relation between dielectric constant and dielectric susceptibility. **(C.S.V.T.U., 2009)**
3. With usual notations show that $P = \epsilon_0(\epsilon_r - 1)E$.
4. Explain dielectric polarization.
5. What is meant by dipole moment?
6. What are non-polar dielectrics?
7. What are polar dielectrics?
8. Distinguish between electronic, ionic and orientation polarization and discuss the effect of temperature on each of them.
9. Explain the behaviour of dielectrics under static electric fields. Derive the relation between the polarization P and the external field E .
10. Explain Gauss's law for dielectrics and derive relationship between \mathbf{D} , \mathbf{E} and \mathbf{P} vectors. **(RGPV, 2007)**
11. Define dielectric susceptibility and polarizability of a dielectric. Derive the relation connecting the two.
12. (a) Explain 'polarizability'.

- (b) Discuss the dependence of electronic polarizability on the frequency of the applied field.
 (c) Explain the frequency dependence of relative permittivity. **(Andhra Univ.)**
13. Explain electronic polarizability and show that electronic polarizability for a monoatomic gas increases as the size of the atoms becomes larger.
14. Explain ionic polarizability and derive an expression for ionic polarizability.
15. Define dipole moment and classify dielectric materials on its basis.
16. What are different mechanisms of polarization in a dielectric?
17. Explain briefly the various types of polarization in dielectrics. **(VTU, 2007)**
18. Explain what is meant by a permanent dipole moment. Obtain an expression for orientation polarization.
19. Write one difference between polar and non-polar dielectrics. **(C.S.V.T.U., 2009)**
20. Discuss the frequency dependence of various polarization processes in dielectric materials.
21. Obtain an expression for the internal field seen by an atom in an infinite array of atoms subjected to an external field.
22. Explain the meaning of internal field in solids. Incorporating internal field in the expression for polarization, derive Clausius-Mosotti relation for elemental solid dielectrics.
23. What is meant by local field in a dielectric and how it is calculated for a cubic structure. Deduce the Clausius-Mosotti relation. **(Anna Univ., 2005, 2007)**
24. What are polar and non-polar dielectrics? Derive Clausius-Mosotti equation for a solid dielectric exhibiting electronic polarizability. **(C.S.V.T.U., 2005)**
25. Derive Clausius-Mosotti relation in dielectrics. **(RGPV, 2007)**
26. Explain the concept of internal field in solids and hence obtain an expression for the static dielectric constant of elemental solid dielectric.
27. Derive an expression for internal field in case of liquids and solids. **(VTU, 2007)**
28. Explain the term internal field. Derive an expression for internal field in the case of one dimensional array of atoms in dielectric solids. **(VTU, 2008)**
29. Derive Clausius-Mosotti equation for non-polar solids having cubic crystal structure. **(C.S.V.T.U., 2008)**
30. What do you mean by internal field? Derive Clausius-Mosotti relationship for cubic solids. **(C.S.V.T.U., 2006)**
31. Write short notes on dielectric loss. Show that dielectric loss is given by $\tan \delta = \epsilon''_r / \epsilon'_r$.
32. Explain loss tangent and loss factor.
33. What are the causes behind the dielectric losses occurring in r.f., infrared and visible regions of the electromagnetic spectrum.
34. Discuss in detail the various dielectric breakdown mechanisms. **(Anna Univ., 2003)**
35. (a) Obtain the relevant mathematical expressions for :
 (i) Electronic polarizability and
 (ii) Ionic polarizability
 (b) Distinguish between ferroelectrics and piezoelectrics. **(JNTU, 2010)**
36. What are linear dielectrics? Why are they called passive dielectrics?
37. What are nonlinear dielectrics? Why are they called active dielectrics?
38. Explain the origin of direct piezoelectric effect and inverse piezoelectric effect.
39. Define piezoelectric effect. Discuss some of the important applications of the piezoelectrics.
40. What is meant by spontaneous polarization?
41. Explain the phenomenon of ferroelectricity with particular reference to barium titanate.
42. What are the important characteristics of ferroelectric materials?
43. How does the dielectric constant of a ferroelectric vary with temperature? Mention some of the uses of ferroelectric materials.

44. Describe the ionic displacement theory and show how it explains the ferroelectric nature of barium titanate.
45. Explain the important requirements of insulators.
46. Explain the characteristics and function of transformer oil in transformers.
47. What is electrical poling? Why is it done?
48. Discuss the applications of piezoelectric and ferroelectric materials.

PROBLEMS

1. A parallel plate capacitor, has an area $6.45 \times 10^{-4} \text{ m}^2$ and plate separation of $2 \times 10^{-3} \text{ m}$, and across the plates a potential of 12 V is applied. If a material having a dielectric constant 5 is placed within the region between the plates, calculate the polarization. [Ans: $5.8 \times 10^{-11} \text{ F.m}^2$]
2. Carbon tetra chloride contains 74 electrons in its molecule. Its relative permittivity is 2.26 when its density is $1.68 \times 10^3 \text{ kg/m}^3$. If the field acting on the liquid is $5 \times 10^6 \text{ V/m}$, what is its electronic polarizability and average electron displacement.
[Ans: $8.57 \times 10^{-33} \text{ C.m}$; $0.72 \times 10^{-15} \text{ m}$]
3. A monoatomic gas contains $3 \times 10^{25} \text{ atoms/m}^3$ at a certain temperature at one atmosphere pressure. The radius of the atom is 0.19 nm. What is the relative permittivity of the gas at the given pressure and temperature? What is the polarizability of the atom?
[Ans: 1.00^{26} ; $7.66 \times 10^{-40} \text{ F.m}$]
4. A water molecule has a dipole moment of $6.2 \times 10^{-30} \text{ C.m}$. What is the polarization of a water drop of 0.1 cm radius polarized in the same direction? [Ans: $26.4 \times 10^{-10} \text{ C/m}^2$]
5. The centers of two identical atoms of polarizability $\alpha = 2 \times 10^{-40} \text{ F.m}^2$ are separated by a distance of $5 \times 10^{-10} \text{ m}$. A uniform electric field is applied in a direction parallel to the line joining the centres of the two atoms. Calculate the ratio between the internal field, and E.
6. An elemental dielectric material has $\epsilon_r = 12$ and it contains $5 \times 10^{28} \text{ atoms/m}^3$. Calculate its electronic polarizability assuming Lorentz field.
7. Find the total polarizability of CO_2 , if its susceptibility is 0.985×10^{-3} and density is 1.977 kg/m^3 .
[Ans: $3.24 \times 10^{-40} \text{ F.m}^2$]
8. A solid elemental dielectric having density of $3 \times 10^{28} \text{ atoms/m}^3$ shows an electronic polarizability of 10^{-40} F.m^2 . Assuming the internal electric field to be a Lorentz field, find the dielectric constant of the material.
[Ans: 1.339]
9. A parallel plate capacitor of area 650 mm^2 and a plate separation of 4 mm has a charge of $2 \times 10^{-10} \text{ C}$ on it. When a material of dielectric constant 3.5 is introduced between the plates, what is the resultant voltage across the capacitor?
[Ans: 13.9V]
10. The relative permittivity of sulphur is 4. Calculate its atomic polarizability. Given that sulphur is in cubic form and has a density of $2.08 \times 10^3 \text{ kg/m}^3$ and atomic weight of 32.
[Ans: $3.39 \times 10^{-40} \text{ F.m}^2$]
11. Three identical atoms in a string are subjected to a uniform electric field $E \text{ V/m}$ along the line joining their centers. If the polarizability is $2.5 \times 10^{-40} \text{ F.m}^2$ and the center to center spacing is 0.3 nm, find (E_i/E) at the position of the center of the atom.
12. Carbon tetrachloride contains 74 electrons in its molecule. Its relative permittivity is 2.26 when its density is $1.68 \times 10^3 \text{ kg/m}^3$. If the field acting on the liquid is $5 \times 10^6 \text{ V/m}$, what is its electronic polarizability and average electron displacement? [Ans: $8.57 \times 10^{-33} \text{ F.m}^2$, $7.2 \times 10^{-16} \text{ m}$]
13. A water molecule has a dipole moment of $6.2 \times 10^{-30} \text{ C.m}$. What is the polarization of a water drop of 0.1 cm radius polarized in the same direction?
[Ans: $2.64 \times 10^{-9} \text{ C/m}^2$]

CHAPTER

34

Magnetic Materials

34.1 INTRODUCTION

Magnetic materials play a prominent role in modern technology. They are widely used in industrial electronics, entertainment electronics and computer industry. The methods of information storage and retrieval are based on magnetic storage techniques. Magnetic materials are substances, which upon being introduced into an external magnetic field, change so that they themselves become sources of an additional magnetic field. In 1845 Michael Faraday discovered that the magnetic materials can be broadly classified into three groups, namely diamagnetic, paramagnetic and ferromagnetic materials. The diamagnetic and paramagnetic materials are weakly magnetic and at any temperature they interact weakly with a magnetic field. Ferromagnetic materials interact strongly with a field at definite temperatures. The development of quantum physics helped us understand the phenomenon of magnetism to a great extent. A large number of devices utilize mainly two magnetic phenomena, ferromagnetism and ferrimagnetism. The magnetic materials are classified into soft and hard materials. Soft magnetic materials are easily magnetized and demagnetized and are therefore used in ac applications. Hard magnetic materials retain magnetism on a permanent basis and are used in producing permanent magnets. These materials play an important role in information storage devices. A basic understanding of the magnetic phenomena is essential to appreciate the operating principles of the various magnetic devices.

34.2 TERMS AND DEFINITIONS

When a solid is placed in magnetic field, it gets magnetized. The following terms are of great assistance in understanding the behaviour of materials in the presence of magnetic field.

(i) Magnetic field, \mathbf{H} :

The magnetic field in which a material is kept is called *magnetizing field*. The strength (or intensity) of the magnetic field is denoted by \mathbf{H} . The units of \mathbf{H} are ampere-turns per metre (A/m) in SI system. Magnetic field is produced by permanent magnets such as a horse shoe magnet and temporarily by electromagnets or superconductor-magnets.

(ii) Magnetization, \mathbf{M} :

The magnetic moment per unit volume developed inside a solid is called *magnetization* and is denoted by \mathbf{M} . In SI system, \mathbf{M} is measured in amperes per

metre (A/m). Since the magnetization is induced by the field, we may assume that \mathbf{M} is proportional to \mathbf{H} . Thus,

$$\mathbf{M} \propto \mathbf{H}$$

or

$$\mathbf{M} = \chi \mathbf{H} \quad (34.1)$$

where χ is the proportionality constant and is known as magnetic susceptibility.

(iii) Magnetic Susceptibility, χ :

The *magnetic susceptibility* of a material is a measure of the ease with which the material can be magnetized. It is defined as magnetization produced in the material per unit applied magnetic field. Thus,

$$\chi = \frac{\mathbf{M}}{\mathbf{H}} \quad (34.2)$$

In general, the vectors \mathbf{M} and \mathbf{H} can have different directions and χ is a tensor. However, in isotropic media, \mathbf{M} and \mathbf{H} point in the same direction and χ is a scalar quantity. Materials having high susceptibility are easily magnetized.

(iv) Magnetic Induction, \mathbf{B} :

A magnetic field is schematically represented by lines of magnetic induction. It is described either by magnetic field strength \mathbf{H} or by the *magnetic induction* (or *magnetic flux density*), \mathbf{B} . The lines of induction are collectively called **flux**. The number of field lines passing through a unit area of cross-section is called the magnetic flux density. That is,

$$\mathbf{B} = \frac{\text{Magnetic flux}}{\text{area}} = \frac{\Phi}{A} \quad (34.3)$$

The quantity B is measured in weber per square metre (Wb/m^2) or tesla (T). The cgs unit for magnetic induction is the gauss (G).

$$1 \text{ G} = 10^{-4} \text{ T}$$

Relationship between \mathbf{B} and \mathbf{H} :

When a material is kept in a magnetic field, two types of induction arise: one due to the magnetizing field, \mathbf{H} and the other as a consequence of the magnetization, \mathbf{M} of the material itself. The magnetic induction, \mathbf{B} , produced inside the material is given by

$$\mathbf{B} = \mu_0 (\mathbf{H} + \mathbf{M}) \quad (34.4)$$

where μ_0 is known as the *permeability of the free space*. It is equal to $4\pi \times 10^{-7}$ henry per metre (H/m).

Using equ. (34.1) in (34.4), we get

$$\mathbf{B} = \mu_0 (1 + \chi) \mathbf{H} \quad (34.5)$$

or

$$\mathbf{B} = \mu \mathbf{H} \quad (34.6)$$

where μ is called the absolute permeability of the medium. Like χ , μ is in general a tensor. In isotropic medium, it is a scalar quantity.

In case of free space, $M = 0$ and equ. (34.4) reduces to

$$\mathbf{B} = \mu_0 \mathbf{H} \quad (34.7)$$

(v) Absolute Permeability, μ :

When a magnetic material is placed in a magnetic field, the magnetic field lines are redistributed and tend to pass more (or less in some cases) through the material. The absolute permeability of the material is a measure of the degree of which the field lines penetrate or permeate the material. It is defined as the ratio of the magnetic induction, \mathbf{B} , in the medium to the magnetizing field, \mathbf{H} . Thus,

$$\mu = \frac{\mathbf{B}}{\mathbf{H}} \quad (34.8)$$

The unit of absolute permeability is henry per metre (H/m).

(vi) Relative Permeability, μ_r :

The relative permeability of a material is defined as the ratio of the absolute permeability of that material to the permeability of free space. That is,

$$\mu_r = \frac{\mu}{\mu_0} \quad (34.9)$$

μ_r is only a number and has no units. Its value for air or vacuum is equal to unity. Thus,

$$\mu_r = 1$$

34.3 RELATION BETWEEN μ_r AND χ

Comparing equ. (34.5) and (34.6), we find that

$$\mu = \mu_0(1 + \chi) \quad (34.10)$$

Using equ. (34.9) into the above equation, we obtain

$$\mu_r = (1 + \chi) \quad (34.11)$$

The above equation (34.11) relates the permeability to susceptibility of the material.

Example 34.1. A magnetic material has a magnetization of 2300 A/m and produces a flux density of 0.00314 Wb/m². Calculate magnetizing force and relative permeability of the material.

Solution. Magnetizing force, $H = \frac{B}{\mu_0} - M = \frac{0.00314 \text{ Wb / m}^2}{12.57 \times 10^{-7} \text{ H/m}} - 2300 \text{ A/m} = 198 \text{ A/m}$.

Relative Permittivity, $\mu_r = \frac{B}{\mu_0 H} = \frac{0.00314 \text{ Wb / m}^2}{(12.57 \times 10^{-7} \text{ H / m})(198 \text{ A / m})} = 12.56$.

34.4 ORIGIN OF MAGNETIZATION

The magnetic properties of solids arise due to electrons undergoing different motions in the atoms, which give rise to magnetic dipole moments. These magnetic dipole moments are responsible for the magnetic properties of materials. In general, the magnetic dipole moment of the atom arises from three sources:

1. **The orbital motion of electrons:** The atom of any material consists of a central nucleus and the electrons move around the nucleus in specific orbits. Each electron orbit is equivalent to a tiny current loop and behaves as an elementary magnet having a magnetic dipole moment. The total orbital magnetic moment of an atom is the sum of orbital magnetic moments of individual electrons.
2. **The electron spin:** Each electron is spinning about an axis through itself and this spin also gives rise to a magnetic dipole moment.
3. **The nuclear spin:** In addition to electronic contribution, nuclear spin also contributes to magnetic moment of atoms. The magnetic moment of the nucleus is about 1/2000 of the magnetic moment of electron. Therefore, in studying magnetic properties of solids, the magnetic moment due to nuclear spin is neglected.

In general, the resulting magnetic moment of an atom is the sum of the orbital and spin magnetic moments of its electrons. The major contribution to atomic magnetic moment comes from the spin of unpaired valence electrons. A number of such magnetic moments may align

themselves in different directions to generate a net non-zero magnetic moment. When the substance is placed in a magnetic field, the atomic dipoles are aligned with their directions of magnetic moment along the direction of the external field. Thus, the material is magnetized.

34.4.1 Bohr Magneton

Bohr magneton is the elementary electron magnetic moment. It is elementary in the sense that no electron can have a magnetic moment below it. It is the natural unit for the measurement of atomic magnetic moments. It is denoted by μ_B . It has a value

$$\mu_B = \frac{e\hbar}{4\pi m} = 9.28 \times 10^{-24} \text{ A.m}^2.$$

34.5 CLASSIFICATION OF MAGNETIC MATERIALS

Solids are classified into three groups basing on the magnitude and sign of *relative permeability*, μ_r , exhibited by them. Thus,

Diamagnetic materials: $\mu_r < 1$

Paramagnetic materials: $\mu_r > 1$

Ferromagnetic materials: $\mu_r \gg 1$.

Alternatively, it has been found that the solids can be divided into two broad groups, on the basis of magnetic dipole moments. The atoms of one group do not possess permanent magnetic dipole moment whereas atoms of the other group carry permanent magnetic dipole moment. **Diamagnetic materials** are the materials consisting of atoms with zero magnetic dipole moment. Basing on the interaction of permanent magnetic dipoles, the second group is further subdivided into the following four groups.

- (i) Paramagnetic materials
- (ii) Ferromagnetic materials
- (iii) Antiferromagnetic materials and
- (iv) Ferrimagnetic materials.

Materials composed of atoms or molecules having permanent magnetic moment are classified into four categories depending on the interaction between the atomic magnetic dipoles. If the interaction between the atomic magnetic dipoles is negligible, the material is **paramagnetic**. If the magnetic dipoles interact

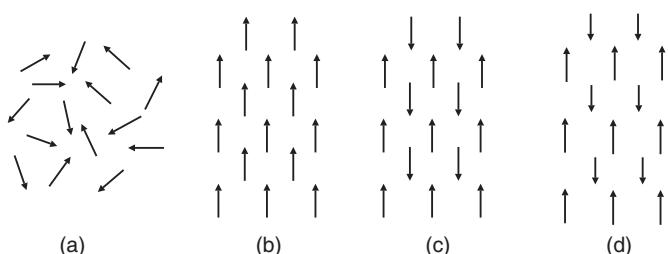


Fig. 34.1. Schematic illustration of the orientation of spins in (a) paramagnetic, (b) ferromagnetic, (c) antiferromagnetic and (d) ferrimagnetic materials

in such a way that they tend to orient in the same direction, the material is **ferromagnetic**. If neighbouring dipoles orient in opposite directions and if the dipoles are of equal magnitude, the material is **antiferromagnetic**. If the neighbouring dipoles are of different magnitude and orient antiparallel, the material is **ferrimagnetic**. These four kinds of orientations of atomic dipoles are illustrated in Fig 34.1.

34.6 DIAMAGNETIC MATERIALS

When placed in a magnetic field, diamagnetic materials acquire feeble magnetism in a direction opposite to that of field. Inert gases, a majority of metals, and many organic

compounds are diamagnetic substances. Hydrogen, air, water, gold, silver, and bismuth are examples. The salient features of diamagnetic materials are as follows:

1. Diamagnetic materials exhibit negative magnetic susceptibility. The magnetization \mathbf{M} in diamagnetic materials is directed opposite to the direction of the applied magnetic field, \mathbf{H} and hence the susceptibility is negative. The absolute value of susceptibility is small and is of the order of 10^{-6} .
2. As diamagnetic susceptibility is negative, the relative permeability μ_r is slightly less than unity ($\mu_r < 1$).
3. Diamagnetic materials are substances that are repelled by a magnetic field. If a diamagnetic body is placed in an inhomogeneous magnetic field, it tends to be pushed into the regions of weaker field. When the material is taken in the form of a small rod and is freely suspended in the magnetic field, it turns to a position perpendicular to the field lines. Diamagnetic materials push aside the field lines, as illustrated in Fig. 34.2 (b).

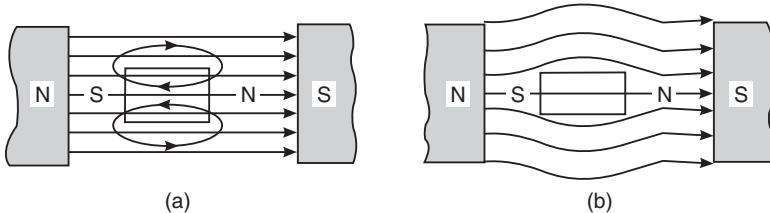


Fig. 34.2

4. The magnetic susceptibility of diamagnetic materials is practically independent of temperature. Pierre Curie discovered the independence of diamagnetic susceptibility on temperature in 1895.
5. The magnetization, M is a linear function of magnetic field H , as illustrated in Fig. 34.3.

The above salient features of the diamagnetic materials may be explained as follows.

Diamagnetism occurs in those materials whose atoms consist of an even number of electrons. The electrons of such atoms are paired. The electrons in each pair have orbital motions as well as spin motions in opposite sense. The magnetic moments cancel each other and hence the resultant magnetic moment of the atom is zero. When a diamagnetic material is placed in an external magnetic field, alignment of magnetic dipoles does not occur, as there are no dipoles in the material. However, the external magnetic field modifies the motion of electrons in the orbits. Following Lenz's law one of the electrons in each pair is accelerated while the other is decelerated. Hence, now each electron pair gives rise to a resultant magnetic moment. Consequently, an effective magnetic moment is *induced* in the atom as a whole in a direction opposite to that of the external magnetic field (Fig. 34.4). As a result the material becomes magnetized.

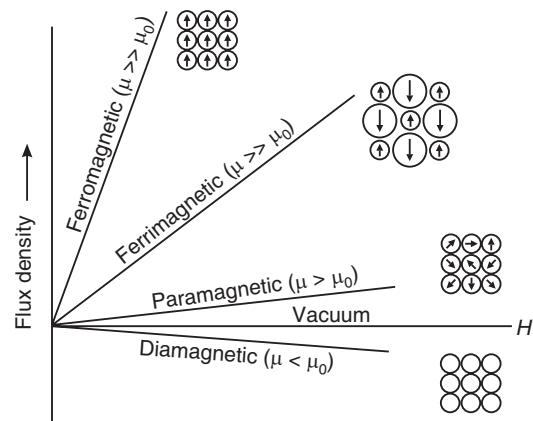


Fig. 34.3

The strength of the induced magnetic moment is proportional to the applied field and hence the magnetization of the material varies linearly with the strength of the magnetic field. The induced dipoles and magnetization vanish as soon as the applied field is switched off. The induced dipoles and their orientation in external field

are not at all influenced by the temperature, with the result that the diamagnetic susceptibility does not depend on the temperature.

In general, diamagnetic materials do not have significant engineering applications. There is one special group known as *superconducting materials*, which are perfect diamagnetic materials. Their susceptibility is given by

$$\chi_{\text{dia}} = \frac{M}{H} = -1 \quad (34.12)$$

which is very large. This strong diamagnetism finds application in making frictionless bearings leading to a levitation effect.

34.6.1 Langevin's Theory of Diamagnetism

In 1905 Langevin explained diamagnetism considering orbital motion of electrons in the atoms. Let us consider an electron revolving in a stationary orbit around the nucleus in an atom. The electrostatic force F_C exerted by the nucleus is the centripetal force acting on the electron. Thus,

$$F_C = m\omega_0^2 r$$

where ω_0 is the angular velocity of the electron in its orbit. Now, let an external magnetic field \mathbf{B} be applied such that it acts perpendicular to the plane of the orbit. It produces an additional force F_m , which will also be centripetal force acting on the electron. The net centripetal force on the electron is

$$F_C \pm F_m$$

depending upon whether the force F_m acts radially inward or outward. Then,

$$F_C \pm F_m = m\omega^2 r \quad (34.13)$$

where ω is the new angular velocity. Here, we assume that the magnetic field will change only the velocity of the electron but not radius of its orbit. We know that

$$F_m = evB = e\omega rB$$

Therefore, we can rewrite eqn. (34.13) as

$$m\omega_0^2 r \pm e\omega rB = m\omega^2 r$$

or

$$m\omega_0^2 \pm e\omega B = m\omega^2$$

or

$$\pm e\omega B = m(\omega^2 - \omega_0^2) = m(\omega - \omega_0)(\omega + \omega_0)$$

Writing $(\omega - \omega_0) = \Delta\omega$ and $(\omega + \omega_0) \approx 2\omega$ into the above equation, we get

$$\pm e\omega B = m\Delta\omega(2\omega)$$

or

$$\Delta\omega = \pm \frac{eB}{2m} \quad (34.14)$$

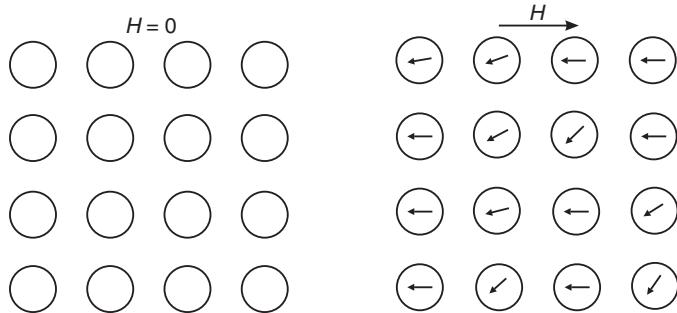


Fig. 34.4

The electron revolving in an orbit behaves as a current loop and possesses a magnetic moment. The magnetic moment associated with a current loop is given by

$$\mu_m = IA = ev A$$

where I is the current in the loop, A the area of the loop and v is the frequency of revolution of electron in the orbit.

As $A = \pi r^2$ and $v = \omega/2\pi$, we obtain

$$\mu_m = e \left(\frac{\omega}{2\pi} \right) \pi r^2 = \frac{1}{2} e \omega r^2$$

A change in angular velocity produces a change in magnetic moment and using equ. (34.14), the magnitude of **induced magnetic moment** is given by

$$\Delta \mu_m = \frac{1}{2} e r^2 \Delta \omega = \frac{1}{2} e r^2 \left(\frac{eB}{2m} \right)$$

or

$$\Delta \mu_m = \frac{e^2 r^2 B}{4m} \quad (34.15)$$

If there are i electrons in an atom, then the total induced magnetic moment is

$$\sum_i \Delta \mu_m = \frac{e^2 \sum r_i^2}{4m} B$$

If there are N atoms per unit volume of the material, then the magnitude of the magnetization, is given by

$$M = N \sum_i \Delta \mu_m = \frac{Ne^2 \sum r_i^2}{4m} B = \frac{Ne^2 \sum r_i^2}{4m} \mu_0 H$$

The vectors μ_m and \mathbf{H} are opposite in direction (due to Lenz's law). Therefore, the above equation in vector form is written as

$$\mathbf{M} = - \frac{\mu_0 Ne^2 \sum r_i^2}{4m} \mathbf{H}$$

The diamagnetic susceptibility is thus,

$$\chi_{\text{dia}} = \frac{\mathbf{M}}{\mathbf{H}} = - \frac{\mu_0 Ne^2 \sum r_i^2}{4m}$$

As all the i electron orbits may not be perpendicular to the applied magnetic field, the right hand side of the above equation is multiplied by a factor of $(2/3)$. Thus, the diamagnetic susceptibility is given by

$$\chi_{\text{dia}} = - \frac{\mu_0 Ne^2 \sum r_i^2}{6m} \quad (34.16)$$

It is seen from the above equ. (34.16) that

- the diamagnetic susceptibility is negative,
- the diamagnetic susceptibility is independent of temperature,
- the diamagnetic susceptibility is directly proportional to the atomic number,

- the diamagnetic susceptibility is larger, if the atom is bigger.

In fact, the induced dipole effect is present in all materials. Therefore, *diamagnetism is a universal property of matter* and all substances have a diamagnetic contribution to their susceptibility. However, in a majority of the materials, diamagnetism is overshadowed by other magnetic phenomenon.

Example 34.2. Diamagnetic Al_2O_3 is subjected an external magnetic field of $10^5 A/m$. Evaluate magnetization and magnetic flux density in Al_2O_3 . (Susceptibility of $Al_2O_3 = -5 \times 10^{-5}$).

Solution: Magnetization $M = \chi H = (-5 \times 10^{-5})(10^5 A/m) = -5 A/m$

$$\begin{aligned} \text{Magnetic flux density, } B &= \mu_0(H + M) = (12.57 \times 10^{-7} H/m)(10^5 A/m - 5 A/m) \\ &= 0.126 \text{ wb/m}^2. \end{aligned}$$

34.7 PARAMAGNETIC MATERIALS

Paramagnetic materials are substances which when placed in a magnetic field acquire feeble magnetism in the direction of the magnetic field. Oxygen, solutions of iron salts, copper chloride, chromium and platinum are examples of paramagnetic materials. The salient features of paramagnetic materials are as follows:

- Paramagnetic materials exhibit positive magnetic susceptibility as the magnetization coincides in direction with the magnetic field H . The susceptibility is of the order of 10^{-6} .
- The relative permeability μ_r is slightly more than unity ($\mu_r > 1$) for paramagnetic materials. Field lines are pulled towards the material and permeate through it when it is placed in a magnetic field, as illustrated in Fig. 34.5.

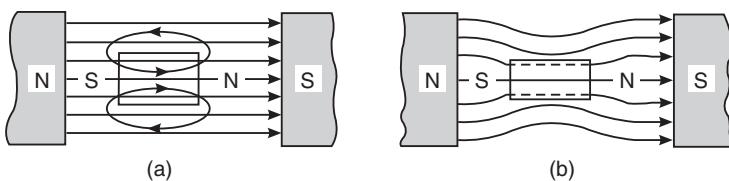


Fig. 34.5

- When a rod of paramagnetic material is freely suspended in a magnetic field, it aligns itself along the lines of induction. In a nonuniform field, the paramagnetic substances are attracted towards stronger region of magnetic field. If a paramagnetic liquid is taken in a U-shaped glass tube and one leg is placed in a magnetic field, the level of the liquid rises in this leg.
- The magnetization, M is a linear function of the magnetic field H , when the field is not too strong, as shown in Fig. 34.3.
- The paramagnetic susceptibility is strongly dependent on temperature. Pierre Curie discovered in 1895 that the susceptibility of a paramagnetic substance varies inversely with the temperature.

Thus,

$$\chi_{\text{para}} = \frac{C}{T} \quad \text{Curie Law} \quad (34.17)$$

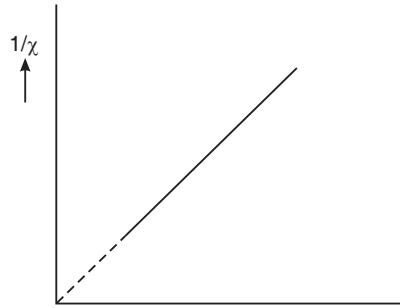


Fig. 34.6

where C is a constant known as Curie constant. The above relation (34.17) is called **Curie's Law**.

The temperature dependence of the paramagnetic susceptibility is conveniently illustrated by $1/\chi_{para}$ plotted as a function of T , as in Fig. 34.6.

Molecules of paramagnetic materials possess a net permanent magnetic moment even in the absence of an external magnetic field. The magnetic moments are randomly oriented in the absence of an external magnetic field. Therefore, the net magnetization of the material is equal to zero (Fig. 34.7 a). When the material is subjected to the influence of magnetic field, the magnetic dipoles tend to align in the direction of the field (Fig. 34.7 b) and the material becomes magnetized. The thermal agitation tends to counteract the orientation of dipoles. When the temperature is increased, the thermal agitation increases and the alignment of dipoles becomes more and more difficult. Therefore, the magnetization and hence susceptibility of paramagnetic materials decrease with an increase in the temperature.

34.7.1 Langevin's Theory of Paramagnetism

The French physicist, Paul Langevin in 1905, developed the classical theory of paramagnetism. The Langevin theory attempts to explain the experimental observations of Curie and Weiss. It is considered in the theory that the magnetic moments arising from the orbital motion of electrons causes paramagnetic behaviour and that a paramagnetic material is conveniently represented by "a gas of magnetic needles". The word gas denotes the fact that the interaction between magnetic moments is neglected. In the absence of an external magnetic field, the magnetic axes of molecules are uniformly distributed in all directions, and the sum of projections of magnetic moments in any chosen direction is zero. Therefore, there is no net magnetization at the macroscopic level.

The potential energy corresponding to a magnetic dipole in a magnetic field of induction \mathbf{B} is given by

$$W = -\mu_m \cdot \mathbf{B} = \mu_m B \cos \theta \quad (34.18)$$

where θ is the angle between the magnetic field and the magnetic dipole. The potential energy attains its minimum value when the direction of dipole orientation coincides with the direction of the magnetic field. Since a system is stable when its potential energy is minimum, the magnetic dipole moments tend to orient along the external magnetic field direction. Therefore, when a paramagnetic material is subjected to a magnetic field, the action of the field turns the dipoles so that they are aligned with the field. Thermal agitation however counteracts the ordering action

of the magnetic field. The degree of orientation of magnetic dipoles in a magnetic field is therefore determined both by the field strength and the temperature of the substance.

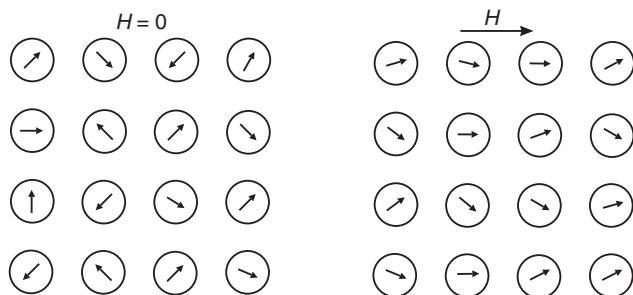


Fig. 34.7

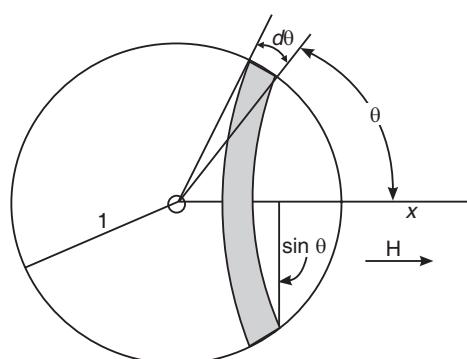


Fig. 34.8

Let us consider a system of N dipoles each carrying a dipole moment μ_m . They are randomly oriented in all directions in the absence of an external magnetic field. Therefore, all directions of orientation are equally probable as a result of which the net magnetic moment of a unit volume of material is zero.

Let us assume that the system is in thermal equilibrium at a temperature T . At any instant, the number of magnetic moments $N(\theta) d\theta$ that lie within a solid angle $d\Omega$ is proportional to the solid angle, which is equal to $2 \pi \sin \theta \cdot d\theta$. Thus,

$$N(\theta)d\theta \propto 2 \pi \sin \theta \cdot d\theta.$$

Let us now suppose that the gas of magnetic moments is subjected to a magnetic field of induction \mathbf{B} . The magnetic field causes orientation of dipoles along its direction. According to Maxwell-Boltzmann statistics, the number of dipoles whose potential energy is W , is proportional to the Boltzmann factor $e^{-W/kT}$. Therefore,

$$\left. \begin{array}{l} \text{The number of dipoles } dN \\ \text{which have potential energy } W \\ \text{and lying within the solid angle } d\Omega \end{array} \right\} = A' d\Omega e^{-W/kT}$$

$$\therefore dN = A'(2\pi \sin \theta d\theta) e^{\mu_m B \cos \theta / kT}$$

$$\text{or } dN = A e^{\mu_m B \cos \theta / kT} \sin \theta d\theta, \quad \text{where } A = 2\pi A'.$$

$$\text{Denoting } \beta = \frac{\mu_m B}{k T}, \text{ we write } dN = A e^{\beta \cos \theta} \sin \theta d\theta \quad (34.19)$$

The total number of dipoles N lying within a unit volume and having potential energy W is given by

$$N = \int_0^\pi dN = A \int_0^\pi e^{\beta \cos \theta} \sin \theta d\theta \quad (34.20)$$

If a magnetic dipole μ_m inclined at an angle θ to the applied magnetic field, the component of its moment in the field direction is $\mu_m \cos \theta$. The magnetic moment contributed by dN dipoles is given by $dN \cdot \mu_m \cos \theta$. The total magnetic moment due to N molecules per m^3 is given by

$$p_m = \int_0^\pi dN \mu_m \cos \theta$$

$$\text{But } p_m \text{ is the same as magnetization. Therefore, } M = \int_0^\pi dN \mu_m \cos \theta.$$

$$\text{or } M = A \mu_m \int_0^\pi e^{\beta \cos \theta} \cos \theta \sin \theta d\theta$$

$$\text{From equation (34.20), we have } A = \frac{N}{\int_0^\pi e^{\beta \cos \theta} \sin \theta d\theta}$$

$$\therefore M = N \mu_m \frac{\int_0^\pi e^{\beta \cos \theta} \cos \theta \sin \theta d\theta}{\int_0^\pi e^{\beta \cos \theta} \sin \theta d\theta}$$

Setting $\cos \theta = y$, we get $dy = -\sin \theta d\theta$. We can write the above equation as

$$\begin{aligned}
 M &= N\mu_m \frac{\int_{-1}^{+1} e^{\beta y} y (-dy)}{\int_{-1}^{+1} e^{\beta y} (-dy)} = N\mu_m \frac{\left[\frac{e^{\beta y}}{\beta} \cdot y \right]_{-1}^{+1} - \left[\frac{e^{\beta y}}{\beta^2} \right]_{-1}^{+1}}{\left[\frac{e^{\beta y}}{\beta} \right]_{-1}^{+1}} \\
 &= N\mu_m \frac{\left[-\frac{e^{-\beta}}{\beta} - \frac{e^\beta}{\beta} \right] - \left[-\frac{e^{-\beta}}{\beta^2} - \frac{e^\beta}{\beta^2} \right]}{\left[\frac{e^{-\beta}}{\beta} - \frac{e^\beta}{\beta} \right]} = N\mu_m \frac{-\frac{2}{\beta} \left[\left(\frac{e^\beta + e^{-\beta}}{2} \right) - \left(\frac{e^\beta - e^{-\beta}}{2\beta} \right) \right]}{-\frac{2}{\beta} \left(\frac{e^\beta - e^{-\beta}}{2} \right)} \\
 &= N\mu_m \frac{\cosh \beta - \frac{1}{\beta} \sinh \beta}{\sinh \beta} = N\mu_m \left(\coth \beta - \frac{1}{\beta} \right)
 \end{aligned} \tag{34.21}$$

$$\text{or } M = N\mu_m L(\beta) \tag{34.22}$$

The function $L(\beta) = \left(\coth \beta - \frac{1}{\beta} \right)$ is known as the **Langevin function**.

Case(i): $\mu_m \beta \gg kT$ and $\beta \gg 1$.

$$\begin{aligned}
 \text{Then } |L(\beta)|_{\beta \rightarrow \infty} &= \left| \left(\coth \beta - \frac{1}{\beta} \right) \right|_{\beta \rightarrow \infty} = \left| \frac{e^\beta + e^{-\beta}}{e^\beta - e^{-\beta}} - \frac{1}{\beta} \right|_{\beta \rightarrow \infty} \\
 &= \left| \frac{1 + e^{-2\beta}}{1 - e^{-2\beta}} - \frac{1}{\beta} \right|_{\beta \rightarrow \infty} = 1
 \end{aligned}$$

∴

$$M_s = N\mu_m$$

where M_s denotes the saturation value of magnetization. It will be attained either when the applied field \mathbf{B} is very large or the temperature is very low. It corresponds to the condition where all the dipoles are completely aligned in the field direction.

Case(ii): $\mu_m \beta \ll kT$ and $\beta \ll 1$.

The magnetic moment μ_m of an atom has a magnitude of the order of 10^{-23} J/T. At moderate fields of the order of 1 Tesla, $\mu_m B = 10^{-23}$ J. The factor kT is about 4×10^{-21} J at room temperature. It follows that $\mu_m \beta \ll kT$ for not very strong magnetic fields. For values of $\beta \ll 1$, the Langevin function is given by

$$\therefore L(\beta) = \left[\frac{1}{\beta} + \frac{\beta}{3} - \frac{\beta^3}{45} + \frac{2\beta^5}{945} + \dots \right] - \frac{1}{\beta} \approx \frac{1}{\beta} + \frac{\beta}{3} - \frac{1}{\beta} = \frac{\beta}{3}.$$

∴

$$L(\beta) = \frac{\mu_m \beta}{3kT}$$

Substituting the above value of $L(\beta)$ into equ. (34.22), we obtain

$$M = \frac{N\mu_m \beta}{3} = \frac{N\mu_m^2 B}{3kT}$$

or

$$M = \frac{N\mu_0 \mu_m^2 H}{3kT} \quad (34.23)$$

$$\therefore \chi_{\text{para}}^{\text{orb}} = \frac{M}{H} = \frac{N\mu_0 \mu_m^2}{3kT} = \frac{C}{T} \quad (34.24)$$

where the Curie constant C is given by

$$C = \frac{N\mu_0 \mu_m^2}{3k} \quad (34.25)$$

Equation (34.24) is the Curie's law.

Example 34.3. The susceptibility of paramagnetic FeCl_3 is 3.7×10^{-3} at 27°C . What will be the value of its relative permeability at 200°K and 500°K ?

Solution. Curie constant $C = \chi T = (3.7 \times 10^{-3})(300\text{K}) = 1.11 \text{ K}$

$$\chi_{200\text{K}} = \frac{C}{T} = \frac{1.11\text{K}}{200\text{K}} = 5.55 \times 10^{-3}$$

$$\chi_{500\text{K}} = \frac{C}{T} = \frac{1.11\text{K}}{500\text{K}} = 2.22 \times 10^{-3}.$$

34.8 FERROMAGNETIC MATERIALS

Ferromagnetic materials are metallic crystals which when placed in a magnetic field become strongly magnetized in the direction of the field. Iron, nickel and some steels are examples of ferromagnetic materials. The experiments of Einstein and de Haas showed that the spin magnetic moments of electrons are responsible for ferromagnetism. In definite conditions, the magnetic moments of adjacent atoms become aligned parallel to one another even in the absence of external magnetic field. Consequently, ferromagnetic materials exhibit spontaneous magnetic moment even in the absence of applied magnetic field. The internal magnetic field has an induction that is hundreds and thousands of times greater than the induction of external magnetic field producing the magnetization. It means that $M \gg H$ and we may write equ. (34.4) as

$$B \approx \mu_0 M$$

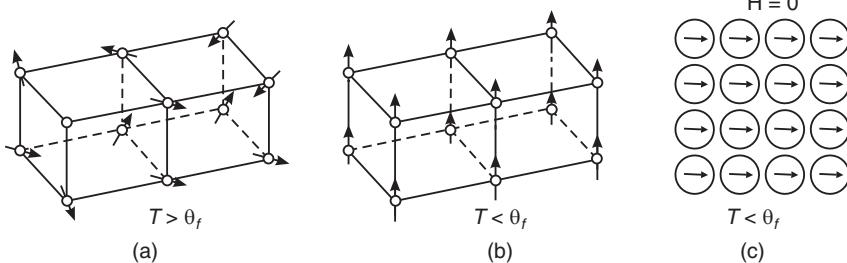
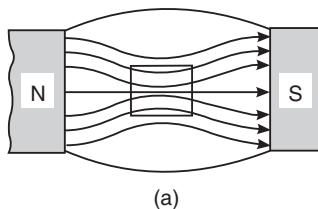


Fig. 34.9

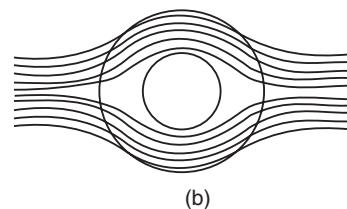
The salient features of ferromagnetic materials are as follows.

1. Ferromagnetic materials exhibit very high values of magnetic susceptibility and relative permeability. Susceptibilities as large as 10^6 and relative permeability of the order of a few thousands are common.

2. When a ferromagnetic material is kept in a magnetic field, the field lines crowd into the material, as



(a)



(b)

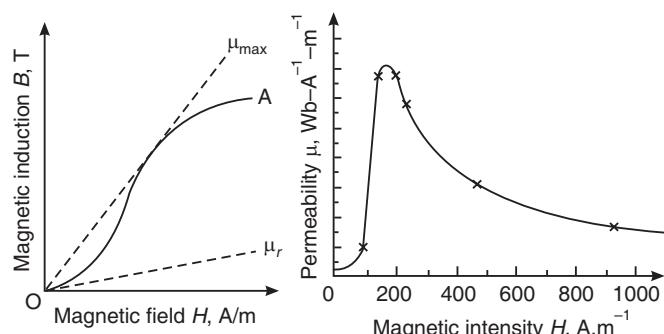
Fig. 34.10

shown in Fig. 34.10(a). If the material is taken in a shell form, the magnetic lines are concentrated mainly in the shell without penetrating into the cavity, as shown in Fig. 34.10(b). It means that a shell made of a ferromagnetic material acts as a **magnetic shield** which does not allow the magnetic field to penetrate into the space enclosed by the shell. Thus, ferromagnetic materials **conduct magnetic flux** much as metals conduct electric current.

3. Magnetization of a ferromagnetic material does not vary linearly with the applied field \mathbf{H} . It is a very complex nonlinear function of the field strength. It is illustrated in Fig. 34.11(a). Initially, \mathbf{M} (therefore \mathbf{B}) increases very fast with a small increase in \mathbf{H} . Then its increase slows down and at large enough values of \mathbf{H} , magnetic saturation is reached in which magnetization becomes practically independent of the strength of the magnetizing field.

The magnetization proceeds along the curve OA (Fig. 34.11 a) in case of an initially non-magnetized ferromagnetic substance. The curve OA is known as the **initial magnetizing curve**.

4. Because of the nonlinear relationship

**Fig. 34.11**

between \mathbf{B} and \mathbf{H} , the permeability of a ferromagnetic material does not have a constant value. The relation (34.6) is not valid. Therefore, in practice, the magnetic permeability of a ferromagnetic is measured by either its initial permeability μ_i or its maximum permeability μ_{\max} . The permeability calculated from the curve OA is plotted as a function of H in Fig. 34.11(b). It is seen that μ rises with increasing H , reaches its maximum value and then rapidly decreases as the magnetic saturation is attained.

5. The specific properties of a ferromagnetic material manifest only in the crystalline state at temperatures lower than a certain temperature. In contrast, diamagnetism and paramagnetism are displayed in any state of aggregation. In the liquid and

gaseous states, ferromagnetic materials behave like ordinary paramagnetic substances.

6. The ferromagnetic properties of crystals are found to be dependent on the direction of magnetization. The magnetization is more readily obtained along certain crystallographic axes than along others. The direction in which the magnetization is the strongest for a given value of the field is called the **direction of easy magnetization** while the direction corresponding to the lowest magnetization for a given field is called the **direction of hard magnetization**. For example iron belongs to BCC lattice. If the magnetic field is applied along <111>, <110> and <111> directions, the crystal is easily magnetized along the cube edge <100> and it is not easy to magnetize iron crystal along the body diagonal direction <111>.

It requires a much higher field to obtain the same magnetization. The magnetization along the different directions for BCC iron is shown in Fig. 34.12.

7. Ferromagnetic materials are characterized by a definite temperature T_C , called the **Curie temperature**, above which ferromagnetic behaviour disappears. The saturation magnetization is a maximum at 0 K in the material. It diminishes gradually with increasing

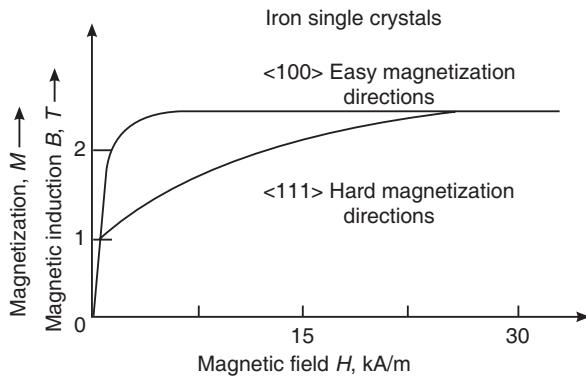


Fig. 34.12

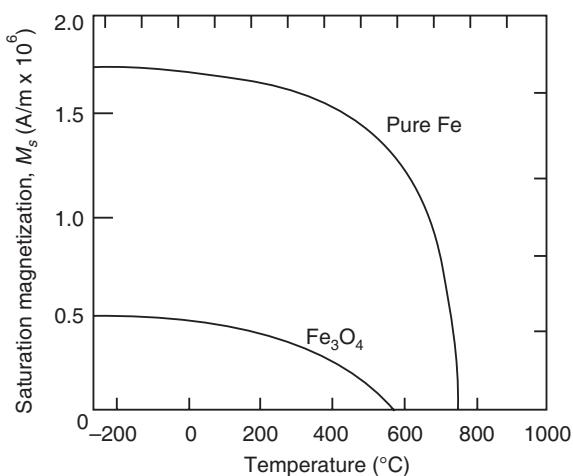


Fig. 34.13

temperature and abruptly drops to zero at the Curie temperature, as shown in Fig. 34.13. The ferromagnetic behaviour vanishes beyond T_C and the material transforms into a paramagnetic material. It is a reversible transition. When the material is cooled back through T_C , the ferromagnetic character reappears. The disappearance of ferromagnetic character at T_C is not accompanied by changes in the crystal lattice. The lattice is not broken up but changes its type of symmetry.

Well above the Curie temperature T_C , the magnetic susceptibility of the paramagnetic state obeys Curie –Weiss law.

$$\chi = \frac{C}{T - \theta}; \quad T > T_C \quad (34.26)$$

Fig. 34.14 shows the variation of $1/\chi$ with temperature above T_C . The ferromagnetic behaviour was first recognized in iron and hence materials, which show similar behaviour are all called **ferromagnetics**.

8. Hysteresis, retentivity, and coercivity:

A typical property of ferromagnetic materials is **hysteresis**. **Hysteresis may be defined as the lag in the changes of magnetization behind variations of the magnetic field.** Because of hysteresis, the magnetization of ferromagnetic material depends not only on the strength of the magnetizing field at the given instant but also on the magnetization history of the material.

If an initially unmagnetized specimen of a ferromagnetic material is subjected to increasing or decreasing magnetic fields, the magnetic field induction B varies as a function of H along a closed loop, called the **hysteresis loop**. The curve (Fig. 34.15) begins at the origin O. As H is increased, the field B begins to increase slowly, then more rapidly and finally attaining a saturation value and becoming independent of H . The maximum value of B is the saturation flux density B_s and the corresponding magnetization is the saturation magnetization M_s .

If H is now decreased, B also decreases but following a path AC instead of the original path AO (Fig. 34.15). Thus, B lags behind H . When H becomes zero, B does not become zero but has a value equal to OC (B_r). This magnetic flux density remaining in the material is called the **residual magnetism**. It indicates that the material remains magnetized even in the absence of an external applied field H . The power of retaining the magnetism is called the **retentivity** or **remanence** of the material. Thus, **the retentivity of a material is a measure of the magnetic flux density remaining in the material when the magnetizing field is removed**.

If the magnetic field H is now increased in the reverse direction, the value of B decreases along the path CD. It becomes zero, when H attains a value equal to OD. It means that to reduce magnetic induction within the material to zero, a field of magnitude H_c must be externally applied in a direction opposite to that of the original magnetizing field. H_c is called the **coercivity** or the **coercive force**. Thus, **coercivity is a measure of the magnetic field strength required to destroy the residual magnetism in the material**.

As the applied field H is increased further in the negative direction, saturation is ultimately reached in the reverse direction (point E on Fig. 34.15).

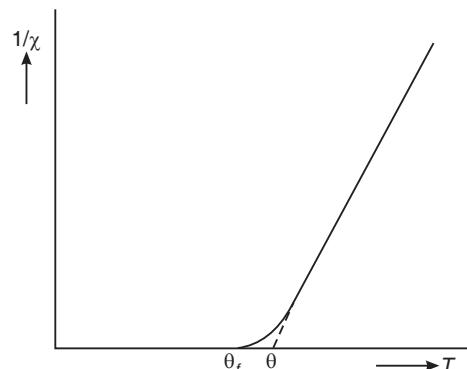


Fig. 34.14

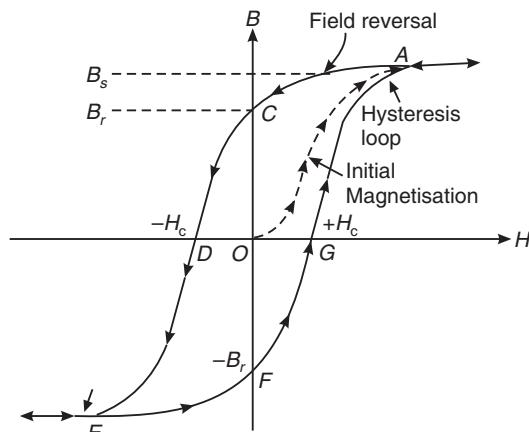


Fig. 34.15: Hysteresis loop

On reversing the variation of the field H , a curve similar to ACDE is traced through points EFGA, yielding a negative remanence ($-B_r$) and a positive coercivity $+H_c$. At points C and F where the specimen is magnetized in the absence of any external magnetic field, it is said to become a **permanent magnet**.

The closed curve ACDEFGA represents a **cycle of magnetization** of the specimen and is known as the **hysteresis loop** of the specimen.

The following Table presents a brief comparison of the three magnetic materials.

Sl.No.	Ferromagnetics	Paramagnetics	Diamagnetics
1.	Solids and possess crystalline structure	Solid, liquid or gas	Solid, liquid or gas
2.	Strongly attracted towards magnetic field	Feeble attraction towards magnetic field	Feeble repulsion from the magnetic field
3.	Field lines are concentrated in the material	More number of field lines pass through the material than outside	Less number of field lines pass through the material than outside
4.	Set along the direction of the magnetic field	Tend to align along the magnetic field direction	Tend to align perpendicular to the magnetic field direction
5.	Susceptibility, χ is large and positive	Susceptibility, χ is <1 but positive	Susceptibility, χ is small but negative
6.	Relative permeability, μ_r is greater than unity	Relative permeability, μ_r is slightly greater than unity	Relative permeability, μ_r is less than unity
7.	Susceptibility, χ decreases with temperature in a complex manner	Obeys Curie law, i.e. $\chi = \frac{1}{T}$	χ is independent of temperature
8.	Have definite Curie point above which they become paramagnetic	No Curie point	No Curie point
9.	B and M vary with H but not linearly and ultimately attain saturation	B and M vary with H linearly at low temperature and at high field tend towards saturation	B and M vary with H linearly but no saturation is reached
10.	Exhibit phenomenon of Hysteresis	Hysteresis is not exhibited	Hysteresis is not exhibited
11.	Possess retentivity	No retentivity	No retentivity
Examples	Iron, steel, cobalt, nickel	Platinum, Chromium, Aluminium, salts of Fe & Ni, oxygen	Bismuth, Mercury, Silver, Copper, water, air

34.8.1 Weiss Theory of Ferromagnetism

In 1907, P. Weiss proposed the molecular field theory to explain qualitatively the salient features of ferromagnetism. He postulated that the internal molecular field sets up spontaneous magnetization and that a macroscopic ferromagnetic specimen is divided up into small regions called **domains**. Each domain is at all times spontaneously magnetized to saturation and has a definite magnetic moment. In a non-magnetized specimen, the directions of the magnetic moments of the domains are distributed randomly so that the material as a whole will have a zero magnetization.

- Spontaneous Magnetization:** Weiss postulated that an *internal molecular field* causes a parallel alignment of magnetic dipoles (34.1 b) and sets up spontaneous magnetization in a ferromagnetic material. In 1928, Heisenberg explained the origin of internal molecular field as due to quantum **exchange interactions** between the electrons. The spontaneous magnetization exists in the material only below the Curie temperature, T_C . The magnetization in the material is a maximum temperature at 0K and decreases with increase in temperature. It falls off rapidly to zero value at T_C . If the material is cooled from above the Curie temperature, the spontaneous magnetization reappears at T_C . At temperatures above Curie point, a ferromagnetic material goes into paramagnetic state (Fig. 34.16 a).
- Ferromagnetic Domains:** Ferromagnetic materials, though spontaneously magnetized, do not show macroscopically observable magnetization. In order to explain why a virgin sample of ferromagnetic material has no net magnetic moment, Weiss put forward domain hypothesis. According to this hypothesis, the entire ferromagnetic volume splits into a large number of small regions of spontaneous magnetization (Fig. 34.16 a). The regions are called the *domains*. Each domain is at all times spontaneously magnetized to saturation and has a definite magnetic moment. In the absence of an external magnetic field, the magnetic moment vectors of the separate domains are oriented in all probable directions so that the net magnetic moment of the entire body equals to zero.

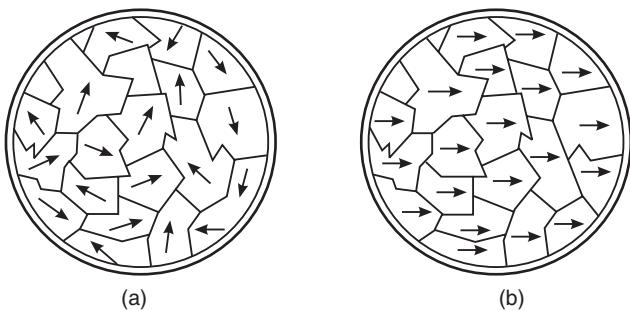


Fig. 34.16

When a magnetic field is applied, the magnetization of the material takes place through two processes. In weak fields, domains, which are favourably oriented with respect to the applied field, increase their volume (grow in size) at the expense of unfavourably oriented domains. In strong fields, domains rotate in an attempt to align their magnetic moments with the field direction (Fig. 34.16 b).

Weiss postulated that in ferromagnetic materials the internal molecular field is responsible for lining up of dipoles in the same direction. The internal field H_i is given by

$$H_i = H + \gamma M \quad (34.27)$$

where H is the applied field, and γM is a measure for the tendency of the environment to align a given dipole parallel to the magnetization already existing. The factor γ is the **molecular field constant** which is also known as **Weiss constant**. It determines the intensity of interaction between the dipoles.

The equation (34.27) explains the Curie-Weiss law and occurrence of spontaneous magnetization by assuming that ferromagnetics are essentially paramagnetic materials having a very large molecular field.

For paramagnetic materials, the susceptibility is given by the expression

$$\chi_m = \frac{C}{T}$$

In case of the paramagnetic state of a ferromagnetic material,

$$\chi_{\text{para}} = \frac{M}{H_i} = \frac{C}{T}$$

Using equ. (34.27) into the above equation, we find that

$$\begin{aligned} \frac{M}{H + \gamma M} &= \frac{C}{T} \\ \text{or } M &= \frac{HC}{T - \gamma C} \\ \therefore \chi &= \frac{C}{T - \gamma C} = \frac{C}{T - \theta} \end{aligned} \quad (34.28)$$

or

Equation (34.28) is identical in form with the Curie-Weiss law, where $\theta = \gamma C$. It is obvious from the equation (34.28) that the value of magnetization tends to infinity at $T = \theta$ (i.e. T_C). It means that the interactions of the individual magnetic moments reinforce each other causing them to align parallel at $T = \theta$.

From measurements of the susceptibility as a function of temperature, the internal field constant γ is of the order of 10^3 . This value is thousand times larger than the value theoretically calculated basing on the assumption that the internal field arises due to the interaction of atomic dipoles. This discrepancy arises because of the fact that the interaction between dipoles is considered to be the classical magnetic interaction. In fact, the interaction is of a quantum mechanical origin and the forces operating between the magnetic dipoles are exchange forces.

Equation (34.28) suggests nonvanishing magnetization for $T = \theta$ even in the absence of magnetic field ($H = 0$). It means that the magnetization at temperature $T \leq \theta$, is spontaneous. We confirm it by the following considerations. When $H = 0$, the magnetic dipoles are subjected only to the molecular field γM . The Langevin variable β may be written now as

$$\begin{aligned} \beta &= \frac{\mu_m H}{kT} = \frac{\mu_m \gamma M}{kT} \\ M &= \frac{kT}{\gamma \mu_m} \beta \end{aligned} \quad (34.29)$$

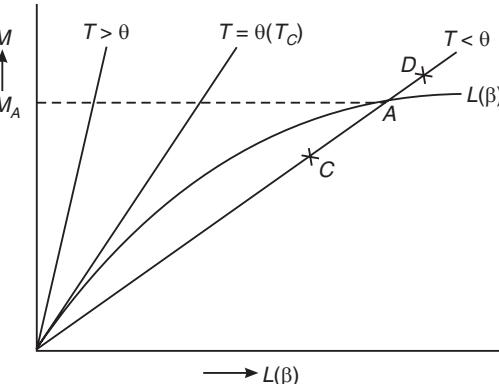


Fig. 34.17

The magnetization M is thus a linear function of β . Magnetization M is also given by the expression (34.21).

$$M = N \mu_m \left[\coth \beta - \frac{1}{\beta} \right]$$

Therefore, the equations (34.21) and (34.29) simultaneously determine the condition of spontaneous magnetization.

Fig. 34.17 shows the plots drawn between M and β . The straight lines depict the linear relationship represented by (34.29) at different temperatures whereas the curve corresponds to the Langevin curve corresponding to the equation (34.21). The intersection point A, of a given temperature line with the Langevin curve, represents the finite spontaneous magnetization at that temperature. If the dipoles assume state C, then the local magnetization is less than the corresponding equilibrium state A. The magnetization and Langevin variable β will increase until the state A is reached. On the other hand, if the dipoles assume state D, then the local magnetization is more than the equilibrium value. Now, the magnetization M and the value of β decrease, until the state A is reached.

With increasing temperature, the straight lines increase in slope according to equation (34.29), thus bringing down the point of intercept A and the value of spontaneous magnetization M . Finally at the Curie temperature T_C , there is no intercept and there is no more spontaneous magnetization. At T_C , the slope of the line ($kT_C / \gamma\mu_m$) is identical to the slope of the Langevin curve near the origin. It is equal to

$$\begin{aligned} \therefore \frac{k\theta}{\gamma\mu_m} &= \frac{N\mu_m}{3} = \frac{M_s}{3} \\ \therefore \gamma &= \frac{3k\theta}{\mu_m M_s} \\ \therefore \gamma M_s &= \frac{3k\theta}{\mu_m} \end{aligned} \quad (34.30)$$

which is the internal molecular field or Weiss field.

$$\begin{aligned} \text{For iron, } \mu_m &= 2.2\mu\beta = 9.273 \times 10^{-24} \text{ Am}^2 \quad \text{and} \quad \theta = 1043 \text{ K} \\ \therefore \gamma M_s &= \frac{3 \times 1.38 \times 10^{-23} J / K \times 1043 K}{2.2 \times 9.273 \times 10^{-24} \text{ Am}^2} = 2117 \text{ T} \end{aligned}$$

The condition for stable spontaneous magnetization is given by

$$\begin{aligned} T &< \frac{\mu_m M_s \gamma}{3k} \\ \therefore T &< \theta \end{aligned}$$

Hence below the Curie point θ (i.e. T_C), the material is **spontaneously** magnetized to a degree depending upon temperature. The magnetization approaches saturation value as the temperature approaches 0 K. Above the Curie point θ , spontaneous magnetization does not occur and the ferromagnetic material transforms into paramagnetic material.

Explanation of Hysteresis Effect on the basis of domains

Hysteresis effect observed in ferromagnetic materials can be explained on the basis of ferromagnetic domains. In the absence of a magnetic field, the domains in the material are randomly directed and the resultant magnetic moment is equal to zero. When the material is placed in a magnetic field H , the orientations of the magnetization vectors of various domains with respect to the magnetic field direction are different. The situation is illustrated Fig. 34.18. It is seen that the magnetization vector of the first domain forms the smallest angle H . With an increase in H , the growth of the most favorably oriented domain 1 becomes energetically advantageous. Therefore, domain 1 grows at the expense of its neighbour domains by sidewise motion of domain walls. Domain 1 grows until the whole crystal becomes a single domain.

Fig. 34.18 shows the magnetization curve for a single crystal. As the applied field is increased, magnetization proceeds along the curve OAB. The portion OA of the curve corresponds to the stage of wall motion. A further increase in the field strength causes the onset of rotation of domain moments into the field direction. The process of magnetization goes on more slowly in this direction. The magnetization at this stage reaches technical saturation. It occurs along the AB portion of the curve.

Beyond B, the growth of magnetization with H is very slow. In the single domain formed under the action of the magnetic field, the individual magnetic moment alignment would not be complete due to the disturbing effect of temperature. An increasing H prevails over the thermal disturbance and tends to orient atomic dipoles.

When the magnetic field is reduced back to zero, the magnetization

curve does not retrace the initial curve BAO. The failure to retrace the original path is due to irreversible processes occurring in the course of demagnetization. Various defects such as impurity atoms, dislocations and nonmagnetic inclusions present in the material have a strong effect on the domain wall motion. They may aid or hinder the domain wall motion. During the course of demagnetization some walls may get stuck up. Consequently, the domains magnetized along the field direction retain their former orientation even after the applied field is made zero. It leads to remanent induction B_r . Elimination of the remanent magnetization requires the application of $H (= H_C)$ in opposite direction.

Example 34.4. Find the relative permeability of the ferromagnetic material if a magnetic field of strength 220 A/m produces magnetization of 3300 A/m in it.

$$\text{Solution. } \mu_r = (1 + \chi) = 1 + \frac{M}{H} = 1 + \frac{3300 \text{ A/m}}{220 \text{ A/m}} = 16.$$

Example 34.5. The saturation magnetic induction of nickel is 0.65 T. If the density of nickel is 8906 kg/m^3 and atomic weight is 58.7, find out the magnetic moment of the nickel atom in Bohr magneton.

$$\text{Solution. Number of atoms / m}^3 N = \frac{\rho N_A}{M} = \frac{8906 \text{ kg/m}^3 \times 6.023 \times 10^{26} \text{ atoms/k.mol}}{58.7}$$

$$= 9.14 \times 10^{28} \text{ m}^{-3}$$

$$\text{Magnetic moment } \mu_m = \frac{B}{N\mu_0} = \frac{0.65T}{9.14 \times 10^{28} / \text{m}^3 \times 4\pi \times 10^{-7}} = 5.66 \times 10^{-24} \text{ A.m}^2$$

$$= \frac{5.66 \times 10^{-24}}{9.27 \times 10^{-24}} = 0.61 \mu_B$$

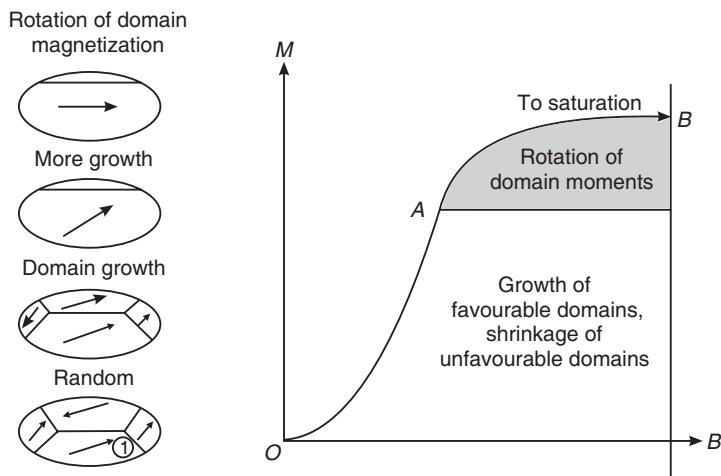


Fig. 34.18

34.9 MAGNETOSTRICTION

When a ferromagnetic material is magnetized, its dimensions change slightly and the sample being magnetized either expands or contracts in the direction of magnetization. The material reverts to the original dimensions on removal of the magnetic field. This magnetically induced reversible elastic strain ($\Delta l/l$) is called **magnetostriiction**. Since the process is reversible, the change in the magnetization can be produced when the dimensions are changed as the effect of an external force. Magnetostriiction is caused by the rotation of domains of a ferromagnetic material under the action of a magnetic field. The rotation of domains gives rise to internal strains in the material causing its contraction or expansion. The magnitude of magnetostriiction for a given material depends on the initial orientation of domains which can be controlled by heat treatment and cold work.

Ferromagnetic materials may show both positive and negative magnetostriiction, varying in magnitude according to the intensity of the magnetic field. Materials showing negative magnetostriiction contract when the magnetic field increases in strength and expand when it decreases. The converse is true for materials showing positive values of magnetostriiction. Iron shows positive magnetostriiction in weak magnetic fields and negative magnetostriiction in strong magnetic fields. Nickel and cobalt show negative magnetostriiction increasing in magnitude with the strength of the magnetic field (Fig. 34.19).

A humming noise ($f=100$ Hz), is produced by transformers due to the periodic change in dimensions caused by the alternating magnetic field ($f=50$ Hz).

34.10 ANTIFERROMAGNETISM

Antiferromagnetic materials are crystalline materials, which exhibit a small positive susceptibility of the order of 10^{-3} and 10^{-5} . The variation of susceptibility with temperature follows a peculiar pattern in these materials. The susceptibility increases with increasing temperature and reaches a maximum at a certain temperature called **Neel temperature**, T_N (Fig. 34.20 a).

a). With a further increase in temperature, the material goes into paramagnetic state. The material is antiferromagnetic below T_N . The transition temperature T_N lies far below room temperature for most of the materials.

In the paramagnetic state, the variation of inverse susceptibility

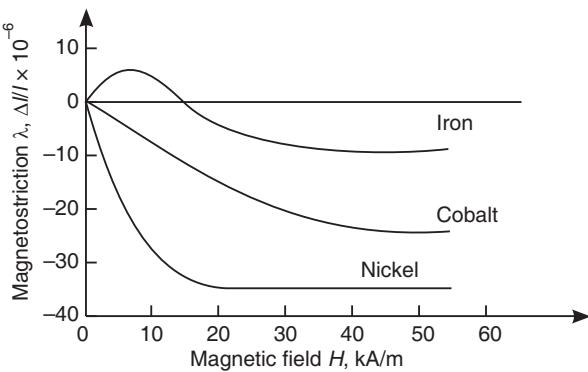


Fig. 34.19: Magnetostriuctive behaviour of iron, cobalt and nickel crystals.

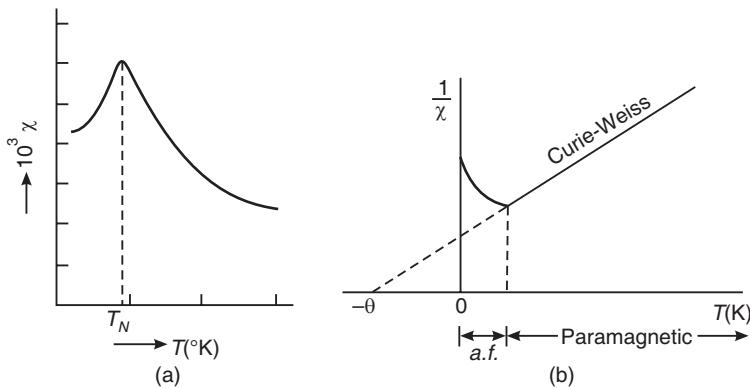


Fig. 34.20

$(1/\chi)$ with temperature is linear, as shown in Fig. 34.20 (b). The extrapolation of the paramagnetic line, in Fig. 34.20 (b) to $1/\chi = 0$ yields a negative θ . The variation of susceptibility with temperature obeys therefore a modified Curie-Weiss law.

$$\chi_{\text{a.f.}} = \frac{C}{T - (\theta)} = \frac{C}{T + (\theta)} \quad T > T_N \quad (34.31)$$

where θ is called the paramagnetic Curie temperature and C the Curie constant.

The elements, manganese and chromium exhibit antiferromagnetism at room temperature. Most of the antiferromagnetic materials are ionic compounds. MnO , MnS , Cr_2O_3 , NiCr are some of the compounds which exhibit antiferromagnetism. Antiferromagnetic materials are of little practical interest.

Antiferromagnetism arises in materials in which the magnetic moments are equal, but adjacent magnetic moments point in opposite directions (Fig. 34.21 a). An antiferromagnetic unit cell can

be considered as composed of two interpenetrating ferromagnetic sublattices, say A and B. The atoms of one of these sublattices have their spins oriented in one direction and the atoms of the other sublattice have their spins oriented in opposite direction, as shown in Fig. 34.21 (b). When there is no external magnetic field acting on them, the magnetization of the unit cell is zero since the magnetic moments of the sublattices are mutually compensated giving a zero net magnetization. When the material is placed in an external magnetic field and the temperature is 0K, the magnetization is still zero. As the temperature is increased in the presence of an external field, the antiparallel arrangement of spins is slowly disturbed and magnetization (and hence the susceptibility) increases (Fig. 34.20 a). The susceptibility reaches a maximum at Neel temperature, T_N . At the Neel temperature, the spin ordering is lost completely. The susceptibility decreases with further rise in temperature and the material transforms into a paramagnetic above T_N .

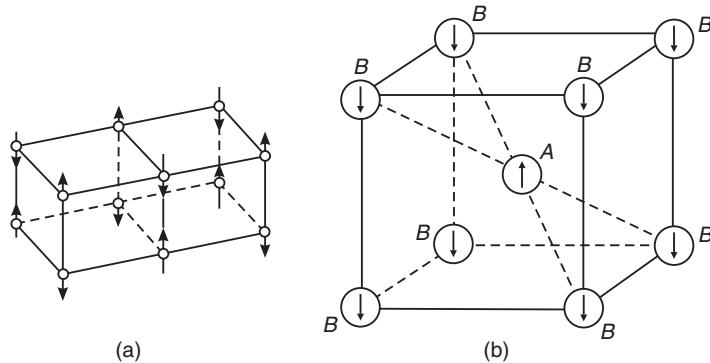


Fig. 34.21

34.11 FERRIMAGNETISM

When the magnetic moments of sublattices in a crystal unit cell are equal in magnitude but opposite in direction, they cancel each other giving rise to antiferromagnetism. But there are certain crystals, in which the magnetic moments of the two sublattices are not exactly equal in magnitude and they are oriented in opposite directions. The net effect is a resultant magnetic moment. A typical example of ferrimagnetic ordering is shown in Fig. 34.1(d). Such crystals possess spontaneous magnetization and exhibit most of the properties of ferromagnets. This uncompensated antiferromagnetism is known as *ferrimagnetism*.

Ferrimagnetic materials are very much similar to ferromagnetic materials in their macroscopic magnetic characteristics. The saturation magnetization in these materials is not as high as for ferromagnetic materials. The saturation magnetization decreases with increasing temperature until it vanishes at a Curie temperature T_C . Above T_C , they exhibit

paramagnetic state, the Curie-Weiss linear relationship is not valid, and the susceptibility is a nonlinear function of temperature. The materials exhibit hysteresis character. They have a residual magnetization, are characterized by a coercive force and so on. They possess small domains in which electron spins are spontaneously aligned. However, the spins are antiparallel. Ferrimagnetism is therefore referred to as uncompensated antiferromagnetism. Ferrimagnetic materials are widely used in high frequency applications where ferromagnetic materials cannot be utilized.

34.12 FERRITES

Ferrimagnetic materials are ceramic materials and are therefore good electrical insulators. The most important group of these materials is ferrites. Ferrites consist chiefly of ferric oxide Fe_2O_3 , combined with one or more oxides of divalent metals. They are represented by a general formula written as $M\text{Fe}_2\text{O}_4$ or $M\cdot\text{O}.\text{Fe}_2\text{O}_3$ in which M represents any one of several metallic elements. An example is magnetite, $\text{FeO}\cdot\text{Fe}_2\text{O}_3$, where Fe^{++} is the divalent metal. The name of the divalent or monovalent metal, whose oxide is in the composition of ferrite, is given to the ferrite. Thus a ferrite is called a zinc ferrite or nickel ferrite depending on whether the composition contains ZnO or NiO .

Composite ferrites, which are solid solutions of one simple ferrite in another show the best magnetic properties. Depending upon the composition and treatment, ferrites with a wide variety of magnetic properties can be produced. Ni – Zn ferrite is an example of a composite ferrite.

Garnets constitute another important group of ferrimagnetic materials. The garnets have a very complicated crystal structure which may be represented by the general formula $M_3\text{Fe}_5\text{O}_{12}$ where M represents a rare earth ion such as samarium, europium, gadolinium or yttrium. Yttrium iron garnet ($\text{Y}_3\text{Fe}_5\text{O}_{12}$) denoted as YIG is the most common material of this type.

Characteristics:

The salient characteristics of ferrimagnetic materials are as follows.

1. They exhibit nonlinear magnetization curve, that is, hysteresis loop. There is a characteristic retentivity and coercivity for each ferrite. Ferrites are polycrystalline samples and the coercivity, retentivity and permeability depend on the grain size.
2. Ferrites possess high permeability values normally. But some ferrites can be made to have low permeability values. The range of permeabilities extends from 15 to 20 for nickel ferrite to several thousand for some manganese-zinc ferrites.
3. Ferrites have high resistivity ranging from 10^2 to 10^{10} ohm-m.

34.13 HYSTERESIS LOSS

When a ferromagnetic specimen is magnetized by keeping it in an external magnetic field, those domains in the specimen that are favourably oriented with respect to the external field grow in size, and also each domain rotates as a single unit such that its direction of magnetization becomes aligned with the field direction. Energy is consumed in this process of domain growth and rotation. When the external field is removed, the domain boundaries do not move back to their original positions and the specimen is left with some magnetization. It means that the energy supplied to the specimen during magnetization is not fully recovered. The difference of energy is lost as heat in the material. This is known as *hysteresis loss*. It is calculated that

Energy dissipated per cycle of magnetization

$$= \oint H dB (J/m^3) = \text{area of the B-H curve.} \quad (34.32)$$

A ferromagnetic or ferrimagnetic material that exhibits a narrow B-H curve produces smaller hysteresis loss. Such a material will be better suited in fabrication of the cores of the transformers and armatures of motors.

34.14 SOFT AND HARD MAGNETIC MATERIALS

Ferromagnetic materials are known as **metallic magnets** and ferrimagnetic materials as **ceramic magnets**. The magnetic flux produced by an electric current is proportional to the relative permeability of the material through which the flux passes. As ferromagnetic and ferrimagnetic materials possess high permeabilities, they produce required flux densities with relatively small currents. Therefore, both these materials are widely used as flux paths and flux multipliers in electrical machines. Transformers, motors and generators etc rotating electrical machines use these magnetic materials as core materials. Basing on the area of the hysteresis loop, these magnetic materials are broadly categorized into two types, namely, **soft magnetic materials and hard magnetic materials**.

The area within a hysteresis loop represents a magnetic energy loss per unit volume of material per magnetization-demagnetization cycle. The energy loss is manifested as heat that is generated within the specimen and raises its temperature. Therefore, in such applications where the material is subjected to ac fields, materials having narrow and small area hysteresis loop are to be used. Materials having narrow hysteresis loop are known as *soft magnetic materials*. Cores of rotating electrical machines are made from soft magnetic materials. In some applications, the magnetization of the material is to be retained on a permanent basis. Materials with large area hysteresis loop are more suitable for this purpose. Materials with large area hysteresis loop are called *hard magnetic materials*. Permanent magnets, magnetic tapes and disks used in entertainment and computer industries are made from hard magnetic materials.

34.14.1 Soft Magnetic Materials

Magnetic materials, which are easily magnetized and demagnetized are known as soft magnetic materials. They are characterized by thin hysteresis loop (Fig. 34.22). It is imperative that a soft magnetic material should have a high initial permeability and a low coercivity. In view of these properties, the material reaches its saturation magnetization with a relatively low applied field, and exhibits low hysteresis energy losses. This becomes an important consideration when the material is used in alternating current applications, since the area of hysteresis loop represents the energy lost as heat during a cycle. The smaller the area, the lower are the power losses and the greater the possibility of using the material at higher frequencies.

A low value of coercivity corresponds to the easy movement of domain walls as the applied magnetic field changes in magnitude and direction. Structural defects impede the domain wall motion and increase coercivity. Therefore, a soft magnetic material must be free of such structural defects.

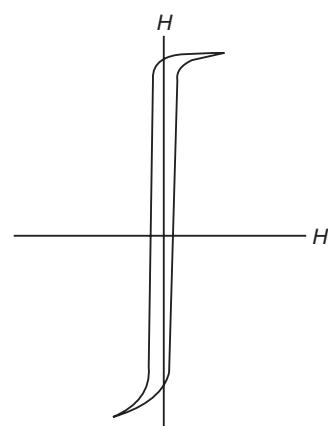


Fig. 34.22: Hysteresis loop for soft magnetic material

When the magnetic flux through a conducting material changes, voltages are induced which cause the flow of eddy currents. Eddy currents constitute the second major contributor to energy loss. It is, therefore, necessary to minimize eddy current losses in soft magnetic materials. Increasing the electrical resistivity of the material can reduce eddy current losses. The resistivity of ferromagnetic materials is very low in their pure form. Forming solid solutions increases their resistivity. For example, iron containing a few percent of silicon has higher resistivity instead of pure iron. Using thin laminations of ferromagnetic material, which are insulated from each other, can further increase the resistivity of the material. Further, eddy current losses increase with the square of the frequency. Therefore, the laminations are made extremely thin of the order of 1 to $0.025\text{ }\mu\text{m}$ thickness. In very high frequency applications, even laminated sheets cause excessive power losses. Therefore powdered materials are used. Very fine metallic or ceramic magnet powders of a 2 to $10\text{ }\mu\text{m}$ particle size are mixed with a binder that insulates the particles from one another. The material is fabricated into the required shape and size.

The soft magnetic material can be broadly divided into four groups:

- Heavy duty flux multipliers:** These are the cores of transformers, motors and generators. Electrical steel is used for manufacturing cores.
- Light duty flux multipliers:** These are cores of small special purpose transformers, inductors etc used in communication equipment. Ni-Fe alloys and soft ferrites are used in these applications.
- Square loop materials:** They are used in magnetic amplifiers, saturable core devices, computers etc. For these applications also, nickel-iron alloys and soft ferrites are used.
- Microwave system components:** Soft ferrites and garnets are used in these applications.

34.14.2 Hard Magnetic Materials

Hard magnetic materials are those, which have a high resistance to demagnetization. A high remanence, high permeability, a high coercive field and a large hysteresis loop characterize the hard magnetic materials (Fig. 34.23). A desirable requirement is immunity of the material to loss of magnetization by ac fields, mechanical illtreatment and changes in temperature.

The hard magnetic materials are magnetized in a magnetic field strong enough to orient the magnetic moments of their domains in the direction of the applied field. Part of the energy of the applied field is converted into potential energy, which is stored in the permanent magnet produced. A permanent magnet in the fully magnetized condition is thus in a relatively high-energy state as compared to a demagnetized state. The power or external energy of a hard magnetic material is directly related to the size of its hysteresis loop. The magnetic potential energy of a hard material is measured by its maximum energy product $(BH)_{\max}$.

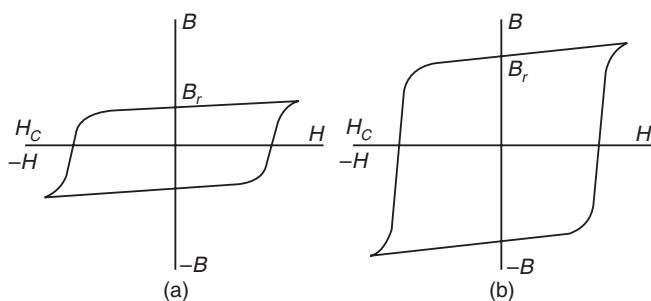


Fig. 34.23: Hysteresis loops for hard magnetic materials

A permanent magnet in the fully magnetized condition is thus in a relatively high-energy state as compared to a demagnetized state. The power or external energy of a hard magnetic material is directly related to the size of its hysteresis loop. The magnetic potential energy of a hard material is measured by its maximum energy product $(BH)_{\max}$.

Energy product

The second and fourth quadrants of the B-H curve represent the magnetizing curves of the material, and are related to the energy required to demagnetize the magnet. It is customary to choose the second quadrant as the demagnetizing curve. The energy product versus the induction B curve is plotted either separately or along with the demagnetizing curve. Fig. 34.24 shows the demagnetizing curve and the energy product (BH) curve for a hypothetical hard material.

It is seen that the energy product is zero at point H_c and at point B_r and it attains a maximum at an intermediate point, known as the maximum energy product $[BH]_{\max}$. Basically, the maximum energy product of hard material is the area occupied by the largest rectangle that can be inscribed in the demagnetizing curve.

The hysteresis behavior is related to the ease with which the magnetic domain boundaries move. Through changes caused in the microstructure of the material, domain wall motion can be hindered and the energy product value may be enhanced. Usually, the alloys used in making permanent magnets are subjected to precipitation hardening in such a way that microscopically heterogeneous structures composed of acicular particles oriented in a particular direction are produced. The size of the particles is of the order of $1\mu\text{m}$. Each particle constitutes a magnetically saturated domain. The reversal of magnetization in this case can take place only by a complete rotation of the domain. It requires a very high field. In view of each particle being a domain, the displacement of domain walls is suppressed and consequently the coercive force becomes large.

Example 34.6. The area of a hysteresis loop drawn between B and H is 100 m^2 . Each unit space along the vertical axis represent 0.01 wb/m^2 and each unit space along the horizontal axis represents 40 A/m . Determine the hysteresis loss per cycle.

$$\begin{aligned}\text{Solution: } \text{The hysteresis loss per cycle} &= \text{Area of the hysteresis loop} \times \text{Value of unit length along B-axis} \times \text{Value of unit length along H-axis} \\ &= 100 \text{ m}^2 \times 0.01 \text{ wb/m}^2 \times 40 \text{ A/m} = 40 \text{ J/m}^3.\end{aligned}$$

34.15 MAGNETIC MATERIALS AND THEIR APPLICATIONS

34.15.1 Soft Magnetic Materials

Soft magnetic materials are used in a wide variety of electrical machines in daily use, such as power transformers, output transformers, motors, generators, electromagnets etc. Electrical steels are used as core materials of these machines. Hard magnetic materials are used in fabrication of permanent magnets, which are required to retain their magnetic field indefinitely.

34.15.1.1 Low Carbon Steel

Pure iron has a higher permeability but its higher electrical conductivity causes more eddy current losses. Low carbon steel (Fe-0.05% C) has a relatively small permeability and a higher resistivity. It is the least expensive core material and is used where low cost is more important than other factors.

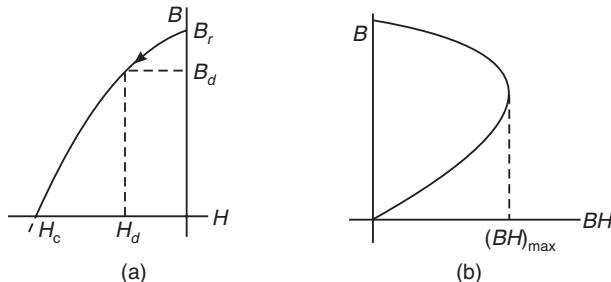


Fig. 34.24

34.15.1.2 Iron-silicon alloys

The addition of about 3-4% silicon to iron produces iron-silicon alloys with improved characteristics. Silicon increases the electrical resistivity of low carbon steel and thus reduces the eddy current losses. It increases the magnetic permeability and lowers hysteresis losses. It decreases magnetostriction and therefore reduces transformer noise known as hum. However, silicon addition tends to decrease the ductility of iron and makes the material brittle. The maximum limit for addition of silicon is only about 4 wt%.

Grain orientation

The permeability can be substantially increased and hysteresis losses can be decreased by making use of favourable grain orientation in the material. In case of iron crystal, $<100>$ direction is the easy direction and the spin moments in a virgin crystal are aligned along $<100>$ directions. Further, when sheets of iron or iron-alloys are manufactured by rolling and annealing, the $<100>$ direction is parallel to the rolling direction. Therefore, cold rolled grain oriented (CRGO) steels possess better magnetic properties in a direction parallel to the direction of rolling. Consequently, the cores require less material.

Often, aluminium and manganese of less than 1% are added in iron-silicon alloys to improve the grain orientation and reduce the hysteresis losses.

34.15.1.3 Nickel-iron alloys

A pure nickel-iron alloy with nickel content of about 25% is practically non-magnetic. With the increase in nickel content, a wide range of magnetic properties is obtained. The magnetic permeability of iron-silicon alloys is relatively low at low fields. Low initial permeability is not important for power applications such as transformer cores. For communications applications much higher permeabilities at low fields are required. Nickel-iron alloys are used for these applications. These alloys can be broadly divided into three groups: 36% nickel, 50% nickel and 77% nickel.

36% nickel alloys have high resistivity and low permeabilities. They are used for high frequency devices such as high-speed relays, wideband transformers and inductors. They are comparatively cheaper.

The 50% nickel alloy have moderate permeability ($\mu_{\max} = 25,000$) and high saturation induction ($B_{\text{sat}} = 1.6\text{T}$). They are used where low loss and small size are required, such as in relays, small motors and synchros, etc.

The 79% nickel alloy has high permeability ($\mu_{\max} = 10^6$) but lower saturation induction ($B_{\text{sat}} = 0.8\text{ T}$). They are used in recording heads, pulse transformers, sensitive relays etc.

34.15.1.4 Mumetal

The highest permeabilities of the order of 10^5 to 10^6 are found in multicomponent nickel-iron alloys, such as Permalloy, Supermalloy etc. Mumetal having a composition of 77% nickel, 16% iron, 5% copper and 2% chromium can be rolled into thin sheets and is used to shield electronic equipment from stray magnetic fields.

34.15.1.5 Soft ferrites

A remarkable feature of ferrites is that they have a high electrical resistance, 10^5 to 10^{15} times the resistance of metallic ferromagnets. Ferrites possess the electrical properties of dielectrics combined with the magnetic properties of ferromagnetic materials. Therefore, they can be used for applications at high frequencies, without the eddy current losses.

Some of the most important uses of soft ferrites are for low signal memory core, audiovisual and recording head applications. At low signal levels, soft ferrite cores are used

for transformers and low energy inductors. A large use is for deflection yoke cores, flyback transformers and convergence coils for television receivers.

Mn-Zn and Ni-Zn ferrites are widely used for these applications. Mn-Zn ferrites are used for operation upto 500 kHz whereas Ni-Zn ferrites can be used for high frequency operation upto 100 MHz.

Materials having nearly square-shaped hysteresis loop are earlier used as memory or logic operation devices in computers, as switching devices, and in information storage. Magnesium-manganese ferrite, manganese-copper ferrite and lithium-nickel ferrite are used for these applications. These ferrites can exist in two stable but different magnetic states capable of storing a ‘bit’ of information. Usually they are made in the form of tiny rings called ‘cores’ which are assembled using external connections into a large matrix of wires containing cores at each junction. Microwave devices in the frequency range 1 GHz to 100 GHz rely for their operation on the Faraday rotation. The Faraday effect consists in rotation of the plane of polarization of a plane-polarized electromagnetic wave as it travels through a medium in the direction of an applied magnetic field.

34.15.2 Magnetic Storage Materials

Magnetic materials are widely used for storage of information. The entertainment electronics and computer industries heavily rely on magnetic tapes for the storage and reproduction of audio, video and digital sequences.

The magnetic storage medium consists of magnetic oxide particles distributed in a binder, which strongly bonds to a plastic film. The oxide particles are needle shaped and each particle is single domain magnetized to saturation along its major axis. The most commonly used magnetic particles are $\gamma\text{-Fe}_2\text{O}_3$ crystallites. Videotapes use cobalt-doped $\gamma\text{-Fe}_2\text{O}_3$, which has a higher coercivity.

34.15.3 Hard Magnetic Materials

Hard magnetic materials have a high resistance to demagnetization. A high remanence, high permeability, a high coercive field and a large hysteresis loop characterize the hard magnetic materials.

34.15.3.1 Alnico alloys

Permanent magnets are made mainly using alnico alloys. Alnico alloys contain various amounts of aluminium, nickel, cobalt and iron along with some minor constituents such as copper and titanium. Alnico alloys are mechanically hard and brittle. Their magnetic properties are highly stable against variations in temperature.

34.15.3.2 Hard Ferrites

Hard ferrites are used in making permanent magnets. The most important hard ferrites are the barium ferrite (BaFe_2O_3) and strontium ferrite (SrFe_2O_3). They are widely used in generators, relays, loud speakers, telephone ringers, toys etc. Hard ferrite powders are often mixed with plastic materials to form flexible magnets for door closer and other holding devices.

34.16 MAGNETIC DEVICES

1. Transformer Cores

Transformers are comprised of a number of parts, each working in conjunction with the others to ensure the safe and effective transmission of energy. The core makes up the bulk of a transformer. A **magnetic core** is a piece of magnetic material with a high permeability used to confine and guide magnetic fields in electrical devices. The composition of a transformer core

depends on such factors as voltage, current, and frequency. It is usually made of ferromagnetic metal such as iron, or ferrimagnetic compounds such as ferrites. The high permeability, relative to the surrounding air, causes the magnetic field lines to be concentrated in the core material. The magnetic field is often created by a coil of wire around the core that carries a current. The presence of the core can increase the magnetic field of a coil by a factor of several thousand over what it would be without the core. A range of cores exist, such as steel laminated, solid, toroidal, pot and planar cores, as well as variations of each within their respective categories.

Steel Laminated Cores: Steel laminated cores have high level of permeability and are used for transmitting voltage at the audio frequency level. Unlaminated steel cores have a high level of eddy current loss, and result in core heating. The presence of several steel laminations, protected by a non-conducting insulator material between layers, contain the eddy currents and reduce magnetizing effects. Although thin laminations are harder to manufacture and are more expensive, they are effective in high frequency transformers.

Several designs are available for steel laminated transformers, each offering its own advantages. An E shaped core is affordable to manufacture, but tends to exhibit more energy loss. A C type core, on the other hand, offers reduced resistance because the metal grains run parallel to the energy flux.

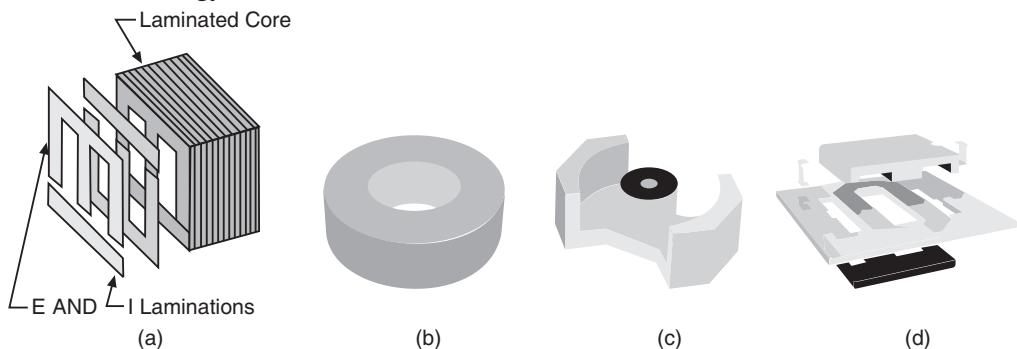


Fig. 34.25. Different types of transformer cores- (a) Steel laminated core (b) toroidal core (c) pot core (d) planar core

Solid Cores: Solid cores, particularly the powdered iron cores used in circuits, have high magnetic permeability as well as electrical resistance. When used in circuits, they tend to work best for transmission levels above main frequencies. For frequencies that tend to range even higher, such as those beyond the VHF (very high frequency) band, powdered iron is replaced by ferrites.

Toroidal Cores: A range of materials are available for use in toroidal cores, including steel, coiled permalloys, powdered iron, or ferrites. These cores can be circular in structure, with the rest of the transformer built around the core ring—the lack of an opening in the core ring means no air gaps—or they can be a long strip of material. The advantage of using a strip is reduced resistance as a result of properly aligned grain boundaries. With a circular core, the windings are generally wound around the core, covering the surface in its entirety.

Toroidal cores are more efficient at handling the same kind of energy load than steel laminated E shape cores, and can be made smaller, lighter, and with a lower magnetic field. However, windings tend to be more expensive for toroidal cores.

Pot core : The shape of a pot core is round with an internal hollow that almost completely encloses the coil. Usually a pot core is made in two halves which fit together around a coil former. This design of core has a shielding effect which reduces electromagnetic interference.

Planar core: A planar core consists of two flat pieces of magnetic material, one above and one below the coil. It is typically used with a flat coil that is part of a printed circuit board. This design is excellent for mass production and allows a high power, small volume transformer to be constructed for low cost. It is not as ideal as either a pot core or toroidal core but costs less to produce.

2. Magnetic Storage

Data storage devices are the essential requirements of entertainment electronics and computer systems. There are two types of storage devices available, namely semiconductor memories in which flip-flops are the storage elements and magnetic memories in which the magnetic domains are the storage elements. The semiconductor memories are *volatile* whereas the magnetic memories are permanent and *non-volatile*. The most common form of storage technology is magnetic storage. The data and programs required for computer operations are stored on magnetic bulk storage devices like tapes and disks. Magnetic storage and magnetic recording means storage of information on a magnetized medium. Magnetic storage was first suggested by Oberlin Smith in 1888. The first working magnetic recorder was invented by Valdemar Poulsen in 1898 and in 1928, Fritz Pfleumer developed the first magnetic tape recorder.

In the field of computing, the term *magnetic storage* is preferred and in the field of audio and video production, the term *magnetic recording* is more commonly used. The distinction is less technical and more a matter of preference. The storage media is typically called a **disk** or a **cartridge**. The process of storing data in a memory unit is called **writing** and the process of retrieving data from memory is called **reading**. The medium used in magnetic-storage devices is coated with **iron oxide**. This oxide is a **ferromagnetic** material. In most cases, magnetic storage uses a drive, which is a mechanical device that connects to the computer. The media, which is the part that actually stores the information, is inserted into the drive. For example, 1.44-MB floppy-disk drives use 3.5-inch diskettes. The drive uses a motor to rotate the media at a high speed, and it accesses (reads) the stored information using small devices called **heads**.

Each head has a tiny electromagnet, which consists of an iron core wrapped with wire. The electromagnet applies a **magnetic flux** to the oxide on the media, and the oxide permanently “remembers” the flux it sees. During writing, the data signal is sent through the coil of wire to create a magnetic field in the core (Fig. 34.26). At the gap, the magnetic flux forms a fringe pattern. This pattern bridges the gap, and the flux magnetizes the oxide on the media. When the data is read by the drive, the **read head** pulls a varying magnetic field across the gap, creating a varying magnetic field in the core and therefore a signal in the coil. This signal is then sent to the computer as binary data.

Hard disks: Magnetic disk memories provide large storage capabilities with moderate operating speed. A magnetic disk is a flat, circular plate called **platter** which has a surface

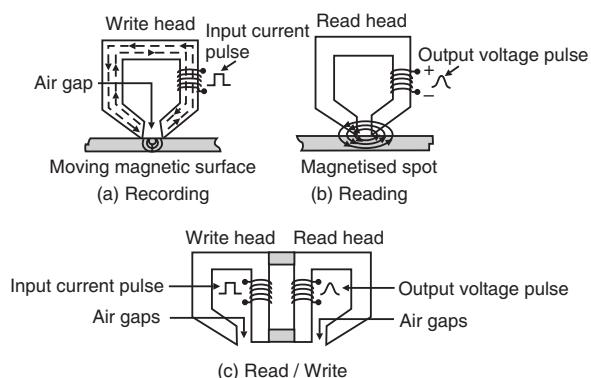


Fig. 34.26

that is coated with magnetic iron oxide particles. It also has a read-write head that hovers over the surface to read data. A *hard disk* is one or more platters and their associated read-write heads. The platters rotate at very high speed of the order of 3600 rpm. The information is stored on their surface by magnetic heads mounted on access arms. Information is recorded in the form of bands.

Each band of information on a particular disk is called a **track**. The tracks are divided into **sectors** (see Fig. 34.27). Tracks and sectors are created when the hard disc is first **formatted** and this must take place before the disc can be used. The tracks are arranged in **concentric rings** so the software can jump from “file 1” to “file 19” without having to fast forward through files 2 through 18. There will be several thousand data tracks on one side of a disk in which the bits are recorded in a track at a density of the order of 20,000 to 1,00,000 bits/inch. A typical outer track contains more bits than the inner tracks since the circumference of an outer track is greater. The disk or cartridge spins like a record and the heads move to the correct track, providing what is known as **direct-access storage**.

The total time taken by the head to begin reading or to begin writing on a selected track is called **access time**. The time taken to position a head on the selected track is called **seek time** which is of the order of milliseconds. The time required for the desired data to reach the magnetic head after the positioning of the head is called **rotational delay**. The time requirement will be of the order of milliseconds. The total access time for a disk is the sum of the seek time and rotational delay. The number of bits transferred per second once reading or writing begins is called transfer rate of disk. A schematic diagram of a flying head is shown in Fig. 34.28.

When the magnetic disk rotates, a thin but resilient layer of air rotates along with the disk. The shape of the head is designed in such a way that it rides on this layer of air causing the disk to maintain separation from the head. It floats on a cushion of air a fraction of a millimetre above the surface of the disk. The drive is inside a sealed unit because even a speck of dust could cause the heads to crash.

The storage capacity in the modern disk memories is increased by mounting several disks on a common drive unit known as disk pack as shown in Fig. 34.29. During the operation of the memory, the disks are rotated at a uniform speed by a disk drive

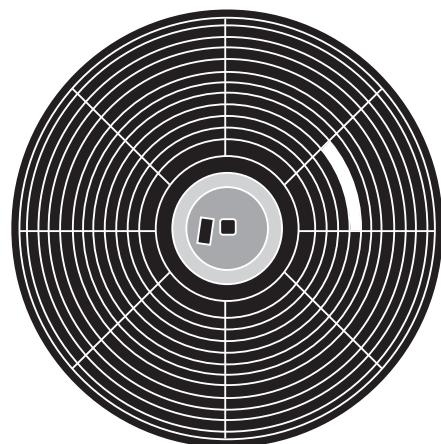


Fig. 34.27

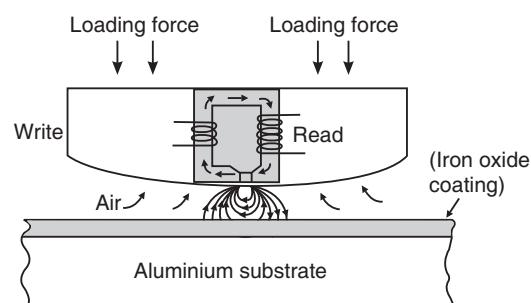


Fig. 34.28

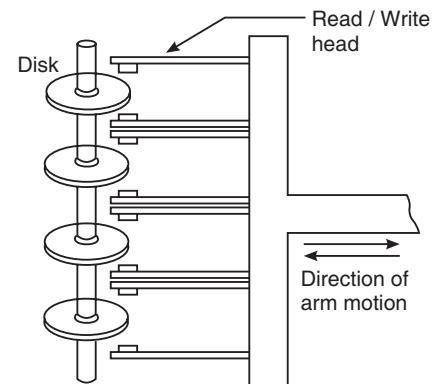


Fig. 34.29: Hard disk

unit. Each recording surface consists of one read-write head. Using the movable arm connected with the disks, a particular set of track for reading/writing can be selected. The magnetic disk memories are relatively inexpensive, have low access time and high disk transfer rate. Hard disks of size 80 to 250 GB have become common in latest personal computers.

Advantages:

Very fast access to data. Data can be read directly from any part of the hard disc (**random access**). The access speed is about **1000 KB per second**.

Floppy disks

The floppy disk is a small flexible direct access magnetic storage device. It is made of very thin and flexible mylar (plastic) material. The surface of the disk is coated with a thin magnetic film. It is permanently housed in a square jacket for protective purpose as illustrated in

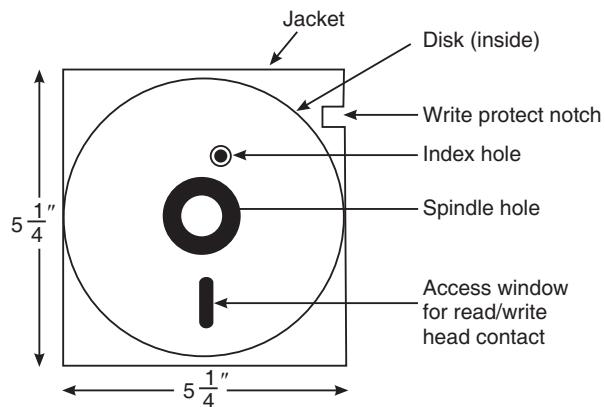


Fig. 34.30: Floppy disk

Fig. 34.30. The index hole establishes a reference point for all the tracks on the disc. The outer circle shows a hole in the jacket while the inner circle shows the index hole in the disk. During data processing the disk rotates and when the two holes are aligned a beam of light shining on one side of the disc is sensed from the other side and is used for timing functions. The surface of the floppy disc is divided into a number of concentric circles called tracks and the information is recorded on the tracks. Tiny magnetic spots are used to record the logic states 1 or 0 in such a way that for 1 it is magnetized in one direction and for 0, it is magnetized in the opposite direction.

As the disk rotates at 360 rpm within the fixed jacket, the read/write head makes contact through the access window and moves to specified position along the length of the slot. The write protect notch is used to protect the stored information. A typical 5 1/4 inch floppy disk is organized into 77 tracks and is divided into 26 sectors. Thus, each track is divided into 26 equal-sized sectors. The 5 1/4 inch disk is available with double density and high density with storage capacity of 360 KB and 1.2 MB.

Floppy disks have become obsolete because of their low capacity range and so now most of the latest computers are being made without a floppy disk drive.

Advantages

They are very cheap to buy and floppy disc drives are very common.

Disadvantages

They are easily physically damaged if unprotected and magnetic fields can damage the data.

They are relatively slow to access because floppy discs rotate far more slowly than hard discs, at only six revolutions per second, and only start spinning when requested. The access speed is about 36 KB per second.

Magnetic Tapes

A tape is a type of backup storage device used to copy data on a hard disk. The tapes are generally thin flexible plastic tapes with a thin coating of hard magnetic material on one side. The magnetic material must have high remanent magnetization and least sensitivity for

self-demagnetization. As such materials with rectangular hysteresis loop are used in making the coating. Just like the tape in a tape-recorder, the data is written to or read from the tape as it passes the magnetic heads.

Data on a tape is arranged as a long sequence, beginning at one end of the tape and stretching to the other end. Therefore a tape is a sequential storage device whereas hard disks and floppies are random access storage devices. Access time on a tape is measured in seconds, unlike hard disk drives which are measured in milliseconds. Tape drives have good storage capacity, being able to hold as much information as a hard disk. Most tape drives can back up 1GB of data in 15-20 minutes. Tape storage is not suitable for daily tasks because it is too slow to be the computers main storage device.

Fig. 34.31 shows a high speed start-stop tape mechanism which uses a set of tension arms around which the tape is laced. The tension arms are movable and when the tape is suddenly driven past the heads by the capstan, the mechanism provides a buffering supply of tape. A servo mechanism is used to drive the upper and lower reels to maintain sufficient tape between capstan and tape reels.

Output signals from read heads are generally in the range of 0.1 to 0.5 V. the recording density varies from few hundred bits/inch to several thousand bits/inch. Data are recorded on magnetic tape by using some coding system. Usually one character is stored per row along the tape.

Advantages

Magnetic tape is relatively cheap and tape cassettes can store very large quantities of data (*typically 26 GB*).

Disadvantages

Accessing data is very slow and we cannot go directly to an item of data on the tape as we can do with a disc. It is necessary to start at the beginning of the tape and search for the data as the tape goes past the heads (serial access).

3. Magneto-optical Recording:

Compact discs are classified based on the storage techniques and capabilities as CD-ROM (Read only Memory), CD-WORM (Write Once Read Many), CD-R/W (Read/Write) etc. They use different materials but write and read the data using a laser beam. In case of CD(R/W), we can write the data, read and rewrite after erasure. Erasability implies that the recording media can undergo a very large number of write/erase operations without any loss in recording / reading quality. There are two main media designs for rewritable optical systems: MO (magneto-optical) media and phase-change media.

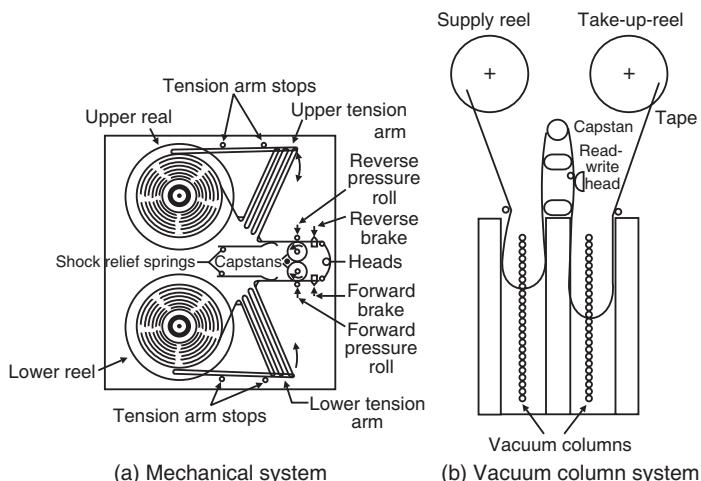


Fig. 34.31

MO systems write magnetically and read optically.

Principle

A laser beam is used for writing, reading and erasing of data on a MO medium. Even a relatively weak laser generates high local temperatures when focused at a small spot on the medium. All magnetic materials have a characteristic temperature, called the Curie temperature. For the magnetic materials used in MO systems, the Curie temperature is of the order of 200°C. Above the Curie temperature the magnetic material loses

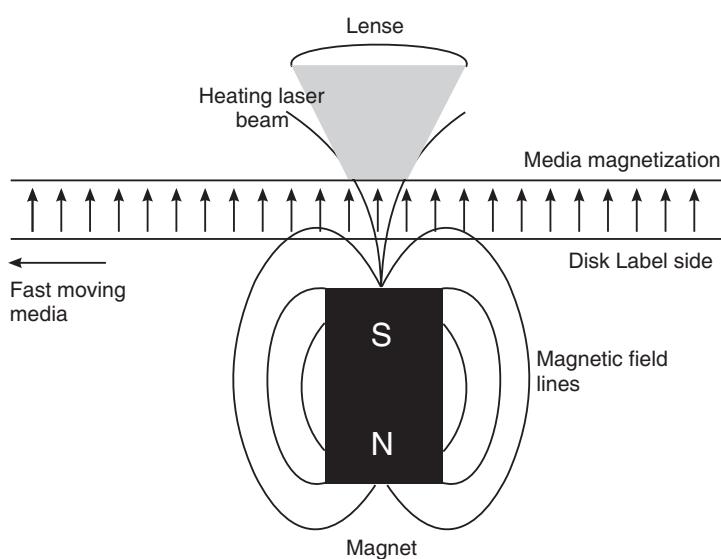


Fig. 34.32. Schematic of the magneto-optic recording process
magnetization due to a complete disordering of magnetic domains in it. Further, the coercivity of material drops at higher temperatures. Therefore, when the material is heated, its coercivity is low, and the magnetization of the media can be changed by applying a weak magnetic field from the magnet. When the material is cooled to room temperature, its coercivity rises back to such a high level that the magnetic data cannot be easily affected by the magnetic fields we encounter in our regular daily activity.

Writing

The basic schematic of the recording process is illustrated in Fig. 34.32. When the disk is inserted into the drive, the label side will face the magnet, and the transparent side will face the laser. The direction of magnetization in the thin magnetic film is perpendicular to the surface. That is, the atomic magnets are oriented in the upward direction perpendicular to the film. When focused laser pulse falls on the film, strong localized heating occurs. At the high temperature, domain reversal takes place due to the action of the downward magnetic field and the atomic magnets turn downward (Fig. 34.32). The magnetized areas cannot be seen in regular light, but can be seen only in polarized light.

Reading

MO systems use polarized light to read the data from the disk. The changes in light polarization occur due to the presence of a magnetic field on the surface of the disk. This is known as the Kerr effect. If a beam of polarized light is reflected from a magnetized surface, the polarization of the reflected beam will change slightly. If the magnetization is reversed, the change in polarization is reversed too. The change in direction of magnetization could be associated with numbers 0 or 1, making this technique useful for binary data storage.

MO Erasing

To erase the data, magnetic field is applied in the upward direction. The focused laser pulse causes local heating which assists the atomic magnets to turn in the upward direction.

Thus, data is written, read, erased and rewritten on the MO disc.

Design of MO disk

A thin film of amorphous Terbium iron cobalt (TbFeCo) magnetic film is coated on a substrate. It is the active medium. The active magneto-optical layer is in fact enclosed between thin dielectric layers. The layers act as ‘anti-reflection layers’ and increase light absorption by the active layer. On top of these layers, a thin aluminium layer is coated. The aluminum layer acts as a light reflector and a heat sink to minimize lateral heating of the active layer.

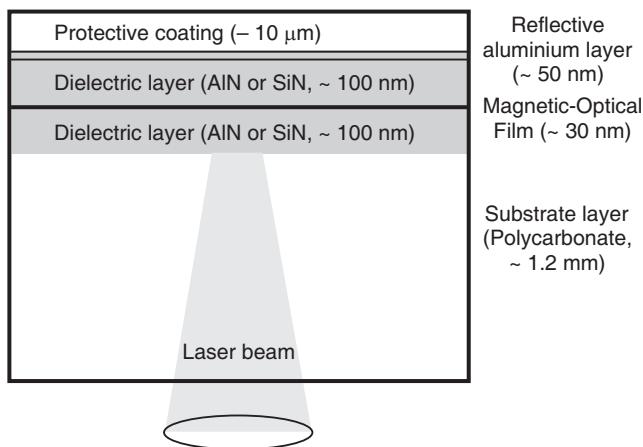


Fig. 34.33. Design of a quadrilayer magneto-optical disk

The materials for MO recording should meet the following major criteria:

- Have amorphous structure (smooth surface and domain’s boundaries to decrease system’s noise)
- Low thermal conductivity (to limit lateral heating to the recording layer itself)
- High melting point at about $200^\circ - 300^\circ\text{C}$ (media stability, accidental data loss prevention)
- Rapid drop of coercivity near the Curie temperature (sharp recording threshold)
- High coercivity at room temperature (media stability, accidental data loss prevention)
- Vertical anisotropy (perpendicular magnetic recording)
- Chemical stability (constant material’s properties under repeated heating-cooling)

QUESTIONS

1. Explain magnetic flux density, B , magnetic field strength, H , and magnetization, M . How are they related to each other?
2. Define magnetic susceptibility and permeability. Obtain the relation between them. **(M.G.Univ., 2006)**
3. Why diamagnetic materials have negative susceptibility?
4. What happens to paramagnetic susceptibility when the temperature is increased?
5. Explain how substances are classified according to their magnetic behaviour. **(M.G.Univ., 2005)**
6. Describe diamagnetic, paramagnetic and ferromagnetic materials. Explain their classification on the basis of permanent magnetic moment. **(C.S.V.T.U., 2005)**
7. How do you distinguish between diamagnetic, paramagnetic and ferromagnetic materials? **(M.G.Univ., 2006)**
8. What are ferromagnetic materials? Discuss the importance of hysteresis curve. How would you use the hysteresis curve for selecting the material for the construction of a permanent magnet? **(C.S.V.T.U., 2006)**
9. (i) Define magnetic susceptibility.

- (ii) Distinguish between paramagnetic, diamagnetic and ferromagnetic substances. Also discuss briefly the terms antiferromagnetism and ferrimagnetism on the basis of magnetic dipoles of the atoms. **(C.S.V.T.U., 2008)**
10. Explain the significance of Curie temperature for a ferromagnetic material.
 11. State the Curie law of paramagnetism. What is the Curie temperature?
 12. Explain important magnetic properties of ferromagnetic materials.
 13. Explain the terms diamagnetism, paramagnetism, ferromagnetism, antiferromagnetism and ferrimagnetism on the basis of magnetic dipoles of the atoms.
 14. What is spontaneous magnetization?
 15. What are domain?
 16. What is hysteresis loop? What does it represent? What is the significance?
 17. Explain ferromagnetic hysteresis on the basis of domains.
 18. What are the losses in ferromagnetic materials? How are they reduced?
 19. What is meant by hysteresis? Explain hysteresis loss. How would you use the hysteresis curves to select material for the construction of permanent magnets? **(UPTU, Lucknow)**
 20. Discuss the variation of spontaneous magnetization with temperature for ferromagnetic materials.
 21. Compare and contrast ferromagnetism with ferrimagnetism.
 22. What are ferrites? In what respect they are superior to ferromagnetic materials?
 23. What are ferrites? What is their importance?
 24. Why are ferrites preferred over ferromagnetic materials at high frequency as core material?
 25. What are soft and hard magnetic materials? Give their characteristic properties and applications.
 26. Si-Fe and Ni-Fe alloys have important engineering applications. Discuss in brief about these alloys and their applications.
 27. Show that $B = H + 4\pi I$ and give an expression for energy loss in a complete hysteresis cycle. **(M.G.Univ., 2005)**
 28. What are hard and soft magnetic materials? Indicate the properties sought in each case. Give their applications. **(C.S.V.T.U., 2009)**
 29. (a) What are hard and soft magnetic materials – distinguish.
(b) Write down the characteristics and requirements of microwave ferrites. **(Andhra Univ.)**
 30. Power transformers produce a loud hum when they operate. Why?
 31. Laminated iron sheet is commonly used for transformer cores. Why are not castings used?
 32. How can the permeability of a ferrite be improved?
 33. Explain the difference between the terms ‘Curie temperature’ and ‘Neel temperature’.
 34. A polycrystalline iron rod does not show any magnetization. Explain.
 35. Write a short note on materials used for magnetic tape.
 36. Distinguish between ferro-, ferri-, and antiferromagnetic materials. Out of these, which materials are used as permanent magnets? Why?
 37. A ferromagnetic material is placed in a magnetic field of intensity H . What is the energy stored in the material?
 38. What should be the characteristics of permanent magnetic materials?
 39. Discuss the applications of soft ferrites?
 40. What are hard ferrites?
 41. What are different types of compact discs? How are data stored, read and rewritten?

PROBLEMS

1. A magnetic material has a magnetization of 2300 A/m and produces a flux density of 31.4 gauss. Calculate magnetizing force and relative permeability of the material.
 2. A magnetic field of 1800 A/m produces a magnetic flux of 3×10^{-5} Wb in an iron bar of cross-sectional area 0.2 cm^2 . Determine the permeability of iron. [Ans: $8.3 \times 10^{-3} \text{ H/m}$]
 3. A magnetic material has a magnetization of 3000 A/m and a flux density of 0.005 wb/m^2 . Calculate the magnetic force and the relative permeability of the material. [Ans: 980.9 A/m , **4.06**]
 4. Find the relative permeability of a ferromagnetic material if field of strength 220 A/m produces a magnetization of 3300 A/m in it. [Ans: **16**]
 5. The molecular weight of a paramagnetic salt is 168.5 and its density is 4370 kg/m^3 at 27°C . Calculate its susceptibility and the magnetization produced in it in a field of $2 \times 10^5 \text{ A/m}$. Assume that the contribution to paramagnetism is two Bohr magnetons per molecule.
- [Ans: 5.4×10^{-4} , 108.5 A/m]**
6. Assume that iron atoms have magnetic moment of two Bohr magnetons. Calculate the Curie constant if its density is 7150 kg/m^3 and atomic weight is 55.84. [Ans: **0.209**]
 7. Assuming a bcc lattice for iron with lattice constant $a = 3\text{\AA}$, calculate the Curie constant and Weiss constant. Given that the Curie temperature of iron is 1050 K. [Ans: **0.193, 5440**]
 8. A paramagnetic system of electron spin magnetic dipole moment is placed in an external field of 10^5 A/m . Calculate the average magnetic moment per dipole at 300 K. [Ans: $2.6 \times 10^{-27} \text{ A.m}^2$]
 9. A paramagnetic salt contains 10^{28} ions/ m^3 with magnetic moment of one Bohr magneton. Calculate the paramagnetic susceptibility and the magnetization produced in a uniform magnetic field of 10^6 A/m when the temperature is 27°C .
 10. An electron in an atom of hydrogen revolves in an orbit of radius 0.51 \AA . Calculate the change in magnetic moment for this electron if a magnetic field of induction 2 T acts at right angles to the plane of the orbit.
[Ans: $3.66 \times 10^{-29} \text{ A.m}^2$]

CHAPTER

35

Superconductivity

35.1 INTRODUCTION

Superconductivity is a state of matter exhibited usually at very low temperatures where the resistivity of the material drops to zero. The superconducting state is a state in which quantum mechanics operates on a macroscopic scale of the order of many atomic distances rather than the usual atomic and subatomic scale. The superconducting state is influenced by temperature, current and magnetic field. There exist critical values for these three parameters, above which the material passes into normal state. Besides being of immense theoretical interest, the phenomenon has many possible practical applications. Superconducting magnets are designed for use in the fields of medicine and particle physics. It is being vigorously tried for use in transportation and transmission of power. Until recent times the main hurdle in the way of their use is the requirement of extremely low temperatures of the order of 20 K and less. Recent discoveries of high temperature ceramic superconductors raised the hopes of using superconductors in making more efficient and smaller electrical and electronic devices at normal temperatures. Extensive research is being carried out all over the world to improve the properties of ceramic superconductors to make them suitable for various practical applications.

35.2 SUPERCONDUCTIVITY

Phenomenon

Metals are good conductors of electricity as they contain a tremendous number of free electrons. The low resistance offered by them to the flow of current is attributed to the scattering of the free electrons by vibrating ions of the lattice. When temperature increases, the amplitude of the lattice vibrations increase and cause more scattering of electrons leading to more resistance. Even at 0 K, metals offer finite resistance, called *residual resistance*, which is attributed to the scattering of electrons by impurities and crystal defects present in the material. The variation of resistivity of normal metals with temperature is shown in Fig. 35.1, which indicates the existence of residual resistance, ρ_r . H.K.Onnes was verifying the behaviour of metals at very low temperatures and in 1911, he discovered that the electrical resistance of highly purified mercury dropped abruptly to zero at a temperature of 4.15 K (see Fig. 35.2). The sudden drop in resistivity was quite unexpected and Onnes recognized it to be an entirely new phenomenon. Onnes also found that the transition was reversible. When heated above the transition temperature 4.15 K, mercury regained its resistivity. Onnes named the phenomenon as *superconductivity*. Subsequently, superconductivity was discovered in lead, zinc, aluminium, and other metals as well as in a number of alloys. Superconductivity

was strictly a low temperature phenomenon till 1980's, when certain ceramic materials were found to exhibit superconductivity at higher temperatures of about 120 K.

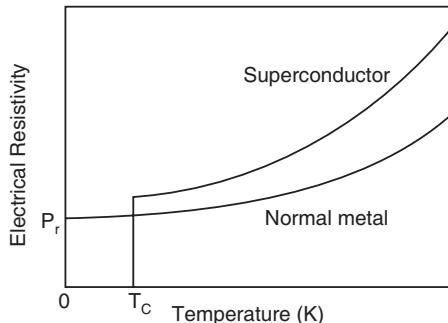


Fig. 35.1

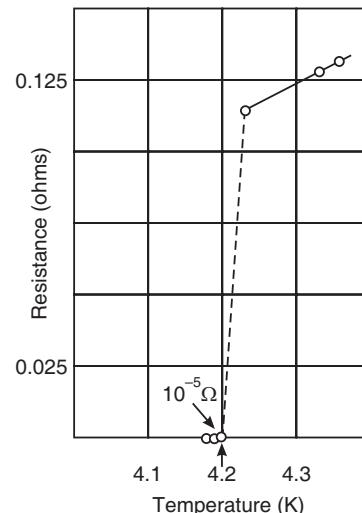


Fig. 35.2

Superconductors

Superconductivity is the phenomenon in which electrical resistance of materials suddenly disappears below a certain temperature. The materials that exhibit superconductivity and which are in the superconducting state are called **superconductors**.

Transition temperature

The temperature at which a normal material abruptly changes into a superconductor is called transition temperature, T_C . It is also known as the critical temperature.

35.3 MATERIALS (LOW T_c MATERIALS)

Since the discovery of superconductivity in the metal mercury, many metals have been found to exhibit superconductivity at liquid helium temperatures. About 40 elements from the

Fig. 35.3

periodic table are found to exhibit superconductivity (Fig. 35.3). Apart from these, thousands of alloys display superconductivity.

Good conductors of electric current such as silver, copper and gold are not superconductors. In fact, the superconductors are relatively poor conductors at room temperatures. Non-transition metals such as Nb, Mo and Zn exhibit superconductivity. Semiconductors like Si, Ge, Se and Te transform to a metallic phase when subjected to high pressure and then become superconducting at low temperatures. The ferromagnetic materials like Fe, Co, Ni do not show superconductivity. Table-1 lists some of the superconducting materials and their transition temperatures T_C .

Table 1: Transition Temperatures and Critical Field of Some Superconducting Materials

Material		Transition Temperature (K)	Critical Field at 0K (Wb/m ²)
Type I. Superconductors			
Tungsten	(W)	0.015	0.0001
Titanium	(Ti)	0.390	0.0100
Cadmium	(Cd)	0.560	0.0030
Zinc	(Zn)	0.850	0.0054
Molybdenum	(Mo)	0.920	0.0095
Indium	(In)	3.408	0.0281
Tin	(Sn)	3.722	0.0305
Mercury	(Hg)	4.153	0.0411
Vanadium	(V)	5.380	0.1420
Lead	(Pb)	7.193	0.0803
Niobium	(Nb)	9.460	0.1980
Type II. Superconductors			
Nb Ti		10.0	15.7
Nb NN		15.7	1.5
B ₃ Al		18.7	32.4
Nb ₃ Ge		23.2	38.4
V ₃ G		14.8	2.1
V ₃ Si		16.9	2.35
La – Ba – Cu – O		40	–
Y Ba ₂ Cu ₃ O ₇		92	–
Bi ₂ Sr ₂ Ca ₂ Cu ₃ O ₁₀ + δ		105	–
Th ₂ Ca Ba ₂ Cu ₂ O ₁₀ + δ		125	–

The following points may be noted from Table-1.

- Among the pure metals, niobium has the highest transition temperature ($T_C = 9.46$ K) and tungsten the lowest ($T_C = 0.015$ K).
- Metal alloys have higher transition temperatures. Nb₃Ge alloy has the highest transition temperature of 23.2 K.
- The superconducting binary compounds and alloys may be formed from elements where both of them are superconducting or one of them is superconducting or neither of them is superconducting. For example, CuS is a superconductor where neither copper nor sulphur is a superconductor.

- Crystalline structure is not a necessary condition for super conductivity. Amorphous and polycrystalline samples also display superconductivity.

35.4 PROPERTIES OF SUPERCONDUCTORS

Superconductors exhibit many unusual and interesting properties.

35.4.1 Zero Electrical Resistance

A super conductor is characterized by zero electrical resistivity. It is not fundamentally possible to test experimentally whether the resistance is zero. A method devised by Onnes consists of measuring the decrease of the current in a closed ring of superconducting wire. The superconducting ring is kept in a magnetic field and it is cooled to below the critical temperature so that it goes into the superconducting state. When the external magnetic field is switched off, a current is induced in the ring. If the ring had a finite resistance, R , the current circulating in the ring would decrease according to the equation

$$I(t) = I(0)e^{-Rt/L} \quad (35.1)$$

where L is the inductance of the ring. The decay current is monitored by a change in the magnetic flux through a test coil held close to the superconducting ring. Any change in the magnetic flux of the superconducting ring will induce an emf in the test coil. Careful measurements established that the resistivity of superconductors could be taken as zero.

35.4.2 Persistent Current

Once a current is started in a closed loop of superconducting material, it will continue to keep flowing, of its own accord, around the loop as long as the loop is held below the critical temperature (see Fig. 35.4 b). Such a steady current, which flows without diminishing in strength, is called a **persistent current**. The persistent current does not need external power to maintain it because there do not exist I^2R losses. Calculations show that once the current flow is initiated, it persists for more than 10^5 years. Persistent current is one of the most important properties of a superconductor. Superconductor coils with persistent current flowing through them produce magnetic fields and can therefore act as magnets. Such a superconducting magnet does not require power supply to maintain its magnetic field.

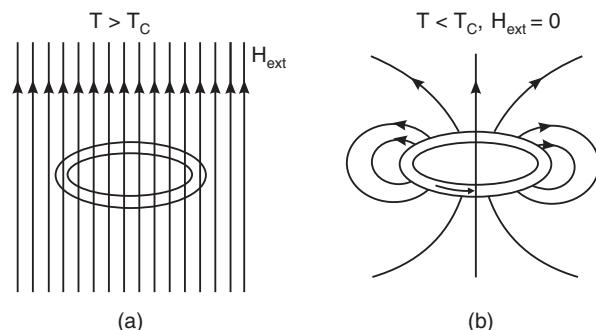


Fig. 35.4

35.4.3 Critical Temperature

When a superconducting material is cooled below a certain temperature, it goes into the superconducting state from normal state. The temperature at which a material in normal state goes into superconducting state is known as the **critical temperature**, T_C (see Fig. 35.1). Different materials have different critical temperatures. The transition is reversible. When the temperature of the material is increased above the critical temperature, it passes into the normal state. The transition is a thermodynamic phase transition. Just as the order in the arrangement of atoms increases at the transition of a material from liquid to solid state, a rearrangement of conduction electrons takes place leading to an increase in the order at the transition from normal to superconducting state. The superconducting transition is sharp for

a chemically pure and structurally perfect specimen while the transition range is broad for specimens which are structurally imperfect or which contain impurities.

35.4.4 Critical Magnetic Field

Superconducting state depends on the strength of the magnetic field in which the material is placed. Superconductivity vanishes if a sufficiently strong magnetic field is applied. The minimum magnetic field, which is necessary to regain the normal resistivity, is called the **critical magnetic field**, H_C . When the applied magnetic field exceeds the critical value H_C , the superconducting state is destroyed and the material goes into normal state.

The value of H_C varies with temperature. Fig. 35.5 shows the dependence of H_C on temperature in a typical superconductor. At temperatures below T_C , in the absence of magnetic field, the material is in superconducting state. When a magnetic field is applied and as its strength reaches the critical value H_C , the superconductivity in the material disappears. At any temperature $T < T_C$ the material remains superconducting until a corresponding critical magnetic field is applied. When the magnetic field exceeds the critical value, the material goes into normal state. The critical field required to destroy the superconducting state decreases progressively with increasing temperature. The dependence of critical field on temperature is governed by the following relation.

$$H_C(T) = H_C(0) \left[1 - \left(\frac{T}{T_C} \right)^2 \right] \quad (35.2)$$

where $H_C(0)$ is the critical field at 0 K.

Example 35.1. The transition temperature for Pb is 7.2 K. However, at 5 K it loses the superconducting property if subjected to magnetic field of 3.3×10^4 A/m. Find the maximum value of H which will allow the metal to retain its superconductivity at 0 K.

$$\text{Solution. } H_C(0) = \frac{H_C(T)}{1 - (T^2 / T_C^2)} = \frac{3.3 \times 10^4 \text{ A/m}}{1 - (25 / 51.28)} = 6.37 \times 10^4 \text{ A/m.}$$

Example 35.2. The critical field of niobium is 1×10^5 A/m at 8 K and 2×10^5 A/m at 0 K. Calculate the transition temperature of the element.

$$\text{Solution. } T_C = \frac{T}{\left[1 - \frac{H_C(T)}{H_C(0)} \right]^{\frac{1}{2}}} = \frac{8 \text{ K}}{\left[1 - \frac{1 \times 10^5 \text{ A/m}}{2 \times 10^5 \text{ A/m}} \right]^{\frac{1}{2}}} = 11.3 \text{ K}$$

Example 35.3. The transition temperature for lead is 7.26 K. The maximum critical field for the material is 8×10^5 A/m. Lead has to be used as a superconductor subjected to a magnetic field of 4×10^4 A/m. What precaution will have to be taken?

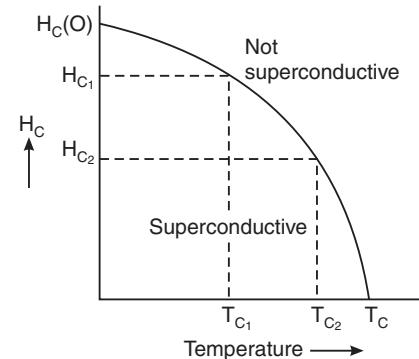


Fig. 35.5 : Schematic representation of the critical magnetic field as a function of temperature.

Solution. $T = T_C \left[1 - \frac{H_C(T)}{H_C(0)} \right]^{\frac{1}{2}} = 7.26K \left[1 - \frac{4 \times 10^4 \text{ A/m}}{8 \times 10^5 \text{ A/m}} \right]^{\frac{1}{2}} = 7.08 \text{ K.}$

Therefore, the temperature of the metal should be held below 7.08 K.

35.4.5 Critical Current Density

The critical magnetic field required to destroy superconductivity need not be necessarily applied externally. An electric current flowing through the superconducting material itself may produce magnetic field of requisite strength. Thus, if a superconducting ring carries a current I , it gives rise to its own magnetic field. As the current increases to a critical value, I_C , the associated magnetic field increases to H_C and the superconductivity disappears. The maximum current density at which the superconductivity disappears is called the **critical current density**, J_C . For any value of $J < J_C$, the current can sustain itself whereas for values $J > J_C$, the current cannot sustain itself. This effect was observed in 1916 by Silsbee and is known as *Silsbee effect*.

A superconducting ring of radius R ceases to be a superconductor when the current is

$$I_C = 2\pi R H_C \quad (35.3)$$

Thus, the existence of a critical current sets a definite limit to the size of the current that can flow through a superconducting coil without disturbing its superconducting state. The maximum current that a superconductor can carry decreases as the temperature is raised and falls to zero at the transition temperature of the material. Since the critical current falls with temperature, the critical magnetic field will also decrease as the transition temperature is approached. The variation of critical current density J_C and critical magnetic field H_C with temperature is shown in Fig. 35.6.

Example 35.4. The critical magnetic field at 5 K is $2 \times 10^3 \text{ A/m}$ in a superconductor ring of radius 0.02 m. Find the value of critical current.

Solution. $I_C = 2\pi R H_C = 2 \times 3.143 \times 0.02 \text{ m} \times 2 \times 10^3 \text{ A/m} = 251.4 \text{ A}$

Example 35.5. Calculate the critical current for a wire of lead having a diameter of 1 mm at 4.2 K. The critical temperature for lead is 7.18 K and $H_C(0) = 6.5 \times 10^4 \text{ A/m}$.

Solution. $H_C(T) = H_C(0) \left[1 - \left(\frac{T}{T_C} \right)^2 \right] = 6.5 \times 10^4 \text{ A/m} \left[1 - \left(\frac{4.2 \text{ K}}{7.18 \text{ K}} \right)^2 \right] = 4.28 \times 10^4 \text{ A/m}$

The critical current $I_C = 2\pi r H_C = \pi d H_C = 1 \times 3.14 \times 10^{-3} \text{ m} \times 4.28 \times 10^4 \text{ A/m} = 134.5 \text{ A.}$

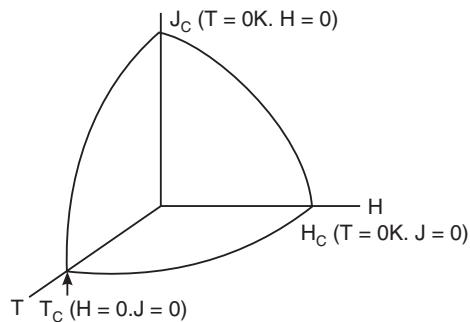


Fig. 35.6

35.4.6 Perfect Diamagnetism – Meissner Effect

In 1933 W.Hans Meissner and Robert Ochsenfeld found that when superconductors are cooled below their critical temperature in the presence of a magnetic field, the magnetic flux is expelled from the interior of the specimen and the superconductor becomes a perfect diamagnetic. This phenomenon is known as **Meissner effect** (see Fig. 35.7). Meissner and

Ochsenfeld found that as the temperature of the specimen is lowered to T_C , the magnetic flux is suddenly and completely expelled from it. The flux expulsion continues for $T < T_C$. The effect is reversible. When the temperature is raised from below T_C , the flux suddenly penetrates the specimen at $T = T_C$ and the material returns to the normal state.

The magnetic induction inside the specimen is given by

$$B = \mu_0(H + M) = \mu_0(1 + \chi)H$$

where H is the magnetic field applied externally and M is the magnetization produced within the specimen.

At $T < T_C$

$$B = 0 \text{ and therefore } \mu_0(H + M) = 0$$

It follows that

$$M = -H$$

The susceptibility of the material is

$$\chi = \frac{M}{H} = -1 \quad (35.4)$$

The specimen is therefore diamagnetic and the state in which magnetization cancels the external magnetic field completely is referred to as *perfect diamagnetism*.

The Meissner effect contradicts the fundamental principles of electromagnetism. The condition of perfect diamagnetism cannot be explained from the simple definition that superconductivity is a state of zero resistivity. Meissner effect shows that

in the superconductor not only $\frac{dB}{dt} = 0$

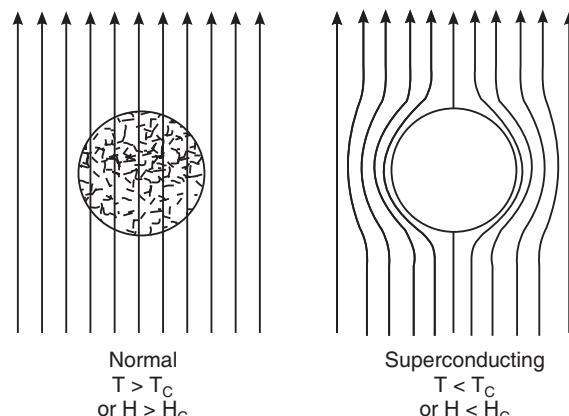


Fig. 35.7

but also $B = 0$. Thus, two mutually independent properties, namely zero resistivity and perfect diamagnetism are the essential properties that characterize the superconducting state.

Applications of Meissner effect

- The Meissner effect is the standard test used to conclusively prove whether a particular material is a superconductor or not.
- A material in superconducting state is a perfect dia-magnet and hence strongly repels external magnets. A smaller magnet repelled by a bigger superconductor hovers in air. This is known as **levitation effect**. In a similar way, a small chip of superconducting material hangs on to a bigger magnet and this effect is known as **suspension effect**. The levitation effect is utilized in the operation of Maglev trains.

35.4.7 London Penetration Depth

When a magnetic field is applied to a superconductor, the applied field does not suddenly drop to zero at the surface. Instead the field decays exponentially according to the formula

$$H(x) = H(0)e^{-x/\lambda} \quad (35.5)$$

where $H(0)$ is the field applied at the surface at $x = 0$ and x is the distance from the surface. The length λ is called the **London penetration depth**. It may be defined as the effective depth to

which a magnetic field penetrates a superconductor. The penetration depth λ ranges from 300 to about 5000 Å depending on the material. It is independent of frequency of the magnetic field but it strongly depends on temperature.

The temperature dependence of λ is given by the relation

$$\lambda(T) = \frac{\lambda(0)}{\sqrt{1 - \left(\frac{T}{T_C}\right)^4}} \quad (35.6)$$

where $\lambda(T)$ and $\lambda(0)$ are the penetration depths at T and 0 K.

It follows from equ. (35.6) that λ increases with the increase in temperature and at the critical temperature, it becomes infinite. At $T = T_C$ the material goes into the normal state and hence the magnetic field penetrates the whole specimen.

Example 35.6. Calculate the penetration depth of lead at 5.2 K if the London penetration depth at 0 K is 37 nm. The critical temperature of lead is 7.193 K.

$$\text{Solution. } \lambda(T) = \lambda(0) \left[1 - \left(\frac{T}{T_C} \right)^4 \right]^{-1/2} = 37 \text{ nm} \left[1 - \left(\frac{5.2 \text{ K}}{7.193 \text{ K}} \right)^4 \right]^{-1/2} = 43.4 \text{ nm.}$$

35.4.8 Flux Quantization

In 1957 A.A. Abrikosov predicted the existence of magnetic flux quanta. According to him, a closed superconducting loop can enclose magnetic flux only in integral multiples of a fundamental quantum of flux. The magnetic flux enclosed by a superconducting ring is given by

$$\phi = n \frac{h}{2e} = n\phi_0 \quad n = 1, 2, 3, \dots$$

where

$$\phi_0 = \frac{h}{2e} \quad (35.7)$$

is the flux quantum and is called a **fluxon**. Its value is

$$\phi_0 = 2.07 \times 10^{-15} \text{ weber}$$

The quantization of magnetic flux has been confirmed experimentally in 1961 by Deaver and Fairbank.

35.4.9 Entropy

Entropy is a measure of the disorder of a system. Generally, entropy decreases with decreasing temperature. In superconducting materials, the entropy decreases linearly up to critical temperature and more remarkably below the critical temperature (see Fig. 35.9).

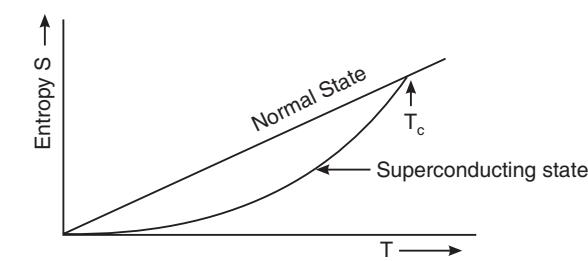


Fig. 35.9

It indicates that a superconducting state is more ordered than the normal state.

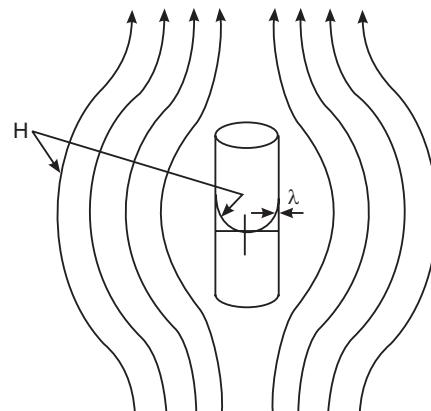


Fig. 35.8

35.4.10 Heat Capacity

The transition of a metal from its normal state to superconducting state does not involve a change of crystallographic structure. It is only a thermodynamic phase transition where the specific heat changes discontinuously at the transition temperature T_C . The specific heat of a normally conducting metal is composed of a lattice part and an electronic part. The electronic contribution varies smoothly at low temperatures. In superconductors it is found that the electronic part decreases exponentially, as shown in Fig. 35.10, and it is given by

$$C_v = e^{-b T_C/T} \quad (35.8)$$

where b is a constant.

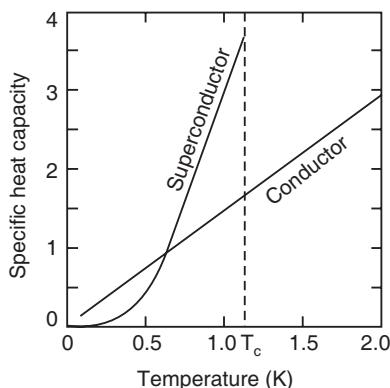


Fig. 35.10

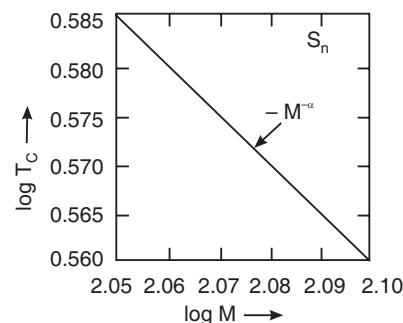


Fig. 35.11

35.4.11 Isotope Effect

In 1950, C.A.Reynolds and E.Maxwell found that the critical temperature decreases with increasing isotopic mass M . The variation is given by the relation

$$T_C \propto M^{-\alpha}$$

or $M^\alpha T_C = \text{constant}$ (35.9)

where α is a constant and is approximately equal to $\frac{1}{2}$. The phenomenon of decrease of

critical temperature with increasing atomic mass is called **isotope effect**. Fig. 35.11 shows the experimental results for tin. Since the mean square of amplitude of atomic vibrations is proportional to \sqrt{M} at low temperatures, the equ. (35.9) suggests that the lattice vibrations are involved in causing superconductivity.

The value of α can be expressed as
$$\alpha = -\frac{\partial(\ln T_C)}{\partial(\ln M)} \quad (35.10)$$

Change in the isotopic mass does not change the electronic structure. Dependence of the transition temperature on the isotopic mass suggests involvement of electron-lattice interaction.

Example 35.7. In a superconducting material isotopic mass is 199.5 amu and critical temperature is 5K. Calculate isotopic mass at 5.1 K.

Solution. $T_C \propto \frac{1}{\sqrt{M}}$. Therefore, $T_1 \sqrt{M_1} = T_2 \sqrt{M_2}$

or

$$M_2 = M_1 \left[\frac{T_1}{T_2} \right]^2 = 199.5 \text{ amu} \left[\frac{5\text{K}}{5.1\text{K}} \right]^2 = 191.68 \text{ amu.}$$

35.5 OTHER EXTERNAL FACTORS THAT AFFECT SUPERCONDUCTIVITY

- Stress:** The transition temperature can be changed by application of stress. Usually, stress which increases the dimensions increases the transition temperature. Highly sensitive methods are required to detect the change.
- Impurities:** Addition of chemical impurities modifies nearly all the superconductivity properties and in particular the magnetic properties.
- Size:** If the size of the specimen is reduced below about $100 \mu\text{m}$, its superconducting properties are modified in many respects. The magnetic permeability of the small specimen is no longer zero and varies with temperature while the critical magnetic field becomes greater than that of bulk material.
- Frequency:** The zero resistance of superconductors is modified at very high frequencies of alternating current. Up to 10 MHz the resistance is still zero but for frequencies of the order of GHz, it is found that considerable resistance is observed even below T_C .

35.6 TYPE-I AND TYPE-II SUPERCONDUCTORS

Superconductors are divided into two categories depending on the way in which the transition from superconducting to normal state proceeds when the externally applied magnetic field exceeds H_C .

Type-I Superconductors:

In type I superconductors, the transition from superconducting state to normal state in the presence of magnetic field occurs sharply at the critical value of H_C , as shown in Fig. 35.12 (a). Type I superconductors are perfectly diamagnetic below H_C and completely expel the magnetic field

from the interior of the superconducting phase. Up to the critical field strength, the magnetization of the material grows in proportion to the external field and then abruptly drops to zero at the transition to the normal state, as shown in Fig. 35.12 (b). The magnetic field can penetrate only the surface layer and current can flow only in this layer. Consequently, type I superconductors are poor carriers of electrical current.

Aluminium, lead and indium are examples of Type I superconductors. The critical field is relatively low for type I superconductors. They would generate magnetic fields of about 100 to 2000 G only. Hence, they are not of much use in production of high magnetic fields. Type I superconductors are also called **soft superconductors**. We summarize here the characteristics of Type I superconductors.

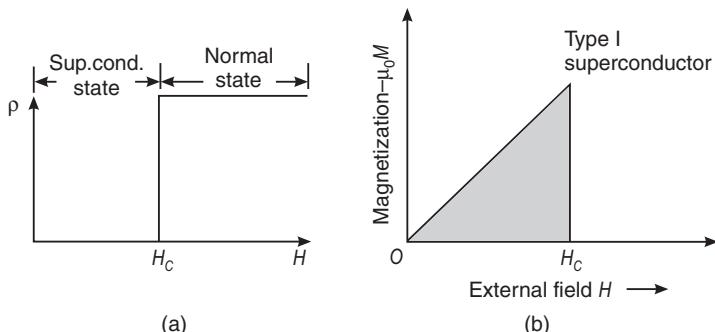


Fig. 35.12

Characteristics of Type-I superconductors

1. They are perfectly diamagnetic and exhibit complete Meissner effect.
2. They have only one critical field. At the critical field the magnetization drops to zero.
3. The maximum critical field for type I superconductor is of the order of 0.1 Wb/m^2 .
4. The transition at H_C is reversible. Below H_C the material behaves as a superconductor, and above H_C it behaves as a normal conductor.

Disadvantages

Type I superconductors cannot carry large currents and hence are not of much use in producing high magnetic fields.

Type-II Superconductors

Type II superconductivity was discovered by Schubnikov in 1930s and was explained by Abrikosov in 1957. Type II superconductors are characterized by two critical fields H_{C1} and

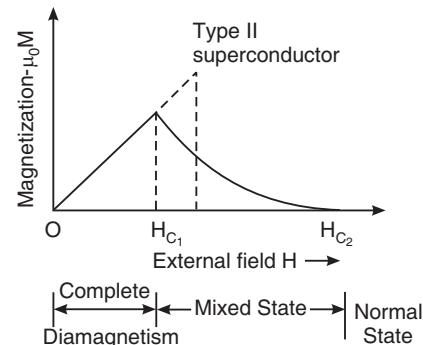
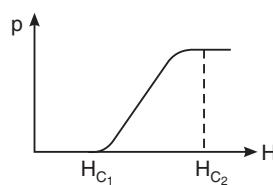


Fig.35.13

H_{C2} . The transition from superconducting state to normal state occurs gradually as the magnetic field is increased from H_{C1} to H_{C2} , as shown in Fig. 35.13. The magnetization of the material grows in proportion to the external field up to the lower critical field H_{C1} . The external magnetic flux is expelled from the interior of the material till then. At H_{C1} the magnetic field lines begin penetrating the material.

As the magnetic field increases further, the magnetic flux through the material increases. At the upper critical field H_{C2} , the magnetization vanishes completely and the external field has completely penetrated and destroyed the superconductivity. In the region between H_{C1} and H_{C2} , the material is in a magnetically mixed state but electrically it is a superconductor. H_{C2} can be as high as 20 to 50 Wb/m^2 and the retention of superconductivity in such high magnetic fields make type II materials very useful in applications of creating very high magnetic fields.

Transition metals and alloys consisting of niobium, silicon and vanadium exhibit type II superconductivity. Ceramic superconductors also belong to this category.

A distinguishing feature of type II superconductors is that super currents arising in an external magnetic field can flow not only over the surface of a conductor but also in the bulk. Above the lower critical field H_{C1} , it becomes energetically more favourable to admit a

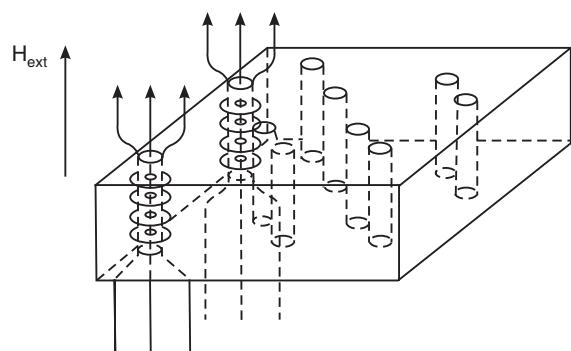


Fig. 35.14

single flux quantum rather than have the superconductor exclude H_{C1} . The superconductor passes into a mixed state where the bulk of the material is superconducting but is threaded by very thin filaments of normal material (see Fig. 35.14). The thin filaments of normal material serve as the paths along which the magnetic field penetrates. In the center of the filament superconductivity is absent. These normal regions are surrounded by vortices of super currents. A flux line together with its current vortex is called a **fluxoid**. At H_{C1} , fluxoids appear in the material and increase in number as the magnetic field is increased. An increase in the magnetic field will not cause an increase of the flux in each vortex line but will cause an increase in the number of fluxoids threading the superconductor. At H_{C2} the fluxoids fill the entire specimen and superconductivity disappears.

Type II superconductors can carry larger currents when the magnetic field lies between H_{C1} and H_{C2} . Type II superconductors are called **hard superconductors**. We summarize here the characteristics of Type II superconductors.

Characteristics of Type-II superconductors

1. They have two critical magnetic fields, H_{C1} and H_{C2} .
2. The material is perfect diamagnetic below the *lower critical field*, H_{C1} . Meissner effect is complete in this region. Above the *upper critical field*, H_{C2} , magnetic flux enters the specimen.
3. Above H_{C1} they do not show complete Meissner effect and therefore do not behave as perfect diamagnetic materials.
4. They exist in an *intermediate state* in between the critical fields, H_{C1} and H_{C2} . The intermediate state is a mixture of the normal and superconducting states, magnetically but electrically the material is a superconductor.
5. At H_{C2} the magnetization vanishes and the specimen returns to normal conducting state.
6. The upper critical field is very high and is of the order of 30 Wb/m^2 .

Applications

They are used in applications of generating very high magnetic fields.

COMPARISON BETWEEN TYPE I AND TYPE II SUPERCONDUCTORS

SL.No.	Type-I Superconductors	Type-II Superconductors
1.	They exhibit complete Meissner effect	They do not exhibit complete Meissner effect
2.	They show perfect diamagnetic behaviour	They do not show perfect diamagnetic behaviour
3.	They have only one critical magnetic field, H_C	They have two critical magnetic fields, lower critical magnetic field, H_{C1} and upper critical magnetic field, H_{C2}
4.	There is no mixed state or intermediate state in case of these materials	Mixed state or intermediate state is present in these materials
5.	The material loses magnetization abruptly	The material loses magnetization gradually
6.	Highest value for H_C is about 0.1 Wb/m^2	Upper critical field is of the order of 30 Wb/m^2
7.	They are known as soft superconductors	They are known as hard superconductors
8.	Lead, tin, mercury are examples	Nb-Sn, Nb-Ti, Nb-Zr, Va-Ga are examples

35.7 BCS THEORY

In 1957, the American physicists J. Bardeen, L.N. Cooper and J.R. Schrieffer developed the quantum theory of superconductivity, which came to be known as BCS theory. Starting from

the two experimental results, namely the isotope effect and the variation of electronic specific heat with temperature, the BCS theory assumed interaction of two electrons through quanta of lattice vibrations. It successfully explained the effects like zero resistivity, Meissner effect etc. The two principal features of BCS theory are

1. Electrons form pairs, called **Cooper pairs**, which propagate throughout the lattice and
2. Such propagation is without resistance because the electrons move in resonance with phonons.

To appreciate the formation of Cooper pair, let us consider the model in Fig. 35.15, in which two electrons propagate along a single lattice row. Each electron experiences an attraction towards its nearest positive ion. When the electrons get very close to each other in the region between ions, they repel each other due to Coulomb force. In an equilibrium condition, a balance between attraction and repulsion is established and the two electrons combine to form a Cooper pair. At normal temperatures, the attractive force is too small and pairing of electrons does not take place. However, at lower temperatures, such pairing is energetically advantageous. In a typical superconductor, the dense cloud of Cooper pairs form a collective state and the motion of all the Cooper pairs is correlated. As such the pairs drift cooperatively through the material. Thus, the superconducting state is an *ordered state of the conduction electrons*. Since the density of Cooper pairs is very high, even large currents require only a small velocity. The small velocity of ordered Cooper pairs minimize collision processes and leads to zero resistivity.

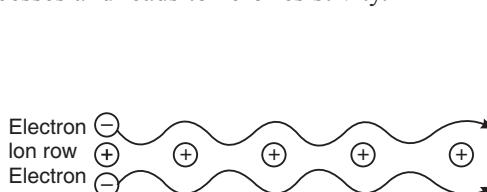


Fig. 35.15

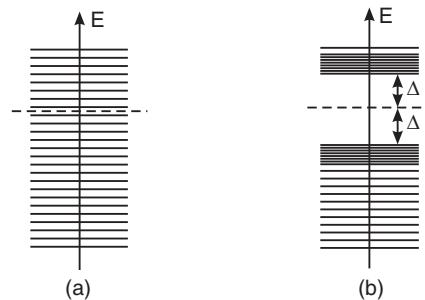


Fig. 35.16

The electrons of a Cooper pair have a lower energy than two unpaired electrons. The theory predicted the existence of an energy gap between the ground state (superconducting state) and first excited state (see Fig. 35.16). The energy gap represents the energy required to break up a Cooper pair. Hence, larger energy gaps correspond to more stable superconductors. According to BCS theory, the energy gap at 0K is given by

$$E_g(0) = 2\Delta \approx 3.52kT_C \quad (35.11)$$

The energy gap is generally of the order of 10^{-3} eV. The existence of energy gap can be proved experimentally. One such experiment involves studying the absorption of microwaves by a superconductor. At temperatures close to absolute zero, a superconductor does not absorb energy until the energy quanta of incident radiation is equal to greater than 2Δ . The absorption then grows fast to a value typical for the normal metal, because electrons can now absorb photons and go to higher energy states that lie above the energy gap.

The BCS theory also predicted flux quantization.

35.8 JOSEPHSON EFFECT

In 1962, Brian Josephson predicted that Cooper pairs could tunnel through an insulating layer, which separates two superconductors. The superconductor-insulator-superconductor layer

constitutes the **Josephson junction**, as shown in Fig. 35.17 (b). The insulating layer is of the order of 1 nm thickness. Josephson predicted that the tunneling can occur without any resistance, giving rise to a direct current when the voltage applied across the junction is zero and an alternating current when the applied voltage is a dc voltage.

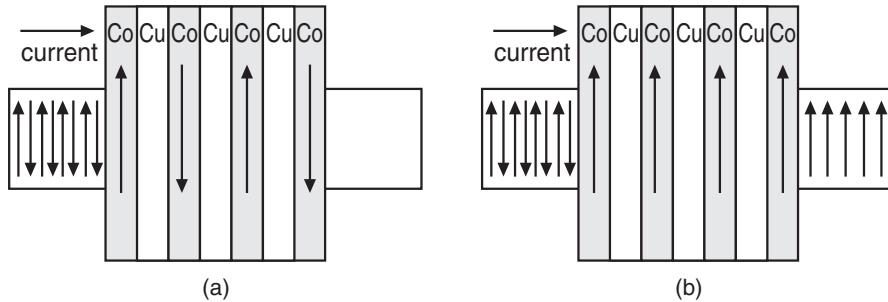


Fig. 35.17

35.8.1 The dc Josephson Effect

Two superconductors separated by a thick insulating layer, say of 10 nm thickness, are essentially two independent superconductors without any joint properties. When the insulator layer is thin, say 1 nm thick, they become a system of coupled conductors. The cooper pairs tunnel through the barrier (insulating layer) as a single unit.

Consider a Josephson junction consisting of two superconducting metal films separated by a thin oxide barrier of 10 to 20 Å thick. Let it be connected in a circuit as shown in Fig. 35.17 (c). The Cooper pairs in a superconductor can be represented by a wave function, which is the same for all pairs. The Cooper pairs tunnel from one side of the junction to the other side easily. The effect of the insulating layer is that it introduces a phase difference between the wave function of Cooper pairs on one side of the insulating layer and the wave function of the pairs on the other side. Because of this phase difference, a super current appears across the junction even though the applied voltage is zero. This is known as the dc Josephson effect. Josephson showed that the super current through the junction is given by

$$I_S = I_C \sin \varphi_0 \quad (35.12)$$

where φ_0 is the phase difference between the wave functions describing Cooper pairs on both sides of the barrier, and I_C is the critical current at zero voltage condition. I_C depends on the thickness and width of the insulating layer and the temperature.

35.8.2 The ac Josephson Effect

If we apply a dc voltage across the Josephson junction, it introduces an additional phase on Cooper pairs during tunneling. As a result a strikingly new phenomenon will be observed. The dc voltage generates an alternating current I given by

$$I = I_C \sin(\varphi_0 + \Delta\varphi) \quad (35.13)$$

Because of the dc voltage V applied across the barrier, the energies of Cooper pairs on both sides of the barrier differ in energy by $2eV$. Using the quantum mechanical calculations, it can be shown that

$$\Delta\varphi = 2\pi t \left(\frac{2eV}{h} \right) \quad (35.14)$$

$$\therefore I = I_C \sin \left[\varphi_0 + 2\pi t \left(\frac{2eV}{h} \right) \right] \quad (35.15)$$

The current given by eq. (35.15) represents an alternating current of frequency

$$\nu = \frac{2eV}{h} \quad (35.16)$$

Equ. (35.16) shows that a photon of frequency ν is emitted or absorbed when a Cooper pair crosses the junction. Thus, when a dc voltage is applied across a Josephson junction, an ac current is produced by the junction. This is known as the **ac Josephson effect**. At $V = 1 \mu V$, ac current of frequency 483.6 MHz is produced.

Example 35.8. A Josephson junction with a voltage difference of $650 \mu V$ radiates electromagnetic radiation. Calculate its frequency.

$$\text{Solution. } \nu = \frac{2eV}{h} \text{ Hz} = \frac{2(1.602 \times 10^{-19} C)(650 \times 10^{-6} V)}{6.626 \times 10^{-34} J} = 3 \times 10^{11} \text{ Hz.}$$

35.9 HIGH SUPERCONDUCTORS

Superconductors are divided into low T_C and high T_C superconductors based on their transition temperature. Broadly, materials having T_C below 24 K are regarded as low T_C superconductors and those having T_C above 27 K are regarded as high T_C superconductors. However, in practice, materials for which liquid nitrogen cooling can cause transition to superconducting state may be regarded as high T_C superconductors while those that require liquid helium coolant are considered as low T_C superconductors. The maximum transition temperature that could be achieved before 1980's was 23.2 K in Nb_3Ge , which is a metallic alloy. Therefore, it was hoped that metallic alloy systems could be made to have higher transition temperatures but such systems

were not discovered. Therefore, the focus has shifted to ceramic oxides, which are insulating materials at normal temperatures. In 1986 Bednortz and Muller discovered superconductivity in ceramic materials. They found that the mixed metallic oxide of lanthanum-barium-copper ($La_1Ba_2Cu_3O_7$) exhibited superconductivity at about 30 K. The superconductivity of the oxide was linked with the deficiency of oxygen ions in the oxide compound. When this deficiency of oxygen was carefully controlled, by keeping the samples in oxygen atmosphere, it was found that the material would exhibit superconductivity in the temperature range 30 to 40 K. In 1987 Chu and coworkers replaced lanthanum with yttrium and prepared $YBa_2Cu_3O_7$ with transition temperature of about 95 K. This was a major breakthrough, as this can be maintained in the superconducting state with far less expensive liquid nitrogen coolant (nitrogen exists in liquid state below 77 K) and marks the beginning of preparation of high T_C superconductors. This oxide also contains a deficiency of oxygen with the chemical formula $YBa_2Cu_3O_{7-\delta}$ where δ indicates the deficiency of oxygen and is in the range 0 to 0.1. $YBa_2Cu_3O_7$ is the most extensively studied high T_C superconductor and exhibits a defective perovskite structure with three perovskite cubic unit cells stacked on top of each other

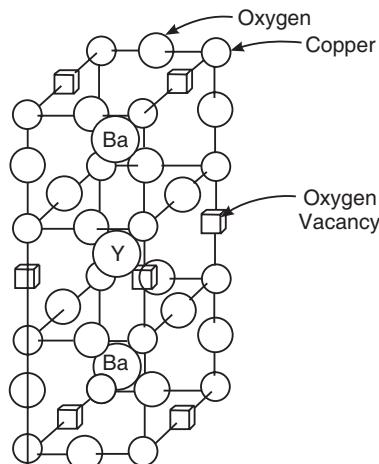


Fig. 35.18

(Fig. 35.18). For an ideal stack of three perovskite cubic unit cells, the $\text{YBa}_2\text{Cu}_3\text{O}_x$ compound should have the composition $\text{YBa}_2\text{Cu}_3\text{O}_9$, in which x would be equal to 9. However, analysis shows that x ranges from 6.65 to 6.90 for this material to exhibit superconductivity. At $x = 6.90$, its T_C is highest (≈ 90 K) and at $x = 6.65$, superconductivity disappears. Thus, oxygen vacancies are found to play a key role in the superconducting behaviour of ceramic oxides. If the cell contains one atom of rare earth metal, two barium atoms, three copper atoms and seven oxygen atoms, then such compounds are called 1-2-3 superconductors. These high T_C copper oxide superconductors belong to type II and their upper critical field is very high (of the order of 150 to 200 tesla). Apart from $\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$ compounds, the compounds based on bismuth like Bi-Sr-Cu-O or Ba-Ca-Cu-O systems are also superconducting. For example, the compound $\text{Bi}_2\text{CaSr}_2\text{Cu}_2\text{O}_{8+x}$ exhibits superconductivity and has the transition temperature 85 K and $\text{Bi}_2\text{Ca}_2\text{Sr}_2\text{Cu}_3\text{O}_{10+x}$ has the transition temperature 110 K. In 1993, a still higher of 133 K was achieved in mercury based copper oxide $\text{HgBa}_2\text{Ca}_2\text{Cu}_3\text{O}_{1+x}$.

The above discussed mixed oxide materials may be arranged into four major families.

- (i) LBCO (mixed oxide of lanthanum, barium and copper)
- (ii) YBCO (mixed oxides of yttrium, barium and copper)
- (iii) BSCCO (mixed oxides of bismuth, strontium, calcium and copper)
- (iv) TBCCO (mixed oxides of thallium, barium, calcium and copper)

Continuous search is going on to discover materials that may exhibit superconducting state around room temperature. Interesting practical applications of superconductors are visualized in many fields, but they could not be used on large scale because of the requirement of liquid helium, which is highly expensive. Therefore, they remained as a laboratory curiosity for a long time. The discovery of high temperature materials opened up possibilities of putting superconductors to large-scale use. They require only liquid nitrogen, which is easily available and also cheaper.

Properties of high T_C superconductors

Some of the properties of high T_C superconductors are as follows:

1. The high T_C superconductors are brittle in nature.
2. The properties of the normal state of these materials are highly anisotropic.
3. The Hall coefficient is positive indicating that the charge carriers are holes.
4. Their behaviour cannot be explained by BCS theory.
5. The isotope effect is almost absent in these materials.
6. The magnetic properties of these materials are highly anisotropic.
7. The effect of pressure is different on different materials. For example the application of pressure increases the critical temperature of LBCO compounds but decreases the critical temperature of YBCO compounds.

35.10 APPLICATIONS

Utilization of superconductivity in practical applications is severely limited by the very low temperatures required to maintain the superconducting material in the superconducting state. Till 1986 the highest critical temperature known was about 27 K. Only using liquid helium as the coolant, which is very costly, one can attain such low temperatures. In the last two decades certain high T_C ceramic materials have been discovered. These materials are brittle, difficult to be drawn into wires and cannot carry large currents. Vigorous research is going on around the world to overcome the drawbacks of high materials and to gainfully utilize them in different applications. We discuss here some of the interesting applications.

1. The most obvious application of superconductors is in power transmission. If the national grid were made of superconductors rather than aluminium, then the savings would be enormous - there would be no need to transform the electricity to a higher voltage (this lowers the current, which reduces energy loss to heat) and then back down again.
2. Superconducting coils in transformers and electrical machines generate much stronger magnetic fields than magnetic circuits employing ferromagnetic materials produce. The normal eddy current losses and hysteresis losses will not be present in superconducting devices and hence the size of motors and generators will be drastically reduced. Thus, superconductors are likely to revolutionize the whole range of rotating electrical machines, making them smaller, lighter and highly efficient. For example, a superconducting generator about half the size of a copper wire generator is about 99% efficient; typical generators are around 50% efficient.
3. High magnetic fields are required in many areas of research and diagnostic equipments in medicine. The electromagnets are cumbersome being very big, demand large electrical power and require continuous cooling. Superconducting solenoids produce very strong magnetic fields. They are small in size and are less cumbersome. They do not need either large power supplies or the means of removing heat. The only power required is to bring the solenoid into the superconducting state and maintain it in that state. The low power requirement and simple cooling technique leads to a large saving in cost. The development of superconductors has improved the field of MRI, as the superconducting magnet can be smaller and more efficient than an equivalent conventional magnet.
4. Type II superconductors can be used as very fast electronic switches due to the way in which a magnetic field can penetrate into the superconductor - this has allowed researchers to build a 4-bit computer microchip operating at about 500 times the speed of current processors, where heat output is currently a major problem with typical speeds approaching the 1GHz mark.
5. **Cryotrons:** The application of a magnetic field greater than its critical magnetic field changes the superconducting state of a superconducting material to normal state and removal of the field brings the material back from normal state to the superconducting state. This fact is used in developing **cryotron switches**.

Fig. 35.19 (a) shows a schematic diagram of a simple cryotron. It consists of a superconducting material (control) coil, B, wound around another superconducting (gate) wire, A. For example, the control coil is of niobium ($T_C = 9.3\text{K}$) or lead ($T_C = 7.2\text{K}$)

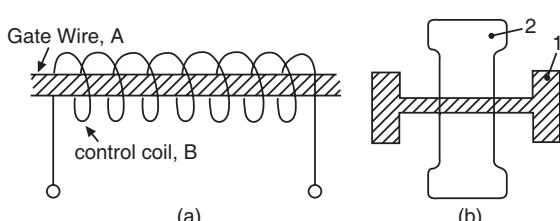


Fig. 35.19

whereas the gate wire is of tantalum. The value of $[H_C]_A$ of material A is less than the value of $[H_C]_B$ of material B. Initially, let the temperature of the device be below the transition temperature of the two materials A and B. In such a condition the wire B does not offer resistance to the flow of current through it as it is in the superconducting state. Now, at the operating temperature, the magnetic field produced by the

coil B may exceed the critical field of the core wire A. It makes the material to go into normal state because the critical field of A is less than that of B. However, B will not go to normal state at the critical field of A because $[H_C]B > [H_C]A$. Hence, the current in the central wire A can be controlled by the current in the coil B. Thus, whenever the current passing through the coil B exceeds the critical current value, the wire A will become a normal conductor exhibiting a finite resistance. This closes the gate for the flow of current through the core wire, A. Removal of the current reopens the gate. Thus the system acts as a fast acting relay or switching element and is highly suitable as fast acting memory element in computers.

The speed of the cryotron switch is dependent on its time constant $\tau = L/R$, where L is the inductance of the control coil B and R is the resistance of the gate coil A in the normal state. Wire wound cryotrons have τ of the order of 10^{-3} s. To reduce the value of τ , R is to be increased and L is to be decreased as far as possible. This objective is achieved by depositing two crosses strips on a substrate and separating them by a thin dielectric layer (Fig. 35.19 b). Strip 1 acts as the gate and is usually made of tin ($T_C = 3.7$ K) and strip 2 acts as the control element usually made of lead ($T_C = 7.2$ K). Varying the current the strip 2 enables switching the strip 1 from the superconducting state to the normal state and vice versa. Thus opening and closing of the circuit is achieved. Through an appropriate design, the switching can be made faster as the value of τ can be decreased up to 10^{-7} s.

6. **MagLev Trains:** The most spectacular application would be the so-called ‘MagLev’ trains. The coaches of the train do not slide over steel rails but float on a four inch air cushion above the track using superconducting magnets; this eliminates friction and energy loss as heat, allowing the train to reach high speeds of the order of 500 km/hr. Such magnetic levitation trains would make train travel much faster, smoother, and more efficient due to the lack of friction between the tracks and train.

Operation

A typical plan of Maglev train is shown in Fig. 35.20. The train has superconducting magnets built into the base of its carriages. An aluminium guideway is laid on the ground and carries electric current. The repulsion between the two powerful magnetic fields, namely the field produced by the superconductor magnet and the field produced by the electric current in the aluminium guideway causes magnetic levitation of the train. A levitation of about 10 to 15 cm is achieved. The train is fitted with retractable wheels, which act in a way similar to the wheels of an airplane. Once the train is levitated in air, the wheels are retracted into the body and the train glides forward on the air cushion. When the train is to be halted, the wheels are drawn out and the train descends slowly onto the guideway and runs forward till it stops.

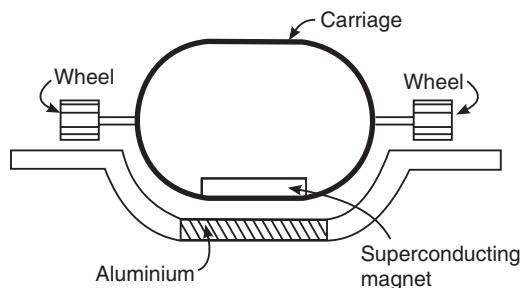


Fig. 35.20

7. SQUIDS

A superconducting quantum interference device (SQUID) is a device used to measure extremely weak magnetic flux. Thus, it is basically a sensitive magnetometer. The heart of a SQUID is a superconducting ring, which contains one or more Josephson junctions. There

are two main types of SQUID: DC SQUID and RF (or AC) SQUID. The DC SQUID was invented in 1964 by Robert Jaklevic, John Lambe, Arnold Silver, and James Mercereau. The RF SQUID was invented in 1965 by J.E.Zimmernan and Arnold Silver at Ford. An RF SQUID is made of one Josephson junction and a dc SQUID consists of two Josephson junctions in parallel and relies on the interference of the currents from each junction. We study here the dc SQUID. DC SQUIDS are more difficult and expensive to produce, but DC SQUID magnetometers are much more sensitive.

Fabrication

SQUIDs are usually fabricated from lead or pure niobium. The lead is taken in the form of an alloy with 10% gold or indium. A thin niobium layer deposited on to it acts as the base electrode of the SQUID and the tunnel barrier is oxidized onto this niobium surface. The top electrode is a layer of lead alloy deposited on top of the other two, forming a sandwich arrangement. The entire device is then cooled to within a few degrees of absolute zero with liquid helium.

The schematic of a two-junction SQUID [direct current (DC) SQUID] is shown in Fig. 35.21 (a). It consists of two Josephson junctions arranged in parallel so that electrons tunneling through the junctions demonstrate quantum interference.

Working

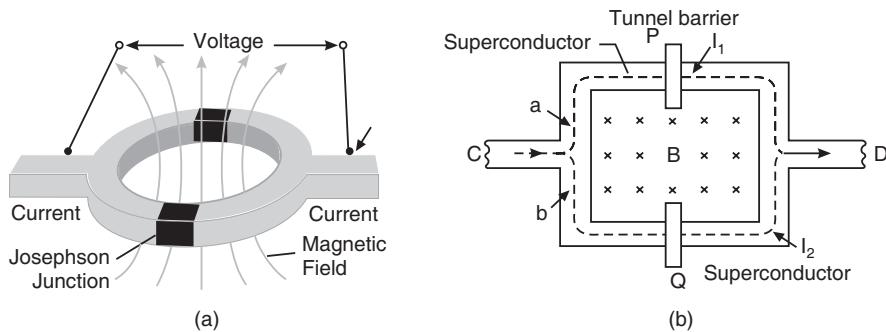


Fig. 35.21

A dc supercurrent is applied to the SQUID (Fig. 35.21 b). This current, known as *bias* current, enters the device through the arm C. It is divided along two paths *a* and *b* and again merge into one and leaves through the arm D. P and Q are the Josephson junctions and the insulating layers at P and Q are of different thickness. I_1 and I_2 are the currents tunneling through the junctions P and Q respectively. In a superconductor, a single wave function describes all the Cooper pairs. The wave function experiences a phase shift at the junctions P and Q. Let the phase difference between points C and D taken on a path through junction P be δ_a and the phase difference between points C and D taken on a path through junction Q be δ_b . In the absence of magnetic field these two phases are equal. That is, $\delta_b - \delta_a = 0$. When a magnetic field B is applied perpendicular to the loop, the flux passes through the loop, and changes the quantum mechanical phase difference across each of the two junctions. The wave functions at the two Josephson junctions interfere with each other. In other words, the supercurrents flowing along the paths *a* (PD) and *b* (QD) interfere. Hence, the device is named SQUID. The interference closely resembles the optical interference observed with Young's double slit. In the case of light, the phase difference between light waves is due to the difference in optical path lengths. In case of supercurrent interference, the waves are the de Broglie waves of Cooper pairs, and the phase difference is caused by the applied magnetic

field. According to Josephson's theory, the phase difference between the reunited currents is directly proportional to the magnetic flux, Φ , through the ring. It can be shown that the total current through two parallel Josephson junctions is given by

$$I_T = 2(I_0 \sin \delta_0) \cos \frac{e\Phi}{\hbar c} \quad (35.17)$$

The above relation indicates that a progressive increase or decrease of the magnetic flux, causes the current to oscillate between a maximum and a minimum value. Maxima in the current occur whenever the magnetic flux increases by one flux quantum. Thus, the period of these oscillations is one flux quantum Φ_0 .

$$\Phi_0 = \frac{\hbar}{2e} = 2.06 \times 10^{-15} \text{ webers}$$

Fig. 35.22 shows the variation of the current through a pair of Josephson junctions as a function of the magnetic flux applied.

In practice, instead of the current, we measure the voltage across the SQUID, which also oscillates with the changing magnetic field. Thus, the SQUID is a flux-to-voltage transducer, converting a tiny change in magnetic flux into voltage. The flux Φ is related to the magnetic field B through the relation

$$\Phi = BA$$

where A is the area of the ring. With the help of this relation, flux measurements are converted into magnetic field measurements.

SQUID is a very sensitive magnetometer, which can measure very weak magnetic fields of the order of 10^{-13} Wb/m^2 . The sensitivity of a SQUID to magnetic fields can be enhanced by using a flux transformer (Fig. 35.23). A flux transformer consists of a loop of superconducting material, which is coupled to the SQUID. An external magnetic field produces a persistent supercurrent in the loop and this current induces a flux in the SQUID. As the loop encloses a much larger area than can a SQUID, the sensitivity of the device gets enhanced.

Applications of Squids

SQUIDS are used to measure very small magnetic fields. Since the current is sensitive to very small changes in the magnetic field, the SQUID acts as a very sensitive magnetometer. As ordinary magnetometers, SQUIDS are capable of measuring magnetic fluctuations of the order of 10^{-13} T . Because of their extreme sensitivity, SQUIDS find applications in many fields, engineering, medicine and many other fields. For example, geologists use them for measuring rock magnetism and continental drift. Physical processes, such as muscular or neural activity, in humans (and other animals) create magnetic fields as small as a thousand billionth of a tesla (as a comparison, a fridge magnet generates about a tenth of a tesla). Human heart generates magnetic fields of about 10^{-14} wb/m^2 and the human brain generates

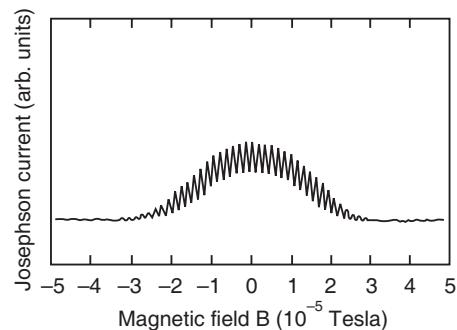


Fig. 35.22

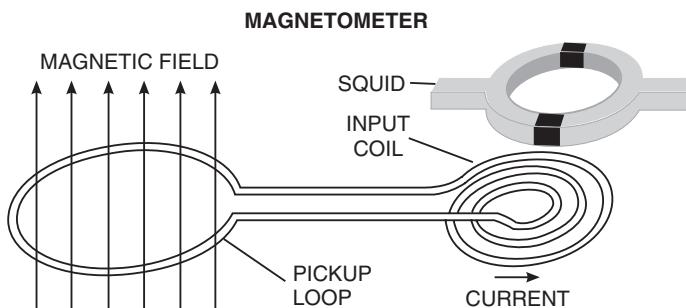


Fig. 35.23

magnetic fields of about 10^{-14} Wb/m². SQUIDs can detect these feeble fields and an array of SQUIDs is used in *magnetoencephalography* (MEG), the process of brain imaging. The SQUIDs are also used in nondestructive testing. In testing for corrosion of aluminum sheets riveted together in aircraft, the SQUID measures the influence of the aircraft skin on an applied oscillating magnetic field; a change in electrical conductivity reveals the defects.

Example 35.9. For a certain metal the critical magnetic field is 5×10^3 A/m at 6K and 2×10^4 A/m at 0K. Determine its transition temperature.

Solution. $T_C = \frac{T}{\left[1 - \frac{H_C(T)}{H_C(0)}\right]^{\frac{1}{2}}} = \frac{6\text{ K}}{\left[1 - \frac{5 \times 10^3 \text{ A/m}}{2 \times 10^4 \text{ A/m}}\right]^{\frac{1}{2}}} = 6.93 \text{ K.}$

QUESTIONS

1. The dc resistance of a superconductor is practically zero. What about its ac resistance?
2. How can you change a superconductor from type I to type II?
3. What is the importance of isotope effect in superconductivity?
4. What are Cooper pairs?
5. What is coherence length?
6. How is Josephson tunneling different from single particle tunneling?
7. A superconducting wire and a copper wire are connected in parallel. Does the copper wire carry current when a potential difference is applied?
8. Give an account of the phenomenon of superconductivity.
9. What is the significance of critical temperature, critical magnetic field and critical current density for superconductors?
10. Explain the Silsbee effect.
11. What is Meissner effect? (M.G.Univ., 2005)
12. Explain Meissner effect and isotope effect. (Calicut Univ., 2007)
13. Compare the dependence of resistance on temperature of a superconductor with that of a normal conductor.
14. Define superconductivity. Explain the effect of isotopes on superconductors. (M.G.Univ., 2005)
15. Discuss the relation between isotopic mass and transition temperature.
16. What is flux quantization?
17. What do you mean by "perfect diamagnetism" of a superconductor?
18. Describe how Cooper pairs are formed and explain the salient features of superconductivity. (V.T.U., 2007)
19. Give a short account of high temperature superconductivity.
20. Explain the term high temperature superconductivity. Give the various applications of superconductors. (M.G.Univ., 2005)
21. Explain high temperature superconductivity. Explain its advantage. (M.G.Univ., 2005)
22. Distinguish between dc and ac Josephson effects. (Calicut Univ., 2007)
23. What is superconductivity? Explain Meissner effect. Describe type-I and type-II superconductors. (C.S.V.T.U., 2005, 2009)
24. What are type I and type II superconductors? Explain B.C.S. theory with key note of Cooper pairs. (M.G.Univ., 2006)
25. Explain the BCS theory with key note of Cooper pairs. Distinguish between type I and II superconductors. (Calicut Univ., 2005)
26. Write short notes on BCS theory of superconductivity. (Calicut Univ., 2005, 2007), (C.S.V.T.U., 2006, 2008)
27. What is superconductivity? Describe type I and type II superconductors. (V.T.U., 2008)

28. Explain in brief type I and type II superconductors. How does a superconductor differ from a normal conductor? (V.T.U., 2007)
29. Differentiate between type I and type II superconductors. (Calicut Univ., 2007)
30. Explain BCS theory of superconductors. Explain the magnetic behaviour of type I and type II superconductors. (M.G.Univ., 2005)
31. Why are type I superconductors poor current carrying conductors?
32. Explain the various properties and important applications of superconducting materials. (M.G.Univ., 2005)
33. Explain ac and dc Josephson's effect. (Calicut Univ., 2006)
34. Explain the theory of d.c.Josephson's effect. (Calicut Univ., 2007)
35. Explain the Josephson's effect in superconductivity. Describe the principle of SQUID and mention its applications. (M.G.Univ., 2005)
36. Explain what is meant by Quantum interference?
37. Discuss the principle and working of SQUID. (M.G.Univ., 2005)
38. Write short notes on SQUIDs. (Calicut Univ., 2005, 2007)
39. What is SQUID? Explain its working. Explain Josephson effect. (M.G.Univ., 2006)
40. What is meant by SQUIDs? Mention anyone of its applications. (Calicut Univ., 2007)
41. Give any five applications of superconductors. (Calicut Univ., 2006)
42. (i) Explain Meissner effect and magnetic levitation.
(ii) Discuss the applications of superconductor. (Anna Univ., 2005)
43. Describe the following practical applications of superconducting materials.:
(i) Superconducting magnet (ii) MAGLEV vehicles (iii) SQUIDs.
44. (a) Explain in detail, the properties of superconducting materials.
(b) What is BCS theory ? Enumerate the important results of BCS theory.
(c) Describe the applications of Superconductors in various fields. (JNTU, 2010)

PROBLEMS

- For a specimen of superconductor, the critical fields are respectively $1.4 \times 10^5 \text{ A/m}$ and $4.2 \times 10^5 \text{ A/m}$ for temperatures 14 K and 13 K respectively. Calculate the transition temperature and the critical fields at 0 K and 4.2 K.
- Estimate the current through a tin wire of diameter 2mm at 2K, if T_C and critical field at 0 K is 0.0305 T.
- A Josephson junction has a voltage of $8.5 \mu\text{V}$ across its terminals. Calculate the frequency of the e.m. waves generated by it. [Ans: 410 GHz]
- Calculate the critical current for a wire of Pb having a diameter of 3mm at 5K. The critical temperature for Pb is 8 K and critical magnetic field is $5 \times 10^4 \text{ A/m}$ at 0K. [Ans: 286.9 A]
- The transition temperature of mercury with an average atomic mass of 200.59 amu is 4.153 K. Determine the transition temperature of one of its isotopes $^{80}\text{Hg}^{204}$. [Ans: 4.118 K]
- The critical temperature of a superconductor at zero magnetic field is T_C . Determine the temperature at which the critical field becomes half of its value at 0K. [Ans: 0.707 T_C]
- A long thin superconducting wire of a metal produces a magnetic field of $105 \times 10^3 \text{ A/m}$ on its surface due to the current through it at a certain temperature T. The critical magnetic field of the metal is $150 \times 10^3 \text{ A/m}$ at 0K. The critical temperature T_C of the metal is 9.2 K. What is the value of T? [Ans: 4.3 K]
- A lead wire has a critical magnetic field of $6.5 \times 10^3 \text{ A/m}$ at 0K. The critical temperature is 7.18 K. At what temperature the critical field would drop to $4.5 \times 10^3 \text{ A/m}$. The diameter of the wire is 2mm. What is the critical current density at that temperature? [Ans: 5.19 K, $9.6 \times 10^6 \text{ A/m}^2$]
- The London penetration depth of mercury at 3.5 K is 75 nm. Estimate the penetration depth at 0K. [Ans: 51.9 nm]
- The penetration depths for lead at 3 K and 7.1 K are 39.6 nm and 173 nm respectively. Calculate the critical temperature for lead. [Ans: 7.193 K]

CHAPTER

36

Modern Engineering Materials

36.1 INTRODUCTION

The progress of human civilization is closely linked to the developments in materials. Science and technology have made amazing developments in the design of electronics and machinery using standard materials. Standard materials have been traditionally divided into three basic groups, namely metals, ceramics and polymers. This classification is primarily based on atomic structure and chemical makeup. Later, three other groups of engineering materials are added which are composites, semiconductors and biomaterials. There is a continuous search for materials with improved properties such as good mechanical strength, high stability, large electrical conductivity etc. A number of new materials such as amorphous metals, liquid crystals, smart materials, biomaterials etc have been discovered for high-tech applications. These materials have properties that scientists can manipulate.

36.2 METALLIC GLASSES

It has been found in 1970s that some of the metals can be produced in the amorphous state if they are cooled rapidly from the liquid state to solid state. Thus, one can produce amorphous metallic solids on cooling from liquid phase to the solid phase at high cooling rates of the order of 10^6°C/s . The amorphous metallic solids thus formed are known as **metallic glasses** or **met glasses**. Thus, *metallic glasses are amorphous metals with non-crystalline structures that are produced by solidification of liquid alloys*.

Metals are malleable, ductile whereas glass is brittle, transparent and non-magnetic. The metallic glasses share the properties of both metals and glasses. They are strong, ductile, malleable and opaque like a metal, and in addition, they have amorphous structure, brittleness and high corrosion resistance like glass. As they are amorphous solids, they are also known as **amorphous metals**.

Solids are divided basically into two types: crystals and amorphous solids. Amorphous solids are generally referred to as **super cooled liquids**, since they do not possess long-range order. The arrangement of atoms within amorphous materials is not periodic and regular. Glass is a familiar example of amorphous solids while metals are examples of crystalline solids. Glass is obtained from liquid glass when cooling rate is high enough to bypass crystallization. Similarly, liquid alloys form metallic glasses on solidification under cooling rates rapid enough to suppress the crystal phase.

Glass transition temperature

The temperature at which the frozen liquid reaches the glassy state is known as **glass transition temperature**, T_g . The glass transition temperature for metallic alloys is about 20°C to 300°C. The glass- to-liquid transition is reversible and upon heating, metallic glasses exhibit the transition at T_g .

36.2.1 Types of Metallic Glasses

Metallic glasses are of two types, namely metal-metal glasses and metal – metalloid glasses.

- (i) **Metal-metal glasses:** These are formed through a combination of metals, for example, Cu-Zr, Ni-Nb, Mg-Zn etc.
- (ii) **Metal-metalliod glasses:** These are produced from a combination of metals Fe, Co, Ni and metalloids B, Si, C, P. The composition ranges are found to be varied for each system. In addition, there are rare-earth transition metal glasses like GdCo, GdFe, GdTbFe which have good magnetic properties. Both the types of glasses are stable at room temperature and above. Some are capable of withstanding temperatures of the order of 773 K without crystallization.

36.2.2 Preparation of Metallic Glass

Metallic glasses are prepared by cooling molten metal to the glassy state at a cooling rate high enough to bypass crystallization. The cooling rate for glass formation varies from material to material. Pure metals are difficult to form in the glassy state because they require very fast rates of cooling. They are produced by atomic deposition technique like sputtering. In general, metallic glasses consist of two or more elements, as they require cooling rates slower than those required for pure metals.

There are several techniques available for producing metallic glass. We discuss here briefly two important techniques.

1. Melt Spinning

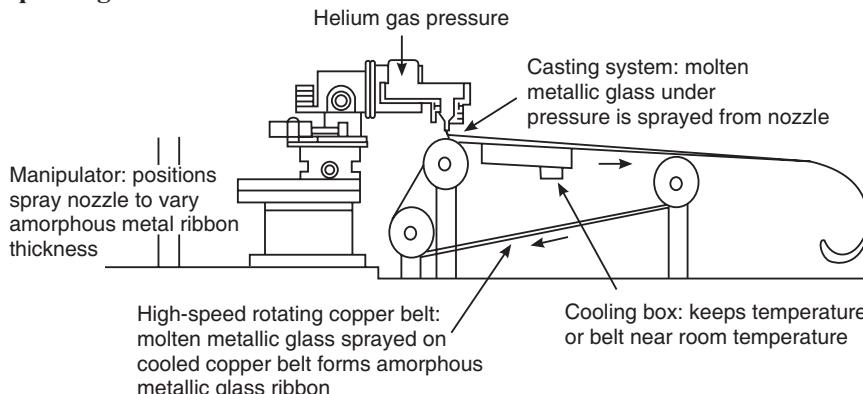


Fig. 36.1

Melt spinning is one of the common methods used to prepare metallic glass and is used in its large-scale production. The alloy is filled in a refractory tube having a tapered end (Fig. 36.1). The crucible is heated by an r.f. induction heater. The alloy is melted under inert gas atmosphere and when the gas pressure is increased, it is sprayed through the fine nozzle at the end of the refractory tube.

The molten metal is directed on to a copper belt, located below the refractory tube and rotating at a very high speed. The molten alloy falling on the cooled copper belt is solidified and forms a thin ribbon of about 0.0025 cm thick and 15 cm wide, which is wound on a

spool. The main advantage of this method is that the metallic glass is obtained in the form of continuous ribbon having uniform cross-section and reproducible properties.

2. Sputtering

Sputtering is another common technique, used especially for the formation of amorphous semiconductors and metals. Sputtering is a vacuum deposition process in which atoms are released from a target under the bombardment of positive ions and deposited on a substrate. A sputtering unit consists of a high power diode operating at r.f. frequencies of the order of 14 MHz. The sputtering chamber (Fig. 36.2) is evacuated and a steady flow of ultra high pure sputtering gas is introduced into the chamber. The target (cathode) consists of a base metal partially covered with another metal piece. The substrate is a mirror polished silicon single crystal attached to the anode. When the unit is switched on, the sputtering gas is ionized and the positive ions bombard the surface of the cathode releasing atoms or groups of atoms from it. These atoms fly toward the substrate and form a thin layer of metallic glass on it.

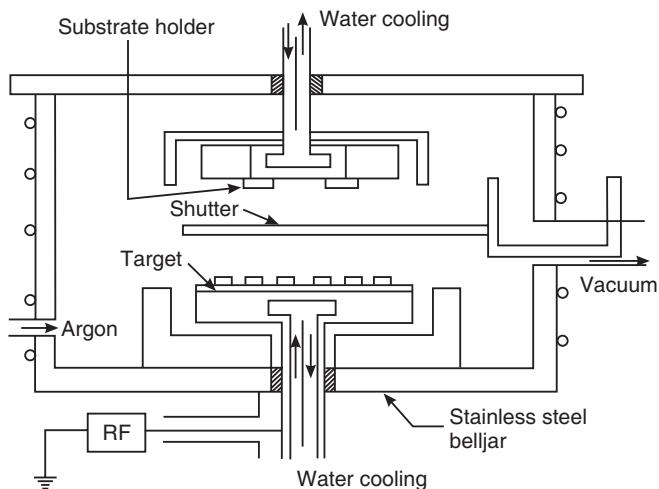


Fig. 36.2

36.2.3 Properties of Metallic Glasses

The metallic glasses exhibit a number of superior properties compared to their crystalline counterparts. Metallic glasses exhibit high mechanical strength, good ductility, magnetic behavior, very high elastic limit, and resistance to wear and corrosion, which set them apart from conventional crystalline materials. As they are non-crystalline in nature, they lack the grain boundaries, dislocation defects and segregated characteristics which are typical of crystalline materials.

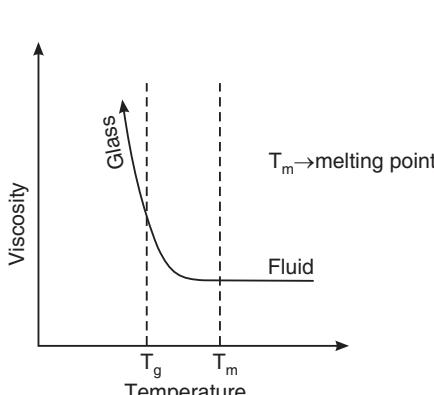


Fig. 36.3. Temperature dependence of viscosity of a supercooled melt

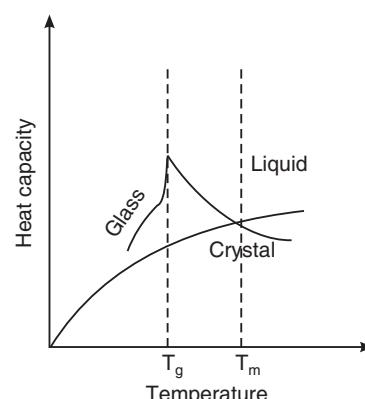


Fig. 36.4. Heat capacity variation as a function of temperature

Thermodynamic properties

Glass transition is a phenomenon that occurs when a liquid undergoes configurational freezing to become a solid. During this transition, atomic mobility is reduced to a large extent and there is a dramatic increase in viscosity as shown in Fig. 36.3. Hence, the material loses its fluidity.

The heat capacity vanishes at very low temperatures, increases over an intermediate temperature range. Near T_g , an anomalous increase is observed. Fig. 36.4 shows the heat capacity variation with temperature. The thermal conductivity of metallic glass at room temperature is generally of the same order of magnitude as that of crystalline alloys.

Structural Properties

- The metallic glasses do not possess grain boundaries and dislocations.
- Studies on the structure of state showed that the atomic packing is of *tetrahedral* close packing. These materials do not possess long-range anisotropy.
- They can be easily made into ribbons.

Mechanical Properties

- Because of the absence of defects and dislocations in their structure, metallic glasses are stronger than metals and alloys.
- Metallic glasses have better corrosion resistance than the crystalline metal.
- Metallic glasses are highly ductile.
- In general, the filaments of metallic glass can be used as low-cost high-performance structural reinforcement elements as they are stronger and stiffer.
- The elastic constants of metallic glasses are smaller than those observed in the corresponding crystalline materials. By contrast, values of yield strength (σ_y) and hardness (H) are typically much greater than those of crystalline solids. A common measure of the relative strength of solid is the ratio of yield strength to Young's modulus (σ_y/E). For metallic glasses, this ratio is typically 0.02 which approaches the theoretical maximum attainable yield strength. Hence, metallic glasses are among the strongest of known solids. The hardness of metallic glasses corresponds to the yield strength.
- However, as metallic glasses are very susceptible to brittle fracture, their practical applications are limited.

Electrical Properties

- Electrical resistivity of metallic glasses is high and it does not vary much with temperature.
- Eddy current losses are very small in metallic glasses due to their high resistivity.
- The Hall co-efficient of metallic glasses is found to have both positive and negative signs. This indicates that the current through metallic glasses is due to the conduction of electrons and holes as in case of crystalline states.
- Some metallic glasses exhibit superconductivity; for example metal-metalloid glasses based on molybdenum, ruthenium, rhenium and niobium. They are extreme Type II superconductors with very small lower critical fields (H_{c1}) and very large upper critical fields (H_{c2}).

Magnetic Properties

- The amorphous metallic magnetic glasses consist of various combinations of ferromagnetic Fe, Co, and Ni with the metalloids B and Si. These materials have superior soft magnetic properties.

- Eddy currents are minimized by the smaller thickness and high electrical resistivity of the metallic glasses. Thus, the core losses are very less.
- They exhibit very narrow hysteresis loops, as shown in Fig. 36.5, and thus have very low hysteresis energy losses.
- These materials do not possess long-range anisotropy and grain boundaries. Therefore, domain walls move with extreme ease in them.
- These materials possess high maximum permeabilities and hence they can be magnetized and demagnetized very easily.
- The Curie temperature of metallic glasses is generally slightly lower than for crystalline phase because of the dilution effect of metalloid elements and atomic disorder in metallic glasses.
- Similarly, the coercivity of ferromagnetic metallic glasses is lower because of random fluctuations in the amorphous phase.

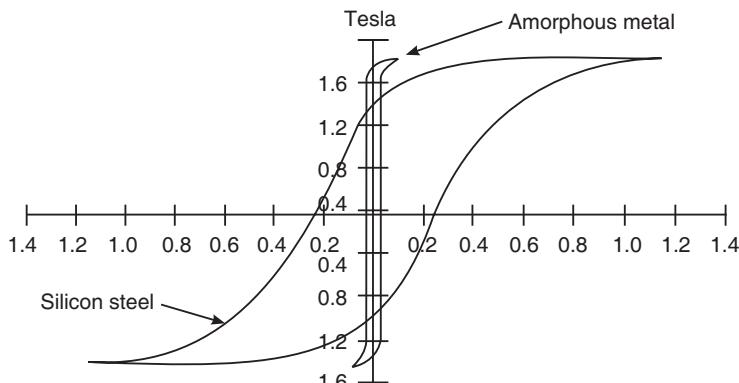


Fig. 36.6

Chemical properties

- The absence of extended defects like grain boundary or dislocation in the amorphous metals offers a more uniform surface which leads to interesting chemical properties. Due to the formation of protective oxide film chromium containing metallic glasses have excellent corrosion resistance in a variety of chemically hostile environments.
- Apart from the absence of extended defects, the amorphous metals possess structural fluctuations on the scale of atomic distances. Each local atomic environment is distinct and varies over the surface. Hence, the surface exhibits a distribution of chemically active sites which leads to catalytic properties. Therefore, the amorphous state is more active than the crystalline state. Amorphous iron-nickel-metalloid alloys are used for catalytic synthesis of hydrocarbons by hydrogenation of carbon monoxide.

36.2.4 Applications of Metallic Glasses

Soft magnetic materials

The homogeneity, low coercive force and relatively high permeability of metallic glasses are suitable for many magnetic applications. Thin sheets of metallic glasses are used as the core materials for power distribution transformers. Eddy current loss is proportional to the electrical conductivity and inversely proportional to thickness of the sheet. Hence, metallic glasses are used to reduce eddy current losses as they have lower electrical conductivity. The low coercive force reduces the core loss. Hence, the utilization of metallic glasses in transformers reduces the energy loss and thereby increases the efficiency.

Metallic glasses are also increasingly used in inductive components such as reactors, inverter transformers, chokes for magnetic switches, tape recorder heads, magnetic shields etc. Applications in the areas of magnetic sensors and transducers are under development.

The chief advantage of metallic glasses is that they are fabricated using inexpensive metals and they can be fabricated with relative ease in the form of thin tapes.

- (i) Metallic glasses possess good tensile strength and are superior to common steels. They are used as reinforcing elements in concrete, plastic and rubber.
- (ii) Metallic glasses possess high ductility, good corrosion resistance and hence used in making springs for different applications.
- (iii) One of the most important applications of metallic glasses is as inductors in transformers, recording devices, magnetic shielding, motors and others.
- (iv) A transformer having a core made from metallic glass is found to have eight times less core loss and requires about twenty times less current for excitation, compared to an equal sized commercial transformer having a crystalline silicon steel core.
- (v) Magnetic recording heads made of metallic glasses are superior in overall performance to those made of ferrites and permalloys, because of their high flux density and high wear resistance.
- (vi) Metallic glasses offer opportunities for increasing the efficiency of magnets, motors, and transformers used in energy-conversion devices. They would reduce costly losses from power-distribution systems and corrosion damage.

36.3 LIQUID CRYSTALS

Liquid crystals are substances that exhibit a phase of matter that has properties between those of a conventional liquid and those of a crystal. Liquid crystals have long rod-shaped organic molecules of about 25 Angstroms in length. In the liquid crystal state, the molecules can move about and, therefore, lack **positional order**. However, the molecules strongly interact and are arranged and oriented in a crystal-like way. Thus, they possess long-range **orientational order**. The common axis along which the molecules tend to orient is called the **director**.

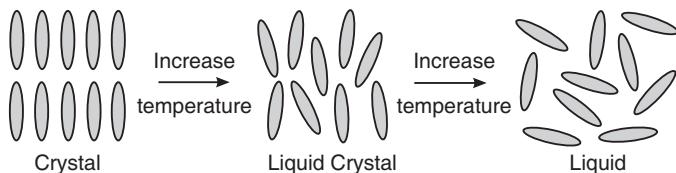


Fig. 36.6

The tendency of the liquid crystal molecules to point along the director leads to a condition known as **anisotropy**. This term means that the properties of a material depend on the direction in which they are measured. The anisotropic nature of liquid crystals is responsible for the unique optical properties exploited in a variety of applications.

Liquid crystal phases can be divided into two classes: **thermotropic** and **lyotropic** liquid crystals. Thermotropic liquid crystals pass into the liquid crystal phase as temperature is increased. The liquid crystal state exists within some temperature range, $T_m < T < T_c$, where T_m is temperature of melting from solid state into liquid crystal state, and T_c is clearing temperature, when the liquid crystal transforms into an isotropic liquid. In the solid state, the centers of gravity of molecules possess long-range positional order, and, also, the molecules orientation points in the same direction providing the long-range orientational order. When

solid melts into a liquid crystal at T_m , the positional order is lost although some orientational order of the molecular long axes remains. At still higher temperature T_C , liquid crystal melts into an isotropic liquid with no positional and orientational order.

An example of a compound displaying thermotropic liquid crystal behavior is methoxybenzilidene butylaniline (MBBA), which is a nematic liquid crystal between 21°C and 48°C.

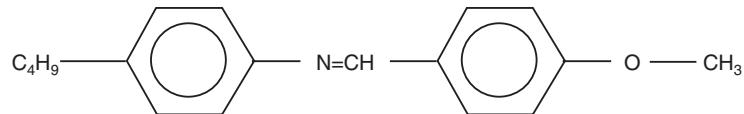


Fig. 36.7. Methoxybenzilidene Butylaniline ("MBBA")

Lyotropic liquid crystals exhibit phase transitions as a function of concentration of the liquid crystal in a solvent (typically water) as well as temperature.

36.3.1 Liquid Crystal Phases

Liquid crystal phases are broadly divided into three types, depending upon the positional order and orientational order of the molecules in the phase. They are

- (i) Nematic phase
- (ii) Smectic phase and
- (iii) Cholesteric phase.

(i) Nematic phase: The nematic liquid crystal phase is characterized by molecules that have no positional order but tend to point in the same direction (along the director). In the following diagram, notice that the molecules point vertically but are arranged with no particular order.

In the nematic phase, the molecules flow and their center of mass positions are randomly distributed as in a liquid, but they all point in the same direction within a small domain. Nematics have fluidity similar to that of ordinary (isotropic) liquids but they can be easily aligned by an external magnetic or electric field. An aligned nematic has the optical properties of a uniaxial crystal and this makes them extremely useful in liquid crystal displays.

The characteristic properties of nematic liquid crystals may be summed up as follows:

- In nematic phase the molecules have no positional order; the location of their centres of gravity is irregular as in a liquid.
- The molecules have long range orientational order; the axes of the molecules are parallel to each other.
- The nematic liquid crystal consists of smaller regions each having its molecules aligned parallel to a unique axis, called **director**. This axis varies in direction in different regions of the crystal. Therefore, nematic liquid crystal appears turbid.
- The ordering of the molecules varies with temperature.
- The molecular orientation can be controlled with applied electric and magnetic fields.
- When an external field orients the molecules in the same direction, the liquid crystal appears transparent.



Fig. 36.8. Schematic of molecular ordering in a nematic phase

- Special treatment of the surfaces supporting nematic liquid crystal causes alignment of the molecules in a desired way: either parallel to the surface or perpendicular to the surface.
- If the molecules are aligned parallel to the surface, the alignment is said to be **planar**. If the molecules are aligned perpendicular to the surface, the alignment is said to be **homeotropic**.

(ii) Smectic Phase: In the smectic state, the molecules maintain the general orientational order of nematics, but also tend to align themselves in layers or planes. Motion is restricted to within these planes, and separate planes are observed to flow past each other. The increased order means that the smectic state is more “solid-like” than the nematic.

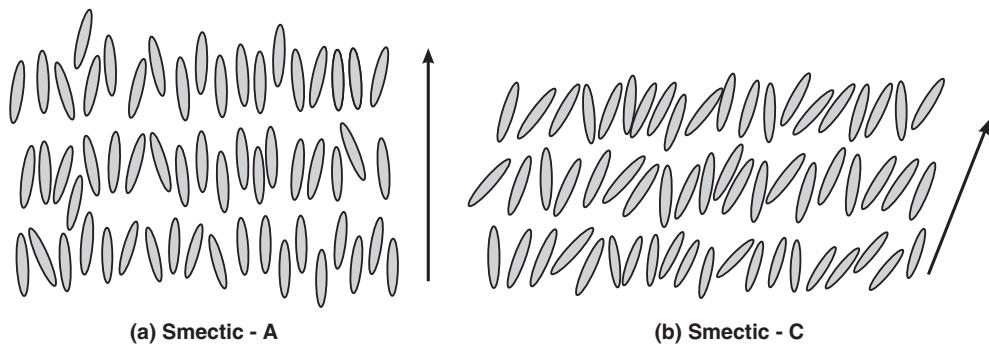


Fig. 36.9. Schematic of molecular alignment in the smectic phases. (a) The smectic A phase has molecules organized into layers. (b) In the smectic C phase, the molecules are tilted inside the layers

Many compounds are observed to form more than one type of smectic phase. About 12 of these variations have been identified. Three of the more often observed variations are known as smectic-A, smectic-B and smectic-C.

The characteristic properties of smectic liquid crystals may be summed up as follows:

- These phases occur at lower temperatures than nematic phase.
- Smectic liquid crystals have a layer structure. Molecules align themselves parallel in each layer. The layers can slide over one another causing fluidity to the liquid crystal.
- In the smectic-A phase, the director is perpendicular to the smectic plane, and there is no particular positional order in the layer.
- Similarly, the smectic-B molecules orient perpendicular to the smectic plane, but the molecules are arranged into a network of hexagons within the layer.
- In the smectic-C phase, molecules are arranged as in the smectic-A phase, but the director is at a constant tilt angle measured normally to the smectic plane.
- The value of the angle varies with temperature and can be altered by an external electric field.

(iii) Cholesteric phase: This phase is called the cholesteric phase because it was first observed for cholesterol derivatives. Only chiral molecules (i.e., those that lack inversion symmetry) can give rise to this phase. Therefore, the phase is also called

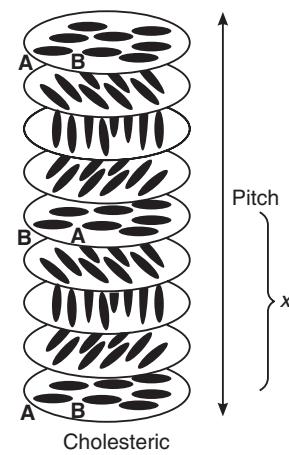


Fig. 36.10. Schematic of molecular alignment in the cholesteric phase

chiral nematic phase. The phase exhibits a twisting of the molecules perpendicular to the director, with the molecular axis parallel to the director. The finite twist angle between adjacent molecules is due to their asymmetric packing, which results in longer-range chiral order.

When the molecules align in layers, this causes the director orientation to rotate slightly between the layers, eventually bringing the molecules back into the original orientation. Thus, the distance it takes for the director to rotate one full turn in the helix is known as the **pitch** which is an important characteristic of the cholesteric phase. Note that the structure of the chiral nematic phase repeats itself every half-pitch, since in this phase directors at 0° and $\pm 180^\circ$ are equivalent. The pitch changes when the temperature is altered or when other molecules are added to the host. A byproduct of the helical structure of the chiral nematic phase is its ability to selectively reflect light of wavelengths equal to the pitch length, so that a color will be reflected when the pitch is equal to the corresponding wavelength of light in the visible spectrum. Mixtures of various types of cholesteric liquid crystals are often used to create sensors with a wide variety of responses to temperature change. Such sensors are used as thermometers often in the form of heat sensitive films to detect flaws in circuit board connections, fluid flow patterns, condition of batteries, the presence of radiation, or in novelties such as “mood” rings.

The characteristic properties of cholesteric liquid crystals may be summed up as follows:

- The phase exhibits twisting of molecules in successive layers, the molecular axis being parallel to the director.
- In going from one layer to the next layer, the director turns by a certain angle and it describes a helix with a pitch of about 0.2 to $20\ \mu\text{m}$.
- Though the chiral pitch refers to the distance it takes for the director to rotate one full (360°) turn, the structure of the chiral nematic phase repeats itself every half-pitch.
- The periodicity of the structure leads to Braggs reflection of light at a wavelength equal to the pitch divided by the refractive index of the cholesteric liquid crystal.
- The pitch varies with temperature. It also varies when other molecules are added to the liquid crystal.
- Since the pitch varies with temperature, the wavelength of the reflected light and the colour of the liquid crystal vary with temperature.
- Cholesteric liquid crystals exhibit high optical activity and cause the plane of polarization of light to turn through angles of the order of 6000° to $7000^\circ/\text{mm}$.

36.3.2 Electric and Magnetic Field Effects

The response of liquid crystal molecules to an electric field is the major characteristic utilized in industrial applications. The ability of the director to align along an external field is caused by the electric nature of the molecules. When an external electric field is applied to the liquid crystal, the dipole molecules tend to orient themselves along the direction of the field.

36.3.3 Liquid Crystal Displays (LCDs)

The most common application of liquid crystal technology is liquid crystal displays (LCDs). They come in two different formats – (i) segment displays and (ii) matrix displays. A **liquid crystal display (LCD)** is a thin, flat cell filled with liquid crystal material and arrayed in front of a light source. It is often utilized in battery-powered electronic devices because it uses very small amounts of electric power.

A segment LCD consists of two thin glass plates having transparent electrically conducting coatings of indium tin oxide on their inner surfaces. These two plates will also serve as the electrodes. The conducting coating on one of the glass plates is etched in the

form of a digit or character, as illustrated in Fig. 36.11 (a). The surfaces of the electrodes that are in contact with the liquid crystal material are treated so as to align the liquid crystal molecules in a direction that is parallel to the glass and also to each other. In case of a twisted nematic LCD, the surfaces are treated in mutually perpendicular directions. When the cell is filled with liquid crystal material, the molecular layers spiral a quarter of a turn about the twist axis normal to the glass plates. The molecules are aligned vertically on one electrode and gradually they are turned until they are horizontal on the other electrode. This results in a **twisted nematic device**. The cell is then sandwiched between two Polaroid sheets held in crossed configuration, as shown in Fig. 36.11 (b). The rear Polaroid sheet is backed with a reflecting film. With no actual liquid crystal between the polarizing filters, light passing through the first filter would be blocked by the second (crossed) polarizer and appears dark. On the other hand, a cell filled with liquid crystal rotates the plane of polarization of the incident light and hence light is not blocked. As a result the device appears grey.

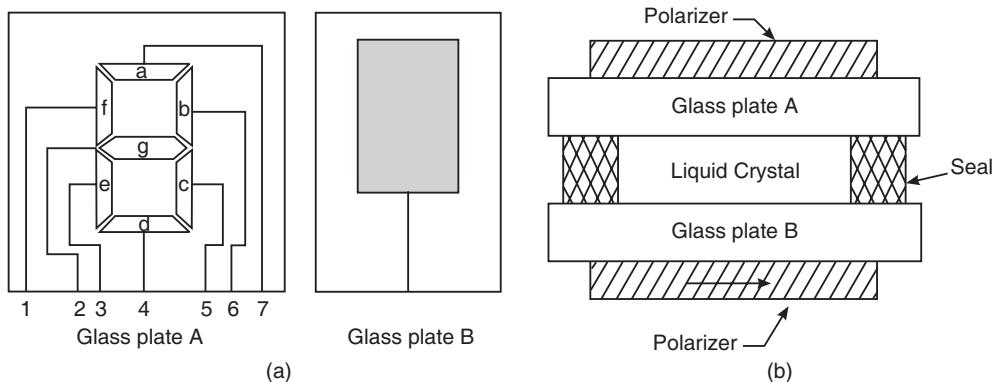


Fig. 36.11

When a voltage is applied across the cell, the liquid crystal molecules turn into the direction of the electric field. If the applied voltage is large enough, the liquid crystal molecules in the center of the layer are almost completely untwisted and the polarization of the incident light is not rotated as it passes through the liquid crystal layer. The light will then be mainly polarized perpendicular to the second filter, and thus gets blocked (Fig. 36.12 b). As a result, the areas where electrodes are present in the form of digits or character, appears dark (see Fig. 36.13).

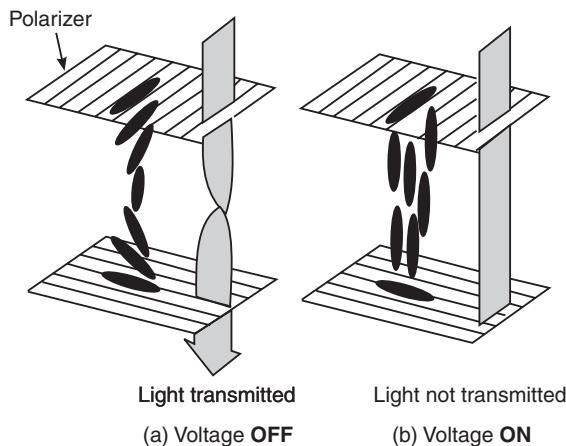


Fig. 36.12

36.3.4 Flat Panel Liquid Crystal Displays

Small liquid crystal displays such as used in calculators and other devices have direct driven segments. Voltage can be applied across one segment without interfering with other segments

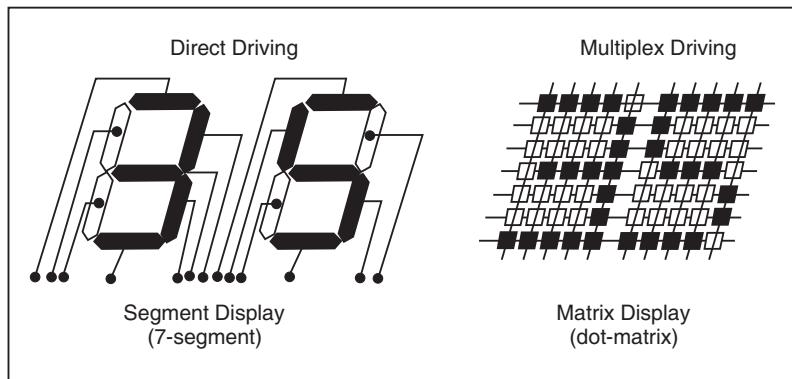


Fig. 36.13

of the display. This is impractical for a large display with a large number of picture elements (**pixels**). A **flat panel liquid crystal display (LCD)** is made up of a matrix of pixels filled with liquid crystal material and arrayed in front of a light source. There are mainly two types of flat panel LCD displays - **passive matrix** and **active matrix**.

Passive-matrix LCD Panels

Passive-matrix LCDs (Fig. 36.14 a) use a simple grid to supply the voltage to a particular pixel on the display. It has two glass layers called **substrates**. One substrate is given columns and the other is given rows made from a transparent conductive material, usually **indium-tin oxide**. The liquid crystal material is sandwiched between the two glass substrates, and a polarizing film is added to the outer side of each substrate. The rows or columns are connected to integrated circuits that control when a voltage is applied to a particular column or row. The point of intersection of the row and column represents the designated pixel on the LCD panel to which a voltage is applied to untwist the liquid crystal at that pixel to control the passage of light.

In passive-matrix LCDs (PMLCDs) there are no switching devices, and each pixel is addressed for more than one

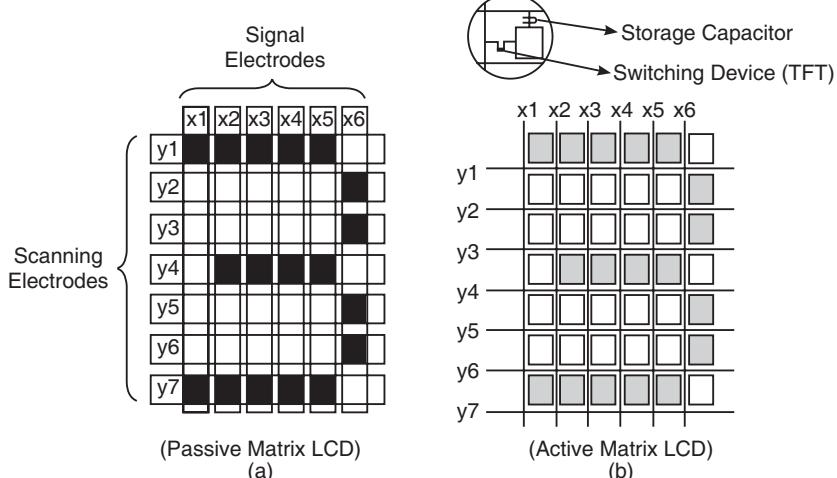


Fig. 36.14

frame time. The effective voltage applied to the LC must average the signal voltage pulses over several frame times, which results in a slow response time of greater than 150 msec and a reduction of the maximum contrast ratio.

Passive matrix LCD displays are simple to manufacture, and therefore cheap, but they have a *slow response time* - in the order of a few hundred milliseconds - and a relatively

imprecise voltage control. These characteristics render images that are somewhat fuzzy and lacking in contrast. Passive matrix LCD displays are therefore unsuitable for most of the high speed, high resolution video applications.

Active Matrix LCD Panels

A switching device and a storage capacitor are integrated at the each cross point of the electrodes in active-matrix LCDs (AMLCDs) (Fig. 36.14 b). For their integrated switching devices they use transistors made of deposited thin films, which are therefore called thin-film transistors (TFTs).

TFTs are arranged in a matrix on a glass substrate to control each picture element (or pixel). Switching on one of the TFTs will activate the associated pixel.

The use of an active switching device embedded onto the display panel itself to control each picture element helps reduce cross-talk between adjacent pixels while drastically improving the display response.

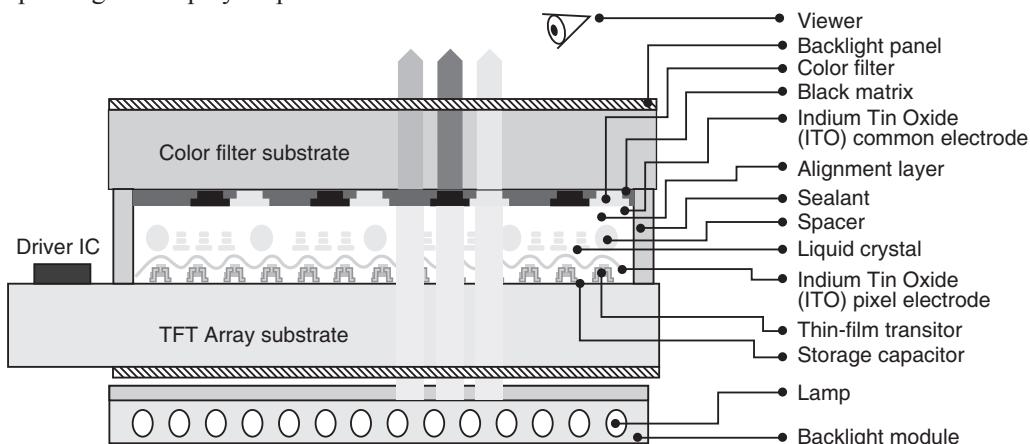


Fig. 36.15. Structure of TFT LCD

An active matrix display does not suffer from many of the limitations of the passive display. It can be viewed at an angle of up to 45 degrees and has a contrast of 40:1, meaning that the brightness of an “on” pixel is 40 times greater than an “off” pixel. It does, however, require a more intense back lighting system because the TFT’s and the gate and source lines are not very transparent and therefore block a fraction of the light.

36.3.5 Color Displays

The techniques discussed so far have only been able to describe a simple two color display. In order to achieve color, it is first necessary to have a display which is black in one state and white in the other. This distinction is made because some displays may have a yellow on blue appearance which will not be able to produce the full range of colors. In a white display, all wavelengths pass through and therefore, all wavelengths can be manipulated to create the desired color. To get full color, each individual pixel is divided into three subpixels: red, green and blue (RGB). That is to say that for each full color pixel, three distinct pixels are employed. These subpixels are created by applying color filters which only allow certain wavelengths to pass through them while absorbing the rest. With a combination of red, blue and green subpixels of various intensities, a pixel can be made to appear any number of different colors. This is analogous to a color cathode ray tube (CRT) like a television or computer monitor in which different phosphors glow red, green or blue when excited by an electron beam. The

number of colors that can be made by mixing red, green and blue subpixels depends on the number of distinct gray scales (intensities) that can be achieved by the display. The structure of a TFT LCD is shown in Fig. 36.15.

36.3.6 Specifications

- **Resolution:** The horizontal and vertical size expressed in pixels (e.g., 1024×768). An LCD panel has a fixed number of liquid crystal cells and can display only one resolution at full-screen size using one cell per pixel. Lower resolutions can be displayed by using only a proportion of the screen. For example, a 1024×768 panel can display a resolution of 640×480 by using only 66% of the screen.
- **Dot pitch:** The distance between the centers of two adjacent pixels. The smaller the dot pitch size, the less granularity is present, resulting in a sharper image. Dot pitch may be the same both vertically and horizontally, or different (less common).
- **Viewable size:** The size of an LCD panel measured on the diagonal (more specifically known as active display area).
- **Response time:** Response time is the time it takes for an applied voltage to effect the liquid crystals' alignment and register a change on the screen. It is measured in milliseconds and clearly the lower the value the better for the screen. Very fast changes, 3ms or less, will give fluid on-screen motion and a clear picture. Slower response times, above 12ms, will likely lead to problems with motion blur and ghosting. An AMLCD has a much better response time than a PMLCD.
- **Refresh rate:** The number of times per second in which the monitor draws the data it is being given. Since activated LCD pixels do not flash on/off between frames, LCD monitors exhibit no refresh-induced flicker, no matter how low the refresh rate. Many high-end LCD televisions now have a 120 Hz or 200 Hz refresh rate.
- **Matrix type:** Active matrix or Passive matrix.
- **Viewing angle:** Viewing angle describes the direction of observation of (a point on) a visual display.
- **Color support:** The number of types of colors that are supported. Through the careful control and variation of the voltage applied, the intensity of each subpixel can range over **256 shades**. Combining the subpixels produces a possible palette of **16.8 million colors** (256 shades of red \times 256 shades of green \times 256 shades of blue).
- **Brightness:** The amount of light emitted from the display, more specifically known as luminance. Luminance is often used to characterize emission or reflection from flat, diffuse surfaces. The luminance indicates how much luminous power will be perceived by an eye looking at the surface from a particular angle of view. Luminance is thus an indicator of how bright the surface will appear. A typical computer display emits between 50 and 300 cd/m².
- **Contrast ratio:** Contrast ratio is a measure of how much brighter a pure white output is compared to a pure black output. The higher the contrast the sharper the image and the more pure the white will be. The ratio of the intensity of the brightest bright to the darkest dark.

36.4 SHAPE MEMORY ALLOYS

A **shape memory alloy** (SMA) is an alloy that "remembers" its shape, and can be returned to that shape after being deformed, by applying heat to the alloy. When an SMA is cold, or below its transformation temperature, it has a very low yield strength and can be deformed quite easily into any new shape, which it will retain. However, when the material is heated above its transformation temperature it undergoes a change in crystal structure which causes it to return to its original shape. In other words they "remember" their original shape. This effect is

known as the **shape memory effect** (SME), and the alloys are named **shape memory alloys** (SMA). The shape recovery process occurs due to a reversible solid-solid phase transformation in particular metal alloys. A solid-solid phase change is that in which a molecular rearrangement occurs, but the molecules remain closely packed so that the substance remains a solid.

If the SMA encounters any resistance during this transformation, it can generate extremely large forces. Therefore, when the shape memory effect is correctly harnessed, this material becomes a lightweight, solid-state alternative to conventional actuators such as hydraulic, pneumatic, and motor-based systems.

36.4.1 Definition

Shape memory alloys (SMAs) are metallic alloys that when severely deformed at some relatively low temperature, regain their original shape after a thermal (heating/cooling) cycle.

36.4.2 Two Phases

There are two stable solid phases which occur in shape memory alloys. They are

- The high-temperature phase, called **austenite** and
- The low-temperature phase, called **martensite**.

These two phases have geometrically different crystallographic elementary cells (see Fig. 36.16 b).

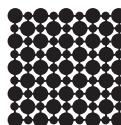
The martensite can be in one of two forms: **twinned** and **detwinned**, as shown in Fig. 36.16 (a).

Martensite is a crystallographically less-ordered phase and therefore is the softer and easily deformed phase of shape memory alloys. This phase exists at lower temperatures. The martensite structure is self-accommodating, and the deformation on transformation to martensite is zero.

Austenite is a more-ordered phase and hence the stronger phase of shape memory alloys. It occurs at higher temperatures. The shape of the austenite structure is cubic. The twinned martensite phase is of the same size and shape as the cubic austenite phase on a macroscopic scale. Change in size or shape becomes visible in shape memory alloys when the twinned martensite is detwinned, that is deformed.

Austenite

- High temperature phase
- Cubic Crystal Structure



Martensite

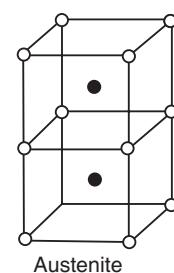
- Low temperature phase
- Monoclinic Crystal Structure



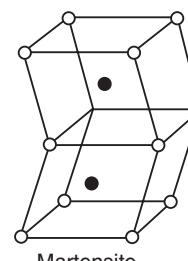
Twinned Martensite



Detwinned Martensite



Austenite



Martensite

(a)

(b)

Fig. 36.16. Different phases of an SMA

A phase transformation which occurs between these two phases upon heating/cooling is the basis for the unique properties of the SMAs. The phase transformation may occur due to the temperature, or mechanical load applied or due to both temperature and load. The key effects associated with the phase transformation are **shape memory effect** and **pseudoelasticity**.

Let us now study the effect of temperature and mechanical load on the phases of shape memory alloys.

1. Temperature-induced phase change without mechanical load: If an SMA is cooled in the *absence of applied load* the material changes from austenite into twinned martensite. Observable macroscopic shape change does not occur during the transformation. Now if the material in the martensitic phase is heated, a reverse phase transformation takes place and the material transforms to austenite. The above process is shown in Fig. 36.17.

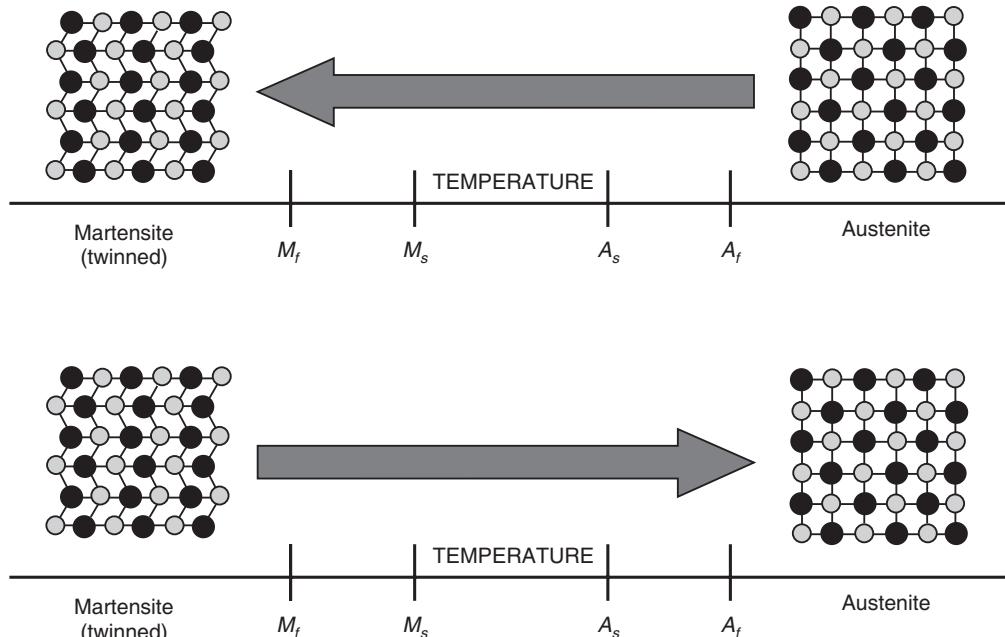


Fig. 36.17. Temperature-induced phase transformation of an SMA without mechanical loading.

Characteristic temperatures

There are four characteristic temperatures characterizing the phase transformation. The four characteristic temperatures are defined in Fig. 36.17.

- (i) Martensite start temperature, M_s , is the temperature at which the material starts transforming from austenite to martensite. The transformation proceeds with further cooling and is complete at the martensite finish temperature, M_f . Below M_f , the entire body is in the martensite phase, and a specimen typically consists of many regions each containing a different variant of martensite. The boundaries between the variants are mobile under small applied loads.
- (ii) Martensite finish temperature, M_f , is the temperature at which the transformation is complete and the material is fully in the martensitic phase;
- (iii) With heating, the austenite start temperature, A_s , is the temperature at which austenite first appears in the martensite. With further heating, more and more of the body transforms back into austenite and

- (iv) Austenite finish temperature, A_f , at which the reverse phase transformation is completed and the material is in the austenitic phase. Above A_f , the specimen is in the original undistorted state.

2. Temperature-Induced Transformation with Applied Mechanical Load: At a temperature above M_s , the specimen is entirely in the austenite phase. When the specimen is cooled below M_f , it transforms entirely to the twinned martensite state, but the macroscopic volume of the specimen does not change - a condition known as *self-accommodation*. By applying small loads the specimen can be easily *detwinned* or deformed (Fig. 36.18), and the deformed shape remains after removing the loads.

Now, if the specimen in the detwinned state is heated to above the temperature A_f , the original un-deformed shape may be recovered (Fig. 36.18).

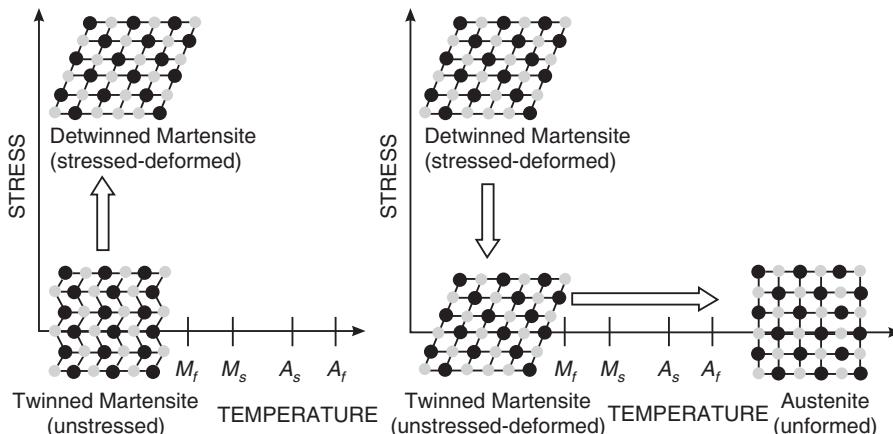


Fig. 36.18. Deforming of martensite phase under the action of applied load

3. Shape Memory Effect: When a piece of shape memory alloy is cooled below the temperature M_f and the material is applied a constant load, it gets deformed. When cooled, the SMA goes into twinned martensite phase and when it is loaded in this state, the SMA goes into the state of deformed (detwinned) martensite. The original shape can be recovered simply by heating the piece above the temperature A_f . When heated the deformed martensite is transformed to the cubic austenite phase (Fig. 36.19). The heat transferred to the material is the power driving the molecular rearrangement of the alloy, similar to heat melting ice into water, but the alloy remains solid.

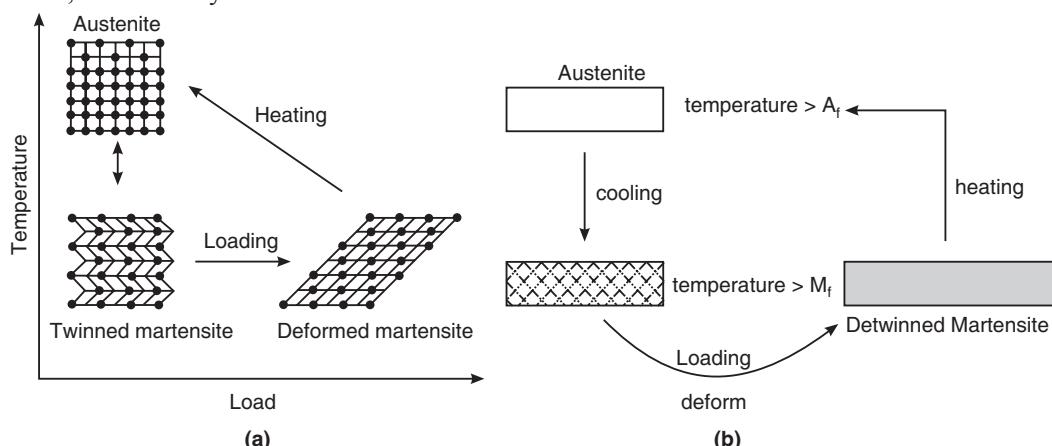


Fig. 36.19. Schematics of Shape Memory Effect

Note that an SMA undergoes phase changes while remaining a solid. The phase changes occur below its melting point. These phase changes “involve the rearrangement of the position of particles within the crystal structure of the solid”. Thus, the alloy can retain its shape without melting. Under the transition temperature, the alloy is in the martensite phase and can be bent into various shapes. To get the initial shape, the metal must be held in position and heated to about 500 °C. The high temperature causes the atoms to arrange themselves into a high symmetry, often cubic arrangement known as the austenite phase.

4. Load-induced phase change at constant temperature -Pseudo-elasticity: A phase transformation may be induced in a sample purely due to mechanical load and without a change in its temperature. Let the alloy be in the austenite phase. Referring to Fig. 36.19, the load on the shape memory alloy is increased until the austenite becomes transformed into martensite simply due to the loading. When the loading is decreased, the martensite begins to transform back to austenite since the temperature of the wire is still above A_f , and the wire springs back to its original shape. Thus, the material behavior resembles elasticity. This effect is known as pseudoelastic Effect and the phenomenon **pseudo-elasticity or super elasticity**. The pseudoelastic stress-strain diagram is shown in Fig. 36.20.

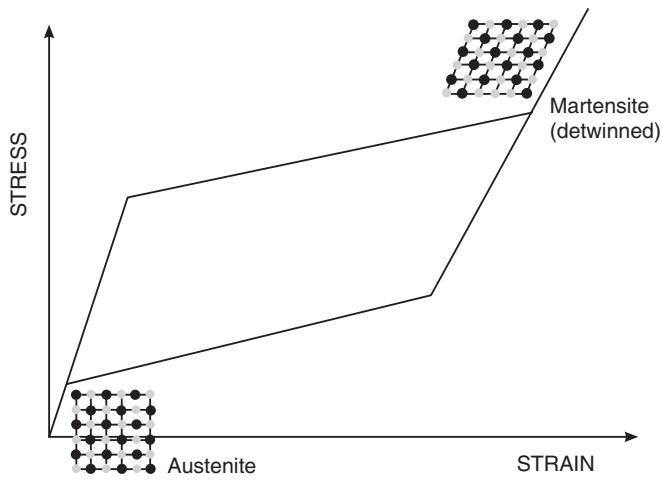


Fig. 36.20. Pseudoelastic stress-strain diagram

36.4.3 Characteristics of SMA

1. Transformation hysteresis

The transition from the *martensite* phase to the *austenite* phase is only dependent on temperature and stress. The transformation of SMA from one solid phase to another solid phase does not occur sharply at a particular temperature, but occurs over a range of temperature, as shown in Fig. 36.21. In this figure, $\xi(T)$ represents the martensite fraction. Most of the transformation occurs over a narrow temperature range. Further, the heating and cooling transformations exhibit hysteresis.

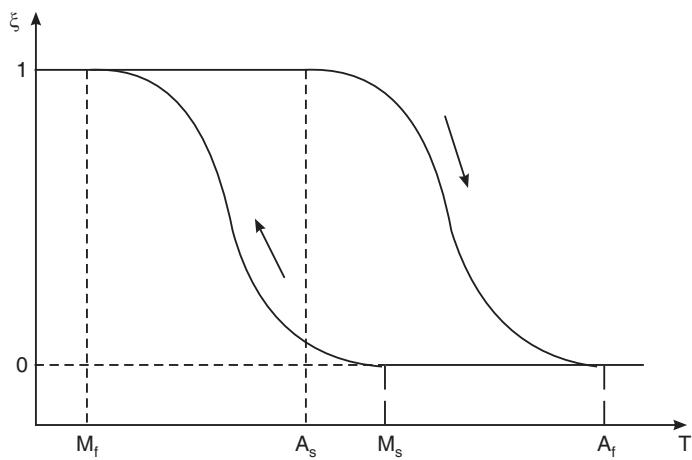


Fig. 36.21

That is, the transformation on heating does not overlap on the transformation on cooling (Fig. 36.21). The magnitude of the hysteresis varies with the alloy type and is typically in the range $10^\circ - 50^\circ\text{C}$.

2. Variation of Phase Change Temperature on Loading

Fig. 36.22 shows the characteristic temperatures of a shape memory alloy. M_s and M_f are the temperatures of martensite phase start and martensite phase finish respectively; whereas A_s and A_f are respectively the temperatures of austenite phase start and austenite phase finish. The initial values of these four variables are dependent on the composition of the alloy material (i.e. what amounts of each element are present).

Secondly, the values of the temperatures vary with the load applied on the SMA.

3. The Superelasticity

The mechanical properties of SMAs vary over the temperature range spanning their transformation. At low temperatures, the material exists as martensite and is deformed by a relatively small applied force. It has a rubbery feel and can be deformed by a small force. At high temperatures, the material exists as austenite, which is not easily deformed; it behaves like normal metals and any induced strain is not recoverable because there is no phase change.

Superelasticity is a mechanical type of shape memory as opposed to thermally induced shape memory. In this case, a small force induces considerable deformation but when the force is removed, the material automatically recovers its original shape without the need for heating. Shape memory alloys show a super elastic behaviour if deformed at a temperature which is slightly above their transformation temperatures. The applied stress transforms the austenite to martensite and the material exhibits increasing strain at constant applied stress, i.e. considerable deformation occurs for a relatively small applied stress. When the stress is removed, the martensite reverts to austenite and the material recovers its original shape. This effect, which makes the alloy appear extremely elastic, is known as **superelasticity** or **pseudoelasticity**.

36.4.4 DIFFERENT KINDS OF SHAPE MEMORY EFFECT

Shape memory alloys may have different kinds of shape memory effect. The two most common memory effects are the one-way and two-way shape memory.

1. The one-way Shape Memory Effect

A schematic view of the effect is given in Fig. 36.23. When a shape memory alloy is in its cold state (below A_s), the metal can be bent or stretched into a variety of new shapes and will hold that shape until it is heated above the transition temperature. Upon heating, the shape changes back to its original shape, regardless of the shape it was when cold. When the metal cools again it will remain in the hot shape, until deformed again. Note that the shape recovery is achieved only during heating.

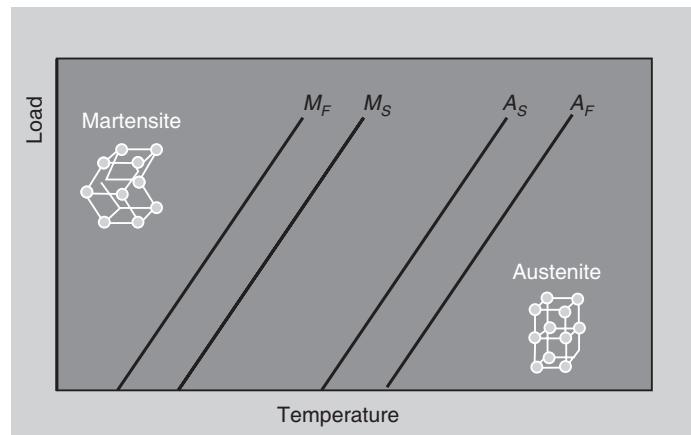


Fig. 36.22. The Dependency of Phase Change Temperature on Loading

With the one-way effect, cooling from high temperatures does not cause a macroscopic shape change. A deformation is necessary to create the low-temperature shape. On heating, transformation starts at A_s and is completed at A_f (typically 2 to 20°C or hotter, depending on the alloy or the loading conditions).

When the sample is cooled below the martensite finish temperature, the martensitic units produced from austenite take many different orientations in an effort to assume minimum energy configuration. Though the transformation of the phase into the martensite is basically a deformation process, the overall macroscopic deformation upon transformation is zero. If the material is now deformed, a particular orientation of the various self-accommodating units – that most favourably oriented with respect to the applied stress – grows at the expense of others, eventually leading to a single oriented martensite. If we heat the material above the austenite start temperature, the martensite units convert totally into austenite and the accumulated deformation is totally recovered. The schematic of a stress-strain-temperature curve is shown in Fig. 36.24.

2. The two-way Shape Memory Effect

The two-way shape memory effect is the effect in which the material remembers two different shapes: one at low temperatures, and one at the high temperature shape. A schematic view of the effect is given in Fig. 36.25. A material that shows a shape memory effect during both heating and cooling is called two-way shape memory. This can also be obtained without the application of an external force. The reason the material behaves so differently in these situations lies in training. Training implies that a shape memory can “learn” to behave in a certain way. Under normal circumstances, a shape memory alloy “remembers” its high-temperature shape, but upon heating to recover the high-temperature shape, immediately “forgets” the low-temperature shape. However, it can be “trained” to “remember” to leave some reminders of the deformed low-temperature condition in the high-temperature phases. There are several ways of doing this.

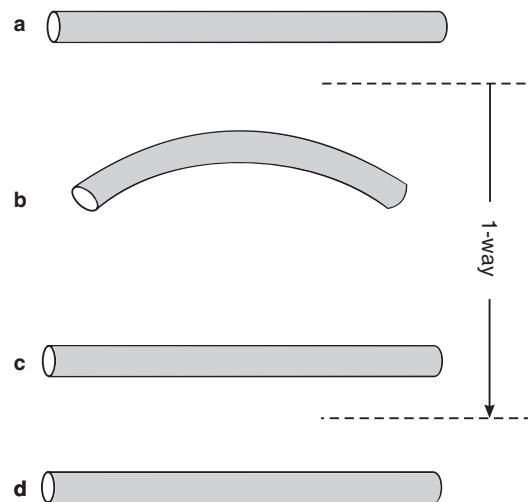


Fig. 36.23. One-way memory effect: (a) Material in martensite phase, (b) material deformed, (c) sample heated and (d) sample cooled again.

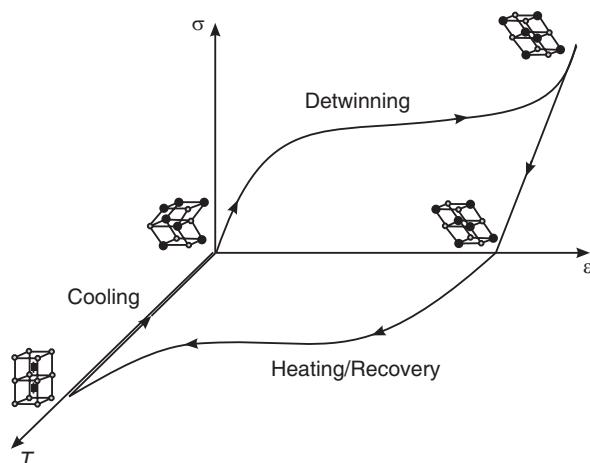


Fig. 36.24. Schematic of a stress-strain-temperature curve showing the shape memory effect.

For example, a specimen of SMA is “programmed” by means of thermo-mechanical treatments producing micro-stresses in the parent phase, which in turn programs the specimen to behave as a stress-induced martensitic transformation. That is, the micro-stresses favour only one orientation of the martensitic phase upon subsequent cooling, which produce a spontaneous and reversible deformation just upon heating and cooling cycles, even with no stress applied. Thus, a two-way shape memory effect is obtained. The two-way SME can be repeated indefinitely too (see Fig. 36.26), as opposed to the one-way SME, which is a one time only operation.

36.4.5 Manufacture

There are various ways to manufacture shape memory materials. Some of the techniques of producing nickel-titanium alloys are vacuum melting techniques such as electron-beam melting, vacuum arc melting or vacuum induction melting. These are specialist techniques used to keep impurities in the alloy to a minimum and ensure the metals are well mixed. The ingot is then hot-rolled into longer sections and then drawn to turn it into wire. Hot working to this point is done at temperatures between 700°C and 900°C . There is also a process of cold working of Ni-Ti alloys. Carbide and diamond dies are used in the process to produce wires ranging from 0.075mm to 1.25mm in diameter.

36.4.6 Characterization Techniques

There are a number of methods to characterize the transformation of SMAs. Two methods of characterizing SMAs are discussed below:

1. A direct method called Differential Scanning Calorimetry (DSC) uses heat energy measurements to characterize SMA. The heat absorbed or given out by a small unstressed sample of SMA is measured during heating and cooling through the transformation temperature range. The endotherm and exotherm peaks are measured for the beginning and the end of phase change.
2. The most direct method employed to characterize SMA is strain measurements. At constant stress, the sample is allowed to undergo the cycle of transformation. The strain experienced by the sample is measured simultaneously. The transformation hysteresis shown in Fig. 36.21 is the direct result obtained from this information. The martensite starts M_s and austenite finish A_f , transformation temperatures obtained through this method are slightly higher than those obtained from DSC method. The strain measurement is made by applying stress over the sample. On the other hand,

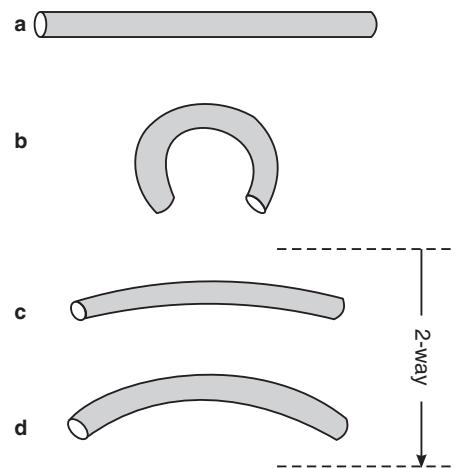


Fig. 36.25. Two-way memory effect: (a) Material in martensite phase, (b) material deformed, (c) sample heated and (d) sample cooled again.

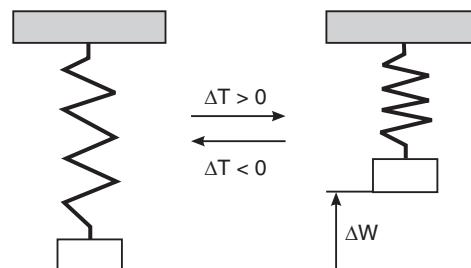


Fig. 36.26. A two-way SME spring. This device is able to produce work (ΔW) by thermal cycles.

DSC test is made with no stress applied. The increase in the stress is attributed to the increase in transformation temperatures M_s and A_f . This test is the direct indicative of shape memory effect in any material. Difficulty in specimen preparation and the dependence of results on the way the test is conducted are the disadvantages of this method.

36.4.7 Engineering SMAs

The shape memory effect was first observed in AuCd in 1951 and since then it has been observed in numerous other alloy systems. However, copper base alloys and nickel-titanium alloys are found suitable for commercial exploitation. They are regarded as engineering materials. These compositions can be manufactured to almost any shape and size. The yield strength of shape memory alloys is lower than that of conventional steel, but some compositions have a higher yield strength than plastic or aluminium. The yield stress for NiTi can reach 500 MPa. The high cost of the metal itself and the processing requirements make it difficult and expensive to implement SMAs into a design. As a result, these materials are used in applications where the superelastic properties or the shape memory effect can be exploited. The most common application is in actuation.

1. Nickel – Titanium alloys: The nickel-titanium alloys were first developed in 1962-63 by the Naval Ordnance Laboratory and commercialized under the trade name Nitinol (an acronym for Nickel Titanium Naval Ordnance Laboratories). Their remarkable properties were discovered by accident. A sample that was bent out of shape many times was presented at a laboratory management meeting. One of the associate technical directors, Dr. David S. Muzzey, decided to see what would happen if the sample was subjected to heat and held his pipe lighter underneath it. To everyone's amazement the sample stretched back to its original shape.

Ni-Ti alloy is an extraordinary intermetallic binary compound because it has moderate solubility for excess nickel or titanium. This solubility enables alloying with many other elements thereby getting systems with modified mechanical and transformation properties. Its ductility is comparable to ordinary alloys. It has greater shape memory strain and excellent corrosion resistance compared to copper – base alloys. Also, it tends to be much more thermally stable. Because of the reactivity of titanium, the melting of this alloy must be done in vacuum or in an inert atmosphere. Methods like plasma – arc melting, electron beam melting and vacuum – induction melting are commonly adopted for this purpose. Welding, brazing or soldering of this alloy is generally difficult. The material responds to grinding and shearing. Even punching can be done if thickness is small. The properties of Nitinol are particular to the exact composition of the metal and the way it was processed.

Physical Properties of Nitinol

- Density: 6.45gms/cc
- Melting Temperature: 1240-1310°C
- Resistivity (hi-temp state): 82 $\mu\text{ohm}\cdot\text{cm}$
- Resistivity (low-temp state): 76 $\mu\text{ohm}\cdot\text{cm}$
- Thermal Conductivity: 0.1 W/cm \cdot °C
- Heat Capacity: 0.077 cal/gm \cdot °C
- Latent Heat: 5.78 cal/gm; 24.2 J/gm

Mechanical Properties of Nitinol

- Ultimate Tensile Strength: 754 - 960 MPa
- Typical Elongation to Fracture: 15.5 percent

- Typical Yield Strength (hi-temp): 560 MPa
- Typical Yield Strength (lo-temp): 100 MPa
- Approximate Elastic Modulus (hi-temp): 75 GPa
- Approximate Elastic Modulus (lo-temp): 28 GPa
- Approximate Poisson's Ratio: 0.3

Actuation

- Energy Conversion Efficiency: 5%
- Work Output: ~1 Joule/gram
- Available Transformation Temperatures: -100 to +100° C

2. Copper-Zinc-Aluminium (CuZnAl) alloys: CuZnAl was the first copper based SMA to be commercially exploited and the alloys typically contain 15-30 wt% Zn and 3-7 wt% Al. The useful transformation temperature for this system ranges from -100°C to + 100°C; the actual transformation temperature is a function of both the alloy composition and the thermomechanical treatments applied during its manufacture. The major advantage of the CuZnAl alloys is that they are made from relatively inexpensive metals by conventional metallurgical processes which makes them the cheapest of the commercial SMAs. However, their memory properties are modest with a maximum recoverable strain of about 5%. The ternary alloys also have a very large grain size which makes them brittle but the addition of < 1% of a grain refiner such as titanium or zirconium, limits the grain size and solves the brittleness problem.

Copper base alloys have medium corrosion resistance and they are susceptible to stress corrosion cracking. They can be melted in air and having wider range of potential transformation temperatures. Addition of manganese decreases the transformation temperatures of Cu-Zn-Al. It often replaces aluminium giving better ductility. Aging at low temperatures shift the transformation temperatures. The thermal stability of Cu-Zn-Al alloy is generally limited at higher temperatures.

3. Copper-Aluminium-Nickel (CuAlNi) alloys: Copper-aluminum-nickel (CuAlNi) alloys have undergone extensive development and are now preferred to the CuZnAl alloys. The alloys typically contain 11-14.5% Al and 3-5% Ni and have transformation temperatures in the range 80-200°C dependant on their composition, the transformation temperature is particularly sensitive to the aluminum content. The alloy is made from relatively inexpensive elements but its processing is more difficult since it can only be hot worked and the final heat treatment has to be tightly controlled to produce an alloy with the desired transformation temperature. These processing difficulties have made this alloy system more expensive than CuZnAl but it is still less expensive than NiTi. Some improvement of the mechanical properties can be obtained by reducing the aluminum content below 12%, adding 2% manganese to reduce the transformation temperature and 1% titanium as a grain refiner. The major advantages of the CuAlNi system are its wide range of useful transformation temperatures, its stability at elevated temperature making it the only system that can be used for applications above 100°C, its small hysteresis and its relatively low cost.

36.4.8 Applications of Shape Memory Alloys

The shape memory effect and pseudoelasticity of these alloys are being utilized in wide variety of applications in various fields. Products containing SMAs have been around for many years but consumers are often unaware of their presence because they are usually buried in the mechanism that controls the function of the product. One of the applications is in 'indestructible' spectacle frames; these can be bent and twisted to a remarkable extent and then regain their original shape.

The medical and aerospace and marine industries are the largest consumers of shape memory components. Some of their applications are outlined below.

1. Vena-Cava Filters or Blood – clot filters: Vena-cava filter is a device used to trap blood clots. Inserted as a small cylinder, it reverts to an umbrella shaped filter to trap small blood clots and prevent them from travelling to parts of the body where they may have a detrimental effect. Shape memory effect plays a vital role in inserting the filter into the vein. The under cooled SMA is collapsed and inserted into the vein and the body heat recovers its functional shape.

2. Cryofit hydraulic couplings: These couplings are used to join the metal tubings. They are prepared as cylindrical sleeves having diameters slightly less than that of the metal tubings and are fitted in the appropriate position. Upon warming, the martensite changes to austenite structures and the coupling attain its functional shape and strongly holds the ends. But the tube does not allow the SMA coupling to recover its manufactured shape fully. The stress created as the coupling attempts to recover its shape is great enough to create a joint. Thus, recovery in shape is constrained and is being exploited to get superior coupling system. Such couplings are used widely in aerospace industry. The military has been using Nitinol couplers in F-14 fighter planes since the late 1960s. These couplers join hydraulic lines tightly and easily.

3. Superelastic applications: The pseudoelastic properties of SMAs are being used in many applications such as eye glass frames, guide wire for steering catheters into vessels and arch wires for orthodontic correction. In addition to these because of the biocompatibility, SMAs are used as muscle wires too. Moreover, SMAs are used in surgical tool, coffee pots, space shuttle, thermostats, vascular stents and they even have aeronautical applications.

4. Stents for veins: The stent is a device used to treat coronary disease. It would be inserted in the deformed shape and would expand upon reaching body temperature to open arteries and increase blood flow. Thus it is used for reinforcing weak vein walls and for widening narrow veins. The chilled stent is brought into position through a probe, and expands to its original size when warmed up to body temperature. The stent replaces similar stainless steel stents that are expanded with a little balloon.

5. Dental and Orthodontic Archwires: These devices work similar to a spring. They apply a continuous and gentle force correcting misaligned teeth, as opposed to the periodic and uncomfortable tightening required by stainless steels.

6. Other applications in medical field: Tweezers to remove foreign objects through small incisions were invented by NASA. Anchors with Nitinol hooks to attach tendons to bone were used for shoulder surgery. Nitinol eye-glass frames can be bent totally out of shape and return to their parent shape upon warming. Another successful medical application is Nitinol's use as a guide for catheters through blood vessels.

7. Superelastic glasses: These glasses are made from a superelastic metal alloy. Therefore, they can be bent quite drastically without permanent damage. The glasses utilize the superelastic property of Ni-Ti alloys.

8. Actuators and micromanipulators: SMAs are useful for such things as actuators, which are materials that "change shape, stiffness, position, natural frequency, and other mechanical characteristics in response to temperature or electromagnetic fields".

Nitinol is being used in robotics actuators and micromanipulators to simulate human muscle motion. The main advantage of Nitinol is the smooth, controlled force it exerts upon activation.

Nitinol actuators as engine mounts and suspensions can also control vibration. These actuators can help prevent the destruction of such structures as buildings and bridges.

36.4.9 Advantages and Disadvantages of Shape Memory Alloys

Advantages

- Bio-compatibility
- Diverse Fields of Application
- Good Mechanical Properties (strong, corrosion resistant)

Disadvantages

- Relatively expensive
- Have poor fatigue properties
- Technology is still in infant stage.

36.5 BIOMATERIALS

A biomaterial is any material, natural or man-made, that comprises whole or part of a living structure or biomedical device that performs, augments, or replaces a natural function. To be specific, it is a non-viable material used in medical devices intended to interact with biological systems. Biomaterials are used to make devices that would be in close or direct contact with the body to augment or replace faulty materials. Biomaterials have had a major impact on the contemporary medicine and patient care and in improving the quality of lives of humans. They are used in many of today's medical devices, including, artificial skin, artificial blood

Some common medical devices comprised of biomaterials are illustrated here:

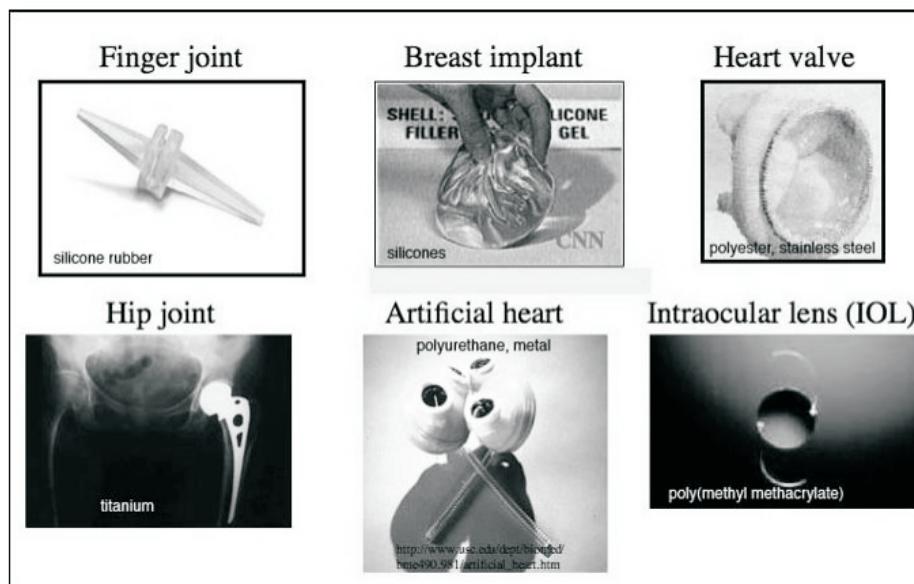


Fig. 36.27

vessels, total artificial hearts, pacemakers, dental fillings, wires, plates and pins for bone repair, total artificial joint replacements. Biomaterials can be metals, ceramics, polymers, glasses, carbons, and composite materials. Such materials are used as molded or machined parts, coatings, fibers, films, foams and fabrics. They are different from other materials in

that they possess a combination of properties, including chemical, mechanical, physical and biological properties that render them suitable for safe, effective and reliable use within a physiological environment. Biomaterials are rarely used on their own but are more commonly integrated into devices or implants. The field of biomaterials uses ideas from medicine, biology, chemistry, materials science and engineering. Engineers play a major role in this field, but they have to work closely with synthetic chemists to optimize materials properties and physicians to ensure that the device is useful in clinical applications.

36.5.1 Criteria

Many different synthetic and modified natural materials are used in biomaterials and some understanding of the criteria they must meet is important.

1. Biocompatibility: Biocompatibility is the ability of a material to perform with an appropriate host response in a specific application. "Appropriate host responses" include lack of blood clotting, resistance to bacterial colonization and normal healing. A hemodialysis (artificial kidney) membrane, a urinary catheter or hip joint prosthesis are examples of applications. It may be noted that the hemodialysis membrane might be in contact with the patient's blood for 3 hours, the catheter may be inserted for a week and the hip joint may be in place for the life of the patient.

2. Toxicity: A biomaterial should not be toxic, unless it is specifically engineered for such requirements.

3. Mechanical and Performance Requirements: Biomaterials and devices have mechanical and performance requirements that originate from the physical (bulk) properties of the material. There are three categories of such requirements: mechanical performance, mechanical durability and physical properties.

First, consider **mechanical performance**. A hip prosthesis must be strong and rigid. A tendon material must be strong and flexible. A heart valve leaflet must be flexible and tough. A dialysis membrane must be strong and flexible, but not elastomeric. An articular cartilage substitute must be soft and elastomeric.

Then, we must consider **mechanical durability**. A catheter may only have to perform for 3 days. A bone plate may fulfill its function in 6 months or longer. A leaflet in a heart valve must flex 60 times per minute without tearing for the lifetime of the patient (realistically, at least for 10 or more years). A hip joint must not fail under heavy loads for more than 10 years.

The bulk physical properties will also determine other aspects of performance. The dialysis membrane has a specified permeability, the articular cup of the hip joint must have high lubricity, and the intraocular lens has clarity and refraction requirements. To meet these requirements, design principles from physics, chemistry, mechanical engineering, chemical engineering, and materials science are invoked.

36.5.2 Metallic Biomaterials

Metals have been used almost exclusively for load-bearing implants, such as hip and knee prostheses and fracture fixation wires, pins, screws and plates. Metals have also been used as parts of artificial heart valves, as vascular stents and as pacemaker leads. For these purposes, alloys are frequently used as they provide improvement in material properties, such as strength and corrosion resistance.

Metallic biomaterials can be divided into three subgroups:

- stainless steels,
- the cobalt-based alloys,
- titanium metals and alloys

These materials have biocompatibility, appropriate mechanical properties, corrosion resistance and reasonable cost. They are very effective in binding the fractured bone, do not corrode and do not release harmful toxins when exposed to body fluids and therefore can be left inside the body for a long period of time.

36.5.3 Ceramic and Glass Biomaterials

Ceramics and glasses are used as components of hip implants, dental implants, middle ear implants, and heart valves. Overall, however, these biomaterials have been used less extensively than either metals or polymers. Some ceramics that have been used for biomedical applications are alumina, zirconia, bioglass, tricalcium phosphate. Although they do not undergo corrosion, ceramics and glasses are susceptible to other forms of degradation when exposed to the physiological environment. The major drawbacks to the use of ceramics and glasses as implants are their brittleness and poor tensile properties. Although they have outstanding strength when loaded in compression, ceramics and glasses fail at low stress when loaded in tension or bending. Only alumina is used as a bearing surface in joint replacements.

36.5.4 Polymeric Biomaterials

Polymers are the most widely used materials in biomedical applications. They are used for making cardiovascular devices and for replacement and augmentation of various soft tissues. Other applications are vascular grafts, heart valves, artificial hearts, breast implants, contact lenses etc. Biodegradable polymers are used for sutures, controlled drug delivery, tissue engineering, and fixture fixation. Compared with metals and ceramics, polymers have much lower strengths and moduli but they can be deformed to a greater extent before failure. Consequently, polymers are generally not used in biomedical applications that bear loads (such as body weight). Ultra – high molecular weight polyethylene is an exception, as it is used as a bearing surface in hip and knee replacements.

Table 1: Example of Biomedical Applications of Polymers

Polymers	Applications
Polyethylene, Polyvinyl chloride, Polyester, Silicone rubber, Polytetrafluoroethylene	Cardovascular implants
Ultra-high-molecular-weight polyethylene, Polymethyl methacrylate	Orthopedic implants
Polylactic acid, Polyglycolic acid	Tissue engineering

36.5.5 Longer-lasting Medical Implants

Currently, medical implants, such as orthopedic implants and heart valves, are made of titanium and stainless steel alloys. These alloys are primarily used in humans because they are bio-compatible, i.e., they do not adversely react with human tissue. In the case of orthopedic implants (artificial bones for hip, etc.), these materials are relatively non-porous. For an implant to effectively mimic a natural human bone, the surrounding tissue must penetrate the implants, thereby affording the implant with the required strength. Since these materials are relatively impervious, human tissue does not penetrate the implants, thereby reducing their effectiveness. Furthermore, these metal alloys wear out quickly necessitating frequent, and often very expensive, surgeries. However, nanocrystalline zirconia (zirconium oxide) ceramic is hard, wear-resistant, corrosion-resistant and bio-compatible. Nanoceramics can also be made porous into aerogels (aerogels can withstand up to 100 times their weight), if

they are synthesized by sol-gel techniques. This results in far less frequent implant replacements, and hence, a significant reduction in surgical expenses. Nanocrystalline silicon carbide (SiC) is used for artificial heart valves primarily due to its low weight, high strength, and extreme hardness, wear resistance, inertness and corrosion resistance.

36.5.6 Titanium and Titanium Alloys

Titanium and some of its alloys are used as biomaterials for dental and orthopaedic applications. The most common grades used are commercially pure titanium and the Ti₆Al₄V alloy, derived from aerospace applications.

These materials are biologically inert biomaterials. Due to their excellent corrosion resistance they remain unchanged when implanted into human bodies. The human body is able to recognize these materials as foreign, and tries to isolate them by encasing them in fibrous tissues. However, they do not produce any adverse reactions and are tolerated well by the human body. Furthermore, they do not induce allergic reactions such as has been observed on occasion with some stainless steels, which have induced nickel hypersensitivity in surrounding tissues.

The favorable performance of titanium implants is attributed to the chemically very stable oxide film that forms on the metal surface. The oxide film protects the underlying metal from corrosion. The dielectric constant ϵ_r of titanium oxide is close to that of water ($\epsilon_r = 78$). The result is that the titanium surface does not lead to excessively strong interaction with proteins in the extracellular matrix; rather the surface is in some way ‘waterlike’, interacting gently with the hydrophilic outer surface of the protein molecules.

The surface of titanium is often modified by coating it with hydroxyapatite. The hydroxyapatite provides a bioactive surface which actively participates in bone bonding, such that bone cements and other mechanical fixation devices are often not required.

Titanium and its alloys possess suitable mechanical properties such as strength, bend strength and fatigue resistance to be used in orthopaedics and dental applications. Due to the mechanical properties of pure titanium, its use in implants is restricted to applications which involve moderate mechanical stress, such as dental implants. In applications where high mechanical strength is necessary, like orthopedic implants, it is appropriate to employ titanium-based alloys, which have better properties than pure titanium.

Other specific properties that make titanium a desirable biomaterial are density and elastic modulus. In terms of density, it has a significantly lower density than other metallic biomaterials, meaning that the implants will be lighter than similar items fabricated out of stainless steel or cobalt chrome alloys.

Ti-6Al-4V alloy has been used for decades as an implant material due to its excellent properties. The material was originally developed as a high temperature aerospace alloy. However, vanadium is a potentially toxic element and as an alternative the Ti-6Al-7Nb alpha-beta alloy, Protasul-100 was developed between 1978 and 1982, and proved to be highly biocompatible.

QUESTIONS

1. Distinguish between an amorphous and a crystalline solid.
2. What do you mean by metallic glasses? (G.T.U., 2009), (Anna Univ., 2003)
3. What are metallic glasses? Why is it easier to form metallic glass from alloys than from pure metals?
4. What is meant by glass transition temperature?
5. Explain briefly two important techniques that are used to produce metallic glass.
6. State the salient properties of metallic glasses. (Anna Univ., 2005)
7. Explain the electrical properties of metallic glasses.
8. What are the types of metallic glasses? Mention a few metallic glasses. (Anna Univ., 2005)
9. Discuss the properties, types and applications of metallic glasses. (G.T.U., 2009)
10. How does the hysteresis of a metallic glass differ from that of a crystalline material?
11. Why are metallic glasses used in transformers?
12. What is the advantage of using metallic glasses as transformer core material? (Anna Univ., 2005)
13. What are metallic glasses? How are they prepared? Explain their properties and applications. (Anna Univ., 2003)
14. What are the applications of metallic glasses? (Anna Univ., 2005)
15. What are shape memory alloys? (Anna Univ., 2006)
16. Define shape memory effect. (Anna Univ., 2004)
17. Explain briefly the main features of shape memory alloys.
18. Explain one-way shape memory effect.
19. Explain two-ways shape memory effect.
20. Explain pseudoelasticity.
21. What are the properties of shape memory alloys? (Anna Univ., 2004)
22. What are the applications of shape memory alloys? (Anna Univ., 2003)
23. What are shape memory alloys? Write their characteristics. List out any four applications of shape memory alloys. (Anna Univ., 2002)
24. What are liquid crystals? How are they different from crystals and liquids?
25. Explain positional and orientational order in liquid crystals.
26. How are the liquid crystals classified?
27. What are liquid crystals? How do they differ from crystalline state and liquid state? Give applications of liquid crystals. (Univ. of Pune, 2007)
28. What is a twisted nematic display? How is it made?
29. Explain the construction and working of a segment LCD.
30. What is an LCD? Explain with neat sketch the twisted nematic display. (Calicut Univ., 2005)
31. What are matrix displays and how are they classified?
32. Explain the construction and working of a passive matrix LCD.
33. Explain the construction and working of an active matrix LCD.
34. Explain the construction and working of a TFT LCD.
35. What are biomaterials? State some of their applications. (Anna Univ., 2007)
36. What are the types of biomaterials? Mention a few biomaterials and their applications. (Anna Univ., 2005)
37. What are the criteria that the biomaterials have to satisfy for being used in devices?
38. Classify biomaterials and discuss their applications.
39. Explain biomaterials and their modern applications in the field of medicine. (Anna Univ., 2005)
40. Describe the salient properties of titanium and its alloys that make them superior biomaterials.

CHAPTER

37

Non Destructive Testing

37.1 INTRODUCTION

An industrial product is manufactured with an aim to perform a certain function and the user expects that it performs the assigned function well. However, the product may contain certain defects and imperfections, which impair its performance level. **Non Destructive Testing** (NDT) refers to the entire range of test methods that detect the harmful defects in the finished products without affecting their future usefulness.

Defects in a product can arise during manufacturing stage, or during assembly, installation, commissioning or during in-service. In the pre-service stage, the defects may be present in the raw material or may be introduced during machining, fabrication, heat treatment, assembling. The pre-service quality can be achieved essentially by good engineering practice i.e. by way of selecting suitable quality raw materials and by ensuring that harmful defects are not produced during the subsequent stages of fabrication and assembly, prior to putting the part/component into service.

However, even with the highest quality of materials and workmanship, the occurrence of some form of imperfections during manufacture is inevitable and there will be a typical distribution of imperfection sizes associated with a particular manufacturing process and quality. In the in-service stage, defects will be generated due to deterioration of the component/structure as a result of one or combination of the operating conditions like elevated temperature, pressure, stress, hostile chemical environment etc. **The goal of NDT is to detect the defects and give information about their distribution.** Therefore, NDT plays a vital role in modern engineering practice for achieving the required standards of quality in manufacturing.

37.2 TYPES OF DEFECTS

The defects that are found in manufactured components may be broadly divided into two types, namely external defects and internal defects. Some of the commonly occurring defects are as follows:

External Defects

- Blow holes in castings
- Forging defects
- Metallic inclusions on surfaces
- Surfaces finish
- Micro-cracks

Internal Defects

- Overstressed parts
- Fatigue blow holes
- Porosity in casting
- Micro-segregation
- Axial segregation

37.3 METHODS OF NDT

There are a large number of nondestructive tests in use. We study here are some of the most common methods. The common methods are

1. Visual Inspection
2. Liquid Penetrant Testing
3. Magnetic Particle Testing
4. Eddy Current Testing
5. Ultrasonic Testing
6. Radiography

Depending on the requirements, the above tests are used singly or in conjunction with one another.

37.4 VISUAL INSPECTION

Visual NDT is mainly used for examination and detection of surface defects. In this, the test component is first adequately cleaned. Then, it is illuminated with light and examined with unaided eye or with the help of optical devices. Visual inspection can reveal the information on

- (i) general condition of the component
- (ii) presence or absence of corrosive product on the surface
- (iii) presence or absence of cracks
- (iv) orientation and position of cracks, if present
- (v) surface porosity etc.

In many cases, the results of visual examination will be of great assistance to other tests.

Magnifying devices or lighting aids are used wherever their necessity is felt. Microscopes, borescopes, endoscopes and flexiscopes are some of the optical aids used in the visual testing. A microscope detects minute defects and details of fine structure. A *borescope* is an instrument used to inspect the inside of a narrow tube, bore, or chamber. The *endoscope* is much like a borescope with a superior optical system and high intensity light source. The *flexiscope* is a flexible fibre-optic borescope that can be used to inspect inaccessible regions and through passages with several bends. In recent times, holography is used to inspect surfaces of precision components.

37.5 LIQUID/DYE PENETRANT TESTING

This method is used for detecting minute discontinuities such as cracks and surface openings. **Liquid penetrant testing** can be applied to any non-porous clean material, metallic or non-metallic, but is unsuitable for dirty or very rough surfaces. Though it is applicable for both magnetic and non-magnetic materials, primarily it is applied for non-magnetic materials. It is simple to perform and relatively cheaper than other testing methods. Cracks, as narrow as 150 nanometers can be detected with this method.

37.5.1 Basic Principle

The basic principle of penetrant inspection is that when a liquid penetrant is applied over a clean surface to be inspected, the penetrant seeps into defect by the combined action of surface tension and capillary action; when developed, by the blotting action of the developer powder, gives a recognizable indication of the defect (a crack or surface opening etc).

In this method, a liquid penetrant is applied to the surface of a component for a certain predetermined time. The penetrant seeps through any surface opening defect by capillary action. Subsequently, the excess penetrant is removed from the surface. The surface is then

dried and a developer is applied to it. The developer absorbs the penetrant that remained in the discontinuity and indicates the presence as well as the location, size and nature of the discontinuity.

37.5.2 Basic Processing Steps of Liquid Penetrant Inspection

The following are the basic stages of liquid penetrant method:

1. Surface Preparation: One of the most important steps of a liquid penetrant inspection is the initial cleaning and drying of the surface area of the test component. The surface must be free of oil, grease, water, or other contaminants that may prevent penetrant from entering flaws.

2. Penetrant Application: Once the surface has been thoroughly cleaned and dried, the penetrant material is applied by spraying, brushing, etc methods. The liquid should spread uniformly over the surface and seep into the crack.

3. Penetrant Dwell: The penetrant is left on the surface for a sufficient time to allow as much penetrant as possible to seep into a defect. Penetrant dwell time is the total time that the penetrant is in contact with the part surface. The penetrant producers usually recommend dwell times. The times vary depending on the application, penetrant materials used, the component being inspected, and the type of defect being inspected. Minimum dwell times typically range from 5 to 60 minutes.

4. Removal of Excess Penetrant:
The excess penetrant is then carefully removed from the surface of the sample. Insufficient cleaning leaves penetrant on the surface

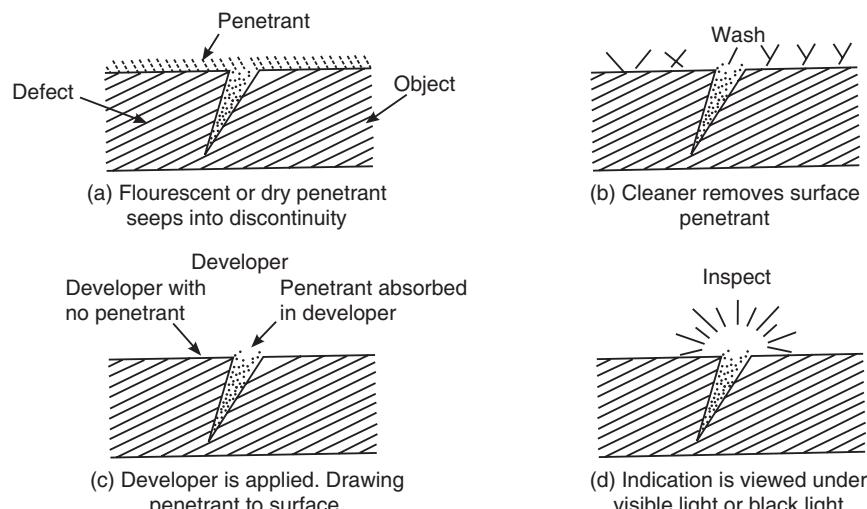


Fig. 37.1

and leads to overlooking of the defect. On the other hand over cleaning may remove penetrant from defects. Therefore, both insufficient cleaning and over cleaning must be avoided. The cleaning of the surface is carried out using a solvent and then rinsing with water, or first treating with an emulsifier and then rinsing with water.

5. Developer Application: In the next step, a thin layer of developer is applied to the surface. The developer acts as a blotter and draws trapped penetrant out of imperfections open to the surface and makes them visible. Developers come in a variety of forms that may be applied by dusting (dry powdered), dipping, or spraying (wet developers).

6. Indication Development: The developer is allowed to stand on the part surface for a period of time sufficient to permit the extraction of the trapped penetrant out of any surface flaws. This development time is usually a minimum of 10 minutes and significantly longer times may be necessary for tight cracks.

7. Inspection: Inspection is then performed under appropriate lighting to detect indications from any flaws, which may be present.

8. Clean Surface: The final step in the process is to thoroughly clean the part surface to remove the developer from the parts that were found to be acceptable.

The sequence of steps in liquid penetrant testing is shown in Fig. 37.1.

37.5.3 Materials used in Penetrant Testing

A variety of materials are used in penetrant testing method. They are penetrant materials, cleaners, emulsifiers and developer materials.

Penetrants

The penetrant materials are light, oil-like liquids, which have dissolved coloured dyes and may be classified as fluorescent, visible dye or dual-purpose materials. The role of the dye is to give colour contrast with respect to the surrounding background under white light or UV light illumination so that the defect becomes clearly visible. The penetrant materials may also be classified as (i) water-washable, (ii) solvent removable, and (iii) post-emulsifiable. Penetrants that are water washable can be removed from the surface by ordinary tap water. Other penetrants are removed with special solvents. For petroleum-based penetrants, an emulsifier is used that reacts with the oil-based penetrant to form a water-soluble substance, which is then washed away by water washing. The penetrant must be chemically stable, have low viscosity, chemical inertness, ease of removal, and sufficient brightness of colour.

Developers

The developer is required to enhance the conspicuity of the indication. Two types of developer materials are used in the technique: (i) dry developers and (ii) wet developers. Dry developers are dry light coloured powdery materials such as amorphous silica powder. Wet developers consist of a powdered material suspended in a suitable liquid. The developer must be highly absorptive to draw penetrant from the defect, provide a contrast background to indicate the location of the defect, and must be easy to apply.

37.5.4 Advantages and Limitations

The advantages of liquid penetrant testing method are as follows.

- No elaborate setup is required.
- It is simple to apply and cheaper in cost.
- It is portable and fast.
- Results are easy to interpret.
- It can be used to inspect any non-porous material.
- Defects of any size, shape and orientation can be detected using this method.

The limitations are that

- It can be used to detect surface-breaking defects only.
- Surfaces must be clean
- It cannot be applied to porous materials.

37.6 MAGNETIC PARTICLE TESTING

This is a method of detecting the presence of cracks, inclusions, and other similar discontinuities in ferromagnetic materials such as iron and steel. The method can be used to detect surface and subsurface. It is not applicable to nonmagnetic materials.

When a specimen is magnetized, the magnetic lines of force flow inside the ferromagnetic material. Wherever there is a flaw, magnetic lines of force are interrupted and some

of the lines leak out of the specimen. The points of exit and reentry of leakage lines form opposite magnetic poles. If minute magnetic particles are sprinkled onto the specimen, the particles are attracted by the magnetic poles and give a visual indication of the size and shape of the defect.

37.7 EDDY CURRENT TESTING

Eddy current testing is used to inspect electrically conducting materials for defects, irregularities in structure, and variations in composition. In eddy current testing, a varying magnetic field is produced if a source of alternating current is connected to a coil. When this field is placed near a conducting test specimen, eddy currents will be induced in the test specimen. The eddy currents, in turn will produce a magnetic field of their own. The detector measures the new magnetic field and converts the signal into a voltage that can be displayed on a CRO. This test will detect surface and sub-surface defects.

37.8 ULTRASONIC INSPECTION METHOD

Ultrasonic testing is a versatile and widely used NDT method. It was suggested by Sokolovin in 1935 and applied by Firestonein in 1940, and was further developed subsequently.

Principle: The principle of ultrasonic testing is based on the fact that solid materials are good conductors of sound waves. The waves are not only reflected at the interfaces but also by internal flaws (material separations, inclusions etc.). The interaction effect of sound waves with the material is stronger the smaller the wavelength, or the higher the frequency of the wave. Therefore, ultrasonic waves of high frequencies with frequency ranging between about 0.5 MHz and 25 MHz and having wavelength in mm are generally used. With lower frequencies, the interaction effect of the waves with internal flaws would be so small that detection becomes questionable.

A typical UT inspection system consists of several functional units, such as the pulse transmitter, transducer, and display devices.

The ultrasonic waves are generated with the help of piezoelectric devices. A pulse transmitter is an electronic device that can produce high voltage electrical pulses. When these bursts of alternating voltage are applied to the transducer, the transducer emits high frequency ultrasonic energy. The ultrasonic beam is then transmitted from the transducer into the specimen under testing which propagates through the specimen in the form of waves. When there is a discontinuity (such as a crack) in the wave path, part of the energy will be reflected back from the flaw surface. The reflected wave signal is transformed into an electrical signal by the transducer. This signal is displayed on a screen of CRT. The characteristics of the pulses produced by the transducer are used for interpretation of the nature of the defect in the specimen.

37.8.1 Transducer Probes

Piezoelectric transducers generate ultrasonic waves. They convert high frequency electrical signals into mechanical vibrations. The three common piezoelectric materials used in probes are quartz, lithium sulphate and ceramics such as barium titanate and lead zirconate titanate. Different types of transducer probes are in use. We describe here three probes, namely normal beam probe, angle beam probe and transmitter-receiver probe.

- (i) **Normal beam probe:** The schematic of a normal beam probe is shown in Fig. 37.2
 - (a). Electrodes are fitted on both faces of the piezoelectric crystal and wires from these electrodes lead to the connector socket of the probe. A damping material is

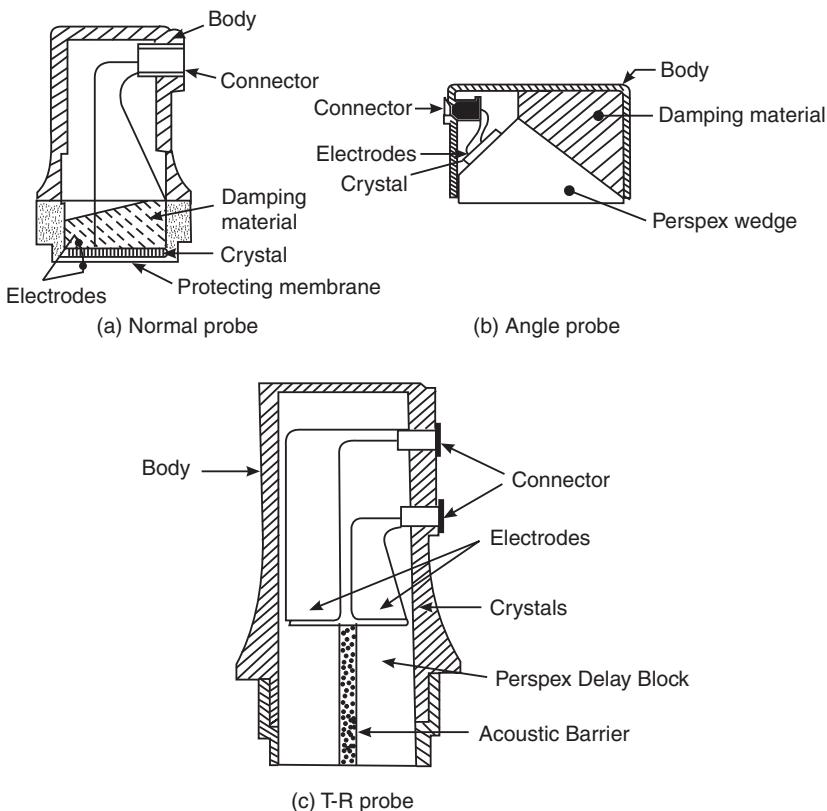


Fig. 37.2

fitted at the back of the crystal. A wear plate is attached to the outer face of the crystal, which protects the crystal from wear and tear. The entire assembly is mounted in a casing, which protects the transducer from mechanical damage.

- (ii) **Angle beam probe:** In an angle beam probe, the transducer is mounted on a perspex wedge, as shown in Fig. 37.2 (b). The ultrasonic beam reaches the bottom flat surface of the probe at an angle and refracts into the material to be tested. The part of the beam that is reflected from the perspex wedge is damped with a suitable damping material.
- (iii) **Transmitter-receiver probe:** Transmitter-receiver probe consists of separate transmitter and receiver incorporated in a single housing. The transmitter and receiver transducers are of the same frequency. They are arranged such that they are inclined slightly towards each other (Fig. 37.2 c). Delay blocks of perspex are added to both the transducers and an acoustic barrier is kept between them to prevent any cross-talk.

37.8.2 Coupling Materials

Acoustic impedance mismatch occurs if the transducer probe is directly placed on the test object, because of the air layer between the probe and the object. A major portion of the energy is reflected back and very less energy gets transmitted into the test specimen. Therefore, a thin film of oil, or glycerin is applied to the surface of the object before placing the transducer on it. Then, the ultrasonic energy is transmitted into the test piece more effectively. Thus, oil or glycerin film acts as a couplant to couple ultrasonic energy from the probe to the test object.

37.8.3 Ultrasonic Testing Methods

Several methods have been developed for the ultrasonic testing. Among them, pulse-echo methods are most popular and widely used. These methods can be divided into two broad testing techniques, namely contact method and immersion method.

In the **contact method**, the transducer is placed in direct contact with the test object (see Fig. 37.3). A very thin film of suitable coupling material is applied between the probe and the test object for maximum transfer of ultrasonic energy into the object under testing. This method can be used for bigger test objects.

In the **immersion method**, the probe head and test object are both kept immersed in water. The probes used for immersion testing are therefore made waterproof. The probe head is kept at a distance from the face of the test object (see Fig. 37.4).

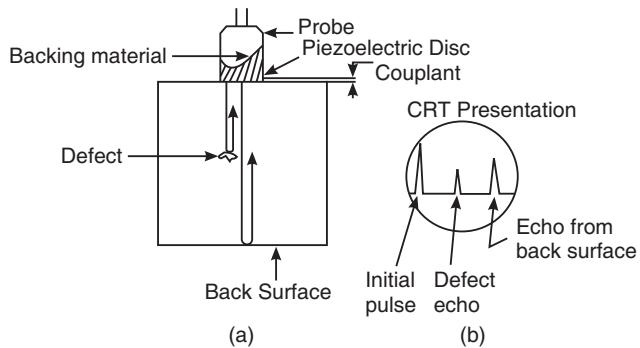


Fig. 37.3

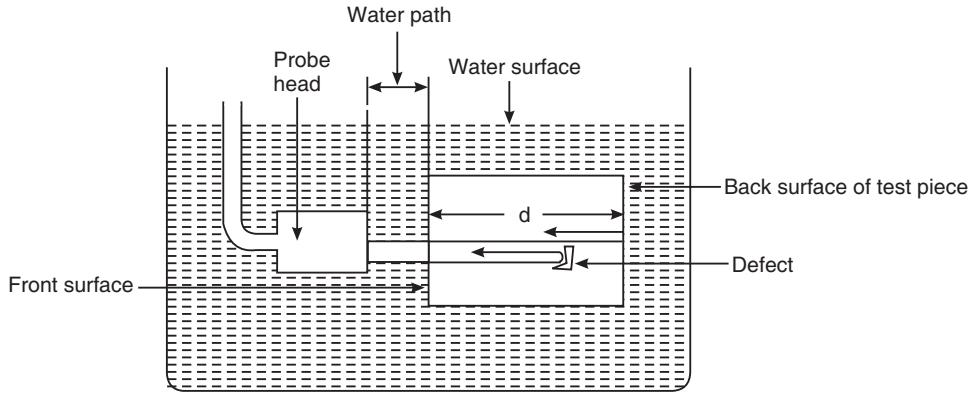


Fig. 37.4

The above methods may again be grouped into three common testing methods. Two of them utilize normal beam probes and are known as (i) *Normal beam pulse-echo testing* and (ii) *Normal beam pulse through-transmission testing*, respectively. In these methods, normal beam probes are used in which a transducer crystal is fixed parallel to the bottom plate of the probe. The

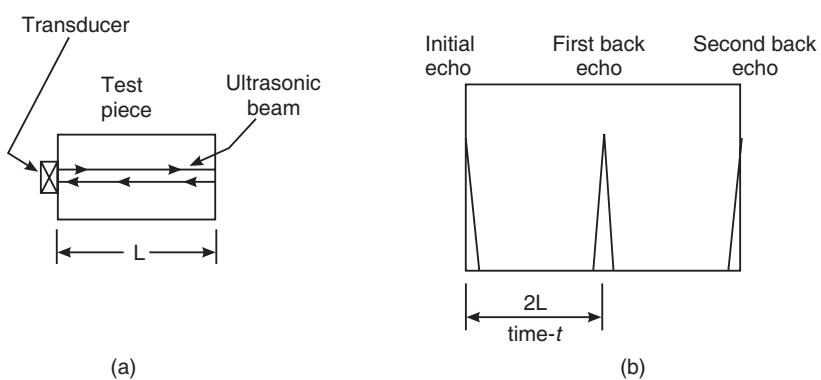


Fig. 37.5

ultrasonic beam produced by the probe propagates into the object perpendicular to the surface of contact and travel in the material in the form of longitudinal waves.

37.8.3.1 Normal beam pulse-echo testing

The pulse echo method may use either a single crystal unit or two crystal units. In single transducer method, only one probe is used which acts both as transmitter and receiver (Fig. 37.5 a). The ultrasonic beam is incident normally on the surface of the specimen under test. The beam travels through the specimen and is reflected back from the rear surface of the specimen. If a defect is present in its path, part of the energy will be reflected back from the defect and the remaining part travels forward in the material. The initial pulse, the echo pulse reflected from the defect and the echo pulse reflected from the opposite face of the specimen are displayed on a CRO screen (see Fig. 37.5 b). Since the pips on the oscilloscope screen measure the elapsed time between reflection of the pulse from the front and back surfaces, the distance between the pips is a measure of the thickness of the specimen (Fig. 37.5 b). The location of a defect may therefore be accurately determined from the pips on the screen.

In case of two transducers method, one transducer acts as transmitter and the other as receiver. Both the units are placed on the same side of the specimen. The transmitter sends an ultrasonic pulse into the specimen. The echoes reflected by any defects and from the back surface of the specimen are received by the receiver and are displayed on CRO screen.

37.8.3.2 Normal beam pulse through-transmission testing

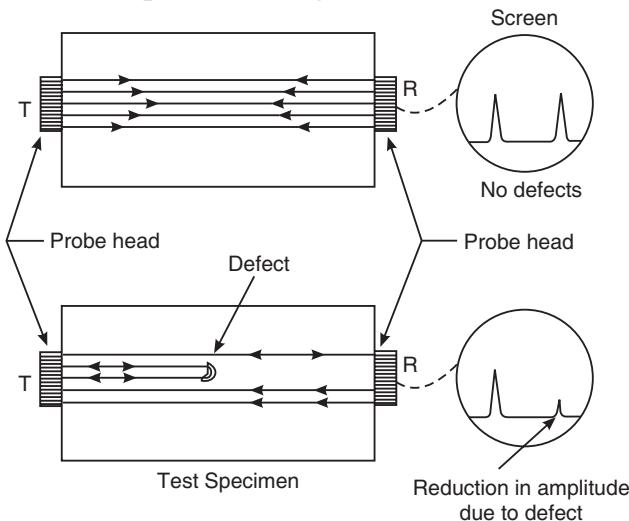


Fig. 37.6

In certain cases, the pulse-echo technique may not provide required information. It happens whenever a defect does not provide a suitable reflection surface or whenever its orientation is not favourable for detection. In such cases, the through-transmission method is adopted. The method uses two ultrasonic transducers on each side of the specimen being inspected. Fig. 37.6 and Fig. 37.7 show the arrangements in contact and immersion methods respectively. If an electrical pulse of the desired frequency is applied to the transmitting transducer, ultrasonic waves are produced. They travel through the specimen to the other side. The receiver transducer on the opposite side receives the vibrations and converts them into an electrical signal, which may be displayed on an oscilloscope. If the ultrasonic pulse travels

through specimen without encountering any defect, the signal received will be relatively large. If there is a defect in the path of the ultrasonic beam, part of the energy is reflected and hence the signal received at the opposite end will be reduced (see Fig. 37.6).

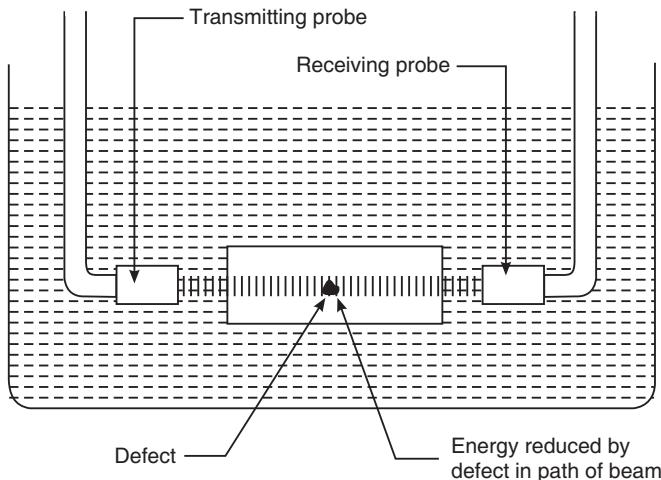


Fig. 37.7

37.8.3.3 Angle beam pulse echo testing

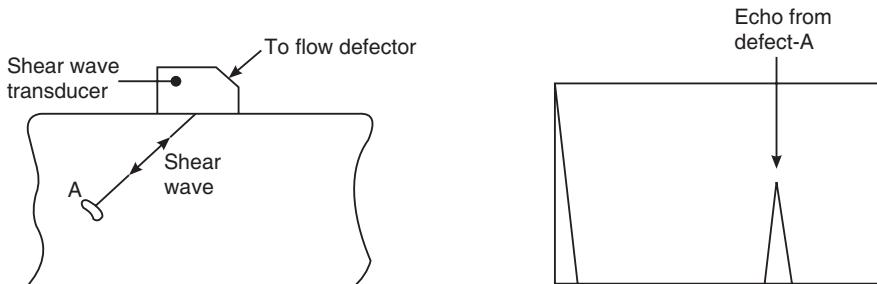


Fig. 37.8

If an angle beam probe is employed for inspection, the ultrasonic beam enters the specimen at an oblique incidence. Therefore, the method is known as *angle beam pulse echo testing*. The ultrasonic beam produced by the probe propagates in the material in the form of shear waves (transverse waves).

Angle beam transducers provide access to areas that are inaccessible to normal beam probes. Fig. 37.8 shows the test arrangement. This method of testing is generally used for the inspection of sheet or plate, pipe or tubing and test pieces having shapes that prevent access for normal beam.

37.8.4 Pulse Echo System

Of the several methods developed for the ultrasonic testing, pulse-echo method is one of the most popular and widely used methods.

Fig. 37.9 shows the block diagram of ultrasonic pulse echo testing equipment. The essential units in the equipment are the pulse transmitter, clock or timer, receiver amplifier and cathode ray oscilloscope. The time base generator used in the CRO here differs from that

used in an ordinary laboratory oscilloscope. After the end of a sawtooth wave, the next cycle starts only after about 4 to 5 times the sweep time; thus it waits till the reverberations in the specimen diminish.

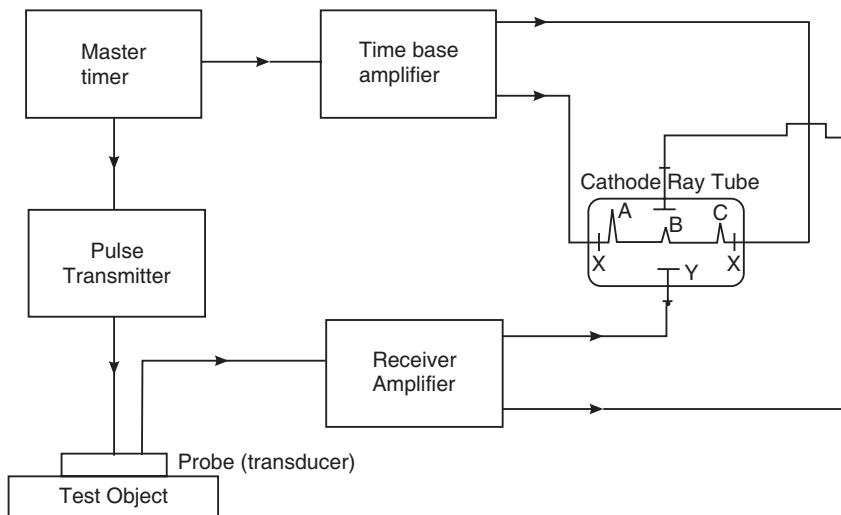


Fig. 37.9. Ultrasonic pulse echo testing equipment

From the transmitter, the electric pulse is fed to the transducer probe. The piezoelectric transducer is excited by the electric pulse and vibrates at its resonant frequency. It produces a short ultrasonic pulse, which is propagated into the test object through the couplant layer. The electric pulse also triggers the time-base generator, so that the pulse of ultrasound starts to move through the object at the same time as the luminous spot moves across the CRO screen. The output of the transducer is applied to the Y-plates of CRO through an amplifier. It produces a transmission signal (A), which represents the initial ultrasonic pulse (see Fig. 37.9). The luminous spot continues to move across the screen, as the ultrasonic pulse travels through the object. If the ultrasonic pulse encounters a defect, part of the energy is reflected back from the defect surface. The reflected part of the ultrasound returns to the transducer. Under the action of this reflected energy, the transducer vibrates and produces a small voltage pulse. This induced voltage is fed to the Y-plates of CRO through the amplifier and produces signal (B), the echo pulse from the defect. The ultrasonic energy in the transmitted pulse travels further to the bottom surface of the object and gets reflected there back to the transducer. The transducer produces an electric voltage pulse (C), which is much smaller than the transmitted pulse (A). The signal (C) constitutes the bottom surface echo.

As the Y-axis of the display on CRO screen represents time, one can determine the location of the defect in the object. The time interval between the transmitted pulse and echo pulse from the defect equals the time, taken by the energy to travel from the transducer to the defect surface and back to the transducer. If v is the velocity of ultrasonic waves in the material, then the distance, d of the defect from the top surface of the object is given by

$$d = \frac{vt}{2} \quad (37.1)$$

37.8.5 Data Presentation

Ultrasonic data can be collected and displayed in a number of different formats. The three most common formats are known as **A-scan**, **B-scan** and **C-scan** presentations. Each presentation

mode provides a different way of looking at and evaluating the region of material being inspected. Modern computerized ultrasonic scanning systems can display data in all three presentation forms simultaneously.

A-scan presentation: A-scan display is the most used mode of display in ultrasonic testing. In this mode of display, the X-axis represents time taken by the pulse to the reflecting surface and return back to the transducer. Y-axis represents the amplitude of the echoes. The location of the defect is estimated by the position of the echo given by it on the horizontal axis and size of the defect from the relative amplitude of the echo. The information that is available in A-scan is one-dimensional. A typical A-scan echo pattern is shown in Fig. 37.10.

B-scan presentation: B-scan display gives a cross-sectional view of the test object and shows the position, orientation and depth of defects in the specimen. In this mode of display, Y-axis represents elapsed time while X-axis represents the position of the transducer along a line on the surface of the test object relative to the starting position of the transducer. Thus, the probe movement

is displayed in x-direction while the distance of the defect is displayed in y-direction. Echo amplitude is indicated by the relative brightness of echo indications. If a storage oscilloscope is used, the whole picture will be displayed, which reveals the depth of the defect beneath the surface and its size in the lateral direction (see Fig. 37.11).

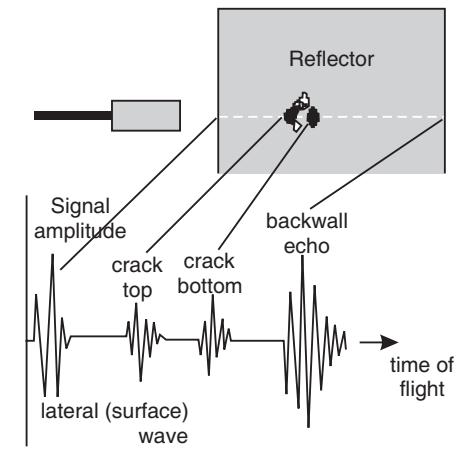


Fig. 37.10. A-scan presentation

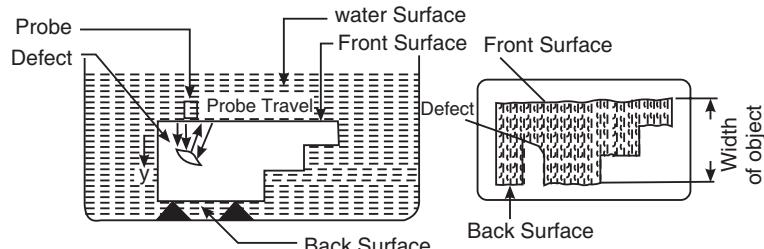


Fig. 37.11. B-scan presentation

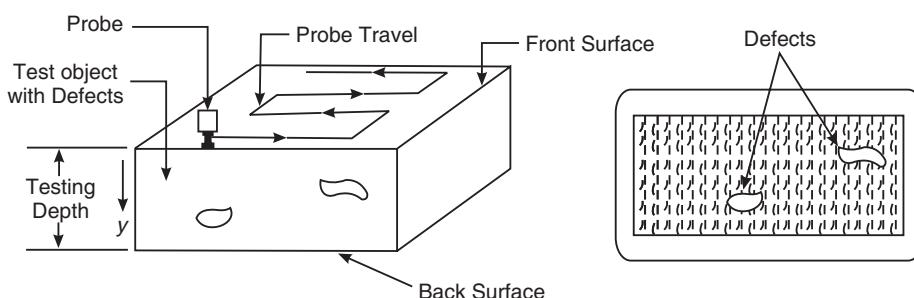


Fig. 37.12. C-scan presentation

C-scan: The depth of defects is not relevant in some testing problems, but information about their distribution parallel to the test surface is required. In the C-scan mode, the transducer is moved over the surface of the test piece and the echo intensity is recorded as

a variation in line shading (Fig. 37.12). The image shows the plan of the object as viewed from the top and is a true-to-scale reproduction of the defect in the object. C-scan presentations are produced with an automated data acquisition system, such as a computer controlled immersion scanning system.

37.9 ADVANTAGES

1. The ultrasonic test is fast and dependable.
2. It is easy to operate and also lends itself to automation.
3. Results of tests are immediately known.
4. It can detect both surface and subsurface defects.
5. Relatively portable equipment.
6. No special safety precautions are needed.
7. High-sensitivity which helps in detection of minute defects.
8. High penetrating power so that objects of very large thickness can be tested.
9. High accuracy in finding the location, size and shape of the defects.
10. Single side access is sufficient for conducting the test.
11. It can also be used to find out the thickness of the object.

37.9.1 Limitations

1. It is not suited for testing of objects that are rough and irregular in shape, very small parts, very thin and inhomogeneous objects.
2. Penetration is not good in coarse structure materials like castings and stainless steel.
3. Line shaped defects lying parallel to the sound beam may escape detection.
4. Interpretation of readings requires good technical knowledge.

37.10 X-RAY RADIOGRAPHY

X-ray radiography is a nondestructive technique of detecting the presence and location of internal cavities or other discontinuities in materials. It is based on the principle of dissimilar transmission of X-rays through different materials and utilized to create an image of various contrasts. Radiation is directed through a part and onto film or other imaging media. Defects are indicated as density changes on the film (Fig. 37.13) in the same manner as medical X-ray shows broken bones. X-ray examination of materials may be carried out by processes known as radiography and fluoroscopy.

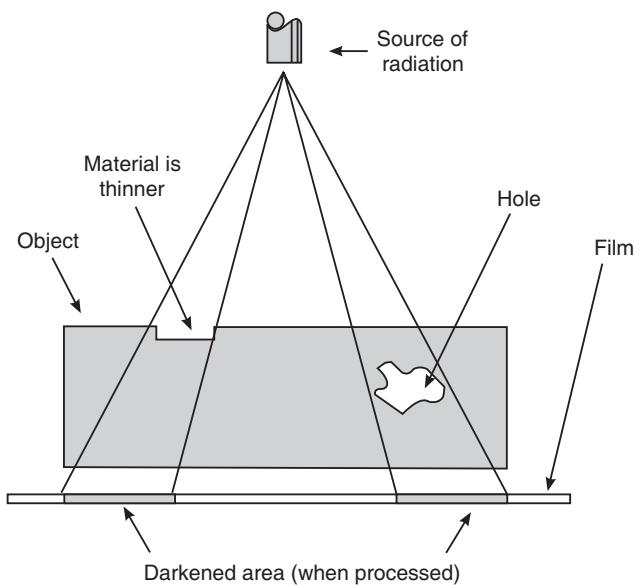


Fig. 37.13

They differ in the manner in which the radiation emerging from the test object is detected or recorded.

If a beam of *X*-rays is directed normal to the surface of an object and if the object is of uniform thickness, the transmitted beam will be of equal intensity. However, if the object contains a cavity, groove or a similar discontinuity, a variation in *X*-ray transmission will result at these points. As *X*-rays are invisible, the effect of differential absorption caused by discontinuities has to be shown by recording the transmitted beam on a photosensitive film. The image of a cavity inside the object will appear darker than the surrounding area because of the greater *X*-ray transmission at this point. Thus, radiography is based on the principle of shadow projection and such a shadow picture is called a **radiograph**. The position or depth of the cavity from the surface has negligible effect upon the appearance of the *X*-ray image.

The basic setup essentially consists of a source of radiation, the object to be radiographed and a detector that is a photographic film.

37.10.1 X-ray Source

X-rays are produced when fast moving electrons are suddenly brought to rest by colliding with matter. The source of *X*-rays is an *X*-ray tube (see Fig. 37.14). The *X*-ray tube consists of a glass bulb under vacuum enclosing an anode and a cathode. The cathode is a filament which, when heated by a current, emits electrons. The filament is located in a recess in the cathode, called the focusing cup, which helps to produce a narrow, well-defined beam of electrons. The electrons are accelerated towards the anode under the action of the high voltage on the anode. The target is embedded in the anode and when the electron beam strikes the target, *X*-rays are produced. At the anode, only a small portion (1 to 10%) of the energy of the electrons is converted to *X*-rays and the rest is converted into heat. The target is usually a tungsten disc, which has a very high melting point. The anode is made of copper, which has high thermal conductivity, and a coolant such as water or oil further cools it. The anode face is angled and is hooded to spread and restrict the *X*-ray beam appropriately. The effective width of the source of *X*-rays is considerably smaller than the area of the target and

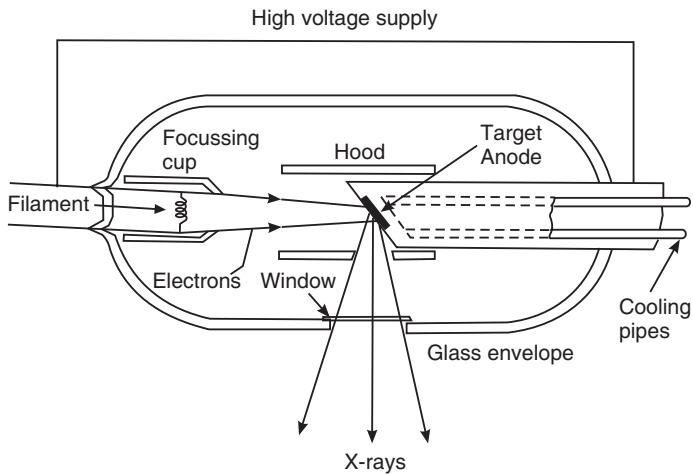


Fig. 37.14

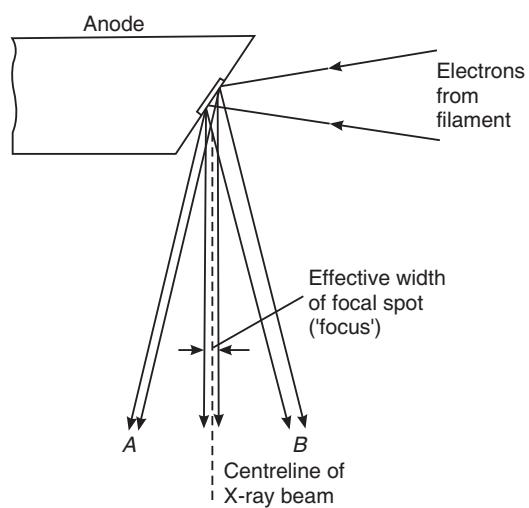


Fig. 37.15

is about 1 to 5 mm. This effective width is usually called *focal spot size* or *focus size* (see Fig. 37.15).

37.10.2 Photographic Film

A radiographic film is very much similar to ordinary photographic film. It consists of a base that is made of thin transparent plastic sheet coated with a recording medium. The recording medium is an emulsion of a silver halide with gelatin binder. It is applied to both the sides of the base. A protective layer covers the emulsion. When radiation strikes the emulsion, a change takes place in the emulsion. The latent image on the film is then developed to give a visible image that is a two-dimensional shadow picture of the object.

37.10.3 Displacement method

Fig. 37.13 is the schematic of a single exposure method. In this method, a beam of X-rays irradiates the test object and the portion of the radiation that is not absorbed by the object is transmitted and is incident on a sheet of photographic film. Absorption of radiation is less in the region of the defect as the density of the material at that place is less. When the film is developed, the defect is seen as a darker spot in the image. This single exposure method indicates the presence of the defect but its depth in the material cannot be calculated.

The depth of the defect in the interior of the object can be determined if we take two exposures from two different positions by moving the X-ray tube. This is known as displacement method. The schematic of the displacement method is shown in Fig. 37.16. If one of the exposures is obtained from position A and then from position B, the displacement of the tube is the distance l . Let the distance of the photographic film from the source be L and x be the distance between the images of the defect on the photographic film. Then, the location of the flaw from the bottom surface of the object, d , is given by

$$d = \frac{L \cdot x}{(x + 1)} - H \quad (37.2)$$

where H is the height of the object from the film.

37.10.4 Advantages

1. Provides permanent record.
2. Works well on thin sections.
3. Highly sensitive.

37.10.5 Limitations

1. Compared to other NDT methods, radiography is expensive.
2. Trained technicians are required to carryout the radiographic inspection.
3. Cracks cannot be detected unless they are parallel to the radiation beam.
4. Minute discontinuities such as inclusions, microporosity, microfissures etc cannot be detected unless they are sufficiently large in size.
5. Inspection of thick specimens is a time consuming process.
6. It is almost impossible to detect lamination type defects in metals.
7. Safety precautions are to be taken in the use of radiographic testing.

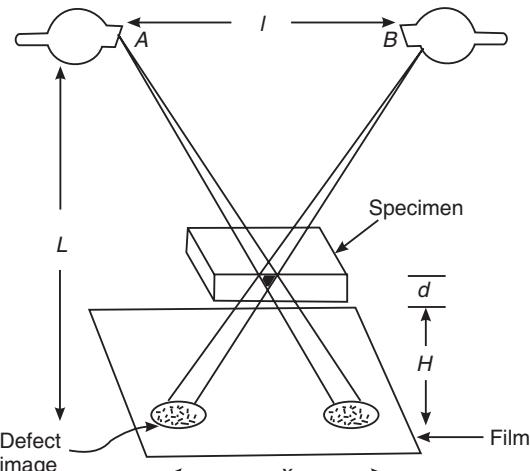


Fig. 37.16

37.11 X-RAY FLUOROSCOPY

If a fluorescent screen is held behind the test object instead of a photographic film, the X-rays are converted into visible light and an image of the object can be seen on this screen. This technique is known as **fluoroscopy**. The image obtained on the fluorescent screen is usually very faint and it is difficult to ascertain the details of defects. However, using a suitably sensitive closed-circuit television (CCTV) camera focused on the fluorescent screen, a brighter image can be obtained on the television monitor screen. Since the image is produced on a television monitor, which can be kept remote from the X-ray equipment, the radiation hazards can be eliminated.

The television image can be converted into digital data and using digital image processing technique, image enhancement can be achieved. This method is known as **real time radiography**. Nowadays the term fluoroscopy is used for the real-time radiography. A basic system is shown in Fig. 37.17. It consists of a source of radiation, X-ray image intensifier, video camera, and computer with image processing arrangement and a display monitor. The X-rays from the source pass through the test object and strike the double layer screen in the image intensifier. The fluorescent layer of the screen converts the x-rays incident on it into light. The adjoining photo-cathode layer generates an electron image of the visible image produced by the fluorescent layer. The electrons from the double layer input screen are accelerated and focused on to a smaller output screen. The output screen reconverts the electron image into visible image. Thus, the dim image from the input screen is made 100 to 10,000 times more intense. The video camera views the image, which is converted into digital data by the analogue-to-digital converter. The computer processes the digital output and the real time image of the object is presented on the display monitor.

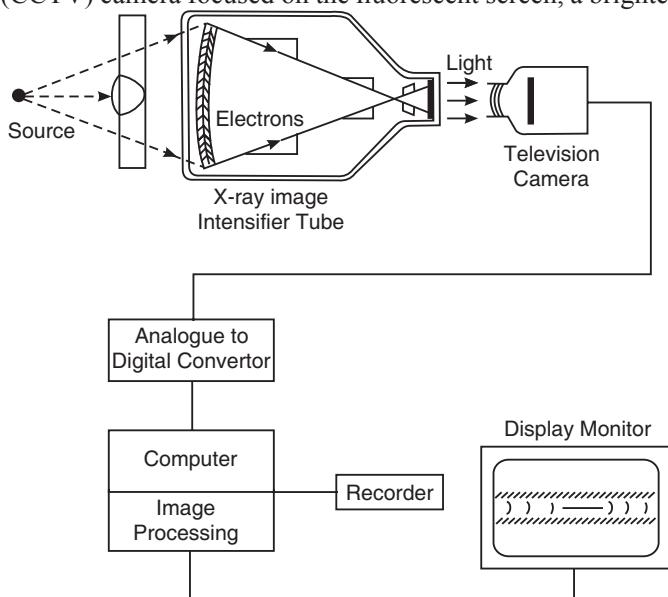


Fig. 37.17

37.12 COMPARISON OF CONVENTIONAL AND REAL-TIME RADIOGRAPHY

1. In conventional radiography, we obtain a negative image of the object on the photographic film. In the real time radiography, we see the positive image of the object.
2. In conventional radiography the image is viewed in static mode. In real time radiography, the image is viewed in dynamic mode and is interpreted at the same time as the radiation passes through the object.

QUESTIONS

1. What is NDT? Name some important NDT methods.
2. What are the benefits nondestructive testing of products?
3. Define NDT. Discuss the liquid penetrate method of NDT in detail. (G.T.U., 2009)
4. What is the principle of liquid penetrant method?
5. Explain in detail how liquid penetrant method is used in NDT. What are its advantages and limitations?
6. What is the basic principle of radiographic testing? Explain the displacement method for finding the depth of the defect in an object.
7. What is meant by fluoroscopy? Draw the schematic diagram of a real time radiographic testing apparatus.
8. Distinguish between radiography and fluoroscopy.
9. What is the principle of ultrasonic testing?
10. Discuss in detail the ultrasonic flaw detection. (G.T.U., 2009)
11. What is non-destructive testing? Explain with principle how flaw in a solid can be detected by non-destructive method using ultrasonics. (V.T.U., 2007)
12. Draw a block diagram of ultrasonic testing equipment. Explain the three different scan modes used for presentation of data.
13. What are the advantages and limitations of ultrasonic testing?
14. What is the basic principle of radiographic testing? Explain the displacement method for finding the depth of the defect in an object.

CHAPTER

38

Vacuum Technology

38.1 INTRODUCTION

The first major use of vacuum technology in industry occurred about 1900 in the manufacture of electric light bulbs. Other devices requiring a vacuum for their operation followed, such as the various types of vacuum tubes, cathode ray tube and electron microscope. Furthermore, it was discovered that certain processes carried out in a vacuum achieved superior results that are actually unattainable under normal atmospheric conditions. Such developments included the “blooming” of lens surfaces to increase the light transmission, the preparation of blood plasma for blood banks, and the production of reactive metals such as titanium. The advent of nuclear energy in the 1950s provided impetus for development of vacuum equipment on a large scale. High to ultra high vacuum is used in thin film deposition in semiconductor industry and other industries. Increasing applications for vacuum processes were steadily discovered, as in space simulation and microelectronics.

38.2 VACUUM

The term **vacuum** refers to the condition of a closed space which is devoid of all gases or other material content. It is defined as a diluted gas with its pressure or density lower than that of the ambient surrounding atmosphere. In order to achieve a vacuum in a vessel it is necessary to remove from it the molecules that make up the atmosphere (mostly nitrogen, oxygen, carbon dioxide) and generate a pressure that is lower than the ambient pressure. Removal of molecules continues until the desired level of vacuum is obtained and the removal process is to be continued further to maintain the desired vacuum level while the work that requires the vacuum is performed.

However, it is not experimentally feasible to achieve a “perfect” vacuum, although one can approach this condition extremely closely. It is possible routinely to obtain a vacuum of 10^{-6} Torr. With more sophisticated techniques 10^{-10} Torr and by special techniques a vacuum of 10^{-15} Torr, or about 30 molecules per cubic centimeter can be attained.

38.3 UNITS OF VACUUM

Traditionally, the pressure in a system is stated in terms of the height of a column of mercury that may be supported by the pressure in the system. At one standard atmosphere the force is 1.03 kg/sq. cm. This pressure will support a mercury column 760 millimeter high (as in a barometer). Thus, one standard atmosphere equals 760 mm Hg. A smaller unit torr is

introduced to express below-normal atmospheric pressures. A **torr** is defined as the pressure equivalent of a manometer reading of 1 mm of liquid mercury (1 torr = 1 mm Hg). Hence

One standard atmosphere = 760 torr.

The term ‘torr’ was replaced in 1971 by SI unit ‘**pascal**’ which is defined as the newton per square metre (N/m^2).

One pascal = 7.5×10^{-3} torr.

$$\therefore 1 \text{ Torr} = 1/760 \text{ atm} = 133.3 \text{ Pa.}$$

38.4 VACUUM RANGES

Table - 1

S.No.	Vacuum	Range
1.	Low vacuum	< 25 Torr
2.	Medium vacuum	25 to 10^{-3} Torr
3.	High vacuum	10^{-3} to 10^{-9} Torr
4.	Ultrahigh vacuum	10^{-9} to 10^{-12} Torr

38.5 PRODUCTION OF VACUUM

A vacuum is created in a chamber or vessel with the help of a vacuum pump commonly known as an exhaust pump. An **exhaust pump** is a device to remove or exhaust air, gas or vapour from a vessel. An exhaust pump consists of two sides. One side, known as **inlet** is connected to the vessel to be evacuated and the other known as **outlet** or **exhaust side** which expels the air, gas or vapour drawn from the vessel. The pressure at which the gas or air is expelled out on the exhaust side is called **exhaust pressure**. This pressure would be equal to the atmospheric pressure if the outlet is open to the atmosphere; or it would be less than that if the outlet is connected to an auxiliary pump called **backing pump**. In order to attain very low pressure or higher order vacuum in the chamber, the exhaust pressure should be very small. To achieve this, the backing pump and high vacuum pumps are arranged in series. The air or gas in the chamber are drawn in at the inlet of high vacuum pump and expelled at its outlet into the inlet of backing pump, which finally expels the gas into the atmosphere. The minimum pressure that can be produced in a given vessel is called **attainable vacuum**. This pressure depends largely on the exhaust pressure and varies from pump to pump.

38.6 CLASSIFICATION OF VACUUM PUMPS

Exhaust pumps can be classified into the following four categories:

- (i) Oil pumps
- (ii) Mercury pumps
- (iii) Molecular pumps and
- (iv) Diffusion pumps.

Oil pumps or mercury pumps can be of piston type or rotatory type. Piston type oil pumps produce low pressures of only 10^{-2} Torr and are used as backing pumps. Piston type mercury pumps are very slow and tedious in action though they produce vacuum of the order of 10^{-5} Torr. We describe here some of the pumps that are commonly used.

38.7 ROTARY OIL PUMPS

Principle: Rotary oil pumps are based on the principle of displacement of air. A portion of the air inside a vessel is isolated by a rotating disc and is then compressed till it attains

sufficiently high pressure and gets discharged to the atmosphere. As the air is expelled, more air comes to occupy the space in the vessel which in turn is also removed. The process continues and the pressure in the vessel keeps on decreasing.

Construction: Fig. 38.1 shows a typical rotary vane oil-seal pump. It consists of a hollow steel chamber which is known as **stator**. Stator acts as the working chamber. A massive cylindrical shaft known as a rotor is installed inside the stator. The rotor is electrically driven and rotates eccentrically inside the stator such that it is always in contact with a certain peripheral point of the stator. The rotor carries two sliding rods known as the **vanes**

that move radially under spring force. Rotor and vanes divide the working chamber into two separate spaces having variable volumes. The inlet port is connected to the vessel that is to be evacuated. A spring operated valve is fitted to the outlet port. This outlet valve is oil-sealed.

Working: The working principle of the pump is depicted in Fig. 38.2. As the rotor turns in the direction indicated, the space between rotor and stator on the inlet side keeps on increasing while the space between

rotor and the outlet side goes on decreasing. As a result, air or gas from the vessel to be evacuated is sucked into the enlarging suction chamber until the inlet is sealed off by the vane. As the rotor moves further, the gas enclosed in the space between rotor and the outlet side is compressed. When the pressure of the compressed gas becomes high enough, it forces open the discharge valve at the outlet and escapes out. All these operations are completed in

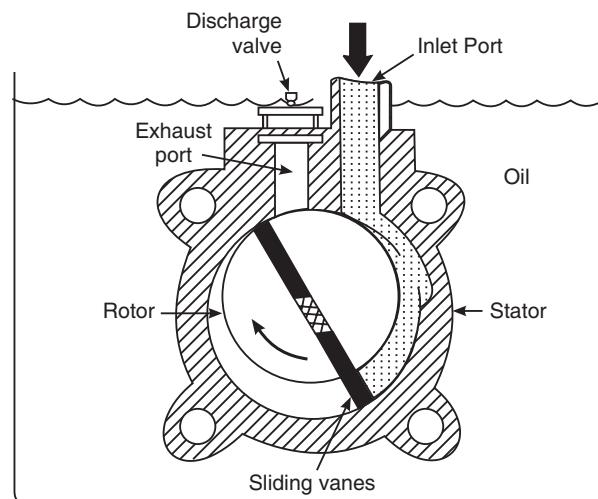


Fig. 38.1. Chief components of a typical rotary oil-seal pump.

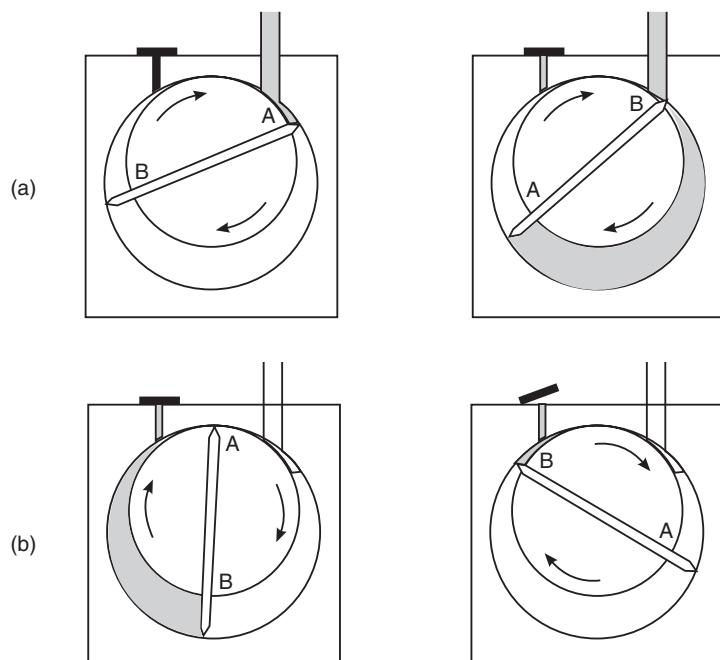


Fig. 38.2. Principle of operation of oil-sealed rotary vane pump: (a) gas from the vacuum system is expanded into the pump and (b) the gas is pushed through the pump exhaust.

one full rotation of the rotor. The process goes on repeating till a pressure of the order of 10^{-3} torr is produced in the vessel being evacuated.

Since each revolution sweeps out a fixed volume, this pump is called a *constant-displacement pump*.

38.7.1 Multi-stage Pumps

Rotary vane vacuum pumps are built in single- and two-stage versions. Two-stage pumps (Fig. 38.2) achieve lower ultimate pressures than single-stage pumps.

Applications

Typical applications of this pump are in food packaging, high-speed centrifuges, and ultraviolet spectrometers. It is also widely used as a forepump or a roughing pump, or both, for most of the other pumps that produce higher vacuum.

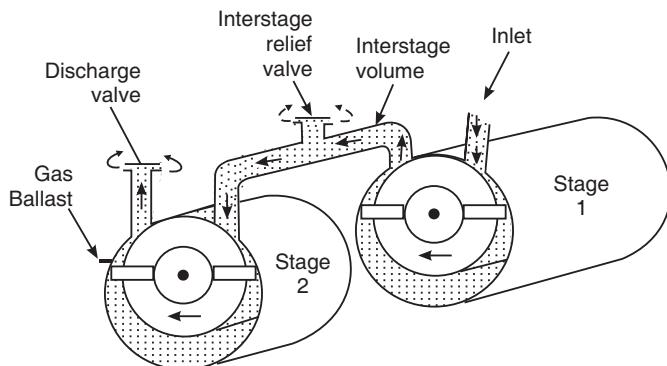


Fig. 38.3. Two stage Rotary Mechanical Pump

38.8 DIFFUSION PUMP

The most common type of pump for use in high vacuum applications is the **diffusion pump** (or, more properly, vapor jet pump). The only justification for calling them diffusion pumps is the observation that the molecules of the pumped gas penetrate some distance into the vapor of line jet in a manner resembling diffusion of one gas into another.

Construction:

A vacuum diffusion pump is basically a stainless steel chamber containing vertically stacked cone-shaped jet assemblies. Typically there are three jet assemblies of diminishing sizes, with the largest at the bottom. At the base of the chamber is a pool of a specialized type of oil having a low vapour pressure. There are several types of oil,

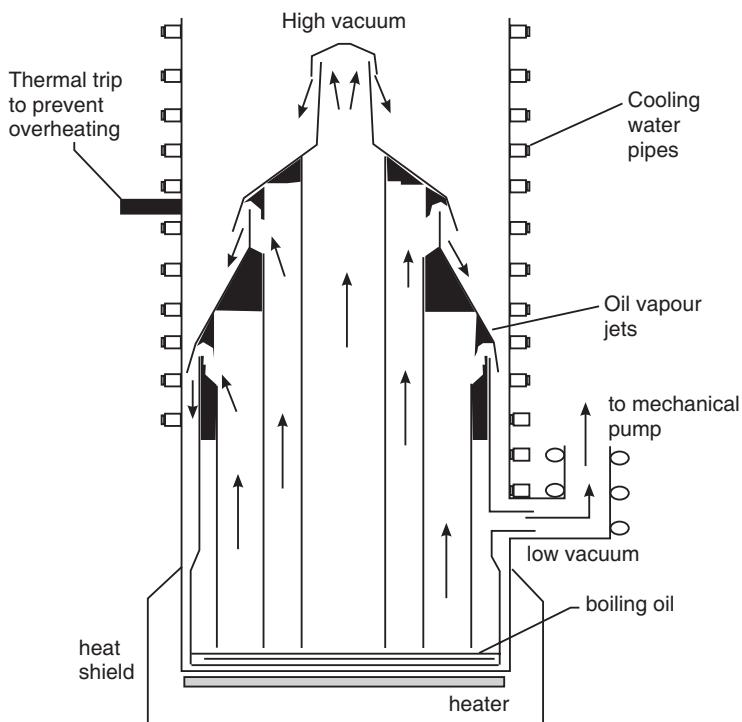


Fig. 38.4. Schematic diagram of an oil diffusion pump. The arrows indicate the movement of boiling oil vapour.

based variously on silicones, hydrocarbons, esters, perfluorals, and polyphenyl ethers that can be used. An electric heater is installed beneath the floor of the chamber to boiling the oil. The temperature at the base of the chamber, where the oil is being vaporized, ranges from about 190°C to about 280°C. The upper part of the pump wall is wrapped around with coils which circulate cold water through them. Water circulated through coils on the outside of the chamber cool the chamber and permit operation over long periods of time.

Working: A diffusion pump operates by boiling oil at the base of the pump. Oil molecules escape to produce a supersonic vapor (400-400 m/s) that is directed into a funnel-shaped set of baffles and jettied towards the sides. Air molecules that have diffused into the vicinity are compressed against the jets and entrained. When the oil vapour strikes the upper pump wall, it condenses into liquid and sinks to the bottom of the pump taking the adsorbed air molecules with it. At the bottom, the heater reboils the oil, releasing the air molecules, producing a build up of air molecules in the lower region. A port that is attached to a mechanical pump removes these air molecules. The effect of removing molecules is to create a high vacuum in the upper portion of the chamber. It is this part of the chamber that is connected to the sample chamber where a high vacuum is needed. The diffusion pump produces a working pressure of 10^{-5} to 10^{-6} torr.

A vacuum of 10^{-3} Torr has removed a great many of the atmospheric molecules, but this is actually the level at which a vacuum diffusion pump begins to operate. A vacuum diffusion pump cannot begin its work with full atmospheric pressure inside the chamber. Instead, an ancillary mechanical roughing pump (or forepump), capable of a modest level of pumping, first brings the pressure inside the vacuum diffusion chamber down to about 10^{-3} Torr. At this point, the vacuum diffusion pump takes over to create a vacuum ranging from 10^{-3} to 10^{-10} Torr. Since the diffusion pump cannot exhaust directly to atmospheric pressure, the forepump is used to maintain proper discharge pressure conditions.

Advantages: They are reliable, they are simple in design, they run without noise or vibration, and they are relatively inexpensive to operate and maintain. In fact, diffusion pumping is still the most economical means of creating high vacuum environments. These pumps also tolerate operating conditions such as excess particles and reactive gases that would destroy other types of high vacuum pumps. Since the chamber itself has no moving parts aside from the oil droplets, a vacuum diffusion pump can operate with stability over long periods.

In all diffusion pumps, a small amount of backstreaming occurs. By definition, backstreaming is the migration of minute levels of oil that moves in the opposite direction - toward the inlet of the pump and into the process stream, which may be the stage of an electron microscope or a welding chamber. For this reason, some systems add a liquid nitrogen cryotrap to remove oil particles before they can reach the process stream.

38.9 TURBOMOLECULAR PUMPS

Turbomolecular pumps use no oil and operate like jet engines with multiple, angled blades rotating at very high speed. The pump compresses gas by using a high-speed rotating surface to give momentum and direction to gas molecules and literally push the air molecules to the outlet of the pump.

Construction: Turbomolecular pumps employ a series of rotor/stator pairs mounted in series. The rotors rotate at high speed (9000-90,000 rpm). The pump rotor contains a number of slotted disks (sets of rotor blades, typically 10 to 14). These rotating sets of disks are spaced interposed between fixed disks (sets of stator blades). The incidence angles of the slots in the fixed disks oppose those of the rotating disks.

Working: As gas molecules enter through the inlet, the rotor, consisting of a number of angled blades, strikes the molecules, imparting energy to them. Gas molecules are propelled into gas transfer holes in a plate below the blades called a “stator”. Gas captured by the first rotor is sent into the lower rotor/stator pairs and is successively compressed until it reaches a pressure where it can be removed by a mechanical backing pump. The blades are as thin as possible and slightly bent for maximum compression. Turbo pumps can reach 10^{-7} to 10^{-10} torr.

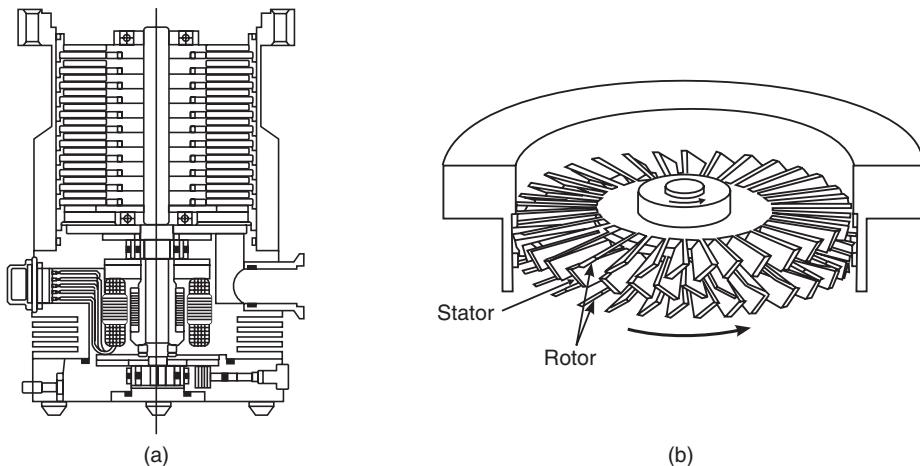


Fig. 38.5. (a) Turbo molecular Pump (b) Rotors and Stators of Turbo molecular Pump

Unfortunately, because of the high rotation speeds (10-20,000 rpm), turbo molecular pumps have shorter life spans than oil diffusion or sealed-oil mechanical pumps.

38.10 CRYOPUMPS

A **cryogenic pump** removes gas from a container by condensing the gas molecules on an extremely cold surface in the container.

A schematic of a typical cryopump is shown in Fig. 38.6. Cryogenic high-vacuum pumps, or cryopumps, create high vacuum by freezing molecules of air onto cryogenically cooled surfaces inside the pump. The vacuum typically ranges from 10^{-4} to 10^{-9} Torr. Cryopumps have no moving parts that are exposed to vacuum. Because they require no oil, they are inherently clean and cannot contaminate products created in the vacuum they produce. Nothing but cryogenically cooled surfaces are exposed to the process chamber.

Working: Cryopumps usually consist of two internal stages, or pump areas, which freeze specific gas species at set cryogenic temperatures. The stages are connected to a sealed cryogenic refrigerator run by an external compressor. The refrigerator cools the stages.

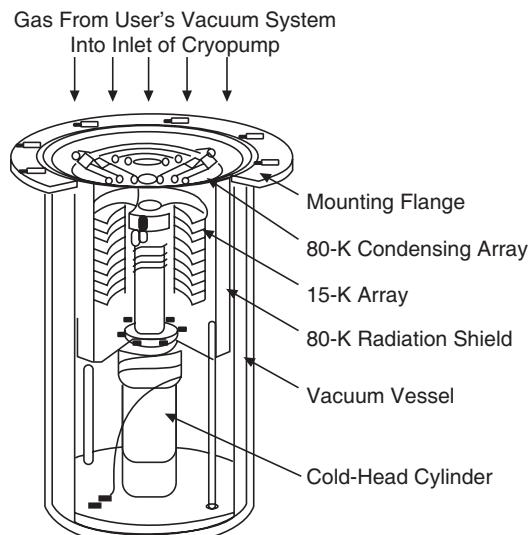


Fig. 38.6. Schematic of a cryopump

The first stage of the pump, the primary pumping surface of which is the inlet array, is generally operated at temperatures between 65 and 100 K. Its main function is to pump or capture water vapour. The second stage consists of a series of metal pumping surfaces, which are arranged in patterns designed for particular applications. Generally operated at temperatures ranging from 10 to 20 K, this stage can pump gases such as nitrogen and argon. The metal pumping surfaces are partially covered with charcoal granules. Gases such as hydrogen and helium, which cannot be frozen at typical second-stage temperatures, are adsorbed by the charcoal granules and thereby removed from the vacuum chamber.

38.11 VACUUM GAUGES

A **vacuum gauge** determines the pressure in an evacuated apparatus by the measurement of some physical property of the residual gases, such as viscosity of the gas, thermal conductivity of the gas or the electrical properties of the gas when ionized. No single gauge can measure pressure from atmosphere to high vacuum. Different gauges are used for measuring different ranges of vacuum.

There are two fundamental classes of vacuum measurement devices:

(a) **Direct reading vacuum gauges** are based on fundamental principles of gas compressibility. These include Bourdon tube, McLeod, and other liquid manometers, and aneroid (and of course, mercury) barometers.

(b) **Indirect reading vacuum gauges** sense some parameter of the residual gas such as its thermal conductivity or ion conduction and translate this to a pressure indication. These include thermocouple, Pirani, and ion gauges.

We describe here the indirect reading gauges which are used to measure high vacuum.

38.11.1 Thermocouple Gauge

The **thermocouple gauge** is one of the more common gauges for vacuum pressure measurement in the 1 Torr to 1 milliTorr range. Thermocouple gauge does not measure pressure directly. Instead, it depends on changes of the thermal conductivity of the gas.

Construction: The thermocouple gauge contains two elements: a heater (filament) and a thermocouple junction which contacts the filament. These gauges consist of the gauge tube itself, a power supply for the filament, and a moving coil meter for displaying the pressure (see Fig. 38.7).

Working: When the power supply is switched on current flows through the filament coil. The filament current is held constant during the operation of the gauge. As the evacuation of the sample chamber starts, the pressure within the gauge tube decreases. The filament becomes hotter because the number of gas molecules hitting the wire and conducting heat away from the wire decreases. As temperature rises, the thermocouple voltage increases. The thermocouple junction voltage is measured by a sensitive meter that has previously been calibrated against a manometer to determine the pressure. Each type of thermocouple tube has its own calibration curve.

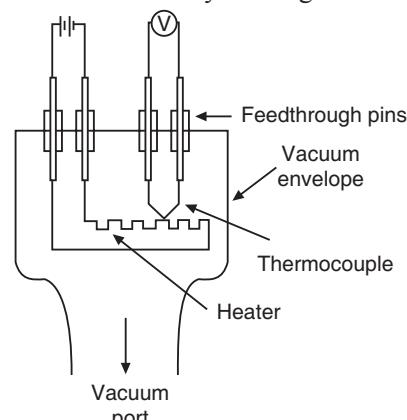


Fig. 38.7. Schematic of a thermocouple gauge

38.11.2 Pirani Gauge

The **Pirani gauge** consists of a heated filament of platinum, tungsten, or some other metal with a high temperature coefficient of electrical resistance suspended in a tube which is connected to the system where vacuum is to be measured.

Principle: The principle of this instrument is based on the measurement of the changes in resistance of the heated wire placed into the vacuum system. The resistance depends on the temperature of the wire. At high pressures the wire will be cooled by collisions with the residual gas molecules, i.e. by warming up the residual gas. As the pressure is reduced this cooling mechanism gets less effective and the temperature of the wire rises. The resistance of the wire can be calibrated as function of the pressure.

Working: The filament, mounted in a bulb fitted with a connecting tube (Fig. 38.8 a), is connected to one arm of a Wheatstone bridge and is heated by a constant current as shown in Fig.

38.8 (b). It is balanced with an identical compensating filament mounted in an adjacent arm of the bridge. This auxiliary bulb is evacuated and sealed off at a very low pressure. The use of an auxiliary bulb serves to make the gauge insensitive to variations in room temperature. Changes in the over-all temperature of one bulb are the same as changes in the other, so that the galvanometer does not respond to these changes but only to the changes produced by the residual gas in the one bulb.

If the gas is at high pressure, gas molecules collide frequently with the filament and absorb energy from the filament which results in cooling of the filament. As the pressure of the gas molecules decreases the number of gas molecules inside the chamber also goes down resulting in fewer collisions with the filament. As a result the temperature of the filament increases because of decreased cooling. Electrical resistance of a wire varies with temperature. If the bridge is balanced at one temperature of the filament, a change of its temperature caused by a change in the heat conductivity of the residual gases will unbalance it. Thus, the deflection of the bridge galvanometer indicates the pressure of the residual gases.

This gauge is used to measure the pressure between 0.5 Torr to 10^{-4} Torr. For higher vacuum measurement other instruments such as a Penning gauge are used.

38.11.3 BAYARD-ALPERT GAUGE

There are two types of ionization gauges which are used for pressure measurements between 10^{-2} and 10^{-10} Torr. They are known as hot cathode gauges and cold cathode gauges.

In the **hot cathode gauge** (HCG) electrons emitted from a thermionic cathode are accelerated by suitable electrodes into an ionizing space where they cause ionization. The "Bayard-Alpert" gauge is the most common and somewhat standardized variety of hot cathode gauges. The Bayard-Alpert gauge was invented by R.T. Bayard and D. Alpert in 1950.

Construction: The Bayard-Alpert ion gauge tube (sensor) resembles a triode in construction and consists of a tungsten or coated filament located outside of a helical grid (Fig. 38.8). The grid cage is the ionization space where ions are produced due to collisions of electrons with the residual gas atoms. A fine straight wire arranged in the center of the helical cage acts as the anode and collects the ions that are produced in the ionization region.

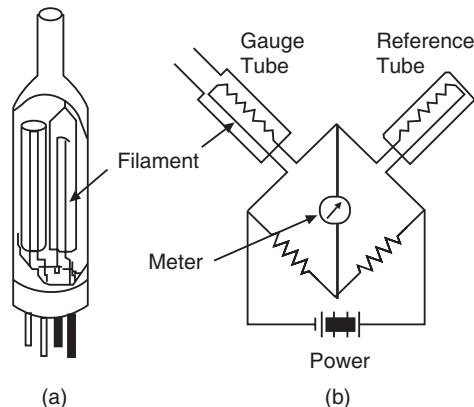


Fig. 38.8. Schematic diagram of a Pirani Gauge

The grid is maintained at a positive voltage (typically +180 VDC), with the filament at a lower but still positive voltage (typically +30 VDC). With this constant potential difference, emission current is related to filament temperature. An electronic feedback loop maintains emission current at a constant value (typically 0.6 to 2.4 mA). The collector electrode is maintained at ground potential (0 V).

Working: When heated, the filament emits electrons. They are accelerated towards the cylindrical cage which is maintained at a positive potential. As the electrons travel in the space enclosed by the grid, collide with residual gas molecules and ionize some of them. Positively charged ions produced in the process are attracted by the collector electrode which is at a negative potential with respect to the grid. The number of ions produced per unit volume depends on the gas pressure, and hence the current also depends on gas pressure. At constant temperature, the collector current is directly proportional to the gas pressure.

The useful operating range of a conventional BAG extends between 10^{-2} and 10^{-10} Torr.

38.11.4 Penning Gauge

“Penning” gauge is one of the varieties of **cold cathode gauges**. In this gauge, ionization is caused by circulating electron plasma trapped in crossed electric and magnetic fields.

Construction: A schematic of the Penning gauge is shown in Fig. 38.10. It consists of two electrodes, cathode and anode arranged between the pole pieces of a horse-shoe magnet that can produce a magnetic field of the order of 1-2 kG. The cathodes are two flat plates while the anode is a small cylindrical tube. A high voltage of the order of 2-6 kV is applied between the electrodes. The magnetic field is arranged such that the magnetic field lines cross the electric field lines.

Working: A cosmic ray, field emission, a photon, radioactivity or some other event causes the random release of an electron at the cathode, as shown in Fig. 38.11 (a). A discharge slowly builds electron plasma inside the ionization volume. Due to the action of the magnetic field, the electrons move in cycloidal jumps, circling about the anode [Fig. 38.11 (b)], and during part of each jump they have sufficient energy to ionize residual

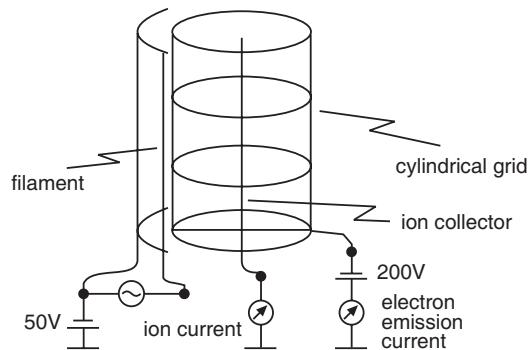


Fig. 38.9. Schematic of Bayard-Alpert ion gauge

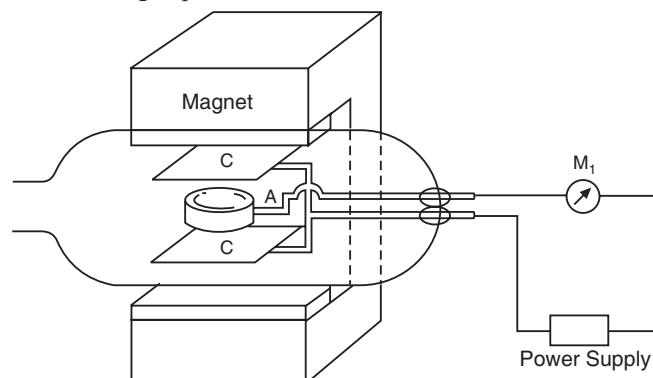


Fig. 38.10. Schematic of a Penning gauge

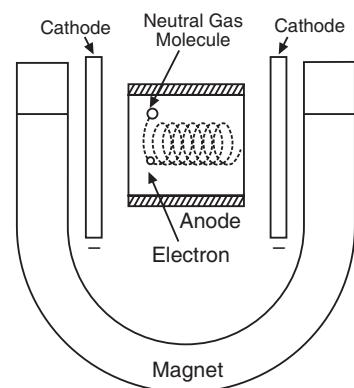


Fig. 38.11

gas molecules through electron impact ionization. The probability of collision is proportional to the gas density. As the paths of the electrons along the spiral path are long, their collision probability with gas molecules is sufficiently large to maintain the discharge on the formation of the required number of charge carriers. The probability of collision is proportional to the gas density. The positive and negative charge carriers produced by collisions move to the corresponding electrodes. The current generated by this ion collection process is measured and used as an indirect indication of gas density and pressure. The typical operating range of a Penning gauge is between 10^{-2} and 10^{-9} Torr.

38.12 VACUUM TECHNOLOGY

Vacuum Technology is the term applied to all processes and physical measurement carried out under conditions of below-normal atmospheric pressure.

A process or physical measurement is generally performed under vacuum for one of the following reasons:

- to remove the constituents of the atmosphere that could cause a physical or chemical reaction during the process (e.g., vacuum melting of reactive metals such as titanium)
- to disturb an equilibrium condition that exists at normal room conditions, such as the removal of occluded or dissolved gas or volatile liquid from the bulk of material (e.g., degassing of oils, freeze-drying) or desorption of gas from surfaces (e.g., the cleanup of microwave tubes and linear accelerators during manufacture)
- to extend the distance that a particle must travel before it collides with another, thereby helping the particles in a process to move without collision between source and target (e.g., in vacuum coating, particle accelerators, television picture tubes).
- to reduce the number of molecular impacts per second, thus reducing chances of contamination of surfaces prepared in vacuum (e.g., in clean-surface studies and preparation of pure, thin films).

38.13 APPLICATIONS OF VACUUM

Industrial vacuum applications range from mechanical handling (such as the manipulation of heavy and light items by suction pads) to the deposition of integrated electronic circuits on silicon chips. Obviously, vacuum requirements are as widely varied as the particular processes using vacuum. In the rough vacuum range from about one torr to near atmosphere, typical applications are mechanical handling, vacuum packing and forming, gas sampling, filtration, degassing of oils, concentration of aqueous solutions, impregnation of electrical components, distillation, and steel stream degassing.

At lower pressures down to about 10^{-4} torr, many metallurgical processes such as melting, casting, sintering, heat treatment, and brazing can derive benefit. Chemical processes such as vacuum distillation and freeze-drying also need this range of vacuum. Freeze-drying is used extensively in the pharmaceutical industry to prepare vaccines and antibiotics and to store skin and blood plasma. The food industry freeze-dries coffee mainly, although most foods can be stored without refrigeration after freeze-drying, and the technique is receiving widespread acceptance.

The pressure range down to about 10^{-6} torr is used for cryogenic (low-temperature) and electrical insulation. It is used in the production of lamps; television picture tubes, X-ray tubes; decorative, optical, and electrical thin-film coatings; and mass spectrometer leak detectors.

In thin-film coating, a metal or compound is evaporated under high vacuum from a source onto a base material or substrate. The base material is generally plastic for decorative coatings;

glass for optical coatings; and glass ceramic, or silica for electrical coatings. Thickness of the film can vary from about 1/4 wavelength of visible light to 0.001 inches or more. In the optical field, antireflection coatings are deposited on lenses for cameras, telescopes, eyeglasses, and other optical devices, considerably reducing the amount of light reflected by the lenses and thus giving a brighter transmitted image.

Almost every research laboratory uses vacuum directly in its experiments or employs equipment that depends on vacuum for its operation. The lowest pressures are obtained in the research laboratories, where equipment is generally similar to but smaller than that used by industry.

Typical of the research equipment using vacuum down to about 10^{-6} torr are the electron microscope, analytical mass spectrometer, particle accelerator, and large space simulation equipment. Particle accelerators range from small van de Graaff machines to large proton synchrotrons.

In space simulation, large units that simulate space around a complete vehicle require a vacuum of 10^{-6} torr or below. Such vessels incorporate a complete shroud at liquid nitrogen temperature and a port through which high-intensity light can be beamed to simulate the sun's radiation.

In the pressure region down to and below 10^{-9} torr, research applications include electrical insulation, thermonuclear energy conversion experiments, microwave tubes, field ion microscopes, field emission microscopes, storage rings for particle accelerators, specialized space simulator experiments, and clean-surface studies. In many experiments it is not only necessary to reach such pressures of 10^{-9} torr but to reduce the hydrocarbons in the residual gases to an absolute minimum. Even small traces of hydrocarbons can make results unreliable. To achieve a vacuum of this order the vacuum vessel and the equipment inside must be cleared of residual gas (degassed) to the greatest extent possible. A common solution is to bake the whole apparatus for a number of hours at about 350°C while maintaining a vacuum in the 10^{-5} torr region. Baking at this temperature requires the use of all-metal sealing rings. To eliminate hydrocarbons, the unit is pumped down to about 10^{-3} torr using sorption pumps; and from there, sputter ion pumps and titanium sublimation pumps complete the task down to 10^{-9} torr or below.

38.14 HIGH VACUUM SYSTEMS

A **high vacuum system** consists of a vacuum chamber, vacuum pumps, and valves (Fig. 38.12). In between various combinations of tubing, fittings and valves are employed. These are required for the system to operate. We do not have such a powerful pump that can exhaust air in the chamber from the level of atmosphere to high

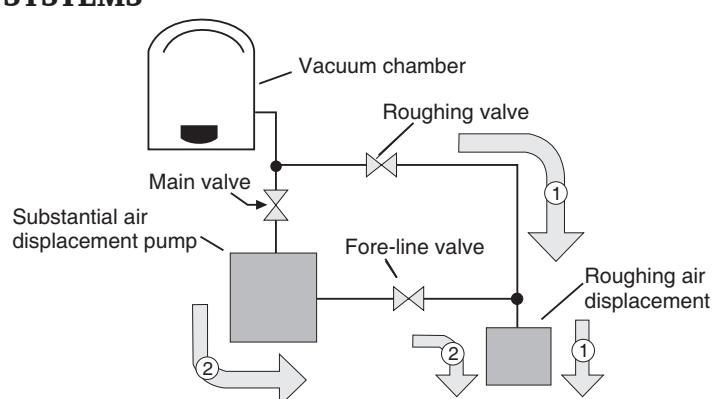


Fig. 38.12. Schematic of a typical vacuum system

vacuum at once. Therefore, two vacuum pumps are connected to a sample chamber. One of the vacuum pumps, known as **roughing pump**, produces medium vacuum, and the second

pump, a high vacuum pump, produces the high vacuum required for a specific process. At the beginning, only the roughing valve is open, and the roughing pump works to exhaust air. Then, the rough valve is closed, and the remaining two valves are opened to activate both the roughing and diffusion pumps connected in series. By doing this, the pressure inside the chamber reaches the required level. A vacuum gauge is usually connected to the system to monitor pressure in the sample chamber.

The minimum configuration of a vacuum system is dependent upon the application.

38.15 THIN FILM DEPOSITION

The formation of a thin film requires high vacuum condition. If it is attempted to form a film in the atmosphere or in a low vacuum, the particles of the film material try to move toward the object on the ceiling of the chamber but are obstructed by vapour, oxygen, nitrogen, and carbon dioxide in the air. Thus, they seldom reach the object and cannot form a film. Furthermore, even if the particles reach the object, it causes a number of problems where the adhesiveness of the film is weak or the film material bonds with other substances in the air. The solution is to maintain a high vacuum in the chamber (Fig. 38.13) to reduce unwanted substances. This produces a quality thin film of high degree of purity and adhesiveness.

There are different methods of forming a thin film in the vacuum thin-film coating technology. Physical Vapour Deposition (PVD) is one of the processes used. In PVD techniques, resistance heating, and electron beam heating are the processes most widely used for producing thin films.

A resistance heated evaporation source is relatively simple and inexpensive. Resistance heating is done inside a vacuum chamber where the material, usually in a boat is heated typically to its melting point and the substrate to be deposited on is positioned facing the source at a distance from it. A high current flowing through the boat heats it up and causes evaporation of the film material. A crystal monitor is mounted close to the substrate, which provides an estimate of how much and how fast the material is being deposited.

A vacuum thin-film coating system applies a thin coat on the object in a vacuum chamber. Although the thickness of the film varies from product to product, the average is 0.1 to several dozen micrometers, which is thinner than aluminum foil for

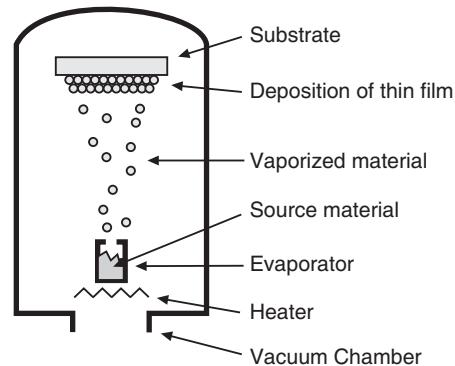


Fig. 38.13. The concept of thin film deposition

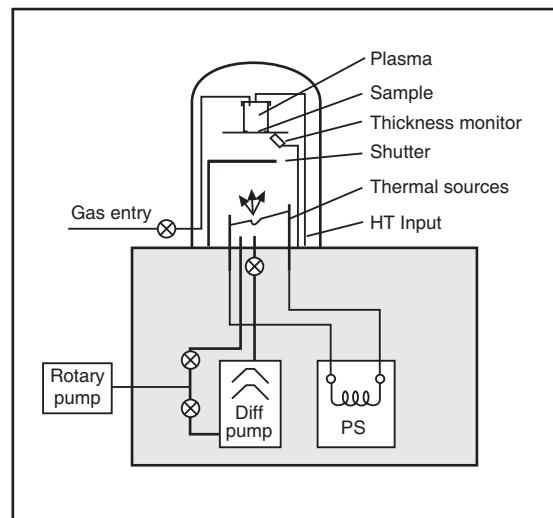


Fig. 38.14. Schematic of a vacuum thin-film coating system

household use (several dozen micrometer). Fig. 38.14 shows the schematic of simple vacuum thin-film coating system.

The thin films are utilized in a wide variety of fields and many of them exist around us. Thin film deposition of metallic, insulating, conductive and dielectric materials plays an important role in a large number of manufacturing, production and research applications.

QUESTIONS

1. What is a vacuum? How it is measured? It is possible to have prefect vacuum.
2. What do you understood by vacuum technology?
3. Give the theory and working of a rotary oil pump. What is the limit of pressure that can be achieved with the pump?
4. How does diffusion pump generate vacuum? What is the order of vacuum produced by a diffusion pump?
5. Explain the construction and working of a turbo molecular pump.
6. Explain the principle and working of a cryopump.
7. Explain units of pressures, and measuring instruments used in various ranges.
8. Explain the principle and working of a thermocouple gauge.
9. Explain the principle and working of a Pirani gauge.
10. Explain the principle and working of a Bayard-Alpert gauge.
11. List the applications of vacuum technology in detail.
12. Give an outline of a high vacuum system.
13. Draw the schematic of a thin-film coating system and explain its working.

CHAPTER

39

Nanotechnology

39.1 INTRODUCTION

One of the outstanding features of technological progress in the last century is the continuous miniaturization of technical devices and components. A strong impact for the development arose from the necessity to reduce the geometry of integrated circuits in order to follow the demands for a higher complexity of circuits and devices. It resulted in the development of various micro-technological processes. The developments of microelectronics indicate a fundamental approach known as **top down approach**. The top down approach starts with a large-scale object or pattern and progressively reduces its dimensions. The process is similar to sculpture a statue from a big a rock. Nature adopts the opposite approach in building the materials. Simple atoms and molecules are collected, consolidated and fashioned into a complex structure. This is known as **bottom-up** approach. Bottom up approach is a widespread process in biology where, for example, enzymes assemble amino acids to construct living tissue that forms the organs of the body. On 29th December 1959, Physicist and Nobel Laureate, Richard Feynman presented a visionary and prophetic lecture at the annual meeting of American Physical Society, entitled “There is Plenty of Room at the Bottom” where he speculated that the bottom-up approach could some day be adopted by scientists for preparation of materials and fabrication of devices to atomic specifications: he said “the principle of physics, as far as I can see, do not speak against the possibility of maneuvering things atom by atom.” Feynman’s sparkling discussion of the problems and promise of miniaturisation was the starting point for the original definition of *nano*technology (or *molecular manufacturing*): the use of nanoscale machines to build complex products including other nanomachines. However, it was not until the 1980s that a notable research activity occurred and significant developments resulted. The advent of scanning probe microscopes which permitted observation of individual atoms and molecules, made it possible to manipulate and move atoms and molecules to form new structures and thus design new materials that are built from simple atomic level constituents. The ability to carefully arrange



Ancient Greek Vase with Gold-Containing Ruby Glass

atoms provides opportunities to develop mechanical, electrical, magnetic, and other properties in materials that are not otherwise possible.

In recent years the term nanotechnology has increasingly been used to describe any nanoscale products, from thin coatings to tiny particles, and any tiny objects in general. Whole areas of materials-science, biotechnology and chemistry are labeled and marketed as nanotechnology.

Thus, nanotechnology has been developing along two different lines. One of them is leading to technological developments on the nanometer scale, usually 0.1-100nm. Much of current interest and enthusiasm is focused on this. The other line anticipates designing and building atom by atom nanomachines and devices which would bring in the next industrial revolution.

39.2 NANOSCALE

The word “nano” is derived from a Greek word meaning dwarf or extremely small and means a billionth (10^{-9}) part of a unit. A *nanometer* or nm is one thousand millionth of a metre, i.e., $1 \text{ nm} = 10^{-9} \text{ m} = 10^{-3} \mu\text{m} = 10 \text{ \AA}$. One nanometer spans 3 to 5 atoms lined up in a row. For comparison, a single human hair is about 80,000 nm wide and a red blood cell is approximately 7,000 nm wide. Scientists and engineers are nowadays interested in the nanoscale, which may be taken as 100 nm to 0.2 nm approximately. Below this lies the atomic scale 0.1nm. Therefore, the nano-world is a borderland between the quantum world and the macro world.

Some examples of size from molecular to macro

Size (nm)	Examples	Terminology
0.1-0.5	Individual chemical bonds	Molecular/atomic
0.5-1.0	Small molecules	Molecular
1-1000	Proteins, DNA, inorganic nanoparticles	Nano
10^3 - 10^4	Devices on a silicon chip, living cells (bacteria: 1 μm ; yeast: 5 μm), human hair (50 μm)	Micro
$>10^4$	Normal bulk matter	Macro

39.3 THE SIGNIFICANCE OF THE NANOSCALE

Many of the properties of solids depend on the size of the solid. Microscopic details become averaged out in bulk materials. In case of bulk materials, the properties such as density, elastic modulus, resistivity, dielectric constant etc are averaged properties. Many properties of materials change in the micrometer or nanometer range. As materials become smaller and smaller, eventually a point is reached where the averaging no longer works. The properties of materials can be different at the nanoscale for the following reasons:

1. First, nanomaterials have a relatively larger surface area when compared to the same mass of material produced in a larger form. This can make materials more chemically reactive (in some cases materials that are inert in their larger form are reactive when produced in their nanoscale form), and affect their strength or electrical properties.
2. Second, quantum effects can begin to dominate the behaviour of matter at the nanoscale affecting the optical, electrical and magnetic behaviour of materials.
3. Physical properties of materials are generally characterized by critical lengths. For example, the electrical conductivity of a metal is determined by the mean free path which is the distance that electrons travel between collisions with the vibrating atoms or impurities of the solid. If the size or one of the dimensions of nanomaterial is

smaller than the critical length, the nanomaterial exhibits properties that are different from the corresponding bulk material.

4. Further, the bulk properties of materials often change dramatically with nano ingredients. Composites made from particles of nano-size ceramics or metals smaller than 100 nanometers can suddenly become much stronger than predicted by existing materials-science models. For example, metals with a so-called grain size of around 10 nanometers are as much as seven times harder and tougher than their ordinary counterparts with grain sizes in the hundreds of nanometers.

39.4 NANOTECHNOLOGY

The term nanotechnology was first coined in 1974 by Norio Taniguchi of the Tokyo Science University. In 1986, K. Eric Drexler wrote “Engines of Creation” and popularized the term nanotechnology. One of the problems facing nanotechnology is the confusion about its definition. Originally, nanotechnology meant building things from the bottom up starting from molecular level while it implies nowadays the study and control of phenomena and materials at length scales below 100 nm.

(i) Molecular Nanotechnology: Nanotechnology, in its traditional sense, means building things from the bottom up, with atomic precision, as was envisioned by the renowned physicist Richard Feynman. It is the postulated ability to manufacture objects and structures with atomic precision, literally atom by atom. This is very much similar to the abilities of living cells, which do exactly the same thing, although based on evolution and not by intent and design. Nanotechnology aims at building machines on the scale of molecules, a few nanometers wide—motors, robot arms, and even whole computers, far smaller than a cell. The original meaning is now more properly labeled “molecular nanotechnology” (MNT), or “molecular manufacturing.” MNT is a hypothetical advanced form of nanotechnology that is believed will be developed far into the future.

(ii) Nanoscale Bulk Technology: Much of the work being done today that carries the name ‘nanotechnology’ is not nanotechnology in the original meaning of the word. The term “nanotechnology” has undergone a change in its meaning over the years to imply “anything smaller than microtechnology with novel properties,” such as nano-powders, and other things that are nanoscale in size. As nanotechnology became an accepted concept, the meaning of the word shifted to encompass the simpler kinds of nanometer-scale technology. Their definition includes anything smaller than 100 nanometers. Nanotechnology comprises any technological developments on the nanometer scale, usually 0.1-100nm. The broadened version of the technology may be more properly called “nanoscale bulk technology”. It exploits the strange properties of nano-scale materials to produce new useful products. **Nanotechnology is the design, characterization, production and application of structures, devices and systems by controlling shape and size at the nanometer scale.**

The term nanotechnology is often used interchangeably with molecular nanotechnology (MNT). However, it is very important to note that **molecular manufacturing** is fundamentally different from **nanoscale technologies** in many ways, both in approach and implications.

39.5 WHAT IS MOLECULAR NANOTECHNOLOGY?

“I want to build a billion tiny factories, models of each other, which are manufacturing simultaneously. . . The principles of physics, as far as I can see, do not speak against the possibility of maneuvering things atom by atom. It is not an attempt to violate any laws; it is something, in principle, that can be done; but in practice, it has not been done because we are too big.” — Richard Feynman, Nobel Prize winner in physics.

Thus, Feynman's vision is of miniature factories using nanomachines to build complex products, by manipulating atoms individually and place them in a pattern to produce a desired structure. Nature has perfected the science of manufacturing matter molecularly. For instance, our bodies are assembled in a specific manner from millions of living cells. Cells are nature's nanomachines. At the atomic scale, elements are at their most basic level. On the nanoscale, we can potentially put these atoms together to make almost anything.

Today's manufacturing methods are very crude at the molecular level. Though we can manufacture complex mechanical machinery, electronic devices and computers, we cannot make such devices with the precision with which the chemist can synthesize a crystal, a bio-polymer, or a relatively small molecule. With chemistry we can make precise molecular structures and compounds, but we have not been able to scale up that success to other macroscopic products. Thus there is a wide gap in our synthetic abilities.

At the most basic technical level, MNT is "large-scale mechanosynthesis based on positional control of chemically reactive molecules" building, with intent and design, and molecule by molecule,

- (i) incredibly advanced and extremely capable nano-scale and micro-scale machines and computers, and
- (ii) ordinary size objects, using other incredibly small machines called **assemblers**.

Production of goods through MNT involves the following three steps:

1. **Manipulate individual atoms:** A suitable technique to grab single atoms and move them to desired positions will be developed. In 1990, IBM researchers showed that it is possible to manipulate single atoms. They positioned 35 xenon atoms on the surface of a nickel crystal using atomic force microscopy instruments.
2. **To develop nanoscopic machines:** Machines that can be programmed to manipulate atoms and molecules at will are called **assemblers**. Trillions of assemblers will be needed to develop products in a viable time frame. Nanomachines, called **replicators**, will be programmed to build more assemblers.
3. **Replacement of traditional methods:** Assemblers and replicators will work together like hands and automatically construct products. It will vastly decrease manufacturing costs, thereby making consumer goods plentiful, cheaper and stronger.

Once the above envisioned molecular machinery is created, it will usher in the next industrial revolution. Nanotechnology offers better products with exacting specifications and characteristics and along with a vastly improved manufacturing process. It will offer better built, longer lasting, cleaner, safer, and smarter products for the home, for communications, for medicine, for transportation, for agriculture, and for industry in general.

This advanced form of nanotechnology is expected to be achieved by around 2040.

39.6 NANOTECHNOLOGIES IN THE PAST

Artisans had taken advantage of nanosized materials since a long time. In the 4th century A.D. Roman glass makers were fabricating glasses containing nanosized metals. Silver and copper nanoparticles were used as far back as the 9th century in Mesopotamia for generating a glittering effect on the surface of pots. Even these days, pottery from the Middle Ages and Renaissance often retain a distinct gold or copper colored metallic glitter. This so called luster is caused by a metallic film that was applied to the transparent surface of a glazing.

Michael Faraday provided the first description, in scientific terms, of the optical properties of nanometer-scale metals in his classic 1857 paper. In 1908, Gustav Mie explained the dependence of the glasses on metal size and kind.

Photography is an example of “old” nanotechnologies. Photography was developed in 18th and 19th centuries, which depends on production of silver nanoparticles sensitive to light. Photographic film is an emulsion, a thin layer of gelatin containing silver bromide. The light decomposes the silver bromide producing nanoparticles of silver, which are the pixels of an image.

Over the last several decades, photo-lithographic patterning of matter on the 1000 nm length scale has led to the revolution in microelectronics and helped to create tiny features on computer chips for the past 20 years.

During the last century, investigations at nanoscale were not pursued much as compared to molecular and bulk length scales because significant developments of the corresponding investigative tools have not been made till recently.

However, chemists have developed the ability to control the arrangement of small numbers of atoms inside molecules (length scale of less than 1.5 nm), leading to revolutions in drug design, plastics, and many other areas.

39.7 FOUR GENERATIONS OF NANOTECHNOLOGY DEVELOPMENT

The nanotechnology as it is evolving nowadays has been passing through the following stages. Starting with bulk nanoscale technology, we ultimately end up in MNT.

1. The first phase is that of passive nanostructures, materials designed to perform one task.
2. The second phase introduces active nanostructures for multitasking; for example, actuators, drug delivery devices, and sensors.
3. The third phase will feature nanosystems with thousands of interacting components.
4. A few years after that, the first integrated nanosystems, functioning much like a mammalian cell with hierarchical systems within systems, are expected to be developed.

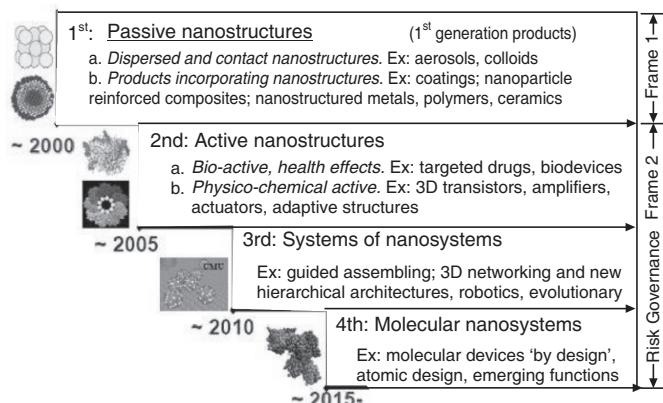


Fig. 39.1.

39.8 WHY NANOTECHNOLOGY?

With nanotechnology, it is possible to control matter on every important length scale, enabling tremendous new power in materials design. Furthermore, by tailoring the structure of materials in the range about 10^{-9} to 10^{-7} m one can systematically and significantly change specific properties at larger scales: material behavior can be engineered. Larger systems constructed of nanometer-scale components can have entirely new properties that have never before been identified in nature. It is also possible to produce composites that combine the most desirable properties of very different materials to obtain characteristics that are greatly improved over those that nature supplies or that appear in combinations nature does not produce.

As on date, nanotechnology is still in its infancy. It has a great potential for producing improvements and innovations in many areas of our lives, such as new and improved health

treatments; reduced use of harmful or scarce resources; cleaner, faster and safer manufacturing; quicker and smaller devices; increased life-cycle of products and many other improvements to existing products. It is particularly important because it involves little labour, land or maintenance; it is highly productive and inexpensive; and it requires only modest amounts of materials and energy.

The possibilities to create new things appear limitless. The principal domains which will be affected by developments in nanotechnology are:

- (i) **Materials:** New materials harder, more durable and resistant, lighter and less expensive.
- (ii) **Electronics:** Electronic components will become smaller and smaller, allowing the design of more powerful computers.
- (iii) **Energy:** A vast increase in the potential of solar energy generation
- (iv) **Health and biotechnology:** Great expectations are held in the areas of prevention, diagnostics and treatment.

Developments in these domains would impact a broad range of industries such as pharmaceuticals, cosmetics, consumer appliances, communications, security and safety, and space exploration.

39.9 PRODUCTION TECHNIQUES

Two main techniques are used in nanotechnology. They are (i) bottom-up technique and (ii) top-down technique.

(i) **Bottom-up technique** is a technique in which materials and devices are built up atom by atom. Bottom-up manufacturing would provide components made of single molecules, which are held together by covalent forces that are far stronger than the forces that hold together macro-scale components. Furthermore, the amount of information that could be stored in devices build from the bottom-up approach would be enormous.

(a) Molecular self-assembly

Molecular self-assembly is the assembly of molecules without guidance or management from an outside source. Self assembly is a manufacturing method used to construct things at the microscale, which is comprised of structures with atleast one dimension that is less than 100 μm . Many biological systems use self assembly technique to assemble various molecules and structures. Living systems are able to live because of the vast amount of highly ordered **molecular machinery** from which they are built. The central dogma of molecular biology states that the information required for building a living cell or organism is stored in the DNA. This information is transferred from the DNA to the proteins by the processes called **transcription** and **translation**. These processes are all executed by various biomolecular components, mostly protein and nucleic acids. Since the goal of nanotechnology is molecular and atomic precision, nanotechnology has much to learn from nature. Copying, borrowing and learning tricks from nature is one of the primary techniques used by nanotechnology and has been termed **biomimetics**. In self-assembly the final desired structure is encoded in the shape and properties of the molecules that are used. The synthesis of molecules for self-assembly often involves a chemical process called **convergent synthesis**.

(b) Positional Assembly

Drexler and other researchers have proposed that, instead of biomimetics, advanced nanotechnology ultimately could be based on mechanical engineering principles, namely, a manufacturing technology based on the mechanical functionality of these components (such as gears, bearings, motors, and structural members) that would enable programmable,

positional assembly to atomic specification. The basic steps involved in this technology are already mentioned in Art.39.5.

The positional assembly implies development of molecular robotics e.g., robotic devices that are molecular both in their size and precision. These molecular scale positional devices are likely to resemble very small versions of their everyday macroscopic counterparts.

One robotic arm assembling molecular parts is going to take a long time to assemble anything large. Therefore, lots of robotic arms will be needed. This is known as **massive parallelism**. While earlier proposals achieved massive parallelism through self-replication, future molecular manufacturing systems are expected to be designed to use some form of **convergent assembly**. Convergent assembly means that smaller parts can be assembled into larger parts; larger parts can be assembled into still larger parts, and so forth. In the present case, vast numbers of small parts are assembled by vast numbers of small robotic arms into larger parts; those larger parts are assembled by larger robotic arms into still larger parts, and so forth. If the size of the parts doubles at each iteration, we can go from one nanometer parts to one meter parts in only 30 steps.

The first step would be to develop nanoscopic machines, called **assemblers**, that scientists can program to manipulate atoms and molecules at will.

In order for molecular manufacturing to be practical, you would need trillions of assemblers working together simultaneously. Eric Drexler believes that assemblers could first replicate themselves, building other assemblers. Each generation would build another, resulting in exponential growth until there are enough assemblers to produce objects.

Trillions of assemblers and replicators could fill an area smaller than a cubic millimeter, and could still be too small for us to see with the naked eye. Assemblers and replicators could work together to automatically construct products, and could eventually replace all traditional labor methods. This could vastly decrease manufacturing costs, thereby making consumer goods plentiful, cheaper and stronger.

(ii) **Top-down technique** is a technique in which devices are fabricated by removing existing material from larger entities. The current top-down method for manufacturing involves the construction of parts through methods such as cutting, carving and molding. Using these methods we have been able to fabricate a remarkable variety of machinery and electronic devices. However, the sizes at which we can make these devices are severely limited by our ability to cut, carve and mold. We describe here three different techniques used in making nanostructures.

(1) **Lithography:** Optical lithography is a technique that generates patterns on the surface which is used in fabricating integrated circuits. To fabricate devices smaller than 100 nm, UV light is required. Lithography is the transfer of a pattern to a photosensitive material by selective exposure to a radiation source such as light. If we selectively expose a photosensitive material to radiation (e.g. by masking some of the radiation) the pattern of the radiation on the material is transferred to the material exposed, as the properties of the exposed and unexposed regions differ.

In lithography, the photosensitive material used is typically a photoresist. When resist is exposed to UV radiation source, the chemical resistance of the resist to developer solution changes. If the resist is placed in a developer solution after selective exposure to the light source, it will etch away one of the two regions (exposed or unexposed). If the exposed material is etched away by the developer and the unexposed region is resilient, the material is considered to be a positive resist (shown in Fig. 39.2a). If the exposed material is resilient to the developer and the unexposed region is etched away, it is considered to be a negative resist (shown in Fig. 39.2b).

Electron Beam Lithography

Electron beam lithography is used in making nanostructures. A finely focused beam of electrons is used in this method to scan a specific pattern over the surface of a material. It can produce a structure having 10 nm resolution. This process is not amenable for large scale production, since this process requires the electron beam to hit the surface of the material point by point in a serial manner.

(2) Nanoimprint Lithography

(NIL): Nanoimprint

Lithography is a low-cost, high-production rate technology. Unlike conventional lithography, nanoimprint lithography patterns a resist by physically deforming the resist shape with a mold having a nanopattern on it. The process is illustrated in Fig. 39.3.

A mold having a nanopattern on it is pressed into a thin resist coating on a substrate (Fig. 39.3 a). It creates the nanopattern in the resist. Then the mold is lifted off. The resist material in the compressed regions is etched away (Fig. 39.3 c).

Nanoimprint lithography can produce patterns having 10 nm resolution at low cost because it does not require the use of sophisticated radiation beams for generating patterns.

(3) Dip Pen Nanolithography (DPN): Dip Pen Nanolithography (DPN) is a lithography technique where photoresist is not used. In DPN the tip of an atomic force microscope cantilever acts as a “pen,” which is coated with a chemical

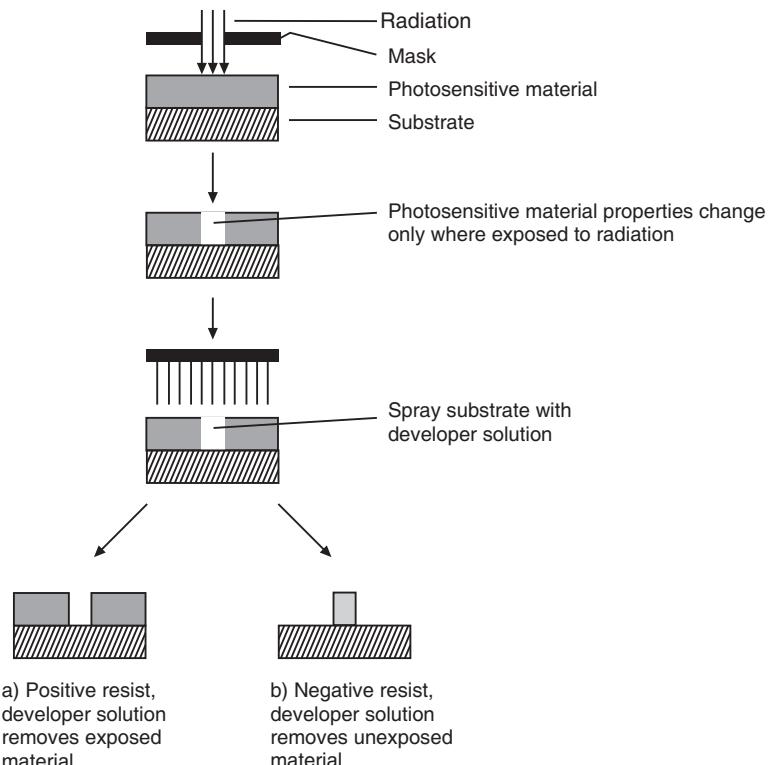


Fig. 39.2. (a) Pattern definition in positive resist, (b) Pattern definition in negative resist.

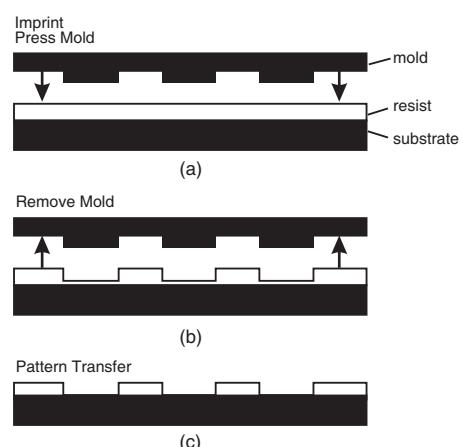


Fig. 39.3: Schematics of steps in a nanoimprint lithography process: (a) a mold of a hard material made by electron beam lithography is pressed into a resist, a softer material, to make an imprint; (b) the mold is then lifted off; (c) the remaining soft material is removed by etching from the bottom of the grooves.

compound or mixture acting as an “ink,” and put in contact with a substrate, the “paper”. The atomic force microscope tip transfers molecules to the surface of the substrate via a solvent meniscus (Fig. 39.4). It is a single step process to create nanopatterns on a substrate. This technique allows surface patterning on scales of under 100 nanometres.

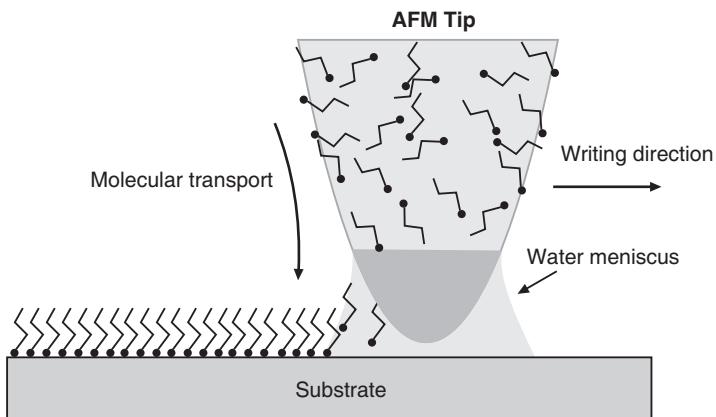


Fig. 39.4

39.10 TOOLS

The scanning tunnelling electron microscope (STM) and the atomic force microscope (AFM) are the scanning probes that enabled development of nanotechnology. There are other types of scanning probe microscopy, all based on the idea of the STM that make it possible to observe structures at the nanoscale. The term ‘microscope’ in these names is actually a misnomer because it implies looking, while in fact the information is gathered by “feeling” the surface with a mechanical probe.

1. Scanning tunnelling electron microscope (STM): The instrument was invented in the early 1979 by Gerd Binnig and Heinrich Rohrer, who were awarded the 1986 Nobel prize in physics for their work. The scanning tunnelling electron microscope uses electron tunnelling to produce images of surfaces down to the scale of individual atoms. If two conducting samples are brought in close proximity, with a small but finite distance between them, electrons from one sample flow into the other if the distance is of the order of the spread of the electronic wave into space. One says that the electrons “tunnel” through the barrier into the adjacent sample. For electrons, the barrier width which may be overcome via a tunneling process is of the order of nm, i.e. of the order of several atomic spacings. The probability of an electron to get through the tunneling barrier decreases exponentially with the barrier width, i.e. the so called tunneling current is extremely sensitive measure of the distance between two conducting samples. The STM makes use of this sensitivity.

Working: The schematic diagram of a scanning tunnelling electron microscope is shown in Fig. 39.5 (a). In the scanning tunnelling microscope the sample is scanned by a very fine metallic tip. The tip is mechanically connected to the scanner, an XYZ positioning device. The sharp metal needle is brought close to the surface to be imaged. The distance is of the order of a few angstroms. A bias voltage is applied between the sample and the tip. When the needle is at a positive potential with respect to the surface, electrons can tunnel across the gap and set up a small “tunneling current” in the needle. The tunneling current is exponentially dependent on the junction width and increases by a decade per Angstrom as the tip is brought closer to the surface. In typical systems a tip-sample separation of 0.5nm will produce currents of $\sim 1\text{nA}$ for biases of 1V. This feeble tunneling current is amplified and measured. With the help of the tunneling current the feedback electronics keeps the distance between tip and sample constant. By scanning the tip along the surface and monitoring the current, one can resolve the surface topography directly underneath the tip. The tip motion is controlled using piezoelectric materials, which can position the tip with sub-Angstrom resolution. The

sensitivity of the STM is so large that electronic corrugation of surface atoms and the electron distribution around them can be detected.

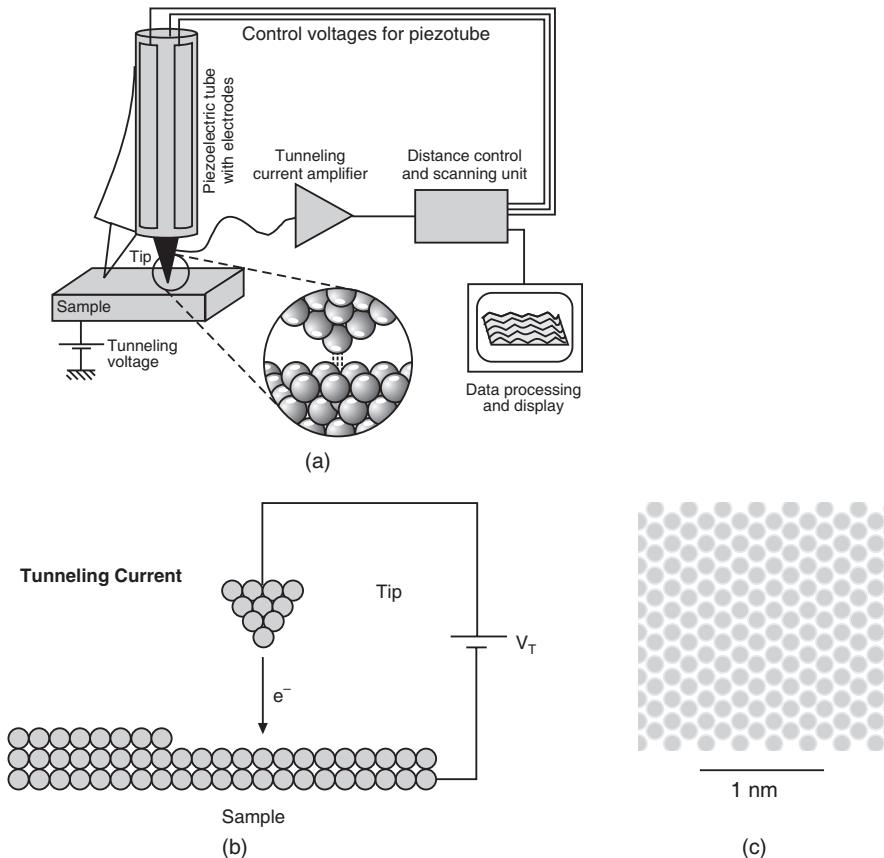


Fig. 39.5. (a) The schematic diagram of STM (b) The tunneling tip is scanned over the specimen, producing an image of the tunneling current (c) An STM image of the Ag (001) surface; each bright spot corresponds to a single Ag atom.

An image of the Ag (001) surface taken in this manner is shown in Fig. 39.5 (c). Each bright spot in the image corresponds to a single Ag atom.

Molecular Manipulation

Another powerful STM capability is the ability to move atoms and molecules. An adsorbed atom is held on the surface by chemical bonds with the atoms of the surface. When the tip of the STM is placed close enough to the surface, the interaction of the tip and the atom becomes greater than that between the atom and the surface. The atom can be dragged along by the tip wherever it is moved along the surface. At any point in the scan the atom can be reattached to the surface by increasing the separation between the tip and the surface (Fig. 39.6). In this way adsorbed atoms can be rearranged on the surface of materials and structures can be built on the surfaces atom by atom. The manipulation is to be done under ultra-high-vacuum conditions in order to keep the surface of the material clean. Secondly, the surface of the material has to be held at liquid helium temperature in order to reduce the thermal vibrations which disturb the arrangement of atoms on the surface.

2. Atomic Force Microscope: The atomic force microscope (AFM) was invented in 1986 by Binnig, Quate and Gerber. The Atomic Force Microscope was developed to overcome a basic drawback with STM - that it can only image conducting or semiconducting surfaces. The AFM, however, has the advantage of imaging almost any type of surface, including polymers, ceramics, composites, glass, and biological samples.

The AFM consists of a microscale cantilever shaped much like a diving board with a sharp tip (probe) at its end which is used to scan the sample surface (Fig.39.7). The cantilever is typically made of silicon or silicon nitride. The radius of curvature of the tip is of the order of nanometers. A laser is positioned such that its light strikes at an oblique angle at the very end of the cantilever. When the tip is brought into proximity of a sample surface, the tip is repelled by or attracted to the surface. These forces between the tip and the sample lead to a deflection of the cantilever according to Hooke's law. As the cantilever bends the light from the laser is reflected onto an array of photodiodes. A plot of the laser deflection versus tip position on the sample surface provides the resolution of the hills and valleys that constitute the topography of the surface.

Using an interferometer technique, changes in the bending of the cantilever can be measured. Since the cantilever obeys Hooke's Law for small displacements, the interaction force between the tip and the sample can be found.

The AFM can work with the tip touching the sample (contact mode), or the tip can tap across the surface (tapping mode) much like the cane of a blind person.

Advantages:

- Because its operation does not require a current between the sample surface and the tip, the AFM can move into potential regions inaccessible to the Scanning Tunnelling Microscope (STM) or image fragile samples which would be damaged irreparably by the STM tunnelling current.
- Since the atomic force microscope does not depend on a current, it can be used to visualize the surfaces of conductors as well as non-conducting materials. Insulators, organic materials, biological

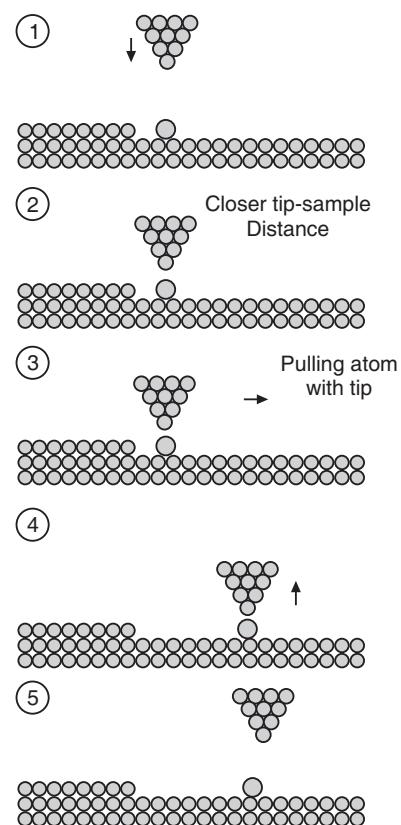


Fig. 39.6. Schematic diagram of molecular manipulation with the STM

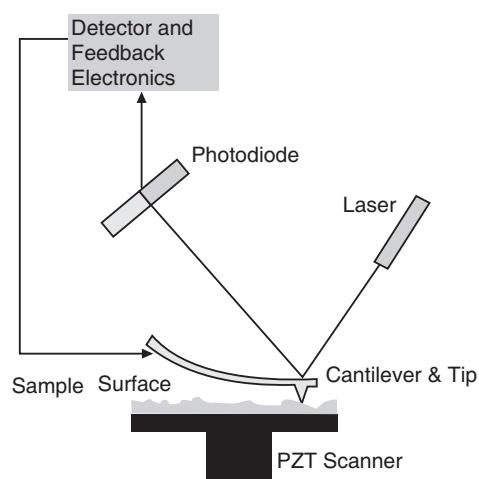


Fig. 39.7. Beam deflection system, using a laser and photodetector to measure the beam position.

macromolecules, polymers, ceramics and glasses are some of the many materials which can be imaged in different environments, such as liquids, vacuum, and low temperatures.

Other measurements can be made using modifications of the AFM. These include variations in surface microfriction with a lateral force microscope (LFM), orientation of magnetic domains with a magnetic force microscope (MFM), and differences in elastic modulii on the micro-scale with a force modulation microscope (FMM). A very recent adaptation of the SPM has been developed to probe differences in chemical forces across a surface at the molecular scale. This technique has been called the chemical force microscope (CFM). The AFM and STM can also be used to do electrochemistry on the microscale.

39.11 NANOMATERIALS

Having understood the goals and tools of nanotechnology, we shall now look at the materials that go into the nanotechnology products. Materials at atomic, molecular and macromolecular scales are known as nano-scale materials which exhibit properties that differ significantly from those of bulk materials. Nanotechnology exploits the strange properties of the nano-scale materials to produce new useful products.

We categorize nanomaterials as those which have structured components with at least one dimension less than 100nm. Materials that have one dimension in the nanoscale (and are extended in the other two dimensions) are **nanolayers**, such as thin films or surface coatings. Some of the features on computer chips come in this category. Materials that are nanoscale in two dimensions (and extended in one dimension) include **nanowires** and **nanotubes**. Materials that are nanoscale in three dimensions are **nanoparticles**, for example precipitates, colloids and quantum dots. Nanocrystalline materials, made up of nanometre-sized grains, also fall into this category. Some of these materials have been available for some time; others are genuinely new.

Nanoparticle research is currently an area of intense scientific interest due to a wide variety of potential applications in biomedical, optical and electronic fields.

According to Siegel, Nanostructured materials are classified as Zero dimensional, one dimensional, two dimensional, three dimensional nanostructures (Fig.39.8).

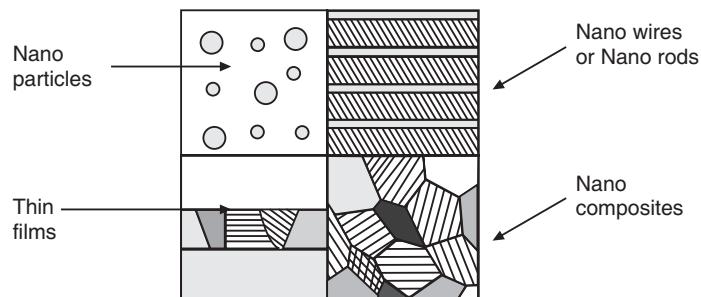


Fig. 39.8. Classification of nanomaterials following Siegel

39.12 NANOLAYERS

One-dimensional nanomaterials, such as thin films and engineered surfaces, have been developed and used for decades in fields such as electronic device manufacture, chemistry and engineering. In the silicon integrated-circuit industry, for example, many devices rely on thin films for their operation, and control of film thicknesses approaching the atomic level is routine. Monolayers, that are one atom or molecule deep, are also routinely made and used in chemistry. The formation and properties of these layers are reasonably well understood from the atomic level upwards, even in quite complex layers such as lubricants. Advances are being made in the control of the composition and smoothness of surfaces, and the growth of films.

39.12.1 Production of Nanolayers

Vapour deposition refers to any process in which materials in a vapour state are condensed through condensation, chemical reaction, or conversion to form a solid material. There are two categories of vapour deposition processes: physical vapour deposition (PVD) and chemical vapour deposition (CVD). In PVD processes, the workpiece is subjected to plasma bombardment. In CVD processes, thermal energy heats the gases in the coating chamber and drives the deposition reaction.

1. Physical vapour deposition (PVD)

There are different methods of physical vapour deposition in use. The methods are evaporation, sputtering, ion plating and laser ablation. We study here one of the methods, namely evaporation method.

(a) Evaporation: In evaporation the substrate is placed inside a vacuum chamber, in which a source of the material to be deposited is also located. The source material is then heated to the point where it starts to boil and evaporate. The vacuum is required to allow the molecules to evaporate freely in the chamber, and they subsequently condense on all surfaces. This principle is

the same for all evaporation technologies, only the method used to heat (evaporate) the source material differs. There are two popular evaporation technologies, which are e-beam evaporation and resistive evaporation each referring to the heating method.

(i) Resistive evaporation: The resistive evaporation is carried out in an evacuated bell jar (Fig. 39.9). The material to be deposited is taken in a crucible around which a high resistance wire is wrapped. When a current is passed through the wire, it heats up the material and vaporizes it. The process is carried out at pressure of less than 0.1 Pa (1 m Torr) and in vacuum levels of 10 to 0.1 MPa. The substrate temperature ranges from ambient to 500°C. For vacuum deposition, a reasonable deposition rate can be obtained if the vaporization rate is fairly high. A useful deposition rate is obtained at a vapour pressure of 1.3 Pa (0.01 Torr). Vapour phase nucleation can occur in dense vapour cloud by multibody collisions. The atoms are passed through a gas to provide necessary collision and cooling for nucleation. The advantages associated with vacuum deposition process are high deposition rates and economy.

(ii) e-beam evaporation: In e-beam evaporation, an electron beam is aimed at the source material causing local heating and evaporation. In resistive evaporation, a tungsten boat, containing the source material, is heated electrically with a high current to make the material evaporate. Many materials are restrictive in terms of what evaporation method can be used (i.e. aluminum is quite difficult to evaporate using resistive heating), which typically relates to the phase transition properties of that material. A schematic diagram of a typical system for e-beam evaporation is shown in Fig.39.10.

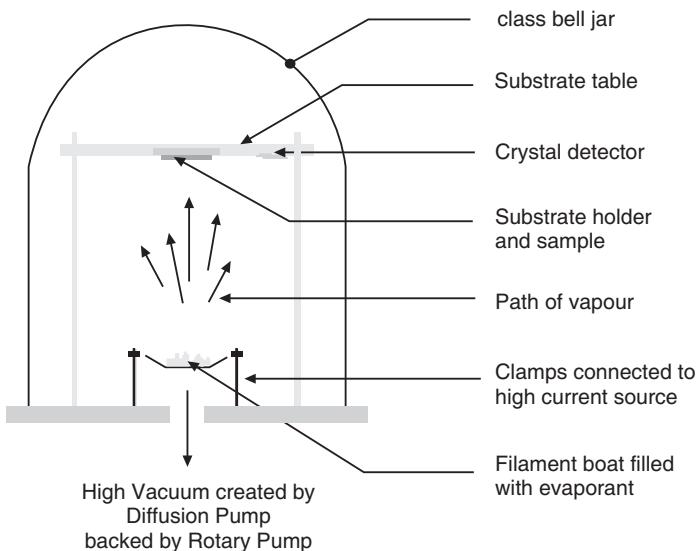


Fig. 39.9. Typical system for resistive evaporation of materials

2. Sputtering: Sputtering is a technology in which the material is released from the source at much lower temperature than evaporation. The substrate is placed in a vacuum chamber with the source material, named a target, and an inert gas (such as argon) is introduced at low pressure. A gas plasma is struck using an RF power source, causing the gas to become ionized. The ions are accelerated towards the surface of the target, causing atoms of the source material to break off from the target in vapour form and condense on all surfaces including the substrate. As for evaporation, the basic principle of sputtering is the same for all sputtering technologies. The differences typically relate to the manner in which the ion bombardment of the target is realized. A schematic diagram of a typical RF sputtering system is shown in Fig. 39.11.

3. Chemical Vapour Deposition (CVD)

CVD is a well known process in which a solid is deposited on a heated surface via a chemical reaction from the vapour or gas phase. CVD reaction requires activation energy to proceed. In thermal CVD the reaction is activated by a high temperature. A typical apparatus comprises of gas supply system, deposition chamber and an exhaust system. The deposition chamber is an evacuated chamber (Fig.39.12). A wafer is kept on a carrier and heated to a temperature between 350 and 800°C. One or several species of gases are admitted into the chamber through an inlet till a medium gas pressure is built up in the chamber. Now a dissociation or reaction between two species takes place. In both cases, a newly formed molecule adheres to the wafer surface and participates in the formation of a nanolayer. As an example let us take silane, SiH_4 . The gas dissociates into elementary silicon, which partly adheres to the wafer surface and partly to hydrogen which is removed by the exhaust pump.

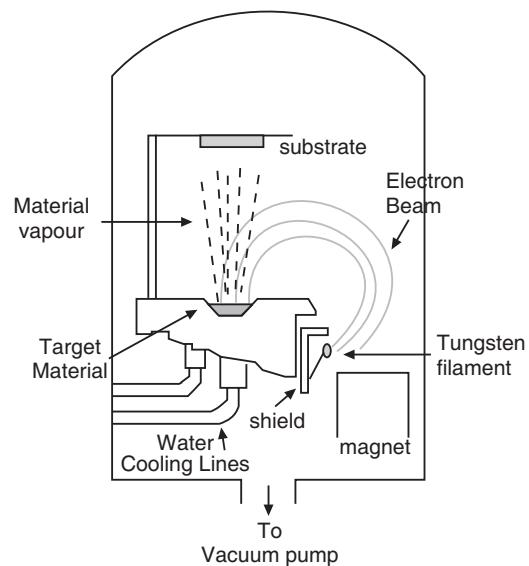


Fig. 39.10. Typical system for e-beam evaporation of materials.

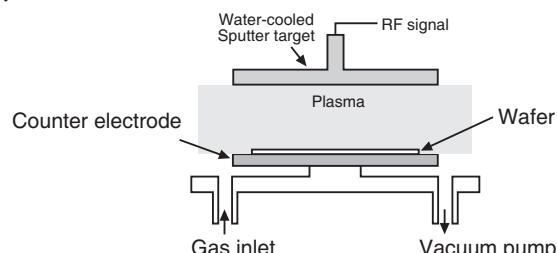


Fig. 39.11. Typical RF sputtering system

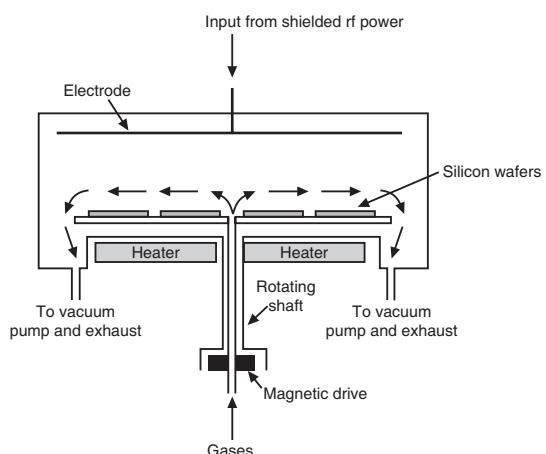


Fig. 39.12

4. Vapour Phase Epitaxy (VPE)

Epitaxy technology is quite similar to CVD processes. The technology is primarily used for deposition of silicon. If the substrate is an ordered semiconductor crystal (i.e. silicon, gallium arsenide), it is possible with this process to continue building on the substrate with the same crystallographic orientation with the substrate acting as a seed for the deposition. If an amorphous/polycrystalline substrate surface is used, the film will also be amorphous or polycrystalline.

There are several technologies for creating the conditions inside a reactor needed to support epitaxial growth, of which the most important is Vapour Phase Epitaxy (VPE). In this process, a number of gases are introduced in an induction heated reactor where only the substrate is heated. The temperature of the substrate typically must be at least 50% of the melting point of the material to be deposited.

Epitaxy is a widely used technology for producing silicon on insulator (SOI) substrates. A schematic diagram of a typical vapour phase epitaxial reactor is shown in the figure below.

5. Molecular Beam Epitaxy

Molecular beam epitaxy is a technique for epitaxial growth via the interaction of one or several molecular or atomic beams that occurs on a surface of a heated crystalline substrate. In Fig. 39.14, a scheme of a typical MBE system is shown. The solid sources materials are placed in evaporation cells to provide an angular distribution of atoms or molecules in a beam. The substrate is heated to the necessary temperature and, when needed, continuously rotated to improve the growth homogeneity.

In ultra-high vacuum, a beam of atoms or, more general, a beam of molecules is directed towards a crystalline substrate such that the atoms or molecules stick at the substrate's surface forming a new layer of deposited material. The difference between MBE and other material deposition methods as such e.g. thermal vacuum evaporation is as follows:

MBE does not only deposit material like it is done by conventional evaporation techniques, but using the very low rates of impinging atoms, migration on

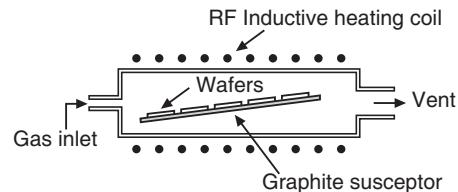


Fig. 39.13. Typical vapour phase epitaxial reactor.

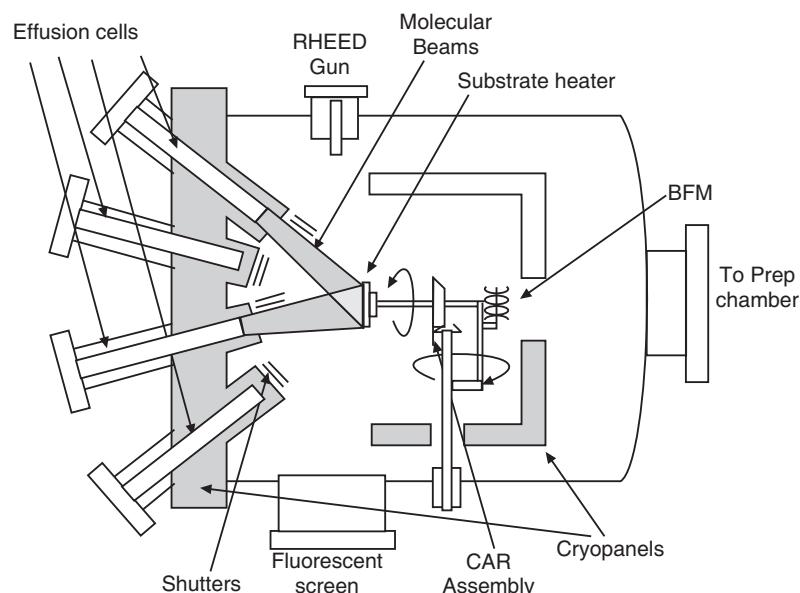


Fig. 39.14. A typical MBE system

the surface and subsequent surface reactions ensure the controlled epitaxial growth of a new layer. Every atom reaching the surface of the heated substrate has enough time to migrate around and find its place to build up a new crystal lattice.

At the left hand side there are the effusion cells to provide the molecular beam for either the bulk constituents or the dopants. These cells can be thermal evaporation cells (Knudsen cells), cells for gaseous media or plasma sources as well. In front of them is a shutter, this means a plate which could be brought into the beam for ‘switching’ the beam on and off. Opposite to the cells and the shutters is the substrate, mounted on a heatable and rotatable substrate holder.

The whole system is in ultra-high vacuum environment to guarantee formation of a molecular ‘beam’ (without vacuum, the atoms or molecules leaving the effusion cells will be scattered at residual gas molecules and never form a beam directed towards substrate) and purity and therefore quality of the grown layer. Often there are instruments for in-situ analysis like RHEED attached to the growth chamber. A cryopanel around the sample and the cells absorbs residual gases and provides a clean substrate environment.

39.12.2 Nanolayer Devices

A. Giant Magnetoresistance (GMR)

Magnetoresistance is a phenomenon in which the application of a magnetic field changes the resistance of a material. The magnetoresistance effect occurs in metals only at very high magnetic fields and at low temperatures. For example, in pure copper at a 4K a field of 10 Wb/m² produces a factor of 10 change in the resistance. Because of the large fields and low temperatures, magnetoresistance in metals had limited practical applications. In 1988 a new effect called **giant magnetoresistance (GMR)** was discovered by A. Fert and P. Grünberg in synthetically fabricated magnetic metallic multilayers of nanometer

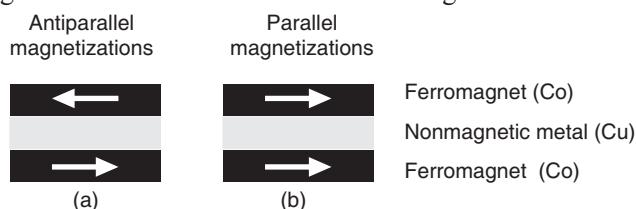


Fig. 39.15

thickness. Alternate layers of nanometer thickness of a ferromagnetic material and a nonferromagnetic metal such as Fe/Cr and Co/Cu are deposited on a substrate. The application of a magnetic field to such a multilayer results in a significant reduction of the electrical resistance of the multilayer. This effect was found to be much larger than other magnetoresistive effects that had ever been observed in metals and was, therefore, called “giant magnetoresistance”.

A schematic of a three-layered structure is illustrated in Fig. 39.15 (a). The bottom and top layers are Co layers which are ferromagnetic layers and the middle layer is a Cu layer which is a conductive, nonmagnetic interlayer. In ferromagnetic materials, the spins from unpaired electrons spontaneously align parallel to give the material a total magnetic moment. In the absence of the magnetic field the magnetizations of the ferromagnetic layers are antiparallel. Copper is normally an excellent conductor, but when it is only a few atoms thick, electron scattering causes its resistance to increase significantly. Resistance to the flow of current in a direction perpendicular to such a multilayer system is high.

When an external magnetic field is applied parallel to the layers, the magnetic moments in

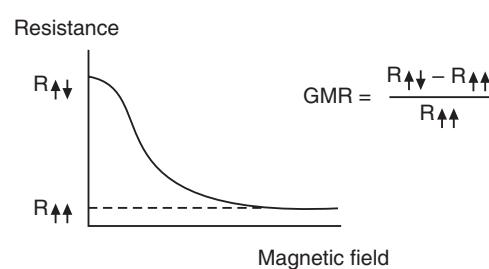


Fig. 39.16

successive ferromagnetic layers align in the direction of the magnetic field (see Fig. 39.15b) and the resistance of the multilayer system decreases, as shown in Fig. 39.16.

The GMR effect occurs because of the dependence of electron mean free path on the orientation of the electron spin with respect to the direction of magnetization. A conduction electron moving through a ferromagnetic material passes through more easily if it encounters electron spins oriented in the same direction as its own spin orientation. Therefore, electrons whose spins are aligned along the direction of magnetization of ferromagnetic layer will travel further in the layer without being scattered (Fig. 39.17 b).

In a multilayer system, the first magnetic layer polarizes the electron spins. The

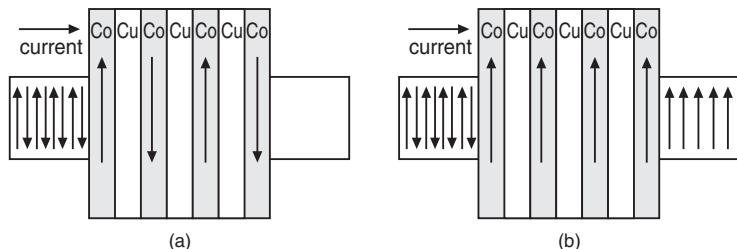


Fig. 39.17: Electrical conduction of multilayered magnetoresistive structure

second layer scatters the spins strongly if its moment is not aligned with the moment of the first layer. If the moment of the second layer is aligned, it allows the spins to pass (Fig. 39.17). The resistance therefore changes depending on whether the moments of the magnetic layers are antiparallel (high resistance) or parallel (low resistance).

Optimal layer thicknesses enhance magnetic-layer antiparallel coupling, which is necessary to keep the system in the high-resistance state when no field is applied. When an external field overcomes the antiparallel coupling, the moments in the magnetic layers align and reduce the resistance. For spin-dependent scattering to be a significant part of the total resistance, the layers must be thinner (to a magnitude of several nanometers) than the mean free path of electrons. In a typical GMR medical sensor the thickness of a conducting layer is approximately 3 nm. For reference, that is less than 10 atomic layers of copper, and less than one ten-thousandth the thickness of a piece of tissue paper.

Applications of GMR

The largest technological application of GMR is in the data storage industry. IBM was first to put on the market hard disks based on GMR technology, and nowadays all disk drives make use of this technology.

B. Vertical Cavity Surface Emitting Laser (VCSEL)

One-dimensional nanotechnology techniques involving precise growth of very thin semiconductor layers were developed during 1990s. Such nanostructuring resulted in the production of efficient VCSELs. VCSEL is a specialized laser diode which is used as optical source in optical communication. It has nanoscale layers of compound semiconductors grown into their structure – alternating dielectric layers as mirrors and quantum wells. Quantum wells confine the charge carriers in well-defined regions and provide the energy conversion into light at desired wavelengths. The details of construction and working of VCSEL are discussed in Chapter 26, Art. 26.13.2(b).

39.13 NANOPARTICLES

Nanoparticles are generally considered to be an aggregate of atoms or molecules bonded together with a radius of <100 nm. A cluster of 1 nm radius has approximately 25 atoms in it; but most of the atoms are on the surface of the cluster.

A bulk material should have constant physical properties regardless of its size, but at the nano-scale size-dependent properties are often observed. Two principal factors cause

the properties of nanomaterials to differ significantly from other materials: increased relative surface area, and quantum effects. These factors can change or enhance properties such as reactivity, strength and electrical characteristics. For bulk materials larger than one micrometer (or micron), the percentage of atoms at the surface is insignificant in relation to the number of atoms in the bulk of the material. As a nanoparticle decreases in size, a greater proportion of atoms are found at the surface compared to those inside. For example, a particle of size 30 nm has 5% of its atoms on its surface, at 10 nm 20% of its atoms, and at 3 nm 50% of its atoms. Thus nanoparticles have a much greater surface area per unit mass compared with larger particles. As growth and catalytic chemical reactions occur at surfaces, this means that a given mass of material in nanoparticulate form will be much more reactive than the same mass of material made up of larger particles. The interesting and sometimes unexpected properties of nanoparticles are therefore largely due to the large surface area of the material, which dominates the contributions made by the small bulk of the material.

An excellent example of this is the absorption of solar radiation in photovoltaic cells, which is much higher in materials composed of nanoparticles than it is in thin films of continuous sheets of material. In this case, the smaller the particles, the greater the solar absorption.

Moreover nanoparticles have been found to impart some extra properties to various day to day products. For example, the presence of titanium dioxide nanoparticles imparts what we call the self-cleaning effect, and the size being nanorange, the particles cannot be observed. Zinc oxide particles have been found to have superior UV blocking properties compared to its bulk substitute. This is one of the reasons why it is often used in the preparation of sunscreen lotions.

We study here about the properties of nanoparticles with particular reference to metal nanoclusters.

39.13.1 Metal Nanoclusters

Condensed “hard”matter nanoparticles are generally termed nanoclusters. A nanocluster is a nanometer sized particle made up of equal subunits. These subunits can be atoms of a single element, molecules or even combinations of atoms of several elements in subunits with equal stoichiometries (alloys, etc.).

E.g.: Nan , $(\text{SF}_6)_n$, $(\text{H}_2\text{O})_n$, $(\text{Cu}_3\text{Au})_n$, $(\text{ClCH}_3\text{C}_6\text{H}_3\text{CO}_2\text{H})_n$, $(\text{TiO}_2)_n$, . . .

The properties of nanoclusters are solely guided by the number of subunits they contain.

39.13.2 Properties of Metal Nanoclusters

1. Magic Numbers: Metal nanoparticles often show a size preference i.e. there are **magic numbers** of metal atoms. When lead metal nanoclusters were formed, magic numbers were observed at 7 and 10 atoms. It means that formation of clusters of 7 and 10 atoms is more likely than other clusters and these clusters are more stable than clusters of other sizes. The existence of magic numbers suggests that clusters can be viewed as superatoms.

2. Jellium Model: The **jellium model** regards a cluster of atoms as a large atom. The positive nuclear charge of each atom is assumed to be smeared out over the entire volume of the cluster while the valence electrons are free to move within this homogeneously distributed, positively charged background. The electron energy levels can be obtained by applying the Schrödinger equation for this system in a manner analogous to that for the hydrogen atom. The results showed the magic numbers correspond to those clusters having a size in which all the energy levels are filled.

3. Geometric structure: Usually the crystal structure of a large nanocluster is the same as the bulk structure of the material, with somewhat different lattice parameters (in general clusters are slightly contracted as compared to bulk). E.g. Cu clusters tend to have an FCC

structure. On the other hand, smaller clusters of Cu (e.g. Cu_{55} , Cu_{147}, \dots) have perfect icosahedral structures.

Though their bulk solid counterparts may form close packed crystal structures, metal nanoclusters may prefer icosahedral or decahedral structures, depending on their size (Fig. 39.18). This appears to be due to a size preference i.e **magic numbers** of metal atoms. For example a magic number for aluminium is 13. Instead of a face centered cubic structure, the Al_{13} cluster is an icosahedron of 12 aluminium atoms with an additional aluminium atom in the center.

4. Inter-particle spacing: A decrease in inter-particle spacing with decreasing size is typical of metal clusters. The effect of decreasing cluster size is to create more surface sites. This changes the surface pressure and result in change in the inter-particle spacing. This effect is illustrated in Fig. 39.19 for the case of Cu_n particles. It is seen that the inter-particle spacing decreases with size due to competition between the long-range electronic forces and the short range core-core repulsion.

5. Melting temperatures: A major fraction of the atoms in a nanometer-size particle is located at the surface. Therefore, the surface to volume ratio in particles is larger compared to the bulk material. The change in inter-particle spacing and

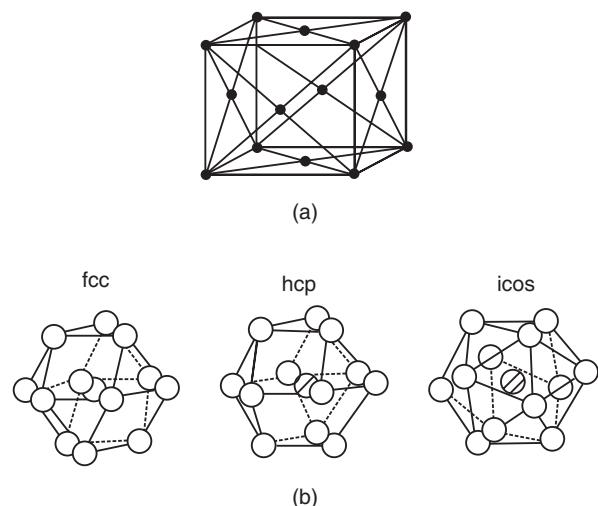


Fig. 39.18. (a) The unit cell of bulk aluminium (b) three possible structures of Al_{13} : a face centered cubic (FCC) structure, an hexagonal close packed (HCP) structure and an icosahedral (ICOS) structure

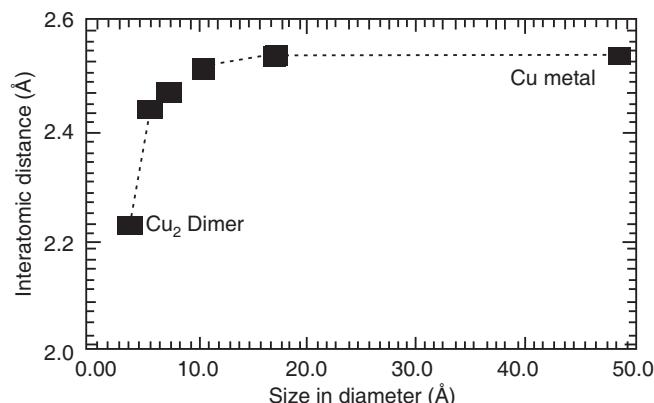


Fig. 39.19. Interatomic distance in Cu_n as a function of size

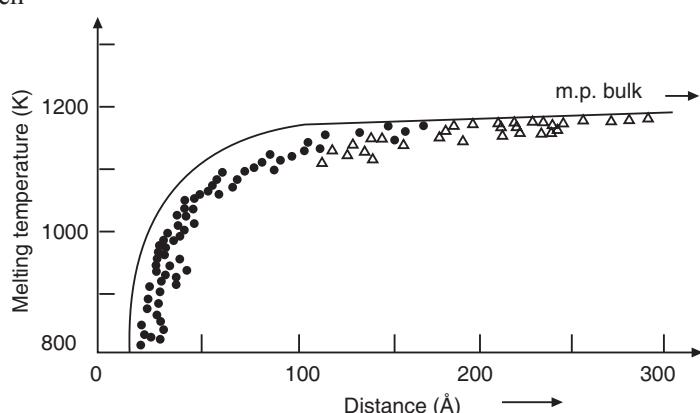


Fig. 39.20. Melting point of small Au_n particles as a function of size

the large surface to volume ratio in particles have a combined effect on material properties. This affects, for example, the thermodynamic properties of which the simplest example is the melting point. Fig. 39.20 shows the melting point of Au_n particles as a function of size. The melting point decreases with size and the rate of decrease increases substantially at very small sizes.

6. Reactivity and catalytic properties: In general, the reactivity of all nanomaterials is high as compared to that of bulk, due to their higher surface to volume.

Atoms and molecules have discrete energy levels or orbitals. However, in solid materials, these orbitals overlap to form wide energy bands. The electrons in the bands are delocalized and no longer belong to a particular atom. As the size is reduced from the bulk, the electronic bands in metals become narrower and the delocalized electronic states are transformed to more localized molecular bonds. One of the electronic quantities that is affected due to this change is the ionization potential. The ionization potential at small sizes is higher than the bulk work function and show marked fluctuations as a function of size. Fig. 39.21 shows this for the case of Fe_n clusters reacting with hydrogen.

The large surface to volume ratio and the variations in geometry and electronic structure have a strong effect on catalytic properties.

7. Electronic structure:

Atoms and molecules have discrete energy levels or orbitals. However, in solid materials, these orbitals overlap to form bands. In the case of a metal, one of these bands is partially filled. This makes a metal electrically-conductive. As the bulk shrink in size there is a dramatic change in these bands, as the continuous density of states in bulk is replaced with a set of discrete energy levels, as illustrated in Fig. 39.22.

8. Optical properties: The energy level separations in a nanocluster are dependent on the size of the clusters, and material, which therefore affect the energies needed for the transitions of electrons to excited states.

Clusters of different sizes will therefore have different absorption spectra but they lie generally in the visible range of the spectrum. Hence, the clusters of different sizes exhibit different colours. For e.g., nanoscale gold particles can be orange, purple, red or greenish depending on the size of the cluster.

9. Magnetic properties: A normal ferromagnet contains domains each

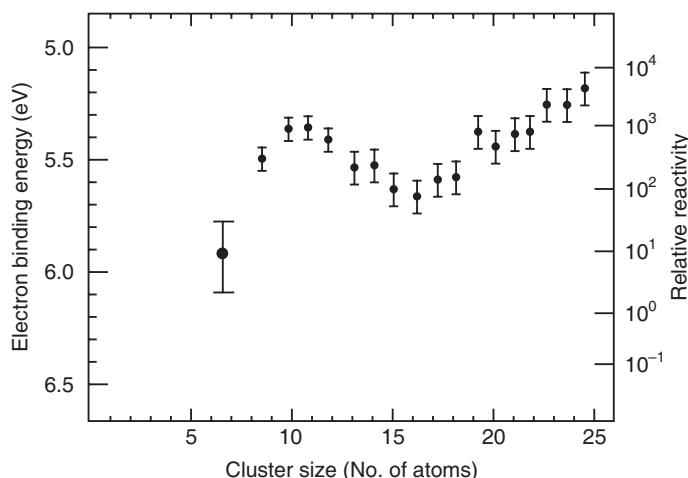


Fig. 39.21. Ionization potential and reactivity of Fe_n clusters

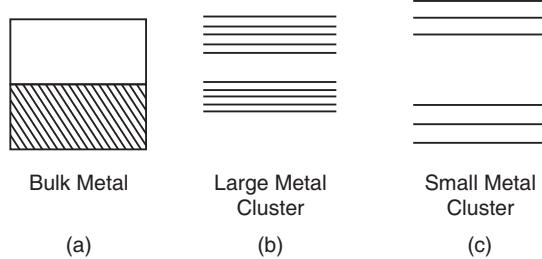


Fig. 39.22. Changes in energy band structure as the size of the cluster decreases

containing several thousand atomic spins. Within the domains the spins are aligned in a single direction while different domains point in different directions. However, nanoclusters of magnetic materials exhibit totally new class of magnetic properties. In a cluster the magnetic moment of each atom will interact with the moments of the other atoms, and can force all the moments to align in one direction with respect to some symmetry axis of the cluster – the cluster will have a net magnetic moment and thus it will be magnetized. As cluster size decreases it therefore becomes easier for them to exhibit ferromagnetic behaviour. In some cases, even clusters made up of nonmagnetic atoms can have a net magnetic moment. For instance rhenium clusters show a pronounced increase in their magnetic moment when they contain less than 20 atoms. For clusters with less than 15 atoms these moments are fairly large.

Fig. 39.23 shows the plot of the magnetic moment per atom of a cluster on the number of atoms in the cluster. It is seen that the magnetic moment increases as the cluster size decreases. Thus, nanoclusters are more magnetic than the bulk material. At small sizes, the clusters become spontaneously magnetic and the magnetism disappears in clusters containing more than 80 atoms. Even the clusters of nonmagnetic solids are found to be magnetic.

10. Mechanical properties: Most metals are made up of small crystalline grains; the boundaries between the grains slow down or arrest the propagation of defects when the material is stressed, thus giving it strength. If these grains can be made very small, or even nanoscale in size, the interface area within the material greatly increases, which enhances its strength. For example, nanocrystalline nickel is as strong as hardened steel.

The intrinsic elastic moduli of nanostructured materials are essentially the same as those for conventional grain size materials until the grain size becomes very small, e.g. < 5 nm, such that the number of atoms associated with the grain boundaries and triple junctions become very large as shown in Fig. 39.24 for noncrystalline Fe.

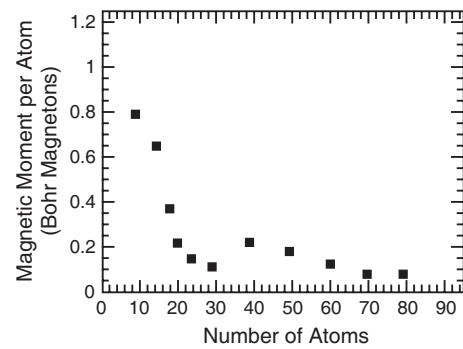


Fig. 39.23: Plot of the magnetic moment per atom of rhenium nanoparticles versus the number of atoms in the particle.

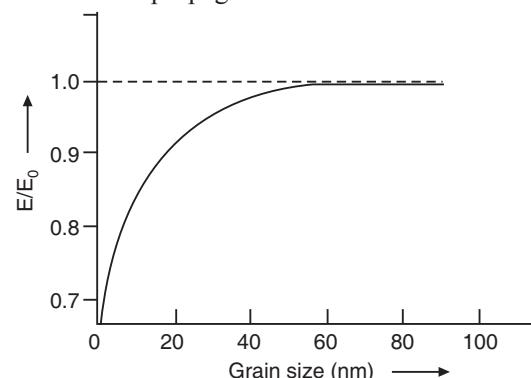


Fig. 39.24: Ratio of Young's modulus (E) of nanocrystalline materials to (E_0) of conventional grain size materials as a function of grain size.

39.13.3 Synthesis of Nanoparticles

There are a wide variety of methods to produce nanoparticles. A few commonly used methods are outlined here.

1. High-energy Ball milling: High energy ball milling is a *top-down approach* technique. Coarse grained materials are crushed mechanically in rotating drums by hard steel and tungsten carbide balls. The grain size in powder samples are reduced to nanometer range by mechanical deformation produced by ball milling process. Magnetic and catalytic nanoparticles are usually produced by this method.

In this method, a container is filled with stainless steel balls of a few millimeters in diameter (Fig. 39.25). The material to be crushed is added in the form of a powder of about 50 μm diameter grain size.

After filling the container with liquid nitrogen, a rotating shaft grinds the material. The grinding periods are within the range of minutes to some 100 hours. This process is simple. However, the difficulty in this technique is that we cannot be sure that all the particles are broken down to the required particle size. Further, during mechanical attrition, contamination by the milling tools (Fe) and atmosphere (trace elements of O₂, N₂, in rare gases) can be

a problem. The use of tungsten carbide component and inert atmosphere and /or high vacuum processes has reduced impurity levels to within acceptable limits. Common drawbacks include low surface, highly poly disperse size distribution, and partially amorphous state of the powder. The main advantage of this top-down approach is high production rates of nanopowders.

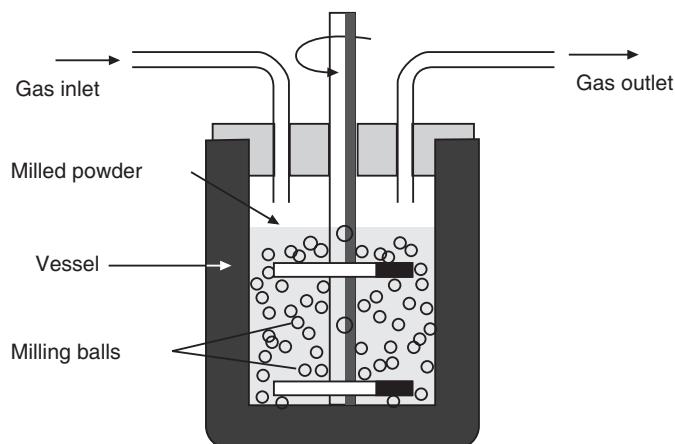


Fig. 39.25

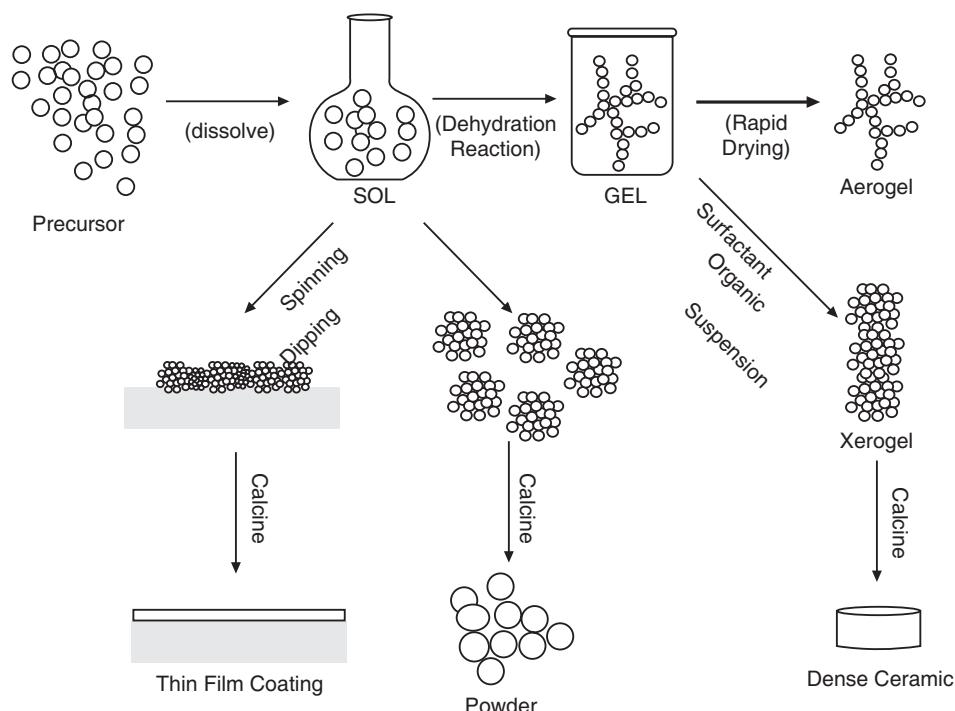


Fig. 39.26: Schematic representation of sol-gel process of synthesis of nanomaterials

2. Sol-gel process: Sol-gel method of synthesizing nanomaterials is very popular amongst chemists and is widely employed to prepare oxide materials. A *sol* is a solution with particles suspended in it. When the particles in the sol form long *polymers* (chains) that span the entire sol, a *gel* is formed. The sol-gel process is a *bottom-up* approach technique. In this process, the starting material is processed to form a dispersible oxide and a colloidal suspension (sol) of the particles of the metal compound is prepared first and then converted into a gel. The gel so formed is a network in a continuous liquid phase. Removal of the liquid from the sol yields the gel, and the sol/gel transition controls the particle size and shape. Calcination of the gel produces the oxide.

The sol-gel formation occurs in four stages.

- Hydrolysis
- Condensation
- Growth of particles
- Agglomeration of particles

Production of SiO_2 is an example of this process. The sol-gel process may be summarized in Fig.39.26.

Step 1: A stable solution of the alkoxide or solvated metal precursor (the sol) is formed.

Step 2: An oxide- or alcohol- bridged network (the gel) forms by a polycondensation or polyesterification reaction.

Step 3: The polycondensation reactions continue until the gel transforms into a solid mass, accompanied by contraction of the gel network and expulsion of solvent from gel pores.

Step 4: Drying of the gel, when water and other volatile liquids are removed from the gel network. If the solvent (such as water) is extracted under supercritical or near super critical conditions, the product is an *aerogel*. If isolated by thermal evaporation, the resulting product is termed a *xerogel*.

Step 5: Calcining the xerogel at temperatures up to 800°C stabilizes the gel against rehydration.

Aerogels are porous and extremely light, but they can withstand 100 times their weight. If the gelled spheres are calcined, one obtains powder. If the gel is collected on a surface, a thin film is obtained.

The interest in the sol-gel synthesis method arises due to the possibility of synthesizing nonmetallic inorganic materials like glasses, glass ceramics or ceramic materials at very low temperatures compared to the high temperature process required by melting glass or firing ceramics.

The major technical difficulties to overcome in developing a successful bottom-up approach is controlling the growth of the particles and then stopping the newly formed particles from agglomerating.

Sol-gel synthesis is superior of all the available processes because it can produce large quantities of nanomaterials at relatively low cost. It synthesizes almost any material, co-synthesize two or more materials simultaneously, coat one or more materials onto other materials (metal or ceramic particulates, and three-dimensional objects), produce extremely homogeneous alloys and composites, synthesize ultra-high purity (99.9999%) materials, tailor the composition very accurately even in the early stages of the process, precisely control the microstructure of the final products, and precisely control the physical, mechanical, and chemical properties of the final products.

3. Inert gas condensation: Gas condensation was the first technique used to synthesize nanocrystalline metals and alloys. This technique was pioneered by Gleiter and co-workers. In this technique, a metallic or inorganic material is vaporized using resistive heating in an atmosphere of 1-50 mbar He (or another inert gas like Ar, Ne, Kr). In gas evaporation, a high residual gas pressure causes the formation of ultra fine particles (100 nm) by gas phase collision. The ultrafine particles are formed by collision of evaporated

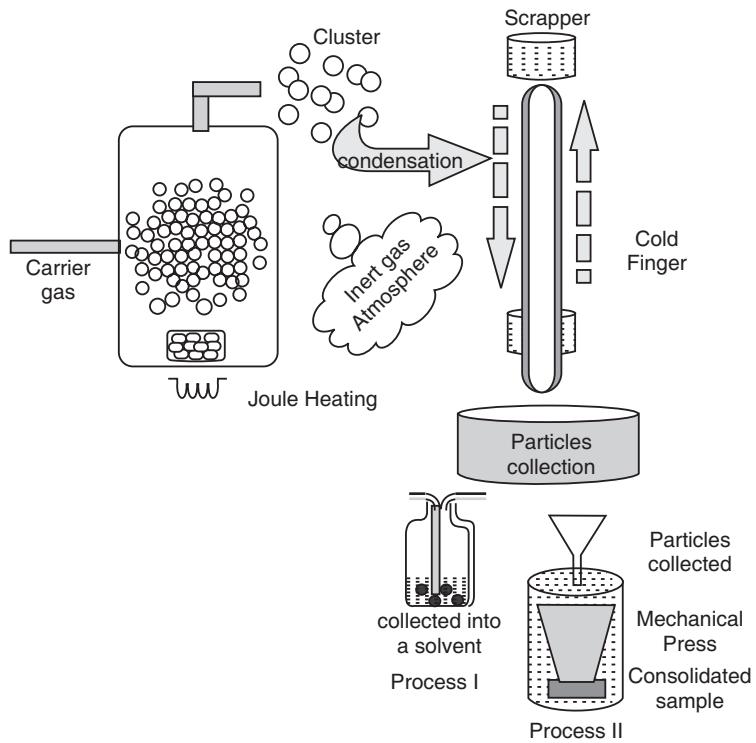


Fig. 39.27. Schematic representation of typical set-up for gas condensation synthesis of nanomaterials

atoms with residual gas molecules. Clusters rapidly condense in the inert gas in the form of small crystallites. A rotating cylindrical device cooled with liquid nitrogen is employed for the particle collection on a cold finger maintained at liquid nitrogen temperature. Subsequently, the nanoparticles are removed from the surface of the cylinder by means of a scraper in the form of a metallic plate (Fig. 39.27). Evaporation is to be done from W, Ta or Mo refractory metal crucibles. This technique is very useful to produce composite materials by mixing two or more evaporation sources. The particle size can be controlled by the evaporation rate and condensation gas pressure.

4. Laser ablation: Laser ablation has been extensively used for the preparation of nanoparticles and particulate films. In this process, a laser beam is used for generating clusters directly from a solid sample in a wide variety of applications. In this method, a high energy pulsed laser with an intensity flux exceeding 10^7 W/cm^2 is focused on a target containing the material to be made into clusters. The resulting plasma causes highly efficient vaporization, as pulsed lasers can generate temperatures greater

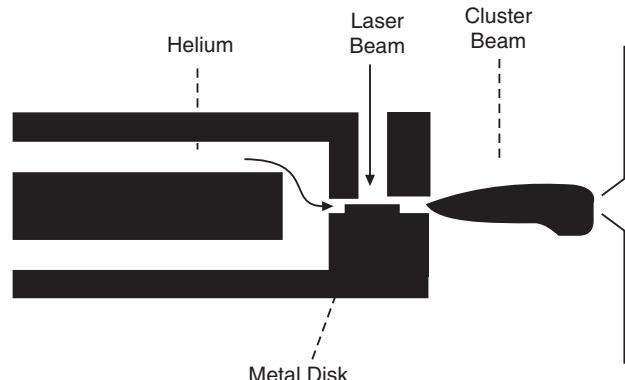


Fig. 39.28

than 10^4 K at the target material. This high temperature vaporizes all known substances so quickly that the rest of the source can operate at room temperature. The hot metal vapour is entrained in a pulsed flow of carrier gas (typically He) and expanded through a nozzle into a vacuum. The cool, high-density helium flowing over the target serves as a buffer gas in which clusters of the target material form, thermalize to near room temperature and then cool to a few K in the subsequent supersonic expansion. A typical laser vapourization source is shown in Fig. 39.28.

5. RF Plasma:

In this method of nanoparticle synthesis, the starting metal is kept in a pestle which is in turn placed in an evacuated chamber (Fig. 39.29). RF coils are wrapped around the evacuated system in the vicinity of the pestle. When high voltage is applied to RF coils heat is generated and the metal is heated above its evaporation point. Helium gas is then allowed to enter the system which forms high temperature plasma in the region of the coils. The metal vapour nucleates on the He gas atoms and diffuses up to a colder collector. Nanoparticles form at the collector. The particles are generously passivated by the introduction of oxygen.

6. Thermolysis: Nanoparticles can be made by decomposing solids at high temperature having metal cations and molecular anions or metal organic compounds. The process is called *thermolysis*. For example, small lithium particles can be made by decomposing lithium azide, LiN_3 . The material is placed in an evacuated quartz tube and heated to 400°C in the apparatus (Fig. 39.30). At about 370°C the LiN_3 decomposes, releasing N_2 gas, which is observed by

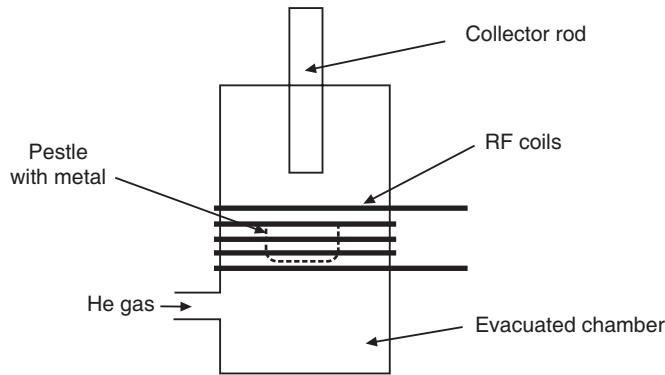


Fig. 39.29: Production of nanoparticles using RF plasma

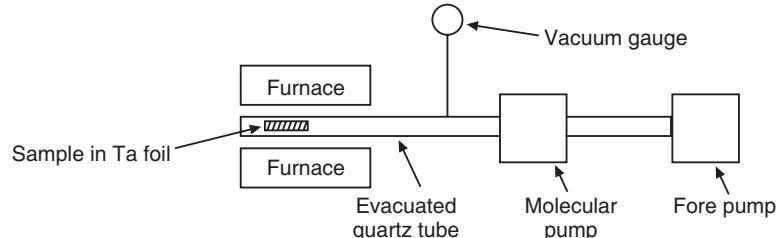


Fig. 39.30: Production of nanoparticles by thermolysis

an increase in the pressure on the vacuum gauge. In a few minutes the pressure drops back to its original value, indicating that all the N_2 gas has been removed. The remaining lithium atoms coalesce to form small colloidal metal particles. Particles less than 5 nm have been made by this method.

39.14 APPLICATIONS OF NANOMATERIALS

Since nanomaterials possess unique, beneficial chemical, physical, and mechanical properties, they can be used for a wide variety of applications. Only some of the applications are described here.

1. Tougher and harder cutting tools: Cutting tools made of nanocrystalline materials, such as tungsten carbide, tantalum carbide, and titanium carbide, are much harder, much

more wear-resistant, erosion-resistant, and last longer than their conventional (large-grained) counterparts. They also enable the manufacturer to machine various materials much faster, thereby increasing productivity and significantly reducing manufacturing costs. Also, for the miniaturization of microelectronic circuits, the industry needs microdrills (drill bits with diameter less than the thickness of an average human hair or 100 μm) with better wear resistance. Since nanocrystalline carbides are much stronger, harder, and wear-resistant, they are currently being used in these microdrills.

2. Better insulation materials: Nanocrystalline materials synthesized by the sol-gel technique results in foam like structure called an “aerogel.” These aerogels are porous and extremely lightweight; yet, they can withstand 100 times their weight. Aerogels are composed of three-dimensional, continuous networks of particles with air (or any other fluid, such as a gas) trapped at their interstices. Since they are porous and air is trapped at the interstices, aerogels are currently being used for insulation in offices, homes, etc. Aerogels are advantageous insulators compared to polyurethane foams (PUF), which emit chloro-fluoro-carbon (CFC) gases that damage the UV protecting ozone layer. They are also being used as materials for “smart” windows, which darken when the sun is too bright and they allow more light when the sun is not shining too brightly. Aerogels have also found their application as acoustic impedance matching in ultrasonic distance sensors. Due to their smallest impedance aerogels are useful to boost the efficiency of transducers, which are common components of cameras and of robotic systems that gauge distances by emitting ultrasonic waves.

3. Ductile, machinable ceramics: Ceramics are very hard, brittle, and hard to machine. These characteristics of ceramics have discouraged the potential users from exploiting their beneficial properties. However, with a reduction in grain size, these ceramics have increasingly been used. Zirconia, a hard, brittle ceramic, has even been rendered superplastic, i.e., it can be deformed to great lengths (up to 300% of its original length). However, these ceramics must possess nanocrystalline grains to be superplastic. In fact, nanocrystalline ceramics, such as silicon nitride (Si_3N_4) and silicon carbide (SiC), have been used in such automotive applications as high-strength springs, ball bearings, and valve lifters, because they possess good formability and machinability combined with excellent physical, chemical, and mechanical properties. They are also used as components in high-temperature furnaces.

4. Low-cost flat-panel electrochromic displays: An electrochromic device consists of materials in which an optical absorption band can be introduced, or an existing band can be altered by the passage of current through the materials, or by the application of an electric field. Nanocrystalline materials, such as tungstic oxide ($\text{WO}_3 \cdot x\text{H}_2\text{O}$) gel, are used in very large electrochromic display devices. These devices are primarily used in public billboards and sticker boards to convey information. Electrochromic devices are similar to liquid-crystal displays (LCD) commonly used in calculators and watches. However, electrochromic devices display information by changing color when a voltage is applied. When the polarity is reversed, the color is bleached. The resolution, brightness, and contrast of these devices greatly depend on the tungstic oxide gel’s grain size.

5. Elimination of pollutants: Nanocrystalline materials possess extremely large grain boundaries relative to their grain size. Hence, nanomaterials are very active in terms of their chemical, physical, and mechanical properties. Due to their enhanced chemical activity, nanomaterials can be used as catalysts to react with such noxious and toxic gases as carbon monoxide and nitrogen oxide in automobile catalytic converters and power generation equipment to prevent environmental pollution arising from burning gasoline and coal.

6. High-power magnets: The strength of a magnet is measured in terms of coercivity and saturation magnetization values. These values increase with a decrease in the grain size and an increase in the specific surface area (surface area per unit volume of the grains) of the grains. It has been shown that magnets made of nanocrystalline yttrium-samarium-cobalt grains possess very unusual magnetic properties due to their extremely large surface area. Typical applications for these high-power rare-earth magnets include quieter submarines, automobile alternators, land-based power generators, motors for ships, ultra-sensitive analytical instruments, and magnetic resonance imaging (MRI) in medical diagnostics.

7. High energy density batteries: Conventional and rechargeable batteries are used in almost all applications that require electric power. These applications include automobiles, laptop computers, cellular phones, cordless phones, toys, watches etc. The storage capacity of these batteries is quite low and they require frequent recharging. The life of conventional and rechargeable batteries is also low. Nanocrystalline materials synthesized by sol-gel techniques can be used for separator plates in batteries because of their foam-like structure, which can hold considerably more energy than conventional ones. Furthermore, nickel-metal hydride (Ni-MH) batteries made of nanocrystalline nickel and metal hydrides are expected to require far less frequent recharging and to last much longer, because of their large grain boundary area and enhanced physical, chemical, and mechanical properties.

8. High-sensitivity sensors: Sensors employ their sensitivity to the changes in various parameters they are designed to measure. The measured parameters include electrical resistivity, chemical activity, magnetic permeability, thermal conductivity, and capacitance. All of these parameters depend greatly on the grain size of the materials employed in the sensors. A change in the sensor's environment is manifested by the sensor material's chemical, physical, or mechanical characteristics, which is exploited for detection. For instance, a carbon monoxide sensor made of zirconium oxide uses its chemical stability to detect the presence of carbon monoxide. In the event of carbon monoxide's presence, the oxygen atoms in zirconium oxide react with the carbon in carbon monoxide to partially reduce zirconium oxide. This reaction triggers a change in the sensor's characteristics, such as conductivity and capacitance. The rate and the extent of this reaction are greatly increased by a decrease in the grain size. Hence, sensors made of nanocrystalline materials are extremely sensitive to the change in their environment. Typical applications for sensors made out of nanocrystalline materials are smoke detectors, ice detectors on aircraft wings, automobile engine performance sensor, etc.

9. Aerospace components with enhanced performance characteristics: Due to the risks involved in flying, aircraft manufacturers strive to make the aerospace components stronger, tougher, and last longer. One of the key properties required of the aircraft components is the fatigue strength, which decreases with the component's age. By making the components out of stronger materials, the life of the aircraft is greatly increased. The fatigue strength increases with a reduction in the grain size of the material. Nanomaterials provide such a significant reduction in the grain size over conventional materials that the fatigue life is increased by an average of 200-300%. Furthermore, components made of nanomaterials are stronger and can operate at higher temperatures; aircrafts can fly faster and more efficiently for the same amount of aviation fuel. In spacecrafts, elevated-temperature strength of the material is crucial because the components (such as rocket engines, thrusters, and vectoring nozzles) operate at much higher temperatures than aircrafts and higher speeds.

From the above examples, it is evident that nanomaterials outperform their conventional counterparts because of their superior chemical, physical, and mechanical properties and

of their exceptional formability. The examples given above are only a few applications of nanomaterials. Many new applications are being discovered almost daily.

10. Sunscreen: Many sunscreens contain nanoparticles of zinc oxide or titanium oxide. Older sunscreen formulas use larger particles; hence most sunscreens are whitish color. Smaller particles are less visible; therefore, when the sunscreen is rubbed into skin, it does not give whitish tinge.

11. Self-cleaning glass: Some companies offer a product which uses nanoparticles to make glass **photocatalytic** and **hydrophilic**. The photocatalytic effect means that when UV radiation from light hits the glass, nanoparticles become energized and begin to break down and loosen dirt molecules on the glass. Hydrophilic means that when water makes contact with the glass, it spreads across the glass evenly, which helps wash the glass clean. Thus, the glass is a self-cleaning glass.

12. Clothing: Scientists are using nanoparticles to enhance clothing. By coating fabrics with a thin layer of zinc oxide nanoparticles, manufacturers can create clothes that give better protection from UV radiation. Some clothes have nanoparticles in the form of little hairs or whiskers that help repel water and other materials, making the clothing stain-resistant.

13. Scratch-resistant coatings: Engineers discovered that adding aluminum silicate nanoparticles to scratch-resistant polymer coatings made the coatings more effective, increasing resistance to chipping and scratching. Scratch-resistant coatings are common on everything from cars to eyeglass lenses.

39.15 CARBON NANOMATERIALS

In 1980, we knew of only three forms of carbon, namely diamond, graphite, and amorphous carbon. Diamond and graphite have different physical structures and properties; however their atoms are both arranged in covalently bonded networks. Fullerenes and carbon nanotubes (CNTs) are the two important nanomaterials. The **fullerenes and carbon nanotubes are a class of allotropes of carbon** which conceptually are graphene sheets rolled into spheres or tubes respectively. Fig.39.31 shows the four different types of carbon structures: diamond, fullerene, graphite, and nanotube structure.

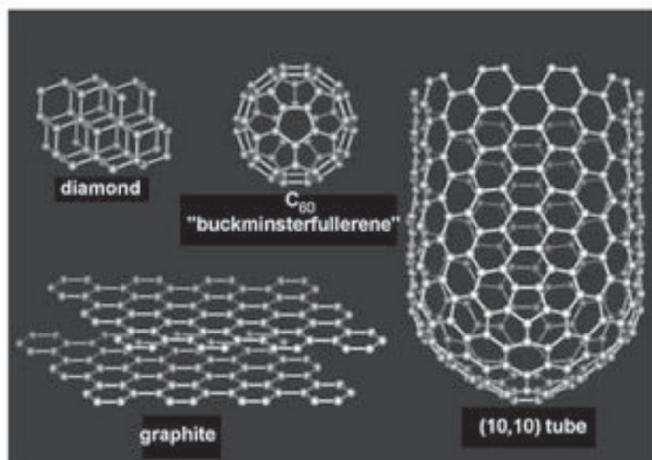


Fig. 39.31. The four different types of carbon structures

39.16 FULLERENES

Fullerene commonly refers to a molecule with 60 carbon atoms, C_{60} . It was discovered in 1985 by Harry Kroto and Richard Smalley. They named it "**buckminsterfullerene**", in recognition of the architect Buckminster Fuller, who was well-known for building geodesic domes. Later, the name is shortened to **fullerene**. C_{60} is a hollow, spherical molecule about 1nm in diameter,

comprising 60 carbon atoms (Fig. 39.32). It is also known as *buckyball* and has 32 faces. These ball-like molecules bind with each other in the solid state to form a crystal lattice having a FCC structure. After the discovery of C_{60} , other related molecules (C_{36} , C_{70} , C_{76} and C_{84}) composed of only carbon atoms, but possess different geometric structures, were also discovered. This new class of carbon molecules came to be known as the fullerenes. There are now thirty or more forms of fullerenes. The term fullerene was given to any *closed carbon cage*.

Properties

The 60 carbon atoms in C_{60} are located at the vertices of a regular truncated icosahedron and every carbon site on C_{60} is equivalent to every other site. The average nearest neighbor C-C distance in C_{60} (1.44 Å) is almost identical to that in graphite (1.42 Å). Each carbon atom in C_{60} is trigonally bonded to other carbon atoms, the same as that in graphite, and most of the faces on the regular truncated icosahedron are hexagons. There are 20 hexagonal faces and 12 additional pentagonal faces in each C_{60} molecule, which has a molecule diameter of 7.10 Å.

Synthesis

Fullerenes are usually synthesized by using an arc discharge (Fig. 39.33) between graphite electrodes in approximately 200 torr of He gas. The heat generated at the contact point between the electrodes evaporates carbon to form soot and fullerenes, which condense on the water cooled walls of the reactor. This discharge produces carbon soot that contains up to ~15% fullerenes: C_{60} (~13%) and C_{70} (~2%). The fullerenes are next separated from the soot, according to their mass, by use of liquid chromatography and using a solvent such as toluene. The following figure shows the schematic diagram of fullerene soot production chamber.

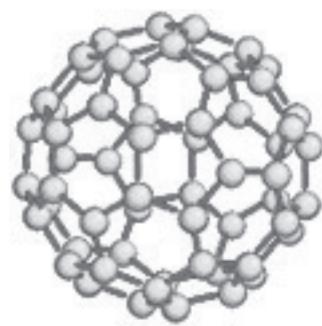
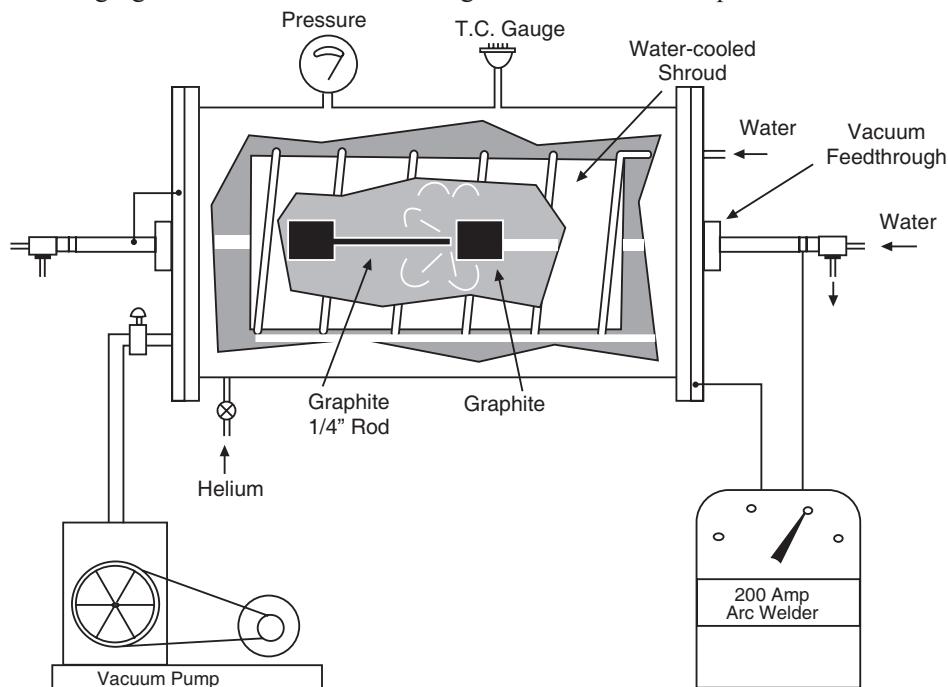


Fig. 39.32. The C_{60} molecule

Applications

Some of the potential applications of fullerenes are discussed here.

1. As fullerenes are very large graphitic systems, they can easily accommodate extra electrons. When three electrons are added to C_{60} , ionic solids of the general formula A_3C_{60} are obtained; A is any metal in Group I (lithium, sodium, potassium, rubidium, cesium). These materials are actually metals, and display superconductivity at some what low temperatures. Current research is aimed at getting the maximum superconducting temperature (or T_c) to higher values.
2. Useful dopant atoms can be placed inside the hollow fullerene ball. Atoms contained within the fullerene are said to be endohedral.
3. C_{60} has the right size to fit into the active cavity of HIV Protease, an enzyme important to the activity of the virus which causes AIDS. Cramming a buckyball into the active cavity would deactivate the enzyme and kill the virus. Ways of getting the molecule to the enzyme are under investigation
4. Possible applications of interest to industry include optical devices; chemical sensors and chemical separation devices; production of diamonds and carbides as cutting tools or hardening agents; batteries and other electrochemical applications, including hydrogen storage media; drug delivery systems and other medical applications; polymers, such as new plastics; and catalysts. Catalysts, in fact, appear to be a natural application for fullerenes, given their combination of rugged structure and high reactivity. Experiments suggest that fullerenes which incorporate alkali metals possess catalytic properties resembling those of platinum.
5. The C_{60} molecule can also absorb large numbers of hydrogen atoms—almost one hydrogen for each carbon—without disrupting the buckyball structure. This property suggests that fullerenes may be a better storage medium for hydrogen than metal hydrides, the best current material, and hence possibly a key factor in the development of new batteries and even of non-polluting automobiles based on fuel cells.
6. A thin layer of the C_{70} fullerene, when deposited on a silicon chip, seems to provide a vastly improved template for growing thin films of diamond.
7. Several other important applications are envisaged for fullerenes, such as miniature ‘ball bearings’ to lubricate surfaces, drug delivery vehicles and in electronic circuits.

39.17 CARBON NANOTUBES

In 1991, S. Iijima discovered that carbon atoms can form long cylindrical tubes. These tubes are known as carbon nanotubes or CNTs for short.

A **carbon nanotube** is a cylindrical rolled up sheet of graphene, which is a single layer of graphite atoms arranged in a hexagonal pattern, as shown in Fig.39.34. Each nanotube is a single molecule composed of millions of atoms. The length ($\approx 100 \mu\text{m}$) of a carbon nanotube (CNT) is much greater than its diameter ($\approx 2 \text{ nm}$). Their hexagonal structure gives them great tensile strength and elastic properties. The tubes are tough and when bent or squeezed, they spring back to their original shape. They are 600 times stronger than steel and 6 times lighter than it. They exhibit many other remarkable properties.

Structure

In diamond the carbon atoms link into four sided tetrahedral (Fig.39.31). Graphite consists of graphene sheets (graphene is an individual graphite layer) stacked on top of each other,

while the carbon atoms in each sheet are arranged in a honeycomb structure. During the formation of graphite, sp^2 hybridization takes place where three-hybrid sp^2 orbital are formed at 120° to each other within a plane. The in-plane s-bond is a strong covalent bond that strongly binds the atoms in the plane. The remaining p-bond is out of-plane (perpendicular to the plane) and much weaker than the in-plane bond.

A carbon nanotube (CNT) is formed when one single layer of graphite is wrapped onto itself to make a seamless cylinder. Just a nanometre across, the cylinder can be tens of microns long, and each end is “capped” with half of a fullerene molecule (Fig.39.34).

A graphene sheet can be rolled more than one way, producing different types of carbon nanotubes. There are three distinct ways in which a graphene sheet can be rolled into a tube. Depending on folding of graphite sheet, they are identified as **armchair**, **zigzag** or **chiral type** nanotubes.

Nanotubes are composed entirely of sp^2 bonds, similar to graphite. Stronger than the sp^3 bonds found in diamond, this bonding structure provides them with their unique strength. Nanotubes naturally align themselves into “ropes” held together by Van der Waals forces. Under high pressure, nanotubes can merge together, trading some sp^2 bonds for sp^3 bonds, giving great possibility for producing strong, unlimited-length wires through high-pressure nanotube linking.

39.17.1 The Two Types of Carbon Nanotubes

Carbon nanotubes are of two types:

- single-walled nanotubes (SWNT) and
- multi-walled nanotubes (MWNT).

A perfect **single-walled nanotube** (SWNT) consists of a graphene sheet rolled on leading to a single cylinder (Fig.39.35 a). Two halves of a Fullerene molecule closes the structure at both ends. Single-wall nanotubes can be thought of as the fundamental cylindrical structure, and these form the building blocks of both multi-wall nanotubes and the ordered arrays of single-wall nanotubes called *ropes*. The SWNTs have diameters from 1nm to 5 nm and are usually well over 1 mm in length. Single Walled Nanotubes (SWNT) can be considered as long wrapped graphene sheets. The length to diameter ratio of SWNTs is generally about 1000, so that they can be considered as nearly one-dimensional structures.

A **multi-walled carbon nanotube** (MWNT) is arrangement of several coaxial tubes of graphene sheets forming a tube-like structure (Fig.39.35 b & c). Each MWNT has from 2 to 50 such tubes. The separation between neighbouring tubes is roughly equal to that between the layers in natural graphite and is of 0.34-0.36nm. MWNTs have inner diameters from 1.5

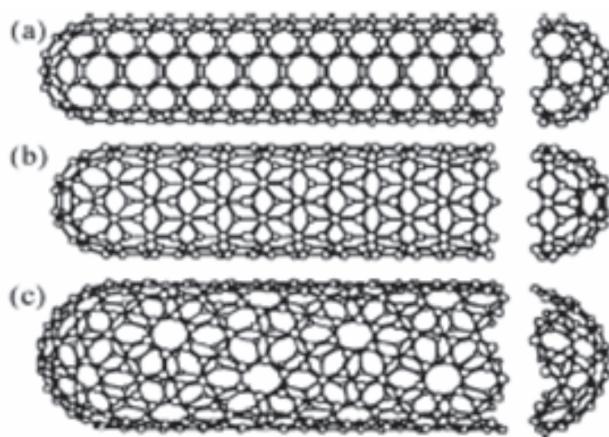


Fig. 39.34. Models of different single wall nanotubes:
(a) armchair structure (b) zigzag structure and (c) chiral structure. The difference in structure is easily seen at the open end of the tubes.

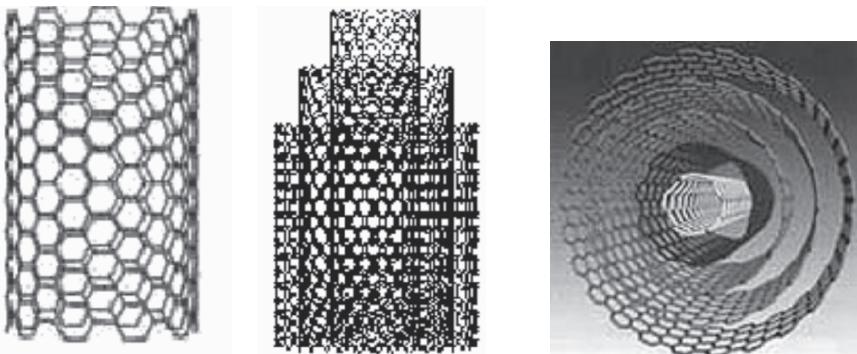


Fig. 39.35. (a) A single-walled carbon nanotube (SWNT) (b) & (c) Multi-walled carbon nanotubes (MWNT)

to 15 nm and outer diameters from 2.5 to 30 nm. MWNTs are usually well over 1 mm in length.

Thus, the carbon nanotubes are typically a few nanometres in diameter and several micrometres to centimetres long.

39.17.2 Synthesis of Carbon Nanotubes

The fabrication of carbon nanotubes is not a difficult task, since they can be found also in common environments such as the flame of a candle. But it is very difficult to control their size, orientation and structure, in order to be able to use them for technological purposes. They were synthesized for the first time in 1991 by Sumio Iijima. There are a number of methods of making CNTs and fullerenes: arc evaporation, pulsed laser deposition and Chemical Vapour Deposition (CVD) etc.

1. Arc-evaporation method or plasma arcing

The carbon arc discharge method is the most common and perhaps easiest way to produce CNTs. The method is also called plasma arcing. In this method, two carbon rods of 5-20 μm diameters are placed end to end, in an enclosure that is usually filled with inert gas at low pressure (Fig.39.36). The electrodes are separated by approximately 1mm. A potential of 20-25V is applied across the carbon electrodes. A direct current of 50 to 100 A, driven by the potential difference, creates a high temperature discharge between the two electrodes. The discharge vaporizes the surface of the positive electrode, and forms a small rod-shaped nanotubes deposit on the negative electrode. As the tubes form the length of the positive electrode decreases and a carbon deposit forms on the negative electrode. This plasma-based process is analogous to the more familiar electroplating process in a liquid medium. To produce SWNT a small amount of cobalt, nickel or iron is used as a catalyst in the central region of the positive electrode. If no catalyst is used the tubes are MWNT type. The carbon arc method can produce SWNT of diameters 1-5 nm with a length of 1 μm . Producing CNTs in high yield depends on the uniformity of the plasma arc, and the temperature of the deposit forming on the carbon electrode.

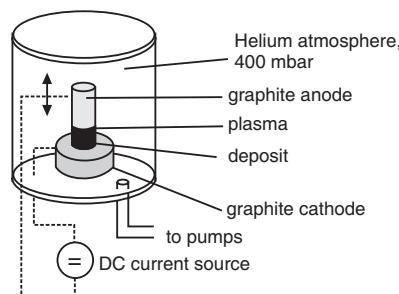


Fig. 39.36. Experimental set-up of an arc discharge apparatus

Drawback

This technique produces a complex mixture of components, and requires further purification.

2. Laser Ablation

In 1996 CNTs were first synthesized using a dual-pulsed laser technique. In this method, a quartz tube containing a graphite target with a 50:50 catalyst mixture of Cobalt and Nickel are heated to 1200°C in flowing argon (Fig. 39.37).

An intense pulsed laser beam is incident on the target, evaporating carbon from the graphite. The argon then sweeps the carbon atoms from the high temperature zone to the cold copper collector on which they condense into nanotubes. The initial laser vaporization pulse was followed by a second pulse, to vaporize the target more uniformly. The use of two successive laser pulses minimizes the amount of carbon deposited as soot. The second laser pulse breaks up the larger particles ablated by the first one, and feeds them into the growing nanotube structure. Nanotubes of 10-20nm in diameter and 100μm long can be made by this method. By varying the growth temperature, the catalyst composition, and other process parameters, the average nanotube diameter and size distribution can be varied.

The material produced by this method appears as a mat of “ropes”, 10-20nm in diameter and up to 100μm or more in length. Each rope is found to consist primarily of a bundle of single walled nanotubes, aligned along a common axis. By varying the growth temperature, the catalyst composition, and other process parameters, the average nanotube diameter and size distribution can be varied.

Drawbacks

Arc-discharge and laser vaporization are currently the principal methods for obtaining small quantities of high quality CNTs. However, both methods suffer from drawbacks.

- The first is that both methods involve evaporating the carbon source, so it has been unclear how to scale up production to the industrial level using these approaches.
- The second is that vaporization methods grow CNTs in highly tangled forms, mixed with unwanted forms of carbon and/or metal species. The CNTs thus produced are difficult to purify, manipulate, and assemble for building nanotube-device architectures for practical applications.

3. Chemical Vapour Deposition (CVD)

Chemical vapour deposition method involves decomposing a hydrocarbon gas such as methane at 1100°C. As the gas decomposes, carbon atoms are produced which condense on a cooler substrate that contains various catalysts such as iron. This method produces tubes

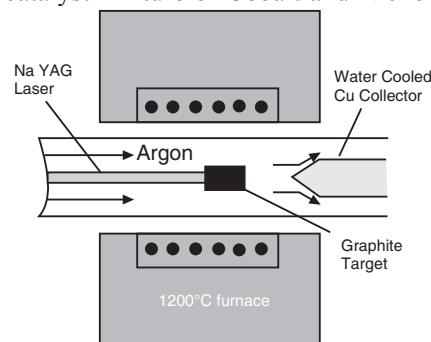


Fig. 39.37. Schematic drawing of a laser ablation apparatus

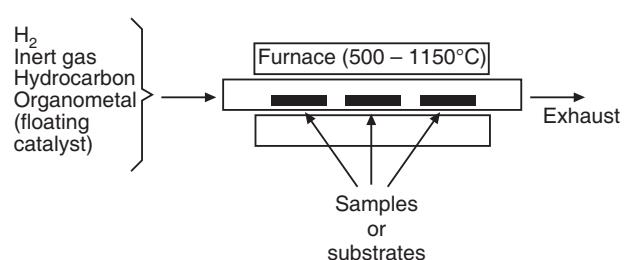


Fig. 39.38: Preparation of CNT by chemical vapour deposition

with open ends with no caps. Large amounts of CNTs can be formed by catalytic CVD of acetylene over Cobalt and iron catalysts supported on silica or zeolite. The carbon deposition activity seems to relate to the cobalt content of the catalyst, whereas the CNTs' selectivity seems to be a function of the pH in catalyst preparation.

Advantages

This method allows continuous fabrication and is suitable for scale up and production.

39.17.3 Properties of Carbon Nanotubes

A. Tensile strength

The strength of the sp^2 carbon-carbon bonds gives carbon nanotubes amazing mechanical properties.

- The tensile strength, or breaking strain of a carbon nanotube can be up to 63 GPa, around $50 \times$ higher than steel.
- High tensile strength is due to the carbon-carbon bonds and the fact that each carbon nanotube is one large molecule.
- Carbon nanotubes are elastic despite their high tensile strength. Therefore they can be bent like a rubber tube.
- The stiffness of a material is measured in terms of its Young's modulus, the rate of change of stress with applied strain. The Young's modulus of the best nanotubes can be as high as 1000 GPa which is approximately $5 \times$ higher than steel.
- These properties, coupled with the lightness of carbon nanotubes, give them great potential in applications such as aerospace. It has even been suggested that nanotubes could be used in the "space elevator", an Earth-to-space cable.

B. Thermal conductivity

- The thermal conductivity of carbon nanotubes is 10 times that of silver.
- Carbon nanotubes conduct heat by vibrations of the covalent bonds between the carbon atoms. The atoms wiggle around themselves and transmit heat through the material.
- Because the bonds in the molecule are elastic like spring the vibrations occur.
- These vibrations transmit quickly through the tube due to stiffness of the tube.

C. Electrical conductivity

- Carbon nanotubes can be metallic or semiconducting depending on their structure.
- Some nanotubes have conductivities higher than that of copper, while others behave more like silicon.
- The differences in conducting properties are caused by the molecular structure that results in a different band structure and thus a different band gap.
- The differences in conductivity can easily be derived from the graphene sheet properties.

D. Chemical reactivity

- A SWNT consists of two separate regions with different physical and chemical properties. The first is the sidewall of the tube and the second is the end cap of the tube. The end cap structure is similar to or derived from a smaller fullerene, such as C_{60} . C-atoms placed in hexagons and pentagons form the end cap structures.
- The other structure of which a SWNT is composed is a cylinder. It is generated when a graphene sheet of a certain size that is wrapped in a certain direction. As the result is cylinder symmetric we can only roll in a discreet set of directions in order to form a closed cylinder.
- For these reasons, a smaller nanotube diameter results in increased reactivity.

39.17.4 Applications of CNTs

1. Energy storage

Graphite, carbonaceous materials and carbon fibre electrodes are commonly used in fuel cells, batteries and other electrochemical applications. The advantages of using nanotubes for energy storage are their small dimensions, smooth surface topology and perfect surface specificity. The efficiency of fuel cells is determined by the electron transfer rate at the carbon electrodes, which is the fastest on nanotubes.

2. Hydrogen storage

The advantage of hydrogen as energy source is that its combustion product is water. In addition, hydrogen can be easily regenerated. For this reason, a suitable hydrogen storage system is necessary, satisfying a combination of both volume and weight limitations. The two commonly used means to store hydrogen are gas phase and electrochemical adsorption. Because of their cylindrical and hollow geometry, and extremely small diameters, carbon nanotubes can store hydrogen in the inner cores through a capillary effect.

3. Electrochemical supercapacitors

Supercapacitors have a high capacitance and potentially applicable in electronic devices. Typically, they are comprised two electrodes separated by an ionic insulating material. The capacity of an electrochemical supercap inversely depends on the separation between the charge on the electrode and the counter charge in the electrolyte. Because this separation is about a nanometer for nanotubes in electrodes, very large capacities result from the high nanotube surface area. In this way, a large amount of charge injection occurs if only a small voltage is applied. This charge injection is used for energy storage in nanotube supercapacitors.

4. Field emitting devices

If a solid is subjected to a sufficiently high electric field, electrons tunnel through the surface potential barrier of the solid. This emission current depends on the strength of the local electric field at the emission surface and its work function. The applied electric field must be very high in order to extract an electron. This condition is fulfilled for carbon nanotubes, because their elongated shape ensures very large field amplification.

Examples of potential applications for nanotubes as field emitting devices are flat panel displays, gas discharge tubes in telecom networks, electron guns for electron microscopes, atomic force microscope (AFM) tips and microwave amplifiers.

5. Transistors

The field-effect transistor – a three-terminal switching device – can be constructed of only one semiconducting SWNT. By applying a voltage to a gate electrode, the nanotube can be switched from a conducting to an insulating state. Such carbon nanotube transistors can be coupled together, working as a logical switch, which is the basic component of computers.

6. Nanoprobes and sensors

Because of their flexibility, nanotubes can be used in scanning probe instruments. Since MWNT tips are conducting, they can be used in scanning tunneling microscope (STM) and atomic force microscope (AFM) instruments.

39.18 NANOWIRES

Nanowires are wires with a very small diameter, sometimes as small as 1 nanometer. Nanowires are 2-D materials. They have an aspect ratio (length-to-width ratio) of 1000 or more. Scientists hope to use them to build tiny transistors for computer chips and other electronic devices. In the last couple of years, carbon nanotubes have overshadowed nanowires.

Types of nanowires:

- (i) Metallic nanowires (e.g., Ni, Pt, Au)
- (ii) Semiconducting nanowires (e.g., Si, InP, GaN)
- (iii) Insulating nanowires (e.g., SiO_2 , TiO_2)
- (iv) Molecular nanowires composed of repeating molecular units either organic (e.g., DNA) or inorganic (e.g., $\text{Li}_2\text{Mo}_6\text{Se}_6$)

Properties of nanowires:

Nanowires have many interesting properties that are not seen in bulk materials. This is because electrons in nanowires are quantum confined laterally and occupy energy levels that are different from the traditional energy bands in bulk materials.

The quantum confinement manifests in discrete values of electrical conductance. The discrete values are often referred to as the quantum of conductance and are integer values of $\frac{2e^2}{h} \approx 12.9 \text{ } k\Omega^{-1}$. They are the inverse of the well known resistance unit $\frac{h}{e^2} \approx 25812.8 \text{ ohms}$ which is known as the von Klitzing constant R_K .

Production of nanowires

Nanowires can be suspended, deposited or synthesized. A suspended nanowire is a wire in a vacuum chamber held at the extremities. A deposited nanowire is a wire deposited on a surface of different nature.

- (i) A suspended nanowire can be produced by chemical etching of a bigger wire, or bombarding a bigger wire with high energetic particles.
- (ii) A common technique for producing a nanowire is the Vapour-Liquid-Solid (VLS) synthesis. This technique uses as source material either laser ablated particles or a feed gas such as silane. The source is then exposed to a catalyst.

The catalysts used are liquid metal nanoclusters. The source enters these nanoclusters and begins to saturate it. Once supersaturation is reached, the source solidifies and grows outward from the nanocluster. The length of the nanowire can be adjusted by turning off the source.

This process often produces crystalline nanowires in case of semiconductor materials.

- (iii) Inorganic nanowires are synthesized in a single-step vapour phase reaction at elevated temperature.

Applications

Nanowires can be used to create active electronic elements such as p-n junctions, logic gates. Ultimately they can be used to build the next generation of computers. Nanowires have potential applications in high-density data storage, either as magnetic read heads or as storage media, and electronic and optoelectronic nanodevices, for metallic interconnects of quantum devices and nanodevices.

39.19 QUANTUM DOTS

Quantum dots are nano-meter-scale “boxes” for selectively holding or releasing electrons. A quantum dot holds a discrete number of electrons. Generally speaking, atoms are quantum dots, however, adding a number of molecules together in small space, produce the quantum dots effects. Dots have been made ranging from 30 nm to 1 micron in size, and holding from zero to hundreds of electrons.

Addition or removal of an electron changes the properties of a quantum dot. The number of electrons in a dot may be adjusted by changing the electrostatic environment of the dot. Adjusting electric fields in the neighbourhood of the dot, for example by applying a voltage to a nearby metal gate, can change this number. Of course, since quantum dots are fabricated in solids, not in vacuum, there are many electrons in them. However, almost all of these are tightly bound to atoms in the solid. The few electrons spoken of are extra ones beyond those that are tightly bound. These extra electrons could roam free in a solid were they are not confined in a quantum dot.

In a quantum dot the energies, at which electrons and holes can exist, are limited in the particles. As energy is related to wavelength, this means that the optical properties of the particle can be finely tuned depending on its size. Thus, particles can be made to emit or absorb specific colours of light, merely by controlling their size. By using an external UV light on nano-crystals made from semiconductor materials such as zinc sulphide, cadmium selenide, indium phosphide or lead sulphide, the nano-crystal will absorb the light. And then, as a result of the crystal being stimulated by the absorbed light, it will re-emit the light, usually of a certain colour, depending on the size of the quantum dot.

Further, the position of a single electron in a quantum dot might attain several states, so that a quantum dot could represent a byte of data. Alternatively, a quantum dot might be used in more than one computational instruction at a time.

Applications

Recently, quantum dots have found applications in composites, solar cells and fluorescent biological labels. The quantum dots are expected to be used in various other applications such as the following:

1. Quantum computers, 2. Domestic and office lighting applications, 3. Television screens and monitors.

39.20 DENDRIMERS

A **dendrimer** is a tree-like highly branched polymer molecule. Dendrimers are synthesized from monomers with new branches added in discrete steps to form a tree-like architecture. There are many types of dendrimer; the smallest is several nanometres in size.

Dendrimers are used in conventional applications such as coatings and inks, but they also have a range of interesting properties which could lead to useful applications. For example, dendrimers can act as nanoscale carrier molecules and as such could be used in drug delivery. Environmental clean-up could be assisted by dendrimers as they can trap metal ions, which could then be filtered out of water with ultra-filtration techniques.

Dendrimers are of particular interest for cancer applications because it is easy to attach a variety of other molecules to the surface of a dendrimer. Dendrimers go through the vascular pores and into tissue more efficiently than larger carriers. They have high drug-carrying capacity that can release a heavy payload without damaging tissue. A single dendrimer can carry a molecule that recognizes cancer cells, a therapeutic agent to kill those cells and a molecule that recognizes the signal of cell death.

39.21 NANOCOMPOSITES

Nanocomposites are materials with a nanoscale structure that improve the macroscopic properties of products. Typically, nanocomposites are clay, polymer or carbon, or a combination of these materials with nanoparticle building blocks.

Composite materials

An important use of carbon nanotubes is in composites, materials that combine one or more separate components and which are designed to exhibit overall the best properties of each component. Currently, carbon fibres and bundles of multi-walled CNTs are used in polymers to control or enhance conductivity. A particular type of nanocomposite is where nanoparticles act as fillers in a matrix; for example, carbon black used as a filler to reinforce car tyres. Because of the stiffness of carbon nanotubes, they are ideal candidates for structural applications. For example, they may be used as reinforcements in high strength, low weight, and high performance composites.

39.22 SCALING LAWS

Many phenomena in nature exhibit the remarkable property of self-similarity and they reproduce themselves as scales change, subject to *scaling laws*. **Scaling laws** are extremely simple observations about how physics works at different sizes. They provide a very simple approach to understanding the nanoscale. Machines can be characterized by simple measures and ratios; for example, a manufacturing system can handle its own mass of materials in a certain number of seconds, and a motor will handle a certain amount of power per unit volume. Broadly speaking, these numbers vary in predictable ways according to the size of the system. These relationships are called “scaling laws.” Most physical magnitudes characterizing nanoscale systems differ enormously from those familiar in macroscale systems. Some of these magnitudes can, however, be estimated by applying scaling laws to the values for macroscale systems. Detailed engineering requires more intricate calculations. But basic scaling law calculations, used with appropriate care, can show why technology based on nanoscale devices is expected to be extremely powerful. Scaling laws show what we can expect once we develop the ability to build nanoscale systems: performance vastly higher than we can achieve with today’s large-scale machines.

39.22.1 Scale Factor

If we put the question, ‘how much bigger is one tree than another?’, it has no unique answer. We have to specify on what basis the comparison is to be made - for example, on their heights, on the cross-sectional area of their trunks or on the volume of wood in their trunks. If the trees have roughly the same shape, a comparison can be made in terms of a measure based on length - the **scale factor**, L . Then if one tree were L times taller than the other, the radius of its trunk would be L times bigger, the cross-sectional area of its trunk L^2 times bigger and the volume of its trunk L^3 times bigger than the other. Thus, our *basic* comparison is being made on corresponding lengths. However L is a *ratio* of lengths; it has no unit; it is a *scale factor*, not a *scale length*. A scale factor can be assigned to any class of objects, which have essentially the same shape.

Nature behaves differently on large and small scales. Galileo showed that this results fundamentally from the way area and volume scale. Area scales as the second power of length,

$$A \propto L^2$$

while volume scales as length to the third power,

$$V \propto L^3$$

39.22.2 Application of Scaling Laws

Let us look at an example. A flea can jump dozens of times its height, while an elephant cannot jump at all. Scaling laws tell us that this is a general rule: smaller things are less affected by gravity. Let us analyze this example. As a muscle shrinks, its strength decreases

with its cross-sectional area. Therefore, strength is proportional to L^2 . But the weight of the muscle is proportional to its volume. Therefore, $\text{weight} \propto L^3$. *Strength versus weight* is a crude indicator of how high an organism can jump. It is proportional to area divided by volume, which is L^2 divided by L^3 or $1/L$. It implies that strength-per-weight gets ten times better when an organism gets ten times smaller. Strength and mass are completely different kinds of thing, and cannot be directly compared. But they both affect the performance of systems, and they both scale in predictable ways.

A flea can move its legs back and forth far faster than an elephant. The speed of a leg while it is moving may be about the same in each animal, but the distance it has to travel is a lot less in the flea. Therefore, *frequency of operation* $\propto 1/L$. It implies that as a machine shrinks, its frequency of operation will increase proportionally. A machine in a factory might join or cut ten things per second. In contrast, the fastest biochemical enzymes can perform about a million chemical operations per second.

Thus, scaling laws can compare the relative performance of systems at different scales, and the technique works for any systems with the relevant properties—the strength of a steel cable scales the same as a muscle. Any property that can be summarized by a scaling factor can be used in this kind of calculation. And most importantly, properties can be combined: just as strength and weight are components of a useful strength-per-weight measure, other quantities like power and volume can be combined to form useful measures like power density.

Power density is an important aspect of machine performance. Power is force times speed. Force is basically the same as strength, which is $\propto L^2$ and we assume that the speed is constant. So power $\propto L^2$. Thus, an engine 10 times bigger than another engine will have 100 times more power. But volume $\propto L^3$, so power per volume or power density $\propto 1/L$. Thus, power density varies inversely with machine size. Suppose an engine measuring 10 cm on one of its sides, produces 1,000 watts of power. Then an engine measuring 1 cm on one of its sides should produce 10 watts of power, i.e., $1/100$ of the ten-times-larger engine. Then, one thousand (1000) 1-cm engines would take the same volume as one 10-cm engine, but produce 10,000 watts. So according to scaling laws, by building 1,000 times as many parts, and making each part 10 times smaller, we can get 10 times as much power out of the same mass and volume of material. Thus, when the design was shrunk by a factor of ten, power output increased ten-fold.

There is another scaling law regarding functional density $\propto 1/L^3$. If we can build machine parts at nanoscale, then we can pack 10^{18} more parts into the same volume. If the parts can be built using a massively parallel process like chemistry, and if reliability is high and the design is fault-tolerant so that the collection of parts will last for the life of the product, then it would be very much worth doing.

Friction and wear are important factors in mechanical design. Friction is proportional to force. Thus, friction $\propto L^2$. This implies that frictional power is proportional to the total power used, regardless of scale.

In systems that are subject to wear, the lifetime decreases proportionally with the size. However, due to atomic granularity and the properties of certain interfaces, atomically precise molecular manufacturing systems need not be subject to incremental wear. Assuming unchanging pressure and speed, the rate of erosion is independent of scale. However, the thickness available to erode decreases as the system shrinks: wear life $\propto L$. So it is expected that a nanoscale system subjected to conventional wear mechanisms might have a lifetime of only a few seconds. Fortunately, a *non-scaling mechanism* comes to the rescue here. Chemical

covalent bonds are far stronger than typical forces between sliding surfaces. As long as the surfaces are built smooth, run at moderate speed, and can be kept perfectly clean, there should be no wear, since there will never be a sufficient concentration of heat or force to break any bonds. Calculations and preliminary experiments have shown that some types of atomically precise surfaces can have near-zero friction.

Thus, performance of machines improves as machines shrink in size. Scaling laws show what we can expect once we develop the ability to build nanoscale systems: performance vastly higher than we can achieve with today's large-scale machines.

Scaling laws are explored in detail in Chapter 2 of *Nanosystems* written by E.Drexler. Some of them are reproduced here for the benefit of the readers.

39.22.3 Scaling of Classical Mechanical Systems

Nanomechanical systems are fundamental to molecular manufacturing and are useful in many of its products and processes. We examine here how different physical magnitudes depend on the size of a system (defined by a length parameter L) if all shape parameters and material properties (e.g., strengths, moduli, densities, coefficients of friction) are held constant.

39.22.4 Magnitudes and Scaling

Given constancy of stress and material strength, both the strength of a structure and the force it exerts scale with its cross-sectional area

$$\text{total strength} \propto \text{force} \propto \text{area} \propto L^2 \quad (39.1)$$

Nanoscale devices accordingly exert only small forces: a stress of 10^{10} N/m² equals 10^{-8} N/nm², or 10 nN/nm². Stiffness in shear, like stretching stiffness, depends on both area and length.

$$\text{shear stiffness} \propto \text{stretching stiffness} \propto \frac{\text{area}}{\text{length}} \propto L \quad (39.2)$$

and varies less rapidly with scale; a cubic nanometer block of $E = 10^{12}$ N/m² has a stretching stiffness of 1000 N/m. The bending stiffness of a rod scales in the same way

$$\text{bending stiffness} \propto \frac{\text{radius}^4}{\text{length}^3} \propto L \quad (39.3)$$

Given the above scaling relationships, the magnitude of the deformation under load

$$\text{deformation} \propto \frac{\text{force}}{\text{stiffness}} \propto L \quad (39.4)$$

is proportional to scale, and hence the shape of deformed structures is scale invariant.

The assumption of constant density makes mass scale with volume,

$$\text{mass} \propto \text{volume} \propto L^3 \quad (39.5)$$

and the mass of a cubic nanometer block of density $\rho = 3.5 \times 10^3$ kg/m³ equals 3.5×10^{-24} kg. The above expressions yield the scaling relationship

$$\text{acceleration} \propto \frac{\text{force}}{\text{mass}} \propto L^{-1} \quad (39.6)$$

A cubic-nanometer mass subject to a net force equaling the above working stress applied to a square nanometer experiences an acceleration of $\sim 3 \times 10^{15}$ m/s². Accelerations in nanomechanisms commonly are large by macroscopic standards, but aside from special cases (such as transient acceleration during impact and steady acceleration in a small flywheel) they

rarely approach the value just calculated. (Terrestrial gravitational accelerations and stresses usually have negligible effects on nanomechanisms.)

Modulus and density determine the acoustic speed, a scale-independent parameter [along a slim rod, the speed is $(E/\rho)^{1/2}$; in bulk material, somewhat higher]. The vibrational frequencies of a mechanical system are proportional to the acoustic transit time

$$\text{frequency} \propto \frac{\text{acoustic speed}}{\text{length}} \propto L^{-1} \quad (39.7)$$

The acoustic speed in diamond is $\sim 1.75 \times 10^4$ m/s. Some vibrational modes are more conveniently described in terms of lumped parameters of stiffness and mass,

$$\text{frequency} \propto \sqrt{\frac{\text{stiffness}}{\text{mass}}} \propto L^{-1} \quad (39.8)$$

but the scaling relationship is the same. The stiffness and mass associated with a cubic nanometer block yield a vibrational frequency characteristic of a stiff, nanometer-scale object: $[(1000 \text{ N/m})/(3.5 \times 10^{-24} \text{ kg})]^{1/2} \approx 1.7 \times 10^{13} \text{ rad/s}$.

Characteristic times are inversely proportional to characteristic frequencies

$$\text{Time} \propto \text{frequency}^{-1} \propto L \quad (39.9)$$

Most mechanical systems use bearings to support moving parts. Macromechanical systems frequently use liquid lubricants, but (as noted by Feynman, 1961), this poses problems on a small scale. The above scaling law ordinarily holds speeds and stresses constant, but reducing the thickness of the lubricant layer increases shear rates and hence viscous shear stresses:

$$\text{viscous stress at constant speed} \propto \text{share rate} \propto \frac{\text{speed}}{\text{thickness}} \propto L^{-1} \quad (39.10)$$

In Newtonian fluids, shear stress is proportional to shear rate. Molecular simulations indicate that liquids can remain nearly Newtonian at shear rates in excess of 100 m/s across a 1 nm layer (e.g., in the calculations of Ashurst and Hoover, 1975), but they depart from bulk viscosity (or even from liquid behavior) when film thicknesses are less than 10 molecular diameters (Israelachvili, 1992; Schoen et al., 1989), owing to interface-induced alterations in liquid structure. Feynman suggested the use of low-viscosity lubricants (such as kerosene) for micromechanisms (Feynman, 1961); from the perspective of a typical nanomechanism, however, kerosene is better regarded as a collection of bulky molecular objects than as a liquid. If one nonetheless applies the classical approximation to a 1 nm film of low-viscosity fluid ($\eta = 10^{-3} \text{ N} \cdot \text{s/m}^2$), the viscous shear stress at a speed of $1.7 \times 10^3 \text{ m/s}$ is $1.7 \times 10^9 \text{ N/m}^2$; the shear stress at a speed of 1 m/s, 10^6 N/m^2 , is still large, dissipating energy at a rate of 1 MW/m².

The problems of liquid lubrication motivate consideration of dry bearings (as suggested by Feynman, 1961). Assuming a constant coefficient of friction,

$$\text{frictional force} \propto \text{force} \propto L^2 \quad (39.11)$$

and both stresses and speeds are once again scale-independent. The frictional power,

$$\text{frictional power} \propto \text{force} \cdot \text{speed} \propto L^2 \quad (39.12)$$

is proportional to the total power, implying scale-independent mechanical efficiencies. In light of engineering experience, however, the use of dry bearings would seem to present problems (as it has in silicon micromachine research). Without lubrication, efficiencies may be low, and static friction often causes jamming and vibration.

A yet more serious problem for unlubricated systems would seem to be wear. Assuming constant interfacial stresses and speeds (as implied by the above scaling relationships), the

anticipated surface erosion rate is independent of scale. Assuming that wear life is determined by the time required to produce a certain fractional change in shape,

$$\text{wear life} \propto \frac{\text{thickness}}{\text{erosion rate}} \propto L \quad (39.13)$$

and a centimeter-scale part having a ten-year lifetime would be expected to have a 30 s lifetime if scaled to nanometer dimensions.

Design and analysis have shown, however, that dry bearings with atomically precise surfaces need not suffer these problems.

39.22.5 Major Corrections

The above scaling relationships treat matter as a continuum with bulk values of strength, modulus, and so forth. They readily yield results for the behavior of iron bars scaled to a length of 10^{-12} m, although such results are meaningless because a single atom of iron is over 10^{-10} m in diameter. They also neglect the influence of surfaces on mechanical properties, and give (at best) crude estimates regarding small components, in which some dimensions may be only one or a few atomic diameters.

The accuracy of scaling laws to nanoscale systems is good for purely mechanical systems, if the component dimensions substantially exceed atomic dimensions. Scaling principles indicate that mechanical components can operate at high frequencies, accelerations, and power densities. The adverse scaling of wear lifetimes suggests that bearings are a special concern.

39.22.6 Scaling Laws in Electrostatic Forces

The capacitance of a parallel plate capacitor is given by

$$C = \frac{\epsilon_0 \epsilon_r A}{d}$$

where A is the area of plates and d is the distance between them. Hence, scaling for C has the following form.

$$C \propto \frac{(L^0)(L^0)(L^2)}{(L^1)} = (L^1) \quad (39.14)$$

The electrostatic energy stored in a parallel plate capacitor is given by

$$U = \frac{1}{2} CV^2 = \frac{1}{2} \frac{\epsilon_0 \epsilon_r A}{d} V^2$$

V is the potential applied and is equal to $\mathbf{E} \times d$ where \mathbf{E} is the electric field intensity. Keeping the electric field constant, the potential $V \propto d$. Hence scaling for U has the following form.

$$U \propto \frac{(L^0)(L^0)(L^2)(L^1)^2}{(L^1)} = (L^3) \quad (39.15)$$

This suggests that a decrease by a factor of 10 in linear dimensions of the plates reduces the energy stored by 10^3 times. Since the force is a derivative of energy, it follows that the force between the plates decreases by a factor of L^2 , that is 10^2 .

39.22.7 Scaling in Electricity

The resistance of a wire is given by

$$R = \frac{\rho L}{A} \propto L^{-1} \quad (39.16)$$

The resistive power loss is given by

$$P = \frac{V^2}{R} \propto L^1 \quad (39.17)$$

Here the voltage is assumed to be held constant.

Thus, the power loss is linear in scaling. The power supply needed for the device is proportional to volume. Hence, the ratio of power loss to available power is of interest. This ratio is expressed as

$$\frac{P_{loss}}{P_{supply}} = \frac{L^1}{L^3} \propto L^2 \quad (39.18)$$

It shows that a decrease in size of the device by 10 times actually increases the power dissipation by a factor of 100.

39.22.8 Scaling of Electromagnetic Forces

The inductance of the solenoid is given by

$$L = \mu_0 n^2 l A$$

where n is the total number of turns and l is the length of the solenoid and A is the area enclosed by the coil of the solenoid. Assuming that the area enclosed by the coil is constant

$$L \propto (L^1) \quad (39.19)$$

Thus both capacitance and inductance scale as L^1 .

The natural frequency of an LCR circuit would scale as

$$\omega = \frac{1}{\sqrt{LC}} \propto \frac{1}{\sqrt{(L^1)(L^1)}} \propto L^{-1} \quad (39.20)$$

Thus, a reduction in size by a factor of 10 would increase the frequency of an LC circuit by 10 times. Similarly, the Q-factor of LCR circuit would scale as

$$Q \propto \frac{L}{R} \times \omega \propto \frac{(L^1)}{(L^{-1})} \times (L^{-1}) \propto L^1 \quad (39.21)$$

Thus, a reduction in size by a factor of 10 would decrease the Q-factor by an equal factor.

The capacitive time constant is given by

$$\begin{aligned} \tau &= RC \\ \therefore \tau &\propto (L^{-1})(L^1) \propto L^0 \propto \text{constant} \end{aligned} \quad (39.22)$$

It shows that the capacitive time constant is independent of scaling.

The inductive time constant scales as follows:

$$\tau = \frac{L}{R} \propto \frac{(L^1)}{(L^{-1})} \propto L^2 \quad (39.23)$$

Thus, on decreasing the size by a factor of 10, the inductive time constant would decrease by a factor of 100.

The energy stored in an inductor is given by

$$U = \frac{1}{2} L I^2$$

If the radius of the wire forming an inductor is increased the current will increase because the current is proportional to the area of cross-section of the conductor. Thus,

$$U \propto (L)(L^2)^2 \propto L^5 \quad (39.24)$$

Hence, the force which is proportional to the derivative of magnetic stored energy $\propto L^4$.

Therefore, a 10 times reduction in size of inductance would lead to 10^4 times reduction in the electromagnetic force. On the other hand, it is seen that the electrostatic force reduces as L^2 .

39.22.9 Scaling Laws for Thermal Systems

The heat capacity is proportional to mass. Hence,

$$\text{Heat capacity} \propto \text{volume} \propto L^3 \quad (39.25)$$

The thermal conductance is given by

$$\text{Thermal conductance} \propto \frac{\text{area}}{\text{length}} \propto L^1 \quad (39.26)$$

The thermal time constant would scale as follows:

$$\text{Thermal time constant} \propto \frac{\text{heat capacity}}{\text{thermal conductance}} \propto \frac{(L^3)}{(L^1)} \propto L^2 \quad (39.27)$$

39.23 NANO DEVICES AND NANOMACHINES

Nanomachines and nanodevices are in the early stage of development, and are still in conceptual stages. Nano-sized machines do exist in biological systems and the study of biological machines is expected to provide us insights that will help us in the design of mechanical nanomachines. Computer simulation has been used to evaluate the potential of various nanomachine concepts. One example is the idea of making a gear which is an essential part of any sized nanomachine.

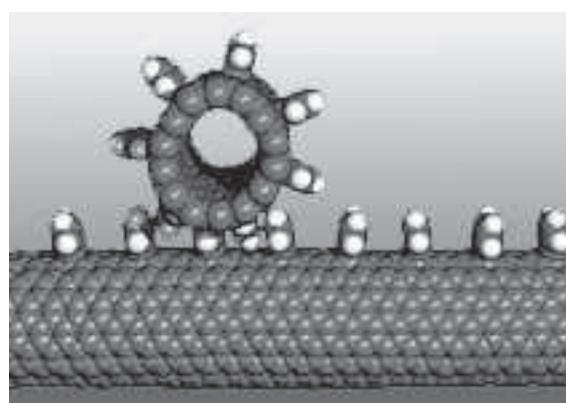


Fig. 39.39: Nanogear

Gears are used to transfer one form of mechanical work into another. Nanogears serve a similar function but on a different scale. Current researchers are developing interactive nanogears fabricated from Fullerene. Synthesis of nanogear systems involves adding a knobby molecule like benzene to the outer walls of a carbon nanotube (Fig.39.39). If some of the teeth or gear molecules are electronically charged they can be made to rotate in response to the applied stimulus.

QUESTIONS

1. What do you mean by Nano?
2. Define Nano Science?
3. Define Nano Technology?
4. What is the difference between Nano Science & Nano Technology?
5. “One nanometer is a magical point on the dimension scale.” Why? Explain.
6. Define nanomaterial. Give classification of nanomaterials.
7. Define top down and bottom up approach.
8. What are the induced effects due to increase in surface area of nanoparticles?
9. Why [surface area/volume] ratio is very large for nanoparticles compared to bulk materials? Explain with a simple example. Highlight any two problems associated with increase in surface area.

10. Write a short note on self-assembly.
11. Explain the working of scanning electron microscope (SEM) with a neat sketch.
12. Explain the working of atomic force microscope (AFM) with a neat sketch.
13. What is the difference between STM & AFM.
14. Explain photolithography (optical lithography) with a neat sketch.
15. Explain Electron beam lithography with a neat sketch.
16. Write an essay on lithographic techniques of fabrication.
17. Classify nanomaterials and give examples for them.
18. Write a short note on Nanomaterials and nanostructures. (RGPV, 2008)
19. Define nanocomposite and classify nanocomposites.
20. With a neat sketch, explain mechanical milling process for synthesis of nano particles. List advantages and disadvantages also.
21. Describe various techniques of physical vapour deposition.
22. Explain gas condensation process for synthesis of nanopowders with a neat sketch?
23. Explain laser ablation process for synthesis of nanopowders with a neat sketch.
24. Explain SOL-GEL synthesis for producing nanomaterials? Explain with the help of a neat sketch.
25. List any four processes to produce nanopowders.
26. List any four day to day live commercial applications of nanotechnology.
27. What are composites? Explain the classification and advantages of composite materials. (V.T.U.,2008)
28. Define bucky ball.
29. Why C₆₀ molecules are called bucky balls? Give reasons.
30. What is the diameter of a bucky ball? How many pentagons and hexagons are there in a bucky ball.
31. List methods for producing bucky balls.
32. What are nanomaterials? Write a note on carbon nanotubes. (V.T.U.,2007, 2008)
33. Define carbon nanotube. What are the types of carbon nanotubes?
34. List the methods for producing carbon nanotubes and explain any one of the method with a neat sketch.
35. Highlight the properties of carbon nanotubes.
36. Give any two excellent properties of carbon nanotubes.
37. List any two applications of bucky balls and carbon nanotubes?
38. Discuss the applications of carbon nanotubes.
39. Discuss the physical properties of carbon nanotubes in relation to their structure.
40. Write a brief note on:
 - (i) Nanotechnology
 - (ii) Carbon nanotubes (V.T.U.,2007)
41. Explain scaling of classical mechanical systems along with two examples and the assumptions involved. (V.T.U.,Karnataka,2007)
42. Explain scaling laws. Explain scaling of classical mechanical systems along with two examples and the assumptions involved in it. (V.T.U.,2007)
43. Discuss mechanical scaling. (V.T.U.,2008)
44. Explain scaling of Electro-magnetic systems.
45. Write notes on nanosystems. (V.T.U.,2008)

CHAPTER

40

Geometrical Optics

40.1 INTRODUCTION

A lens is an image-forming device. It forms an image by refraction of light at its two bounding surfaces. It is expected that the image formed is good and free of aberrations. Ray concept is used in understanding the behaviour of light passing through lenses. As the relationship between the focal length of the lens, object distance and image distance are derived basing on ray approximation and on geometrical relations, this part of optics is known as **geometrical optics**. Geometrical optics played a very important role in the design and fabrication of optical components and instruments.

40.2 THIN LENSES

Lenses are broadly classified into *thin* and *thick* lenses. A lens is said to be **thin** if the thickness of the lens can be neglected when compared to the lengths of the radii of curvature of its two refracting surfaces, and to the distances of the objects and images from it. No lens is actually a thin lens. Yet many simple lenses commonly used can be treated as equivalent to a thin lens.

40.2.1 Lens Equations:

A lens forms an image by refraction of light at its two bounding surfaces. Each surface acts as an image-forming component, and contributes to the final image formed by the lens. If we know the focal length of a lens and the position of an object, the position of the image can be determined either by using graphical construction (ray diagram) or using mathematical relation. It is not always convenient to draw a ray diagram. We study here how the position of the image is determined using mathematical equation.

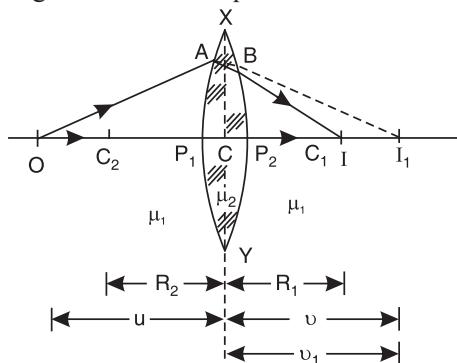


Fig. 40.1

Now let us consider a thin convex lens XY with optical center at C , as shown in Fig. 40.1. Let the absolute refractive index of the lens materials be μ_2 ; and the lens be surrounded on all sides by air or any other rarer medium of refractive index μ_1 . The centers of curvatures of the two refracting surfaces of the lens are C_1 and C_2 . Let the radius of the curvature of the surfaces be R_1 and R_2 respectively. Consider a point object ' O ' situated on the principal axis of the lens.

A ray of light leaving the axial point-object O and travelling along the principal axis passes without deviation. Another ray OA strikes the first surface at A and is refracted in a direction BI_1 . The ray is further refracted by the second surface in a direction BI and meets the ray along the principal axis at I . Therefore, I is the final point-image of the object O formed after refraction by the two surfaces of the lens.

By considering refraction at spherical surfaces we can derive an equation that describes image formation by a lens. The basic idea is that the image formed by the first refracting surface acts as a *virtual object* for the second refracting surface. We can apply **Gauss formula** for refraction at the first surface.

$$\frac{\mu_1}{u} + \frac{\mu_2}{v} = \frac{\mu_2 - \mu_1}{R} \quad \dots(40.1)$$

The first surface, XAP_1Y , forms a real point-image, I_1 . We replace v by v_1 and R by R_1 in equ. (40.1) and apply the sign convention. We see that u is -ve. Therefore, the equation (40.1) is written for the first surface as

$$\frac{\mu_2}{v_1} - \frac{\mu_1}{u} = \frac{\mu_2 - \mu_1}{R_1} \quad \dots(40.2)$$

In the Fig. 40.1, the light leaving the first surface would reach I_1 if the second surface XBP_2Y did not intervene. When the second surface is present, the light does not reach I_1 . I_1 becomes the virtual object for the second surface and the second surface forms the image at I , which is the final image. The position of the final image is again found using Gauss formula using $u = v_1$ and $R = R_2$. Thus, we get

$$\frac{\mu_2}{v_1} + \frac{\mu_1}{v} = \frac{\mu_2 - \mu_1}{R_2}$$

Applying the sign convention, we get

$$-\frac{\mu_2}{v_1} + \frac{\mu_1}{v} = -\frac{\mu_2 - \mu_1}{R_2} \quad \dots(40.3)$$

Adding equations (40.2) and (40.3), we obtain

$$\begin{aligned} \mu_1 \left(\frac{1}{v} - \frac{1}{u} \right) &= (\mu_2 - \mu_1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \\ \left(\frac{1}{v} - \frac{1}{u} \right) &= \left(\frac{\mu_2 - \mu_1}{\mu_1} \right) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \\ \frac{1}{v} - \frac{1}{u} &= \left(\frac{\mu_2}{\mu_1} - 1 \right) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \end{aligned} \quad \dots(40.4)$$

For air, $\mu_1 = 1$. Designating μ_2 as μ , we get

$$\frac{1}{v} - \frac{1}{u} = (\mu - 1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \quad \dots(40.5)$$

Equation (40.5) relates the image distance v of a thin lens to the object distance u and to the thin lens properties namely, refractive index and the radii of curvature. It is to be noted that a lens will not only focus from O to I but between any other pairs of points, as long as the points satisfy the relation (40.5).

40.2.2 Lens Maker's Equation:

If the object is at infinity, the image is formed at the *principal focus* of the lens. When $u = \infty$, $1/u = 0$ and $v = f$. Equ. (40.5) becomes

$$\begin{aligned} \frac{1}{f} - \frac{1}{\infty} &= (\mu - 1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \\ \therefore \frac{1}{f} &= (\mu - 1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \end{aligned} \quad \dots(40.6)$$

Eqn (40.6) is known as the *lens maker's formula*, since it enables one to calculate f from the known properties of the lens. It can also be used to determine the values of R_1 and R_2 needed for a desired focal length of a lens of a given index of refraction.

If the lens is turned around, R_1 and R_2 are interchanged and the sign of each is reversed. Consequently, there is no change in f . Therefore, *for a thin lens, the focal length is independent of the order of the surfaces*.

It is clear from the lens maker's formula that to get a short focal length, the lens has to have a surface of smaller radius of curvature R and be made of a material with a high refractive index.

Comparing equations (40.5) and (40.6), we see that

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f} \quad \dots(40.7)$$

The above equation is known as the **Gauss' formula for a lens**. This is more useful formula than equ. (40.5), because if we know the focal length of the lens, we can relate the position of the object and its image without necessarily knowing the index of refraction or the radii of the lens.

40.2.3 Positions of the principal foci:

Each individual surface of the lens has its own focal points and planes and the lens as a whole has its own pair of focal points and focal planes. The focal points and focal planes of the lens are known as *principal focal points* and *principal focal planes*. We are interested in the knowing the locations of these principal focal points and principal focal planes. They are obtained as follows.

(i) If a point object is placed on the principal axis such that the rays refracted by the lens are parallel to the axis, then the position of the point object is called the **first principal focus** F_1 of the lens (see Fig. 40.2a). The distance at the first principal focus from the optical center C of the lens is called the **first principal focal length**, f_1 . We can find f_1 as follows.

Using $u = f_1$, and $v = \infty$ into equ. (40.6), we get

$$\frac{1}{\infty} - \left(-\frac{1}{f_1} \right) = (\mu - 1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right]$$

$$\text{or } \frac{1}{f_1} = (\mu - 1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \quad \dots(40.8)$$

The plane perpendicular to the axis and passing through the first focal point is known as the **first principal focal plane**.

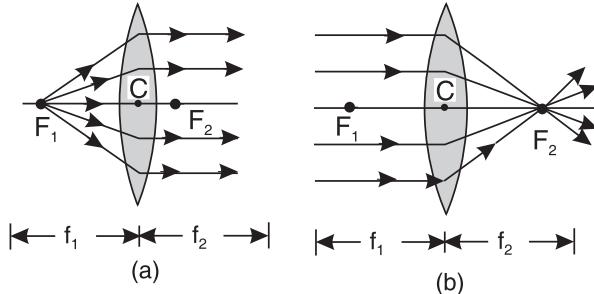


Fig. 40.2

(ii) If the object is situated at infinity, the position of the image on the axis is known as the **second principal focus** F_2 (see Fig. 40.2b). The distance of the second principal focus from the optical center C is called the **second principal focal length**, f_2 .

Using $u = \infty$ and $v = f_2$ into equ. (40.5), we get

$$\begin{aligned} \frac{1}{f_2} - \frac{1}{\infty} &= (\mu - 1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \\ \frac{1}{f_2} &= (\mu - 1) \left[\frac{1}{R_1} - \frac{1}{R_2} \right] \end{aligned} \quad \dots(40.9)$$

The plane perpendicular to the axis and passing through the second focal point is known as the **second principal focal plane**.

It follows from equ. (40.8) and (40.9) that

$$f_1 = f_2 \quad \dots(40.10)$$

Thus, *every thin lens in air has two focal points (F_1 and F_2), one on each side of the lens and equidistant from the centre.*

It will be seen that the second focal length (f_2) of a converging lens is positive and the first (f_1) negative, whilst for a diverging lens the reverse is true. *The two focal lengths of thin lens in air are numerically equal.* Frequently, one refers simply to “focal length” of a lens; it will be assumed that this always refers to the *second focal length* so that the focal length of a converging lens is positive, whilst for a diverging lens it is negative.

40.3 COAXIAL LENS SYSTEMS

Single lenses are rarely used for image formation, as they suffer from various defects. In optical instruments such as cameras, microscopes, telescopes etc, a collection of lenses are employed for forming images of objects. **An optical system consists of a number of lenses placed apart, and having a common principal axis.** The image formed by such a coaxial optical system is good and almost free of aberrations.

40.4 CARDINAL POINTS

In the case of refraction through a thin lens, the thickness of the lens has been neglected in calculating the various formulae. It is then immaterial from which point of the lens the distances are measured. But we cannot apply the above approximation for an optical system consisting of a combination of lenses. One way of calculating the position and size of the image formed by an optical system is to consider refraction at each surface of a lens successively, but the method proves to be more tedious. In 1841, Gauss showed that any number of coaxial lenses could be treated as a single unit, without the necessity of treating the single surfaces of lenses separately. The lens makers' formula can be applied to the system provided the distances are measured from two *hypothetical parallel planes*, fixed with reference to the refracting system. The points of intersection of these planes with the axis are called the **principal points** or **Gauss points**. In fact there are six points in all, which characterize an optical system. They are

- (i) two focal points
- (ii) two principal points and
- (iii) two nodal points.

These six points are known as **cardinal points** of an optical system. The planes passing through these points and which are perpendicular to the principal axis are known as **cardinal planes**. The cardinal points and cardinal planes are intrinsic properties of a particular optical system and determine the image forming properties of the system. If these are known, one can find the image of any object without making a detailed study of the passage of the rays through the system. It is not necessary to consider the refraction of the rays at the various surfaces.

40.5 DEFINITIONS AND PROPERTIES OF CARDINAL POINTS AND PLANES

1. Principal points and Principal planes:

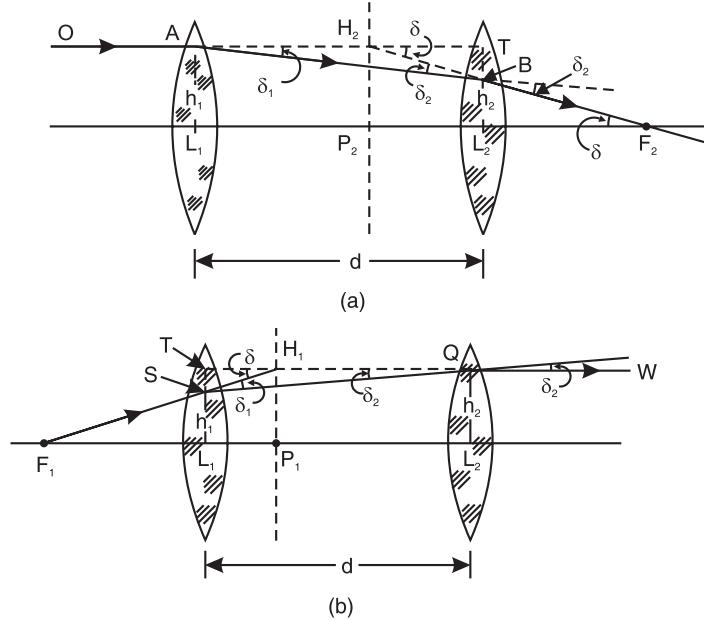


Fig. 40.3

Let us consider an optical system having its principal focal points F_1 and F_2 . A ray OA travelling parallel to the principal axis and incident at A is brought to focus at F_2 in the **image space** of the optical system as shown in Fig. 40.3 (a). The actual ray is refracted at each surface of the optical system and follows the path $OABF_2$. If we extend the incident ray OA forward and the emergent ray BF_2 backward, they meet each other within the optical system at H_2 . Now, we can describe the refraction of the incident ray OA in terms of a *single refraction* at a plane passing through H_2 . A plane drawn through the point H_2 and perpendicular to the axis may be regarded as the surface at which refraction takes place. This plane is called the *principal plane* of the optical system. Thus, the four consecutive deviations of the light rays caused by the four surfaces of the optical system are equivalent to a single refraction at H_2 , taking place at the *principal plane*. We now define the *principal plane* of an optical system as *the loci where we assume refraction to occur without reference as to where the refraction actually occurs*. H_2P_2 is the principal plane in the image space and is called the **second principal plane**. The point P_2 , at which the second principal plane intersects the axis, is called the **second principal point**.

By adopting similar procedure, as shown in Fig. 40.3 (b), we can locate the *first principal plane* H_1P_1 and *first principal point* P_1 in the **object space**. Consider the ray F_1S passing through the first principal focus F_1 such that after refraction it emerges along BW parallel to the axis at the same height as that of the ray OA . The rays F_1S and QW when produced intersect at H_1 . A plane perpendicular to the axis and passing through H_1 is called the **first principal plane**. The point of intersection, P_1 , of the first principal plane with the axis is called the **first principal point**.

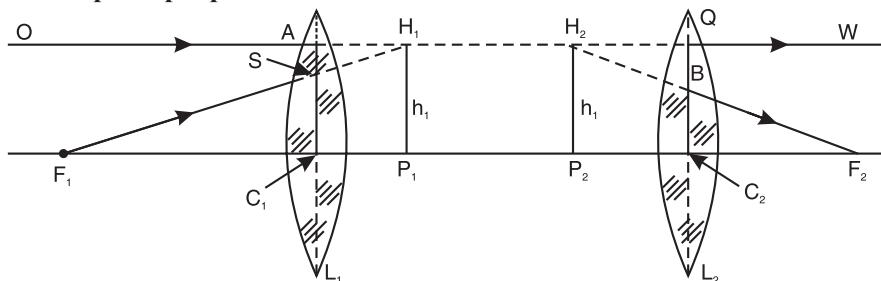


Fig. 40.4

It is seen from Fig. 40.4 that two incident rays are directed towards H_1 and after refraction seem to come from H_2 . Therefore, H_2 is the image of H_1 . Thus, H_1 and H_2 are the *conjugate points* and the planes H_1P_1 and H_2P_2 are a pair of *conjugate planes*. It is also seen that

$$H_2P_2 = H_1P_1.$$

Hence, the lateral magnification of the planes is +1. Thus, the first and second principal planes are planes of unit magnification and are therefore called *unit planes* and the points P_1 and P_2 are called *unit points*.

Note: The principal planes are *conceptual* planes and do not have physical existence within the optical system.

Some remarkable features of Principal planes:

1. Even a complex optical system has *only two principal planes*.
2. Between H_1 and H_2 all rays are *parallel* to the principal axis.
3. The location of the principal planes is characteristic of a given optical system. Their positions do not change with the object and image distance used.

4. The principal planes are conjugate to each other. An object in the first principal plane is imaged in the second principal plane with *unit magnification*. Any ray directed toward a point on the first principal plane emerge from the lens as if it originated at the corresponding point (at the same distance above or below the axis) on the second principal plane. Hence, the name *unit planes*.
5. The principal points H_1 and H_2 provide a set of references from which several system parameters are measured.

2. Focal points and Focal planes:

If a parallel beam of light travelling from the left to the right (in object space) is incident on the optical system, the beam comes together at a point, F_2 , on the other side (in image space) of the optical system. The beam passes through the point F_2 whatever may be its path inside the system. The point, F_2 , is called the **second focal point** of the optical system. A beam of light passing the point F_1 on the axis on the object side is rendered parallel to the axis after emergence through the optical system (Fig. 40.4). The point F_1 is called the **first focal point**.

We can now define the focal points as follows:

The **first focal point** is a point on the principal axis of optical system such that a beam of light passing through it is rendered parallel to the principal axis after refraction through the optical system.

The **second focal point** is a point on the principal axis of the optical system such that a beam of light travelling parallel to the principal axis of the optical system, after refraction through the system, passes through it.

The planes passing through the principal focal points F_1 and F_2 and perpendicular to the axis are called **first focal plane** and **second focal plane** respectively. The main property of the focal planes is that the rays starting from a point in the focal plane in the object space correspond to a set of conjugate parallel rays in the image space. Similarly, a set of parallel rays in the object space corresponds to a set of rays intersecting at a point in the focal plane in the image space.

The distance of the first focal point from the first principal point, *i.e.*, $F_1 P_1$, is called the **first focal length**, f_1 of the optical system and the distance of the second focal point from the second principal point, $F_2 P_2$, is called the **second focal length**, f_2 . f_1 and f_2 are also known as the focal lengths in object space and image space respectively.

When the medium is same on the two sides of the optical system $f_1 = f_2$ (numerically).

3. Nodal Points and Nodal Planes:

Nodal points are points on the principal axis of the optical system where light rays, **without refraction**, intersect the optic axis. In a *thin lens* the nodal point is the centre of the lens. Light passing through the centre of a thin lens does not deviate. In an optical system the centre separates into two nodal points. The planes passing through the nodal points and perpendicular to the principal axis are called the **nodal planes**. Whereas the principal planes are planes where all refraction is assumed to occur, the nodal planes are planes where refraction does *not* take place. Fig. 40.5 represents an optical system with the help of its cardinal planes. It is seen from the Fig. 40.5 that a ray of light, AN_1 , directed towards one of the nodal points, N_1 , after refraction through the optical system, along N_1N_2 , emerges out from the second nodal point, N_2 , in a direction, N_2R , parallel to the incident ray. The distances of the nodal points are measured from the focal points.

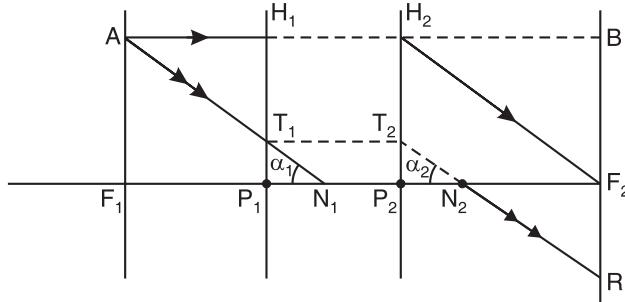


Fig. 40.5

(a) The nodal points are a pair of conjugate points on the axis having unit positive angular magnification.

Let H_1P_1 and H_2P_2 be the first and second principal planes of an optical system. Let AF_1 and BF_2 be its first and second focal planes respectively (Refer to Fig. 40.5). Consider a point A situated on the first focal plane. From A draw a ray AH_1 parallel to the axis. The conjugate ray will proceed from H_2 , a point in the second principal plane such that $H_2P_2 = H_1P_1$ and will pass through the second focus.

Take another ray AT_1 parallel to the emergent ray H_2F_2 and incident on the first principal plane at T_1 . It will emerge out from T_2 , a point on the second principal plane such that $T_2P_2 = T_1P_1$, and will proceed parallel to the ray H_2F_2 . The points of intersection of the incident ray AT_1 and the conjugate emergent ray T_2R with the axis give the positions of the nodal points. It is clear that the two points N_1 and N_2 are a pair of conjugate points and the incident ray AN_1 is parallel to the conjugate emergent ray T_2R .

Further

$$\tan \alpha_1 = \tan \alpha_2$$

The ratio $\frac{\tan \alpha_2}{\tan \alpha_1} = \gamma$ represents the angular magnification.

∴

$$\frac{\tan \alpha_2}{\tan \alpha_1} = 1 \quad \dots(40.11)$$

Therefore, the points N_1 and N_2 are a pair of conjugate points on the axis having unit positive angular magnification.

(b) The distance between two nodal points is always equal to the distance between two principal points.

Referring to Fig. 40.5, we see that in the right angled Δ^{les} $T_1P_1N_1$ and $T_2P_2N_2$

$$T_1P_1 = T_2P_2$$

$$\angle T_1N_1P_1 = \angle T_2N_2P_2 = \alpha$$

Therefore, the two Δ^{les} are congruent.

∴

$$P_1N_1 = P_2N_2$$

Adding N_1P_2 to both the sides, we get

∴

$$P_1N_1 + N_1P_2 = P_2N_2 + N_1P_2$$

∴

$$P_1P_2 = N_1N_2 \quad \dots(40.12)$$

Thus, the distance between the principal points N_1 and N_2 is equal to the distance between the principal points P_1 and P_2 .

(c) The nodal points N_1 and N_2 coincide with the principal points P_1 and P_2 respectively whenever the refractive indices on either side of the lens are the same.

Now consider the two right angled Δ^{les} AF_1N_1 and $H_2P_2F_2$ in Fig. 40.5.

$$AF_1 = H_2P_2$$

$$\angle A N_1 F_1 = \angle H_2 F_2 P_2$$

\therefore The two Δ^{les} are congruent.

$$F_1N_1 = P_2F_2$$

But

$$F_1N_1 = F_1P_1 + P_1N_1$$

\therefore

$$F_1P_1 + P_1N_1 = P_2F_2$$

\therefore

$$P_1N_1 = P_2F_2 - F_1P_1$$

Also

$$P_2F_2 = +f_2 \text{ and } P_1F_1 = -f_1$$

\therefore

$$P_1N_1 = P_2N_2 = (f_1 + f_2)$$

As the medium is the same, say air, on both the sides of the system

$$f_2 = -f_1$$

\therefore

$$P_1N_1 = P_2N_2 = 0$$

...(40.13)

Thus, the principal points coincide with the nodal points when the optical system is situated in the same medium.

40.6 CONSTRUCTION OF IMAGE USING CARDINAL POINTS

From the knowledge of cardinal points of an optical system, the image of an object can be constructed. For this, it is not necessary to know the position and curvature of the refracting surfaces or the nature of the intermediate media.

Let AB be an object placed at some distance from first focal point F_1 as in Fig. 40.6. Let us consider a ray AH_1 parallel the principal axis which meets the first principal plane at H_1 . According to the property of principal planes, the ray emerges from the second principal plane from point H_2 such that $H_1P_1 = H_2P_2$ and it passes through the second focal point F_2 .

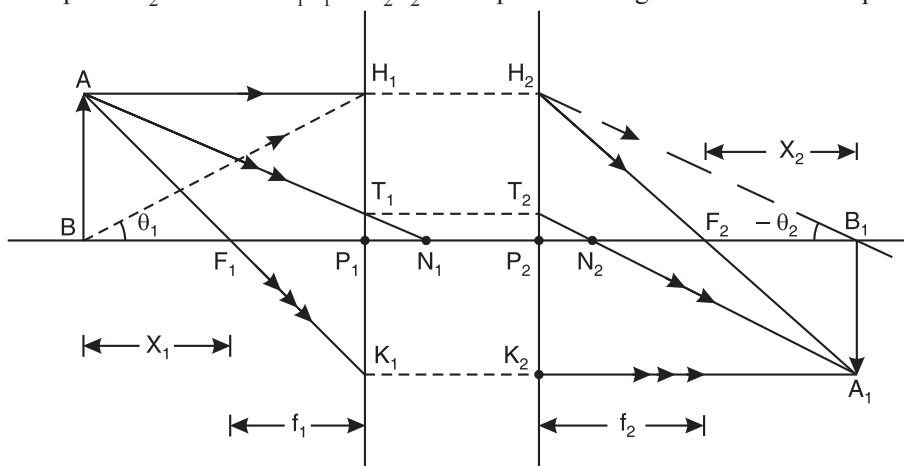


Fig. 40.6

Next let us consider a second ray passing through the first focal point F_1 . It strikes the first focal plane at K_1 and then it emerges parallel to the principal axis. It meets the second principal plane at K_2 such that $H_1K_1 = H_2K_2$.

Again let us consider a third ray AN_1 directed towards first nodal point N_1 . This ray emerges from second nodal point N_2 along N_2A_1 parallel to its original direction.

The three emergent rays meet at A_1 . Therefore, A_1 is the image of A . Similarly the images of other points of AB can be determined which give rise to image A_1B_1 of the object AB .

Newton's formula:

Let the distance of the object from the first focal point F_1 be x_1 and the distance of the image from the second focal point F_2 be x_2 .

Referring to the Fig. 40.6, it is seen that $\Delta^{\text{les}} ABF_1$ and $F_1K_1P_1$ are similar.

$$\begin{aligned} \therefore \quad \frac{K_1P_1}{AB} &= \frac{P_1F_1}{BF_1} \quad \text{But, } K_1P_1 = A_1B_1 \\ \therefore \quad \frac{A_1B_1}{AB} &= \frac{f_1}{x_1} \end{aligned} \quad \dots(40.14)$$

Further, $\Delta^{\text{les}} A_1B_1F_2$ and $H_2P_2F_2$ are similar.

$$\begin{aligned} \frac{A_1B_1}{H_2P_2} &= \frac{B_1F_2}{P_2F_2} \quad \text{But, } H_2P_2 = AB \\ \therefore \quad \frac{A_1B_1}{AB} &= \frac{x_2}{f_2} \end{aligned} \quad \dots(40.15)$$

From equations (40.14) and (40.15), we get

$$\frac{h_2}{h_1} = \frac{f_1}{x_1} = \frac{x_2}{f_2} \quad \dots(40.16)$$

or

$$x_1x_2 = f_1f_2 \quad \dots(40.17)$$

This is the **Newton's formula**. In the foregoing discussion, the distances of the image and the object have been measured from their respective foci. But it is sometimes convenient to measure the conjugate distances from the principal points.

40.6.1 Relationship Between f_1 and f_2

Referring to Fig. 40.6 and using the sign convention,

$$P_1B = H_1A = -u, P_2B_1 = K_2A_1 = +v$$

$$\text{Also, } AB = P_1H_1 = P_2H_2 = +h_1, A_1B_1 = K_1P_1 = K_2P_2 = -h_2$$

$$K_1H_1 = K_1P_1 + P_1H_1 = -h_2 + h_1 \text{ and } K_2H_2 = K_2P_2 + P_2H_2 = -h_2 + h_1$$

$\Delta^{\text{les}} K_1F_1P_1$ and K_1AH_1 are similar.

$$\begin{aligned} \therefore \quad \frac{P_1F_1}{H_1A} &= \frac{K_1P_1}{K_1H_1} \\ \frac{-f_1}{-u} &= \frac{-h_2}{-h_2 + h_1} \\ \text{or} \quad \frac{f_1}{u} &= \frac{-h_2}{-h_2 + h_1} \end{aligned} \quad \dots(40.18)$$

$\Delta^{\text{les}} H_2P_2F_2$ and $H_2K_2A_1$ are similar.

$$\therefore \quad \frac{P_2F_2}{K_2A_1} = \frac{P_2H_2}{K_2H_2}$$

$$\frac{f_2}{v} = \frac{h_2}{-h_2 + h_1} \quad \dots(40.19)$$

Adding the equations (40.18) and (40.19), we get

$$\begin{aligned} \frac{f_1}{u} + \frac{f_2}{v} &= \frac{-h_2 + h_1}{-h_2 + h_1} \\ \therefore \quad \frac{f_1}{u} + \frac{f_2}{v} &= 1 \end{aligned} \quad \dots(40.20)$$

Equ. (40.20) can be rewritten as

$$\frac{v}{u} = \frac{(v - f_2)}{f_1} \quad \dots(40.21)$$

When the system is situated in air $f_2 = -f_1 = f$.

$$\therefore \quad f_1 = -f \quad \text{and} \quad f_2 = f. \quad \dots(40.22)$$

40.7 NODAL SLIDE

A **nodal slide** is a particular type of horizontal metal support for a lens system, which is capable of rotation about a vertical axis (Fig. 40.7). The nodal slide provides a convenient method for locating the focal and nodal points and for determining the focal length of a lens system.

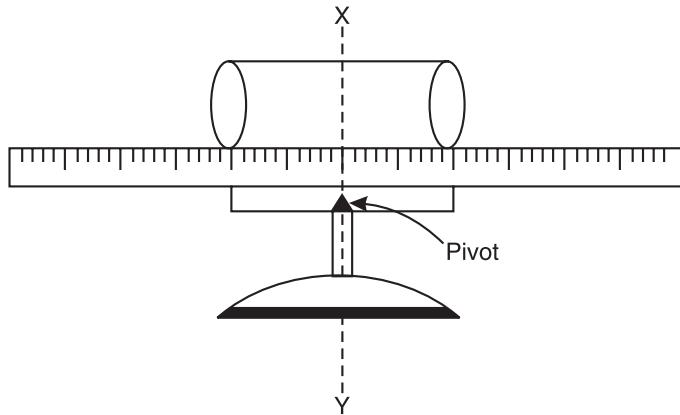


Fig. 40.7

Principle:

Any optical system has two nodal points N_1 and N_2 . An incident light ray directed towards N_1 , after refraction through the system, proceeds from N_2 in a direction parallel to the incident ray. The method of locating the nodal points with the help of the nodal slide involves the following principle.

“If a parallel beam of light is incident on a convergent lens system, it forms an image on a screen held at its second focal plane. When the lens system is rotated through a small angle about a vertical axis through its second nodal point, the image does not shift laterally and remains stationary.”

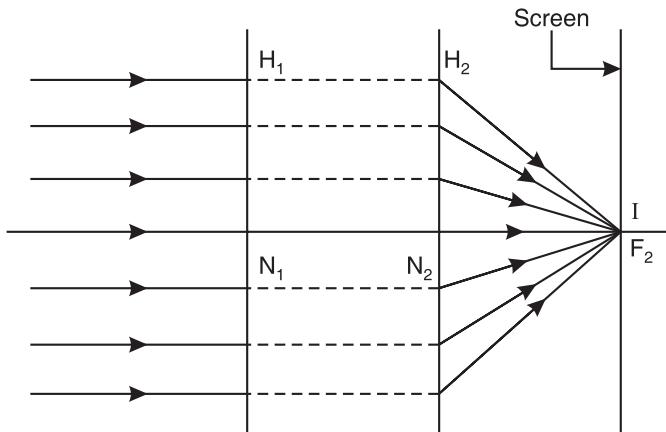


Fig. 40.8

Let us suppose a beam of parallel rays is incident on a coaxial lens system (see Fig. 40.8). The beam passes through the system and converges to the second focus F_2 and a real image is formed on the screen.

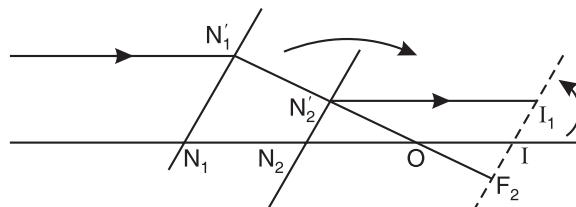


Fig. 40.9

Now let the system be rotated about a perpendicular axis through \$O\$, which lies between \$N_2\$ and \$F_2\$ (Refer to Fig. 40.9). Due to this rotation, the nodal points \$N'_1\$ and \$N'_2\$ shift to the positions \$N'_1\$ and \$N'_2\$ respectively. A ray incident at \$N'_1\$ travels along \$N'_2 I_1\$ parallel to the incident ray (see Fig. 40.9). Since the incident beam is parallel, the image lies on the second focal plane. The point of intersection of the ray \$N'_2 I_1\$ with the focal plane gives the new position of the image. Thus, when the axis of rotation lies between \$N_2\$ and \$F_2\$, a slight rotation of the system changes the position of the image.

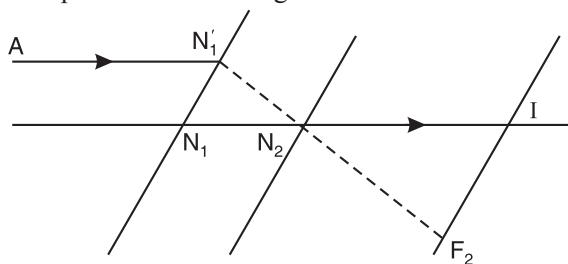


Fig. 40.10

Next let us say the system is rotated through a small angle about an axis passing through \$N_2\$. Then \$N_1\$ shifts to \$N'_1\$ while \$N_2\$ remains fixed. A parallel ray incident at \$N'_1\$ on passing through \$N_2\$ follows the path \$N_2 I\$. Therefore, the image remains stationary, as seen in Fig. 40.10.

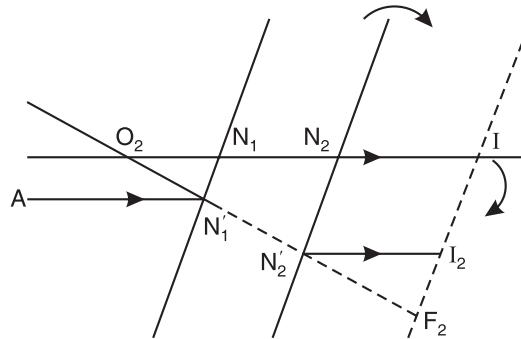


Fig. 40.11

If the axis of rotation lies before N_1 , as in Fig. 40.11, any small amount of rotation displaces N_1 and N_2 and consequently the image position changes.

Thus, the position of the axis of rotation for which there is no displacement of the image can be found. It gives the second nodal point. As the media on both sides of the system being the same, the second nodal point coincides with second principal point.

40.7.1 Determination of nodal points:

The experimental arrangement consists of an optical bench on which four uprights are kept. They carry a plane mirror, the nodal slide, and a screen provided with a slit fitted with cross wires and lamp housing (see Fig. 40.12). Light from the lamp passes through the slit and is incident on the lens system. It is rendered parallel and on passing through the lens system it is reflected back by the vertical plane mirror. The reflected light once again passes through the lens system and is brought to a focus in the plane of the stilt, as shown in Fig. 40.13.

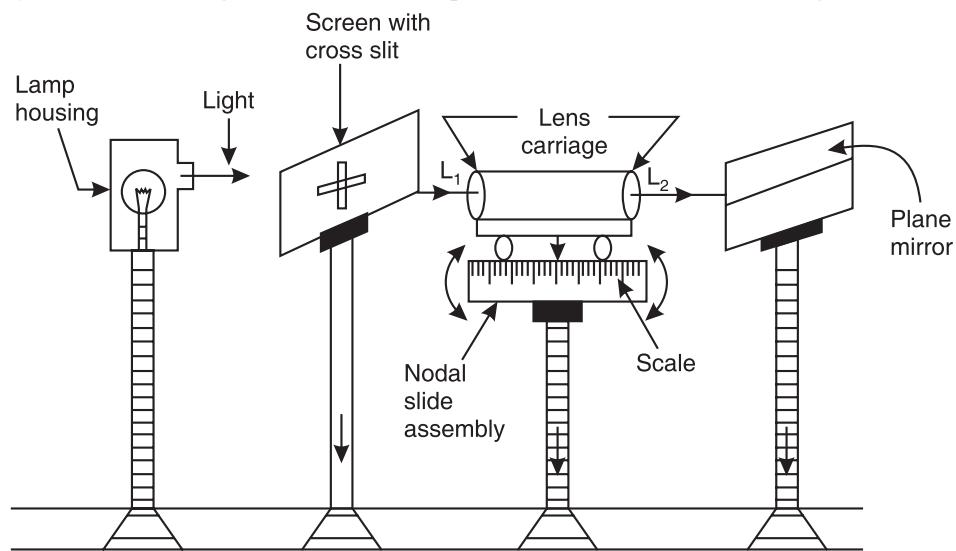


Fig. 40.12

The distance between the optical system and the screen is adjusted in such a way that a well-defined image of the slit is obtained on the screen. The image of the slit is formed slightly to one side of the slit itself. It is obvious that the centre of the slit is at the first focal

point of the lens. The nodal slide carrying the lens system is now rotated through a small angle and it will be found that the image shifts sideward to the right or to the left. The nodal slide and its stand are then adjusted such that the direction of rotation of the image changes its sign and finally the image remains stationary for a slight rotation of the carriage. When this condition is reached, the axis of rotation passes through the second nodal point N_2 . The other focal point nodal point can be determined by turning the nodal slide through 180° and repeating the experiment. Since the medium on both sides of the lens system is the same (air), the nodal points are also the principal points. The distance between the screen and the axis of rotation for the stationary image is an accurate measure of the first focal length of the lens system.

40.8 EQUIVALENT FOCAL LENGTH OF A COAXIAL SYSTEM OF TWO THIN LENSES

We now determine the equivalent focal length of a coaxial optical system located in air medium by assuming that the object is at infinity, as shown in Fig. 40.14.

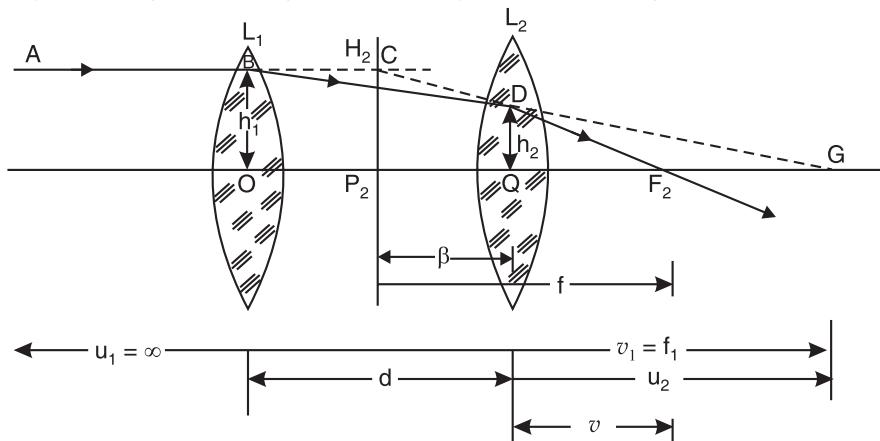


Fig. 40.14

AB is a ray of light coming from an object situated at a very large distance, such that $u_1 = \infty$. The lens L_1 , if alone, would form an image at G . However, because of the presence of the second lens L_2 , G becomes the virtual object for L_2 . The ray BD , instead of going along BDG , refracts along the path DF_2 . When the ray AB is produced forward and the ray DF_2 backward, they intersect at H_2 . The plane H_2P_2 normal to the axis may be considered as the plane at which the refraction occurred and this plane is called **principal plane**.

Now, we can write the expression for the refraction taking place at the surface of first lens as follows.

$$\frac{1}{v_1} - \frac{1}{u_1} = \frac{1}{f_1} \quad \therefore \frac{1}{OG} - \frac{1}{u_1} = \frac{1}{f_1}$$

As $u_1 = \infty$, we obtain $OG = f_1$

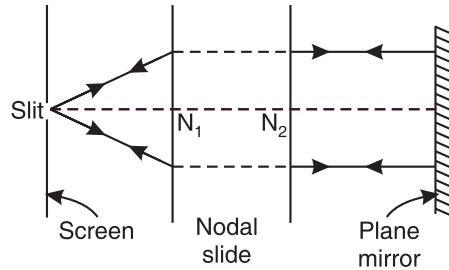


Fig. 40.13

The equation for the refraction at the second lens may be written as

$$\frac{1}{v} - \frac{1}{u_2} = \frac{1}{f_2} \quad \therefore \frac{1}{QF_2} - \frac{1}{QG} = \frac{1}{f_2}$$

or

$$\frac{1}{QF_2} = \frac{1}{f_2} + \frac{1}{f_1 - d} \quad \therefore \frac{1}{QF_2} = \frac{f_1 + f_2 - d}{f_2(f_1 - d)} \quad \dots(40.23)$$

The $\Delta^{\text{les}} BOG$ and DQG are similar and also the $\Delta^{\text{les}} CP_2F_2$ and DQF_2 are similar.

$$\frac{BO}{OG} = \frac{DQ}{QG} \quad \therefore \frac{h_1}{f_1} = \frac{h_2}{(f_1 - d)} \quad \text{or} \quad \frac{h_1}{h_2} = \frac{f_2}{(f_1 - d)} \quad \dots(40.24)$$

$$\frac{CP_2}{P_2F_2} = \frac{DQ}{QF_2} \quad \therefore \frac{h_1}{f} = \frac{h_2}{QF_2} \quad \text{or} \quad \frac{h_1}{h_2} = \frac{f}{QF_2}$$

$$\therefore \frac{h_1}{h_2} = \frac{f(f_1 + f_2 - d)}{f_2(f_1 - d)} \quad \dots(40.25)$$

From equs. (40.24), (40.25) and (40.23), we get

$$\frac{h_1}{h_2} = \frac{f_1}{(f_1 - d)} = \frac{f(f_1 + f_2 - d)}{f_2(f_1 - d)}$$

or

$$f = \frac{f_1 f_2}{f_1 + f_2 - d} \quad \dots(40.26)$$

or

$$f = \frac{f_1 f_2}{\Delta} \quad \dots(40.27)$$

where $\Delta = f_1 + f_2 - d$ is called the *optical separation* or *optical interval* between the two lenses.

Equ. (40.27) represents the **focal length of the equivalent lens**.

Also, because the location of the focal point F_2 is determined by QF_2 , which is known from the equation, the position of the principal plane P_1 is specified by the value of f calculated from the equ. (40.27).

40.9 CARDINAL POINTS OF A COAXIAL SYSTEM OF TWO THIN LENSES

We determine the cardinal points of a coaxial optical system by assuming that the object is at infinity, as shown in Fig. 40.14. In Fig. 40.14 AB is a ray of light coming from an object situated at a very large distance, such that $u_1 = \infty$. The lens L_1 , if alone, would form an image at G . However, because of the presence of the second lens L_2 , G becomes the virtual object for L_2 . The ray BD , instead of going along BDG , refracts along the path DF_2 . When the ray AB is produced forward and the ray DF_2 backward, they intersect at H_1 . The plane H_1P_1 normal to the axis may be considered as the plane at which the refraction occurred and this plane is called **principal plane**.

(i) Second Principal point:

Let us say the second principal plane H_2P_2 is located at a distance of $L_2P_2 = \beta$ from the second lens L_2 . According to sign convention β would be *negative* as it is measured toward the left of the lens.

$$\therefore QF_2 = f - (-\beta) = f + \beta$$

We can determine β using the equation for f into the above relation. Using equ. (40.23) we writes

$$f + \beta = \frac{f_2(f_1 - d)}{f_1 + f_2 - d}$$

$$\beta = -f + \frac{f_2(f_1 - d)}{f_1 + f_2 - d} = \frac{-f_1 f_2}{\Delta} + \frac{f_2 f_1 - f_2 d}{\Delta}$$

where used equ. (40.27) for f and $\Delta = f_1 + f_2 - d$.

$$\therefore \beta = -f_2 \frac{d}{\Delta} \quad \dots(40.28)$$

$$\text{or } \beta = -\frac{f_2 d}{f_1 + f_2 - d} \quad \dots(40.29)$$

But from equ. (40.26), we have $f_1 + f_2 - d = \frac{f_1 f_2}{f}$. Therefore,

$$\therefore \beta = -\frac{f d}{f_1} \quad \dots(40.30)$$

(ii) First Principal Point:

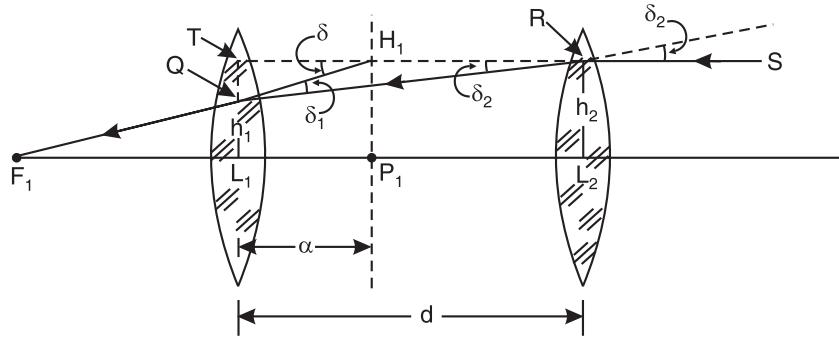


Fig. 40.15

By considering a ray of light parallel to the axis and incident on the second lens L_2 from the right side (see Fig. 40.15), we can show that the distance of first principal plane, $L_1 P_1 = \alpha$ from the first lens L_1 is given by

$$\therefore \alpha = f_1 \frac{d}{\Delta} \quad \dots(40.31)$$

$$\text{Also, } \alpha = +\frac{f d}{f_2} \quad \dots(40.32)$$

Note: In a combination of two lenses, the sequence of the principal planes is in the reverse order- $H_2 P_2$ is to the left of the centre and $H_1 P_1$ is to the right.

(iii) Second Focal point:

Referring to Fig. 40.14, the distance of the second focal point F_2 from the second lens L_2 is given by

$$L_2 F_2 = P_2 F_2 - P_2 L_2$$

$$= f - (-L_2 P_2) = f + \beta$$

$$\begin{aligned}
 &= f + \left(-\frac{f d}{f_1} \right) \\
 \therefore L_2 F_2 &= f \left(1 - \frac{d}{f_1} \right) \quad \dots(40.33)
 \end{aligned}$$

(iv) First Focal point:

The distance of the first focal point F_1 from the first lens L_1 is given by

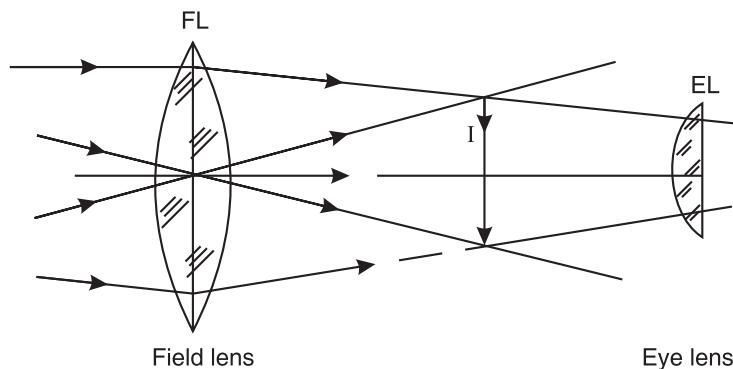
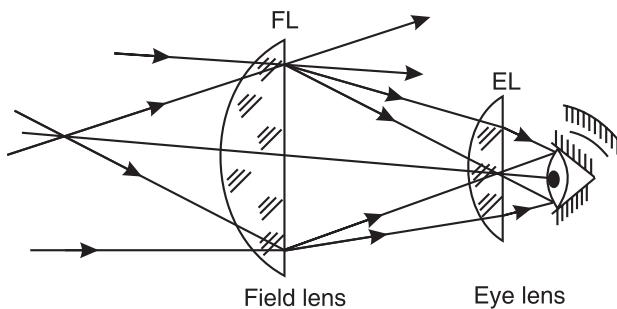
$$\begin{aligned}
 L_1 F_1 &= P_1 F_1 - P_1 L_1 \\
 &= -f - (-L_1 P_1) = -f + \alpha \\
 &= -f + \left(\frac{f d}{f_2} \right) \\
 \therefore L_1 F_1 &= -f \left(1 - \frac{d}{f_2} \right) \quad \dots(40.34)
 \end{aligned}$$

(v) and (vi) Nodal Points:

As the medium on either side of the lenses is air, the nodal points N_1 and N_2 coincide with the principal points P_1 and P_2 respectively. Then $N_1 = P_1$ and $N_2 = P_2$.

40.10 EYEPieces

An optical instrument such as a microscope or telescope is required to produce a magnified image free from aberrations and a bright image covering a wide field of view. If a single lens is used as an eyepiece, the final image will suffer from spherical and chromatic aberrations. Another drawback is that the field of view is small, which becomes lesser and lesser as the magnification of the instrument is increased.

**Fig. 40.16****Fig. 40.17**

An **eyepiece** is a specially designed magnifier that gives a perfect and bright magnified image of an object. It consists of two convex lenses separated by a finite distance. The first lens towards the object is called the **field lens**. It forms a real image of the object under examination. The second lens, called the **eye lens**, enlarges this image further to form a final image and which is then viewed by the eye. The rays passing through the outer portions of the image formed by the field lens are refracted through the peripheral portions of the eye lens and they cannot simultaneously enter the small aperture of the pupil of the eye placed close to the eye lens (Fig. 40.16). Hence, only that part of the image, which is nearer to the axis, will be seen. Therefore, the final image will cover a small field of view. The field of view will progressively decrease as the distance between the field lens and eye lens is increased. The distance is varied in order to increase the magnification. In other words, the greater the magnifying power of the instrument, the smaller the field of view. Therefore field lens is used before the eye lens in the eyepiece to cause all the rays from the image to enter the eye lens. The function of the field lens is to gather in more of the rays from the objective toward the axis of the eyepiece (See Fig. 40.17). The field lens and the eye lens together constitute an **eyepiece**. The two lenses are made and kept in such a way that their combination is achromatic and free from spherical aberrations.

Two common eyepieces are the Huygens and the Ramsden eyepieces.

40.11 HUYGENS EYEPICE

Construction: Huygens' eyepiece consists of two plano-convex lenses of focal lengths $3f$ and f which are separated by a distance $2f$. The lenses are made of the same kind of glass and their curved surfaces face the incident light. The first lens acts as objective and the second lens acts as eye lens.

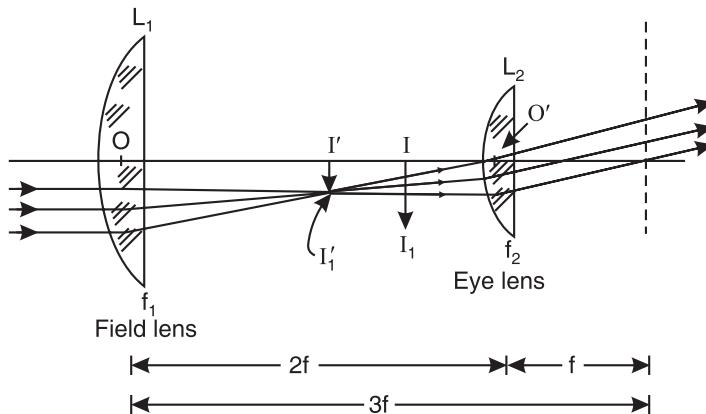


Fig. 40.18

1. *Condition for minimum spherical aberration:* For spherical aberration to be a minimum the distance d between the two lenses must be equal to the difference of their focal lengths, that is

$$d = f_1 - f_2.$$

$$\begin{aligned} \text{In Huygen's eyepiece} \quad f_1 &= 3f, f_2 = f. \\ \therefore \quad d &= 3f - f = 2f. \end{aligned}$$

Thus, it satisfies the condition for minimum spherical aberration.

Also, as the convex surfaces of the field and the eye lenses face the incident ray, the total deviation due to the combination is divided into four parts which makes the combination to have minimum spherical aberration.

2. *Condition for achromatism:* For chromatic aberration to be a minimum the distance between the two lenses must be equal to the mean of their focal lengths

$$d = \frac{f_1 + f_2}{2}.$$

In Huygen's eyepiece $d = \frac{1}{2} (3f + f) = 2f$.

Thus, it satisfies the condition of minimum chromatic aberration.

3. *Working:* The eyepiece works in conjunction with an optical instrument. Referring to the Fig. 40.18, I_1 is the image of the distant object formed by the optical instrument in the absence of the field lens L_1 . With the field lens in position, the rays get refracted on passing through L_1 and the image $I' I_1'$ is formed. When image lies in the focal plane of the eye-lens, the magnified image is seen at infinity by the eye located at the exit pupil.
4. *Equivalent Focal Length:* The equivalent focal length of the eyepiece can be found as follows. If F is the equivalent focal length of the eyepiece, then it is given by

$$\frac{1}{F} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2}$$

Using $f_1 = 3f$, $f_2 = f$ and $d = 2f$ in the above equation, we get

$$\begin{aligned} \frac{1}{F} &= \frac{1}{3f} + \frac{1}{f} - \frac{2f}{3f^2} \\ \therefore \quad \frac{1}{F} &= \frac{1}{f} + \frac{1}{3f} - \frac{2}{3f} = \frac{2}{3f} \\ \therefore \quad F &= 3f/2 \end{aligned}$$

The equivalent lens lies behind field lens at a distance of

$$x = \frac{d \times F}{f} = \frac{2f \times 3f/2}{f} = 3f$$

In other words, the equivalent lens is at a distance of $3f - 2f = f$ behind the eye lens.

5. *Cardinal points of Huygens eyepiece:*

- i. *Position of Principal Points:*

The distance of the first principal point H_1 from the field lens is given by

$$\alpha_1 = \frac{d \times F}{f_2} = \frac{2f \times 3f/2}{f} = 3f$$

Thus, the first principal point H_1 lies at a distance of $3f$ to the right of the field lens.

The distance of the second principal point H_2 from the field lens is given by

$$\alpha_2 = -\frac{d \times F}{f_1} = -\frac{2f \times 3f/2}{3f} = -f$$

Thus, the second principal point H_2 lies at a distance of f to the left of the eye lens.

ii. Position of the Focal Points:

The distance of the first focal point F_1 from the field lens is given by

$$\beta_1 = -F \left(1 - \frac{d}{f_2} \right) = -\frac{3f}{2} \left(1 - \frac{2f}{f} \right) = \frac{3f}{2}$$

Thus, the first focal point F_1 lies at a distance $\frac{3f}{2}$ to the right of the field lens.

The distance of the second focal point F_2 from the field lens is given by

$$\beta_2 = F \left(1 - \frac{d}{f_1} \right) = \frac{3f}{2} \left(1 - \frac{2f}{3f} \right) = -\frac{f}{2}$$

Thus, the second focal point F_2 lies at a distance $\frac{f}{2}$ to the right of the eye lens.

iii. Nodal points: As the medium surrounding the lenses of the eye-piece is air, the nodal points N_1 and N_2 coincide with the principal points P_1 and P_2 respectively.

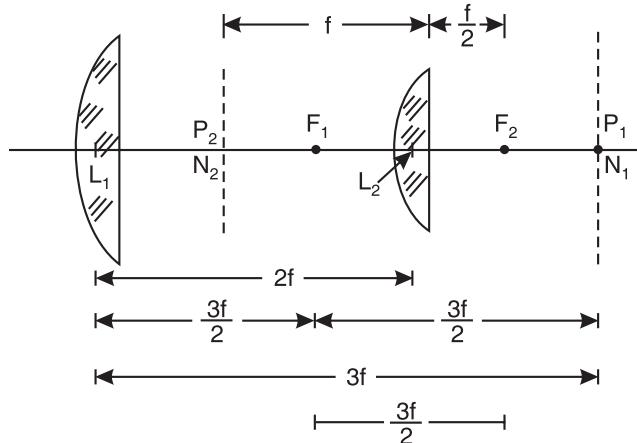


Fig. 40.19

6. Huygen's eye-piece is a negative eye-piece:

In Huygens eyepiece, the first focal plane of the eye-piece lies within the eye-piece and no real object can be placed there. Such an eye-piece is called a negative eye-piece. This eye-piece cannot be used to examine directly an object or a real image formed but it can be used only to examine a virtual image. Hence it is called a *negative eye-piece*. This eye-piece is used in microscopes and other optical instruments using white light only.

7. Position of Cross-wires:

For quantitative measurements cross-hairs or a recticle with a scale is used with the eye-piece. It is required for the recticle to be in focus with the image. Hence, the recticle must be placed in the focal plane of real image. In case of Huygens eye-piece the cross-wires are to be held in between the field and eye-lenses inside the eye-piece. Therefore, they are magnified by the eye-lens only whereas the image is magnified by the field lens as well as the eye-lens. It is not a drawback in simple observations; but in case any measurement the different magnifications of the image and the cross-wire scale leads to erroneous results.

Merits and Demerits

- (i) The Huygens' eyepiece is fully free from chromatic aberration because the distance between the lenses is equal to half the sum of their focal lengths.
- (ii) Spherical aberration is also minimum because the distance between the two lenses is equal to the difference of their focal lengths.
- (iii) The field of view of this eyepiece is smaller than that of Ramsden's eyepiece.

40.12 RAMSDEN EYEPiece

Construction: Ramsden's eyepiece consists of two plano-convex lenses each of focal length f separated by a distance equal to $(2/3)f$. The field lens is a little larger than the intermediate image and is placed close to this image to allow as much light as possible to pass through it. The eye lens has a smaller diameter but carries out the actual magnification.

Theory: The objective forms the real inverted image II' of a distant object. This serves as an object for the field lens, which gives rise to a virtual image $I_1I'_1$. $I_1I'_1$ in turn serves as an object for the eye lens, which gives the final image at infinity, because $I_1I'_1$ is made to lie at its principal focus.

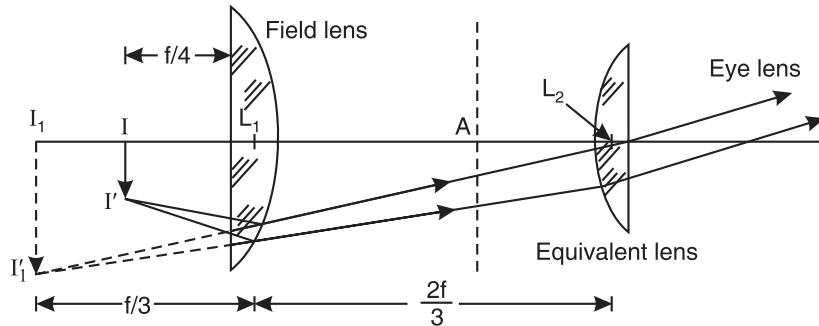


Fig. 40.20

1. *Condition for minimum spherical aberration:* The spherical aberration is reduced by making the both the lenses plano-convex and by keeping them with their curved surfaces facing each other, as shown in Fig. 40.20.
2. *Condition for achromatism:* For chromatic aberration to be a minimum the distance between the two lenses must be equal to the mean of their focal lengths

$$d = \frac{f_1 + f_2}{2}$$

In Ramsden's eye-piece $f_1 = f_2 = f$.

$$\therefore d = \frac{f + f}{2} = f.$$

It means that the field lens should be placed in the focal plane of the eye-lens. However, in this position any dust particle or scratch on the field lens would be magnified and the final image would be spoiled. Therefore, the distance between the two lenses is kept $\frac{2}{3}f$ (a little less than f).

3. *Working:* The eyepiece works in conjunction with an optical instrument. Referring to the Fig. 40.20, II' is the image of the distant object formed by the optical instrument

which acts as an object for the field lens. The rays get refracted on passing through the field lens L_1 and the image $I_1 I'_1$ is formed. When image lies in the focal plane of the eye-lens, the magnified image is seen at infinity by the eye located at the exit pupil.

4. *Equivalent focal length:* The equivalent focal length of the eyepiece can be found as follows. If F denotes the focal length of the equivalent lens, then

$$\begin{aligned}\frac{1}{F} &= \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2} \\ \therefore \frac{1}{F} &= \frac{1}{f} + \frac{1}{f} - \frac{(2/3)f}{f^2} \\ &= \frac{2}{f} - \frac{2}{3f} = \frac{4}{3f} \\ \therefore F &= 3f/4\end{aligned}$$

The final image formed by an eye-piece is at infinity. Therefore, the final image formed by an equivalent lens must be at infinity. For this, II' should be in the focal plane of the equivalent lens. Hence the equivalent lens of focal length $3f/4$ must be placed behind the field lens at a distance

$$x = \frac{F \times d}{f} = \frac{(3/4)f \times (2/3)f}{f} = \frac{f}{2}$$

Thus the equivalent lens lies between the field lens and the eye lens.

5. Cardinal points of Ramsden eyepiece:

i. *Position of Principal Points:*

The distance of the first principal point H_1 from the field lens is given by

$$\alpha_1 = \frac{d \times F}{f_2} = \frac{\frac{2}{3}f \times \frac{3}{4}f}{f} = +\frac{f}{2}$$

Thus, the first principal point H_1 lies at a distance of $f/2$ to the right of the field lens.

The distance of the second principal point H_2 from the field lens is given by

$$\alpha_2 = -\frac{d \times F}{f_1} = -\frac{\frac{2}{3}f \times \frac{3}{4}f}{f} = -\frac{f}{2}$$

Thus, the second principal point H_2 lies at a distance of $f/2$ to the left of the eye lens.

ii. *Position of the Focal Points:*

The distance of the first focal point F_1 from the field lens is given by

$$\beta_1 = -F \left(1 - \frac{d}{f_2} \right) = -\frac{3f}{4} \left(1 - \frac{\frac{2}{3}f}{f} \right) = -\frac{f}{4}$$

Thus, the first focal point F_1 lies at a distance $\frac{f}{4}$ to the left of the field lens.

The distance of the second focal point F_2 from the field lens is given by

$$\beta_2 = F \left(1 - \frac{d}{f_1} \right) = \frac{3f}{4} \left(1 - \frac{\frac{2}{3}f}{f} \right) = +\frac{f}{4}$$

Thus, the second focal point F_2 lies at a distance $\frac{f}{4}$ to the right of the eye lens.

- iii. *Nodal points:* As the medium surrounding the lenses of the eye-piece is air, the nodal points N_1 and N_2 coincide with the principal points P_1 and P_2 respectively.

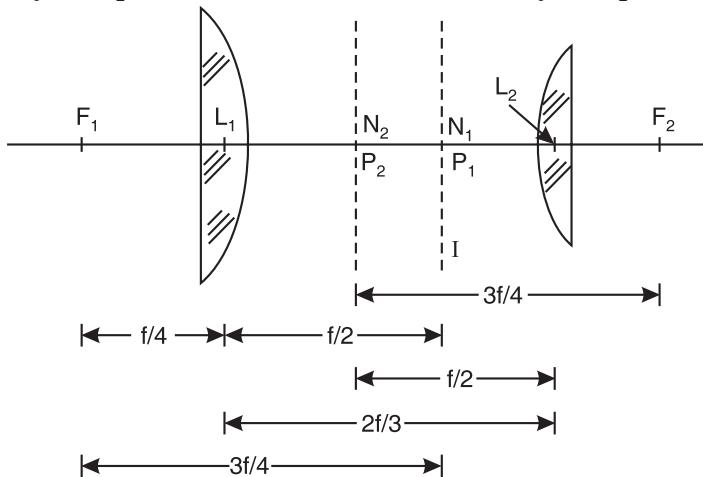


Fig. 40.21

6. *Position of Cross-wires:* The cross-wires should be placed at the position of I . Now, the position of I relative to field lens can be found as follows. If the final image is to be formed at infinity, the image I should lie in the focal plane of the equivalent lens. In other words, the distance F_1N_1 should be equal to $F = 3f/4$. Since $L_1N_1 = f/2$, $F_1L_1 = f/4$. Therefore, the objective should produce the image at a distance of $f/4$ in front of the field lens. A fine scale may be placed here if it is desired to measure the size of the image. Since the scale and image would be magnified equally, the measurement would be trustworthy.

7. *Ramsden eye-piece is a positive eye-piece:* If the first focal plane of an eye-piece lies outside the eye-piece in the object space, then the real object can be placed on first focal plane to be in focus with the final image. Such an eye-piece is called *positive eye-piece*. In case of Ramsden eyepiece, the focal plane of the eyepiece lies to the left of the field lens and is in the object space. Hence it is a positive eyepiece and can be used to examine a real object or a real image.

Merits and Demerits

- (i) The field of view of this eyepiece is fairly wide.
- (ii) It is not entirely free from chromatic aberration since the distance between the two lenses is not equal to half the sum of their focal lengths. However, chromatic aberration is minimised by using an achromatic combination both for the field lens and the eye lens.
- (iii) Spherical aberration is minimised by using two plano-convex lenses thereby spreading deviation over four surfaces.

Ramsden's eyepiece is used practically in all instruments where measurements of the size of the final image are to be made.

40.13 COMPARISON OF RAMSDEN EYEPIECE WITH HUYGENS EYEPIECE

	Ramsden Eyepiece	Huygens Eyepiece
1.	Ramsden's eyepiece is a positive eyepiece. The image formed by the objective lies in front of the field lens. Therefore, cross-wires can be used.	Huygens' eyepiece is a negative eyepiece. The image formed by the objective lies in between the two lenses. Therefore, cross-wires cannot be used.
2.	The condition for minimum spherical aberration is not satisfied. But by spreading the deviations over four surfaces, spherical aberration is minimized.	The condition for minimum spherical aberration is satisfied.
3.	It does not satisfy the condition for achromatism but can be made achromatic by using an achromatic doublet as the eye lens.	It satisfies the condition for achromatism.
4.	It is achromatic for only two chosen colours.	It is achromatic for all colours.
5.	The other types of aberrations are better eliminated. Coma is absent and distortion is 5% less.	Other aberrations like pincushion distortion are not eliminated.
6.	The eye clearance is 5% higher.	The eye clearance is too small and less comfortable.
7.	It is used for quantitative purposes in microscopes and telescopes.	It is used for qualitative purposes in microscopes and telescopes.
8.	Its power is positive.	Its power is negative.
9.	The two principal planes are crossed.	The two principal planes are crossed.
10.	It can be used as a simple microscope because the first principal plane lies to the left of the field lens and the focal plane is real.	It cannot be used as simple microscope because the first focal plane lies to the right of the field lens and the focal plane is virtual.
11.	The nodal points coincide with the principal points.	The nodal points coincide with the principal points.

WORKED-OUT EXAMPLES

Example 40.1: A coaxial lens system placed in air has two lenses of focal lengths $3F$ and F separated by a distance $2F$. Find the positions of the cardinal points.

Solution:

$$f_1 = 3F, f_2 = F, d = 2F$$

$$f = \frac{f_1 f_2}{f_1 + f_2 - d}$$

$$\therefore f = \frac{3F \times F}{3F + F - 2F} = \frac{3}{2}F$$

$$\alpha = +\frac{df}{f_2} = \frac{2F \cdot 3F / 2}{F} = +3F$$

Therefore, the first principal point P_1 is at a distance $3F$ to the right of the first lens, as shown in Fig. 40.22.

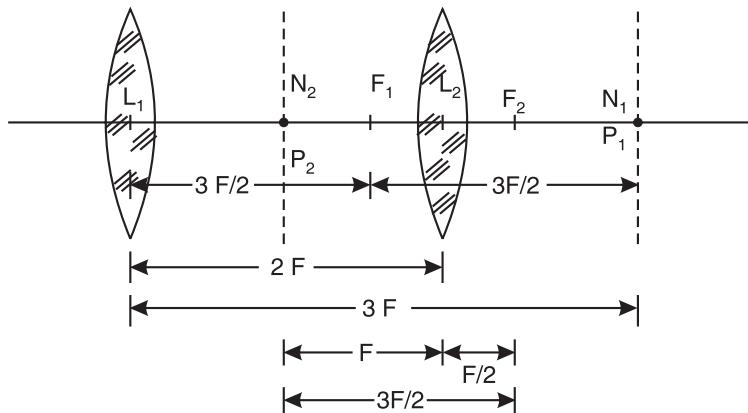


Fig. 40.22

$$\beta = -\frac{df}{f_1} = \frac{-2F \cdot 3F/2}{3F} = -F$$

The second principal point P_2 is at a distance F to the left of the second lens.

The first focal point F_1 is at a distance $3F - 3F/2 = 3F/2$ from the first lens and F_2 is at a distance $3F/2 - F = F/2$ from the second lens. As the medium on the two sides of the lens system is the same, the nodal points N_1 and N_2 coincide with P_1 and P_2 .

Example 40.2: Two thin convex lenses of focal lengths 20 cm and 5 cm are kept coaxially separated by a distance of 10 cm. Plot the positions of the cardinal points for the combination.

Solution: Given that $f_1 = 20$ cm, $f_2 = 5$ cm and $d = 10$ cm

$$f = \frac{f_1 f_2}{f_1 + f_2 - d} = \frac{20 \text{ cm} \times 5 \text{ cm}}{20 \text{ cm} + 5 \text{ cm} - 10 \text{ cm}} = 6.67 \text{ cm}$$

First Principal Point:

$$\alpha = \frac{fd}{f_2} = \frac{6.67 \text{ cm} \times 10 \text{ cm}}{5 \text{ cm}} = 13.33 \text{ cm}$$

Second Principal Point:

$$\beta = -\frac{fd}{f_1} = -\frac{6.67 \text{ cm} \times 10 \text{ cm}}{20 \text{ cm}} = -3.33 \text{ cm}$$

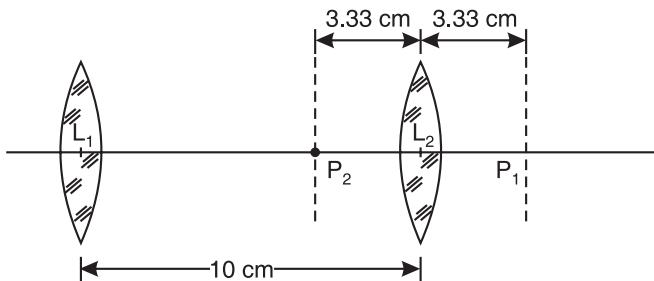


Fig. 40.23

Nodal Points:

As the system is situated in air, the nodal points are same as the principal points.

First Focal Point:

The distance of F_1 from the lens L_1 is

$$L_1 F_1 = -f \left[1 - \frac{d}{f_2} \right] = -6.67 \text{ cm} \left[1 - \frac{10 \text{ cm}}{5 \text{ cm}} \right] = 13.33 \text{ cm}$$

Second Focal Point:

The distance of F_2 from the lens L_2 is

$$L_2 F_2 = +f \left[1 - \frac{d}{f_1} \right] = -6.67 \text{ cm} \left[1 - \frac{10 \text{ cm}}{20 \text{ cm}} \right] = +3.33 \text{ cm}$$

The first principal point P_1 is to the right of the first lens and is at a distance of 13.33 cm from it. The second principal point P_2 is to the left of the second lens and is at a distance of 3.33 cm from it. The cardinal points are plotted in Fig. 40.23.

Example 40.3: Two thin converging lenses L_1 and L_2 of powers 5 D and 4 D are placed coaxially 10 cm apart. Find the focal length f of the combination and positions of principal points P_1 and P_2 .

Solution:

The focal length of the first lens, $f_1 = \frac{1}{P} = \frac{1}{5D} = \frac{1 \text{ m}}{5} = 20 \text{ cm}$

The focal length of the second lens, $f_2 = \frac{1}{P} = \frac{1}{4D} = \frac{1 \text{ m}}{4} = 25 \text{ cm}$

The equivalent focal length of the coaxial system is given by

$$f = \frac{f_1 f_2}{f_1 + f_2 - d}$$

$$\therefore f = \frac{20 \text{ cm} \times 25 \text{ cm}}{20 \text{ cm} + 25 \text{ cm} - 10 \text{ cm}} = 14.29 \text{ cm}$$

$$\alpha = +\frac{d f}{f_2} = \frac{10 \text{ cm} \times (14.29 \text{ cm})}{25 \text{ cm}} = 5.7 \text{ cm}$$

The first principal point P_1 is at a distance 5.7 cm to the right of the first lens.

$$\beta = -\frac{d f}{f_1} = -\frac{-10 \text{ cm} \times 14.29 \text{ cm}}{20 \text{ cm}} = -7.14 \text{ cm}$$

The second principal point P_2 is at a distance of 40 cm to the left of the second lens.

Example 40.4: Two thin convex lenses of focal lengths 30 cm and 10 cm are separated by a distance of 25 cm in air. Calculate the positions of the cardinal points.

Solution: Given that $f_1 = +30 \text{ cm}$, $f_2 = +10 \text{ cm}$ and $d = 25 \text{ cm}$

$$f = \frac{f_1 f_2}{f_1 + f_2 - d} = \frac{30 \text{ cm} \times 10 \text{ cm}}{30 \text{ cm} + 10 \text{ cm} - 25 \text{ cm}} = +20 \text{ cm}$$

First Principal Point:

$$\alpha = +\frac{fd}{f_2} = \frac{20 \text{ cm} \times 25 \text{ cm}}{10 \text{ cm}} = +50 \text{ cm}$$

Second Principal Point:

$$\beta = -\frac{fd}{f_1} = -\frac{20 \text{ cm} \times 25 \text{ cm}}{30 \text{ cm}} = -16.7 \text{ cm}$$

First Focal Point:

The distance of F_1 from the lens L_1 is

$$L_1 F_1 = -f \left[1 - \frac{d}{f_2} \right] = -20 \text{ cm} \left[1 - \frac{25 \text{ cm}}{10 \text{ cm}} \right] = +30 \text{ cm}$$

Second Focal Point:

The distance of F_2 from the lens L_2 is

$$L_2 F_2 = f \left[1 - \frac{d}{f_1} \right] = +20 \text{ cm} \left[1 - \frac{25 \text{ cm}}{30 \text{ cm}} \right] = +3.3 \text{ cm}$$

The first principal point P_1 is to the right of the first lens and is at a distance of 50 cm from it. The second principal point P_2 is to the left of the second lens and is at a distance of -16.7 cm from it. The first focal point F_1 is to the right of the first lens and is at a distance of 30 cm from it. The second focal point F_2 is to the right of the second lens and is at a distance of $+3.3 \text{ cm}$ from it.

Nodal Points:

As the system is situated in air, the nodal points are same as the principal points. The cardinal points are plotted in Fig. 40.24.

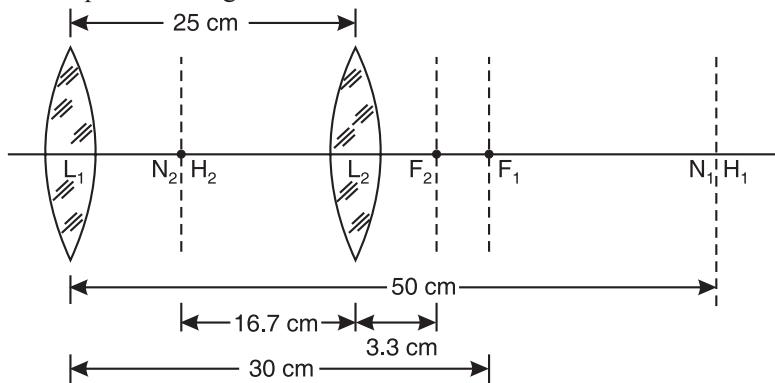


Fig. 40.24

Example 40.5: A convex lens of 10 cm focal length is placed in air from a concave lens of focal length 20 cm, at a distance of 5 cm. Find the distance between the two principal points of the combination.

Solution: Given that $f_1 = +10 \text{ cm}$, $f_2 = -20 \text{ cm}$ and $d = 5 \text{ cm}$

The equivalent focal length of the coaxial system is given by

$$f = \frac{f_1 f_2}{f_1 + f_2 - d}$$

$$\therefore f = \frac{10 \text{ cm} \times (-20 \text{ cm})}{10 \text{ cm} - 20 \text{ cm} - 5 \text{ cm}} = 13.3 \text{ cm}$$

$$\alpha = +\frac{d f}{f_2} = \frac{(5 \text{ cm}) \times (13.3 \text{ cm})}{-20 \text{ cm}} = -3.32 \text{ cm}$$

The first principal point P_1 is at a distance 3.32 cm to the left of the first lens.

$$\beta = -\frac{d f}{f_1} = -\frac{(5 \text{ cm}) \times (13.3 \text{ cm})}{10 \text{ cm}} = -6.65 \text{ cm}$$

The second principal point P_2 is at a distance of 6.65 cm to the left of the second lens.

The distance between the two principal points of the combination

$$= \beta - \alpha - d = [-6.65 - (-3.32) - (-5)] \text{ cm} = 1.67 \text{ cm.}$$

Example 40.6: The focal length of the more convergent lens of a Huygens' eye-piece is 1.5 cm. Calculate the focal length of the eye-piece and locate on a diagram the position of its focal points.

Solution: The focal length of the lenses in Huygens' eye-piece are $3f$ and f respectively and the separation between them is $2f$. The focal length of the more convergent lens is given to be 1.5 cm, i.e., $f_2 = 1.5 \text{ cm}$.

$$\therefore f_1 = 3f = 4.5 \text{ cm}, f_2 = 1.5 \text{ cm} \text{ and } d = 2f = 2 \times 1.5 \text{ cm} = 3 \text{ cm.}$$

Thus in this case the focal lengths of the field and eye-lenses are 4.5 cm and 1.5 cm and the separation between them is 3 cm.

The focal length F of the eye-piece is given by

$$F = \frac{f_1 f_2}{f_1 + f_2 - d} = \frac{4.5 \times 1.5}{4.5 + 1.5 - 3} \text{ cm} = 2.25 \text{ cm.}$$

Focal Points: The distance of the first focal point F_1 from the field lens L_1 is given by

$$\beta_1 = -F \left(1 - \frac{d}{f_2} \right) = -2.25 \left(1 - \frac{3}{1.5} \right) \text{ cm} = +2.25 \text{ cm.}$$

i.e., the first focal point F_1 lies at a distance 2.25 cm to the right of field lens L_1 .

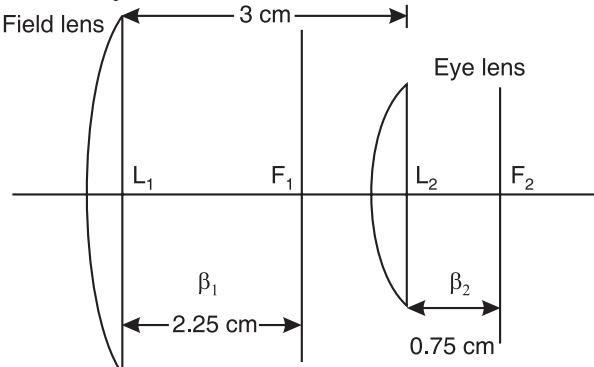


Fig. 40.25

The distance of the second focal point F_2 from the eye-lens L_2 is given by

$$\begin{aligned}\beta_2 &= F \left(1 - \frac{d}{f_1} \right) \\ &= 2.25 \left(1 - \frac{3}{4.5} \right) \text{ cm}\end{aligned}$$

$$= 0.75 \text{ cm.}$$

i.e., the second focal point F_2 lies at a distance 0.75 cm to the right of eye-lens L_2 .

The positions of the focal points F_1 and F_2 are plotted in fig. 40.25.

Example 40.7: From the conditions of 'no chromatic aberration' and 'minimum spherical aberration' of a combination of two separated thin plano-convex lenses, design a combination of equivalent focal length 6 cm. What type of eye-piece this combination forms? Deduce the positions of cardinal points of this eye-piece.

Solution: If f_1 and f_2 are the focal lengths of the lenses of the combination and d is the distance between the lenses, we have for no chromatic aberration.

$$d = \frac{f_1 + f_2}{2} \quad \dots(1)$$

and for minimum spherical aberration,

$$d = f_1 - f_2 \quad \dots(2)$$

From equations (1) and (2), we have

$$f_1 - f_2 = \frac{f_1 + f_2}{2}$$

or

$$f_1 = 3f_2 \quad \dots(3)$$

Substituting this value of f_1 in equation (2), we get

$$d = 3f_2 - f_2 = 2f_2 \quad \dots(4)$$

If F is the focal length of the equivalent lens we have

$$F = \frac{f_1 f_2}{f_1 + f_2 - d}$$

Here $f_1 = 3f_2$, $d = 2f_2$ and $F = 6 \text{ cm}$ (given)

$$\therefore 6 \text{ cm} = \frac{3f_2 \times f_2}{3f_2 + f_2 - 2f_2} = \frac{3}{2} f_2$$

$$\therefore f_2 = 4 \text{ cm}$$

From equation (3), $f_1 = 12 \text{ cm}$.

and from equation (4), $d = 8 \text{ cm}$.

Thus in the required combination, the focal lengths of the two plano-convex lenses are 12 cm and 4 cm respectively and the separation between them is 8 cm obviously this combination forms Huygens' eye-piece.

Cardinal Points

Principal Points: The position of the first principal point H_1 from the field lens L_1 is given by

$$\alpha_1 = \frac{dF}{f_2} = \frac{8 \times 6}{4} \text{ cm} = + 12 \text{ cm}$$

i.e., first principal point H_1 lies at a distance 12 cm to the right of field lens.

The distance of the second principal point H_2 from the eye-lens L_2

$$\alpha_2 = -\frac{dF}{f_1} = -\frac{8 \times 6}{12} \text{ cm} = -4 \text{ cm}$$

i.e., the second principal point H_2 lies at a distance of 4 cm to the left of eye-lens.

Focal Points: The distance of the first focal point F_1 from the field lens L_1 is

$$\beta_1 = -F \left(1 - \frac{d}{f_2}\right) = -6 \left(1 - \frac{8}{4}\right) \text{ cm} = +6 \text{ cm}$$

i.e., first focal point F_1 lie at a distance of 6 cm to the right of field lens.

The distance of the second focal point F_2 from the eye-lens L_2 is

$$\beta_2 = F \left(1 - \frac{d}{f_1}\right) = 6 \left(1 - \frac{8}{12}\right) \text{ cm} = +2 \text{ cm.}$$

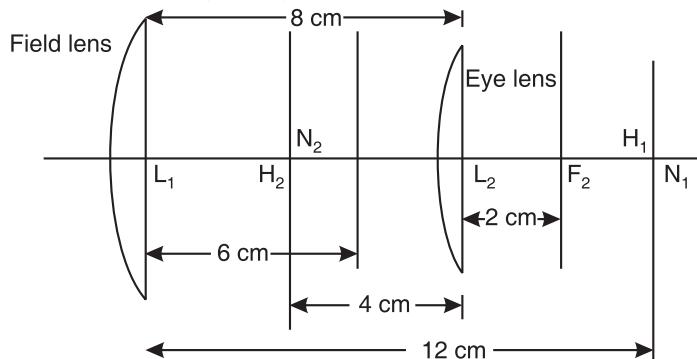


Fig. 40.26

i.e., the second focal point F_2 lies at a distance of 2 cm to the right of eye-lens L_2 .

Nodal Point: As the medium on either side of the combination is same, the nodal points N_1, N_2 coincide with the principal points H_1, H_2 respectively. Cardinal points H_1, H_2, F_1, F_2, N_1 and N_2 are shown in Fig. 40.26+.

Example 40.8: Focal lengths of lenses of Huygens' eye-piece are 4 cm and 12 cm. Find and plot the positions of cardinal points on a diagram. If an object is situated at 6 cm in front of the field lens, find the position of the image formed by the eye-piece.

Solution: The focal lengths of lenses are $f_1 = 12 \text{ cm}, f_2 = 4 \text{ cm}$.

In Huygens' eye-piece, separation between lenses

$$d = f_1 - f_2 = (12 - 4) \text{ cm} = 8 \text{ cm.}$$

The cardinal points of this eye-piece have been calculated in the previous example 40.7 and plotted in Fig. 40.26.

The object is at a distance 6 cm in front of the field lens. If u and v are the distances of the object O and image I from the first and second principal points H_1 and H_2 respectively.

Then $u = H_1 O = -(H_1 L_1 + L_1 O) = -(12 + 6) \text{ cm} = -18 \text{ cm}, v = H_2 I = ?$

$$F = \frac{f_1 f_2}{f_1 + f_2 - d} = \frac{12 \times 4}{12 + 4 - 8} \text{ cm} = \frac{12 \times 4}{8} \text{ cm} = +6 \text{ cm.}$$

Now from thin lens formula $\frac{1}{v} - \frac{1}{u} = \frac{1}{F}$, we have

$$\frac{1}{v} = \frac{1}{u} + \frac{1}{F} = -\frac{1}{18} + \frac{1}{6} = \frac{2}{18}$$

$$\therefore v = H_2 I = +9 \text{ cm.}$$

Hence $L_2 I = H_2 I - H_2 L_2 = (9 - 4) \text{ cm} = +5 \text{ cm.}$

Thus the image lies at a distance of +5 cm to the right (behind) the eye-lens L_2 .

$$= -\frac{3}{4}f + \frac{f}{2} = -\frac{f}{4}$$

i.e., the first focal point F_1 lies at a distance of $f/4$ to the left of field lens L_1 .

The distance of the second focal point F_2 from the eye-lens L_2 ,

$$\begin{aligned}\beta_2 &= F \left(1 - \frac{d}{f_1} \right) \\ &= \left(\frac{3}{4}f \right) \left(1 - \frac{\frac{2}{3}f}{f} \right) = \frac{3}{4}f - \frac{f}{2} = +\frac{f}{4}\end{aligned}$$

i.e., second point F_2 lies at a distance of $f/4$ to the right of eye-lens L_2 .

Nodal Points: As the medium on either side of the eye-piece is same (air), the nodal points N_1, N_2 coincide with the principal points H_1 and H_2 respectively.

Example 40.9: Light from the sun is falling upon a Ramsden's eye-piece. Locate the position of the image thus formed and also find the point from which the distance of the image is to be measured. The focal length of each lens of the eye-piece is $1\frac{1}{2}$ inches.

Solution: The focal lengths of the lenses of Ramsden's eye-piece are equal and the distance between them is $\frac{2}{3}f$.

If F is the focal length of the eye-piece,

$$F = \frac{f_1 f_2}{f_1 + f_2 - d}$$

$$\text{Here } f_1 = f_2 = f = \frac{3''}{2}, \quad d = \frac{2}{3}f = \frac{2}{3} \times \frac{3''}{2} = 1''$$

$$\therefore F = \frac{\frac{3''}{2} \times \frac{3''}{2}}{\frac{3''}{2} + \frac{3''}{2} - 1''} = +\frac{9''}{8}$$

As the rays coming from sun are parallel, the image will be formed by the eye-piece at second focal point F_2 , i.e., the image will be at a distance $\frac{9''}{8}$ from the second principal point H_2 .

The distance of second principal point H_2 from the second lens (eye-lens),

$$\alpha_2 = -\frac{dF}{f_1} = -\frac{1'' \times \frac{9''}{8}}{\frac{3''}{2}} = -\frac{9''}{8} \times \frac{2''}{3} = -\frac{3''}{4},$$

i.e., second principal point lies at a distance $\frac{3''}{4}$ to the left of eye-piece.

The distance of the image from the eye-lens, *i.e.*, the distance of the second focal point F_2 from the eye-lens,

$$\beta_2 = F \left(1 - \frac{d}{f_1} \right) = \frac{9''}{8} \left(1 - \frac{1''}{3/2''} \right) = \frac{9''}{8} \left(1 - \frac{2''}{3} \right) = \frac{3''}{4},$$

i.e., the image is formed at a distance of $\frac{3''}{4}$ to the right of eye-lens.

OBJECTIVE TYPE QUESTIONS

1. The axial points having unit positive lateral magnification in a lens system are called
 (a) principal points (b) focal points (c) nodal points (d) vertices
2. A pair of conjugate points on the axis of a lens having unit positive angular magnification are known as
 (a) focal points (b) nodal points (c) principal points (d) vertices
3. Cardinal points of a lens system consist of
 (a) focal points and principal points
 (b) principal points and nodal points
 (c) nodal points and focal points
 (d) focal points, principal points and nodal points
4. When the medium on the two sides of a lens system is same, the principal points coincide with
 (a) vertices (b) focal points (c) nodal points (d) none of these
5. A convex lens of focal length f_1 and a convex lens of focal length f_2 are placed a distance d apart. The focal length of the combination is
 (a) $\frac{f_1 f_2}{f_1 + f_2 - d}$ (b) $\frac{f_1 f_2}{f_1 + f_2 + d}$ (c) $\frac{f_1 f_2}{f_1 - f_2 + d}$ (d) $\frac{f_1 f_2}{d - f_1 + f_2}$
6. A convex lens of focal length f_1 and a concave lens of focal length f_2 are placed a distance d apart. The focal length of the combination is
 (a) $\frac{f_1 f_2}{f_1 + f_2 - d}$ (b) $\frac{f_1 f_2}{f_1 + f_2 + d}$ (c) $\frac{f_1 f_2}{f_1 - f_2 + d}$ (d) $\frac{f_1 f_2}{d - f_1 + f_2}$
7. Two thin convex lenses of focal lengths f_1 and f_2 are kept coaxially in air at a distance d apart. If the space between the lenses is filled by a liquid of refractive index μ , then the focal length of the combination will be
 (a) $\frac{\mu f_1 f_2}{f_1 + f_2 - d}$ (b) $\frac{f_1 f_2}{f_1 + f_2 - d/\mu}$ (c) $\frac{f_1 f_2}{f_1 + f_2 + d/\mu}$ (d) $\frac{f_1 f_2}{f_1 + f_2 + d}$
8. The position of first focal point of a coaxial system of two thin lenses separated by a distance d is
 (a) $-\frac{fd}{f_1}$ (b) $f \left(1 - \frac{d}{f_1} \right)$ (c) $-f \left(1 - \frac{d}{f_2} \right)$ (d) $\frac{fd}{f_2}$

SHORT ANSWER QUESTIONS

1. What are cardinal points?
2. Why the principal points of a lens system are called unit points?

DESCRIPTIVE QUESTIONS

1. Explain the term cardinal points with reference to a coaxial system.
2. What are principal points and principal planes? Show that the principal planes are the planes of unit linear magnification.
3. What are nodal points and nodal planes? Give their properties. Show that the nodal planes are planes of unit angular magnification.
4. Derive Newton's formula for a convergent system of lenses forming a real image.
5. Describe the construction and working of a nodal slide and show how the nodal points of a system can be located with its help.
6. Show that for a co-axial lens system, $x x' = f f'$ where x and x' are the respective distances of the object and the image from the first and the second focal points and f and f' are the two focal lengths. What form does the expression take when the media on the two sides of the system are the same?
7. Define cardinal points of a system of co-axial lenses. Describe how you would determine experimentally the principal planes of a combination of two thin lenses separated by a distance.
8. What are the properties of cardinal points of a co-axial lens system? Plot the cardinal points of a Huygens eyepiece. How can they be determined experimentally?
9. Explain the term cardinal points with reference to a co-axial lens system.
10. Two thin convex lenses of focal length f_1 cm and f_2 cm are coaxial and separated by d . Show that the focal length f of the combination is given by the relation

$$f = \frac{f_1 f_2}{f_1 + f_2 - d}$$

11. Define cardinal points of a coaxial lens system. Show how the principal planes of such a system can be located using the theory of the method of deviation or otherwise. State the sign convention you use.
12. Derive an expression for the focal length of a system of two thin lenses separated by a finite distance when the space between them is filled with a medium of refractive index μ ($\mu > 1$).
13. Derive expressions for the focal length and the positions of principal points and focal points of a coaxial system of two thin lenses separated by a finite distance.
14. What is an eyepiece and what is its advantage over a single lens?
15. Give the construction and working of a Ramsden eyepiece. How are chromatic and spherical aberrations minimized in this eyepiece?
16. Explain the construction of a Huygens eyepiece. Why cannot a cross-wire be used with it?
17. Give the construction of Huygens eyepiece and calculate the positions of the cardinal points.
18. Give the construction and theory of Huygens eyepiece and show that it is free from spherical and chromatic aberrations.
19. Explain why it is necessary to use an eyepiece consisting of more than one lens.
20. Why is Huygens eyepiece called a negative piece and Ramsden eyepiece called a positive eyepiece?
21. Give the name and construction of the eyepiece, which satisfies the condition for achromatism.
22. Describe and point out the respective merits of Ramsden and Huygens eyepieces.
23. Compare Ramsden eyepiece with Huygens eyepiece.
24. Explain why cross-wires can be used with Ramsden eyepiece but not with Huygens eyepiece.

PROBLEMS FOR PRACTICE

1. Two similar thin convex lenses of focal lengths 10 cm each are coaxial and 5 cm apart. Find the equivalent focal length and position of the principal points. Also find the position of the object for which the image is formed at infinity.

(Ans: $f = 6.67$ cm; $\alpha = +3.33$ cm; $\beta = -3.33$ cm; $u = -3.33$ cm.)

2. Two thin converging lenses of focal lengths 15 cm and 20 cm are placed coaxially 10 cm apart. An object is placed at a distance of 24 cm from the first lens. Find (i) the position of the focal points and principal points and (ii) the position of the image.

(Ans: $f = 12$ cm; $\alpha = +6$ cm; $\beta = -8$ cm; $v = 12$ cm.)

3. Two thin converging lenses each of 30 cm focal length are set coaxially 10 cm apart. An image of an upright pole 100 metres distant and 5 metre high is formed by the combination. Find the position of the unit and focal planes and the image. Also find the size of the image.

(Ans: $f = 18$ cm; $\alpha = +6$ cm; $\beta = -6$ cm; $v = 12.03$ cm, $h_2 = 0.9$ cm.)

4. Determine the positions of the focal points, principal points and nodal points in the case of sphere of radius 12 cm. ($\mu = 1.5$).

(Ans: $f = 18$ cm; $\alpha = +12$ cm; $\beta = -12$ cm. The nodal points and the principal points are at the centre of the sphere.)

5. Two thin converging lenses of focal lengths 20 cm and 40 cm are placed coaxially 20 cm apart. An object is placed at a distance of 50 cm from the first lens. Calculate the positions of the principal points, focal points and the position of the image.

(Ans: $\alpha = +12$ cm; $\beta = -16$ cm. $P_1F_1 = -24$ cm, $P_2F_2 = +24$ cm;

Image is formed at a distance of 23.16 cm to the right of the second lens.)

6. A Huygens' eye-piece has an eye-lens of focal length 4 cm. Sunlight is falling over the eye-piece. Find the position of the image in this condition and the distance of the second principal point from the eye-lens.

(Ans. Image is formed on the right of the eye-lens at a distance of 2 cm, $L_2H_2 = -4$ cm)

[Hint. $f_2 = f = 4$ cm, $f_1 = 3f = 12$ cm]

$$\therefore d = \frac{f_1 + f_2}{2} = \frac{4+12}{2} \text{ cm} = 8 \text{ cm}$$

$$F = \frac{f_1 f_2}{f_1 + f_2 - d} = \frac{12 \times 4}{12 + 4 - 8} \text{ cm} = 6 \text{ cm}]$$

7. The lenses in a Huygens' eye-piece have focal lengths of 2 cm and 4 cm. Find the distance between the lenses and located the cardinal points.

(Ans. $d = (f_1 + f_2)/2$ $L_1H_1 = L_1N_1 = 2$ cm; $L_2H_2 = L_2N_2 = -4$ cm; $L_1F_1 = -\frac{2}{3}$ cm;

$$L_2F_2 = -\frac{4}{3} \text{ cm})$$

8. The focal length of lenses of Huygens' eye-piece are 4 cm and 12 cm respectively. Find the positions of cardinal points. If an object is situated 6 cm in front of the field lens, find the positions of image formed by eye-piece.

(Ans. $\alpha_1 = L_1H_1 = +12$ cm, $\alpha_2 = L_2H_2 = -0.4$ cm, $\beta_1 = L_1F_1 = 6$ cm, $\beta_2 = L_2F_2 = +2$ cm,
Image forms 5 cm behind eye-lens)

9. The focal length of each lens of Ramsden's eye-piece is 4 cm. Calculate the focal length of eye-piece and locate the positions of cardinal points.

(Ans. $F = 3$ cm, $L_1H_1 = L_1N_1 = 2$ cm; $L_2H_2 = L_2N_2 = -2$ cm; $L_1F_1 = -1$ cm; $L_2F_2 = 1$ cm)

10. The equivalent focal length of Huygen's eye-piece is 6 cm. Locate the positions of principal and focal points. **(Ans.** $L_1H_1 = 12$ cm, $L_2H_2 = -4$ cm; $L_1H_1 = 6$ cm, $L_2F_2 = 2$ cm)
11. Deduce the composition of Ramsden's eye-piece of equivalent focal length 4.2 cm.
(Ans. Plano-convex lens each of focal length 5.6 cm placed 3.8 cm. apart.)
12. (a) A Ramsden's eye-piece is to be designed with the help of two plano-convex lenses each of focal length 5 cm. What should be the separation between the lenses? **(Ans.** 10 cm)
- (b) A Huygen's eye-piece is to be designed with the help of two plano-convex lenses of focal lengths 6 cm and 2 cm. What should be the separation between the lenses? **(Ans.** 4 cm)
- (c) If the equivalent focal length of a Huygen's eye-piece is 5 cm calculate the equivalent focal length of the field lens. **(Ans.** 10.0 cm)

ANSWERS TO OBJECTIVE TYPE QUESTIONS

- | | | | | | |
|--------|--------|--------|--------|--------|--------|
| 1. (a) | 2. (b) | 3. (d) | 4. (c) | 5. (a) | 6. (d) |
| 7. (a) | 8. (c) | | | | |