PHYSICS

PART – II

TEXTBOOK FOR CLASS XII

2018-19

PHYSICS

PART – II

TEXTBOOK FOR CLASS XII

राष्ट्रीय शैक्षिक अनुसंधान और प्रशिक्षण परिषद्
NATIONAL COUNCIL OF EDUCATIONAL RESEARCH AND TRAINING

2018-19

ISBN 81-7450-631-4 (Part I)
ISBN 81-7450-671-3 (Part II)

First Edition
March 2007 Chaitra 1928

Reprinted
December 2007 Agrahayana 1929
December 2008 Pausha 1930
December 2009 Pausha 1931
January 2011 Pausha 1932
January 2012 Magha 1933
December 2012 Agrahayana 1934
November 2013 Kartika 1935
December 2014 Pausha 1936
December 2015 Agrahayana 1937
February 2017 Magha 1938
December 2017 Pausha 1939
February 2018 Magha 1939

PD 500T HK

© National Council of Educational Research and Training, 2007

ALL RIGHTS RESERVED

- No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without the prior permission of the publisher.
- This book is sold subject to the condition that it shall not, by way of trade, be lent, re-sold, hired out or otherwise disposed of without the publisher's consent, in any form of binding or cover other than that in which it is published.
- The correct price of this publication is the price printed on this page, Any revised price indicated by a rubber stamp or by a sticker or by any other means is incorrect and should be unacceptable.

OFFICES OF THE PUBLICATION DIVISION, NCERT

NCERT Campus
Sri Aurobindo Marg
New Delhi 110 016 — **Phone : 011-26562708**

108, 100 Feet Road
Hosdakere Halli Extension
Banashankari III Stage
Bengaluru 560 085 — **Phone : 080-26725740**

Navjivan Trust Building
P.O.Navjivan
Ahmedabad 380 014 — **Phone : 079-27541446**

CWC Campus
Opp. Dhankal Bus Stop
Panihati
Kolkata 700 114 — **Phone : 033-25530454**

CWC Complex
Maligaon
Guwahati 781 021 — **Phone : 0361-2674869**

₹ 140.00

Printed on 80 GSM paper with NCERT watermark

Published at the Publication Division by the Secretary, National Council of Educational Research and Training, Sri Aurobindo Marg, New Delhi 110 016 and printed at Amar Ujala Publications Ltd., C-21-22, Sector-59, Noida - 201 301 (U.P.)

Publication Team

Head, Publication Division	:	*M. Siraj Anwar*
Chief Editor	:	*Shveta Uppal*
Chief Business Manager	:	*Gautam Ganguly*
Chief Production Officer (In-charge)	:	*Arun Chitkara*
Assistant Editor	:	*R. N. Bhardwaj*
Production Assistant	:	*Sunil Kumar*

Cover, Layout and Illustrations
Shweta Rao

FOREWORD

The National Curriculum Framework (NCF), 2005 recommends that children's life at school must be linked to their life outside the school. This principle marks a departure from the legacy of bookish learning which continues to shape our system and causes a gap between the school, home and community. The syllabi and textbooks developed on the basis of NCF signify an attempt to implement this basic idea. They also attempt to discourage rote learning and the maintenance of sharp boundaries between different subject areas. We hope these measures will take us significantly further in the direction of a child-centred system of education outlined in the National Policy on Education (NPE), 1986.

The success of this effort depends on the steps that school principals and teachers will take to encourage children to reflect on their own learning and to pursue imaginative activities and questions. We must recognise that, given space, time and freedom, children generate new knowledge by engaging with the information passed on to them by adults. Treating the prescribed textbook as the sole basis of examination is one of the key reasons why other resources and sites of learning are ignored. Inculcating creativity and initiative is possible if we perceive and treat children as participants in learning, not as receivers of a fixed body of knowledge.

These aims imply considerable change in school routines and mode of functioning. Flexibility in the daily time-table is as necessary as rigour in implementing the annual calendar so that the required number of teaching days are actually devoted to teaching. The methods used for teaching and evaluation will also determine how effective this textbook proves for making children's life at school a happy experience, rather than a source of stress or boredom. Syllabus designers have tried to address the problem of curricular burden by restructuring and reorienting knowledge at different stages with greater consideration for child psychology and the time available for teaching. The textbook attempts to enhance this endeavour by giving higher priority and space to opportunities for contemplation and wondering, discussion in small groups, and activities requiring hands-on experience.

The National Council of Educational Research and Training (NCERT) appreciates the hard work done by the textbook development committee responsible for this book. We wish to thank the Chairperson of the advisory group in science and mathematics, Professor J.V. Narlikar and the Chief Advisor for this book, Professor A.W. Joshi for guiding the work of this committee. Several teachers contributed to the development of this textbook; we are grateful to their principals for making this possible. We are indebted to the institutions and organisations which have generously permitted us to draw upon their resources, material and personnel. We are especially grateful to the members of the National Monitoring Committee, appointed by the Department of Secondary and Higher Education, Ministry of Human Resource Development under the Chairpersonship of Professor Mrinal Miri and Professor G.P. Deshpande, for their valuable time and contribution. As an organisation committed to systemic reform and continuous improvement in the quality of its products, NCERT welcomes comments and suggestions which will enable us to undertake further revision and refinement.

Director
National Council of Educational
Research and Training

New Delhi
20 November 2006

Preface

It gives me pleasure to place this book in the hands of the students, teachers and the public at large (whose role cannot be overlooked). It is a natural sequel to the Class XI textbook which was brought out in 2006. This book is also a trimmed version of the textbooks which existed so far. The chapter on thermal and chemical effects of current has been cut out. This topic has also been dropped from the CBSE syllabus. Similarly, the chapter on communications has been substantially curtailed. It has been rewritten in an easily comprehensible form.

Although most other chapters have been based on the earlier versions, several parts and sections in them have been rewritten. The Development Team has been guided by the feedback received from innumerable teachers across the country.

In producing these books, Class XI as well as Class XII, there has been a basic change of emphasis. Both the books present physics to students without assuming that they would pursue this subject beyond the higher secondary level. This new view has been prompted by the various observations and suggestions made in the National Curriculum Framework (NCF), 2005. Similarly, in today's educational scenario where students can opt for various combinations of subjects, we cannot assume that a physics student is also studying mathematics. Therefore, physics has to be presented, so to say, in a stand-alone form.

As in Class XI textbook, some interesting box items have been inserted in many chapters. They are not meant for teaching or examinations. Their purpose is to catch the attention of the reader, to show some applications in daily life or in other areas of science and technology, to suggest a simple experiment, to show connection of concepts in different areas of physics, and in general, to break the monotony and enliven the book.

Features like Summary, Points to Ponder, Exercises and Additional Exercises at the end of each chapter, and Examples have been retained. Several concept-based Exercises have been transferred from end-of-chapter Exercises to Examples with Solutions in the text. It is hoped that this will make the concepts discussed in the chapter more comprehensible. Several new examples and exercises have been added. Students wishing to pursue physics further would find Points to Ponder and Additional Exercises very useful and thoughtful. To provide *resources beyond the textbook* and to encourage *eLearning*, each chapter has been provided with some relevant website addresses under the title *ePhysics*. These sites provide additional materials on specific topics and also provide learners the opportunites for interactive demonstrations/ experiments.

The intricate concepts of physics must be understood, comprehended and appreciated. Students must learn to ask questions like 'why', 'how', 'how do we know it'. They will find almost always that the question 'why' has no answer within the domain of physics and science in general. But that itself is a learning experience, is it not? On the other hand, the question 'how' has been reasonably well answered by physicists in the case of most natural phenomena. In fact, with the understanding of how things happen, it has been possible to make use of many phenomena to create technological applications for the use of humans.

For example, consider statements in a book, like 'A negatively charged electron is attracted by the positively charged plate', or 'In this experiment, light (or electron) behaves like a wave'. You will realise that it is not possible to answer 'why'. This question belongs to the domain of philosophy or metaphysics. But we can answer 'how', we can find the force acting, we can find the wavelength of the photon (or electron), we can determine how things behave under different conditions, and we can develop instruments which will use these phenomena to our advantage.

It has been a pleasure to work for these books at the higher secondary level, along with a team of members. The Textbook Development Team, the Review Team and Editing Teams involved college and university teachers, teachers from Indian Institutes of Technology, scientists from national institutes and laboratories, as well as higher secondary teachers. The feedback and critical look provided by higher secondary teachers in the various teams are highly laudable. Most box items were generated by members of one or the other team, but three of them were generated by friends and well-wishers not part of any team. We are thankful to Dr P.N. Sen of Pune, Professor Roopmanjari Ghosh of Delhi and Dr Rajesh B Khaparde of Mumbai for allowing us to use their box items, respectively in Chapters 3, 4 (Part I) and 9 (Part II). We are very thankful to the members of the Review and Editing Workshops to discuss and refine the first draft of the textbook. We also express our gratitude to Prof. Krishna Kumar, Director, NCERT, for entrusting us with the task of presenting this textbook as a part of the national effort for improving science education. I also thank Prof. G. Ravindra, Joint Director, NCERT, for his help from time-to-time. Prof. Hukum Singh, Head, Department of Education in Science and Mathematics, NCERT, was always willing to help us in our endeavour in every possible way.

We welcome suggestions and comments from our valued users, especially students and teachers. We wish our young readers a happy journey into the exciting realm of physics.

A. W. Joshi
Chief Advisor
Textbook Development Committee

Textbook Development Committee

Chairperson, Advisory Committee for Textbooks in Science and Mathematics

J.V. Narlikar, *Emeritus Professor,* Inter-University Centre for Astronomy and Astrophysics (IUCAA), Ganeshkhind, Pune University Campus, Pune

Chief Advisor

A.W. Joshi, Honorary Visiting Scientist, National Centre for Radio Astrophysics (NCRA), Pune University Campus, Pune (Formerly *Professor* at Department of Physics, University of Pune)

Members

A.K. Ghatak, *Emeritus Professor*, Department of Physics, Indian Institute of Technology, New Delhi

Alika Khare, *Professor*, Department of Physics, Indian Institute of Technology, Guwahati

Anjali Kshirsagar, *Reader*, Department of Physics, University of Pune, Pune

Anuradha Mathur, *PGT* , Modern School, Vasant Vihar, New Delhi

Atul Mody, *Lecturer (S.G.),* VES College of Arts, Science and Commerce, Mumbai

B.K. Sharma, *Professor*, DESM, NCERT, New Delhi

Chitra Goel, *PGT,* Rajkiya Pratibha Vikas Vidyalaya, Tyagraj Nagar, New Delhi

Gagan Gupta, *Reader,* DESM, NCERT, New Delhi

H.C. Pradhan, *Professor,* Homi Bhabha Centre of Science Education (TIFR), Mumbai

N. Panchapakesan, *Professor* (*Retd.*), Department of Physics and Astrophysics, University of Delhi, Delhi

R. Joshi, *Lecturer* (S.G.), DESM, NCERT, New Delhi

S.K. Dash, *Reader,* DESM, NCERT, New Delhi

S. Rai Choudhary, *Professor,* Department of Physics and Astrophysics, University of Delhi, Delhi

S.K. Upadhyay, *PGT*, Jawahar Navodaya Vidyalaya, Muzaffar Nagar

S.N. Prabhakara, *PGT,* DM School, Regional Institute of Education (NCERT), Mysore

V.H. Raybagkar, *Reader*, Nowrosjee Wadia College, Pune

Vishwajeet Kulkarni, *Teacher* (*Grade I*), Higher Secondary Section, Smt. Parvatibai Chowgule College, Margao, Goa

Member-coordinator

V.P. Srivastava, *Reader,* DESM, NCERT, New Delhi

THE CONSTITUTION OF INDIA

PREAMBLE

WE, THE PEOPLE OF INDIA, having solemnly resolved to constitute India into a [1]**[SOVEREIGN SOCIALIST SECULAR DEMOCRATIC REPUBLIC]** and to secure to all its citizens :

JUSTICE, social, economic and political;

LIBERTY of thought, expression, belief, faith and worship;

EQUALITY of status and of opportunity; and to promote among them all

FRATERNITY assuring the dignity of the individual and the [2][unity and integrity of the Nation];

IN OUR CONSTITUENT ASSEMBLY this twenty-sixth day of November, 1949 do **HEREBY ADOPT, ENACT AND GIVE TO OURSELVES THIS CONSTITUTION.**

---

1. Subs. by the Constitution (Forty-second Amendment) Act, 1976, Sec.2, for "Sovereign Democratic Republic" (w.e.f. 3.1.1977)
2. Subs. by the Constitution (Forty-second Amendment) Act, 1976, Sec.2, for "Unity of the Nation" (w.e.f. 3.1.1977)

Acknowledgements

The National Council of Educational Research and Training acknowledges the valuable contribution of the individuals and organisations involved in the development of Physics Textbook for Class XII. The Council also acknowledges the valuable contribution of the following academics for reviewing and refining the manuscripts of this book:

Anu Venugopalan, *Lecturer*, School of Basic and Applied Sciences, GGSIP University, Delhi; A.K. Das, *PGT*, St. Xavier's Senior Secondary School, Delhi; Bharati Kukkal, *PGT*, Kendriya Vidyalaya, Pushp Vihar, New Delhi; D.A. Desai, *Lecturer (Retd.)*, Ruparel College, Mumbai; Devendra Kumar, *PGT*, Rajkiya Pratibha Vikas Vidyalaya, Yamuna Vihar, Delhi; I.K. Gogia, *PGT*, Kendriya Vidyalaya, Gole Market, New Delhi; K.C. Sharma, *Reader*, Regional Institute of Education (NCERT), Ajmer; M.K. Nandy, *Associate Professor*, Department of Physics, Indian Institute of Technology, Guwahati; M.N. Bapat, *Reader*, Regional Institute of Education (NCERT), Mysore; R. Bhattacharjee, *Asstt. Professor*, Department of Electronics and Communication Engineering, Indian Institute of Technology, Guwahati; R.S. Das, *Vice-Principal (Retd.)*, Balwant Ray Mehta Senior Secondary School, Lajpat Nagar, New Delhi; Sangeeta D. Gadre, *Reader*, Kirori Mal College, Delhi; Suresh Kumar, *PGT*, Delhi Public School, Dwarka, New Delhi; Sushma Jaireth, *Reader*, Department of Women's Studies, NCERT, New Delhi; Shyama Rath, *Reader*, Department of Physics and Astrophysics, University of Delhi, Delhi; Yashu Kumar, *PGT*, Kulachi Hans Raj Model School, Ashok Vihar, Delhi.

The Council also gratefully acknowledges the valuable contribution of the following academics for the editing and finalisation of this book: B.B. Tripathi, *Professor (Retd.)*, Department of Physics, Indian Institute of Technology, New Delhi; Dipan K. Ghosh, *Professor*, Department of Physics, Indian Institute of Technology, Mumbai; Dipanjan Mitra, *Scientist*, National Centre for Radio Astrophysics (TIFR), Pune; G.K. Mehta, *Raja Ramanna Fellow*, Inter-University Accelerator Centre, New Delhi; G.S. Visweswaran, *Professor*, Department of Electrical Engineering, Indian Institute of Technology, New Delhi; H.C. Kandpal, *Head*, Optical Radiation Standards, National Physical Laboratory, New Delhi; H.S. Mani, *Raja Ramanna Fellow*, Institute of Mathematical Sciences, Chennai; K. Thyagarajan, *Professor*, Department of Physics, Indian Institute of Technology, New Delhi; P.C. Vinod Kumar, *Professor*, Department of Physics, Sardar Patel University, Vallabh Vidyanagar, Gujarat; S. Annapoorni, *Professor*, Department of Physics and Astrophysics, University of Delhi, Delhi; S.C. Dutta Roy, *Emeritus Professor*, Department of Electrical Engineering, Indian Institute of Technology, New Delhi; S.D. Joglekar, *Professor*, Department of Physics, Indian Institute of Technology, Kanpur; V. Sundara Raja, *Professor*, Sri Venkateswara University, Tirupati.

The Council also acknowledges the valuable contributions of the following academics for refining the text in 2017: A.K. Srivastava, *Assistant Professor*, DESM, NCERT, New Delhi; Arnab Sen, *Assistant Professor*, NERIE, Shillong; L.S. Chauhan, *Assistant Professor*, RIE, Bhopal; O.N. Awasthi, *Professor* (*Retd.*), RIE, Bhopal; Rachna Garg, *Professor*, DESM, NCERT, New Delhi; Raman Namboodiri, *Assistant Professor*, RIE, Mysuru; R.R. Koireng, *Assistant Professor*, DCS, NCERT, New Delhi; Shashi Prabha, *Professor*, DESM, NCERT, New Delhi; and S.V. Sharma, *Professor*, RIE, Ajmer.

Special thanks are due to Hukum Singh, *Professor and Head*, DESM, NCERT for his support.

The Council also acknowledges the support provided by the APC office and the administrative staff of the DESM; Deepak Kapoor, *Incharge*, Computer Station; Inder Kumar, *DTP Operator*; Mohd. Qamar Tabrez and Hari Darshan Lodhi *Copy Editor*; Rishi Pal Singh, *Sr. Proof Reader*, NCERT and Ashima Srivastava, *Proof Reader* in shaping this book.

The contributions of the Publication Department in bringing out this book are also duly acknowledged.

Contents of Physics Part I
Class XII

CHAPTER ONE
Electric Charges and Fields 1

CHAPTER TWO
Electrostatic Potential and Capacitance 51

CHAPTER THREE
Current Electricity 93

CHAPTER FOUR
Moving Charges and Magnetism 132

CHAPTER FIVE
Magnetism and Matter 173

CHAPTER SIX
Electromagnetic Induction 204

CHAPTER SEVEN
Alternating Current 233

CHAPTER EIGHT
Electromagnetic Waves 269

ANSWERS 288

CONTENTS

Foreword *v*
Preface *vii*

CHAPTER NINE
Ray Optics and Optical Instruments

9.1 Introduction 309
9.2 Reflection of Light by Spherical Mirrors 310
9.3 Refraction 316
9.4 Total Internal Reflection 319
9.5 Refraction at Spherical Surfaces and by Lenses 323
9.6 Refraction through a Prism 330
9.7 Dispersion by a Prism 332
9.8 Some Natural Phenomena due to Sunlight 333
9.9 Optical Instruments 335

CHAPTER TEN
Wave Optics

10.1 Introduction 351
10.2 Huygens Principle 353
10.3 Refraction and Reflection of Plane Waves using Huygens Principle 355
10.4 Coherent and Incoherent Addition of Waves 360
10.5 Interference of Light Waves and Young's Experiment 362
10.6 Diffraction 367
10.7 Polarisation 376

CHAPTER ELEVEN
Dual Nature of Radiation and Matter

11.1 Introduction 386
11.2 Electron Emission 387
11.3 Photoelectric Effect 388
11.4 Experimental Study of Photoelectric Effect 389
11.5 Photoelectric Effect and Wave Theory of Light 393
11.6 Einstein's Photoelectric Equation: Energy Quantum of Radiation 393
11.7 Particle Nature of Light: The Photon 395
11.8 Wave Nature of Matter 398
11.9 Davisson and Germer Experiment 403

CHAPTER TWELVE
ATOMS

12.1 Introduction 414
12.2 Alpha-particle Scattering and Rutherford's Nuclear Model of Atom 415
12.3 Atomic Spectra 420
12.4 Bohr Model of the Hydrogen Atom 422
12.5 The Line Spectra of the Hydrogen Atom 428
12.6 DE Broglie's Explanation of Bohr's Second Postulate of Quantisation 430

CHAPTER THIRTEEN
NUCLEI

13.1 Introduction 438
13.2 Atomic Masses and Composition of Nucleus 438
13.3 Size of the Nucleus 441
13.4 Mass-Energy and Nuclear Binding Energy 442
13.5 Nuclear Force 445
13.6 Radioactivity 446
13.7 Nuclear Energy 451

CHAPTER FOURTEEN
SEMICONDUCTOR ELECTRONICS: MATERIALS, DEVICES AND SIMPLE CIRCUITS

14.1 Introduction 467
14.2 Classification of Metals, Conductors and Semiconductors 468
14.3 Intrinsic Semiconductor 472
14.4 Extrinsic Semiconductor 474
14.5 p-n Junction 478
14.6 Semiconductor Diode 479
14.7 Application of Junction Diode as a Rectifier 483
14.8 Special Purpose p-n Junction Diodes 485
14.9 Junction Transistor 490
14.10 Digital Electronics and Logic Gates 501
14.11 Integrated Circuits 505

CHAPTER FIFTEEN
COMMUNICATION SYSTEMS

15.1 Introduction 513
15.2 Elements of a Communication System 513
15.3 Basic Terminology Used in Electronic Communication Systems 515
15.4 Bandwidth of Signals 517
15.5 Bandwidth of Transmission Medium 518
15.6 Propagation of Electromagnetic Waves 519

15.7 Modulation and its Necessity 522
15.8 Amplitude Modulation 524
15.9 Production of Amplitude Modulated Wave 525
15.10 Detection of Amplitude Modulated Wave 526

APPENDICES 532

ANSWERS 534

BIBLIOGRAPHY 552

INDEX 554

2018-19

Cover design

(Adapted from http://nobelprize.org and the Nobel Prize in Physics 2006)

Different stages in the evolution of the universe.

Back Cover

(Adapted from http://www.iter.org and http://www.dae.gov.in)

Cut away view of *International Thermonuclear Experimental Reactor* (ITER) device. The man in the bottom shows the scale.

ITER is a joint international research and development project that aims to demonstrate the scientific and technical feasibility of fusion power.

India is one of the seven full partners in the project, the others being the European Union (represented by EURATOM), Japan, the People's Republic of China, the Republic of Korea, the Russian Federation and the USA. ITER will be constructed in Europe, at Cadarache in the South of France and will provide 500 MW of fusion power.

Fusion is the energy source of the sun and the stars. On earth, fusion research is aimed at demonstrating that this energy source can be used to produce electricity in a safe and environmentally benign way, with abundant fuel resources, to meet the needs of a growing world population.

For details of India's role, see *Nuclear India*, Vol. 39, No. 11-12/ May-June 2006, issue available at Department of Atomic Energy (DAE) website mentioned above.

Chapter Nine

RAY OPTICS AND OPTICAL INSTRUMENTS

9.1 Introduction

Nature has endowed the human eye (retina) with the sensitivity to detect electromagnetic waves within a small range of the electromagnetic spectrum. Electromagnetic radiation belonging to this region of the spectrum (wavelength of about 400 nm to 750 nm) is called light. It is mainly through light and the sense of vision that we know and interpret the world around us.

There are two things that we can intuitively mention about light from common experience. First, that it travels with enormous speed and second, that it travels in a straight line. It took some time for people to realise that the speed of light is finite and measurable. Its presently accepted value in vacuum is $c = 2.99792458 \times 10^8$ m s^{-1}. For many purposes, it suffices to take $c = 3 \times 10^8$ m s^{-1}. The speed of light in vacuum is the highest speed attainable in nature.

The intuitive notion that light travels in a straight line seems to contradict what we have learnt in Chapter 8, that light is an electromagnetic wave of wavelength belonging to the visible part of the spectrum. How to reconcile the two facts? The answer is that the wavelength of light is very small compared to the size of ordinary objects that we encounter commonly (generally of the order of a few cm or larger). In this situation, as you will learn in Chapter 10, a light wave can be considered to travel from one point to another, along a straight line joining

them. The path is called a *ray* of light, and a bundle of such rays constitutes a *beam* of light.

In this chapter, we consider the phenomena of reflection, refraction and dispersion of light, using the ray picture of light. Using the basic laws of reflection and refraction, we shall study the image formation by plane and spherical reflecting and refracting surfaces. We then go on to describe the construction and working of some important optical instruments, including the human eye.

PARTICLE MODEL OF LIGHT

Newton's fundamental contributions to mathematics, mechanics, and gravitation often blind us to his deep experimental and theoretical study of light. He made pioneering contributions in the field of optics. He further developed the corpuscular model of light proposed by Descartes. It presumes that light energy is concentrated in tiny particles called *corpuscles*. He further assumed that corpuscles of light were massless elastic particles. With his understanding of mechanics, he could come up with a simple model of reflection and refraction. It is a common observation that a ball bouncing from a smooth plane surface obeys the laws of reflection. When this is an elastic collision, the magnitude of the velocity remains the same. As the surface is smooth, there is no force acting parallel to the surface, so the component of momentum in this direction also remains the same. Only the component perpendicular to the surface, i.e., the normal component of the momentum, gets reversed in reflection. Newton argued that smooth surfaces like mirrors reflect the corpuscles in a similar manner.

In order to explain the phenomena of refraction, Newton postulated that the speed of the corpuscles was greater in water or glass than in air. However, later on it was discovered that the speed of light is less in water or glass than in air.

In the field of optics, Newton – the experimenter, was greater than Newton – the theorist. He himself observed many phenomena, which were difficult to understand in terms of particle nature of light. For example, the colours observed due to a thin film of oil on water. Property of partial reflection of light is yet another such example. Everyone who has looked into the water in a pond sees image of the face in it, but also sees the bottom of the pond. Newton argued that some of the corpuscles, which fall on the water, get reflected and some get transmitted. But what property could distinguish these two kinds of corpuscles? Newton had to postulate some kind of unpredictable, chance phenomenon, which decided whether an individual corpuscle would be reflected or not. In explaining other phenomena, however, the corpuscles were presumed to behave as if they are identical. Such a dilemma does not occur in the wave picture of light. An incoming wave can be divided into two weaker waves at the boundary between air and water.

9.2 REFLECTION OF LIGHT BY SPHERICAL MIRRORS

We are familiar with the laws of reflection. The angle of reflection (i.e., the angle between reflected ray and the normal to the reflecting surface or the mirror) equals the angle of incidence (angle between incident ray and the normal). Also that the incident ray, reflected ray and the normal to the reflecting surface at the point of incidence lie in the same plane (Fig. 9.1). These laws are valid at each point on any reflecting surface whether plane or curved. However, we shall restrict our discussion to the special case of curved surfaces, that is, spherical surfaces. The normal in

this case is to be taken as normal to the tangent to surface at the point of incidence. That is, the normal is along the radius, the line joining the centre of curvature of the mirror to the point of incidence.

We have already studied that the geometric centre of a spherical mirror is called its pole while that of a spherical lens is called its optical centre. The line joining the pole and the centre of curvature of the spherical mirror is known as the *principal axis*. In the case of spherical lenses, the principal axis is the line joining the optical centre with its principal focus as you will see later.

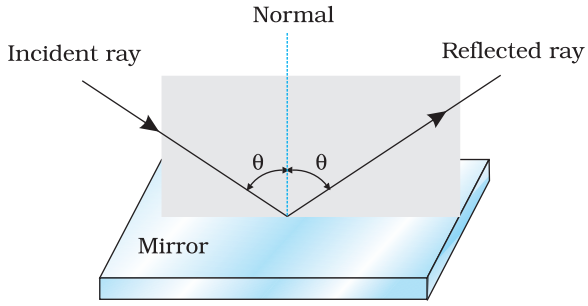


FIGURE 9.1 The incident ray, reflected ray and the normal to the reflecting surface lie in the same plane.

9.2.1 Sign convention

To derive the relevant formulae for reflection by spherical mirrors and refraction by spherical lenses, we must first adopt a sign convention for measuring distances. In this book, we shall follow the *Cartesian sign convention*. According to this convention, all distances are measured from the pole of the mirror or the optical centre of the lens. The distances measured in the same direction as the incident light are taken as positive and those measured in the direction opposite to the direction of incident light are taken as negative (Fig. 9.2). The heights measured upwards with respect to *x*-axis and normal to the principal axis (*x*-axis) of the mirror/lens are taken as positive (Fig. 9.2). The heights measured downwards are taken as negative.

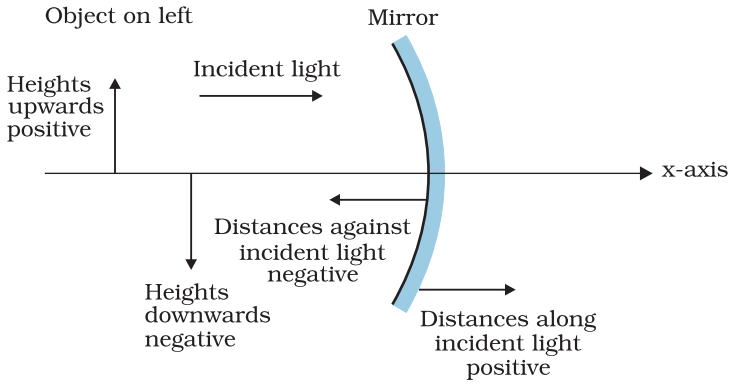


FIGURE 9.2 The Cartesian Sign Convention.

With a common accepted convention, it turns out that a single formula for spherical mirrors and a single formula for spherical lenses can handle all different cases.

9.2.2 Focal length of spherical mirrors

Figure 9.3 shows what happens when a parallel beam of light is incident on (a) a concave mirror, and (b) a convex mirror. We assume that the rays are *paraxial*, i.e., they are incident at points close to the pole P of the mirror and make small angles with the principal axis. The reflected rays converge at a point F on the principal axis of a concave mirror [Fig. 9.3(a)]. For a convex mirror, the reflected rays appear to diverge from a point F on its principal axis [Fig. 9.3(b)]. The point F is called the *principal focus* of the mirror. If the parallel paraxial beam of light were incident, making some angle with the principal axis, the reflected rays would converge (or appear to diverge) from a point in a plane through F normal to the principal axis. This is called the *focal plane* of the mirror [Fig. 9.3(c)].

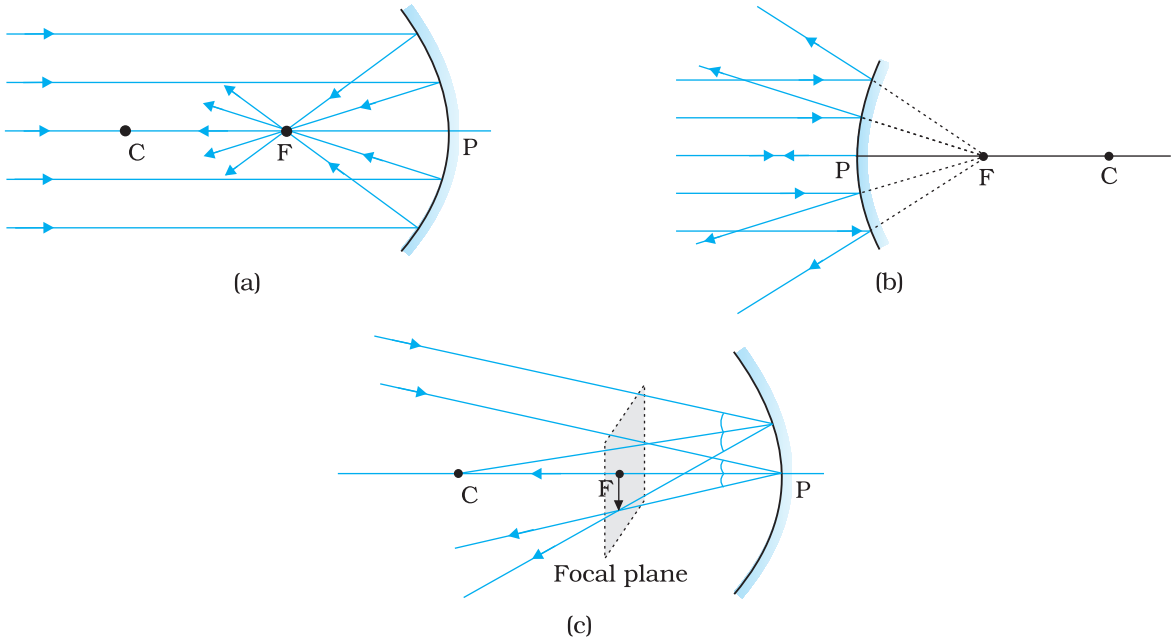


FIGURE 9.3 Focus of a concave and convex mirror.

The distance between the focus F and the pole P of the mirror is called the *focal length* of the mirror, denoted by f. We now show that $f = R/2$, where R is the radius of curvature of the mirror. The geometry of reflection of an incident ray is shown in Fig. 9.4.

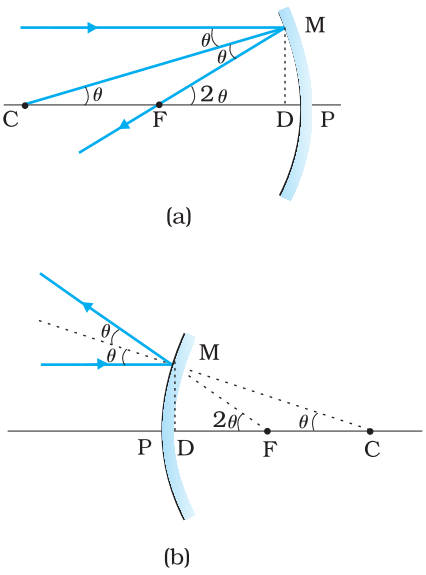


FIGURE 9.4 Geometry of reflection of an incident ray on (a) concave spherical mirror, and (b) convex spherical mirror.

Let C be the centre of curvature of the mirror. Consider a ray parallel to the principal axis striking the mirror at M. Then CM will be perpendicular to the mirror at M. Let θ be the angle of incidence, and MD be the perpendicular from M on the principal axis. Then,

$\angle MCP = \theta$ and $\angle MFP = 2\theta$

Now,

$$\tan\theta = \frac{MD}{CD} \text{ and } \tan 2\theta = \frac{MD}{FD} \qquad (9.1)$$

For small θ, which is true for paraxial rays, $\tan\theta \approx \theta$, $\tan 2\theta \approx 2\theta$. Therefore, Eq. (9.1) gives

$$\frac{MD}{FD} = 2\,\frac{MD}{CD}$$

$$\text{or, } FD = \frac{CD}{2} \qquad (9.2)$$

Now, for small θ, the point D is very close to the point P. Therefore, FD = f and CD = R. Equation (9.2) then gives

$$f = R/2 \qquad (9.3)$$

9.2.3 The mirror equation

If rays emanating from a point actually meet at another point after reflection and/or refraction, that point is called the *image* of the first point. The image is *real* if the rays actually converge to the point; it is

virtual if the rays do not actually meet but appear to diverge from the point when produced backwards. An image is thus a point-to-point correspondence with the object established through reflection and/or refraction.

In principle, we can take any two rays emanating from a point on an object, trace their paths, find their point of intersection and thus, obtain the image of the point due to reflection at a spherical mirror. In practice, however, it is convenient to choose any two of the following rays:

(i) The ray from the point which is parallel to the principal axis. The reflected ray goes through the focus of the mirror.

(ii) The ray passing through the centre of curvature of a concave mirror or appearing to pass through it for a convex mirror. The reflected ray simply retraces the path.

(iii) The ray passing through (or directed towards) the focus of the concave mirror or appearing to pass through (or directed towards) the focus of a convex mirror. The reflected ray is parallel to the principal axis.

(iv) The ray incident at any angle at the pole. The reflected ray follows laws of reflection.

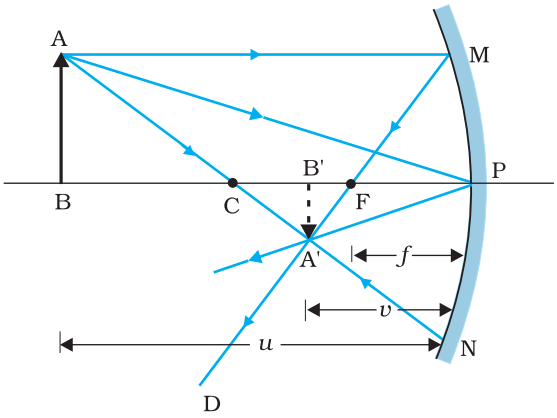


FIGURE 9.5 Ray diagram for image formation by a concave mirror.

Figure 9.5 shows the ray diagram considering three rays. It shows the image A′B′ (in this case, real) of an object AB formed by a concave mirror. It does not mean that only three rays emanate from the point A. An infinite number of rays emanate from any source, in all directions. Thus, point A′ is image point of A if every ray originating at point A and falling on the concave mirror after reflection passes through the point A′.

We now derive the mirror equation or the relation between the object distance (u), image distance (v) and the focal length (f).

From Fig. 9.5, the two right-angled triangles A′B′F and MPF are similar. (For paraxial rays, MP can be considered to be a straight line perpendicular to CP.) Therefore,

$$\frac{B'A'}{PM} = \frac{B'F}{FP}$$

$$\text{or } \frac{B'A'}{BA} = \frac{B'F}{FP} \quad (\because PM = AB) \tag{9.4}$$

Since ∠ APB = ∠ A′PB′, the right angled triangles A′B′P and ABP are also similar. Therefore,

$$\frac{B'A'}{BA} = \frac{B'P}{BP} \tag{9.5}$$

Comparing Eqs. (9.4) and (9.5), we get

$$\frac{B'F}{FP} = \frac{B'P - FP}{FP} = \frac{B'P}{BP} \tag{9.6}$$

Equation (9.6) is a relation involving magnitude of distances. We now apply the sign convention. We note that light travels from the object to the mirror MPN. Hence this is taken as the positive direction. To reach

the object AB, image A′B′ as well as the focus F from the pole P, we have to travel opposite to the direction of incident light. Hence, all the three will have negative signs. Thus,

$$B'P = -v,\ FP = -f,\ BP = -u$$

Using these in Eq. (9.6), we get

$$\frac{-v+f}{-f} = \frac{-v}{-u}$$

or $$\frac{v-f}{f} = \frac{v}{u}$$

$$\frac{v}{f} = 1 + \frac{v}{u}$$

Dividing it by v, we get

$$\frac{1}{v} + \frac{1}{u} = \frac{1}{f} \qquad (9.7)$$

This relation is known as the *mirror equation*.

The size of the image relative to the size of the object is another important quantity to consider. We define linear *magnification* (m) as the ratio of the height of the image (h') to the height of the object (h):

$$m = \frac{h'}{h} \qquad (9.8)$$

h and h' will be taken positive or negative in accordance with the accepted sign convention. In triangles A′B′P and ABP, we have,

$$\frac{B'A'}{BA} = \frac{B'P}{BP}$$

With the sign convention, this becomes

$$\frac{-h'}{h} = \frac{-v}{-u}$$

so that

$$m = \frac{h'}{h} = -\frac{v}{u} \qquad (9.9)$$

We have derived here the mirror equation, Eq. (9.7), and the magnification formula, Eq. (9.9), for the case of real, inverted image formed by a concave mirror. With the proper use of sign convention, these are, in fact, valid for all the cases of reflection by a spherical mirror (concave or convex) whether the image formed is real or virtual. Figure 9.6 shows the ray diagrams for virtual image formed by a concave and convex mirror. You should verify that Eqs. (9.7) and (9.9) are valid for these cases as well.

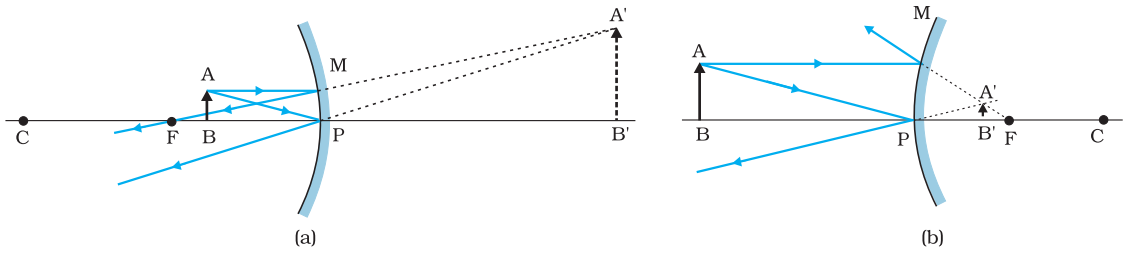


FIGURE 9.6 Image formation by (a) a concave mirror with object between P and F, and (b) a convex mirror.

Example 9.1 Suppose that the lower half of the concave mirror's reflecting surface in Fig. 9.5 is covered with an opaque (non-reflective) material. What effect will this have on the image of an object placed in front of the mirror?

Solution You may think that the image will now show only half of the object, but taking the laws of reflection to be true for all points of the remaining part of the mirror, the image will be that of the whole object. However, as the area of the reflecting surface has been reduced, the intensity of the image will be low (in this case, half).

EXAMPLE 9.1

Example 9.2 A mobile phone lies along the principal axis of a concave mirror, as shown in Fig. 9.7. Show by suitable diagram, the formation of its image. Explain why the magnification is not uniform. Will the distortion of image depend on the location of the phone with respect to the mirror?

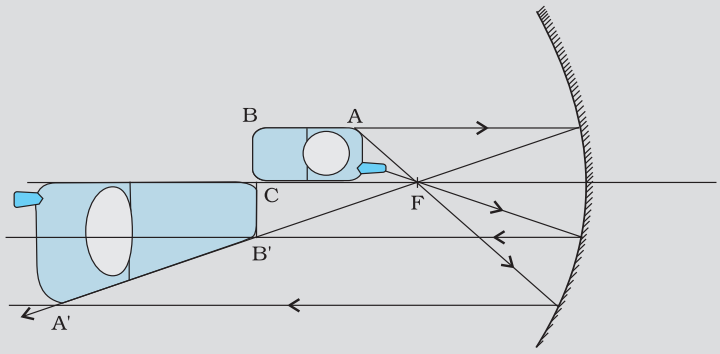


FIGURE 9.7

Solution

The ray diagram for the formation of the image of the phone is shown in Fig. 9.7. The image of the part which is on the plane perpendicular to principal axis will be on the same plane. It will be of the same size, i.e., B′C = BC. You can yourself realise why the image is distorted.

EXAMPLE 9.2

Example 9.3 An object is placed at (i) 10 cm, (ii) 5 cm in front of a concave mirror of radius of curvature 15 cm. Find the position, nature, and magnification of the image in each case.

Solution

The focal length $f = -15/2$ cm $= -7.5$ cm

(i) The object distance $u = -10$ cm. Then Eq. (9.7) gives

$$\frac{1}{v}+\frac{1}{-10}=\frac{1}{-7.5}$$

or $$v=\frac{10\times 7.5}{-2.5}=-30 \text{ cm}$$

The image is 30 cm from the mirror on the same side as the object.

Also, magnification $m = -\frac{v}{u}=-\frac{(-30)}{(-10)}=-3$

The image is magnified, real and inverted.

EXAMPLE 9.3

EXAMPLE 9.3

(ii) The object distance u = –5 cm. Then from Eq. (9.7),

$$\frac{1}{v}+\frac{1}{-5}=\frac{1}{-7.5}$$

or $$v=\frac{5\times 7.5}{(7.5-5)}=15 \text{ cm}$$

This image is formed at 15 cm behind the mirror. It is a virtual image.

Magnification $m = -\frac{v}{u}=-\frac{15}{(-5)}=3$

The image is magnified, virtual and erect.

EXAMPLE 9.4

Example 9.4 Suppose while sitting in a parked car, you notice a jogger approaching towards you in the side view mirror of R = 2 m. If the jogger is running at a speed of 5 m s^{-1}, how fast the image of the jogger appear to move when the jogger is (a) 39 m, (b) 29 m, (c) 19 m, and (d) 9 m away.

Solution

From the mirror equation, Eq. (9.7), we get

$$v=\frac{fu}{u-f}$$

For convex mirror, since R = 2 m, f = 1 m. Then

for u = –39 m, $v = \frac{(-39)\times 1}{-39-1} = \frac{39}{40}$ m

Since the jogger moves at a constant speed of 5 m s^{-1}, after 1 s the position of the image v (for u = –39 + 5 = –34) is (34/35)m.

The shift in the position of image in 1 s is

$$\frac{39}{40}-\frac{34}{35}=\frac{1365-1360}{1400}=\frac{5}{1400}=\frac{1}{280}\text{ m}$$

Therefore, the average speed of the image when the jogger is between 39 m and 34 m from the mirror, is (1/280) m s^{-1}

Similarly, it can be seen that for u = –29 m, –19 m and –9 m, the speed with which the image appears to move is

$\frac{1}{150}$ m s^{-1}, $\frac{1}{60}$ m s^{-1} and $\frac{1}{10}$ m s^{-1}, respectively.

Although the jogger has been moving with a constant speed, the speed of his/her image appears to increase substantially as he/she moves closer to the mirror. This phenomenon can be noticed by any person sitting in a stationary car or a bus. In case of moving vehicles, a similar phenomenon could be observed if the vehicle in the rear is moving closer with a constant speed.

9.3 REFRACTION

When a beam of light encounters another transparent medium, a part of light gets reflected back into the first medium while the rest enters the other. A ray of light represents a beam. The direction of propagation of an obliquely incident (0°< i < 90°) ray of light that enters the other medium,

changes at the interface of the two media. This phenomenon is called *refraction of light*. Snell experimentally obtained the following laws of refraction:

(i) The incident ray, the refracted ray and the normal to the interface at the point of incidence, all lie in the same plane.

(ii) The ratio of the sine of the angle of incidence to the sine of angle of refraction is constant. Remember that the angles of incidence (i) and refraction (r) are the angles that the incident and its refracted ray make with the normal, respectively. We have

$$\frac{\sin i}{\sin r} = n_{21} \qquad (9.10)$$

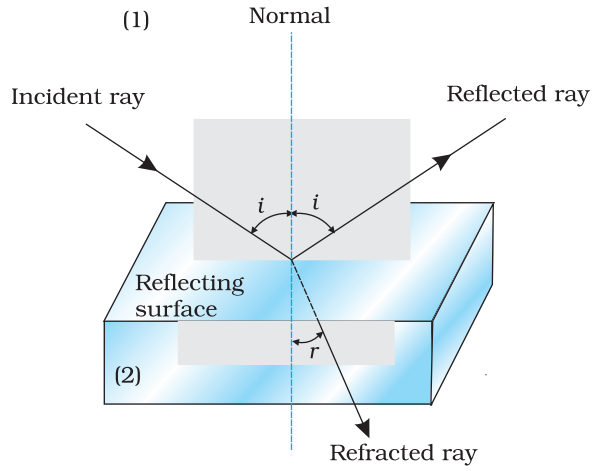


FIGURE 9.8 Refraction and reflection of light.

where n_{21} is a constant, called the *refractive index* of the second medium with respect to the first medium. Equation (9.10) is the well-known Snell's law of refraction. We note that n_{21} is a characteristic of the pair of media (and also depends on the wavelength of light), but is independent of the angle of incidence.

From Eq. (9.10), if $n_{21} > 1$, $r < i$, i.e., the refracted ray bends towards the normal. In such a case medium 2 is said to be *optically denser* (or *denser*, in short) than medium 1. On the other hand, if $n_{21} < 1$, $r > i$, the refracted ray bends away from the normal. This is the case when incident ray in a denser medium refracts into a rarer medium.

Note: *Optical density should not be confused with mass density, which is mass per unit volume. It is possible that mass density of an optically denser medium may be less than that of an optically rarer medium (optical density is the ratio of the speed of light in two media). For example, turpentine and water. Mass density of turpentine is less than that of water but its optical density is higher.*

If n_{21} is the refractive index of medium 2 with respect to medium 1 and n_{12} the refractive index of medium 1 with respect to medium 2, then it should be clear that

$$n_{12} = \frac{1}{n_{21}} \qquad (9.11)$$

It also follows that if n_{32} is the refractive index of medium 3 with respect to medium 2 then $n_{32} = n_{31} \times n_{12}$, where n_{31} is the refractive index of medium 3 with respect to medium 1.

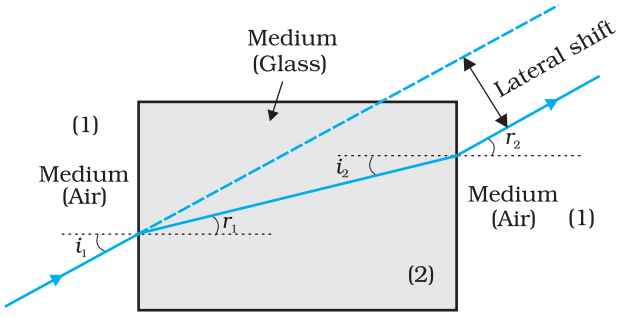


FIGURE 9.9 Lateral shift of a ray refracted through a parallel-sided slab.

Some elementary results based on the laws of refraction follow immediately. For a rectangular slab, refraction takes place at two interfaces (air-glass and glass-air). It is easily seen from Fig. 9.9 that $r_2 = i_1$, i.e., the emergent ray is parallel to the incident ray—there is no

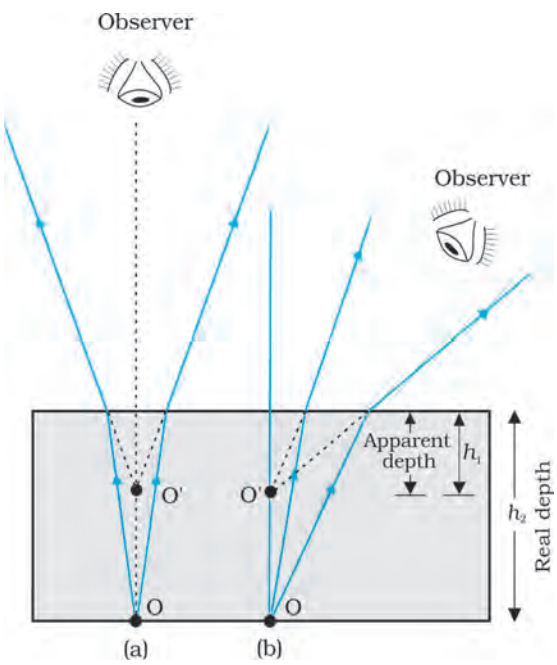


FIGURE 9.10 Apparent depth for (a) normal, and (b) oblique viewing.

deviation, but it does suffer lateral displacement/shift with respect to the incident ray. Another familiar observation is that the bottom of a tank filled with water appears to be raised (Fig. 9.10). For viewing near the normal direction, it can be shown that the apparent depth (h_1) is real depth (h_2) divided by the refractive index of the medium (water).

The refraction of light through the atmosphere is responsible for many interesting phenomena. For example, the Sun is visible a little before the actual sunrise and until a little after the actual sunset due to refraction of light through the atmosphere (Fig. 9.11). By actual sunrise we mean the actual crossing of the horizon by the sun. Figure 9.11 shows the actual and apparent positions of the Sun with respect to the horizon. The figure is highly exaggerated to show the effect. The refractive index of air with respect to vacuum is 1.00029. Due to this, the apparent shift in the direction of the Sun is by about half a degree and the corresponding time difference between actual sunset and apparent sunset is about 2 minutes (see Example 9.5). The apparent flattening (oval shape) of the Sun at sunset and sunrise is also due to the same phenomenon.

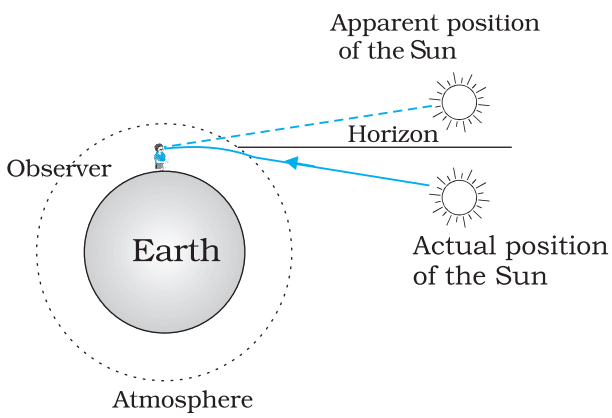


FIGURE 9.11 Advance sunrise and delayed sunset due to atmospheric refraction.

EXAMPLE 9.5

Example 9.5 The earth takes 24 h to rotate once about its axis. How much time does the sun take to shift by 1° when viewed from the earth?

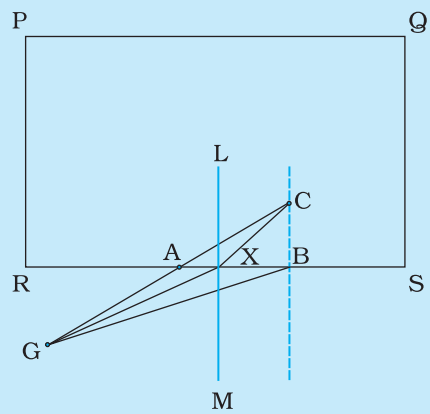
Solution

Time taken for 360° shift = 24 h

Time taken for 1° shift = 24/360 h = 4 min.

THE DROWNING CHILD, LIFEGUARD AND SNELL'S LAW

Consider a rectangular swimming pool PQSR; see figure here. A lifeguard sitting at G outside the pool notices a child drowning at a point C. The guard wants to reach the child in the shortest possible time. Let SR be the side of the pool between G and C. Should he/she take a straight line path GAC between G and C or GBC in which the path BC in water would be the shortest, or some other path GXC? The guard knows that his/her running speed v_1 on ground is higher than his/her swimming speed v_2.



Suppose the guard enters water at X. Let GX $= l_1$ and XC $= l_2$. Then the time taken to reach from G to C would be

$$t = \frac{l_1}{v_1} + \frac{l_2}{v_2}$$

To make this time minimum, one has to differentiate it (with respect to the coordinate of X) and find the point X when t is a minimum. On doing all this algebra (which we skip here), we find that the guard should enter water at a point where Snell's law is satisfied. To understand this, draw a perpendicular LM to side SR at X. Let $\angle GXM = i$ and $\angle CXL = r$. Then it can be seen that t is minimum when

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2}$$

In the case of light v_1/v_2, the ratio of the velocity of light in vacuum to that in the medium, is the refractive index n of the medium.

In short, whether it is a wave or a particle or a human being, whenever two mediums and two velocities are involved, one must follow Snell's law if one wants to take the shortest time.

9.4 TOTAL INTERNAL REFLECTION

When light travels from an optically denser medium to a rarer medium at the interface, it is partly reflected back into the same medium and partly refracted to the second medium. This reflection is called the *internal reflection.*

When a ray of light enters from a denser medium to a rarer medium, it bends away from the normal, for example, the ray AO_1B in Fig. 9.12. The incident ray AO_1 is partially reflected (O_1C) and partially transmitted (O_1B) or refracted, the angle of refraction (r) being larger than the angle of incidence (i). As the angle of incidence increases, so does the angle of refraction, till for the ray AO_3, the angle of refraction is $\pi/2$. The refracted ray is bent so much away from the normal that it grazes the surface at the interface between the two media. This is shown by the ray AO_3D in Fig. 9.12. If the angle of incidence is increased still further (e.g., the ray AO_4), refraction is not possible, and the incident ray is totally reflected.

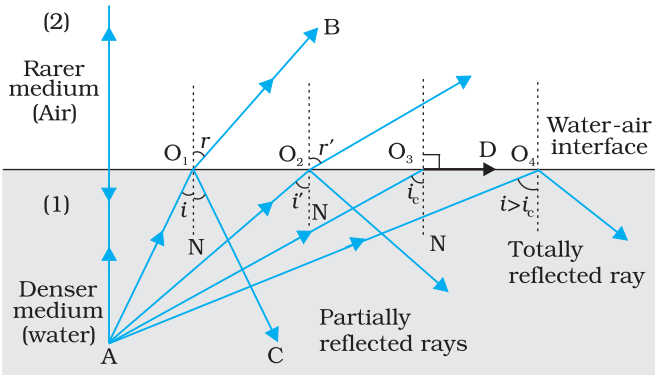


FIGURE 9.12 Refraction and internal reflection of rays from a point A in the denser medium (water) incident at different angles at the interface with a rarer medium (air).

This is called *total internal reflection*. When light gets reflected by a surface, normally some fraction of it gets transmitted. The reflected ray, therefore, is always less intense than the incident ray, howsoever smooth the reflecting surface may be. In total internal reflection, on the other hand, no transmission of light takes place.

The angle of incidence corresponding to an angle of refraction 90°, say $\angle AO_3N$, is called the *critical angle* (i_c) for the given pair of media. We see from Snell's law [Eq. (9.10)] that if the relative refractive index is less than one then, since the maximum value of sin r is unity, there is an upper limit to the value of sin i for which the law can be satisfied, that is, $i = i_c$ such that

$$\sin i_c = n_{21} \tag{9.12}$$

For values of i larger than i_c, Snell's law of refraction cannot be satisfied, and hence no refraction is possible.

The refractive index of denser medium 1 with respect to rarer medium 2 will be $n_{12} = 1/\sin i_c$. Some typical critical angles are listed in Table 9.1.

TABLE 9.1 CRITICAL ANGLE OF SOME TRANSPARENT MEDIA WITH RESPECT TO AIR

Substance medium	Refractive index	Critical angle
Water	1.33	48.75
Crown glass	1.52	41.14
Dense flint glass	1.62	37.31
Diamond	2.42	24.41

A demonstration for total internal reflection

All optical phenomena can be demonstrated very easily with the use of a laser torch or pointer, which is easily available nowadays. Take a glass beaker with clear water in it. Add a few drops of milk or any other suspension to water and stir so that water becomes a little turbid. Take a laser pointer and shine its beam through the turbid water. You will find that the path of the beam inside the water shines brightly.

Shine the beam from below the beaker such that it strikes at the upper water surface at the other end. Do you find that it undergoes partial reflection (which is seen as a spot on the table below) and partial refraction [which comes out in the air and is seen as a spot on the roof; Fig. 9.13(a)]? Now direct the laser beam from one side of the beaker such that it strikes the upper surface of water more obliquely [Fig. 9.13(b)]. Adjust the direction of laser beam until you find the angle for which the refraction

above the water surface is totally absent and the beam is totally reflected back to water. This is total internal reflection at its simplest.

Pour this water in a long test tube and shine the laser light from top, as shown in Fig. 9.13(c). Adjust the direction of the laser beam such that it is totally internally reflected every time it strikes the walls of the tube. This is similar to what happens in optical fibres.

Take care not to look into the laser beam directly and not to point it at anybody's face.

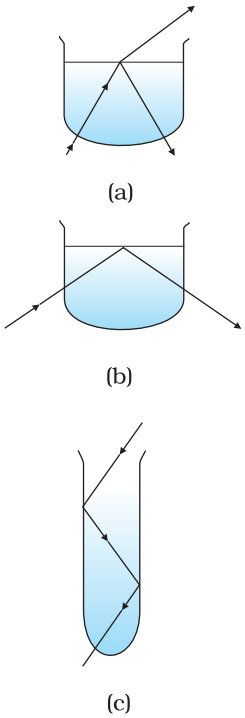


FIGURE 9.13 Observing total internal reflection in water with a laser beam (refraction due to glass of beaker neglected being very thin).

9.4.1 Total internal reflection in nature and its technological applications

(i) *Mirage:* On hot summer days, the air near the ground becomes hotter than the air at higher levels. The refractive index of air increases with its density. Hotter air is less dense, and has smaller refractive index than the cooler air. If the air currents are small, that is, the air is still, the optical density at different layers of air increases with height. As a result, light from a tall object such as a tree, passes through a medium whose refractive index decreases towards the ground. Thus, a ray of light from such an object successively bends away from the normal and undergoes total internal reflection, if the angle of incidence for the air near the ground exceeds the critical angle. This is shown in Fig. 9.14(b). To a distant observer, the light appears to be coming from somewhere below the ground. The observer naturally assumes that light is being reflected from the ground, say, by a pool of water near the tall object. Such inverted images of distant tall objects cause an optical illusion to the observer. This phenomenon is called *mirage*. This type of mirage is especially common in hot deserts. Some of you might have noticed that while moving in a bus or a car during a hot summer day, a distant patch of road, especially on a highway, appears to be wet. But, you do not find any evidence of wetness when you reach that spot. This is also due to mirage.

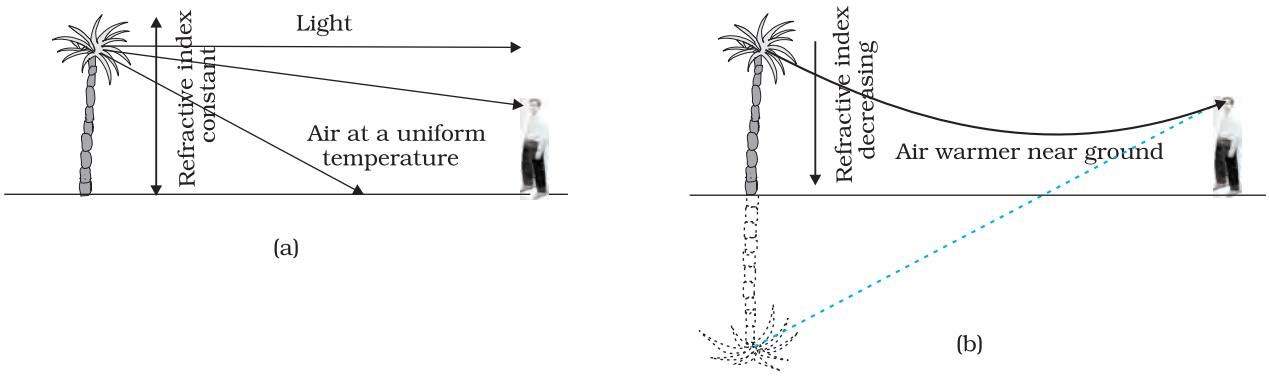


FIGURE 9.14 (a) A tree is seen by an observer at its place when the air above the ground is at uniform temperature, (b) When the layers of air close to the ground have varying temperature with hottest layers near the ground, light from a distant tree may undergo total internal reflection, and the apparent image of the tree may create an illusion to the observer that the tree is near a pool of water.

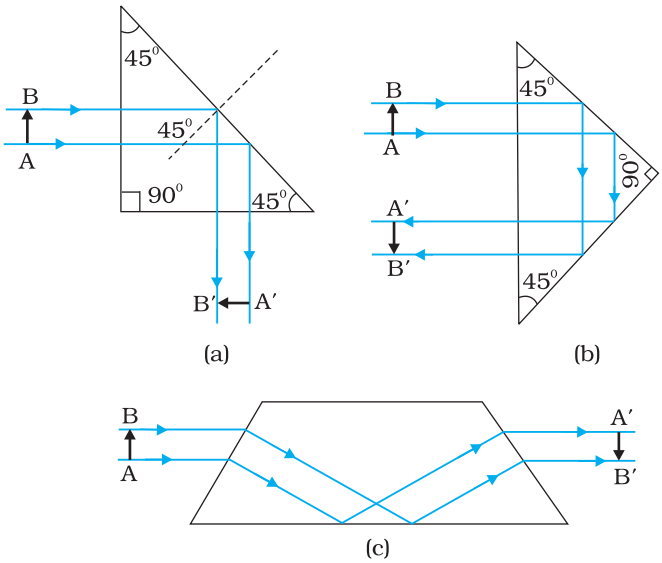


FIGURE 9.15 Prisms designed to bend rays by 90° and 180° or to invert image without changing its size make use of total internal reflection.

(ii) *Diamond*: Diamonds are known for their spectacular brilliance. Their brilliance is mainly due to the total internal reflection of light inside them. The critical angle for diamond-air interface (≅ 24.4°) is very small, therefore once light enters a diamond, it is very likely to undergo total internal reflection inside it. Diamonds found in nature rarely exhibit the brilliance for which they are known. It is the technical skill of a diamond cutter which makes diamonds to sparkle so brilliantly. By cutting the diamond suitably, multiple total internal reflections can be made to occur.

(iii) *Prism*: Prisms designed to bend light by 90° or by 180° make use of total internal reflection [Fig. 9.15(a) and (b)]. Such a prism is also used to invert images without changing their size [Fig. 9.15(c)]. In the first two cases, the critical angle i_c for the material of the prism must be less than 45°. We see from Table 9.1 that this is true for both crown glass and dense flint glass.

(iv) *Optical fibres*: Nowadays optical fibres are extensively used for transmitting audio and video signals through long distances. Optical fibres too make use of the phenomenon of total internal reflection. Optical fibres are fabricated with high quality composite glass/quartz fibres. Each fibre consists of a core and cladding. The refractive index of the material of the core is higher than that of the cladding.

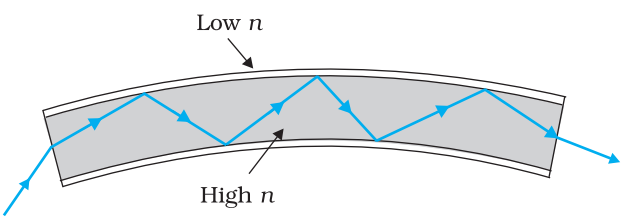


FIGURE 9.16 Light undergoes successive total internal reflections as it moves through an optical fibre.

When a signal in the form of light is directed at one end of the fibre at a suitable angle, it undergoes repeated total internal reflections along the length of the fibre and finally comes out at the other end (Fig. 9.16). Since light undergoes total internal reflection at each stage, there is no appreciable loss in the intensity of the light signal. Optical fibres are fabricated such that light reflected at one side of inner surface strikes the other at an angle larger than the critical angle. Even if the fibre is bent, light can easily travel along its length. Thus, an optical fibre can be used to act as an optical pipe.

A bundle of optical fibres can be put to several uses. Optical fibres are extensively used for transmitting and receiving electrical signals which are converted to light by suitable transducers. Obviously, optical fibres can also be used for transmission of optical signals. For example, these are used as a 'light pipe' to facilitate visual examination of internal organs like esophagus, stomach and intestines. You might have seen a commonly

available decorative lamp with fine plastic fibres with their free ends forming a fountain like structure. The other end of the fibres is fixed over an electric lamp. When the lamp is switched on, the light travels from the bottom of each fibre and appears at the tip of its free end as a dot of light. The fibres in such decorative lamps are optical fibres.

The main requirement in fabricating optical fibres is that there should be very little absorption of light as it travels for long distances inside them. This has been achieved by purification and special preparation of materials such as quartz. In silica glass fibres, it is possible to transmit more than 95% of the light over a fibre length of 1 km. (Compare with what you expect for a block of ordinary window glass 1 km thick.)

9.5 Refraction at Spherical Surfaces and by Lenses

We have so far considered refraction at a plane interface. We shall now consider refraction at a spherical interface between two transparent media. An infinitesimal part of a spherical surface can be regarded as planar and the same laws of refraction can be applied at every point on the surface. Just as for reflection by a spherical mirror, the normal at the point of incidence is perpendicular to the tangent plane to the spherical surface at that point and, therefore, passes through its centre of curvature. We first consider refraction by a single spherical surface and follow it by thin lenses. A thin lens is a transparent optical medium bounded by two surfaces; at least one of which should be spherical. Applying the formula for image formation by a single spherical surface successively at the two surfaces of a lens, we shall obtain the lens maker's formula and then the lens formula.

9.5.1 Refraction at a spherical surface

Figure 9.17 shows the geometry of formation of image I of an object O on the principal axis of a spherical surface with centre of curvature C, and radius of curvature R. The rays are incident from a medium of refractive index n_1, to another of refractive index n_2. As before, we take the aperture (or the lateral size) of the surface to be small compared to other distances involved, so that small angle approximation can be made. In particular, NM will be taken to be nearly equal to the length of the perpendicular from the point N on the principal axis. We have, for small angles,

$$\tan \angle NOM = \frac{MN}{OM}$$

$$\tan \angle NCM = \frac{MN}{MC}$$

$$\tan \angle NIM = \frac{MN}{MI}$$

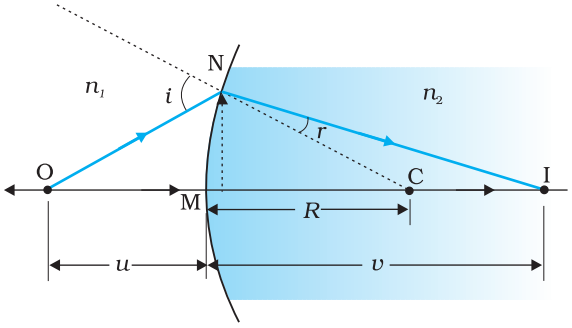


FIGURE 9.17 Refraction at a spherical surface separating two media.

Light sources and photometry

It is known that a body above absolute zero temperature emits electromagnetic radiation. The wavelength region in which the body emits the radiation depends on its absolute temperature. Radiation emitted by a hot body, for example, a tungsten filament lamp having temperature 2850 K are partly invisible and mostly in infrared (or heat) region. As the temperature of the body increases radiation emitted by it is in visible region. The sun with temperature of about 5500 K emits radiation whose energy versus wavelength graph peaks approximately at 550 nm corresponding to green light and is almost in the middle of the visible region. The energy versus wavelength distribution graph for a given body peaks at some wavelength, which is inversely proportional to the absolute temperature of that body.

The measurement of light as perceived by human eye is called *photometry*. Photometry is measurement of a physiological phenomenon, being the stimulus of light as received by the human eye, transmitted by the optic nerves and analysed by the brain. The main physical quantities in photometry are (i) the *luminous intensity* of the source, (ii) the *luminous flux* or flow of light from the source, and (iii) *illuminance* of the surface. The SI unit of *luminous intensity* (I) is candela (cd). The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and that has a radiant intensity in that direction of 1/683 watt per steradian. If a light source emits one candela of luminous intensity into a solid angle of one steradian, the total luminous flux emitted into that solid angle is one *lumen* (lm). A standard 100 watt incadescent light bulb emits approximately 1700 lumens.

In photometry, the only parameter, which can be measured directly is *illuminance*. It is defined as luminous flux incident per unit area on a surface (lm/m^2 or *lux*). Most light meters measure this quantity. The illuminance E, produced by a source of luminous intensity I, is given by $E = I/r^2$, where r is the normal distance of the surface from the source. A quantity named *luminance* (L), is used to characterise the brightness of emitting or reflecting flat surfaces. Its unit is cd/m^2 (sometimes called 'nit' in industry). A good LCD computer monitor has a brightness of about 250 nits.

Now, for ΔNOC, i is the exterior angle. Therefore, $i = \angle NOM + \angle NCM$

$$i = \frac{MN}{OM} + \frac{MN}{MC} \tag{9.13}$$

Similarly,

$$r = \angle NCM - \angle NIM$$

$$\text{i.e., } r = \frac{MN}{MC} - \frac{MN}{MI} \tag{9.14}$$

Now, by Snell's law

$$n_1 \sin i = n_2 \sin r$$

or for small angles

$$n_1 i = n_2 r$$

Substituting i and r from Eqs. (9.13) and (9.14), we get

$$\frac{n_1}{\text{OM}} + \frac{n_2}{\text{MI}} = \frac{n_2 - n_1}{\text{MC}} \tag{9.15}$$

Here, OM, MI and MC represent magnitudes of distances. Applying the Cartesian sign convention,

OM = $-u$, MI = $+v$, MC = $+R$

Substituting these in Eq. (9.15), we get

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R} \tag{9.16}$$

Equation (9.16) gives us a relation between object and image distance in terms of refractive index of the medium and the radius of curvature of the curved spherical surface. It holds for any curved spherical surface.

Example 9.6 Light from a point source in air falls on a spherical glass surface (n = 1.5 and radius of curvature = 20 cm). The distance of the light source from the glass surface is 100 cm. At what position the image is formed?

Solution

We use the relation given by Eq. (9.16). Here

u = – 100 cm, v = ?, R = + 20 cm, n_1 = 1, and n_2 = 1.5.

We then have

$$\frac{1.5}{v} + \frac{1}{100} = \frac{0.5}{20}$$

or v = +100 cm

The image is formed at a distance of 100 cm from the glass surface, in the direction of incident light.

EXAMPLE 9.6

9.5.2 Refraction by a lens

Figure 9.18(a) shows the geometry of image formation by a double convex lens. The image formation can be seen in terms of two steps: (i) The first refracting surface forms the image I_1 of the object O [Fig. 9.18(b)]. The image I_1 acts as a virtual object for the second surface that forms the image at I [Fig. 9.18(c)]. Applying Eq. (9.15) to the first interface ABC, we get

$$\frac{n_1}{\text{OB}} + \frac{n_2}{\text{BI}_1} = \frac{n_2 - n_1}{\text{BC}_1} \tag{9.17}$$

A similar procedure applied to the second interface* ADC gives,

$$-\frac{n_2}{\text{DI}_1} + \frac{n_1}{\text{DI}} = \frac{n_2 - n_1}{\text{DC}_2} \tag{9.18}$$

* Note that now the refractive index of the medium on the right side of ADC is n_1 while on its left it is n_2. Further DI_1 is negative as the distance is measured against the direction of incident light.

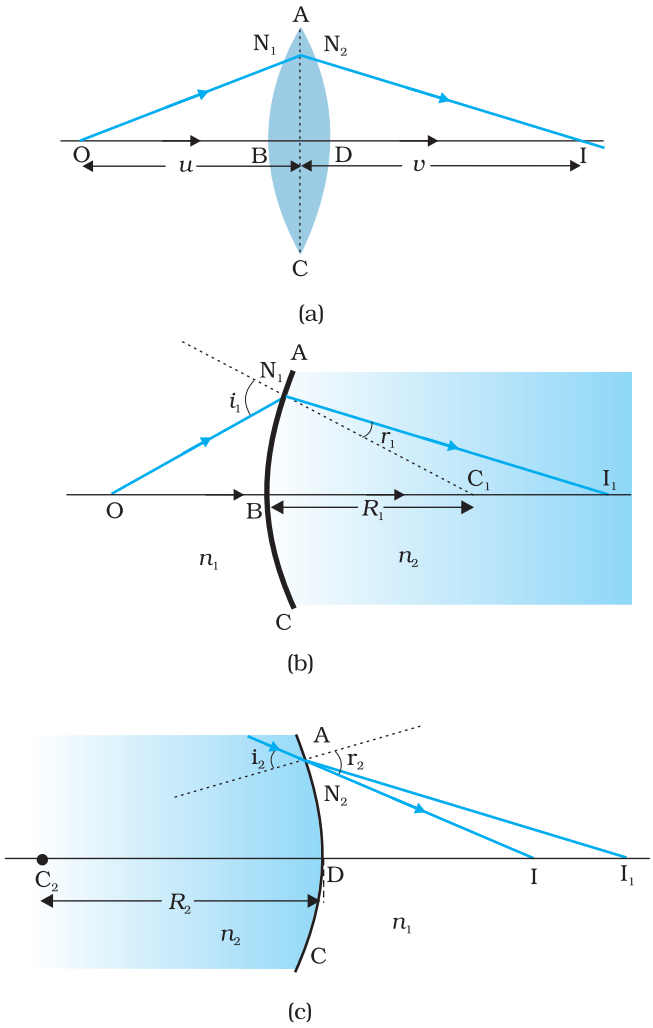


FIGURE 9.18 (a) The position of object, and the image formed by a double convex lens, (b) Refraction at the first spherical surface and (c) Refraction at the second spherical surface.

For a thin lens, $BI_1 = DI_1$. Adding Eqs. (9.17) and (9.18), we get

$$\frac{n_1}{OB}+\frac{n_1}{DI}=(n_2-n_1)\left(\frac{1}{BC_1}+\frac{1}{DC_2}\right) \qquad (9.19)$$

Suppose the object is at infinity, i.e., $OB \to \infty$ and $DI = f$, Eq. (9.19) gives

$$\frac{n_1}{f}=(n_2-n_1)\left(\frac{1}{BC_1}+\frac{1}{DC_2}\right) \qquad (9.20)$$

The point where image of an object placed at infinity is formed is called the *focus* F, of the lens and the distance f gives its *focal length*. A lens has two foci, F and F′, on either side of it (Fig. 9.19). By the sign convention,

$BC_1 = +R_1$,

$DC_2 = -R_2$

So Eq. (9.20) can be written as

$$\frac{1}{f}=(n_{21}-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right) \quad \left(\because n_{21}=\frac{n_2}{n_1}\right) \qquad (9.21)$$

Equation (9.21) is known as the *lens maker's formula*. It is useful to design lenses of desired focal length using surfaces of suitable radii of curvature. Note that the formula is true for a concave lens also. In that case R_1 is negative, R_2 positive and therefore, f is negative.

From Eqs. (9.19) and (9.20), we get

$$\frac{n_1}{OB}+\frac{n_1}{DI} = \frac{n_1}{f} \qquad (9.22)$$

Again, in the thin lens approximation, B and D are both close to the optical centre of the lens. Applying the sign convention,

$BO = -u$, $DI = +v$, we get

$$\frac{1}{v}-\frac{1}{u}=\frac{1}{f} \qquad (9.23)$$

Equation (9.23) is the familiar *thin lens formula*. Though we derived it for a real image formed by a convex lens, the formula is valid for both convex as well as concave lenses and for both real and virtual images.

It is worth mentioning that the two foci, F and F′, of a double convex or concave lens are equidistant from the optical centre. The focus on the side of the (original) source of light is called the *first focal point*, whereas the other is called the *second focal point*.

To find the image of an object by a lens, we can, in principle, take any two rays emanating from a point on an object; trace their paths using

the laws of refraction and find the point where the refracted rays meet (or appear to meet). In practice, however, it is convenient to choose any two of the following rays:

(i) A ray emanating from the object parallel to the principal axis of the lens after refraction passes through the second principal focus F′ (in a convex lens) or appears to diverge (in a concave lens) from the first principal focus F.

(ii) A ray of light, passing through the optical centre of the lens, emerges without any deviation after refraction.

(iii) A ray of light passing through the first principal focus (for a convex lens) or appearing to meet at it (for a concave lens) emerges parallel to the principal axis after refraction.

Figures 9.19(a) and (b) illustrate these rules for a convex and a concave lens, respectively. You should practice drawing similar ray diagrams for different positions of the object with respect to the lens and also verify that the lens formula, Eq. (9.23), holds good for all cases.

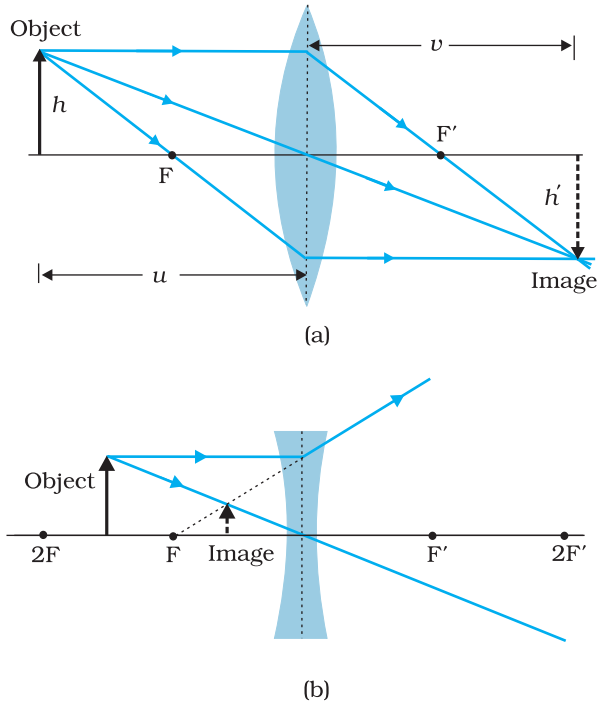


FIGURE 9.19 Tracing rays through (a) convex lens (b) concave lens.

Here again it must be remembered that each point on an object gives out infinite number of rays. All these rays will pass through the same image point after refraction at the lens.

Magnification (m) produced by a lens is defined, like that for a mirror, as the ratio of the size of the image to that of the object. Proceeding in the same way as for spherical mirrors, it is easily seen that for a lens

$$m = \frac{h'}{h} = \frac{v}{u} \tag{9.24}$$

When we apply the sign convention, we see that, for erect (and virtual) image formed by a convex or concave lens, m is positive, while for an inverted (and real) image, m is negative.

Example 9.7 A magician during a show makes a glass lens with $n = 1.47$ disappear in a trough of liquid. What is the refractive index of the liquid? Could the liquid be water?

Solution

The refractive index of the liquid must be equal to 1.47 in order to make the lens disappear. This means $n_1 = n_2$. This gives $1/f = 0$ or $f \to \infty$. The lens in the liquid will act like a plane sheet of glass. No, the liquid is not water. It could be glycerine.

EXAMPLE 9.7

9.5.3 Power of a lens

Power of a lens is a measure of the convergence or divergence, which a lens introduces in the light falling on it. Clearly, a lens of shorter focal

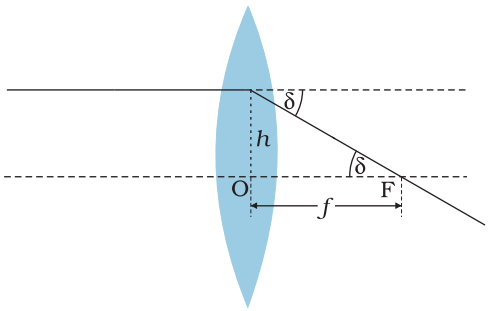


FIGURE 9.20 Power of a lens.

length bends the incident light more, while converging it in case of a convex lens and diverging it in case of a concave lens. The *power P* of a lens is defined as the tangent of the angle by which it converges or diverges a beam of light falling at unit distant from the optical centre (Fig. 9.20).

$\tan\delta = \frac{h}{f}$; if $h = 1$, $\tan\delta = \frac{1}{f}$ or $\delta = \frac{1}{f}$ for small value of δ. Thus,

$$P = \frac{1}{f} \tag{9.25}$$

The SI unit for power of a lens is dioptre (D): 1D = 1m^{-1}. The power of a lens of focal length of 1 metre is one dioptre. Power of a lens is positive for a converging lens and negative for a diverging lens. Thus, when an optician prescribes a corrective lens of power + 2.5 D, the required lens is a convex lens of focal length + 40 cm. A lens of power of – 4.0 D means a concave lens of focal length – 25 cm.

EXAMPLE 9.8

Example 9.8 (i) If f = 0.5 m for a glass lens, what is the power of the lens? (ii) The radii of curvature of the faces of a double convex lens are 10 cm and 15 cm. Its focal length is 12 cm. What is the refractive index of glass? (iii) A convex lens has 20 cm focal length in air. What is focal length in water? (Refractive index of air-water = 1.33, refractive index for air-glass = 1.5.)

Solution

(i) Power = +2 dioptre.

(ii) Here, we have f = +12 cm, R_1 = +10 cm, R_2 = –15 cm.
Refractive index of air is taken as unity.
We use the lens formula of Eq. (9.22). The sign convention has to be applied for f, R_1 and R_2.
Substituting the values, we have

$$\frac{1}{12} = (n-1)\left(\frac{1}{10} - \frac{1}{-15}\right)$$

This gives n = 1.5.

(iii) For a glass lens in air, n_2 = 1.5, n_1 = 1, f = +20 cm. Hence, the lens formula gives

$$\frac{1}{20} = 0.5\left[\frac{1}{R_1} - \frac{1}{R_2}\right]$$

For the same glass lens in water, n_2 = 1.5, n_1 = 1.33. Therefore,

$$\frac{1.33}{f} = (1.5 - 1.33)\left[\frac{1}{R_1} - \frac{1}{R_2}\right] \tag{9.26}$$

Combining these two equations, we find f = + 78.2 cm.

9.5.4 Combination of thin lenses in contact

Consider two lenses A and B of focal length f_1 and f_2 placed in contact with each other. Let the object be placed at a point O beyond the focus of

the first lens A (Fig. 9.21). The first lens produces an image at I_1. Since image I_1 is real, it serves as a virtual object for the second lens B, producing the final image at I. It must, however, be borne in mind that formation of image by the first lens is presumed only to facilitate determination of the position of the final image. In fact, the direction of rays emerging from the first lens gets modified in accordance with the angle at which they strike the second lens. Since the lenses are thin, we assume the optical centres of the lenses to be coincident. Let this central point be denoted by P.

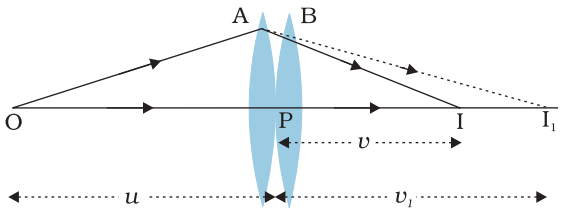


FIGURE 9.21 Image formation by a combination of two thin lenses in contact.

For the image formed by the first lens A, we get

$$\frac{1}{v_1} - \frac{1}{u} = \frac{1}{f_1} \tag{9.27}$$

For the image formed by the second lens B, we get

$$\frac{1}{v} - \frac{1}{v_1} = \frac{1}{f_2} \tag{9.28}$$

Adding Eqs. (9.27) and (9.28), we get

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f_1} + \frac{1}{f_2} \tag{9.29}$$

If the two lens-system is regarded as equivalent to a single lens of focal length f, we have

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

so that we get

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} \tag{9.30}$$

The derivation is valid for any number of thin lenses in contact. If several thin lenses of focal length $f_1, f_2, f_3, \ldots$ are in contact, the effective focal length of their combination is given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} + \frac{1}{f_3} + \ldots \tag{9.31}$$

In terms of power, Eq. (9.31) can be written as

$$P = P_1 + P_2 + P_3 + \ldots \tag{9.32}$$

where P is the net power of the lens combination. Note that the sum in Eq. (9.32) is an algebraic sum of individual powers, so some of the terms on the right side may be positive (for convex lenses) and some negative (for concave lenses). Combination of lenses helps to obtain diverging or converging lenses of desired magnification. It also enhances sharpness of the image. Since the image formed by the first lens becomes the object for the second, Eq. (9.25) implies that the total magnification m of the combination is a product of magnification (m_1, m_2, m_3,...) of individual lenses

$$m = m_1 \, m_2 \, m_3 \ \ldots \tag{9.33}$$

Such a system of combination of lenses is commonly used in designing lenses for cameras, microscopes, telescopes and other optical instruments.

Example 9.9 Find the position of the image formed by the lens combination given in the Fig. 9.22.

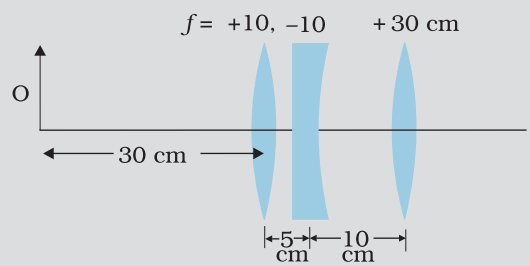


FIGURE 9.22

Solution Image formed by the first lens

$$\frac{1}{v_1} - \frac{1}{u_1} = \frac{1}{f_1}$$

$$\frac{1}{v_1} - \frac{1}{-30} = \frac{1}{10}$$

or $v_1 = 15$ cm

The image formed by the first lens serves as the object for the second. This is at a distance of (15 – 5) cm = 10 cm to the right of the second lens. Though the image is real, it serves as a virtual object for the second lens, which means that the rays appear to come from it for the second lens.

$$\frac{1}{v_2} - \frac{1}{10} = \frac{1}{-10}$$

or $v_2 = \infty$

The virtual image is formed at an infinite distance to the left of the second lens. This acts as an object for the third lens.

$$\frac{1}{v_3} - \frac{1}{u_3} = \frac{1}{f_3}$$

or $$\frac{1}{v_3} = \frac{1}{\infty} + \frac{1}{30}$$

or $v_3 = 30$ cm

The final image is formed 30 cm to the right of the third lens.

EXAMPLE 9.9

9.6 Refraction through a Prism

Figure 9.23 shows the passage of light through a triangular prism ABC. The angles of incidence and refraction at the first face AB are i and r_1, while the angle of incidence (from glass to air) at the second face AC is r_2 and the angle of refraction or emergence e. The angle between the emergent ray RS and the direction of the incident ray PQ is called the *angle of deviation*, δ.

In the quadrilateral AQNR, two of the angles (at the vertices Q and R) are right angles. Therefore, the sum of the other angles of the quadrilateral is 180°.

$\angle A + \angle QNR = 180°$

From the triangle QNR,

$r_1 + r_2 + \angle QNR = 180°$

Comparing these two equations, we get

$$r_1 + r_2 = A \qquad (9.34)$$

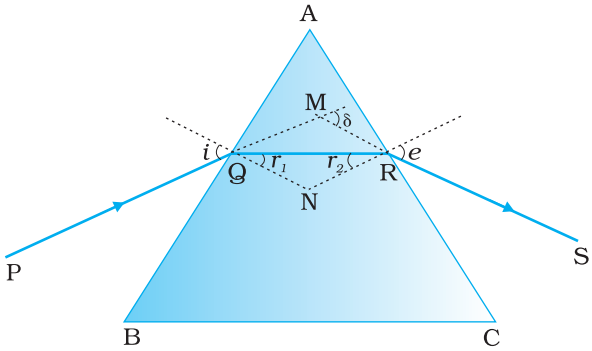


FIGURE 9.23 A ray of light passing through a triangular glass prism.

The total deviation δ is the sum of deviations at the two faces,

$\delta = (i - r_1) + (e - r_2)$

that is,

$$\delta = i + e - A \qquad (9.35)$$

Thus, the angle of deviation depends on the angle of incidence. A plot between the angle of deviation and angle of incidence is shown in Fig. 9.24. You can see that, in general, any given value of δ, except for $i = e$, corresponds to two values i and hence of e. This, in fact, is expected from the symmetry of i and e in Eq. (9.35), i.e., δ remains the same if i and e are interchanged. Physically, this is related to the fact that the path of ray in Fig. 9.23 can be traced back, resulting in the same angle of deviation. At the minimum deviation D_m, the refracted ray inside the prism becomes parallel to its base. We have

$\delta = D_m$, $i = e$ which implies $r_1 = r_2$.

Equation (9.34) gives

$$2r = A \text{ or } r = \frac{A}{2} \qquad (9.36)$$

In the same way, Eq. (9.35) gives

$$D_m = 2i - A, \text{ or } i = (A + D_m)/2 \qquad (9.37)$$

The refractive index of the prism is

$$n_{21} = \frac{n_2}{n_1} = \frac{\sin[(A + D_m)/2]}{\sin[A/2]} \qquad (9.38)$$

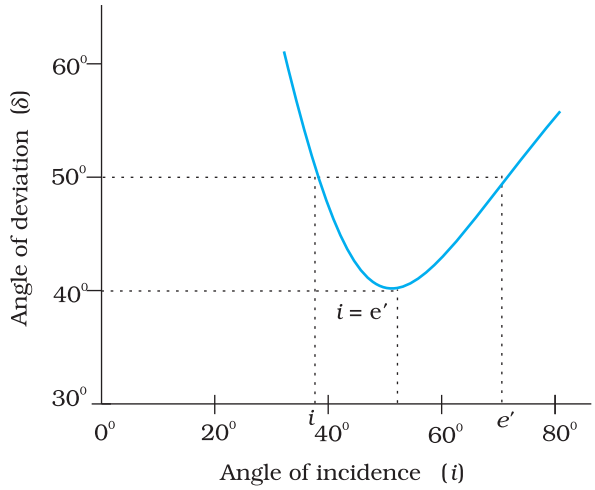


FIGURE 9.24 Plot of angle of deviation (δ) versus angle of incidence (i) for a triangular prism.

The angles A and D_m can be measured experimentally. Equation (9.38) thus provides a method of determining refractive index of the material of the prism.

For a small angle prism, i.e., a thin prism, D_m is also very small, and we get

$$n_{21} = \frac{\sin[(A + D_m)/2]}{\sin[A/2]} \simeq \frac{(A + D_m)/2}{A/2}$$

$D_m = (n_{21} - 1)A$

It implies that, thin prisms do not deviate light much.

9.7 Dispersion by a Prism

It has been known for a long time that when a narrow beam of sunlight, usually called white light, is incident on a glass prism, the emergent light is seen to be consisting of several colours. There is actually a continuous variation of colour, but broadly, the different colours that appear in sequence are: **v**iolet, **i**ndigo, **b**lue, **g**reen, **y**ellow, **o**range and **r**ed (given by the acronym VIBGYOR). The red light bends the least, while the violet light bends the most (Fig. 9.25).

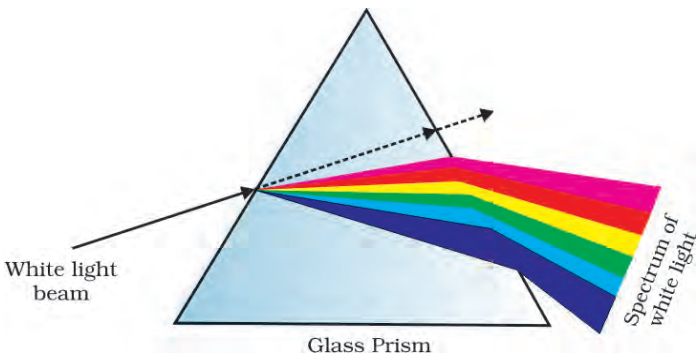


FIGURE 9.25 Dispersion of sunlight or white light on passing through a glass prism. The relative deviation of different colours shown is highly exaggerated.

The phenomenon of splitting of light into different colours is known as *dispersion*. The pattern of colour components of light is called the spectrum of light. The word *spectrum* is now used in a much more general sense: we discussed in Chapter 8 the electro-magnetic spectrum over the large range of wavelengths, from γ-rays to radio waves, of which the spectrum of light (visible spectrum) is only a small part.

Though the reason for appearance of spectrum is now common knowledge, it was a matter of much debate in the history of physics. Does the prism itself create colour in some way or does it only separate the colours already present in white light?

In a classic experiment known for its simplicity but great significance, Isaac Newton settled the issue once for all. He put another similar prism, but in an inverted position, and let the emergent beam from the first prism fall on the second prism (Fig. 9.26). The resulting emergent beam was found to be white light. The explanation was clear— the first prism splits the white light into its component colours, while the inverted prism recombines them to give white light. Thus, white light itself consists of light of different colours, which are separated by the prism.

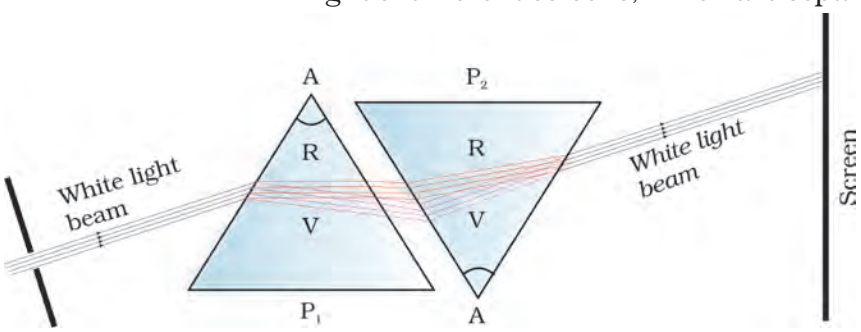


FIGURE 9.26 Schematic diagram of Newton's classic experiment on dispersion of white light.

It must be understood here that a ray of light, as defined mathematically, does not exist. An actual ray is really a beam of many rays of light. Each ray splits into component colours when it enters the glass prism. When those coloured rays come out on the other side, they again produce a white beam.

We now know that colour is associated with wavelength of light. In the visible spectrum, red light is at the long wavelength end (~700 nm) while the violet light is at the short wavelength end (~ 400 nm). Dispersion takes place because the refractive

index of medium for different wavelengths (colours) is different. For example, the bending of red component of white light is least while it is most for the violet. Equivalently, red light travels faster than violet light in a glass prism. Table 9.2 gives the refractive indices for different wavelength for crown glass and flint glass. Thick lenses could be assumed as made of many prisms, therefore, thick lenses show *chromatic aberration* due to dispersion of light. When white light passes through thick lenses, red and blue colours focus at different points. This phenomena is known as chromatic aberration.

TABLE 9.2 REFRACTIVE INDICES FOR DIFFERENT WAVELENGTHS

Colour	Wavelength (nm)	Crown glass	Flint glass
Violet	396.9	1.533	1.663
Blue	486.1	1.523	1.639
Yellow	589.3	1.517	1.627
Red	656.3	1.515	1.622

The variation of refractive index with wavelength may be more pronounced in some media than the other. In vacuum, of course, the speed of light is independent of wavelength. Thus, vacuum (or air approximately) is a non-dispersive medium in which all colours travel with the same speed. This also follows from the fact that sunlight reaches us in the form of white light and not as its components. On the other hand, glass is a dispersive medium.

9.8 SOME NATURAL PHENOMENA DUE TO SUNLIGHT

The interplay of light with things around us gives rise to several beautiful phenomena. The spectacle of colour that we see around us all the time is possible only due to sunlight. The blue of the sky, white clouds, the red-hue at sunrise and sunset, the rainbow, the brilliant colours of some pearls, shells, and wings of birds, are just a few of the natural wonders we are used to. We describe some of them here from the point of view of physics.

9.8.1 The rainbow

PHYSICS

Formation of rainbows
http://www.eo.ucar.edu/rainbows
http://www.atoptics.co.uk/bows.htm

The rainbow is an example of the dispersion of sunlight by the water drops in the atmosphere. This is a phenomenon due to combined effect of dispersion, refraction and reflection of sunlight by spherical water droplets of rain. The conditions for observing a rainbow are that the Sun should be shining in one part of the sky (say near western horizon) while it is raining in the opposite part of the sky (say eastern horizon). An observer can therefore see a rainbow only when his back is towards the Sun.

In order to understand the formation of rainbows, consider Fig. (9.27(a). Sunlight is first refracted as it enters a raindrop, which causes the different wavelengths (colours) of white light to separate. Longer wangelength of light (red) are bent the least while the shorter wavelength (violet) are bent the most. Next, these component rays strike

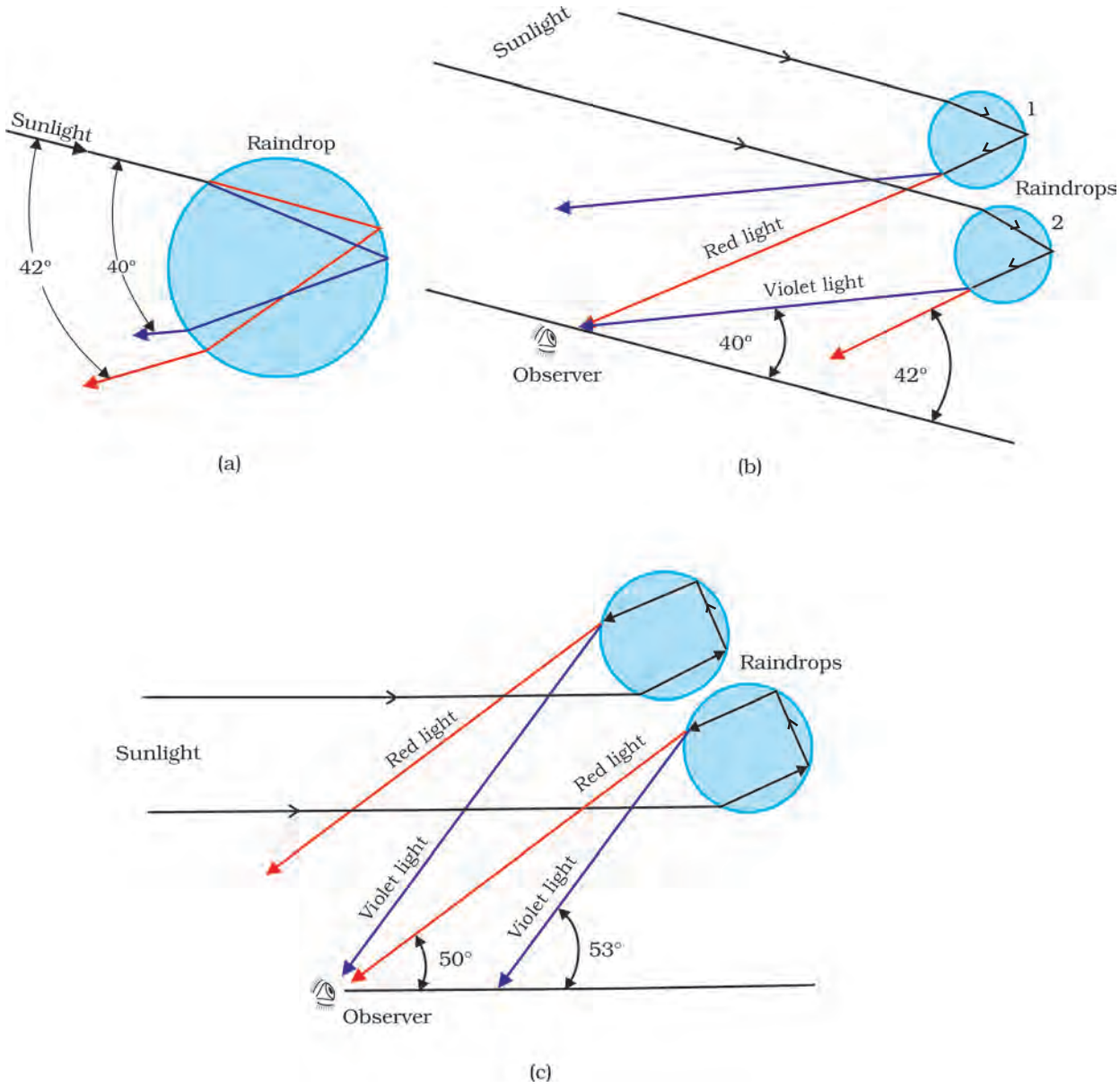


FIGURE 9.27 Rainbow: (a) The sun rays incident on a water drop get refracted twice and reflected internally by a drop; (b) Enlarged view of internal reflection and refraction of a ray of light inside a drop forming primary rainbow; and (c) secondary rainbow is formed by rays undergoing internal reflection twice inside the drop.

the inner surface of the water drop and get internally reflected if the angle between the refracted ray and normal to the drop surface is greater then the critical angle (48° in this case). The reflected light is refracted again when it comes out of the drop, as shown in the figure. It is found that the violet light emerges at an angle of 40° related to the incoming sunlight and red light emerges at an angle of 42°. For other colours, angles lie in between these two values.

Figure 9.27(b) explains the formation of primary rainbow. We see that red light from drop 1 and violet light from drop 2 reach the observer's eye. The violet from drop 1 and red light from drop 2 are directed at level above or below the observer. Thus the observer sees a rainbow with red colour on the top and violet on the bottom. The primary rainbow is a result of three-step process, that is, refraction, reflection and refraction.

When light rays undergoes *two* internal reflections inside a raindrop, instead of *one* as in the primary rainbow, a secondary rainbow is formed as shown in Fig. 9.27(c). It is due to four-step process. The intensity of light is reduced at the second reflection and hence the secondary rainbow is fainter than the primary rainbow. Further, the order of the colours is reversed in it as is clear from Fig. 9.27(c).

9.8.2 Scattering of light

As sunlight travels through the earth's atmosphere, it gets *scattered* (changes its direction) by the atmospheric particles. Light of shorter wavelengths is scattered much more than light of longer wavelengths. (The amount of scattering is inversely proportional to the fourth power of the wavelength. This is known as Rayleigh scattering). Hence, the bluish colour predominates in a clear sky, since blue has a shorter wavelength than red and is scattered much more strongly. In fact, violet gets scattered even more than blue, having a shorter wavelength. But since our eyes are more sensitive to blue than violet, we see the sky blue.

Large particles like dust and water droplets present in the atmosphere behave differently. The relevant quantity here is the relative size of the wavelength of light λ, and the scatterer (of typical size, say, a). For $a << \lambda$, one has Rayleigh scattering which is proportional to $1/\lambda^4$. For $a >> \lambda$, i.e., large scattering objects (for example, raindrops, large dust or ice particles) this is not true; all wavelengths are scattered nearly equally. Thus, clouds which have droplets of water with $a >> \lambda$ are generally white.

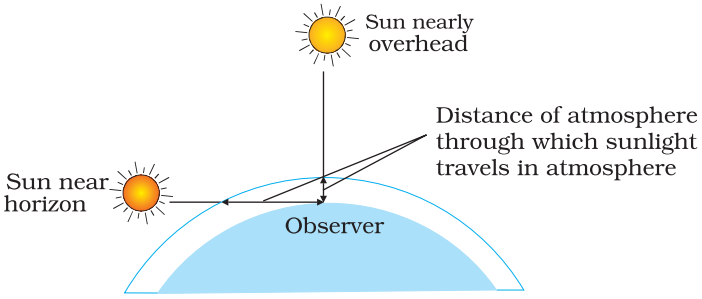


FIGURE 9.28 Sunlight travels through a longer distance in the atmosphere at sunset and sunrise.

At sunset or sunrise, the sun's rays have to pass through a larger distance in the atmosphere (Fig. 9.28). Most of the blue and other shorter wavelengths are removed by scattering. The least scattered light reaching our eyes, therefore, the sun looks reddish. This explains the reddish appearance of the sun and full moon near the horizon.

9.9 Optical Instruments

A number of optical devices and instruments have been designed utilising reflecting and refracting properties of mirrors, lenses and prisms. Periscope, kaleidoscope, binoculars, telescopes, microscopes are some

examples of optical devices and instruments that are in common use. Our eye is, of course, one of the most important optical device the nature has endowed us with. Starting with the eye, we then go on to describe the principles of working of the microscope and the telescope.

9.9.1 The eye

Figure 9.29 (a) shows the eye. Light enters the eye through a curved front surface, the cornea. It passes through the pupil which is the central hole in the iris. The size of the pupil can change under control of muscles. The light is further focussed by the eye lens on the retina. The retina is a film of nerve fibres covering the curved back surface of the eye. The retina contains rods and cones which sense light intensity and colour, respectively, and transmit electrical signals via the optic nerve to the brain which finally processes this information. The shape (curvature) and therefore the focal length of the lens can be modified somewhat by the ciliary muscles. For example, when the muscle is relaxed, the focal length is about 2.5 cm and objects at infinity are in sharp focus on the retina. When the object is brought closer to the eye, in order to maintain the same image-lens distance (≅ 2.5 cm), the focal length of the eye lens becomes shorter by the action of the ciliary muscles. This property of the eye is called *accommodation*. If the object is too close to the eye, the lens cannot curve enough to focus the image on to the retina, and the image is blurred. The closest distance for which the lens can focus light on the retina is called the *least distance of distinct vision*, or the *near point*. The standard value for normal vision is taken as 25 cm. (Often the near point is given the symbol D.) This distance increases with age, because of the decreasing effectiveness of the ciliary muscle and the loss of flexibility of the lens. The near point may be as close as about 7 to 8 cm in a child ten years of age, and may increase to as much as 200 cm at 60 years of age. Thus, if an elderly person tries to read a book at about 25 cm from the eye, the image appears blurred. This condition (defect of the eye) is called *presbyopia*. It is corrected by using a converging lens for reading.

Thus, our eyes are marvellous organs that have the capability to interpret incoming electromagnetic waves as images through a complex process. These are our greatest assets and we must take proper care to protect them. Imagine the world without a pair of functional eyes. Yet many amongst us bravely face this challenge by effectively overcoming their limitations to lead a normal life. They deserve our appreciation for their courage and conviction.

In spite of all precautions and proactive action, our eyes may develop some defects due to various reasons. We shall restrict our discussion to some common optical defects of the eye. For example, the light from a distant object arriving at the eye-lens may get converged at a point in front of the retina. This type of defect is called *nearsightedness* or *myopia*. This means that the eye is producing too much convergence in the incident beam. To compensate this, we interpose a concave lens between the eye and the object, with the diverging effect desired to get the image focussed on the retina [Fig. 9.29(b)].

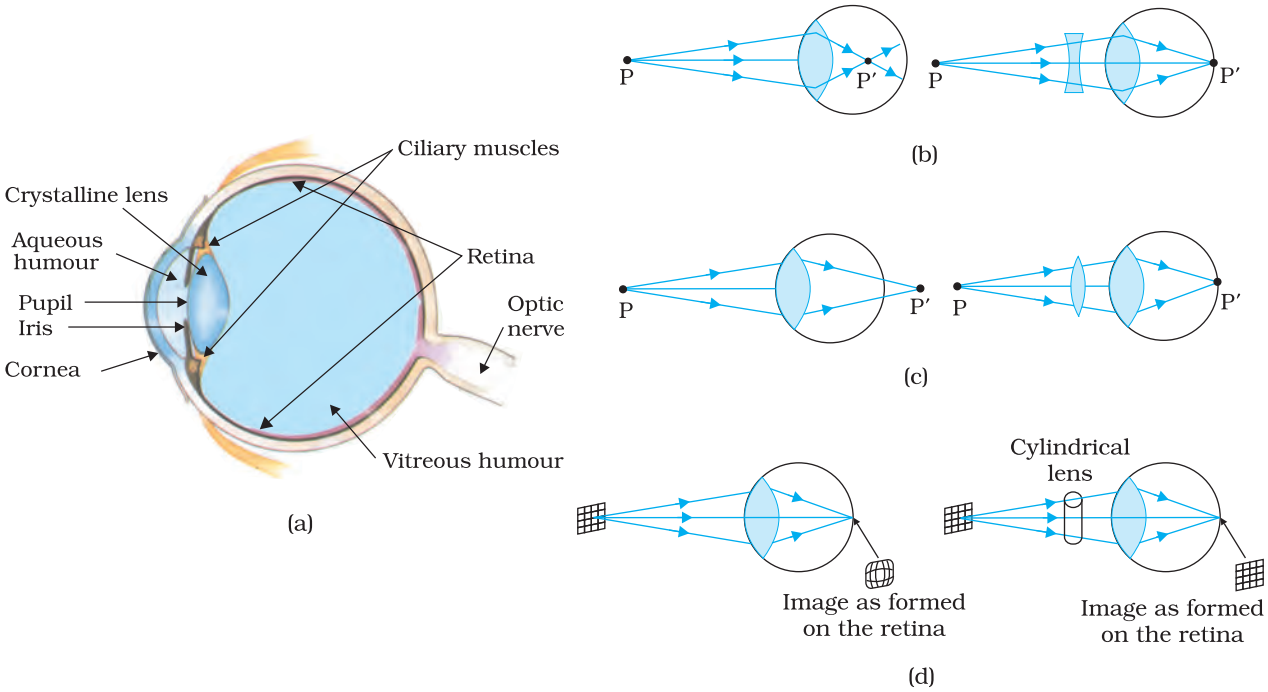


FIGURE 9.29 (a) The structure of the eye; (b) shortsighted or myopic eye and its correction; (c) farsighted or hypermetropic eye and its correction; and (d) astigmatic eye and its correction.

Similarly, if the eye-lens focusses the incoming light at a point behind the retina, a convergent lens is needed to compensate for the defect in vision. This defect is called *farsightedness* or *hypermetropia* [Fig. 9.29(c)].

Another common defect of vision is called *astigmatism.* This occurs when the cornea is not spherical in shape. For example, the cornea could have a larger curvature in the vertical plane than in the horizontal plane or vice-versa. If a person with such a defect in eye-lens looks at a wire mesh or a grid of lines, focussing in either the vertical or the horizontal plane may not be as sharp as in the other plane. Astigmatism results in lines in one direction being well focussed while those in a perpendicular direction may appear distorted [Fig. 9.29(d)]. Astigmatism can be corrected by using a cylindrical lens of desired radius of curvature with an appropriately directed axis. This defect can occur along with myopia or hypermetropia.

Example 9.10 What focal length should the reading spectacles have for a person for whom the least distance of distinct vision is 50 cm?

Solution The distance of normal vision is 25 cm. So if a book is at $u = -25$ cm, its image should be formed at $v = -50$ cm. Therefore, the desired focal length is given by

$$\frac{1}{f} = \frac{1}{v} - \frac{1}{u}$$

$$\text{or } \frac{1}{f} = \frac{1}{-50} - \frac{1}{-25} = \frac{1}{50}$$

or $f = +50$ cm (convex lens).

EXAMPLE 9.10

EXAMPLE 9.11

Example 9.11

(a) The far point of a myopic person is 80 cm in front of the eye. What is the power of the lens required to enable him to see very distant objects clearly?

(b) In what way does the corrective lens help the above person? Does the lens magnify very distant objects? Explain carefully.

(c) The above person prefers to remove his spectacles while reading a book. Explain why?

Solution

(a) Solving as in the previous example, we find that the person should use a concave lens of focal length = – 80 cm, i.e., of power = – 1.25 dioptres.

(b) No. The concave lens, in fact, reduces the size of the object, but the angle subtended by the distant object at the eye is the same as the angle subtended by the image (at the far point) at the eye. The eye is able to see distant objects not because the corrective lens magnifies the object, but because it brings the object (i.e., it produces virtual image of the object) at the far point of the eye which then can be focussed by the eye-lens on the retina.

(c) The myopic person may have a normal near point, i.e., about 25 cm (or even less). In order to read a book with the spectacles, such a person must keep the book at a distance greater than 25 cm so that the image of the book by the concave lens is produced not closer than 25 cm. The angular size of the book (or its image) at the greater distance is evidently less than the angular size when the book is placed at 25 cm and no spectacles are needed. Hence, the person prefers to remove the spectacles while reading.

EXAMPLE 9.12

Example 9.12 (a) The near point of a hypermetropic person is 75 cm from the eye. What is the power of the lens required to enable the person to read clearly a book held at 25 cm from the eye? (b) In what way does the corrective lens help the above person? Does the lens magnify objects held near the eye? (c) The above person prefers to remove the spectacles while looking at the sky. Explain why?

Solution

(a) $u = -25$ cm, $v = -75$ cm

$1/f = 1/25 - 1/75$, i.e., $f = 37.5$ cm.

The corrective lens needs to have a converging power of +2.67 dioptres.

(b) The corrective lens produces a virtual image (at 75 cm) of an object at 25 cm. The angular size of this image is the same as that of the object. In this sense the lens does not magnify the object but merely brings the object to the near point of the hypermetric eye, which then gets focussed on the retina. However, the angular size is greater than that of the same object at the near point (75 cm) viewed without the spectacles.

(c) A hypermetropic eye may have normal far point i.e., it may have enough converging power to focus parallel rays from infinity on the retina of the shortened eyeball. Wearing spectacles of converging lenses (used for near vision) will amount to more converging power than needed for parallel rays. Hence the person prefers not to use the spectacles for far objects.

9.9.2 The microscope

A simple magnifier or microscope is a converging lens of small focal length (Fig. 9.30). In order to use such a lens as a microscope, the lens is held near the object, one focal length away or less, and the eye is positioned close to the lens on the other side. The idea is to get an erect, magnified and virtual image of the object at a distance so that it can be viewed comfortably, i.e., at 25 cm or more. If the object is at a distance f, the image is at infinity. However, if the object is at a distance slightly less than the focal length of the lens, the image is virtual and closer than infinity. Although the closest comfortable distance for viewing the image is when it is at the near point (distance $D \cong 25$ cm), it causes some strain on the eye. Therefore, the image formed at infinity is often considered most suitable for viewing by the relaxed eye. We show both cases, the first in Fig. 9.30(a), and the second in Fig. 9.30(b) and (c).

The linear magnification m, for the image formed at the near point D, by a simple microscope can be obtained by using the relation

$$m = \frac{v}{u} = v\left(\frac{1}{v} - \frac{1}{f}\right) = \left(1 - \frac{v}{f}\right)$$

Now according to our sign convention, v is negative, and is equal in magnitude to D. Thus, the magnification is

$$m = \left(1 + \frac{D}{f}\right) \qquad (9.39)$$

Since D is about 25 cm, to have a magnification of six, one needs a convex lens of focal length, $f = 5$ cm.

Note that $m = h'/h$ where h is the size of the object and h' the size of the image. This is also the ratio of the angle subtended by the image to that subtended by the object, if placed at D for comfortable viewing. (Note that this is not the angle actually subtended by the object at the eye, which is h/u.) What a single-lens simple magnifier achieves is that it allows the object to be brought closer to the eye than D.

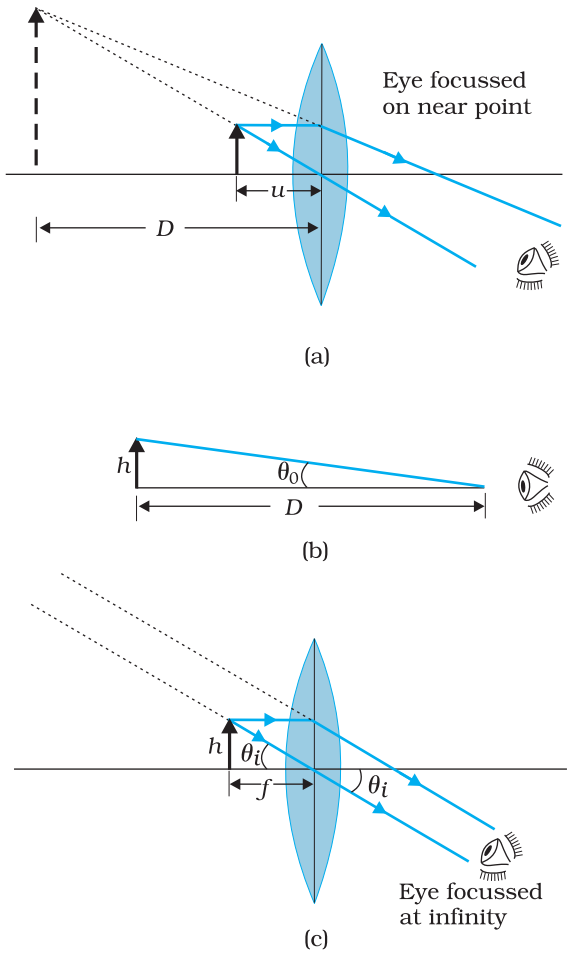


FIGURE 9.30 A simple microscope; (a) the magnifying lens is located such that the image is at the near point, (b) the angle subtanded by the object, is the same as that at the near point, and (c) the object near the focal point of the lens; the image is far off but closer than infinity.

We will now find the magnification when the image is at infinity. In this case we will have to obtained the *angular* magnification. Suppose the object has a height h. The maximum angle it can subtend, and be clearly visible (without a lens), is when it is at the near point, i.e., a distance D. The angle subtended is then given by

$$\tan\,\theta_o = \left(\frac{h}{D}\right) \approx \theta_o \qquad (9.40)$$

We now find the angle subtended at the eye by the image when the object is at u. From the relations

$$\frac{h'}{h} = m = \frac{v}{u}$$

we have the angle subtended by the image

$\tan\theta_i = \frac{h'}{-v} = \frac{h}{-v}\cdot\frac{v}{u} = \frac{h}{-u} \approx \theta$. The angle subtended by the object, when it is at $u = -f$.

$$\theta_i = \left(\frac{h}{f}\right) \tag{9.41}$$

as is clear from Fig. 9.29(c). The angular magnification is, therefore

$$m = \left(\frac{\theta_i}{\theta_o}\right) = \frac{D}{f} \tag{9.42}$$

This is one less than the magnification when the image is at the near point, Eq. (9.39), but the viewing is more comfortable and the difference in magnification is usually small. In subsequent discussions of optical instruments (microscope and telescope) we shall assume the image to be at infinity.

A simple microscope has a limited maximum magnification (≤ 9) for realistic focal lengths. For much larger magnifications, one uses two lenses, one compounding the effect of the other. This is known as a *compound microscope*. A schematic diagram of a compound microscope is shown in Fig. 9.31. The lens nearest the object, called the *objective*, forms a real, inverted, magnified image of the object. This serves as the object for the second lens, the *eyepiece*, which functions essentially like a simple microscope or magnifier, produces the final image, which is enlarged and virtual. The first inverted image is thus near (at or within) the focal plane of the eyepiece, at a distance appropriate for final image formation at infinity, or a little closer for image formation at the near point. Clearly, the final image is inverted with respect to the original object.

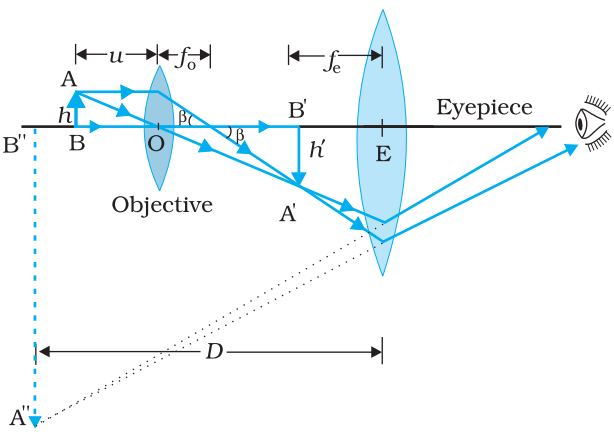


FIGURE 9.31 Ray diagram for the formation of image by a compound microscope.

We now obtain the magnification due to a compound microscope. The ray diagram of Fig. 9.31 shows that the (linear) magnification due to the objective, namely h'/h, equals

$$m_O = \frac{h'}{h} = \frac{L}{f_o} \tag{9.43}$$

where we have used the result

$$\tan\beta = \left(\frac{h}{f_o}\right) = \left(\frac{h'}{L}\right)$$

Here h' is the size of the first image, the object size being h and f_o being the focal length of the objective. The first image is formed near the focal point of the eyepiece. The distance L, i.e., the distance between the second focal point of the objective and the first focal point of the eyepiece (focal length f_e) is called the tube length of the compound microscope.

As the first inverted image is near the focal point of the eyepiece, we use the result from the discussion above for the simple microscope to obtain the (angular) magnification m_e due to it [Eq. (9.39)], when the final image is formed at the near point, is

$$m_e = \left(1 + \frac{D}{f_e}\right) \qquad [9.44(a)]$$

When the final image is formed at infinity, the angular magnification due to the eyepiece [Eq. (9.42)] is

$$m_e = (D/f_e) \qquad [9.44(b)]$$

Thus, the total magnification [(according to Eq. (9.33)], when the image is formed at infinity, is

$$m = m_o m_e = \left(\frac{L}{f_o}\right)\left(\frac{D}{f_e}\right) \qquad (9.45)$$

Clearly, to achieve a large magnification of a *small* object (hence the name microscope), the objective and eyepiece should have small focal lengths. In practice, it is difficult to make the focal length much smaller than 1 cm. Also large lenses are required to make L large.

For example, with an objective with f_o = 1.0 cm, and an eyepiece with focal length f_e = 2.0 cm, and a tube length of 20 cm, the magnification is

$$m = m_o m_e = \left(\frac{L}{f_o}\right)\left(\frac{D}{f_e}\right)$$

$$= \frac{20}{1} \times \frac{25}{2} = 250$$

Various other factors such as illumination of the object, contribute to the quality and visibility of the image. In modern microscopes, multi-component lenses are used for both the objective and the eyepiece to improve image quality by minimising various optical aberrations (defects) in lenses.

9.9.3 Telescope

The telescope is used to provide angular magnification of distant objects (Fig. 9.32). It also has an objective and an eyepiece. But here, the objective has a large focal length and a much larger aperture than the eyepiece. Light from a distant object enters the objective and a real image is formed in the tube at its second focal point. The eyepiece magnifies this image producing a final inverted image. The magnifying power m is the ratio of the angle β subtended at the eye by the final image to the angle α which the object subtends at the lens or the eye. Hence

$$m \approx \frac{\beta}{\alpha} \approx \frac{h}{f_e} \cdot \frac{f_o}{h} = \frac{f_o}{f_e} \qquad (9.46)$$

In this case, the length of the telescope tube is $f_o + f_e$.

PHYSICS

The world's largest optical telescopes
http://astro.nineplanets.org/bigeyes.html

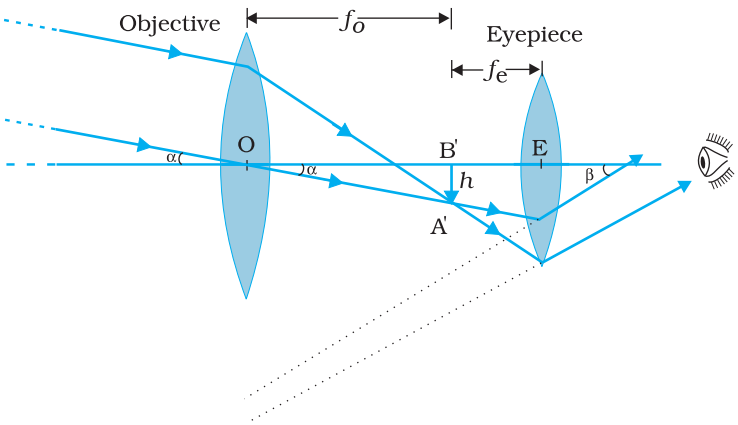


FIGURE 9.32 A refracting telescope.

Terrestrial telescopes have, in addition, a pair of inverting lenses to make the final image erect. Refracting telescopes can be used both for terrestrial and astronomical observations. For example, consider a telescope whose objective has a focal length of 100 cm and the eyepiece a focal length of 1 cm. The magnifying power of this telescope is m = 100/1 = 100.

Let us consider a pair of stars of actual separation 1′ (one minute of arc). The stars appear as though they are separated by an angle of 100 × 1′ = 100′ =1.67°.

The main considerations with an astronomical telescope are its light gathering power and its resolution or resolving power. The former clearly depends on the area of the objective. With larger diameters, fainter objects can be observed. The resolving power, or the ability to observe two objects distinctly, which are in very nearly the same direction, also depends on the diameter of the objective. So, the desirable aim in optical telescopes is to make them with objective of large diameter. The largest lens objective in use has a diameter of 40 inch (~1.02 m). It is at the Yerkes Observatory in Wisconsin, USA. Such big lenses tend to be very heavy and therefore, difficult to make and support by their edges. Further, it is rather difficult and expensive to make such large sized lenses which form images that are free from any kind of chromatic aberration and distortions.

For these reasons, modern telescopes use a concave mirror rather than a lens for the objective. Telescopes with mirror objectives are called *reflecting* telescopes. There is no chromatic aberration in a mirror. Mechanical support is much less of a problem since a mirror weighs much less than a lens of equivalent optical quality, and can be supported over its entire back surface, not just over its rim. One obvious problem with a reflecting telescope is that the objective mirror focusses light inside the telescope tube. One must have an eyepiece and the observer right there, obstructing some light (depending on the size of the observer cage). This is what is done in the very large 200 inch (~5.08 m) diameters, Mt. Palomar telescope, California. The viewer sits near the focal point of the mirror, in a small cage. Another solution to the problem is to deflect the light being focussed by another mirror. One such arrangement using a convex secondary mirror to focus the incident light, which now passes through a hole in the objective primary mirror, is shown in Fig. 9.33. This is known as a *Cassegrain* telescope, after its inventor. It has the advantages of a large focal length in a short

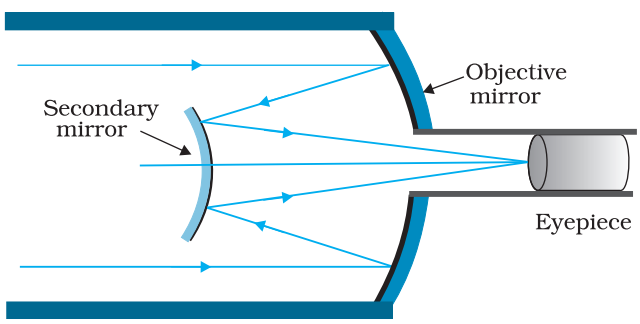


FIGURE 9.33 Schematic diagram of a reflecting telescope (Cassegrain).

telescope. The largest telescope in India is in Kavalur, Tamil Nadu. It is a 2.34 m diameter reflecting telescope (Cassegrain). It was ground, polished, set up, and is being used by the Indian Institute of Astrophysics, Bangalore. The largest reflecting telescopes in the world are the pair of Keck telescopes in Hawaii, USA, with a reflector of 10 metre in diameter.

SUMMARY

1. Reflection is governed by the equation $\angle i = \angle r'$ and refraction by the Snell's law, $\sin i/\sin r = n$, where the incident ray, reflected ray, refracted ray and normal lie in the same plane. Angles of incidence, reflection and refraction are i, r' and r, respectively.
2. The *critical angle of incidence* i_c for a ray incident from a denser to rarer medium, is that angle for which the angle of refraction is 90°. For $i > i_c$, total internal reflection occurs. Multiple internal reflections in diamond ($i_c \cong 24.4°$), totally reflecting prisms and mirage, are some examples of total internal reflection. Optical fibres consist of glass fibres coated with a thin layer of material of *lower* refractive index. Light incident at an angle at one end comes out at the other, after multiple internal reflections, even if the fibre is bent.
3. *Cartesian sign convention*: Distances measured in the same direction as the incident light are positive; those measured in the opposite direction are negative. All distances are measured from the pole/optic centre of the mirror/lens on the principal axis. The heights measured upwards above x-axis and normal to the principal axis of the mirror/lens are taken as positive. The heights measured downwards are taken as negative.
4. *Mirror equation*:

   $$\frac{1}{v} + \frac{1}{u} = \frac{1}{f}$$

   where u and v are object and image distances, respectively and f is the focal length of the mirror. f is (approximately) half the radius of curvature R. f is negative for concave mirror; f is positive for a convex mirror.
5. For a prism of the angle A, of refractive index n_2 placed in a medium of refractive index n_1,

   $$n_{21} = \frac{n_2}{n_1} = \frac{\sin\left[\left(A + D_m\right)/2\right]}{\sin\left(A/2\right)}$$

   where D_m is the angle of minimum deviation.
6. *For refraction through a spherical interface* (from medium 1 to 2 of refractive index n_1 and n_2, respectively)

   $$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R}$$

   Thin lens formula

   $$\frac{1}{v} - \frac{1}{u} = \frac{1}{f}$$

Lens maker's formula

$$\frac{1}{f} = \frac{(n_2 - n_1)}{n_1}\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

R_1 and R_2 are the radii of curvature of the lens surfaces. f is positive for a converging lens; f is negative for a diverging lens. The power of a lens $P = 1/f$.

The SI unit for power of a lens is dioptre (D): 1 D = 1 m^{-1}.

If several thin lenses of focal length f_1, f_2, f_3,.. are in contact, the effective focal length of their combination, is given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} + \frac{1}{f_3} + \ldots$$

The total power of a combination of several lenses is

$P = P_1 + P_2 + P_3 + \ldots$

7. *Dispersion* is the splitting of light into its constituent colours.
8. *The Eye*: The eye has a convex lens of focal length about 2.5 cm. This focal length can be varied somewhat so that the image is always formed on the retina. This ability of the eye is called *accommodation*. In a defective eye, if the image is focussed before the retina (myopia), a diverging corrective lens is needed; if the image is focussed beyond the retina (hypermetropia), a converging corrective lens is needed. Astigmatism is corrected by using cylindrical lenses.
9. *Magnifying power m of a simple microscope* is given by $m = 1 + (D/f)$, where D = 25 cm is the least distance of distinct vision and f is the focal length of the convex lens. If the image is at infinity, $m = D/f$. For a compound microscope, the magnifying power is given by $m = m_e \times m_0$ where $m_e = 1 + (D/f_e)$, is the magnification due to the eyepiece and m_o is the magnification produced by the objective. *Approximately,*

$$m = \frac{L}{f_o} \times \frac{D}{f_e}$$

where f_o and f_e are the focal lengths of the objective and eyepiece, respectively, and L is the distance between their focal points.

10. *Magnifying power m of a telescope* is the ratio of the angle β subtended at the eye by the image to the angle α subtended at the eye by the object.

$$m = \frac{\beta}{\alpha} = \frac{f_o}{f_e}$$

where f_0 and f_e are the focal lengths of the objective and eyepiece, respectively.

POINTS TO PONDER

1. The laws of reflection and refraction are true for all surfaces and pairs of media at the point of the incidence.
2. The real image of an object placed between f and $2f$ from a convex lens can be seen on a screen placed at the image location. If the screen is removed, is the image still there? This question puzzles many, because it is difficult to reconcile ourselves with an image suspended in air

without a screen. But the image does exist. Rays from a given point on the object are converging to an image point in space and diverging away. The screen simply diffuses these rays, some of which reach our eye and we see the image. This can be seen by the images formed in air during a laser show.

3. Image formation needs regular reflection/refraction. In principle, all rays from a given point should reach the same image point. This is why you do not see your image by an irregular reflecting object, say the page of a book.
4. Thick lenses give coloured images due to dispersion. The variety in colour of objects we see around us is due to the constituent colours of the light incident on them. A monochromatic light may produce an entirely different perception about the colours on an object as seen in white light.
5. For a simple microscope, the angular size of the object equals the angular size of the image. Yet it offers magnification because we can keep the small object much closer to the eye than 25 cm and hence have it subtend a large angle. The image is at 25 cm which we can see. Without the microscope, you would need to keep the small object at 25 cm which would subtend a very small angle.

EXERCISES

9.1 A small candle, 2.5 cm in size is placed at 27 cm in front of a concave mirror of radius of curvature 36 cm. At what distance from the mirror should a screen be placed in order to obtain a sharp image? Describe the nature and size of the image. If the candle is moved closer to the mirror, how would the screen have to be moved?

9.2 A 4.5 cm needle is placed 12 cm away from a convex mirror of focal length 15 cm. Give the location of the image and the magnification. Describe what happens as the needle is moved farther from the mirror.

9.3 A tank is filled with water to a height of 12.5 cm. The apparent depth of a needle lying at the bottom of the tank is measured by a microscope to be 9.4 cm. What is the refractive index of water? If water is replaced by a liquid of refractive index 1.63 up to the same height, by what distance would the microscope have to be moved to focus on the needle again?

9.4 Figures 9.34(a) and (b) show refraction of a ray in air incident at 60° with the normal to a glass-air and water-air interface, respectively. Predict the angle of refraction in glass when the angle of incidence in water is 45° with the normal to a water-glass interface [Fig. 9.34(c)].

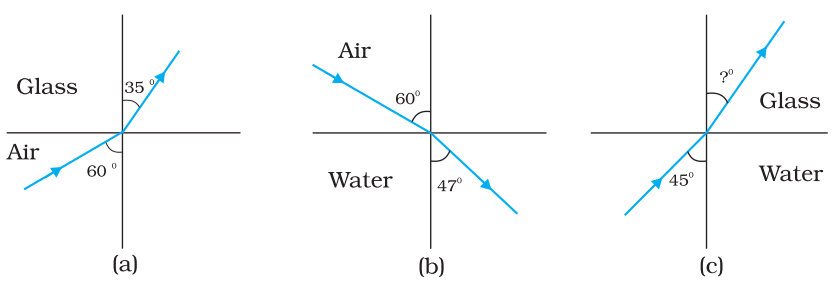


FIGURE 9.34

9.5 A small bulb is placed at the bottom of a tank containing water to a depth of 80 cm. What is the area of the surface of water through which light from the bulb can emerge out? Refractive index of water is 1.33. (Consider the bulb to be a point source.)

9.6 A prism is made of glass of unknown refractive index. A parallel beam of light is incident on a face of the prism. The angle of minimum deviation is measured to be 40°. What is the refractive index of the material of the prism? The refracting angle of the prism is 60°. If the prism is placed in water (refractive index 1.33), predict the new angle of minimum deviation of a parallel beam of light.

9.7 Double-convex lenses are to be manufactured from a glass of refractive index 1.55, with both faces of the same radius of curvature. What is the radius of curvature required if the focal length is to be 20 cm?

9.8 A beam of light converges at a point P. Now a lens is placed in the path of the convergent beam 12 cm from P. At what point does the beam converge if the lens is (a) a convex lens of focal length 20 cm, and (b) a concave lens of focal length 16 cm?

9.9 An object of size 3.0 cm is placed 14 cm in front of a concave lens of focal length 21 cm. Describe the image produced by the lens. What happens if the object is moved further away from the lens?

9.10 What is the focal length of a convex lens of focal length 30 cm in contact with a concave lens of focal length 20 cm? Is the system a converging or a diverging lens? Ignore thickness of the lenses.

9.11 A compound microscope consists of an objective lens of focal length 2.0 cm and an eyepiece of focal length 6.25 cm separated by a distance of 15 cm. How far from the objective should an object be placed in order to obtain the final image at (a) the least distance of distinct vision (25 cm), and (b) at infinity? What is the magnifying power of the microscope in each case?

9.12 A person with a normal near point (25 cm) using a compound microscope with objective of focal length 8.0 mm and an eyepiece of focal length 2.5 cm can bring an object placed at 9.0 mm from the objective in sharp focus. What is the separation between the two lenses? Calculate the magnifying power of the microscope,

9.13 A small telescope has an objective lens of focal length 144 cm and an eyepiece of focal length 6.0 cm. What is the magnifying power of the telescope? What is the separation between the objective and the eyepiece?

9.14 (a) A giant refracting telescope at an observatory has an objective lens of focal length 15 m. If an eyepiece of focal length 1.0 cm is used, what is the angular magnification of the telescope?

(b) If this telescope is used to view the moon, what is the diameter of the image of the moon formed by the objective lens? The diameter of the moon is 3.48×10^6 m, and the radius of lunar orbit is 3.8×10^8 m.

9.15 Use the mirror equation to deduce that:

(a) an object placed between f and $2f$ of a concave mirror produces a real image beyond $2f$.

(b) a convex mirror always produces a virtual image independent of the location of the object.

(c) the virtual image produced by a convex mirror is always diminished in size and is located between the focus and the pole.

(d) an object placed between the pole and focus of a concave mirror produces a virtual and enlarged image.

[*Note:* This exercise helps you deduce algebraically properties of images that one obtains from explicit ray diagrams.]

9.16 A small pin fixed on a table top is viewed from above from a distance of 50 cm. By what distance would the pin appear to be raised if it is viewed from the same point through a 15 cm thick glass slab held parallel to the table? Refractive index of glass = 1.5. Does the answer depend on the location of the slab?

9.17 (a) Figure 9.35 shows a cross-section of a 'light pipe' made of a glass fibre of refractive index 1.68. The outer covering of the pipe is made of a material of refractive index 1.44. What is the range of the angles of the incident rays with the axis of the pipe for which total reflections inside the pipe take place, as shown in the figure.

(b) What is the answer if there is no outer covering of the pipe?

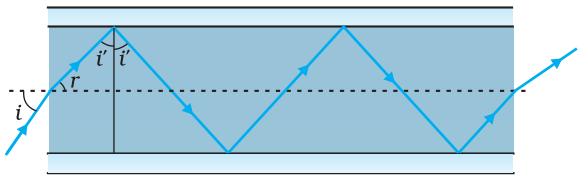


FIGURE 9.35

9.18 Answer the following questions:

(a) You have learnt that plane and convex mirrors produce virtual images of objects. Can they produce real images under some circumstances? Explain.

(b) A virtual image, we always say, cannot be caught on a screen. Yet when we 'see' a virtual image, we are obviously bringing it on to the 'screen' (i.e., the retina) of our eye. Is there a contradiction?

(c) A diver under water, looks obliquely at a fisherman standing on the bank of a lake. Would the fisherman look taller or shorter to the diver than what he actually is?

(d) Does the apparent depth of a tank of water change if viewed obliquely? If so, does the apparent depth increase or decrease?

(e) The refractive index of diamond is much greater than that of ordinary glass. Is this fact of some use to a diamond cutter?

9.19 The image of a small electric bulb fixed on the wall of a room is to be obtained on the opposite wall 3 m away by means of a large convex lens. What is the maximum possible focal length of the lens required for the purpose?

9.20 A screen is placed 90 cm from an object. The image of the object on the screen is formed by a convex lens at two different locations separated by 20 cm. Determine the focal length of the lens.

9.21 (a) Determine the 'effective focal length' of the combination of the two lenses in Exercise 9.10, if they are placed 8.0 cm apart with their principal axes coincident. Does the answer depend on which side of the combination a beam of parallel light is incident? Is the notion of effective focal length of this system useful at all?

(b) An object 1.5 cm in size is placed on the side of the convex lens in the arrangement (a) above. The distance between the object

and the convex lens is 40 cm. Determine the magnification produced by the two-lens system, and the size of the image.

9.22 At what angle should a ray of light be incident on the face of a prism of refracting angle 60° so that it just suffers total internal reflection at the other face? The refractive index of the material of the prism is 1.524.

9.23 You are given prisms made of crown glass and flint glass with a wide variety of angles. Suggest a combination of prisms which will

(a) deviate a pencil of white light without much dispersion,

(b) disperse (and displace) a pencil of white light without much deviation.

9.24 For a normal eye, the far point is at infinity and the near point of distinct vision is about 25 cm in front of the eye. The cornea of the eye provides a converging power of about 40 dioptres, and the least converging power of the eye-lens behind the cornea is about 20 dioptres. From this rough data estimate the range of accommodation (i.e., the range of converging power of the eye-lens) of a normal eye.

9.25 Does short-sightedness (myopia) or long-sightedness (hyper-metropia) imply necessarily that the eye has partially lost its ability of accommodation? If not, what might cause these defects of vision?

9.26 A myopic person has been using spectacles of power –1.0 dioptre for distant vision. During old age he also needs to use separate reading glass of power + 2.0 dioptres. Explain what may have happened.

9.27 A person looking at a person wearing a shirt with a pattern comprising vertical and horizontal lines is able to see the vertical lines more distinctly than the horizontal ones. What is this defect due to? How is such a defect of vision corrected?

9.28 A man with normal near point (25 cm) reads a book with small print using a magnifying glass: a thin convex lens of focal length 5 cm.

(a) What is the closest and the farthest distance at which he should keep the lens from the page so that he can read the book when viewing through the magnifying glass?

(b) What is the maximum and the minimum angular magnification (magnifying power) possible using the above simple microscope?

9.29 A card sheet divided into squares each of size 1 mm^2 is being viewed at a distance of 9 cm through a magnifying glass (a converging lens of focal length 9 cm) held close to the eye.

(a) What is the magnification produced by the lens? How much is the area of each square in the virtual image?

(b) What is the angular magnification (magnifying power) of the lens?

(c) Is the magnification in (a) equal to the magnifying power in (b)? Explain.

9.30 (a) At what distance should the lens be held from the figure in Exercise 9.29 in order to view the squares distinctly with the maximum possible magnifying power?

(b) What is the magnification in this case?

(c) Is the magnification equal to the magnifying power in this case? Explain.

9.31 What should be the distance between the object in Exercise 9.30 and the magnifying glass if the virtual image of each square in the figure is to have an area of 6.25 mm^2. Would you be able to see the squares distinctly with your eyes very close to the magnifier?

[*Note:* Exercises 9.29 to 9.31 will help you clearly understand the difference between magnification in absolute size and the angular magnification (or magnifying power) of an instrument.]

9.32 Answer the following questions:

(a) The angle subtended at the eye by an object is equal to the angle subtended at the eye by the virtual image produced by a magnifying glass. In what sense then does a magnifying glass provide angular magnification?

(b) In viewing through a magnifying glass, one usually positions one's eyes very close to the lens. Does angular magnification change if the eye is moved back?

(c) Magnifying power of a simple microscope is inversely proportional to the focal length of the lens. What then stops us from using a convex lens of smaller and smaller focal length and achieving greater and greater magnifying power?

(d) Why must both the objective and the eyepiece of a compound microscope have short focal lengths?

(e) When viewing through a compound microscope, our eyes should be positioned not on the eyepiece but a short distance away from it for best viewing. Why? How much should be that short distance between the eye and eyepiece?

9.33 An angular magnification (magnifying power) of 30X is desired using an objective of focal length 1.25 cm and an eyepiece of focal length 5 cm. How will you set up the compound microscope?

9.34 A small telescope has an objective lens of focal length 140 cm and an eyepiece of focal length 5.0 cm. What is the magnifying power of the telescope for viewing distant objects when

(a) the telescope is in normal adjustment (i.e., when the final image is at infinity)?

(b) the final image is formed at the least distance of distinct vision (25 cm)?

9.35 (a) For the telescope described in Exercise 9.34 (a), what is the separation between the objective lens and the eyepiece?

(b) If this telescope is used to view a 100 m tall tower 3 km away, what is the height of the image of the tower formed by the objective lens?

(c) What is the height of the final image of the tower if it is formed at 25 cm?

9.36 A Cassegrain telescope uses two mirrors as shown in Fig. 9.33. Such a telescope is built with the mirrors 20 mm apart. If the radius of curvature of the large mirror is 220 mm and the small mirror is 140 mm, where will the final image of an object at infinity be?

9.37 Light incident normally on a plane mirror attached to a galvanometer coil retraces backwards as shown in Fig. 9.36. A current in the coil produces a deflection of 3.5° of the mirror. What is the displacement of the reflected spot of light on a screen placed 1.5 m away?

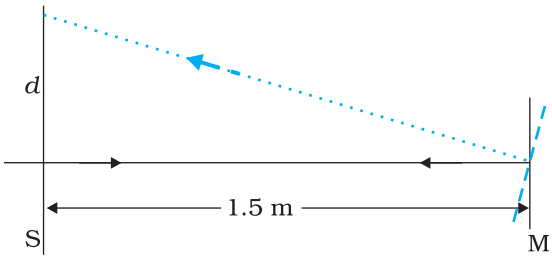


FIGURE 9.36

9.38 Figure 9.37 shows an equiconvex lens (of refractive index 1.50) in contact with a liquid layer on top of a plane mirror. A small needle with its tip on the principal axis is moved along the axis until its inverted image is found at the position of the needle. The distance of the needle from the lens is measured to be 45.0 cm. The liquid is removed and the experiment is repeated. The new distance is measured to be 30.0 cm. What is the refractive index of the liquid?

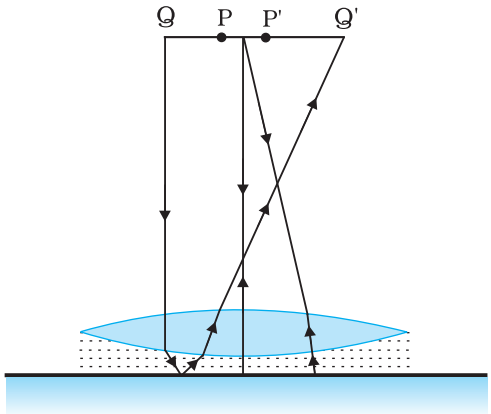


FIGURE 9.37

Chapter Ten

WAVE OPTICS

10.1 Introduction

In 1637 Descartes gave the corpuscular model of light and derived Snell's law. It explained the laws of reflection and refraction of light at an interface. The corpuscular model predicted that if the ray of light (on refraction) bends towards the normal then the speed of light would be greater in the second medium. This corpuscular model of light was further developed by Isaac Newton in his famous book entitled *OPTICKS* and because of the tremendous popularity of this book, the corpuscular model is very often attributed to Newton.

In 1678, the Dutch physicist Christiaan Huygens put forward the wave theory of light – it is this wave model of light that we will discuss in this chapter. As we will see, the wave model could satisfactorily explain the phenomena of reflection and refraction; however, it predicted that on refraction if the wave bends towards the normal then the speed of light would be less in the second medium. This is in contradiction to the prediction made by using the corpuscular model of light. It was much later confirmed by experiments where it was shown that the speed of light in water is less than the speed in air confirming the prediction of the wave model; Foucault carried out this experiment in 1850.

The wave theory was not readily accepted primarily because of Newton's authority and also because light could travel through vacuum

and it was felt that a wave would always require a medium to propagate from one point to the other. However, when Thomas Young performed his famous interference experiment in 1801, it was firmly established that light is indeed a wave phenomenon. The wavelength of visible light was measured and found to be extremely small; for example, the wavelength of yellow light is about 0.6 μm. Because of the smallness of the wavelength of visible light (in comparison to the dimensions of typical mirrors and lenses), light can be assumed to approximately travel in straight lines. This is the field of geometrical optics, which we had discussed in the previous chapter. Indeed, the branch of optics in which one completely neglects the finiteness of the wavelength is called geometrical optics and a ray is defined as the path of energy propagation in the limit of wavelength tending to zero.

After the interference experiment of Young in 1801, for the next 40 years or so, many experiments were carried out involving the interference and diffraction of lightwaves; these experiments could only be satisfactorily explained by assuming a wave model of light. Thus, around the middle of the nineteenth century, the wave theory seemed to be very well established. The only major difficulty was that since it was thought that a wave required a medium for its propagation, how could light waves propagate through vacuum. This was explained when Maxwell put forward his famous electromagnetic theory of light. Maxwell had developed a set of equations describing the laws of electricity and magnetism and using these equations he derived what is known as the wave equation from which he *predicted* the existence of electromagnetic waves*. From the wave equation, Maxwell could calculate the speed of electromagnetic waves in free space and he found that the theoretical value was very close to the measured value of speed of light. From this, he propounded that *light must be an electromagnetic wave*. Thus, according to Maxwell, light waves are associated with changing electric and magnetic fields; changing electric field produces a time and space varying magnetic field and a changing magnetic field produces a time and space varying electric field. The changing electric and magnetic fields result in the propagation of electromagnetic waves (or light waves) even in vacuum.

In this chapter we will first discuss the original formulation of the *Huygens principle* and derive the laws of reflection and refraction. In Sections 10.4 and 10.5, we will discuss the phenomenon of interference which is based on the principle of superposition. In Section 10.6 we will discuss the phenomenon of diffraction which is based on Huygens-Fresnel principle. Finally in Section 10.7 we will discuss the phenomenon of polarisation which is based on the fact that the light waves are *transverse electromagnetic waves*.

* Maxwell had predicted the existence of electromagnetic waves around 1855; it was much later (around 1890) that Heinrich Hertz produced radiowaves in the laboratory. J.C. Bose and G. Marconi made practical applications of the *Hertzian waves*

DOES LIGHT TRAVEL IN A STRAIGHT LINE?

Light travels in a straight line in Class VI; it does not do so in Class XII and beyond! Surprised, aren't you?

In school, you are shown an experiment in which you take three cardboards with pinholes in them, place a candle on one side and look from the other side. If the flame of the candle and the three pinholes are in a straight line, you can see the candle. Even if one of them is displaced a little, you cannot see the candle. *This proves*, so your teacher says, *that light travels in a straight line*.

In the present book, there are two consecutive chapters, one on ray optics and the other on wave optics. Ray optics is based on rectilinear propagation of light, and deals with mirrors, lenses, reflection, refraction, etc. Then you come to the chapter on wave optics, and you are told that light travels as a wave, that it can bend around objects, it can diffract and interfere, etc.

In optical region, light has a wavelength of about half a micrometre. If it encounters an obstacle of about this size, it can bend around it and can be seen on the other side. Thus a micrometre size obstacle will not be able to stop a light ray. If the obstacle is much larger, however, light will not be able to bend to that extent, and will not be seen on the other side.

This is a property of a wave in general, and can be seen in sound waves too. The sound wave of our speech has a wavelength of about 50 cm to 1 m. If it meets an obstacle of the size of a few metres, it bends around it and reaches points behind the obstacle. But when it comes across a larger obstacle of a few hundred metres, such as a hillock, most of it is reflected and is heard as an echo.

Then what about the primary school experiment? What happens there is that when we move any cardboard, the displacement is of the order of a few millimetres, which is much larger than the wavelength of light. Hence the candle cannot be seen. If we are able to move one of the cardboards by a micrometer or less, light will be able to diffract, and the candle will still be seen.

One could add to the first sentence in this box: *It learns how to bend as it grows up*!

10.2 HUYGENS PRINCIPLE

We would first define a wavefront: when we drop a small stone on a calm pool of water, waves spread out from the point of impact. Every point on the surface starts oscillating with time. At any instant, a photograph of the surface would show circular rings on which the disturbance is maximum. Clearly, all points on such a circle are oscillating in phase because they are at the same distance from the source. Such a locus of points, which oscillate in phase is called a *wavefront*; thus *a wavefront is defined as a surface of constant phase*. The speed with which the wavefront moves outwards from the source is called the speed of the wave. The energy of the wave travels in a direction perpendicular to the wavefront.

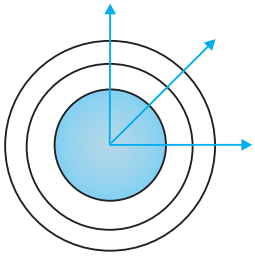


FIGURE 10.1 (a) A diverging spherical wave emanating from a point source. The wavefronts are spherical.

If we have a point source emitting waves uniformly in all directions, then the locus of points which have the same amplitude and vibrate in the same phase are spheres and we have what is known as a *spherical wave* as shown in Fig. 10.1(a). At a large distance from the source, a

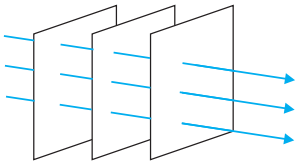


FIGURE 10.1 (b) At a large distance from the source, a small portion of the spherical wave can be approximated by a plane wave.

small portion of the sphere can be considered as a plane and we have what is known as a *plane wave* [Fig. 10.1(b)].

Now, if we know the shape of the wavefront at $t = 0$, then Huygens principle allows us to determine the shape of the wavefront at a later time τ. Thus, Huygens principle is essentially a geometrical construction, which given the shape of the wafefront at any time allows us to determine the shape of the wavefront at a later time. Let us consider a diverging wave and let F_1F_2 represent a portion of the spherical wavefront at $t = 0$ (Fig. 10.2). Now, according to Huygens principle, *each point of the wavefront is the source of a secondary disturbance and the wavelets emanating from these points spread out in all directions with the speed of the wave. These wavelets emanating from the wavefront are usually referred to as secondary wavelets and if we draw a common tangent to all these spheres, we obtain the new position of the wavefront at a later time.*

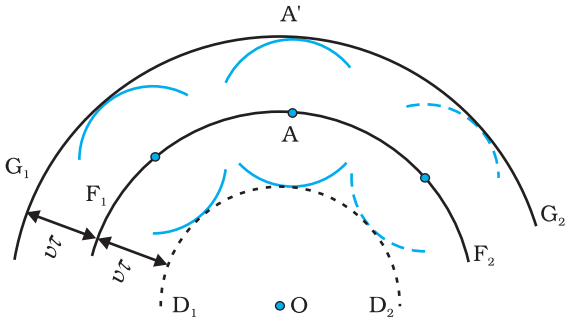


FIGURE 10.2 F_1F_2 represents the spherical wavefront (with O as centre) at $t = 0$. The envelope of the secondary wavelets emanating from F_1F_2 produces the forward moving wavefront G_1G_2. The backwave D_1D_2 does not exist.

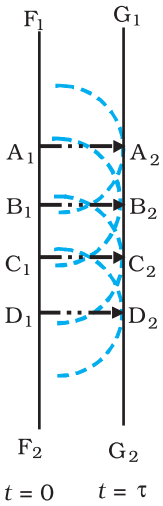


FIGURE 10.3 Huygens geometrical construction for a plane wave propagating to the right. F_1F_2 is the plane wavefront at $t = 0$ and G_1G_2 is the wavefront at a later time τ. The lines A_1A_2, B_1B_2 ... etc., are normal to both F_1F_2 and G_1G_2 and represent rays.

Thus, if we wish to determine the shape of the wavefront at $t = \tau$, we draw spheres of radius $v\tau$ from each point on the spherical wavefront where v represents the speed of the waves in the medium. If we now draw a common tangent to all these spheres, we obtain the new position of the wavefront at $t = \tau$. The new wavefront shown as G_1G_2 in Fig. 10.2 is again spherical with point O as the centre.

The above model has one shortcoming: we also have a backwave which is shown as D_1D_2 in Fig. 10.2. Huygens argued that the amplitude of the secondary wavelets is maximum in the forward direction and zero in the backward direction; by making this adhoc assumption, Huygens could explain the absence of the backwave. However, this adhoc assumption is not satisfactory and the absence of the backwave is really justified from more rigorous wave theory.

In a similar manner, we can use Huygens principle to determine the shape of the wavefront for a plane wave propagating through a medium (Fig. 10.3).

10.3 Refraction and Reflection of Plane Waves using Huygens Principle

10.3.1 Refraction of a plane wave

We will now use Huygens principle to derive the laws of refraction. Let PP′ represent the surface separating medium 1 and medium 2, as shown in Fig. 10.4. Let v_1 and v_2 represent the speed of light in medium 1 and medium 2, respectively. We assume a plane wavefront AB propagating in the direction A′A incident on the interface at an angle i as shown in the figure. Let τ be the time taken by the wavefront to travel the distance BC. Thus,

$$BC = v_1 \tau$$

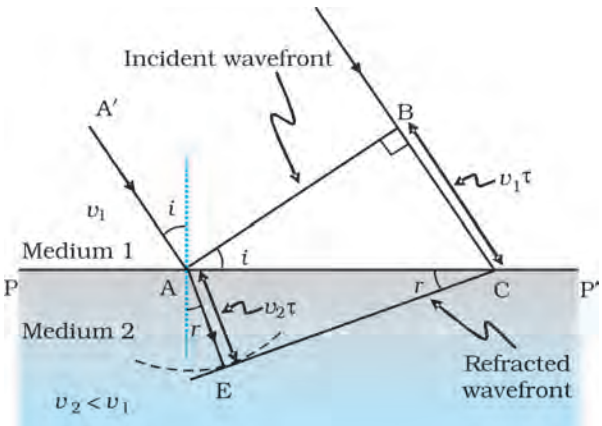


FIGURE 10.4 *A plane wave* AB is incident at an angle i on the surface PP′ separating medium 1 and medium 2. The plane wave undergoes refraction and CE represents the refracted wavefront. The figure corresponds to $v_2 < v_1$ so that the refracted waves bends towards the normal.

Christiaan Huygens (1629 – 1695) Dutch physicist, astronomer, mathematician and the founder of the wave theory of light. His book, *Treatise on light*, makes fascinating reading even today. He brilliantly explained the double refraction shown by the mineral calcite in this work in addition to reflection and refraction. He was the first to analyse circular and simple harmonic motion and designed and built improved clocks and telescopes. He discovered the true geometry of Saturn's rings.

CHRISTIAAN HUYGENS (1629 – 1695)

In order to determine the shape of the refracted wavefront, we draw a sphere of radius $v_2\tau$ from the point A in the second medium (the speed of the wave in the second medium is v_2). Let CE represent a tangent plane drawn from the point C on to the sphere. Then, AE = $v_2 \tau$ and CE would represent the refracted wavefront. If we now consider the triangles ABC and AEC, we readily obtain

$$\sin i = \frac{BC}{AC} = \frac{v_1\tau}{AC} \qquad (10.1)$$

and

$$\sin r = \frac{AE}{AC} = \frac{v_2\tau}{AC} \qquad (10.2)$$

where i and r are the angles of incidence and refraction, respectively.

Thus we obtain

$$\frac{\sin i}{\sin r} = \frac{v_1}{v_2} \tag{10.3}$$

From the above equation, we get the important result that if $r < i$ (i.e., if the ray bends toward the normal), the speed of the light wave in the second medium (v_2) will be less then the speed of the light wave in the first medium (v_1). This prediction is opposite to the prediction from the corpuscular model of light and as later experiments showed, the prediction of the wave theory is correct. Now, if c represents the speed of light in vacuum, then,

$$n_1 = \frac{c}{v_1} \tag{10.4}$$

and

$$n_2 = \frac{c}{v_2} \tag{10.5}$$

are known as the refractive indices of medium 1 and medium 2, respectively. In terms of the refractive indices, Eq. (10.3) can be written as

$$n_1 \sin i = n_2 \sin r \tag{10.6}$$

This is the *Snell's law of refraction*. Further, if λ_1 and λ_2 denote the wavelengths of light in medium 1 and medium 2, respectively and if the distance BC is equal to λ_1 then the distance AE will be equal to λ_2 (because if the crest from B has reached C in time τ, then the crest from A should have also reached E in time τ); thus,

$$\frac{\lambda_1}{\lambda_2} = \frac{\text{BC}}{\text{AE}} = \frac{v_1}{v_2}$$

or

$$\frac{v_1}{\lambda_1} = \frac{v_2}{\lambda_2} \tag{10.7}$$

The above equation implies that when a wave gets refracted into a denser medium ($v_1 > v_2$) the wavelength and the speed of propagation decrease but the *frequency* ν (= v/λ) *remains the same*.

10.3.2 Refraction at a rarer medium

We now consider refraction of a plane wave at a rarer medium, i.e., $v_2 > v_1$. Proceeding in an exactly similar manner we can construct a refracted wavefront as shown in Fig. 10.5. The angle of refraction will now be greater than angle of incidence; however, we will still have $n_1 \sin i = n_2 \sin r$. We define an angle i_c by the following equation

$$\sin i_c = \frac{n_2}{n_1} \tag{10.8}$$

Thus, if $i = i_c$ then $\sin r = 1$ and $r = 90°$. Obviously, for $i > i_c$, there can not be any refracted wave. The angle i_c is known as the *critical angle* and for all angles of incidence greater than the critical angle, we will not have

Demonstration of interference, diffraction, refraction, resonance and Doppler effect
http://www.falstad.com/ripple/

any refracted wave and the wave will undergo what is known as *total internal reflection*. The phenomenon of total internal reflection and its applications was discussed in Section 9.4.

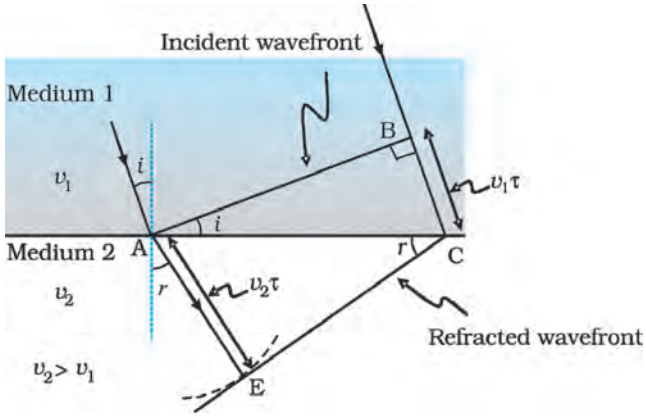


FIGURE 10.5 Refraction of a plane wave incident on a rarer medium for which $v_2 > v_1$. The plane wave bends away from the normal.

10.3.3 Reflection of a plane wave by a plane surface

We next consider a plane wave AB incident at an angle i on a reflecting surface MN. If v represents the speed of the wave in the medium and if τ represents the time taken by the wavefront to advance from the point B to C then the distance

$$BC = v\tau$$

In order to construct the reflected wavefront we draw a sphere of radius $v\tau$ from the point A as shown in Fig. 10.6. Let CE represent the tangent plane drawn from the point C to this sphere. Obviously

$$AE = BC = v\tau$$

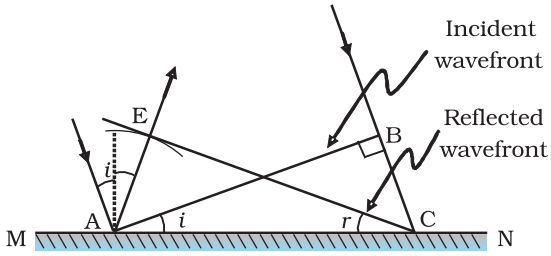


FIGURE 10.6 Reflection of a plane wave AB by the reflecting surface MN. AB and CE represent incident and reflected wavefronts.

If we now consider the triangles EAC and BAC we will find that they are congruent and therefore, the angles i and r (as shown in Fig. 10.6) would be equal. This is the *law of reflection*.

Once we have the laws of reflection and refraction, the behaviour of prisms, lenses, and mirrors can be understood. These phenomena were

discussed in detail in Chapter 9 on the basis of rectilinear propagation of light. Here we just describe the behaviour of the wavefronts as they undergo reflection or refraction. In Fig. 10.7(a) we consider a plane wave passing through a thin prism. Clearly, since the speed of light waves is less in glass, the lower portion of the incoming wavefront (which travels through the greatest thickness of glass) will get delayed resulting in a tilt in the emerging wavefront as shown in the figure. In Fig. 10.7(b) we consider a plane wave incident on a thin convex lens; the central part of the incident plane wave traverses the thickest portion of the lens and is delayed the most. The emerging wavefront has a depression at the centre and therefore the wavefront becomes spherical and converges to the point F which is known as the *focus*. In Fig. 10.7(c) a plane wave is incident on a concave mirror and on reflection we have a spherical wave converging to the focal point F. In a similar manner, we can understand refraction and reflection by concave lenses and convex mirrors.

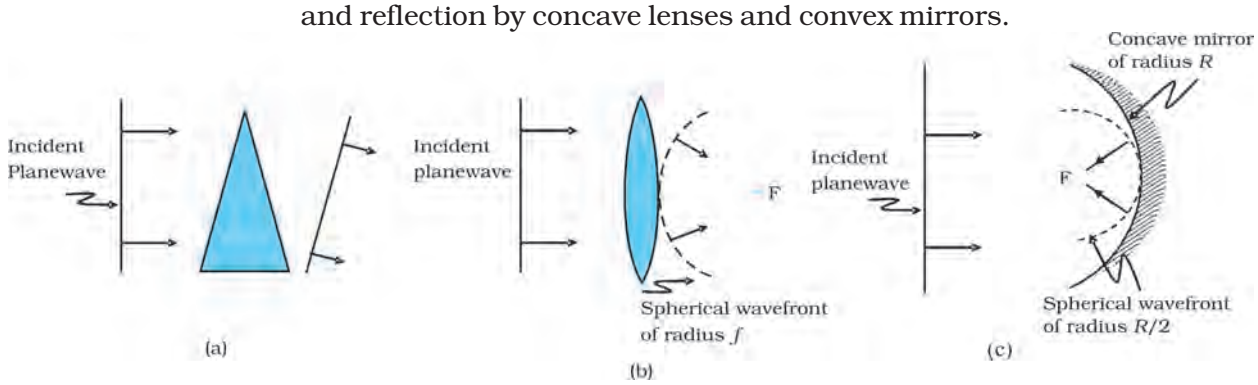


FIGURE 10.7 Refraction of a plane wave by (a) a thin prism, (b) a convex lens. (c) Reflection of a plane wave by a concave mirror.

From the above discussion it follows that the total time taken from a point on the object to the corresponding point on the image is the same measured along any ray. For example, when a convex lens focusses light to form a real image, although the ray going through the centre traverses a shorter path, but because of the slower speed in glass, the time taken is the same as for rays travelling near the edge of the lens.

10.3.4 The doppler effect

We should mention here that one should be careful in constructing the wavefronts if the source (or the observer) is moving. For example, if there is no medium and the source moves away from the observer, then later wavefronts have to travel a greater distance to reach the observer and hence take a longer time. The time taken between the arrival of two successive wavefronts is hence longer at the observer than it is at the source. Thus, when the source moves away from the observer the frequency as measured by the source will be smaller. This is known as the *Doppler effect*. Astronomers call the increase in wavelength due to doppler effect as *red shift* since a wavelength in the middle of the visible region of the spectrum moves towards the red end of the spectrum. When waves are received from a source moving towards the observer, there is an apparent decrease in wavelength, this is referred to as *blue shift*.

You have already encountered Doppler effect for sound waves in Chapter 15 of Class XI textbook. For velocities small compared to the speed of light, we can use the same formulae which we use for sound waves. The fractional change in frequency $\Delta\nu/\nu$ is given by $-v_{radial}/c$, where v_{radial} is the component of the source velocity along the line joining the observer to the source relative to the observer; v_{radial} is considered positive when the source moves away from the observer. Thus, the Doppler shift can be expressed as:

$$\frac{\Delta\nu}{\nu} = -\frac{v_{radial}}{c} \tag{10.9}$$

The formula given above is valid only when the speed of the source is small compared to that of light. A more accurate formula for the Doppler effect which is valid even when the speeds are close to that of light, requires the use of Einstein's special theory of relativity. The Doppler effect for light is very important in astronomy. It is the basis for the measurements of the radial velocities of distant galaxies.

Example 10.1 What speed should a galaxy move with respect to us so that the sodium line at 589.0 nm is observed at 589.6 nm?

Solution Since $\nu\lambda = c$, $\frac{\Delta\nu}{\nu} = -\frac{\Delta\lambda}{\lambda}$ (for small changes in ν and λ). For

$\Delta\lambda = 589.6 - 589.0 = +0.6$ nm

we get [using Eq. (10.9)]

$$\frac{\Delta\nu}{\nu} = -\frac{\Delta\lambda}{\lambda} = -\frac{v_{radial}}{c}$$

or, $$v_{radial} \cong +c\left(\frac{0.6}{589.0}\right) = +3.06\times10^{5}\,\text{m s}^{-1}$$

$$= 306 \text{ km/s}$$

Therefore, the galaxy is moving away from us.

EXAMPLE 10.1

Example 10.2

(a) When monochromatic light is incident on a surface separating two media, the reflected and refracted light both have the same frequency as the incident frequency. Explain why?

(b) When light travels from a rarer to a denser medium, the speed decreases. Does the reduction in speed imply a reduction in the energy carried by the light wave?

(c) In the wave picture of light, intensity of light is determined by the square of the amplitude of the wave. What determines the intensity of light in the photon picture of light.

Solution

(a) Reflection and refraction arise through interaction of incident light with the atomic constituents of matter. Atoms may be viewed as

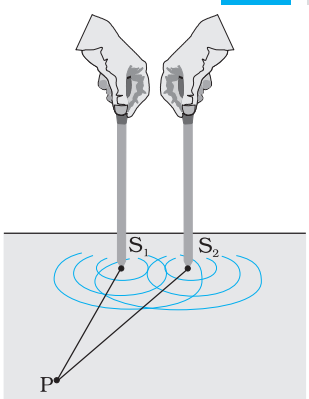
EXAMPLE 10.2

EXAMPLE 10.2

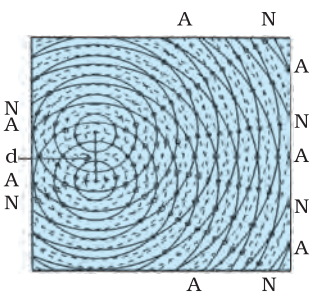
oscillators, which take up the frequency of the external agency (light) causing forced oscillations. The frequency of light emitted by a charged oscillator equals its frequency of oscillation. Thus, the frequency of scattered light equals the frequency of incident light.

(b) No. Energy carried by a wave depends on the amplitude of the wave, not on the speed of wave propagation.

(c) For a given frequency, intensity of light in the photon picture is determined by the number of photons crossing an unit area per unit time.



(a)



(b)

FIGURE 10.8 (a) Two needles oscillating in phase in water represent two coherent sources.
(b) The pattern of displacement of water molecules at an instant on the surface of water showing nodal N (no displacement) and antinodal A (maximum displacement) lines.

10.4 COHERENT AND INCOHERENT ADDITION OF WAVES

In this section we will discuss the interference pattern produced by the superposition of two waves. You may recall that we had discussed the superposition principle in Chapter 15 of your Class XI textbook. Indeed the entire field of interference is based on the *superposition principle* according to which *at a particular point in the medium, the resultant displacement produced by a number of waves is the vector sum of the displacements produced by each of the waves.*

Consider two needles S_1 and S_2 moving periodically up and down in an identical fashion in a trough of water [Fig. 10.8(a)]. They produce two water waves, and at a particular point, the phase difference between the displacements produced by each of the waves does not change with time; when this happens the two sources are said to be *coherent*. Figure 10.8(b) shows the position of crests (solid circles) and troughs (dashed circles) at a given instant of time. Consider a point P for which

$$S_1 P = S_2 P$$

Since the distances S_1 P and S_2 P are equal, waves from S_1 and S_2 will take the same time to travel to the point P and waves that emanate from S_1 and S_2 in phase will also arrive, at the point P, in phase.

Thus, if the displacement produced by the source S_1 at the point P is given by

$$y_1 = a \cos \omega t$$

then, the displacement produced by the source S_2 (at the point P) will also be given by

$$y_2 = a \cos \omega t$$

Thus, the resultant of displacement at P would be given by

$$y = y_1 + y_2 = 2\, a \cos \omega t$$

Since the intensity is proportional to the square of the amplitude, the resultant intensity will be given by

$$I = 4\, I_0$$

where I_0 represents the intensity produced by each one of the individual sources; I_0 is proportional to a^2. In fact at any point on the perpendicular bisector of S_1S_2, the intensity will be $4I_0$. The two sources are said to

interfere constructively and we have what is referred to as *constructive interference*. We next consider a point Q [Fig. 10.9(a)] for which

$$S_2Q - S_1Q = 2\lambda$$

The waves emanating from S_1 will arrive exactly two cycles earlier than the waves from S_2 and will again be in phase [Fig. 10.9(a)]. Thus, if the displacement produced by S_1 is given by

$$y_1 = a \cos \omega t$$

then the displacement produced by S_2 will be given by

$$y_2 = a \cos (\omega t - 4\pi) = a \cos \omega t$$

where we have used the fact that a path difference of 2λ corresponds to a phase difference of 4π. The two displacements are once again in phase and the intensity will again be $4\,I_0$ giving rise to constructive interference. In the above analysis we have assumed that the distances S_1Q and S_2Q are much greater than d (which represents the distance between S_1 and S_2) so that although S_1Q and S_2Q are not equal, the amplitudes of the displacement produced by each wave are very nearly the same.

We next consider a point R [Fig. 10.9(b)] for which

$$S_2R - S_1R = -2.5\lambda$$

The waves emanating from S_1 will arrive exactly two and a half cycles later than the waves from S_2 [Fig. 10.10(b)]. Thus if the displacement produced by S_1 is given by

$$y_1 = a \cos \omega t$$

then the displacement produced by S_2 will be given by

$$y_2 = a \cos (\omega t + 5\pi) = -\, a \cos \omega t$$

where we have used the fact that a path difference of 2.5λ corresponds to a phase difference of 5π. The two displacements are now out of phase and the two displacements will cancel out to give zero intensity. This is referred to as *destructive interference*.

To summarise: If we have two coherent sources S_1 and S_2 vibrating in phase, then for an arbitrary point P whenever the path difference,

$$S_1P \sim S_2P = n\lambda \quad (n = 0, 1, 2, 3, \ldots) \tag{10.10}$$

we will have constructive interference and the resultant intensity will be $4I_0$; the sign ~ between S_1P and S_2P represents the difference between S_1P and S_2P. On the other hand, if the point P is such that the path difference,

$$S_1P \sim S_2P = (n + \frac{1}{2})\,\lambda \quad (n = 0, 1, 2, 3, \ldots) \tag{10.11}$$

we will have *destructive interference* and the resultant intensity will be zero. Now, for any other arbitrary point G (Fig. 10.10) let the phase difference between the two displacements be ϕ. Thus, if the displacement produced by S_1 is given by

$$y_1 = a \cos \omega t$$

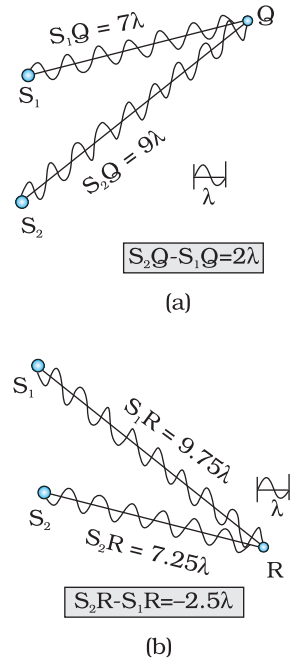


FIGURE 10.9 (a) Constructive interference at a point Q for which the path difference is 2λ. (b) Destructive interference at a point R for which the path difference is $2.5\,\lambda$.

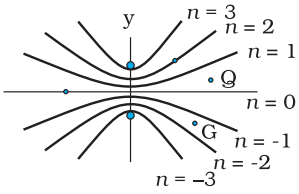


FIGURE 10.10 Locus of points for which $S_1P - S_2P$ is equal to zero, $\pm\lambda$, $\pm 2\lambda$, $\pm 3\lambda$.

then, the displacement produced by S_2 would be

$$y_2 = a \cos(\omega t + \phi)$$

and the resultant displacement will be given by

$$y = y_1 + y_2$$
$$= a\,[\cos \omega t + \cos(\omega t + \phi)]$$
$$= 2\,a \cos(\phi/2) \cos(\omega t + \phi/2)$$
$$\left[\because \cos A + \cos B = 2\cos\left(\frac{A+B}{2}\right)\cos\left(\frac{A-B}{2}\right)\right]$$

The amplitude of the resultant displacement is $2a \cos(\phi/2)$ and therefore the intensity at that point will be

$$I = 4\,I_0 \cos^2(\phi/2) \qquad (10.12)$$

If $\phi = 0, \pm 2\pi, \pm 4\pi, \ldots$ which corresponds to the condition given by Eq. (10.10) we will have constructive interference leading to maximum intensity. On the other hand, if $\phi = \pm\pi, \pm 3\pi, \pm 5\pi \ldots$ [which corresponds to the condition given by Eq. (10.11)] we will have destructive interference leading to zero intensity.

Now if the two sources are coherent (i.e., if the two needles are going up and down regularly) then the phase difference ϕ at any point will not change with time and we will have a stable interference pattern; i.e., the positions of maxima and minima will not change with time. However, if the two needles do not maintain a constant phase difference, then the interference pattern will also change with time and, if the phase difference changes very rapidly with time, the positions of maxima and minima will also vary rapidly with time and we will see a "time-averaged" intensity distribution. When this happens, we will observe an average intensity that will be given by

$$< I >= 4I_0 < \cos^2(\phi/2) > \qquad (10.13)$$

where angular brackets represent time averaging. Indeed it is shown in Section 7.2 that if $\phi(t)$ varies randomly with time, the time-averaged quantity $< \cos^2(\phi/2) >$ will be 1/2. This is also intuitively obvious because the function $\cos^2(\phi/2)$ will randomly vary between 0 and 1 and the average value will be 1/2. The resultant intensity will be given by

$$I = 2\,I_0 \qquad (10.14)$$

at all points.

When the phase difference between the two vibrating sources changes rapidly with time, we say that the two sources are incoherent and when this happens the intensities just add up. This is indeed what happens when two separate light sources illuminate a wall.

10.5 Interference of Light Waves and Young's Experiment

We will now discuss interference using light waves. If we use two sodium lamps illuminating two pinholes (Fig. 10.11) we will not observe any interference fringes. This is because of the fact that the light wave emitted from an ordinary source (like a sodium lamp) undergoes abrupt phase

Ripple Tank experiments on wave interference
http://phet.colorado.edu/en/simulation/legacy/wave-interference
PHYSICS

changes in times of the order of 10^{-10} seconds. Thus the light waves coming out from two independent sources of light will not have any fixed phase relationship and would be incoherent, when this happens, as discussed in the previous section, the intensities on the screen will add up.

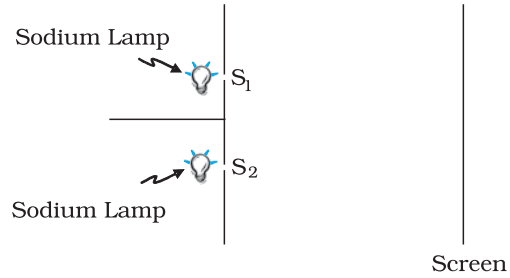
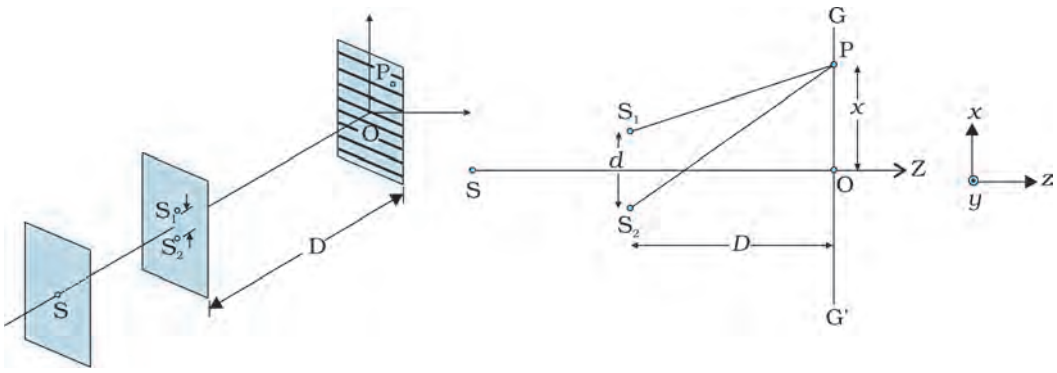


FIGURE 10.11 If two sodium lamps illuminate two pinholes S_1 and S_2, the intensities will add up and no interference fringes will be observed on the screen.

The British physicist Thomas Young used an ingenious technique to "lock" the phases of the waves emanating from S_1 and S_2. He made two pinholes S_1 and S_2 (very close to each other) on an opaque screen [Fig. 10.12(a)]. These were illuminated by another pinholes that was in turn, lit by a bright source. Light waves spread out from S and fall on both S_1 and S_2. S_1 and S_2 then behave like two coherent sources because light waves coming out from S_1 and S_2 are derived from the same original source and any abrupt phase change in S will manifest in exactly similar phase changes in the light coming out from S_1 and S_2. Thus, the two sources S_1 and S_2 will be *locked* in phase; i.e., they will be coherent like the two vibrating needle in our water wave example [Fig. 10.8(a)].



(a) (b)

FIGURE 10.12 Young's arrangement to produce interference pattern.

Thus spherical waves emanating from S_1 and S_2 will produce interference fringes on the screen GG′, as shown in Fig. 10.12(b). The positions of maximum and minimum intensities can be calculated by using the analysis given in Section 10.4 where we had shown that for an arbitrary point P on the line GG′ [Fig. 10.12(b)] to correspond to a maximum, we must have

$$S_2P - S_1P = n\lambda; \quad n = 0, 1, 2 \ldots \tag{10.15}$$

Now,

$$(S_2P)^2 - (S_1P)^2 = \left[D^2 + \left(x + \frac{d}{2}\right)^2\right] - \left[D^2 + \left(x - \frac{d}{2}\right)^2\right] = 2xd$$

THOMAS YOUNG (1773 – 1829)

Thomas Young (1773 – 1829) English physicist, physician and Egyptologist. Young worked on a wide variety of scientific problems, ranging from the structure of the eye and the mechanism of vision to the decipherment of the Rosetta stone. He revived the wave theory of light and recognised that interference phenomena provide proof of the wave properties of light.

where $S_1S_2 = d$ and OP = x. Thus

$$S_2P - S_1P = \frac{2xd}{S_2P+S_1P} \tag{10.16}$$

If x, $d<<D$ then negligible error will be introduced if $S_2P + S_1P$ (in the denominator) is replaced by $2D$. For example, for d = 0.1 cm, D = 100 cm, OP = 1 cm (which correspond to typical values for an interference experiment using light waves), we have

$$S_2P + S_1P = [(100)^2 + (1.05)^2]^{1/2} + [(100)^2 + (0.95)^2]^{1/2}$$
$$\approx 200.01 \text{ cm}$$

Thus if we replace $S_2P + S_1P$ by $2D$, the error involved is about 0.005%. In this approximation, Eq. (10.16) becomes

$$S_2P - S_1P \approx \frac{xd}{D} \tag{10.17}$$

Hence we will have constructive interference resulting in a bright region when $\frac{xd}{D} = n\lambda$ [Eq. (10.15)]. That is,

$$x = x_n = \frac{n\lambda D}{d}; \; n = 0, \pm 1, \pm 2, \ldots \tag{10.18}$$

On the other hand, we will have destructive interference resulting in a dark region when $\frac{xd}{D} = (n+\frac{1}{2})\lambda$ that is

$$x = x_n = (n+\frac{1}{2})\frac{\lambda D}{d}; \quad n = 0, \pm 1, \pm 2 \tag{10.19}$$

Thus dark and bright bands appear on the screen, as shown in Fig. 10.13. Such bands are called *fringes*. Equations (10.18) and (10.19) show that dark and bright fringes are equally spaced and the distance between two consecutive bright and dark fringes is given by

$$\beta = x_{n+1} - x_n \text{ or } \beta = \frac{\lambda D}{d} \tag{10.20}$$

which is the expression for the *fringe width*. Obviously, the central point O (in Fig. 10.12) will be bright because $S_1O = S_2O$ and it will correspond to $n = 0$ [Eq. (10.18)]. If we consider the line perpendicular to the plane of the paper and passing through O [i.e., along the y-axis] then all points on this line will be equidistant from S_1 and S_2 and we will have a bright central fringe which is a straight line as shown in Fig. 10.13. In order to determine the shape of the interference pattern on the screen we note that a particular fringe would correspond to the locus of points with a constant value of $S_2P - S_1P$. Whenever this constant is an integral multiple of λ, the fringe will be bright and whenever it is an odd integral multiple of $\lambda/2$ it will be a dark fringe. Now, the locus of the point P lying in the x-y plane such that $S_2P - S_1P$ (= Δ) is a constant, is a hyperbola. Thus the fringe pattern will strictly be a hyperbola; however, if the distance D is very large compared to the fringe width, the fringes will be very nearly straight lines as shown in Fig. 10.13.

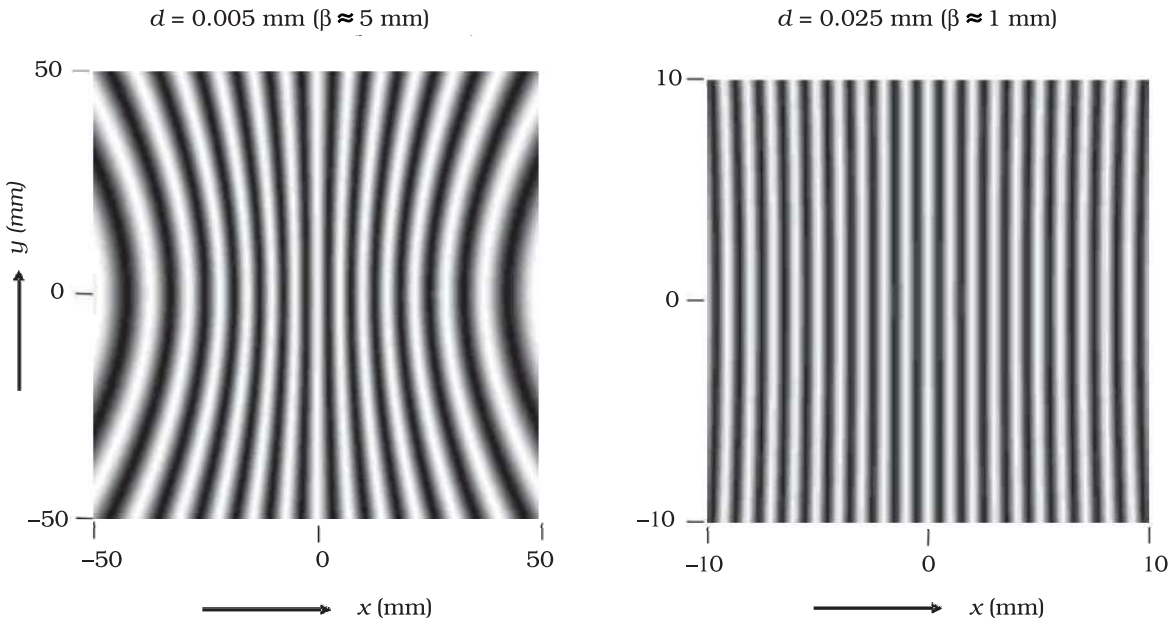
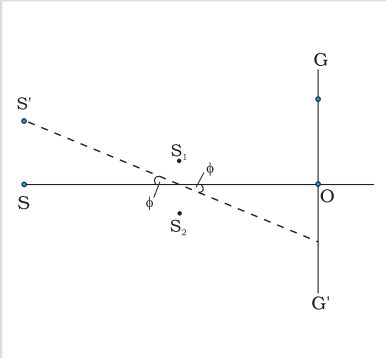


FIGURE 10.13 Computer generated fringe pattern produced by two point source S_1 and S_2 on the screen GG′ (Fig. 10.12); (a) and (b) correspond to d = 0.005 mm and 0.025 mm, respectively (both figures correspond to D = 5 cm and $\lambda = 5 \times 10^{-5}$ cm.) (Adopted from OPTICS by A. Ghatak, Tata McGraw Hill Publishing Co. Ltd., New Delhi, 2000.)

In the double-slit experiment shown in Fig. 10.12(b), we have taken the source hole S on the perpendicular bisector of the two slits, which is shown as the line SO. What happens if the source S is slightly away from the perpendicular bisector. Consider that the source is moved to some new point S′ and suppose that Q is the mid-point of S_1 and S_2. If the angle S′QS is ϕ, then the central bright fringe occurs at an angle $-\phi$, on the other side. Thus, if the source S is on the perpendicular bisector, then the central fringe occurs at O, also on the perpendicular bisector. If S is shifted by an angle ϕ to point S′, then the central fringe appears at a point O' at an angle $-\phi$, which means that it is shifted by the same angle on the other side of the bisector. This also means that the source S′, the mid-point Q and the point O′ of the central fringe are in a straight line.



We end this section by quoting from the Nobel lecture of Dennis Gabor*

The wave nature of light was demonstrated convincingly for the first time in 1801 by Thomas Young by a wonderfully simple experiment. He let a ray of sunlight into a dark room, placed a dark screen in front of it, pierced with two small pinholes, and beyond this, at some distance, a white screen. He then saw two darkish lines at both sides of a bright line, which gave him sufficient encouragement to repeat the experiment, this time with spirit flame as light source, with a little salt in it to produce the bright yellow sodium light. This time he saw a number of dark lines, regularly spaced; the first clear proof that light added to light can produce darkness. This phenomenon is called

* Dennis Gabor received the 1971 Nobel Prize in Physics for discovering the principles of holography.

interference. Thomas Young had expected it because he believed in the wave theory of light.

We should mention here that the fringes are straight lines although S_1 and S_2 are point sources. If we had slits instead of the point sources (Fig. 10.14), each pair of points would have produced straight line fringes resulting in straight line fringes with increased intensities.

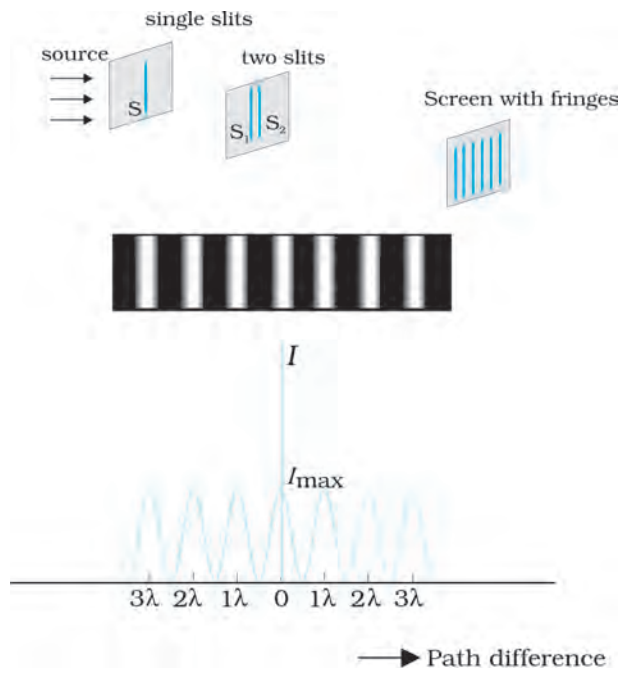


FIGURE 10.14 Photograph and the graph of the intensity distribution in Young's double-slit experiment.

Interactive animation of Young's experiment
http://vsg.quasihome.com/interfer.html

PHYSICS

EXAMPLE 10.3

Example 10.3 Two slits are made one millimetre apart and the screen is placed one metre away. What is the fringe separation when blue-green light of wavelength 500 nm is used?

Solution Fringe spacing $= \dfrac{D\lambda}{d} = \dfrac{1\times5\times10^{-7}}{1\times10^{-3}}$ m

$= 5 \times 10^{-4}$ m = 0.5 mm

EXAMPLE 10.4

Example 10.4 What is the effect on the interference fringes in a Young's double-slit experiment due to each of the following operations:

(a) the screen is moved away from the plane of the slits;
(b) the (monochromatic) source is replaced by another (monochromatic) source of shorter wavelength;
(c) the separation between the two slits is increased;
(d) the source slit is moved closer to the double-slit plane;
(e) the width of the source slit is increased;
(f) the monochromatic source is replaced by a source of white light?

(In each operation, take all parameters, other than the one specified, to remain unchanged.)

Solution

(a) Angular separation of the fringes remains constant (= λ/d). The actual separation of the fringes increases in proportion to the distance of the screen from the plane of the two slits.

(b) The separation of the fringes (and also angular separation) decreases. See, however, the condition mentioned in (d) below.

(c) The separation of the fringes (and also angular separation) decreases. See, however, the condition mentioned in (d) below.

(d) Let s be the size of the source and S its distance from the plane of the two slits. For interference fringes to be seen, the condition $s/S < \lambda/d$ should be satisfied; otherwise, interference patterns produced by different parts of the source overlap and no fringes are seen. Thus, as S decreases (i.e., the source slit is brought closer), the interference pattern gets less and less sharp, and when the source is brought too close for this condition to be valid, the fringes disappear. Till this happens, the fringe separation remains fixed.

(e) Same as in (d). As the source slit width increases, fringe pattern gets less and less sharp. When the source slit is so wide that the condition $s/S \leq \lambda/d$ is not satisfied, the interference pattern disappears.

(f) The interference patterns due to different component colours of white light overlap (incoherently). The central bright fringes for different colours are at the same position. Therefore, the central fringe is white. For a point P for which $S_2P - S_1P = \lambda_b/2$, where λ_b ($\approx$ 4000 Å) represents the wavelength for the blue colour, the blue component will be absent and the fringe will appear red in colour. Slightly farther away where $S_2Q - S_1Q = \lambda_b = \lambda_r/2$ where λ_r ($\approx$ 8000 Å) is the wavelength for the red colour, the fringe will be predominantly blue.

Thus, the fringe closest on either side of the central white fringe is red and the farthest will appear blue. After a few fringes, no clear fringe pattern is seen.

EXAMPLE 10.4

10.6 DIFFRACTION

If we look clearly at the shadow cast by an opaque object, close to the region of geometrical shadow, there are alternate dark and bright regions just like in interference. This happens due to the phenomenon of diffraction. Diffraction is a general characteristic exhibited by all types of waves, be it sound waves, light waves, water waves or matter waves. Since the wavelength of light is much smaller than the dimensions of most obstacles; we do not encounter diffraction effects of light in everyday observations. However, the finite resolution of our eye or of optical

instruments such as telescopes or microscopes is limited due to the phenomenon of diffraction. Indeed the colours that you see when a CD is viewed is due to diffraction effects. We will now discuss the phenomenon of diffraction.

10.6.1 The single slit

In the discussion of Young's experiment, we stated that a single narrow slit acts as a new source from which light spreads out. Even before Young, early experimenters – including Newton – had noticed that light spreads out from narrow holes and slits. It seems to turn around corners and enter regions where we would expect a shadow. These effects, known as *diffraction*, can only be properly understood using wave ideas. After all, you are hardly surprised to hear sound waves from someone talking around a corner!

When the double slit in Young's experiment is replaced by a single narrow slit (illuminated by a monochromatic source), a broad pattern with a central bright region is seen. On both sides, there are alternate dark and bright regions, the intensity becoming weaker away from the centre (Fig. 10.16). To understand this, go to Fig. 10.15, which shows a parallel beam of light falling normally on a single slit LN of width a. The diffracted light goes on to meet a screen. The midpoint of the slit is M.

A straight line through M perpendicular to the slit plane meets the screen at C. We want the intensity at any point P on the screen. As before, straight lines joining P to the different points L,M,N, etc., can be treated as parallel, making an angle θ with the normal MC.

The basic idea is to divide the slit into much smaller parts, and add their contributions at P with the proper phase differences. We are treating different parts of the wavefront at the slit as secondary sources. Because the incoming wavefront is parallel to the plane of the slit, these sources are in phase.

The path difference NP – LP between the two edges of the slit can be calculated exactly as for Young's experiment. From Fig. 10.15,

$$\begin{aligned} NP - LP &= NQ \\ &= a \sin\theta \\ &\approx a\,\theta \text{ (for smaller angles)} \end{aligned} \tag{10.21}$$

Similarly, if two points M_1 and M_2 in the slit plane are separated by y, the path difference $M_2P - M_1P \approx y\theta$. We now have to sum up equal, coherent contributions from a large number of sources, each with a different phase. This calculation was made by Fresnel using integral calculus, so we omit it here. The main features of the diffraction pattern can be understood by simple arguments.

At the central point C on the screen, the angle θ is zero. All path differences are zero and hence all the parts of the slit contribute in phase. This gives maximum intensity at C. Experimental observation shown in

Fig. 10.15 indicates that the intensity has a central maximum at $\theta = 0$ and other secondary maxima at $\theta \approx (n+1/2)\ \lambda/a$, and has minima (zero intensity) at $\theta \approx n\lambda/a$, $n = \pm1, \pm2, \pm3, \ldots$. It is easy to see why it has minima at these values of angle. Consider first the angle θ where the path difference $a\theta$ is λ. Then,

$$\theta \approx \lambda / a. \qquad (10.22)$$

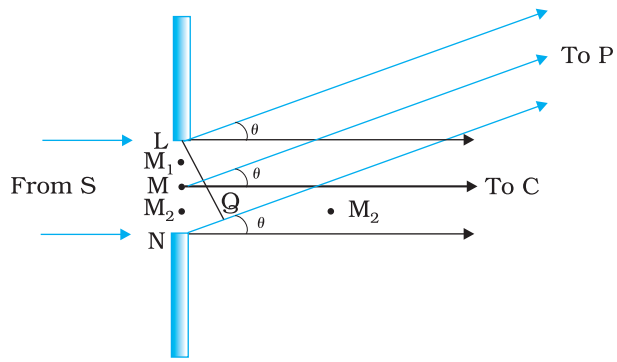


FIGURE 10.15 The geometry of path differences for diffraction by a single slit.

Now, divide the slit into two equal halves LM and MN each of size $a/2$. For every point M_1 in LM, there is a point M_2 in MN such that $M_1M_2 = a/2$. The path difference between M_1 and M_2 at $P = M_2P - M_1P = \theta a/2 = \lambda/2$ for the angle chosen. This means that the contributions from M_1 and M_2 are 180° out of phase and cancel in the direction $\theta = \lambda/a$. Contributions from the two halves of the slit LM and MN, therefore, cancel each other. Equation (10.22) gives the angle at which the intensity falls to zero. One can similarly show that the intensity is zero for $\theta = n\lambda/a$, with n being any integer (except zero!). Notice that the angular size of the central maximum increases when the slit width a decreases.

It is also easy to see why there are maxima at $\theta = (n + 1/2)\ \lambda/a$ and why they go on becoming weaker and weaker with increasing n. Consider an angle $\theta = 3\lambda/2a$ which is midway between two of the dark fringes. Divide the slit into three equal parts. If we take the first two thirds of the slit, the path difference between the two ends would be

$$\frac{2}{3}a \times \theta = \frac{2a}{3} \times \frac{3\lambda}{2a} = \lambda \qquad (10.23)$$

The first two-thirds of the slit can therefore be divided into two halves which have a $\lambda/2$ path difference. The contributions of these two halves cancel in the same manner as described earlier. Only the remaining one-third of the slit contributes to the intensity at a point between the two minima. Clearly, this will be much weaker than the central maximum (where the entire slit contributes in phase). One can similarly show that there are maxima at $(n + 1/2)\ \lambda/a$ with $n = 2, 3$, etc. These become weaker with increasing n, since only one-fifth, one-seventh, etc., of the slit contributes in these cases. The photograph and intensity pattern corresponding to it is shown in Fig. 10.16.

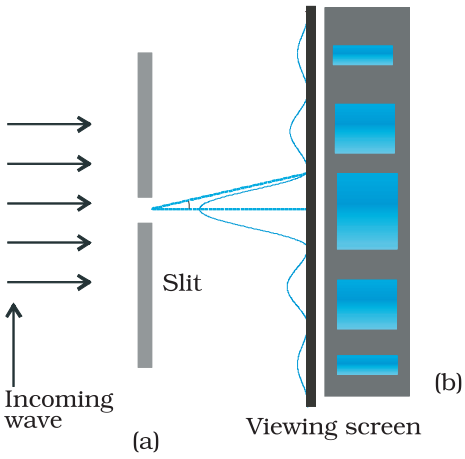


FIGURE 10.16 Intensity distribution and photograph of fringes due to diffraction at single slit.

There has been prolonged discussion about difference between intereference and diffraction among scientists since the discovery of these phenomena. In this context, it is

interesting to note what Richard Feynman* has said in his famous Feynman Lectures on Physics:

> *No one has ever been able to define the difference between interference and diffraction satisfactorily. It is just a question of usage, and there is no specific, important physical difference between them. The best we can do is, roughly speaking, is to say that when there are only a few sources, say two interfering sources, then the result is usually called interference, but if there is a large number of them, it seems that the word diffraction is more often used.*

In the double-slit experiment, we must note that the pattern on the screen is actually a superposition of single-slit diffraction from each slit or hole, and the double-slit interference pattern. This is shown in Fig. 10.17. It shows a broader diffraction peak in which there appear several fringes of smaller width due to double-slit interference. The number of interference fringes occuring in the broad diffraction peak depends on the ratio d/a, that is the ratio of the distance between the two slits to the width of a slit. In the limit of a becoming very small, the diffraction pattern will become very flat and we will obsrve the two-slit interference pattern [see Fig. 10.13(b)].

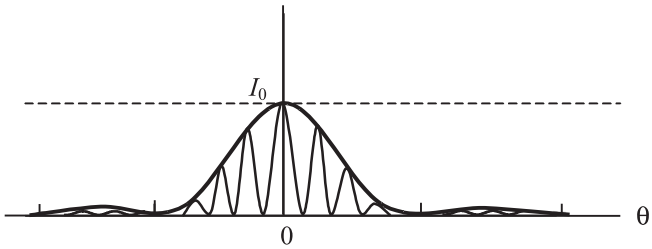


FIGURE 10.17 The actual double-slit interference pattern. The envelope shows the single slit diffraction.

Interactive animation on single slit diffraction pattern
http://www.phys.hawaii.edu/~teb/optics/java/slitdiffr/

PHYSICS

EXAMPLE 10.5

Example 10.5 In Example 10.3, what should the width of each slit be to obtain 10 maxima of the double slit pattern within the central maximum of the single slit pattern?

Solution We want $a\theta = \lambda, \theta = \dfrac{\lambda}{a}$

$$10\frac{\lambda}{d} = 2\frac{\lambda}{a} \quad a = \frac{d}{5} = 0\ 2 \text{ mm}$$

Notice that the wavelength of light and distance of the screen do not enter in the calculation of a.

In the double-slit interference experiment of Fig. 10.12, what happens if we close one slit? You will see that it now amounts to a single slit. But you will have to take care of some shift in the pattern. We now have a source at S, and only one hole (or slit) S_1 or S_2. This will produce a single-

* Richand Feynman was one of the recipients of the 1965 Nobel Prize in Physics for his fundamental work in quantum electrodynamics.

slit diffraction pattern on the screen. The centre of the central bright fringe will appear at a point which lies on the straight line SS_1 or SS_2, as the case may be.

We now compare and contrast the interference pattern with that seen for a coherently illuminated single slit (usually called the single slit diffraction pattern).

(i) The interference pattern has a number of equally spaced bright and dark bands. The diffraction pattern has a central bright maximum which is twice as wide as the other maxima. The intensity falls as we go to successive maxima away from the centre, on either side.

(ii) We calculate the interference pattern by superposing two waves originating from the two narrow slits. The diffraction pattern is a superposition of a continuous family of waves originating from each point on a single slit.

(iii) For a single slit of width a, the first null of the interference pattern occurs at an angle of λ/a. At the same angle of λ/a, we get a maximum (not a null) for two narrow slits separated by a distance a.

One must understand that both d and a have to be quite small, to be able to observe good interference and diffraction patterns. For example, the separation d between the two slits must be of the order of a milimetre or so. The width a of each slit must be even smaller, of the order of 0.1 or 0.2 mm.

In our discussion of Young's experiment and the single-slit diffraction, we have assumed that the screen on which the fringes are formed is at a large distance. The two or more paths from the slits to the screen were treated as parallel. This situation also occurs when we place a converging lens after the slits and place the screen at the focus. Parallel paths from the slit are combined at a single point on the screen. *Note that the lens does not introduce any extra path differences in a parallel beam.* This arrangement is often used since it gives more intensity than placing the screen far away. If f is the focal length of the lens, then we can easily work out the size of the central bright maximum. In terms of angles, the separation of the central maximum from the first null of the diffraction pattern is λ/a. Hence, the size on the screen will be $f\lambda/a$.

10.6.2 Seeing the single slit diffraction pattern

It is surprisingly easy to see the single-slit diffraction pattern for oneself. The equipment needed can be found in most homes — two razor blades and one clear glass electric bulb preferably with a straight filament. One has to hold the two blades so that the edges are parallel and have a narrow slit in between. This is easily done with the thumb and forefingers (Fig. 10.18).

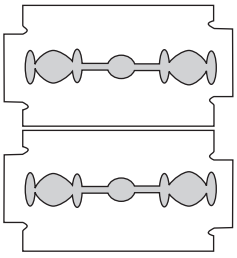


FIGURE 10.18
Holding two blades to form a single slit. A bulb filament viewed through this shows clear diffraction bands.

Keep the slit parallel to the filament, right in front of the eye. Use spectacles if you normally do. With slight adjustment of the width of the slit and the parallelism of the edges, the pattern should be seen with its bright and dark bands. Since the position of all the bands (except the central one) depends on wavelength, they will show some colours. Using a filter for red or blue will make the fringes clearer. With both filters available, the wider fringes for red compared to blue can be seen.

In this experiment, the filament plays the role of the first slit S in Fig. 10.16. The lens of the eye focuses the pattern on the screen (the retina of the eye).

With some effort, one can cut a double slit in an aluminium foil with a blade. The bulb filament can be viewed as before to repeat Young's experiment. In daytime, there is another suitable bright source subtending a small angle at the eye. This is the reflection of the Sun in any shiny convex surface (e.g., a cycle bell). Do not try direct sunlight – it can damage the eye and will not give fringes anyway as the Sun subtends an angle of (1/2)°.

In interference and diffraction, light energy is redistributed. If it reduces in one region, producing a dark fringe, it increases in another region, producing a bright fringe. There is no gain or loss of energy, which is consistent with the principle of conservation of energy.

10.6.3 Resolving power of optical instruments

In Chapter 9 we had discussed about telescopes. The angular resolution of the telescope is determined by the objective of the telescope. The stars which are not resolved in the image produced by the objective cannot be resolved by any further magnification produced by the eyepiece. The primary purpose of the eyepiece is to provide magnification of the image produced by the objective.

Consider a parallel beam of light falling on a convex lens. If the lens is well corrected for aberrations, then geometrical optics tells us that the beam will get focused to a point. However, because of diffraction, the beam instead of getting focused to a point gets focused to a spot of finite area. In this case the effects due to diffraction can be taken into account by considering a plane wave incident on a circular aperture followed by a convex lens (Fig. 10.19). The analysis of the corresponding diffraction pattern is quite involved; however, in principle, it is similar to the analysis carried out to obtain the single-slit diffraction pattern. Taking into account the effects due to diffraction, the pattern on the focal plane would consist of a central bright region surrounded by concentric dark and bright rings (Fig. 10.19). A detailed analysis shows that the radius of the central bright region is approximately given by

$$r_0 \approx \frac{1.22\,\lambda f}{2a} = \frac{0.61\lambda f}{a} \tag{10.24}$$

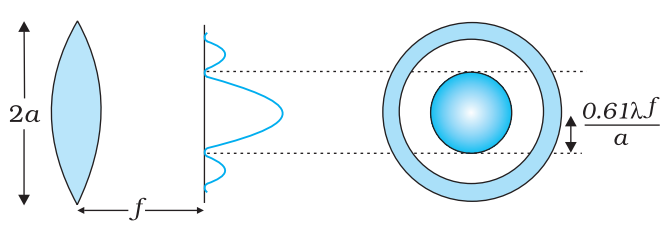


FIGURE 10.19 A parallel beam of light is incident on a convex lens. Because of diffraction effects, the beam gets focused to a spot of radius ≈ 0.61 $\lambda f/a$.

where f is the focal length of the lens and $2a$ is the diameter of the circular aperture or the diameter of the lens, whichever is smaller. Typically if

$$\lambda \approx 0.5\ \mu\text{m},\ f \approx 20\ \text{cm}\ \text{ and }\ a \approx 5\ \text{cm}$$

we have

$$r_0 \approx 1.2\ \mu\text{m}$$

Although the size of the spot is very small, it plays an important role in determining the limit of resolution of optical instruments like a telescope or a microscope. For the two stars to be just resolved

$$f\Delta\theta \approx r_0 \approx \frac{0.61\lambda f}{a}$$

implying

$$\Delta\theta \approx \frac{0.61\lambda}{a} \tag{10.25}$$

Thus $\Delta\theta$ will be small if the diameter of the objective is large. This implies that the telescope will have better resolving power if a is large. It is for this reason that for better resolution, a telescope must have a large diameter objective.

Example 10.6 Assume that light of wavelength 6000Å is coming from a star. What is the limit of resolution of a telescope whose objective has a diameter of 100 inch?

Solution A 100 inch telescope implies that $2a$ = 100 inch = 254 cm. Thus if,

$$\lambda \approx 6000\text{Å} = 6\times10^{-5}\ \text{cm}$$

then

$$\Delta\theta \approx \frac{0.61\times6\times10^{-5}}{127} \approx 2.9\times10^{-7}\ \text{radians}$$

EXAMPLE 10.6

We can apply a similar argument to the objective lens of a microscope. In this case, the object is placed slightly beyond f, so that a real image is formed at a distance v [Fig. 10.20]. The magnification (ratio of image size to object size) is given by $m \approx v/f$. It can be seen from Fig. 10.20 that

$$D/f \approx 2\tan\beta \tag{10.26}$$

where 2β is the angle subtended by the diameter of the objective lens at the focus of the microscope.

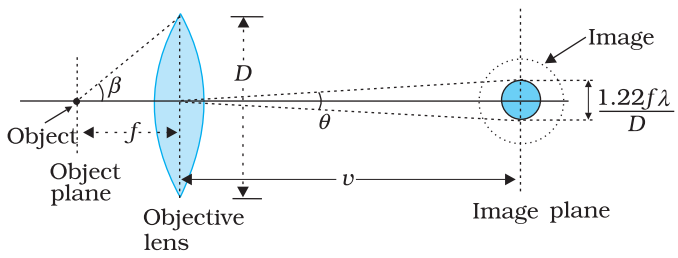


FIGURE 10.20 Real image formed by the objective lens of the microscope.

Determine the resolving power of your eye

You can estimate the resolving power of your eye with a simple experiment. Make black stripes of equal width separated by white stripes; see figure here. All the black stripes should be of equal width, while the width of the intermediate white stripes should increase as you go from the left to the right. For example, let all black stripes have a width of 5 mm. Let the width of the first two white stripes be 0.5 mm each, the next two white stripes be 1 mm each, the next two 1.5 mm each, etc. Paste this pattern on a wall in a room or laboratory, at the height of your eye.

Now watch the pattern, preferably with one eye. By moving away or closer to the wall, find the position where you can just see some two black stripes as separate stripes. All the black stripes to the left of this stripe would merge into one another and would not be distinguishable. On the other hand, the black stripes to the right of this would be more and more clearly visible. Note the width d of the white stripe which separates the two regions, and measure the distance D of the wall from your eye. Then d/D is the resolution of your eye.

You have watched specks of dust floating in air in a sunbeam entering through your window. Find the distance (of a speck) which you can clearly see and distinguish from a neighbouring speck. Knowing the resolution of your eye and the distance of the speck, estimate the size of the speck of dust.

When the separation between two points in a microscopic specimen is comparable to the wavelength λ of the light, the diffraction effects become important. The image of a point object will again be a diffraction pattern whose size in the image plane will be

$$v\,\theta = v\left(\frac{1.22\,\lambda}{D}\right) \tag{10.27}$$

Two objects whose images are closer than this distance will not be resolved, they will be seen as one. The corresponding minimum separation, d_{min}, in the object plane is given by

$$d_{min} = \left[v\left(\frac{1.22\,\lambda}{D}\right)\right]\Big/m$$

$$= \frac{1.22\,\lambda}{D}\cdot\frac{v}{m}$$

or, since $m = \dfrac{v}{f}$

$$= \frac{1.22\,f\,\lambda}{D} \tag{10.28}$$

Now, combining Eqs. (10.26) and (10.28), we get

$$d_{min} = \frac{1.22\,\lambda}{2\tan\beta}$$

$$\simeq \frac{1.22\,\lambda}{2\sin\beta} \tag{10.29}$$

If the medium between the object and the objective lens is not air but a medium of refractive index n, Eq. (10.29) gets modified to

$$d_{\min} = \frac{1.22\,\lambda}{2\,n\sin\beta} \tag{10.30}$$

The product $n\sin\beta$ is called the *numerical aperture* and is sometimes marked on the objective.

The resolving power of the microscope is given by the reciprocal of the minimum separation of two points seen as distinct. It can be seen from Eq. (10.30) that the resolving power can be increased by choosing a medium of higher refractive index. Usually an oil having a refractive index close to that of the objective glass is used. Such an arrangement is called an *'oil immersion objective'*. Notice that it is not possible to make $\sin\beta$ larger than unity. Thus, we see that the resolving power of a microscope is basically determined by the wavelength of the light used.

There is a likelihood of confusion between resolution and magnification, and similarly between the role of a telescope and a microscope to deal with these parameters. A telescope produces images of far objects nearer to our eye. Therefore objects which are not resolved at far distance, can be resolved by looking at them through a telescope. A microscope, on the other hand, magnifies objects (which are near to us) and produces their larger image. We may be looking at two stars or two satellites of a far-away planet, or we may be looking at different regions of a living cell. In this context, it is good to remember that a telescope resolves whereas a microscope magnifies.

10.6.4 The validity of ray optics

An aperture (i.e., slit or hole) of size a illuminated by a parallel beam sends diffracted light into an angle of approximately $\approx \lambda/a$. This is the angular size of the bright central maximum. In travelling a distance z, the diffracted beam therefore acquires a width $z\lambda/a$ due to diffraction. It is interesting to ask at what value of z the spreading due to diffraction becomes comparable to the size a of the aperture. We thus approximately equate $z\lambda/a$ with a. This gives the distance beyond which divergence of the beam of width a becomes significant. Therefore,

$$z \simeq \frac{a^2}{\lambda} \tag{10.31}$$

We define a quantity z_F called the *Fresnel distance* by the following equation

$$z_F \simeq a^2/\lambda$$

Equation (10.31) shows that for distances much smaller than z_F, the spreading due to diffraction is smaller compared to the size of the beam. It becomes comparable when the distance is approximately z_F. For distances much greater than z_F, the spreading due to diffraction

dominates over that due to ray optics (i.e., the size a of the aperture). Equation (10.31) also shows that ray optics is valid in the limit of wavelength tending to zero.

EXAMPLE 10.7

Example 10.7 For what distance is ray optics a good approximation when the aperture is 3 mm wide and the wavelength is 500 nm?

Solution $z_F = \frac{a^2}{\lambda} = \frac{\left(3\times10^{-3}\right)^2}{5\times10^{-7}} = 18 \text{ m}$

This example shows that even with a small aperture, diffraction spreading can be neglected for rays many metres in length. Thus, ray optics is valid in many common situations.

10.7 POLARISATION

Consider holding a long string that is held horizontally, the other end of which is assumed to be fixed. If we move the end of the string up and down in a periodic manner, we will generate a wave propagating in the $+x$ direction (Fig. 10.21). Such a wave could be described by the following equation

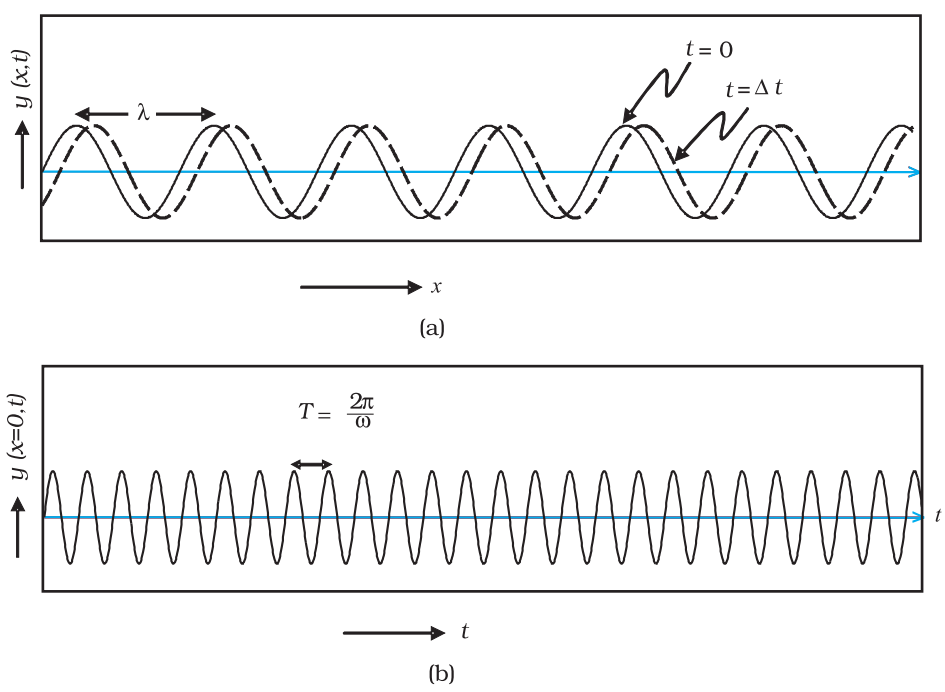


FIGURE 10.21 (a) The curves represent the displacement of a string at $t = 0$ and at $t = \Delta t$, respectively when a sinusoidal wave is propagating in the $+x$-direction. (b) The curve represents the time variation of the displacement at $x = 0$ when a sinusoidal wave is propagating in the $+x$-direction. At $x = \Delta x$, the time variation of the displacement will be slightly displaced to the right.

$$y\,(x,t) = a \sin\,(kx - \omega t) \tag{10.32}$$

where a and $\omega (= 2\pi\nu)$ represent the amplitude and the angular frequency of the wave, respectively; further,

$$\lambda = \frac{2\pi}{k} \tag{10.33}$$

represents the wavelength associated with the wave. We had discussed propagation of such waves in Chapter 15 of Class XI textbook. Since the displacement (which is along the y direction) is at right angles to the direction of propagation of the wave, we have what is known as a *transverse wave*. Also, since the displacement is in the y direction, it is often referred to as a y-polarised wave. Since each point on the string moves on a straight line, the wave is also referred to as a linearly polarised wave. Further, the string always remains confined to the x-y plane and therefore it is also referred to as a *plane polarised wave*.

In a similar manner we can consider the vibration of the string in the x-z plane generating a z-polarised wave whose displacement will be given by

$$z\,(x,t) = a \sin\,(kx - \omega t) \tag{10.34}$$

It should be mentioned that the linearly polarised waves [described by Eqs. (10.33) and (10.34)] are all transverse waves; i.e., the displacement of each point of the string is always at right angles to the direction of propagation of the wave. Finally, if the plane of vibration of the string is changed randomly in very short intervals of time, then we have what is known as an *unpolarised wave*. Thus, for an unpolarised wave the displacement will be randomly changing with time though it will always be perpendicular to the direction of propagation.

Light waves are transverse in nature; i.e., the electric field associated with a propagating light wave is always at right angles to the direction of propagation of the wave. This can be easily demonstrated using a simple polaroid. You must have seen thin plastic like sheets, which are called *polaroids*. A polaroid consists of long chain molecules aligned in a particular direction. The electric vectors (associated with the propagating light wave) along the direction of the aligned molecules get absorbed. Thus, if an unpolarised light wave is incident on such a polaroid then the light wave will get linearly polarised with the electric vector oscillating along a direction perpendicular to the aligned molecules; this direction is known as the *pass-axis* of the polaroid.

Thus, if the light from an ordinary source (like a sodium lamp) passes through a polaroid sheet P_1, it is observed that its intensity is reduced by half. Rotating P_1 has no effect on the transmitted beam and transmitted intensity remains constant. Now, let an identical piece of polaroid P_2 be placed before P_1. As expected, the light from the lamp is reduced in intensity on passing through P_2 alone. But now rotating P_1 has a dramatic effect on the light coming from P_2. In one position, the intensity transmitted

by P_2 followed by P_1 is nearly zero. When turned by 90° from this position, P_1 transmits nearly the full intensity emerging from P_2 (Fig. 10.22).

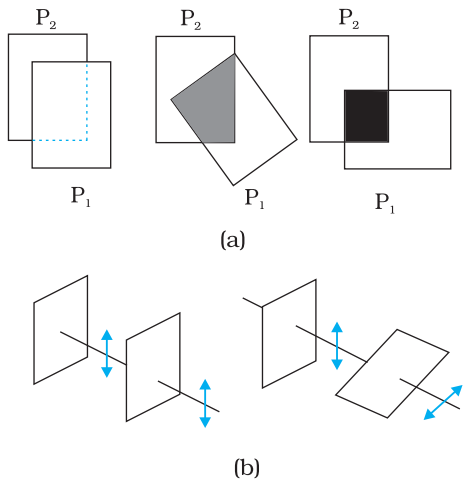


FIGURE 10.22 (a) Passage of light through two polaroids P_2 and P_1. The transmitted fraction falls from 1 to 0 as the angle between them varies from 0° to 90°. Notice that the light seen through a single polaroid P_1 does not vary with angle. (b) Behaviour of the electric vector when light passes through two polaroids. The transmitted polarisation is the component parallel to the polaroid axis. The double arrows show the oscillations of the electric vector.

The above experiment can be easily understood by assuming that light passing through the polaroid P_2 gets polarised along the pass-axis of P_2. If the pass-axis of P_2 makes an angle θ with the pass-axis of P_1, then when the polarised beam passes through the polaroid P_2, the component $E \cos \theta$ (along the pass-axis of P_2) will pass through P_2. Thus, as we rotate the polaroid P_1 (or P_2), the intensity will vary as:

$$I = I_0 \cos^2\theta \qquad (10.35)$$

where I_0 is the intensity of the polarized light after passing through P_1. This is known as *Malus' law*. The above discussion shows that the intensity coming out of a single polaroid is half of the incident intensity. By putting a second polaroid, the intensity can be further controlled from 50% to zero of the incident intensity by adjusting the angle between the pass-axes of two polaroids.

Polaroids can be used to control the intensity, in sunglasses, windowpanes, etc. Polaroids are also used in photographic cameras and 3D movie cameras.

EXAMPLE 10.8

Example 10.8 Discuss the intensity of transmitted light when a polaroid sheet is rotated between two crossed polaroids?

Solution Let I_0 be the intensity of polarised light after passing through the first polariser P_1. Then the intensity of light after passing through second polariser P_2 will be

$$I = I_0 \cos^2\theta,$$

EXAMPLE 10.8

where θ is the angle between pass axes of P_1 and P_2. Since P_1 and P_3 are crossed the angle between the pass axes of P_2 and P_3 will be $(\pi/2-\theta)$. Hence the intensity of light emerging from P_3 will be

$$I = I_0 \cos^2\theta \cos^2\left(\frac{\pi}{2}-\theta\right)$$

$$= I_0 \cos^2\theta \ \sin^2\theta = (I_0/4) \sin^2 2\theta$$

Therefore, the transmitted intensity will be maximum when $\theta = \pi/4$.

10.7.1 Polarisation by scattering

The light from a clear blue portion of the sky shows a rise and fall of intensity when viewed through a polaroid which is rotated. This is nothing but sunlight, which has changed its direction (having been scattered) on encountering the molecules of the earth's atmosphere. As Fig. 10.23(a) shows, the incident sunlight is unpolarised. The dots stand for polarisation perpendicular to the plane of the figure. The double arrows show polarisation in the plane of the figure. (There is no phase relation between these two in unpolarised light). Under the influence of the electric field of the incident wave the electrons in the molecules acquire components of motion in both these directions. We have drawn an observer looking at 90° to the direction of the sun. Clearly, charges accelerating parallel to the double arrows do not radiate energy towards this observer since their acceleration has no transverse component. The radiation scattered by the molecule is therefore represented by dots. It is polarised perpendicular to the plane of the figure. This explains the polarisation of scattered light from the sky.

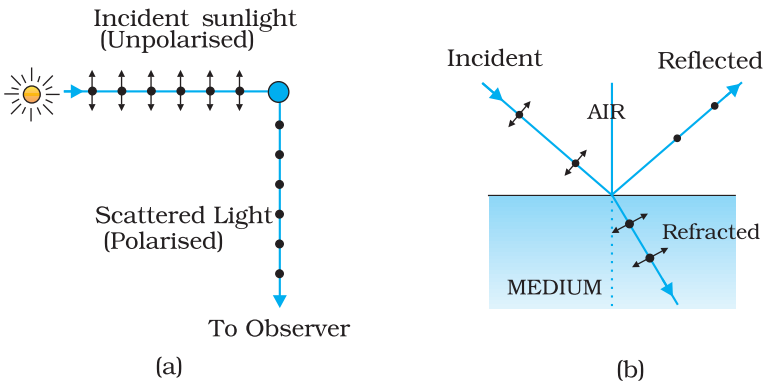
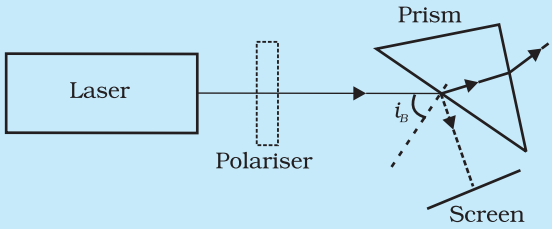


FIGURE 10.23 (a) Polarisation of the blue scattered light from the sky. The incident sunlight is unpolarised (dots and arrows). A typical molecule is shown. It scatters light by 90° polarised normal to the plane of the paper (dots only). (b) Polarisation of light reflected from a transparent medium at the Brewster angle (reflected ray perpendicular to refracted ray).

The scattering of light by molecules was intensively investigated by C.V. Raman and his collaborators in Kolkata in the 1920s. Raman was awarded the Nobel Prize for Physics in 1930 for this work.

A SPECIAL CASE OF TOTAL TRANSMISSION

When light is incident on an interface of two media, it is observed that some part of it gets reflected and some part gets transmitted. Consider a related question: *Is it possible that under some conditions a monochromatic beam of light incident on a surface (which is normally reflective) gets completely transmitted with no reflection?* To your surprise, the answer is *yes*.



Let us try a simple experiment and check what happens. Arrange a laser, a good polariser, a prism and screen as shown in the figure here.

Let the light emitted by the laser source pass through the polariser and be incident on the surface of the prism at the Brewster's angle of incidence i_B. Now rotate the polariser carefully and you will observe that for a specific alignment of the polariser, the light incident on the prism is completely transmitted and no light is reflected from the surface of the prism. The reflected spot will completely vanish.

10.7.2 Polarisation by reflection

Figure 10.23(b) shows light reflected from a transparent medium, say, water. As before, the dots and arrows indicate that both polarisations are present in the incident and refracted waves. We have drawn a situation in which the reflected wave travels at right angles to the refracted wave. The oscillating electrons in the water produce the reflected wave. These move in the two directions transverse to the radiation from wave in the medium, i.e., the *refracted wave*. The arrows are parallel to the direction of the *reflected* wave. Motion in this direction does not contribute to the reflected wave. As the figure shows, the reflected light is therefore linearly polarised perpendicular to the plane of the figure (represented by dots). This can be checked by looking at the reflected light through an analyser. The transmitted intensity will be zero when the axis of the analyser is in the plane of the figure, i.e., the plane of incidence.

When unpolarised light is incident on the boundary between two transparent media, the reflected light is polarised with its electric vector perpendicular to the plane of incidence when the refracted and reflected rays make a right angle with each other. Thus we have seen that when reflected wave is perpendicular to the refracted wave, the reflected wave is a totally polarised wave. The angle of incidence in this case is called *Brewster's angle* and is denoted by i_B. We can see that i_B is related to the refractive index of the denser medium. Since we have $i_B + r = \pi/2$, we get from Snell's law

$$\mu = \frac{\sin i_B}{\sin r} = \frac{\sin i_B}{\sin(\pi/2 - i_B)}$$

$$= \frac{\sin i_B}{\cos i_B} = \tan i_B \qquad (10.36)$$

This is known as *Brewster's law*.

Example 10.9 Unpolarised light is incident on a plane glass surface. What should be the angle of incidence so that the reflected and refracted rays are perpendicular to each other?

Solution For $i + r$ to be equal to $\pi/2$, we should have $\tan i_B = \mu = 1.5$. This gives $i_B = 57°$. This is the Brewster's angle for air to glass interface.

EXAMPLE 10.9

For simplicity, we have discussed scattering of light by 90°, and reflection at the Brewster angle. In this special situation, one of the two perpendicular components of the electric field is zero. At other angles, both components are present but one is stronger than the other. There is no stable phase relationship between the two perpendicular components since these are derived from two perpendicular components of an unpolarised beam. When such light is viewed through a rotating analyser, one sees a maximum and a minimum of intensity but not complete darkness. This kind of light is called *partially polarised*.

Let us try to understand the situation. When an unpolarised beam of light is incident at the Brewster's angle on an interface of two media, only part of light with electric field vector perpendicular to the plane of incidence will be reflected. Now by using a good polariser, if we completely remove all the light with its electric vector perpendicular to the plane of incidence and let this light be incident on the surface of the prism at Brewster's angle, you will then observe no reflection and there will be total transmission of light.

We began this chapter by pointing out that there are some phenomena which can be explained only by the wave theory. In order to develop a proper understanding, we first described how some phenomena like reflection and refraction, which were studied on this basis of Ray Optics in Chapter 9, can also be understood on the basis of Wave Optics. Then we described Young's double slit experiment which was a turning point in the study of optics. Finally, we described some associated points such as diffraction, resolution, polarisation, and validity of ray optics. In the next chapter, you will see how new experiments led to new theories at the turn of the century around 1900 A.D.

SUMMARY

1. Huygens' principle tells us that each point on a wavefront is a source of secondary waves, which add up to give the wavefront at a later time.
2. Huygens' construction tells us that the new wavefront is the forward envelope of the secondary waves. When the speed of light is independent of direction, the secondary waves are spherical. The rays are then perpendicular to both the wavefronts and the time of travel

is the same measured along any ray. This principle leads to the well known laws of reflection and refraction.

3. The principle of superposition of waves applies whenever two or more sources of light illuminate the same point. When we consider the intensity of light due to these sources at the given point, there is an interference term in addition to the sum of the individual intensities. But this term is important only if it has a non-zero average, which occurs only if the sources have the same frequency and a stable phase difference.
4. Young's double slit of separation d gives equally spaced fringes of angular separation λ/d. The source, mid-point of the slits, and central bright fringe lie in a straight line. An extended source will destroy the fringes if it subtends angle more than λ/d at the slits.
5. A single slit of width a gives a diffraction pattern with a central maximum. The intensity falls to zero at angles of $\pm\frac{\lambda}{a}, \pm\frac{2\lambda}{a}$, etc., with successively weaker secondary maxima in between. Diffraction limits the angular resolution of a telescope to λ/D where D is the diameter. Two stars closer than this give strongly overlapping images. Similarly, a microscope objective subtending angle 2β at the focus, in a medium of refractive index n, will just separate two objects spaced at a distance $\lambda/(2n \sin \beta)$, which is the resolution limit of a microscope. Diffraction determines the limitations of the concept of light rays. A beam of width a travels a distance a^2/λ, called the Fresnel distance, before it starts to spread out due to diffraction.
6. Natural light, e.g., from the sun is unpolarised. This means the electric vector takes all possible directions in the transverse plane, rapidly and randomly, during a measurement. A polaroid transmits only one component (parallel to a special axis). The resulting light is called linearly polarised or plane polarised. When this kind of light is viewed through a second polaroid whose axis turns through 2π, two maxima and minima of intensity are seen. Polarised light can also be produced by reflection at a special angle (called the Brewster angle) and by scattering through $\pi/2$ in the earth's atmosphere.

POINTS TO PONDER

1. Waves from a point source spread out in all directions, while light was seen to travel along narrow rays. It required the insight and experiment of Huygens, Young and Fresnel to understand how a wave theory could explain all aspects of the behaviour of light.
2. The crucial new feature of waves is interference of amplitudes from different sources which can be both constructive and destructive, as shown in Young's experiment.
3. Diffraction phenomena define the limits of ray optics. The limit of the ability of microscopes and telescopes to distinguish very close objects is set by the wavelength of light.
4. Most interference and diffraction effects exist even for longitudinal waves like sound in air. But polarisation phenomena are special to transverse waves like light waves.

EXERCISES

10.1 Monochromatic light of wavelength 589 nm is incident from air on a water surface. What are the wavelength, frequency and speed of (a) reflected, and (b) refracted light? Refractive index of water is 1.33.

10.2 What is the shape of the wavefront in each of the following cases:

(a) Light diverging from a point source.

(b) Light emerging out of a convex lens when a point source is placed at its focus.

(c) The portion of the wavefront of light from a distant star intercepted by the Earth.

10.3 (a) The refractive index of glass is 1.5. What is the speed of light in glass? (Speed of light in vacuum is 3.0×10^8 m s^{-1})

(b) Is the speed of light in glass independent of the colour of light? If not, which of the two colours red and violet travels slower in a glass prism?

10.4 In a Young's double-slit experiment, the slits are separated by 0.28 mm and the screen is placed 1.4 m away. The distance between the central bright fringe and the fourth bright fringe is measured to be 1.2 cm. Determine the wavelength of light used in the experiment.

10.5 In Young's double-slit experiment using monochromatic light of wavelength λ, the intensity of light at a point on the screen where path difference is λ, is K units. What is the intensity of light at a point where path difference is $\lambda/3$?

10.6 A beam of light consisting of two wavelengths, 650 nm and 520 nm, is used to obtain interference fringes in a Young's double-slit experiment.

(a) Find the distance of the third bright fringe on the screen from the central maximum for wavelength 650 nm.

(b) What is the least distance from the central maximum where the bright fringes due to both the wavelengths coincide?

10.7 In a double-slit experiment the angular width of a fringe is found to be 0.2° on a screen placed 1 m away. The wavelength of light used is 600 nm. What will be the angular width of the fringe if the entire experimental apparatus is immersed in water? Take refractive index of water to be 4/3.

10.8 What is the Brewster angle for air to glass transition? (Refractive index of glass = 1.5.)

10.9 Light of wavelength 5000 Å falls on a plane reflecting surface. What are the wavelength and frequency of the reflected light? For what angle of incidence is the reflected ray normal to the incident ray?

10.10 Estimate the distance for which ray optics is good approximation for an aperture of 4 mm and wavelength 400 nm.

ADDITIONAL EXERCISES

10.11 The 6563 Å Hα line emitted by hydrogen in a star is found to be red-shifted by 15 Å. Estimate the speed with which the star is receding from the Earth.

10.12 Explain how Corpuscular theory predicts the speed of light in a medium, say, water, to be greater than the speed of light in vacuum. Is the prediction confirmed by experimental determination of the speed of light in water? If not, which alternative picture of light is consistent with experiment?

10.13 You have learnt in the text how Huygens' principle leads to the laws of reflection and refraction. Use the same principle to deduce directly that a point object placed in front of a plane mirror produces a virtual image whose distance from the mirror is equal to the object distance from the mirror.

10.14 Let us list some of the factors, which could possibly influence the speed of wave propagation:

(i) nature of the source.

(ii) direction of propagation.

(iii) motion of the source and/or observer.

(iv) wavelength.

(v) intensity of the wave.

On which of these factors, if any, does

(a) the speed of light in vacuum,

(b) the speed of light in a medium (say, glass or water),

depend?

10.15 For sound waves, the Doppler formula for frequency shift differs slightly between the two situations: (i) source at rest; observer moving, and (ii) source moving; observer at rest. The exact Doppler formulas for the case of light waves in vacuum are, however, strictly identical for these situations. Explain why this should be so. Would you expect the formulas to be strictly identical for the two situations in case of light travelling in a medium?

10.16 In double-slit experiment using light of wavelength 600 nm, the angular width of a fringe formed on a distant screen is 0.1°. What is the spacing between the two slits?

10.17 Answer the following questions:

(a) In a single slit diffraction experiment, the width of the slit is made double the original width. How does this affect the size and intensity of the central diffraction band?

(b) In what way is diffraction from each slit related to the interference pattern in a double-slit experiment?

(c) When a tiny circular obstacle is placed in the path of light from a distant source, a bright spot is seen at the centre of the shadow of the obstacle. Explain why?

(d) Two students are separated by a 7 m partition wall in a room 10 m high. If both light and sound waves can bend around

obstacles, how is it that the students are unable to see each other even though they can converse easily.

(e) Ray optics is based on the assumption that light travels in a straight line. Diffraction effects (observed when light propagates through small apertures/slits or around small obstacles) disprove this assumption. Yet the ray optics assumption is so commonly used in understanding location and several other properties of images in optical instruments. What is the justification?

10.18 Two towers on top of two hills are 40 km apart. The line joining them passes 50 m above a hill halfway between the towers. What is the longest wavelength of radio waves, which can be sent between the towers without appreciable diffraction effects?

10.19 A parallel beam of light of wavelength 500 nm falls on a narrow slit and the resulting diffraction pattern is observed on a screen 1 m away. It is observed that the first minimum is at a distance of 2.5 mm from the centre of the screen. Find the width of the slit.

10.20 Answer the following questions:

(a) When a low flying aircraft passes overhead, we sometimes notice a slight shaking of the picture on our TV screen. Suggest a possible explanation.

(b) As you have learnt in the text, the principle of linear superposition of wave displacement is basic to understanding intensity distributions in diffraction and interference patterns. What is the justification of this principle?

10.21 In deriving the single slit diffraction pattern, it was stated that the intensity is zero at angles of $n\lambda/a$. Justify this by suitably dividing the slit to bring out the cancellation.

Chapter Eleven

DUAL NATURE OF RADIATION AND MATTER

11.1 Introduction

The Maxwell's equations of electromagnetism and Hertz experiments on the generation and detection of electromagnetic waves in 1887 strongly established the wave nature of light. Towards the same period at the end of 19th century, experimental investigations on conduction of electricity (electric discharge) through gases at low pressure in a discharge tube led to many historic discoveries. The discovery of X-rays by Roentgen in 1895, and of electron by J. J. Thomson in 1897, were important milestones in the understanding of atomic structure. It was found that at sufficiently low pressure of about 0.001 mm of mercury column, a discharge took place between the two electrodes on applying the electric field to the gas in the discharge tube. A fluorescent glow appeared on the glass opposite to cathode. The colour of glow of the glass depended on the type of glass, it being yellowish-green for soda glass. The cause of this fluorescence was attributed to the radiation which appeared to be coming from the cathode. These *cathode rays* were discovered, in 1870, by William Crookes who later, in 1879, suggested that these rays consisted of streams of fast moving negatively charged particles. The British physicist J. J. Thomson (1856-1940) confirmed this hypothesis. By applying mutually perpendicular electric and magnetic fields across the discharge tube, J. J. Thomson was the first to determine experimentally the speed and the specific charge [charge to mass ratio (e/m)] of the cathode ray

particles. They were found to travel with speeds ranging from about 0.1 to 0.2 times the speed of light (3×10^8 m/s). The presently accepted value of e/m is 1.76×10^{11} C/kg. Further, the value of e/m was found to be independent of the nature of the material/metal used as the cathode (emitter), or the gas introduced in the discharge tube. This observation suggested the universality of the cathode ray particles.

Around the same time, in 1887, it was found that certain metals, when irradiated by ultraviolet light, emitted negatively charged particles having small speeds. Also, certain metals when heated to a high temperature were found to emit negatively charged particles. The value of e/m of these particles was found to be the same as that for cathode ray particles. These observations thus established that all these particles, although produced under different conditions, were identical in nature. J. J. Thomson, in 1897, named these particles as *electrons*, and suggested that they were fundamental, universal constituents of matter. For his epoch-making discovery of electron, through his theoretical and experimental investigations on conduction of electricity by gasses, he was awarded the Nobel Prize in Physics in 1906. In 1913, the American physicist R. A. Millikan (1868-1953) performed the pioneering oil-drop experiment for the precise measurement of the charge on an electron. He found that the charge on an oil-droplet was always an integral multiple of an elementary charge, 1.602×10^{-19} C. Millikan's experiment established that *electric charge is quantised*. From the values of charge (e) and specific charge (e/m), the mass (m) of the electron could be determined.

11.2 Electron Emission

We know that metals have free electrons (negatively charged particles) that are responsible for their conductivity. However, the free electrons cannot normally escape out of the metal surface. If an electron attempts to come out of the metal, the metal surface acquires a positive charge and pulls the electron back to the metal. The free electron is thus held inside the metal surface by the attractive forces of the ions. Consequently, the electron can come out of the metal surface only if it has got sufficient energy to overcome the attractive pull. A certain minimum amount of energy is required to be given to an electron to pull it out from the surface of the metal. This minimum energy required by an electron to escape from the metal surface is called the *work function* of the metal. It is generally denoted by ϕ_0 and measured in eV (electron volt). One electron volt is the energy gained by an electron when it has been accelerated by a potential difference of 1 volt, so that $1 \text{ eV} = 1.602 \times 10^{-19}$ J.

This unit of energy is commonly used in atomic and nuclear physics. The work function (ϕ_0) depends on the properties of the metal and the nature of its surface. The values of work function of some metals are given in Table 11.1. These values are approximate as they are very sensitive to surface impurities.

Note from Table 11.1 that the work function of platinum is the highest ($\phi_0 = 5.65$ eV) while it is the lowest ($\phi_0 = 2.14$ eV) for caesium.

The minimum energy required for the electron emission from the metal surface can be supplied to the free electrons by any one of the following physical processes:

Table 11.1 Work functions of some metals

Metal	Work function ϕ_0 (eV)	Metal	Work function ϕ_0 (eV)
Cs	2.14	Al	4.28
K	2.30	Hg	4.49
Na	2.75	Cu	4.65
Ca	3.20	Ag	4.70
Mo	4.17	Ni	5.15
Pb	4.25	Pt	5.65

(i) *Thermionic emission:* By suitably heating, sufficient thermal energy can be imparted to the free electrons to enable them to come out of the metal.

(ii) *Field emission*: By applying a very strong electric field (of the order of 10^8 V m^{-1}) to a metal, electrons can be pulled out of the metal, as in a spark plug.

(iii) *Photoelectric emission*: When light of suitable frequency illuminates a metal surface, electrons are emitted from the metal surface. These photo(light)-generated electrons are called *photoelectrons*.

11.3 Photoelectric Effect

11.3.1 Hertz's observations

The phenomenon of photoelectric emission was discovered in 1887 by Heinrich Hertz (1857-1894), during his electromagnetic wave experiments. In his experimental investigation on the production of electromagnetic waves by means of a spark discharge, Hertz observed that high voltage sparks across the detector loop were enhanced when the emitter plate was illuminated by ultraviolet light from an arc lamp.

Light shining on the metal surface somehow facilitated the escape of free, charged particles which we now know as electrons. When light falls on a metal surface, some electrons near the surface absorb enough energy from the incident radiation to overcome the attraction of the positive ions in the material of the surface. After gaining sufficient energy from the incident light, the electrons escape from the surface of the metal into the surrounding space.

11.3.2 Hallwachs' and Lenard's observations

Wilhelm Hallwachs and Philipp Lenard investigated the phenomenon of photoelectric emission in detail during 1886-1902.

Lenard (1862-1947) observed that when ultraviolet radiations were allowed to fall on the emitter plate of an evacuated glass tube enclosing two electrodes (metal plates), current flows in the circuit (Fig. 11.1). As soon as the ultraviolet radiations were stopped, the current flow also

stopped. These observations indicate that when ultraviolet radiations fall on the emitter plate C, electrons are ejected from it which are attracted towards the positive, collector plate A by the electric field. The electrons flow through the evacuated glass tube, resulting in the current flow. Thus, light falling on the surface of the emitter causes current in the external circuit. Hallwachs and Lenard studied how this photo current varied with collector plate potential, and with frequency and intensity of incident light.

Hallwachs, in 1888, undertook the study further and connected a negatively charged zinc plate to an electroscope. He observed that the zinc plate lost its charge when it was illuminated by ultraviolet light. Further, the uncharged zinc plate became positively charged when it was irradiated by ultraviolet light. Positive charge on a positively charged zinc plate was found to be further enhanced when it was illuminated by ultraviolet light. From these observations he concluded that negatively charged particles were emitted from the zinc plate under the action of ultraviolet light.

After the discovery of the electron in 1897, it became evident that the incident light causes electrons to be emitted from the emitter plate. Due to negative charge, the emitted electrons are pushed towards the collector plate by the electric field. Hallwachs and Lenard also observed that when ultraviolet light fell on the emitter plate, no electrons were emitted at all when the frequency of the incident light was smaller than a certain minimum value, called the *threshold frequency*. This minimum frequency depends on the nature of the material of the emitter plate.

It was found that certain metals like zinc, cadmium, magnesium, etc., responded only to ultraviolet light, having short wavelength, to cause electron emission from the surface. However, some alkali metals such as lithium, sodium, potassium, caesium and rubidium were sensitive even to visible light. All these *photosensitive substances* emit electrons when they are illuminated by light. After the discovery of electrons, these electrons were termed as *photoelectrons.* The phenomenon is called *photoelectric effect.*

PHYSICS

Simulate experiments on photoelectric effect
http://www.kcvs.ca/site/projects/physics.html

11.4 Experimental Study of Photoelectric Effect

Figure 11.1 depicts a schematic view of the arrangement used for the experimental study of the photoelectric effect. It consists of an evacuated glass/quartz tube having a photosensitive plate C and another metal plate A. Monochromatic light from the source S of sufficiently short wavelength passes through the window W and falls on the photosensitive plate C (emitter). A transparent quartz window is sealed on to the glass tube, which permits ultraviolet radiation to pass through it and irradiate the photosensitive plate C. The electrons are emitted by the plate C and are collected by the plate A (collector), by the electric field created by the battery. The battery maintains the potential difference between the plates C and A, that can be varied. The polarity of the plates C and A can be reversed by a commutator. Thus, the plate A can be maintained at a desired positive or negative potential with respect to emitter C. When the collector plate A is positive with respect to the emitter plate C, the electrons are

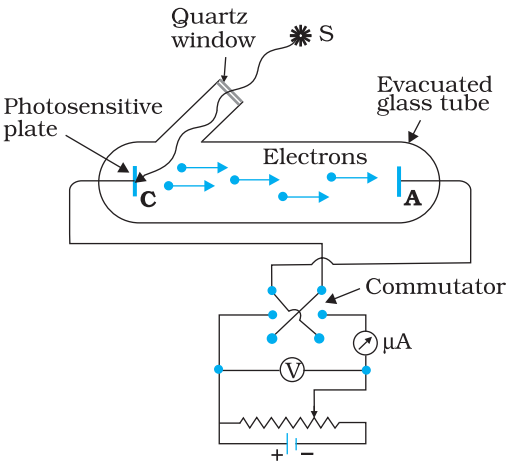


FIGURE 11.1 Experimental arrangement for study of photoelectric effect.

attracted to it. The emission of electrons causes flow of electric current in the circuit. The potential difference between the emitter and collector plates is measured by a voltmeter (V) whereas the resulting photo current flowing in the circuit is measured by a microammeter (μA). The photoelectric current can be increased or decreased by varying the potential of collector plate A with respect to the emitter plate C. The intensity and frequency of the incident light can be varied, as can the potential difference V between the emitter C and the collector A.

We can use the experimental arrangement of Fig. 11.1 to study the variation of photocurrent with (a) intensity of radiation, (b) frequency of incident radiation, (c) the potential difference between the plates A and C, and (d) the nature of the material of plate C. Light of different frequencies can be used by putting appropriate coloured filter or coloured glass in the path of light falling on the emitter C. The intensity of light is varied by changing the distance of the light source from the emitter.

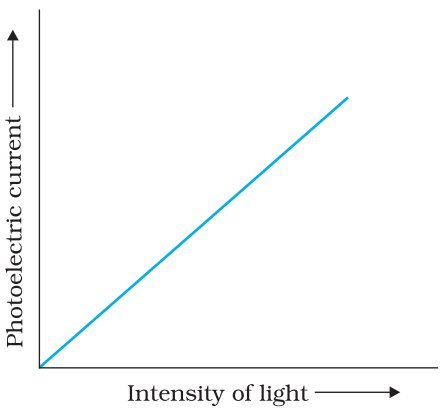


FIGURE 11.2 Variation of Photoelectric current with intensity of light.

11.4.1 Effect of intensity of light on photocurrent

The collector A is maintained at a positive potential with respect to emitter C so that electrons ejected from C are attracted towards collector A. Keeping the frequency of the incident radiation and the potential fixed, the intensity of light is varied and the resulting photoelectric current is measured each time. It is found that the photocurrent increases linearly with intensity of incident light as shown graphically in Fig. 11.2. The photocurrent is directly proportional to the number of photoelectrons emitted per second. This implies that *the number of photoelectrons emitted per second is directly proportional to the intensity of incident radiation.*

11.4.2 Effect of potential on photoelectric current

We first keep the plate A at some positive potential with respect to the plate C and illuminate the plate C with light of fixed frequency ν and fixed intensity I_1. We next vary the positive potential of plate A gradually and measure the resulting photocurrent each time. It is found that the photoelectric current increases with increase in positive (accelerating) potential. At some stage, for a certain positive potential of plate A, all the emitted electrons are collected by the plate A and the photoelectric current becomes maximum or saturates. If we increase the accelerating potential of plate A further, the photocurrent does not increase. This maximum value of the photoelectric current is called *saturation current.* Saturation current corresponds to the case when all the photoelectrons emitted by the emitter plate C reach the collector plate A.

We now apply a negative (retarding) potential to the plate A with respect to the plate C and make it increasingly negative gradually. When the

polarity is reversed, the electrons are repelled and only the most energetic electrons are able to reach the collector A. The photocurrent is found to decrease rapidly until it drops to zero at a certain sharply defined, critical value of the negative potential V_0 on the plate A. For a particular frequency of incident radiation, *the minimum negative (retarding) potential V_0 given to the plate A for which the photocurrent stops or becomes zero is called the cut-off or stopping potential.*

The interpretation of the observation in terms of photoelectrons is straightforward. All the photoelectrons emitted from the metal do not have the same energy. Photoelectric current is zero when the stopping potential is sufficient to repel even the most energetic photoelectrons, with the maximum kinetic energy (K_{max}), so that

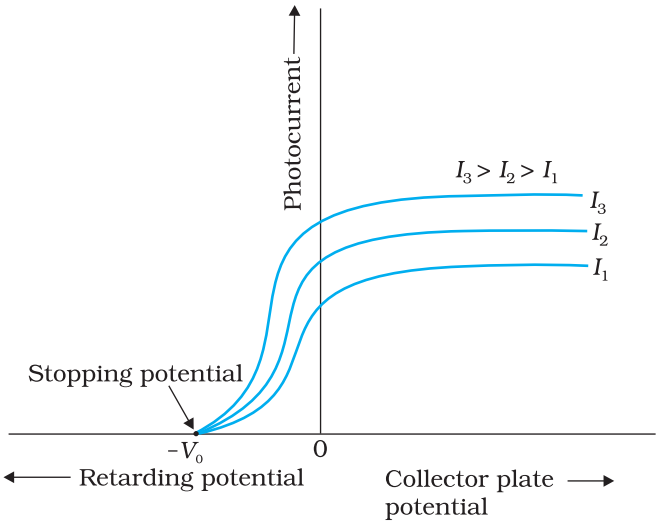


FIGURE 11.3 Variation of photocurrent with collector plate potential for different intensity of incident radiation.

$$K_{max} = e\,V_0 \qquad (11.1)$$

We can now repeat this experiment with incident radiation of the same frequency but of higher intensity I_2 and I_3 ($I_3 > I_2 > I_1$). We note that the saturation currents are now found to be at higher values. This shows that more electrons are being emitted per second, proportional to the intensity of incident radiation. But the stopping potential remains the same as that for the incident radiation of intensity I_1, as shown graphically in Fig. 11.3. Thus, *for a given frequency of the incident radiation, the stopping potential is independent of its intensity.* In other words, the maximum kinetic energy of photoelectrons depends on the light source and the emitter plate material, but is independent of intensity of incident radiation.

11.4.3 Effect of frequency of incident radiation on stopping potential

We now study the relation between the frequency ν of the incident radiation and the stopping potential V_0. We suitably adjust the same intensity of light radiation at various frequencies and study the variation of photocurrent with collector plate potential. The resulting variation is shown in Fig. 11.4. We obtain different values of stopping potential but the same value of the saturation current for incident radiation of different frequencies. The energy of the emitted electrons depends on the frequency of the incident radiations. The stopping potential is more negative for higher frequencies of incident radiation. Note from

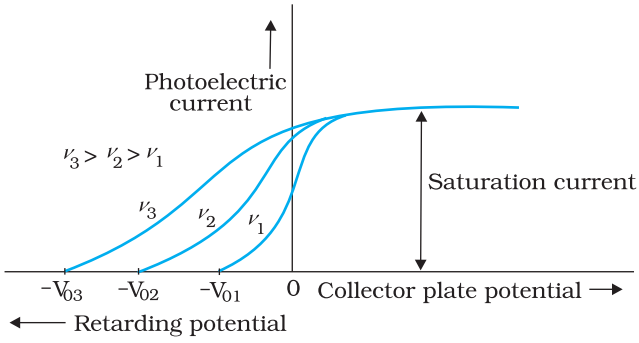


FIGURE 11.4 Variation of photoelectric current with collector plate potential for different frequencies of incident radiation.

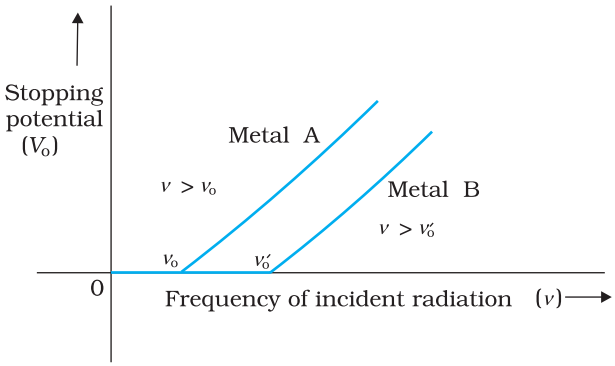


FIGURE 11.5 Variation of stopping potential V_0 with frequency ν of incident radiation for a given photosensitive material.

Fig. 11.4 that the stopping potentials are in the order $V_{03} > V_{02} > V_{01}$ if the frequencies are in the order $\nu_3 > \nu_2 > \nu_1$. This implies that greater the frequency of incident light, greater is the maximum kinetic energy of the photoelectrons. Consequently, we need greater retarding potential to stop them completely. If we plot a graph between the frequency of incident radiation and the corresponding stopping potential for different metals we get a straight line, as shown in Fig. 11.5.

The graph shows that

(i) the stopping potential V_0 varies linearly with the frequency of incident radiation for a given photosensitive material.

(ii) there exists a certain minimum cut-off frequency ν_0 for which the stopping potential is zero.

These observations have two implications:

(i) *The maximum kinetic energy of the photoelectrons varies linearly with the frequency of incident radiation, but is independent of its intensity.*

(ii) *For a frequency ν of incident radiation, lower than the cut-off frequency ν_0, no photoelectric emission is possible even if the intensity is large.*

This minimum, cut-off frequency ν_0, is called the *threshold frequency*. It is different for different metals.

Different photosensitive materials respond differently to light. Selenium is more sensitive than zinc or copper. The same photosensitive substance gives different response to light of different wavelengths. For example, ultraviolet light gives rise to photoelectric effect in copper while green or red light does not.

Note that in all the above experiments, it is found that, if frequency of the incident radiation exceeds the threshold frequency, the photoelectric emission starts instantaneously without any apparent time lag, even if the incident radiation is very dim. It is now known that emission starts in a time of the order of 10^{-9} s or less.

We now summarise the experimental features and observations described in this section.

(i) For a given photosensitive material and frequency of incident radiation (above the threshold frequency), the photoelectric current is directly proportional to the intensity of incident light (Fig. 11.2).

(ii) For a given photosensitive material and frequency of incident radiation, saturation current is found to be proportional to the intensity of incident radiation whereas the stopping potential is independent of its intensity (Fig. 11.3).

(iii) For a given photosensitive material, there exists a certain minimum cut-off frequency of the incident radiation, called the *threshold frequency*, below which no emission of photoelectrons takes place, no matter how intense the incident light is. Above the threshold frequency, the stopping potential or equivalently the maximum kinetic

energy of the emitted photoelectrons increases linearly with the frequency of the incident radiation, but is independent of its intensity (Fig. 11.5).

(iv) The photoelectric emission is an instantaneous process without any apparent time lag (~10^{-9}s or less), even when the incident radiation is made exceedingly dim.

11.5 Photoelectric Effect and Wave Theory of Light

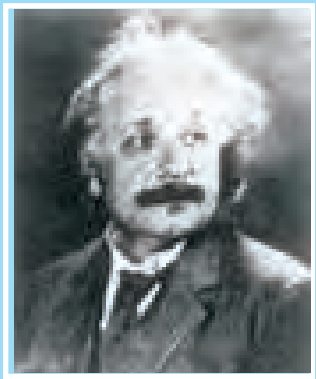
The wave nature of light was well established by the end of the nineteenth century. The phenomena of interference, diffraction and polarisation were explained in a natural and satisfactory way by the wave picture of light. According to this picture, light is an electromagnetic wave consisting of electric and magnetic fields with continuous distribution of energy over the region of space over which the wave is extended. Let us now see if this wave picture of light can explain the observations on photoelectric emission given in the previous section.

According to the wave picture of light, the free electrons at the surface of the metal (over which the beam of radiation falls) absorb the radiant energy continuously. The greater the intensity of radiation, the greater are the amplitude of electric and magnetic fields. Consequently, the greater the intensity, the greater should be the energy absorbed by each electron. In this picture, the maximum kinetic energy of the photoelectrons on the surface is then expected to increase with increase in intensity. Also, no matter what the frequency of radiation is, a sufficiently intense beam of radiation (over sufficient time) should be able to impart enough energy to the electrons, so that they exceed the minimum energy needed to escape from the metal surface . A threshold frequency, therefore, should not exist. These expectations of the wave theory directly contradict observations (i), (ii) and (iii) given at the end of sub-section 11.4.3.

Further, we should note that in the wave picture, the absorption of energy by electron takes place continuously over the entire wavefront of the radiation. Since a large number of electrons absorb energy, the energy absorbed per electron per unit time turns out to be small. Explicit calculations estimate that it can take hours or more for a single electron to pick up sufficient energy to overcome the work function and come out of the metal. This conclusion is again in striking contrast to observation (iv) that the photoelectric emission is instantaneous. In short, the wave picture is unable to explain the most basic features of photoelectric emission.

11.6 Einstein's Photoelectric Equation: Energy Quantum of Radiation

In 1905, Albert Einstein (1879-1955) proposed a radically new picture of electromagnetic radiation to explain photoelectric effect. In this picture, photoelectric emission does not take place by continuous absorption of energy from radiation. Radiation energy is built up of discrete units – the so called *quanta of energy of radiation*. Each quantum of radiant energy



Albert Einstein (1879 – 1955) Einstein, one of the greatest physicists of all time, was born in Ulm, Germany. In 1905, he published three path-breaking papers. In the first paper, he introduced the notion of light quanta (now called photons) and used it to explain the features of photoelectric effect. In the second paper, he developed a theory of Brownian motion, confirmed experimentally a few years later and provided a convincing evidence of the atomic picture of matter. The third paper gave birth to the special theory of relativity. In 1916, he published the general theory of relativity. Some of Einstein's most significant later contributions are: the notion of stimulated emission introduced in an alternative derivation of Planck's blackbody radiation law, static model of the universe which started modern cosmology, quantum statistics of a gas of massive bosons, and a critical analysis of the foundations of quantum mechanics. In 1921, he was awarded the Nobel Prize in physics for his contribution to theoretical physics and the photoelectric effect.

ALBERT EINSTEIN (1879 – 1955)

has energy $h\nu$, where h is Planck's constant and ν the frequency of light. In photoelectric effect, an electron absorbs a quantum of energy ($h\nu$) of radiation. If this quantum of energy absorbed exceeds the minimum energy needed for the electron to escape from the metal surface (work function ϕ_0), the electron is emitted with maximum kinetic energy

$$K_{max} = h\nu - \phi_0 \qquad (11.2)$$

More tightly bound electrons will emerge with kinetic energies less than the maximum value. Note that the intensity of light of a given frequency is determined by the number of photons incident per second. Increasing the intensity will increase the number of emitted electrons per second. However, the maximum kinetic energy of the emitted photoelectrons is determined by the energy of each photon.

Equation (11.2) is known as *Einstein's photoelectric equation*. We now see how this equation accounts in a simple and elegant manner all the observations on photoelectric effect given at the end of sub-section 11.4.3.

- According to Eq. (11.2), K_{max} depends linearly on ν, and is independent of intensity of radiation, in agreement with observation. This has happened because in Einstein's picture, photoelectric effect arises from the absorption of a single quantum of radiation by a single electron. The intensity of radiation (that is proportional to the number of energy quanta per unit area per unit time) is irrelevant to this basic process.
- Since K_{max} must be non-negative, Eq. (11.2) implies that photoelectric emission is possible only if
  $h\nu > \phi_0$
  or $\nu > \nu_0$, where
  $$\nu_0 = \frac{\phi_0}{h} \qquad (11.3)$$
  Equation (11.3) shows that the greater the work function ϕ_0, the higher the minimum or threshold frequency ν_0 needed to emit photoelectrons. Thus, there exists a threshold frequency ν_0 (= ϕ_0/h) for the metal surface, below which no photoelectric emission is possible, no matter how intense the incident radiation may be or how long it falls on the surface.
- In this picture, intensity of radiation as noted above, is proportional to the number of energy quanta per unit area per unit time. The greater the number of energy quanta available, the greater is the number of electrons absorbing the energy quanta and greater, therefore, is the number of electrons coming out of the metal (for $\nu > \nu_0$). This explains why, for $\nu > \nu_0$, photoelectric current is proportional to intensity.

- In Einstein's picture, the basic elementary process involved in photoelectric effect is the absorption of a light quantum by an electron. This process is instantaneous. Thus, whatever may be the intensity i.e., the number of quanta of radiation per unit area per unit time, photoelectric emission is instantaneous. Low intensity does not mean delay in emission, since the basic elementary process is the same. Intensity only determines how many electrons are able to participate in the elementary process (absorption of a light quantum by a single electron) and, therefore, the photoelectric current.

Using Eq. (11.1), the photoelectric equation, Eq. (11.2), can be written as

$$e V_0 = h\nu - \phi_0; \text{ for } \nu \geq \nu_0$$

$$\text{or } V_0 = \left(\frac{h}{e}\right)\nu - \frac{\phi_0}{e} \tag{11.4}$$

This is an important result. It predicts that the V_0 versus ν curve is a straight line with slope = (h/e), independent of the nature of the material. During 1906-1916, Millikan performed a series of experiments on photoelectric effect, aimed at disproving Einstein's photoelectric equation. He measured the slope of the straight line obtained for sodium, similar to that shown in Fig. 11.5. Using the known value of e, he determined the value of Planck's constant h. This value was close to the value of Planck's contant (= 6.626×10^{-34} J s) determined in an entirely different context. In this way, in 1916, Millikan proved the validity of Einstein's photoelectric equation, instead of disproving it.

The successful explanation of photoelectric effect using the hypothesis of light quanta and the experimental determination of values of h and ϕ_0, in agreement with values obtained from other experiments, led to the acceptance of Einstein's picture of photoelectric effect. Millikan verified photoelectric equation with great precision, for a number of alkali metals over a wide range of radiation frequencies.

11.7 Particle Nature of Light: The Photon

Photoelectric effect thus gave evidence to the strange fact that light in interaction with matter behaved as if it was made of quanta or packets of energy, each of energy $h\nu$.

Is the light quantum of energy to be associated with a particle? Einstein arrived at the important result, that the light quantum can also be associated with momentum $(h\nu/c)$. A definite value of energy as well as momentum is a strong sign that the light quantum can be associated with a particle. This particle was later named *photon*. The particle-like behaviour of light was further confirmed, in 1924, by the experiment of A.H. Compton (1892-1962) on scattering of X-rays from electrons. In 1921, Einstein was awarded the Nobel Prize in Physics for his contribution to theoretical physics and the photoelectric effect. In 1923, Millikan was awarded the Nobel Prize in physics for his work on the elementary charge of electricity and on the photoelectric effect.

We can summarise the photon picture of electromagnetic radiation as follows:

(i) In interaction of radiation with matter, radiation behaves as if it is made up of particles called photons.

(ii) Each photon has energy E (=$h\nu$) and momentum p (= $h\nu/c$), and speed c, the speed of light.

(iii) All photons of light of a particular frequency ν, or wavelength λ, have the same energy E (=$h\nu = hc/\lambda$) and momentum p (= $h\nu/c = h/\lambda$), whatever the intensity of radiation may be. By increasing the intensity of light of given wavelength, there is only an increase in the number of photons per second crossing a given area, with each photon having the same energy. Thus, photon energy is independent of intensity of radiation.

(iv) Photons are electrically neutral and are not deflected by electric and magnetic fields.

(v) In a photon-particle collision (such as photon-electron collision), the total energy and total momentum are conserved. However, the number of photons may not be conserved in a collision. The photon may be absorbed or a new photon may be created.

EXAMPLE 11.1

Example 11.1 Monochromatic light of frequency 6.0×10^{14} Hz is produced by a laser. The power emitted is 2.0×10^{-3} W. (a) What is the energy of a photon in the light beam? (b) How many photons per second, on an average, are emitted by the source?

Solution

(a) Each photon has an energy

$E = h\nu = (6.63 \times 10^{-34}\ \text{J s})\ (6.0 \times 10^{14}\ \text{Hz})$

$= 3.98 \times 10^{-19}\ \text{J}$

(b) If N is the number of photons emitted by the source per second, the power P transmitted in the beam equals N times the energy per photon E, so that $P = N E$. Then

$$N = \frac{P}{E} = \frac{2.0 \times 10^{-3}\ \text{W}}{3.98 \times 10^{-19}\ \text{J}}$$

$= 5.0 \times 10^{15}$ photons per second.

EXAMPLE 11.2

Example 11.2 The work function of caesium is 2.14 eV. Find (a) the threshold frequency for caesium, and (b) the wavelength of the incident light if the photocurrent is brought to zero by a stopping potential of 0.60 V.

Solution

(a) For the cut-off or threshold frequency, the energy $h\nu_0$ of the incident radiation must be equal to work function ϕ_0, so that

$$\nu_0 = \frac{\phi_0}{h} = \frac{2.14\ \text{eV}}{6.63 \times 10^{-34}\ \text{J s}}$$

$$= \frac{2.14 \times 1.6 \times 10^{-19}\ \text{J}}{6.63 \times 10^{-34}\ \text{J s}} = 5.16 \times 10^{14}\ \text{Hz}$$

Thus, for frequencies less than this threshold frequency, no photoelectrons are ejected.

(b) Photocurrent reduces to zero, when maximum kinetic energy of the emitted photoelectrons equals the potential energy $e V_0$ by the retarding potential V_0. Einstein's Photoelectric equation is

$$eV_0 = h\nu - \phi_0 = \frac{hc}{\lambda} - \phi_0$$

or, $\lambda = hc/(eV_0 + \phi_0)$

$$= \frac{(6.63\times10^{-34}\,\text{J s})\times(3\times10^{8}\,\text{m/s})}{(0.60\,\text{eV}+2.14\,\text{eV})}$$

$$= \frac{19.89\times10^{-26}\,\text{J m}}{(2.74\,\text{eV})}$$

$$\lambda = \frac{19.89\times10^{-26}\,\text{J m}}{2.74\times1.6\times10^{-19}\,\text{J}} = 454\ \text{nm}$$

EXAMPLE 11.2

Example 11.3 The wavelength of light in the visible region is about 390 nm for violet colour, about 550 nm (average wavelength) for yellow-green colour and about 760 nm for red colour.

(a) What are the energies of photons in (eV) at the (i) violet end, (ii) average wavelength, yellow-green colour, and (iii) red end of the visible spectrum? (Take $h = 6.63\times10^{-34}$ J s and 1 eV = 1.6×10^{-19}J.)

(b) From which of the photosensitive materials with work functions listed in Table 11.1 and using the results of (i), (ii) and (iii) of (a), can you build a photoelectric device that operates with visible light?

Solution

(a) Energy of the incident photon, $E = h\nu = hc/\lambda$

$E = (6.63\times10^{-34}\text{J s})\ (3\times10^{8}\ \text{m/s})/\lambda$

$$= \frac{1.989\times10^{-25}\,\text{J m}}{\lambda}$$

(i) For violet light, λ_1 = 390 nm (lower wavelength end)

Incident photon energy, $E_1 = \dfrac{1.989\times10^{-25}\,\text{J m}}{390\times10^{-9}\,\text{m}}$

$= 5.10\times10^{-19}$J

$= \dfrac{5.10\times10^{-19}\,\text{J}}{1.6\times10^{-19}\,\text{J/eV}}$

= 3.19 eV

(ii) For yellow-green light, λ_2 = 550 nm (average wavelength)

Incident photon energy, $E_2 = \dfrac{1.989\times10^{-25}\,\text{J m}}{550\times10^{-9}\,\text{m}}$

$= 3.62\times10^{-19}$ J = 2.26 eV

(iii) For red light, λ_3 = 760 nm (higher wavelength end)

Incident photon energy, $E_3 = \dfrac{1.989\times10^{-25}\,\text{J m}}{760\times10^{-9}\,\text{m}}$

$= 2.62\times10^{-19}$ J = 1.64 eV

(b) For a photoelectric device to operate, we require incident light energy E to be equal to or greater than the work function ϕ_0 of the material. Thus, the photoelectric device will operate with violet light (with E = 3.19 eV) photosensitive material Na (with ϕ_0 = 2.75 eV), K (with ϕ_0 = 2.30 eV) and Cs (with ϕ_0 = 2.14 eV). It will also operate with yellow-green light (with E = 2.26 eV) for Cs (with ϕ_0 = 2.14 eV) only. However, it will not operate with red light (with E = 1.64 eV) for any of these photosensitive materials.

EXAMPLE 11.3

11.8 Wave Nature of Matter

The dual (wave-particle) nature of light (electromagnetic radiation, in general) comes out clearly from what we have learnt in this and the preceding chapters. The wave nature of light shows up in the phenomena of interference, diffraction and polarisation. On the other hand, in photoelectric effect and Compton effect which involve energy and momentum transfer, radiation behaves as if it is made up of a bunch of particles – the photons. Whether a particle or wave description is best suited for understanding an experiment depends on the nature of the experiment. For example, in the familiar phenomenon of seeing an object by our eye, both descriptions are important. The gathering and focussing mechanism of light by the eye-lens is well described in the wave picture. But its absorption by the rods and cones (of the retina) requires the photon picture of light.

A natural question arises: If radiation has a dual (wave-particle) nature, might not the particles of nature (the electrons, protons, etc.) also exhibit wave-like character? In 1924, the French physicist Louis Victor de Broglie (pronounced as de Broy) (1892-1987) put forward the bold hypothesis that moving particles of matter should display wave-like properties under suitable conditions. He reasoned that nature was symmetrical and that the two basic physical entities – matter and energy, must have symmetrical character. If radiation shows dual aspects, so should matter. De Broglie proposed that the wave length λ associated with a particle of momentum p is given as

$$\lambda = \frac{h}{p} = \frac{h}{mv} \tag{11.5}$$

where m is the mass of the particle and v its speed. Equation (11.5) is known as the *de Broglie relation* and the wavelength λ of the *matter wave* is called *de Broglie wavelength*. The dual aspect of matter is evident in the de Broglie relation. On the left hand side of Eq. (11.5), λ is the attribute of a wave while on the right hand side the momentum p is a typical attribute of a particle. Planck's constant h relates the two attributes.

Equation (11.5) for a material particle is basically a hypothesis whose validity can be tested only by experiment. However, it is interesting to see that it is satisfied also by a photon. For a photon, as we have seen,

$$p = h\nu / c \tag{11.6}$$

Therefore,

$$\frac{h}{p} = \frac{c}{\nu} = \lambda \tag{11.7}$$

That is, the de Broglie wavelength of a photon given by Eq. (11.5) equals the wavelength of electromagnetic radiation of which the photon is a quantum of energy and momentum.

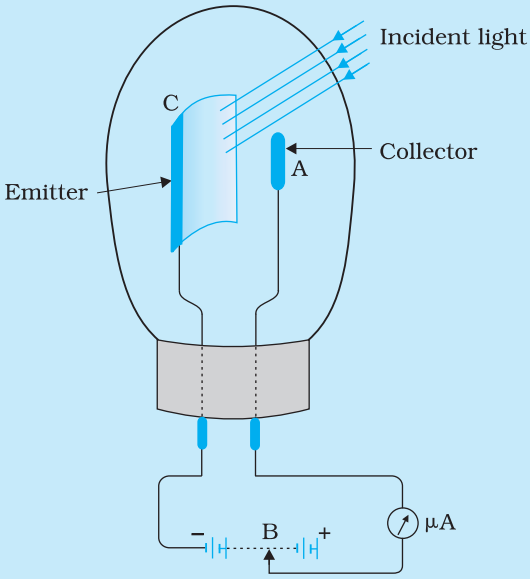
Clearly, from Eq. (11.5), λ is smaller for a heavier particle (large m) or more energetic particle (large v). For example, the de Broglie wavelength of a ball of mass 0.12 kg moving with a speed of 20 m s^{-1} is easily calculated:

PHOTOCELL

A photocell is a technological application of the photoelectric effect. It is a device whose electrical properties are affected by light. It is also sometimes called an electric eye. A photocell consists of a semi-cylindrical photo-sensitive metal plate C (emitter) and a wire loop A (collector) supported in an evacuated glass or quartz bulb. It is connected to the external circuit having a high-tension battery B and microammeter (μA) as shown in the Figure. Sometimes, instead of the plate C, a thin layer of photosensitive material is pasted on the inside of the bulb. A part of the bulb is left clean for the light to enter it.

When light of suitable wavelength falls on the emitter C, photoelectrons are emitted. These photoelectrons are drawn to the collector A. Photocurrent of the order of a few microampere can be normally obtained from a photo cell.

A photocell converts a change in intensity of illumination into a change in photocurrent. This current can be used to operate control systems and in light measuring devices. A photocell of lead sulphide sensitive to infrared radiation is used in electronic ignition circuits.



A photo cell

In scientific work, photo cells are used whenever it is necessary to measure the intensity of light. Light meters in photographic cameras make use of photo cells to measure the intensity of incident light. The photocells, inserted in the door light electric circuit, are used as automatic door opener. A person approaching a doorway may interrupt a light beam which is incident on a photocell. The abrupt change in photocurrent may be used to start a motor which opens the door or rings an alarm. They are used in the control of a counting device which records every interruption of the light beam caused by a person or object passing across the beam. So photocells help count the persons entering an auditorium, provided they enter the hall one by one. They are used for detection of traffic law defaulters: an alarm may be sounded whenever a beam of *(invisible)* radiation is intercepted.

In burglar alarm, (invisible) ultraviolet light is continuously made to fall on a photocell installed at the doorway. A person entering the door interrupts the beam falling on the photocell. The abrupt change in photocurrent is used to start an electric bell ringing. In fire alarm, a number of photocells are installed at suitable places in a building. In the event of breaking out of fire, light radiations fall upon the photocell. This completes the electric circuit through an electric bell or a siren which starts operating as a warning signal.

Photocells are used in the reproduction of sound in motion pictures and in the television camera for scanning and telecasting scenes. They are used in industries for detecting minor flaws or holes in metal sheets.

$$p = m\,v = 0.12 \text{ kg} \times 20 \text{ m s}^{-1} = 2.40 \text{ kg m s}^{-1}$$

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-34} \text{ J s}}{2.40 \text{ kg m s}^{-1}} = 2.76 \times 10^{-34} \text{ m}$$

LOUIS VICTOR DE BROGLIE (1892 – 1987)

Louis Victor de Broglie (1892 – 1987) French physicist who put forth revolutionary idea of wave nature of matter. This idea was developed by Erwin Schródinger into a full-fledged theory of quantum mechanics commonly known as wave mechanics. In 1929, he was awarded the Nobel Prize in Physics for his discovery of the wave nature of electrons.

This wavelength is so small that it is beyond any measurement. This is the reason why macroscopic objects in our daily life do not show wave-like properties. On the other hand, in the sub-atomic domain, the wave character of particles is significant and measurable.

Consider an electron (mass m, charge e) accelerated from rest through a potential V. The kinetic energy K of the electron equals the work done (eV) on it by the electric field:

$$K = e\,V \tag{11.8}$$

Now, $K = \frac{1}{2} m v^2 = \frac{p^2}{2m}$, so that

$$p = \sqrt{2\,m\,K} = \sqrt{2\,m\,eV} \tag{11.9}$$

The de Broglie wavelength λ of the electron is then

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2\,mK}} = \frac{h}{\sqrt{2\,meV}} \tag{11.10}$$

Substituting the numerical values of h, m, e, we get

$$\lambda = \frac{1.227}{\sqrt{V}}\ \text{nm} \tag{11.11}$$

where V is the magnitude of accelerating potential in volts. For a 120 V accelerating potential, Eq. (11.11) gives λ = 0.112 nm. This wavelength is of the same order as the spacing between the atomic planes in crystals. This suggests that matter waves associated with an electron could be verified by crystal diffraction experiments analogous to X-ray diffraction. We describe the experimental verification of the de Broglie hypothesis in the next section. In 1929, de Broglie was awarded the Nobel Prize in Physics for his discovery of the wave nature of electrons.

The matter–wave picture elegantly incorporated the Heisenberg's *uncertainty principle*. According to the principle, it is not possible to measure *both* the position and momentum of an electron (or any other particle) *at the same time* exactly. There is always some uncertainty (Δx) in the specification of position and some uncertainty (Δp) in the specification of momentum. The product of Δx and Δp is of the order of $\hbar$* (with $\hbar = h/2\pi$), i.e.,

$$\Delta x\,\Delta p \approx \hbar \tag{11.12}$$

Equation (11.12) allows the possibility that Δx is zero; but then Δp must be infinite in order that the product is non-zero. Similarly, if Δp is zero, Δx must be infinite. Ordinarily, both Δx and Δp are non-zero such that their product is of the order of $\hbar$.

Now, if an electron has a definite momentum p, (i.e.$\Delta p = 0$), by the de Broglie relation, it has a definite wavelength λ. A wave of definite (single)

* A more rigorous treatment gives $\Delta x\,\Delta p \geq \hbar/2$.

wavelength extends all over space. By Born's probability interpretation this means that the electron is not localised in any finite region of space. That is, its position uncertainty is infinite ($\Delta x \to \infty$), which is consistent with the uncertainty principle.

In general, the matter wave associated with the electron is not extended all over space. It is a wave packet extending over some finite region of space. In that case Δx is not infinite but has some finite value depending on the extension of the wave packet. Also, you must appreciate that a wave packet of finite extension does not have a single wavelength. It is built up of wavelengths spread around some central wavelength.

By de Broglie's relation, then, the momentum of the electron will also have a spread – an uncertainty Δp. This is as expected from the uncertainty principle. It can be shown that the wave packet description together with de Broglie relation and Born's probability interpretation reproduce the Heisenberg's uncertainty principle exactly.

In Chapter 12, the de Broglie relation will be seen to justify Bohr's postulate on quantisation of angular momentum of electron in an atom.

Figure 11.6 shows a schematic diagram of (a) a localised wave packet, and (b) an extended wave with fixed wavelength.

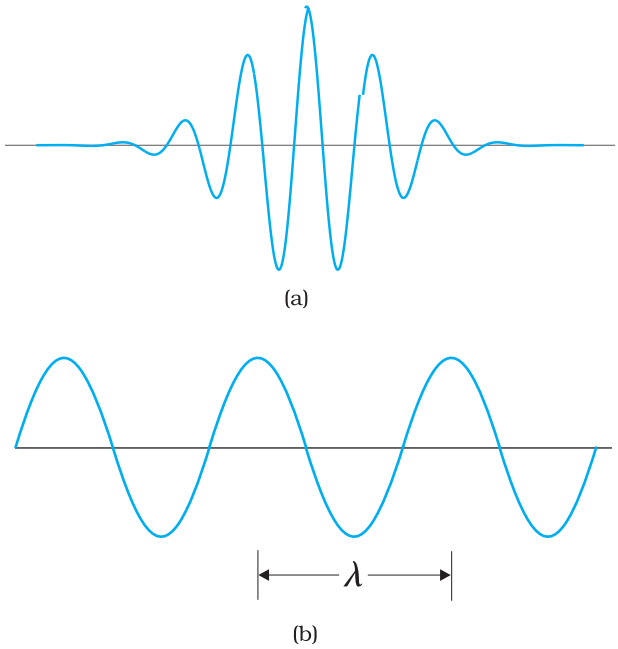


FIGURE 11.6 (a) The wave packet description of an electron. The wave packet corresponds to a spread of wavelength around some central wavelength (and hence by de Broglie relation, a spread in momentum). Consequently, it is associated with an uncertainty in position (Δx) and an uncertainty in momentum (Δp). (b) The matter wave corresponding to a definite momentum of an electron extends all over space. In this case, $\Delta p = 0$ and $\Delta x \to \infty$.

Example 11.4 What is the de Broglie wavelength associated with (a) an electron moving with a speed of 5.4×10^6 m/s, and (b) a ball of mass 150 g travelling at 30.0 m/s?

Solution

(a) For the electron:

Mass $m = 9.11\times10^{-31}$ kg, speed $v = 5.4\times10^6$ m/s. Then, momentum

$p = m\,v = 9.11\times10^{-31}$ (kg) $\times$ 5.4×10^6 (m/s)

$p = 4.92 \times 10^{-24}$ kg m/s

de Broglie wavelength, $\lambda = h/p$

$$= \frac{6.63\times10^{-34}\ \text{J s}}{4.92\times10^{-24}\ \text{kg m/s}}$$

$\lambda = 0.135$ nm

(b) For the ball:

Mass $m' = 0.150$ kg, speed $v' = 30.0$ m/s.

Then momentum $p' = m'\,v' = 0.150$ (kg) $\times$ 30.0 (m/s)

$p' = 4.50$ kg m/s

de Broglie wavelength $\lambda' = h/p'$.

EXAMPLE 11.4

EXAMPLE 11.4

$$= \frac{6.63\times10^{\pm34}\ \text{J s}}{4.50\times\ \text{kg m/s}}$$

$\lambda' = 1.47\times10^{-34}$ m

The de Broglie wavelength of electron is comparable with X-ray wavelengths. However, for the ball it is about 10^{-19} times the size of the proton, quite beyond experimental measurement.

EXAMPLE 11.5

Example 11.5 An electron, an α-particle, and a proton have the same kinetic energy. Which of these particles has the shortest de Broglie wavelength?

Solution

For a particle, de Broglie wavelength, $\lambda = h/p$

Kinetic energy, $K = p^2/2m$

Then, $\lambda = h/\sqrt{2mK}$

For the same kinetic energy K, the de Broglie wavelength associated with the particle is inversely proportional to the square root of their masses. A proton $\left({}^1_1\text{H}\right)$ is 1836 times massive than an electron and an α-particle $\left({}^4_2\text{He}\right)$ four times that of a proton.

Hence, α – particle has the shortest de Broglie wavelength.

PROBABILITY INTERPRETATION TO MATTER WAVES

It is worth pausing here to reflect on just what a matter wave associated with a particle, say, an electron, means. Actually, a truly satisfactory physical understanding of the dual nature of matter and radiation has not emerged so far. The great founders of quantum mechanics (Niels Bohr, Albert Einstein, and many others) struggled with this and related concepts for long. Still the deep physical interpretation of quantum mechanics continues to be an area of active research. Despite this, the concept of matter wave has been mathematically introduced in modern quantum mechanics with great success. An important milestone in this connection was when Max Born (1882-1970) suggested a probability interpretation to the matter wave amplitude. According to this, the intensity (square of the amplitude) of the matter wave at a point determines the probability density of the particle at that point. Probability density means probability per unit volume. Thus, if A is the amplitude of the wave at a point, $|A|^2\ \Delta V$ is the probability of the particle being found in a small volume ΔV around that point. Thus, if the intensity of matter wave is large in a certain region, there is a greater probability of the particle being found there than where the intensity is small.

EXAMPLE 11.6

Example 11.6 A particle is moving three times as fast as an electron. The ratio of the de Broglie wavelength of the particle to that of the electron is 1.813×10^{-4}. Calculate the particle's mass and identify the particle.

Solution

de Broglie wavelength of a moving particle, having mass m and velocity v:

$$\lambda = \frac{h}{p} = \frac{h}{mv}$$

Mass, $m = h/\lambda v$

For an electron, mass $m_e = h/\lambda_e v_e$

Now, we have $v/v_e = 3$ and

$\lambda/\lambda_e = 1.813 \times 10^{-4}$

Then, mass of the particle, $m = m_e \left(\frac{\lambda_e}{\lambda}\right)\left(\frac{v_e}{v}\right)$

$m = (9.11\times10^{-31}\ \text{kg}) \times (1/3) \times (1/1.813 \times 10^{-4})$

$m = 1.675 \times 10^{-27}$ kg.

Thus, the particle, with this mass could be a proton or a neutron.

EXAMPLE 11.6

Example 11.7 What is the de Broglie wavelength associated with an electron, accelerated through a potential differnece of 100 volts?

Solution Accelerating potential V = 100 V. The de Broglie wavelength λ is

$$\lambda = h/p = \frac{1.227}{\sqrt{V}}\ \text{nm}$$

$$\lambda = \frac{1.227}{\sqrt{100}}\ \text{nm} = 0.123\ \text{nm}$$

The de Broglie wavelength associated with an electron in this case is of the order of X-ray wavelengths.

EXAMPLE 11.7

11.9 Davisson and Germer Experiment

The wave nature of electrons was first experimentally verified by C.J. Davisson and L.H. Germer in 1927 and independently by G.P. Thomson, in 1928, who observed diffraction effects with beams of electrons scattered by crystals. Davisson and Thomson shared the Nobel Prize in 1937 for their experimental discovery of diffraction of electrons by crystals.

The experimental arrangement used by Davisson and Germer is schematically shown in Fig. 11.7. It consists of an electron gun which comprises of a tungsten filament F, coated with barium oxide and heated by a low voltage power supply (L.T. or battery). Electrons emitted by the filament are accelerated to a desired velocity

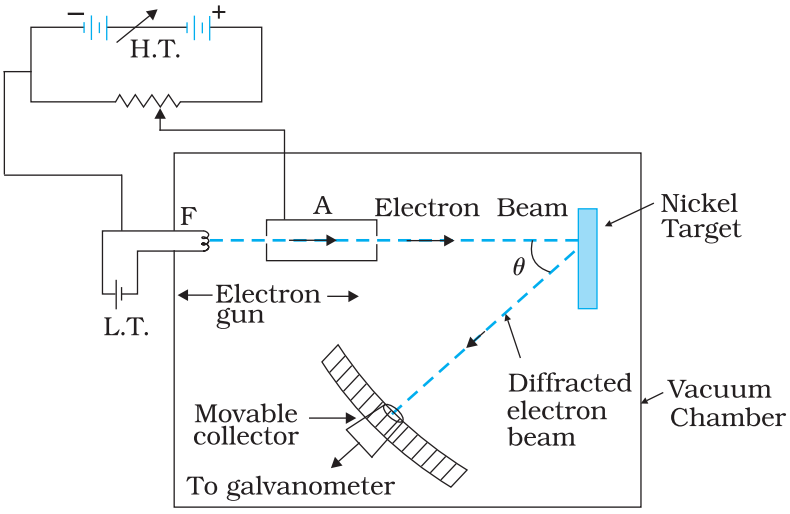


FIGURE 11.7 Davisson-Germer electron diffraction arrangement.

by applying suitable potential/voltage from a high voltage power supply (H.T. or battery). They are made to pass through a cylinder with fine holes along its axis, producing a fine collimated beam. The beam is made to fall on the surface of a nickel crystal. The electrons are scattered in all directions by the atoms of the crystal. The intensity of the electron beam, scattered in a given direction, is measured by the electron detector (collector). The detector can be moved on a circular scale and is connected to a sensitive galvanometer, which records the current. The deflection of the galvanometer is proportional to the intensity of the electron beam entering the collector. The apparatus is enclosed in an evacuated chamber. By moving the detector on the circular scale at different positions, the intensity of the scattered electron beam is measured for different values of angle of scattering θ which is the angle between the incident and the scattered electron beams. The variation of the intensity (I) of the scattered electrons with the angle of scattering θ is obtained for different accelerating voltages.

The experiment was performed by varying the accelarating voltage from 44 V to 68 V. It was noticed that a strong peak appeared in the intensity (I) of the scattered electron for an accelarating voltage of 54V at a scattering angle $\theta = 50°$

The appearance of the peak in a particular direction is due to the constructive interference of electrons scattered from different layers of the regularly spaced atoms of the crystals. From the electron diffraction measurements, the wavelength of matter waves was found to be 0.165 nm.

The de Broglie wavelength λ associated with electrons, using Eq. (11.11), for $V = 54$ V is given by

$$\lambda = h/p = \frac{1\ 227}{\sqrt{V}}\ \text{nm}$$

$$\lambda = \frac{1\ 227}{\sqrt{54}}\ \text{nm} = 0.167\ \text{nm}$$

Thus, there is an excellent agreement between the theoretical value and the experimentally obtained value of de Broglie wavelength. Davisson-Germer experiment thus strikingly confirms the wave nature of electrons and the de Broglie relation. More recently, in 1989, the wave nature of a beam of electrons was experimentally demonstrated in a double-slit experiment, similar to that used for the wave nature of light. Also, in an experiment in 1994, interference fringes were obtained with the beams of iodine molecules, which are about a million times more massive than electrons.

The de Broglie hypothesis has been basic to the development of modern quantum mechanics. It has also led to the field of electron optics. The wave properties of electrons have been utilised in the design of electron microscope which is a great improvement, with higher resolution, over the optical microscope.

Development of electron microscope
http://www.nobelprize.org/nobel_prizes/physics/laureates/1986/presentation-speech.html

SUMMARY

1. The minimum energy needed by an electron to come out from a metal surface is called the work function of the metal. Energy (greater than the work function (ϕ_o) required for electron emission from the metal surface can be supplied by suitably heating or applying strong electric field or irradiating it by light of suitable frequency.
2. Photoelectric effect is the phenomenon of emission of electrons by metals when illuminated by light of suitable frequency. Certain metals respond to ultraviolet light while others are sensitive even to the visible light. Photoelectric effect involves conversion of light energy into electrical energy. It follows the law of conservation of energy. The photoelectric emission is an instantaneous process and possesses certain special features.
3. Photoelectric current depends on (i) the intensity of incident light, (ii) the potential difference applied between the two electrodes, and (iii) the nature of the emitter material.
4. The stopping potential (V_o) depends on (i) the frequency of incident light, and (ii) the nature of the emitter material. For a given frequency of incident light, it is independent of its intensity. The stopping potential is directly related to the maximum kinetic energy of electrons emitted: $e\,V_0 = (1/2)\,m\,v^2_{max} = K_{max}$.
5. Below a certain frequency (threshold frequency) ν_0, characteristic of the metal, no photoelectric emission takes place, no matter how large the intensity may be.
6. The classical wave theory could not explain the main features of photoelectric effect. Its picture of continuous absorption of energy from radiation could not explain the independence of K_{max} on intensity, the existence of ν_o and the instantaneous nature of the process. Einstein explained these features on the basis of photon picture of light. According to this, light is composed of discrete packets of energy called quanta or photons. Each photon carries an energy E (= $h\,\nu$) and momentum p (= h/λ), which depend on the frequency (ν) of incident light and not on its intensity. Photoelectric emission from the metal surface occurs due to absorption of a photon by an electron.
7. Einstein's photoelectric equation is in accordance with the energy conservation law as applied to the photon absorption by an electron in the metal. The maximum kinetic energy $(1/2)m\,v^2_{max}$ is equal to the photon energy ($h\nu$) minus the work function ϕ_0 (= $h\nu_0$) of the target metal:

$$\frac{1}{2}\,m\,v^2_{max} = V_0\,e = h\nu - \phi_0 = h\,(\nu - \nu_0)$$

   This photoelectric equation explains all the features of the photoelectric effect. Millikan's first precise measurements confirmed the Einstein's photoelectric equation and obtained an accurate value of Planck's constant h. This led to the acceptance of particle or photon description (nature) of electromagnetic radiation, introduced by Einstein.
8. Radiation has dual nature: wave and particle. The nature of experiment determines whether a wave or particle description is best suited for understanding the experimental result. Reasoning that radiation and matter should be symmetrical in nature, Louis Victor de Broglie

attributed a wave-like character to matter (material particles). The waves associated with the moving material particles are called matter waves or de Broglie waves.

9. The de Broglie wavelength (λ) associated with a moving particle is related to its momentum p as: $\lambda = h/p$. The dualism of matter is inherent in the de Broglie relation which contains a wave concept (λ) and a particle concept (p). The de Broglie wavelength is independent of the charge and nature of the material particle. It is significantly measurable (of the order of the atomic-planes spacing in crystals) only in case of sub-atomic particles like electrons, protons, etc. (due to smallness of their masses and hence, momenta). However, it is indeed very small, quite beyond measurement, in case of macroscopic objects, commonly encountered in everyday life.
10. Electron diffraction experiments by Davisson and Germer, and by G. P. Thomson, as well as many later experiments, have verified and confirmed the wave-nature of electrons. The de Broglie hypothesis of matter waves supports the Bohr's concept of stationary orbits.

Physical Quantity	Symbol	Dimensions	Unit	Remarks
Planck's constant	h	$[ML^2T^{-1}]$	J s	$E = h\nu$
Stopping potential	V_0	$[ML^2T^{-3}A^{-1}]$	V	$eV_0 = K_{max}$
Work function	ϕ_0	$[ML^2T^{-2}]$	J; eV	$K_{max} = E - \phi_0$
Threshold frequency	ν_0	$[T^{-1}]$	Hz	$\nu_0 = \phi_0/h$
de Broglie wavelength	λ	[L]	m	$\lambda = h/p$

POINTS TO PONDER

1. Free electrons in a metal are free in the sense that they move inside the metal in a constant potential (This is only an approximation). They are not free to move out of the metal. They need additional energy to get out of the metal.
2. Free electrons in a metal do not all have the same energy. Like molecules in a gas jar, the electrons have a certain energy distribution at a given temperature. This distribution is different from the usual Maxwell's distribution that you have learnt in the study of kinetic theory of gases. You will learn about it in later courses, but the difference has to do with the fact that electrons obey Pauli's exclusion principle.
3. Because of the energy distribution of free electrons in a metal, the energy required by an electron to come out of the metal is different for different electrons. Electrons with higher energy require less additional energy to come out of the metal than those with lower energies. Work function is the least energy required by an electron to come out of the metal.

4. Observations on photoelectric effect imply that in the event of matter-light interaction, *absorption of energy takes place in discrete units of hν*. This is not quite the same as saying that light consists of particles, each of energy *hν*.
5. Observations on the stopping potential (its independence of intensity and dependence on frequency) are the crucial discriminator between the wave-picture and photon-picture of photoelectric effect.
6. The wavelength of a matter wave given by $\lambda = \frac{h}{p}$ has physical significance; its phase velocity v_p has no physical significance. However, the group velocity of the matter wave is physically meaningful and equals the velocity of the particle.

EXERCISES

11.1 Find the
(a) maximum frequency, and
(b) minimum wavelength of X-rays produced by 30 kV electrons.

11.2 The work function of caesium metal is 2.14 eV. When light of frequency 6×10^{14}Hz is incident on the metal surface, photoemission of electrons occurs. What is the
(a) maximum kinetic energy of the emitted electrons,
(b) Stopping potential, and
(c) maximum speed of the emitted photoelectrons?

11.3 The photoelectric cut-off voltage in a certain experiment is 1.5 V. What is the maximum kinetic energy of photoelectrons emitted?

11.4 Monochromatic light of wavelength 632.8 nm is produced by a helium-neon laser. The power emitted is 9.42 mW.
(a) Find the energy and momentum of each photon in the light beam,
(b) How many photons per second, on the average, arrive at a target irradiated by this beam? (Assume the beam to have uniform cross-section which is less than the target area), and
(c) How fast does a hydrogen atom have to travel in order to have the same momentum as that of the photon?

11.5 The energy flux of sunlight reaching the surface of the earth is 1.388×10^3 W/m^2. How many photons (nearly) per square metre are incident on the Earth per second? Assume that the photons in the sunlight have an average wavelength of 550 nm.

11.6 In an experiment on photoelectric effect, the slope of the cut-off voltage versus frequency of incident light is found to be 4.12×10^{-15} V s. Calculate the value of Planck's constant.

11.7 A 100W sodium lamp radiates energy uniformly in all directions. The lamp is located at the centre of a large sphere that absorbs all the sodium light which is incident on it. The wavelength of the sodium light is 589 nm. (a) What is the energy per photon associated

with the sodium light? (b) At what rate are the photons delivered to the sphere?

11.8 The threshold frequency for a certain metal is 3.3×10^{14} Hz. If light of frequency 8.2×10^{14} Hz is incident on the metal, predict the cut-off voltage for the photoelectric emission.

11.9 The work function for a certain metal is 4.2 eV. Will this metal give photoelectric emission for incident radiation of wavelength 330 nm?

11.10 Light of frequency 7.21×10^{14} Hz is incident on a metal surface. Electrons with a maximum speed of 6.0×10^{5} m/s are ejected from the surface. What is the threshold frequency for photoemission of electrons?

11.11 Light of wavelength 488 nm is produced by an argon laser which is used in the photoelectric effect. When light from this spectral line is incident on the emitter, the stopping (cut-off) potential of photoelectrons is 0.38 V. Find the work function of the material from which the emitter is made.

11.12 Calculate the
(a) momentum, and
(b) de Broglie wavelength of the electrons accelerated through a potential difference of 56 V.

11.13 What is the
(a) momentum,
(b) speed, and
(c) de Broglie wavelength of an electron with kinetic energy of 120 eV.

11.14 The wavelength of light from the spectral emission line of sodium is 589 nm. Find the kinetic energy at which
(a) an electron, and
(b) a neutron, would have the same de Broglie wavelength.

11.15 What is the de Broglie wavelength of
(a) a bullet of mass 0.040 kg travelling at the speed of 1.0 km/s,
(b) a ball of mass 0.060 kg moving at a speed of 1.0 m/s, and
(c) a dust particle of mass 1.0×10^{-9} kg drifting with a speed of 2.2 m/s?

11.16 An electron and a photon each have a wavelength of 1.00 nm. Find
(a) their momenta,
(b) the energy of the photon, and
(c) the kinetic energy of electron.

11.17 (a) For what kinetic energy of a neutron will the associated de Broglie wavelength be 1.40×10^{-10} m?
(b) Also find the de Broglie wavelength of a neutron, in thermal equilibrium with matter, having an average kinetic energy of $(3/2)\ kT$ at 300 K.

11.18 Show that the wavelength of electromagnetic radiation is equal to the de Broglie wavelength of its quantum (photon).

11.19 What is the de Broglie wavelength of a nitrogen molecule in air at 300 K? Assume that the molecule is moving with the root-mean-square speed of molecules at this temperature. (Atomic mass of nitrogen = 14.0076 u)

ADDITIONAL EXERCISES

11.20 (a) Estimate the speed with which electrons emitted from a heated emitter of an evacuated tube impinge on the collector maintained at a potential difference of 500 V with respect to the emitter. Ignore the small initial speeds of the electrons. The *specific charge* of the electron, i.e., its e/m is given to be 1.76×10^{11} C kg^{-1}.

(b) Use the same formula you employ in (a) to obtain electron speed for an collector potential of 10 MV. Do you see what is wrong ? In what way is the formula to be modified?

11.21 (a) A monoenergetic electron beam with electron speed of 5.20×10^{6} m s^{-1} is subject to a magnetic field of 1.30×10^{-4} T normal to the beam velocity. What is the radius of the circle traced by the beam, given e/m for electron equals 1.76×10^{11}C kg^{-1}.

(b) Is the formula you employ in (a) valid for calculating radius of the path of a 20 MeV electron beam? If not, in what way is it modified?

[**Note:** Exercises 11.20(b) and 11.21(b) take you to relativistic mechanics which is beyond the scope of this book. They have been inserted here simply to emphasise the point that the formulas you use in part (a) of the exercises are not valid at very high speeds or energies. See answers at the end to know what 'very high speed or energy' means.]

11.22 An electron gun with its collector at a potential of 100 V fires out electrons in a spherical bulb containing hydrogen gas at low pressure (~10^{-2} mm of Hg). A magnetic field of 2.83×10^{-4} T curves the path of the electrons in a circular orbit of radius 12.0 cm. (The path can be viewed because the gas ions in the path focus the beam by attracting electrons, and emitting light by electron capture; this method is known as the 'fine beam tube' method.) Determine e/m from the data.

11.23 (a) An X-ray tube produces a continuous spectrum of radiation with its short wavelength end at 0.45 Å. What is the maximum energy of a photon in the radiation?

(b) From your answer to (a), guess what order of accelerating voltage (for electrons) is required in such a tube?

11.24 In an accelerator experiment on high-energy collisions of electrons with positrons, a certain event is interpreted as annihilation of an electron-positron pair of total energy 10.2 BeV into two γ-rays of equal energy. What is the wavelength associated with each γ-ray? (1BeV = 10^{9} eV)

11.25 Estimating the following two numbers should be interesting. The first number will tell you why radio engineers do not need to worry much about photons! The second number tells you why our eye can never 'count photons', even in barely detectable light.

(a) The number of photons emitted per second by a Medium wave transmitter of 10 kW power, emitting radiowaves of wavelength 500 m.

(b) The number of photons entering the pupil of our eye per second corresponding to the minimum intensity of white light that we

humans can perceive (~10^{-10} W m^{-2}). Take the area of the pupil to be about 0.4 cm^2, and the average frequency of white light to be about 6×10^{14} Hz.

11.26 Ultraviolet light of wavelength 2271 Å from a 100 W mercury source irradiates a photo-cell made of molybdenum metal. If the stopping potential is –1.3 V, estimate the work function of the metal. How would the photo-cell respond to a high intensity (~10^5 W m^{-2}) red light of wavelength 6328 Å produced by a He-Ne laser?

11.27 Monochromatic radiation of wavelength 640.2 nm (1nm = 10^{-9} m) from a neon lamp irradiates photosensitive material made of caesium on tungsten. The stopping voltage is measured to be 0.54 V. The source is replaced by an iron source and its 427.2 nm line irradiates the same photo-cell. Predict the new stopping voltage.

11.28 A mercury lamp is a convenient source for studying frequency dependence of photoelectric emission, since it gives a number of spectral lines ranging from the UV to the red end of the visible spectrum. In our experiment with rubidium photo-cell, the following lines from a mercury source were used:

λ_1 = 3650 Å, λ_2= 4047 Å, λ_3= 4358 Å, λ_4= 5461 Å, λ_5= 6907 Å,

The stopping voltages, respectively, were measured to be:

V_{01} = 1.28 V, V_{02} = 0.95 V, V_{03} = 0.74 V, V_{04} = 0.16 V, V_{05} = 0 V

Determine the value of Planck's constant h, the threshold frequency and work function for the material.

[*Note:* You will notice that to get h from the data, you will need to know e (which you can take to be 1.6×10^{-19} C). Experiments of this kind on Na, Li, K, etc. were performed by Millikan, who, using his own value of e (from the oil-drop experiment) confirmed Einstein's photoelectric equation and at the same time gave an independent estimate of the value of h.]

11.29 The work function for the following metals is given:

Na: 2.75 eV; K: 2.30 eV; Mo: 4.17 eV; Ni: 5.15 eV. Which of these metals will not give photoelectric emission for a radiation of wavelength 3300 Å from a He-Cd laser placed 1 m away from the photocell? What happens if the laser is brought nearer and placed 50 cm away?

11.30 Light of intensity 10^{-5} W m^{-2} falls on a sodium photo-cell of surface area 2 cm^2. Assuming that the top 5 layers of sodium absorb the incident energy, estimate time required for photoelectric emission in the wave-picture of radiation. The work function for the metal is given to be about 2 eV. What is the implication of your answer?

11.31 Crystal diffraction experiments can be performed using X-rays, or electrons accelerated through appropriate voltage. Which probe has greater energy? (For quantitative comparison, take the wavelength of the probe equal to 1 Å, which is of the order of inter-atomic spacing in the lattice) (m_e=9.11 × 10^{-31} kg).

11.32 (a) Obtain the de Broglie wavelength of a neutron of kinetic energy 150 eV. As you have seen in Exercise 11.31, an electron beam of this energy is suitable for crystal diffraction experiments. Would a neutron beam of the same energy be equally suitable? Explain. (m_n = 1.675 × 10^{-27} kg)

(b) Obtain the de Broglie wavelength associated with thermal neutrons at room temperature (27 °C). Hence explain why a fast neutron beam needs to be thermalised with the environment before it can be used for neutron diffraction experiments.

11.33 An electron microscope uses electrons accelerated by a voltage of 50 kV. Determine the de Broglie wavelength associated with the electrons. If other factors (such as numerical aperture, etc.) are taken to be roughly the same, how does the resolving power of an electron microscope compare with that of an optical microscope which uses yellow light?

11.34 The wavelength of a probe is roughly a measure of the size of a structure that it can probe in some detail. The quark structure of protons and neutrons appears at the minute length-scale of 10^{-15} m or less. This structure was first probed in early 1970's using high energy electron beams produced by a linear accelerator at Stanford, USA. Guess what might have been the order of energy of these electron beams. (Rest mass energy of electron = 0.511 MeV.)

11.35 Find the typical de Broglie wavelength associated with a He atom in helium gas at room temperature (27 °C) and 1 atm pressure; and compare it with the mean separation between two atoms under these conditions.

11.36 Compute the typical de Broglie wavelength of an electron in a metal at 27 °C and compare it with the mean separation between two electrons in a metal which is given to be about 2×10^{-10} m.

[*Note:* Exercises 11.35 and 11.36 reveal that while the wave-packets associated with gaseous molecules under ordinary conditions are non-overlapping, the electron wave-packets in a metal strongly overlap with one another. This suggests that whereas molecules in an ordinary gas can be distinguished apart, electrons in a metal cannot be distintguished apart from one another. This indistinguishibility has many fundamental implications which you will explore in more advanced Physics courses.]

11.37 Answer the following questions:

(a) Quarks inside protons and neutrons are thought to carry fractional charges [$(+2/3)e$; $(-1/3)e$]. Why do they not show up in Millikan's oil-drop experiment?

(b) What is so special about the combination e/m? Why do we not simply talk of e and m separately?

(c) Why should gases be insulators at ordinary pressures and start conducting at very low pressures?

(d) Every metal has a definite work function. Why do all photoelectrons not come out with the same energy if incident radiation is monochromatic? Why is there an energy distribution of photoelectrons?

(e) The energy and momentum of an electron are related to the frequency and wavelength of the associated matter wave by the relations:

$$E = h\nu,\ p = \frac{h}{\lambda}$$

But while the value of λ is physically significant, the value of ν (and therefore, the value of the phase speed $\nu\lambda$) has no physical significance. Why?

APPENDIX

11.1 The history of wave-particle flip-flop

What is light? This question has haunted mankind for a long time. But systematic experiments were done by scientists since the dawn of the scientific and industrial era, about four centuries ago. Around the same time, theoretical models about what light is made of were developed. While building a model in any branch of science, it is essential to see that it is able to explain all the experimental observations existing at that time. It is therefore appropriate to summarize some observations about light that were known in the seventeenth century.

The properties of light known at that time included (a) rectilinear propagation of light, (b) reflection from plane and curved surfaces, (c) refraction at the boundary of two media, (d) dispersion into various colours, (e) high speed. Appropriate laws were formulated for the first four phenomena. For example, Snell formulated his laws of refraction in 1621. Several scientists right from the days of Galileo had tried to measure the speed of light. But they had not been able to do so. They had only concluded that it was higher than the limit of their measurement.

Two models of light were also proposed in the seventeenth century. Descartes, in early decades of seventeenth century, proposed that light consists of particles, while Huygens, around 1650-60, proposed that light consists of waves. Descartes′ proposal was merely a philosophical model, devoid of any experiments or scientific arguments. Newton soon after, around 1660-70, extended Descartes′ particle model, known as *corpuscular theory*, built it up as a scientific theory, and explained various known properties with it. These models, light as waves and as particles, in a sense, are quite opposite of each other. But both models could explain all the known properties of light. There was nothing to choose between them.

The history of the development of these models over the next few centuries is interesting. Bartholinus, in 1669, discovered double refraction of light in some crystals, and Huygens, in 1678, was quick to explain it on the basis of his wave theory of light. In spite of this, for over one hundred years, Newton's particle model was firmly believed and preferred over the wave model. This was partly because of its simplicity and partly because of Newton's influence on contemporary physics.

Then in 1801, Young performed his double-slit experiment and observed interference fringes. This phenomenon could be explained only by wave theory. It was realized that diffraction was also another phenomenon which could be explained only by wave theory. In fact, it was a natural consequence of Huygens idea of secondary wavelets emanating from every point in the path of light. These experiments could not be explained by assuming that light consists of particles. Another phenomenon of polarisation was discovered around 1810, and this too could be naturally explained by the wave theory. Thus wave theory of Huygens came to the forefront and Newton's particle theory went into the background. This situation again continued for almost a century.

Better experiments were performed in the nineteenth century to determine the speed of light. With more accurate experiments, a value of 3×10^8 m/s for speed of light in vacuum was arrived at. Around 1860, Maxwell proposed his equations of electromagnetism and it was realized that *all* electromagnetic phenomena known at that time could be explained by Maxwell's four equations. Soon Maxwell showed that electric and magnetic fields could propagate through empty space (vacuum) in the form of electromagnetic waves. He calculated the speed of these waves and arrived at a theoretical value of 2.998×10^8 m/s. The close agreement of this value with the experimental value suggested that light consists of electromagnetic waves. In 1887 Hertz demonstrated the generation and detection of such waves. This established the wave theory of light on a firm footing. We might say that while eighteenth century belonged to the particle model, the nineteenth century belonged to the wave model of light.

Vast amounts of experiments were done during the period 1850-1900 on heat and related phenomena, an altogether different area of physics. Theories and models like kinetic theory and thermodynamics were developed which quite successfully explained the various phenomena, except one.

Dual Nature of Radiation and Matter

Every body at any temperature emits radiation of all wavelengths. It also absorbs radiation falling on it. A body which absorbs all the radiation falling on it is called a *black body*. It is an ideal concept in physics, like concepts of a point mass or uniform motion. A graph of the intensity of radiation emitted by a body versus wavelength is called the *black body spectrum*. No theory in those days could explain the complete black body spectrum!

In 1900, Planck hit upon a novel idea. If we assume, he said, that radiation is emitted in packets of energy instead of continuously as in a wave, then we can explain the black body spectrum. Planck himself regarded these quanta, or packets, as a property of emission and absorption, rather than that of light. He derived a formula which agreed with the entire spectrum. This was a confusing mixture of wave and particle pictures – radiation is emitted as a particle, it travels as a wave, and is again absorbed as a particle! Moreover, this put physicists in a dilemma. Should we again accept the particle picture of light just to explain one phenomenon? Then what happens to the phenomena of interference and diffraction which cannot be explained by the particle model?

But soon in 1905, Einstein explained the photoelectric effect by assuming the particle picture of light. In 1907, Debye explained the low temperature specific heats of solids by using the particle picture for lattice vibrations in a crystalline solid. Both these phenomena belonging to widely diverse areas of physics could be explained only by the particle model and not by the wave model. In 1923, Compton's x-ray scattering experiments from atoms also went in favour of the particle picture. This increased the dilemma further.

Thus by 1923, physicists faced with the following situation. (a) There were some phenomena like rectilinear propagation, reflection, refraction, which could be explained by either particle model or by wave model. (b) There were some phenomena such as diffraction and interference which could be explained only by the wave model but *not* by the particle model. (c) There were some phenomena such as black body radiation, photoelectric effect, and Compton scattering which could be explained only by the particle model but *not* by the wave model. Somebody in those days aptly remarked that light behaves as a particle on Mondays, Wednesdays and Fridays, and as a wave on Tuesdays, Thursdays and Saturdays, and we don't talk of light on Sundays!

In 1924, de Broglie proposed his theory of wave-particle duality in which he said that not only photons of light but also 'particles' of matter such as electrons and atoms possess a dual character, sometimes behaving like a particle and sometimes as a wave. He gave a formula connecting their mass, velocity, momentum (particle characteristics), with their wavelength and frequency (wave characteristics)! In 1927 Thomson, and Davisson and Germer, in separate experiments, showed that electrons did behave like waves with a wavelength which agreed with that given by de Broglie's formula. Their experiment was on diffraction of electrons through crystalline solids, in which the regular arrangement of atoms acted like a grating. Very soon, diffraction experiments with other 'particles' such as neutrons and protons were performed and these too confirmed with de Broglie's formula. This confirmed wave-particle duality as an established principle of physics. Here was a principle, physicists thought, which explained all the phenomena mentioned above not only for light but also for the so-called particles.

But there was no basic theoretical foundation for wave-particle duality. De Broglie's proposal was merely a qualitative argument based on symmetry of nature. Wave-particle duality was at best a principle, not an outcome of a sound fundamental theory. It is true that all experiments whatever agreed with de Broglie formula. But physics does not work that way. On the one hand, it needs experimental confirmation, while on the other hand, it also needs sound theoretical basis for the models proposed. This was developed over the next two decades. Dirac developed his theory of radiation in about 1928, and Heisenberg and Pauli gave it a firm footing by 1930. Tomonaga, Schwinger, and Feynman, in late 1940s, produced further refinements and cleared the theory of inconsistencies which were noticed. All these theories mainly put wave-particle duality on a theoretical footing.

Although the story continues, it grows more and more complex and beyond the scope of this note. But we have here the essential structure of what happened, and let us be satisfied with it at the moment. Now it is regarded as a natural consequence of present theories of physics that electromagnetic radiation as well as particles of matter exhibit both wave and particle properties in different experiments, and sometimes even in the different parts of the same experiment.

Chapter Twelve

ATOMS

12.1 Introduction

By the nineteenth century, enough evidence had accumulated in favour of atomic hypothesis of matter. In 1897, the experiments on electric discharge through gases carried out by the English physicist J. J. Thomson (1856 – 1940) revealed that atoms of different elements contain negatively charged constituents (electrons) that are identical for all atoms. However, atoms on a whole are electrically neutral. Therefore, an atom must also contain some positive charge to neutralise the negative charge of the electrons. But what is the arrangement of the positive charge and the electrons inside the atom? In other words, what is the structure of an atom?

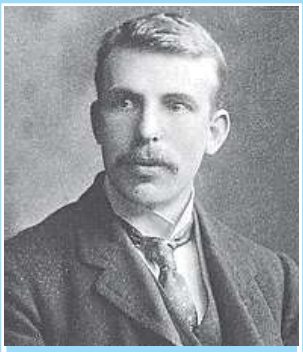
The first model of atom was proposed by J. J. Thomson in 1898. According to this model, the positive charge of the atom is uniformly distributed throughout the volume of the atom and the negatively charged electrons are embedded in it like seeds in a watermelon. This model was picturesquely called *plum pudding model* of the atom. However subsequent studies on atoms, as described in this chapter, showed that the distribution of the electrons and positive charges are very different from that proposed in this model.

We know that condensed matter (solids and liquids) and dense gases at all temperatures emit electromagnetic radiation in which a continuous distribution of several wavelengths is present, though with different intensities. This radiation is considered to be due to oscillations of atoms

and molecules, governed by the interaction of each atom or molecule with its neighbours. *In contrast*, light emitted from rarefied gases heated in a flame, or excited electrically in a glow tube such as the familiar neon sign or mercury vapour light has only certain discrete wavelengths. The spectrum appears as a series of bright lines. In such gases, the average spacing between atoms is large. Hence, the radiation emitted can be considered due to individual atoms rather than because of interactions between atoms or molecules.

In the early nineteenth century it was also established that each element is associated with a characteristic spectrum of radiation, for example, hydrogen always gives a set of lines with fixed relative position between the lines. This fact suggested an intimate relationship between the internal structure of an atom and the spectrum of radiation emitted by it. In 1885, Johann Jakob Balmer (1825 – 1898) obtained a simple empirical formula which gave the wavelengths of a group of lines emitted by atomic hydrogen. Since hydrogen is simplest of the elements known, we shall consider its spectrum in detail in this chapter.

Ernst Rutherford (1871–1937), a former research student of J. J. Thomson, was engaged in experiments on α-particles emitted by some radioactive elements. In 1906, he proposed a classic experiment of scattering of these α-particles by atoms to investigate the atomic structure. This experiment was later performed around 1911 by Hans Geiger (1882–1945) and Ernst Marsden (1889–1970, who was 20 year-old student and had not yet earned his bachelor's degree). The details are discussed in Section 12.2. The explanation of the results led to the birth of Rutherford's planetary model of atom (also called the *nuclear model of the atom*). According to this the entire positive charge and most of the mass of the atom is concentrated in a small volume called the nucleus with electrons revolving around the nucleus just as planets revolve around the sun.



Ernst Rutherford (1871 – 1937) New Zealand born, British physicist who did pioneering work on radioactive radiation. He discovered alpha-rays and beta-rays. Along with Federick Soddy, he created the modern theory of radioactivity. He studied the 'emanation' of thorium and discovered a new noble gas, an isotope of radon, now known as thoron. By scattering alpha-rays from the metal foils, he discovered the atomic nucleus and proposed the plenatery model of the atom. He also estimated the approximate size of the nucleus.

ERNST RUTHERFORD (1871 – 1937)

Rutherford's nuclear model was a major step towards how we see the atom today. However, it could not explain why atoms emit light of only discrete wavelengths. How could an atom as simple as hydrogen, consisting of a single electron and a single proton, emit a complex spectrum of specific wavelengths? In the classical picture of an atom, the electron revolves round the nucleus much like the way a planet revolves round the sun. However, we shall see that there are some serious difficulties in accepting such a model.

12.2 Alpha-particle Scattering and Rutherford's Nuclear Model of Atom

At the suggestion of Ernst Rutherford, in 1911, H. Geiger and E. Marsden performed some experiments. In one of their experiments, as shown in

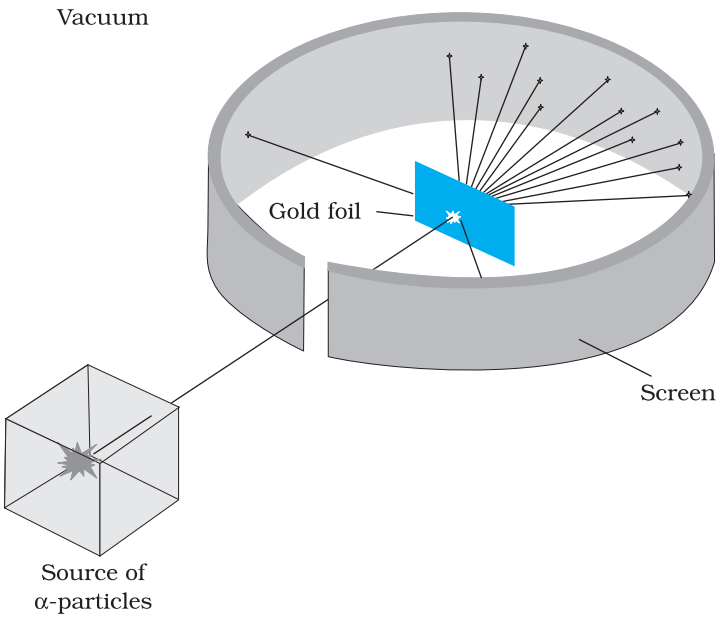


FIGURE 12.1 Geiger-Marsden scattering experiment. The entire apparatus is placed in a vacuum chamber (not shown in this figure).

Fig. 12.1, they directed a beam of 5.5 MeV α-particles emitted from a $^{214}_{83}Bi$ radioactive source at a thin metal foil made of gold. Figure 12.2 shows a schematic diagram of this experiment. Alpha-particles emitted by a $^{214}_{83}Bi$ radioactive source were collimated into a narrow beam by their passage through lead bricks. The beam was allowed to fall on a thin foil of gold of thickness 2.1×10^{-7} m. The scattered alpha-particles were observed through a rotatable detector consisting of zinc sulphide screen and a microscope. The scattered alpha-particles on striking the screen produced brief light flashes or scintillations. These flashes may be viewed through a microscope and the distribution of the number of scattered particles may be studied as a function of angle of scattering.

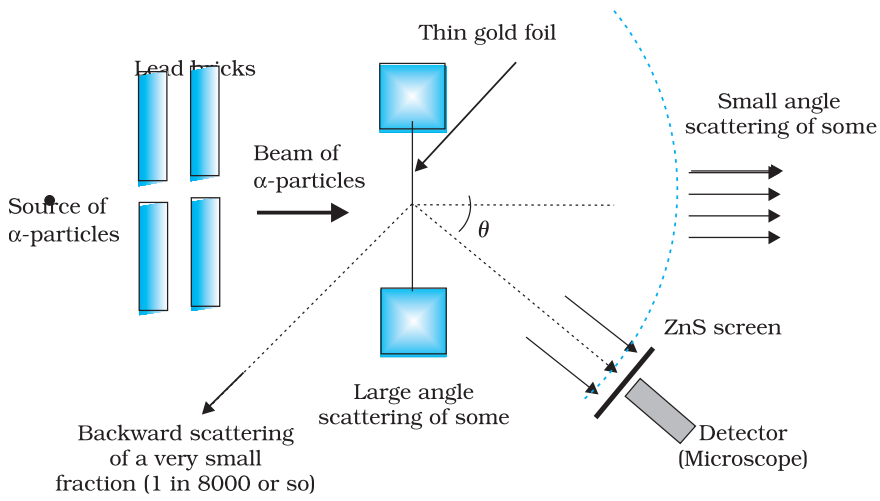


FIGURE 12.2 Schematic arrangement of the Geiger-Marsden experiment.

A typical graph of the total number of α-particles scattered at different angles, in a given interval of time, is shown in Fig. 12.3. The dots in this figure represent the data points and the solid curve is the theoretical prediction based on the assumption that the target atom has a small, dense, positively charged nucleus. Many of the α-particles pass through the foil. It means that they do not suffer any collisions. Only about 0.14% of the incident α-particles scatter by more than 1°; and about 1 in 8000 deflect by more than 90°. Rutherford argued that, to deflect the α-particle backwards, it must experience a large repulsive force. This force could

be provided if the greater part of the mass of the atom and its positive charge were concentrated tightly at its centre. Then the incoming α-particle could get very close to the positive charge without penetrating it, and such a close encounter would result in a large deflection. This agreement supported the hypothesis of the nuclear atom. This is why Rutherford is credited with the *discovery* of the nucleus.

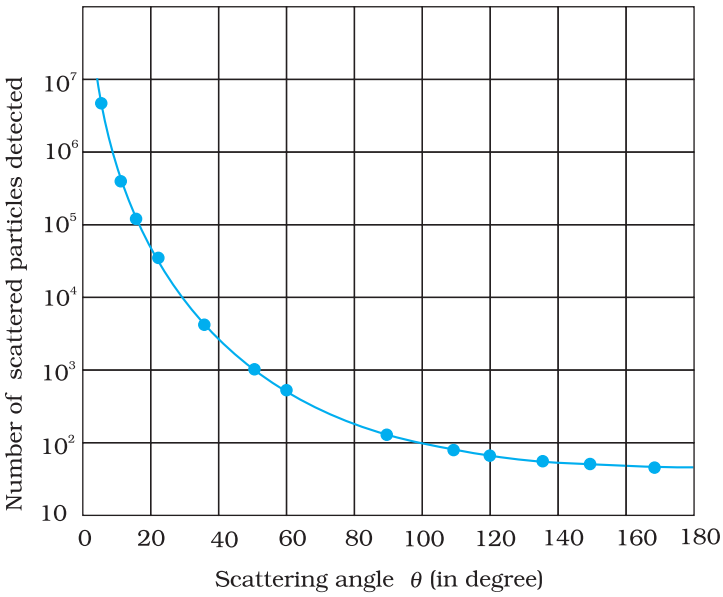


FIGURE 12.3 Experimental data points (shown by dots) on scattering of α-particles by a thin foil at different angles obtained by Geiger and Marsden using the setup shown in Figs. 12.1 and 12.2. Rutherford's nuclear model predicts the solid curve which is seen to be in good agreement with experiment.

In Rutherford's nuclear model of the atom, the entire positive charge and most of the mass of the atom are concentrated in the nucleus with the electrons some distance away. The electrons would be moving in orbits about the nucleus just as the planets do around the sun. Rutherford's experiments suggested the size of the nucleus to be about 10^{-15} m to 10^{-14} m. From kinetic theory, the size of an atom was known to be 10^{-10} m, about 10,000 to 100,000 times larger than the size of the nucleus (see Chapter 11, Section 11.6 in Class XI Physics textbook). Thus, the electrons would seem to be at a distance from the nucleus of about 10,000 to 100,000 times the size of the nucleus itself. Thus, most of an atom is empty space. With the atom being largely empty space, it is easy to see why most α-particles go right through a thin metal foil. However, when α-particle happens to come near a nucleus, the intense electric field there scatters it through a large angle. The atomic electrons, being so light, do not appreciably affect the α-particles.

The scattering data shown in Fig. 12.3 can be analysed by employing Rutherford's nuclear model of the atom. As the gold foil is very thin, it can be assumed that α-particles will suffer not more than one scattering during their passage through it. Therefore, computation of the trajectory of an alpha-particle scattered by a single nucleus is enough. Alpha-particles are nuclei of helium atoms and, therefore, carry two units, $2e$, of positive charge and have the mass of the helium atom. The charge of the gold nucleus is Ze, where Z is the atomic number of the atom; for gold $Z = 79$. Since the nucleus of gold is about 50 times heavier than an α-particle, it is reasonable to assume that it remains stationary throughout the scattering process. Under these assumptions, the trajectory of an alpha-particle can be computed employing Newton's second law of motion and the Coulomb's law for electrostatic force of repulsion between the alpha-particle and the positively charged nucleus.

The magnitude of this force is

$$F = \frac{1}{4\pi\varepsilon_0}\frac{(2e)(Ze)}{r^2} \qquad (12.1)$$

where r is the distance between the α-particle and the nucleus. The force is directed along the line joining the α-particle and the nucleus. The magnitude and direction of the force on an α-particle continuously changes as it approaches the nucleus and recedes away from it.

12.2.1 Alpha-particle trajectory

The trajectory traced by an α-particle depends on the impact parameter, b of collision. The *impact parameter* is the perpendicular distance of the initial velocity vector of the α-particle from the centre of the nucleus (Fig. 12.4). A given beam of α-particles has a distribution of impact parameters b, so that the beam is scattered in various directions with different probabilities (Fig. 12.4). (In a beam, all particles have nearly same kinetic energy.) It is seen that an α-particle close to the nucleus (small impact parameter) suffers large scattering. In case of head-on collision, the impact parameter is minimum and the α-particle rebounds back ($\theta \cong \pi$). For a large impact parameter, the α-particle goes nearly undeviated and has a small deflection ($\theta \cong 0$).

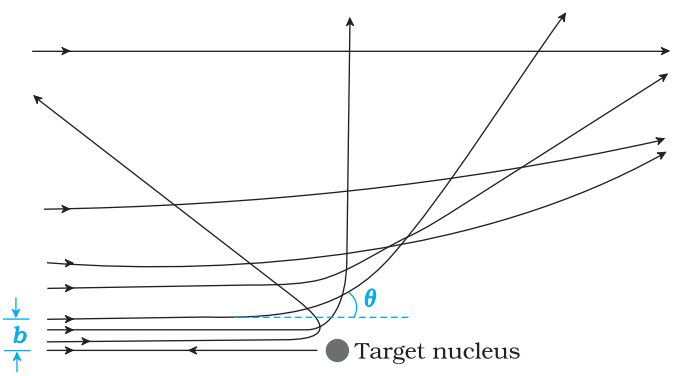


FIGURE 12.4 Trajectory of α-particles in the coulomb field of a target nucleus. The impact parameter, b and scattering angle θ are also depicted.

The fact that only a small fraction of the number of incident particles rebound back indicates that the number of α-particles undergoing head on collision is small. This, in turn, implies that the mass of the atom is concentrated in a small volume. Rutherford scattering therefore, is a powerful way to determine an upper limit to the size of the nucleus.

EXAMPLE 12.1

Example 12.1 In the Rutherford's nuclear model of the atom, the nucleus (radius about 10^{-15} m) is analogous to the sun about which the electron move in orbit (radius ≈ 10^{-10} m) like the earth orbits around the sun. If the dimensions of the solar system had the same proportions as those of the atom, would the earth be closer to or farther away from the sun than actually it is? The radius of earth's orbit is about 1.5×10^{11} m. The radius of sun is taken as 7×10^8 m.

Solution The ratio of the radius of electron's orbit to the radius of nucleus is $(10^{-10}\text{ m})/(10^{-15}\text{ m}) = 10^5$, that is, the radius of the electron's orbit is 10^5 times larger than the radius of nucleus. If the radius of the earth's orbit around the sun were 10^5 times larger than the radius of the sun, the radius of the earth's orbit would be $10^5 \times 7 \times 10^8$ m = 7×10^{13} m. This is more than 100 times greater than the actual orbital radius of earth. Thus, the earth would be much farther away from the sun.

It implies that an atom contains a much greater fraction of empty space than our solar system does.

Example 12.2 In a Geiger-Marsden experiment, what is the distance of closest approach to the nucleus of a 7.7 MeV α-particle before it comes momentarily to rest and reverses its direction?

Solution The key idea here is that throughout the scattering process, the total mechanical energy of the system consisting of an α-particle and a gold nucleus is conserved. The system's initial mechanical energy is E_i, before the particle and nucleus interact, and it is equal to its mechanical energy E_f when the α-particle momentarily stops. The initial energy E_i is just the kinetic energy K of the incoming α- particle. The final energy E_f is just the electric potential energy U of the system. The potential energy U can be calculated from Eq. (12.1).

Let d be the centre-to-centre distance between the α-particle and the gold nucleus when the α-particle is at its stopping point. Then we can write the conservation of energy $E_i = E_f$ as

$$K = \frac{1}{4\pi\varepsilon_0}\frac{(2e)(Ze)}{d} = \frac{2Ze^2}{4\pi\varepsilon_0 d}$$

Thus the distance of closest approach d is given by

$$d = \frac{2Ze^2}{4\pi\varepsilon_0 K}$$

The maximum kinetic energy found in α-particles of natural origin is 7.7 MeV or 1.2×10^{-12} J. Since $1/4\pi\varepsilon_0 = 9.0 \times 10^9$ N m^2/C^2. Therefore with $e = 1.6 \times 10^{-19}$ C, we have,

$$d = \frac{(2)(9.0\times10^9\,\text{Nm}^2/C^2)(1.6\times10^{-19}\,C)^2 Z}{1.2\times10^{-12}\,\text{J}}$$

$$= 3.84 \times 10^{-16}\, Z\,\text{m}$$

The atomic number of foil material gold is $Z = 79$, so that

d (Au) = 3.0×10^{-14} m = 30 fm. (1 fm (i.e. fermi) = 10^{-15} m.)

The radius of gold nucleus is, therefore, less than 3.0×10^{-14} m. This is not in very good agreement with the observed result as the actual radius of gold nucleus is 6 fm. The cause of discrepancy is that the distance of closest approach is considerably larger than the sum of the radii of the gold nucleus and the α-particle. Thus, the α-particle reverses its motion without ever actually *touching* the gold nucleus.

EXAMPLE 12.2

PHYSICS

Simulate Rutherford scattering experiment
http://www-outreach.phy.cam.ac.uk/camphy/nucleus/nucleus6_1.htm

12.2.2 Electron orbits

The Rutherford nuclear model of the atom which involves classical concepts, pictures the atom as an electrically neutral sphere consisting of a very small, massive and positively charged nucleus at the centre surrounded by the revolving electrons in their respective dynamically stable orbits. The electrostatic force of attraction, F_e between the revolving electrons and the nucleus provides the requisite centripetal force (F_c) to keep them in their orbits. Thus, for a dynamically stable orbit in a hydrogen atom

$$F_e = F_c$$

$$\frac{mv^2}{r} = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{r^2} \qquad (12.2)$$

Thus the relation between the orbit radius and the electron velocity is

$$r = \frac{e^2}{4\pi\varepsilon_0 m v^2} \tag{12.3}$$

The kinetic energy (K) and electrostatic potential energy (U) of the electron in hydrogen atom are

$$K = \frac{1}{2}mv^2 = \frac{e^2}{8\pi\varepsilon_0 r} \text{ and } U = -\frac{e^2}{4\pi\varepsilon_0 r}$$

(The negative sign in U signifies that the electrostatic force is in the $-r$ direction.) Thus the total energy E of the electron in a hydrogen atom is

$$E = K + U = \frac{e^2}{8\pi\varepsilon_0 r} - \frac{e^2}{4\pi\varepsilon_0 r}$$

$$= -\frac{e^2}{8\pi\varepsilon_0 r} \tag{12.4}$$

The total energy of the electron is negative. This implies the fact that the electron is bound to the nucleus. If E were positive, an electron will not follow a closed orbit around the nucleus.

EXAMPLE 12.3

Example 12.3 It is found experimentally that 13.6 eV energy is required to separate a hydrogen atom into a proton and an electron. Compute the orbital radius and the velocity of the electron in a hydrogen atom.

Solution Total energy of the electron in hydrogen atom is –13.6 eV = $-13.6 \times 1.6 \times 10^{-19}$ J = -2.2×10^{-18} J. Thus from Eq. (12.4), we have

$$-\frac{e^2}{8\pi\varepsilon_0 r} = -2.2 \times 10^{-18} \text{ J}$$

This gives the orbital radius

$$r = -\frac{e^2}{8\pi\varepsilon_0 E} = -\frac{(9\times10^9 \text{ N m}^2/\text{C}^2)(1.6 \times 10^{-19}\text{C})^2}{(2)(-2.2\times10^{-18} \text{ J})}$$

$$= 5.3 \times 10^{-11} \text{ m.}$$

The velocity of the revolving electron can be computed from Eq. (12.3) with $m = 9.1 \times 10^{-31}$ kg,

$$v = \frac{e}{\sqrt{4\pi\varepsilon_0 m r}} = 2.2\times10^6 \text{ m/s.}$$

12.3 Atomic Spectra

As mentioned in Section 12.1, each element has a characteristic spectrum of radiation, which it emits. When an atomic gas or vapour is excited at low pressure, usually by passing an electric current through it, the emitted radiation has a spectrum which contains certain specific wavelengths only. A spectrum of this kind is termed as emission line spectrum and it

consists of bright lines on a dark background. The spectrum emitted by atomic hydrogen is shown in Fig. 12.5. Study of emission line spectra of a material can therefore serve as a type of "fingerprint" for identification of the gas. When white light passes through a gas and we analyse the transmitted light using a spectrometer we find some dark lines in the spectrum. These dark lines correspond precisely to those wavelengths which were found in the emission line spectrum of the gas. This is called the *absorption spectrum* of the material of the gas.

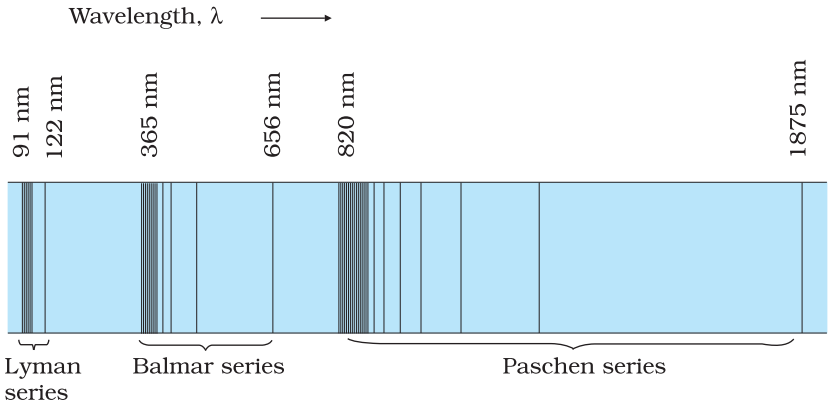


FIGURE 12.5 Emission lines in the spectrum of hydrogen.

12.3.1 Spectral series

We might expect that the frequencies of the light emitted by a particular element would exhibit some regular pattern. Hydrogen is the simplest atom and therefore, has the simplest spectrum. In the observed spectrum, however, at first sight, there does not seem to be any resemblance of order or regularity in spectral lines. But the spacing between lines within certain sets of the hydrogen spectrum decreases in a regular way (Fig. 12.5). Each of these sets is called a *spectral series*. In 1885, the first such series was observed by a Swedish school teacher Johann Jakob Balmer (1825–1898) in the visible region of the hydrogen spectrum. This series is called *Balmer series* (Fig. 12.6). The line with the longest wavelength, 656.3 nm in the red is called H_α; the next line with wavelength 486.1 nm in the blue-green is called H_β, the third line 434.1 nm in the violet is called H_γ; and so on. As the wavelength decreases, the lines appear closer together and are weaker in intensity. Balmer found a simple empirical formula for the observed wavelengths

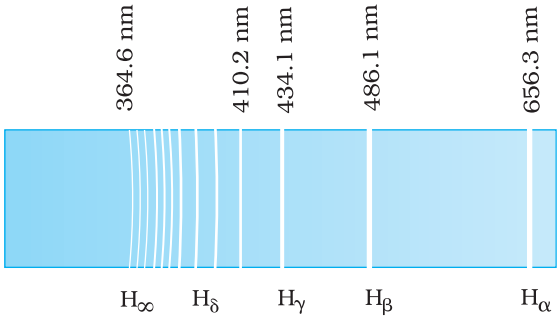


FIGURE 12.6 Balmer series in the emission spectrum of hydrogen.

$$\frac{1}{\lambda} = R\left(\frac{1}{2^2} - \frac{1}{n^2}\right) \qquad (12.5)$$

where λ is the wavelength, R is a constant called the *Rydberg constant*, and n may have integral values 3, 4, 5, etc. The value of R is $1.097 \times 10^7\ \text{m}^{-1}$. This equation is also called Balmer formula.

Taking $n = 3$ in Eq. (12.5), one obtains the wavelength of the H_α line:

$$\frac{1}{\lambda} = 1.097 \times 10^7 \left(\frac{1}{2^2} - \frac{1}{3^2}\right) \text{m}^{-1}$$

$$= 1.522 \times 10^6\ \text{m}^{-1}$$

i.e., $\lambda = 656.3$ nm

For $n = 4$, one obtains the wavelength of H_β line, etc. For $n = \infty$, one obtains the limit of the series, at $\lambda = 364.6$ nm. This is the shortest wavelength in the Balmer series. Beyond this limit, no further distinct lines appear, instead only a faint continuous spectrum is seen.

Other series of spectra for hydrogen were subsequently discovered. These are known, after their discoverers, as Lyman, Paschen, Brackett, and Pfund series. These are represented by the formulae:

Lyman series:

$$\frac{1}{\lambda} = R\left(\frac{1}{1^2} - \frac{1}{n^2}\right) \qquad n = 2,3,4... \tag{12.6}$$

Paschen series:

$$\frac{1}{\lambda} = R\left(\frac{1}{3^2} - \frac{1}{n^2}\right) \qquad n = 4,5,6... \tag{12.7}$$

Brackett series:

$$\frac{1}{\lambda} = R\left(\frac{1}{4^2} - \frac{1}{n^2}\right) \qquad n = 5,6,7... \tag{12.8}$$

Pfund series:

$$\frac{1}{\lambda} = R\left(\frac{1}{5^2} - \frac{1}{n^2}\right) \qquad n = 6,7,8... \tag{12.9}$$

The Lyman series is in the ultraviolet, and the Paschen and Brackett series are in the infrared region.

The Balmer formula Eq. (12.5) may be written in terms of frequency of the light, recalling that

$$c = \nu\lambda$$

or $\frac{1}{\lambda} = \frac{\nu}{c}$

Thus, Eq. (12.5) becomes

$$\nu = Rc\left(\frac{1}{2^2} - \frac{1}{n^2}\right) \tag{12.10}$$

There are only a few elements (hydrogen, singly ionised helium, and doubly ionised lithium) whose spectra can be represented by simple formula like Eqs. (12.5) – (12.9).

Equations (12.5) – (12.9) are useful as they give the wavelengths that hydrogen atoms radiate or absorb. However, these results are empirical and do not give any reasoning why only certain frequencies are observed in the hydrogen spectrum.

12.4 Bohr Model of the Hydrogen Atom

The model of the atom proposed by Rutherford assumes that the atom, consisting of a central nucleus and revolving electron is stable much like sun-planet system which the model imitates. However, there are some fundamental differences between the two situations. While the planetary system is held by gravitational force, the nucleus-electron system being charged objects, interact by Coulomb's Law of force. We know that an

object which moves in a circle is being constantly accelerated – the acceleration being centripetal in nature. According to classical electromagnetic theory, an accelerating charged particle emits radiation in the form of electromagnetic waves. The energy of an accelerating electron should therefore, continuously decrease. The electron would spiral inward and eventually fall into the nucleus (Fig. 12.7). Thus, such an atom can not be stable. Further, according to the classical electromagnetic theory, the frequency of the electromagnetic waves emitted by the revolving electrons is equal to the frequency of revolution. As the electrons spiral inwards, their angular velocities and hence their frequencies would change continuously, and so will the frequency of the light emitted. Thus, they would emit a continuous spectrum, in contradiction to the line spectrum actually observed. Clearly Rutherford model tells only a part of the story implying that the classical ideas are not sufficient to explain the atomic structure.

Niels Henrik David Bohr (1885 – 1962) Danish physicist who explained the spectrum of hydrogen atom based on quantum ideas. He gave a theory of nuclear fission based on the liquid-drop model of nucleus. Bohr contributed to the clarification of conceptual problems in quantum mechanics, in particular by proposing the complementary principle.

NIELS HENRIK DAVID BOHR (1885 – 1962)

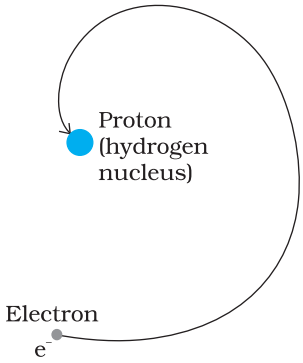


FIGURE 12.7 An accelerated atomic electron must spiral into the nucleus as it loses energy.

Example 12.4 According to the classical electromagnetic theory, calculate the initial frequency of the light emitted by the electron revolving around a proton in hydrogen atom.

Solution From Example 12.3 we know that velocity of electron moving around a proton in hydrogen atom in an orbit of radius 5.3×10^{-11} m is 2.2×10^{-6} m/s. Thus, the frequency of the electron moving around the proton is

$$\nu = \frac{v}{2\pi r} = \frac{2.2 \times 10^{6}\ \text{m s}^{-1}}{2\pi\left(5.3 \times 10^{-11}\ \text{m}\right)}$$

$$\approx 6.6 \times 10^{15}\ \text{Hz}.$$

According to the classical electromagnetic theory we know that the frequency of the electromagnetic waves emitted by the revolving electrons is equal to the frequency of its revolution around the nucleus. Thus the initial frequency of the light emitted is 6.6×10^{15} Hz.

EXAMPLE 12.4

It was Niels Bohr (1885 – 1962) who made certain modifications in this model by adding the ideas of the newly developing quantum hypothesis. Niels Bohr studied in Rutherford's laboratory for several months in 1912 and he was convinced about the validity of Rutherford nuclear model. Faced with the dilemma as discussed above, Bohr, in 1913, concluded that in spite of the success of electromagnetic theory in explaining large-scale phenomena, it could not be applied to the processes at the atomic scale. It became clear that a fairly radical departure from the established principles of classical mechanics and electromagnetism would be needed to understand the structure of atoms and the relation of atomic structure to atomic spectra. Bohr combined classical and early quantum concepts and gave his theory in the form of three postulates. These are :

(i) Bohr's first postulate was that *an electron in an atom could revolve in certain stable orbits without the emission of radiant energy*, contrary to the predictions of electromagnetic theory. According to this postulate, each atom has certain definite stable states in which it can exist, and each possible state has definite total energy. These are called the stationary states of the atom.

(ii) Bohr's second postulate defines these stable orbits. This postulate states that the *electron* revolves around the nucleus *only in those orbits for which the angular momentum is some integral multiple of* $h/2\pi$ where h is the Planck's constant (= 6.6×10^{-34} J s). Thus the angular momentum (L) of the orbiting electron is quantised. That is

$$L = nh/2\pi \tag{12.11}$$

(iii) Bohr's third postulate incorporated into atomic theory the early quantum concepts that had been developed by Planck and Einstein. It states that *an electron might make a transition from one of its specified non-radiating orbits to another of lower energy. When it does so, a photon is emitted having energy equal to the energy difference between the initial and final states. The frequency of the emitted photon is then given by*

$$h\nu = E_i - E_f \tag{12.12}$$

where E_i and E_f are the energies of the initial and final states and $E_i > E_f$.

For a hydrogen atom, Eq. (12.4) gives the expression to determine the energies of different energy states. But then this equation requires the radius r of the electron orbit. To calculate r, Bohr's second postulate about the angular momentum of the electron–the quantisation condition – is used. The angular momentum L is given by

$$L = mvr$$

Bohr's second postulate of quantisation [Eq. (12.11)] says that the allowed values of angular momentum are integral multiples of $h/2\pi$.

$$L_n = mv_n r_n = \frac{nh}{2\pi} \tag{12.13}$$

where n is an integer, r_n is the radius of nth possible orbit and v_n is the speed of moving electron in the nth orbit. The allowed orbits are numbered

1, 2, 3 ..., according to the values of n, which is called the *principal quantum number* of the orbit.

From Eq. (12.3), the relation between v_n and r_n is

$$v_n = \frac{e}{\sqrt{4\pi\varepsilon_0 m r_n}}$$

Combining it with Eq. (12.13), we get the following expressions for v_n and r_n,

$$v_n = \frac{1}{n}\frac{e^2}{4\pi\varepsilon_0}\frac{1}{(h/2\pi)} \tag{12.14}$$

and

$$r_n = \left(\frac{n^2}{m}\right)\left(\frac{h}{2\pi}\right)^2 \frac{4\pi\varepsilon_0}{e^2} \tag{12.15}$$

Eq. (12.14) depicts that the orbital speed in the nth orbit falls by a factor of n. Using Eq. (12.15), the size of the innermost orbit (n = 1) can be obtained as

$$r_1 = \frac{h^2\varepsilon_0}{\pi m e^2}$$

This is called the *Bohr radius*, represented by the symbol a_0. Thus,

$$a_0 = \frac{h^2\varepsilon_0}{\pi m e^2} \tag{12.16}$$

Substitution of values of h, m, ε_0 and e gives $a_0 = 5.29 \times 10^{-11}$ m. From Eq. (12.15), it can also be seen that the radii of the orbits increase as n^2.

The total energy of the electron in the stationary states of the hydrogen atom can be obtained by substituting the value of orbital radius in Eq. (12.4) as

$$E_n = -\left(\frac{e^2}{8\pi\varepsilon_0}\right)\left(\frac{m}{n^2}\right)\left(\frac{2\pi}{h}\right)^2\left(\frac{e^2}{4\pi\varepsilon_0}\right)$$

or $$E_n = -\frac{me^4}{8n^2\varepsilon_0^2 h^2} \tag{12.17}$$

Substituting values, Eq. (12.17) yields

$$E_n = -\frac{2.18\times10^{-18}}{n^2}\text{J} \tag{12.18}$$

Atomic energies are often expressed in electron volts (eV) rather than joules. Since 1 eV = 1.6×10^{-19} J, Eq. (12.18) can be rewritten as

$$E_n = -\frac{13.6}{n^2}\ \text{eV} \tag{12.19}$$

The negative sign of the total energy of an electron moving in an orbit means that the electron is bound with the nucleus. Energy will thus be required to remove the electron from the hydrogen atom to a distance infinitely far away from its nucleus (or proton in hydrogen atom).

The derivation of Eqs. (12.17) – (12.19) involves the assumption that the electronic orbits are circular, though orbits under inverse square force are, in general elliptical. (Planets move in elliptical orbits under the inverse square gravitational force of the sun.) However, it was shown by the German physicist Arnold Sommerfeld (1868 – 1951) that, when the restriction of circular orbit is relaxed, these equations continue to hold even for elliptic orbits.

Orbit vs state (orbital picture) of electron in atom

We are introduced to the Bohr Model of atom one time or the other in the course of physics. This model has its place in the history of quantum mechanics and particularly in explaining the structure of an atom. It has become a milestone since Bohr introduced the revolutionary idea of definite energy orbits for the electrons, contrary to the classical picture requiring an accelerating particle to radiate. Bohr also introduced the idea of quantisation of angular momentum of electrons moving in definite orbits. Thus it was a semi-classical picture of the structure of atom.

Now with the development of quantum mechanics, we have a better understanding of the structure of atom. Solutions of the Schrödinger wave equation assign a wave-like description to the electrons bound in an atom due to attractive forces of the protons.

An orbit of the electron in the Bohr model is the circular path of motion of an electron around the nucleus. But according to quantum mechanics, we cannot associate a definite path with the motion of the electrons in an atom. We can only talk about the probability of finding an electron in a certain region of space around the nucleus. This probability can be inferred from the one-electron wave function called the *orbital*. This function depends only on the coordinates of the electron.

It is therefore essential that we understand the subtle differences that exist in the two models:

- Bohr model is valid for only one-electron atoms/ions; an energy value, assigned to each orbit, depends on the principal quantum number n in this model. We know that energy associated with a stationary state of an electron depends on n only, for one-electron atoms/ions. For a multi-electron atom/ion, this is not true.
- The solution of the Schrödinger wave equation, obtained for hydrogen-like atoms/ions, called the wave function, gives information about the probability of finding an electron in various regions around the nucleus. This *orbital* has no resemblance whatsoever with the *orbit* defined for an electron in the Bohr model.

Example 12.5

Example 12.5 A 10 kg satellite circles earth once every 2 h in an orbit having a radius of 8000 km. Assuming that Bohr's angular momentum postulate applies to satellites just as it does to an electron in the hydrogen atom, find the quantum number of the orbit of the satellite.

Solution

From Eq. (12.13), we have

$m\, v_n\, r_n = nh/2\pi$

Here m = 10 kg and $r_n = 8 \times 10^6$ m. We have the time period T of the circling satellite as 2 h. That is T = 7200 s.
Thus the velocity $v_n = 2\pi\, r_n/T$.
The quantum number of the orbit of satellite
$n = (2\pi\, r_n)^2 \times m/(T \times h)$.
Substituting the values,
$n = (2\pi \times 8 \times 10^6 \text{ m})^2 \times 10/(7200 \text{ s} \times 6.64 \times 10^{-34} \text{ J s})$
$= 5.3 \times 10^{45}$
Note that the quantum number for the satellite motion is extremely large! In fact for such large quantum numbers the results of quantisation conditions tend to those of classical physics.

EXAMPLE 12.5

12.4.1 Energy levels

The energy of an atom is the *least* (largest negative value) when its electron is revolving in an orbit closest to the nucleus i.e., the one for which n = 1. For n = 2, 3, ... the absolute value of the energy E is smaller, hence the energy is progressively larger in the outer orbits. The *lowest* state of the atom, called the *ground state*, is that of the lowest energy, with the electron revolving in the orbit of smallest radius, the Bohr radius, a_0. The energy of this state (n = 1), E_1 is –13.6 eV. Therefore, the minimum energy required to free the electron from the ground state of the hydrogen atom is 13.6 eV. It is called the *ionisation energy* of the hydrogen atom. This prediction of the Bohr's model is in excellent agreement with the experimental value of ionisation energy.

At room temperature, most of the hydrogen atoms are in *ground state*. When a hydrogen atom receives energy by processes such as electron collisions, the atom may acquire sufficient energy to raise the electron to higher energy states. The atom is then said to be in an *excited state*. From Eq. (12.19), for n = 2; the energy E_2 is –3.40 eV. It means that the energy required to excite an electron in hydrogen atom to its first excited state, is an energy equal to $E_2 - E_1$ = –3.40 eV – (–13.6) eV = 10.2 eV. Similarly, E_3 = –1.51 eV and $E_3 - E_1$ = 12.09 eV, or to excite the hydrogen atom from its ground state (n = 1) to second excited state (n = 3), 12.09 eV energy is required, and so on. From these excited states the electron can then fall back to a state of lower energy, emitting a photon in the process. Thus, as the excitation of hydrogen atom increases (that is as n increases) the value of minimum energy required to free the electron from the excited atom decreases.

The energy level diagram* for the stationary states of a hydrogen atom, computed from Eq. (12.19), is given in

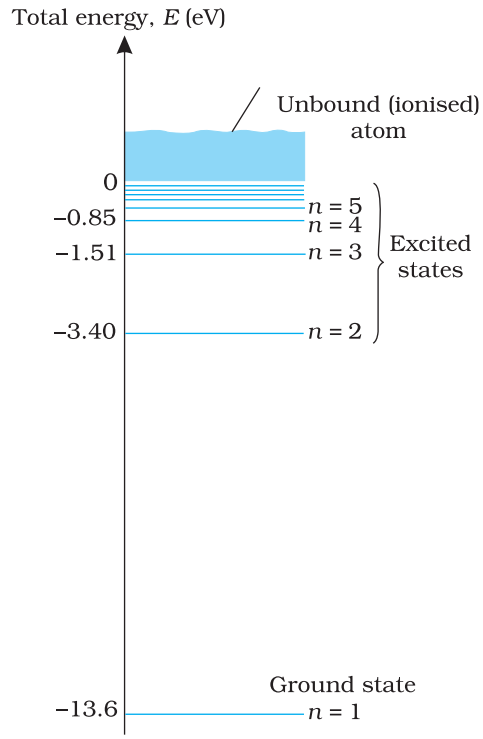


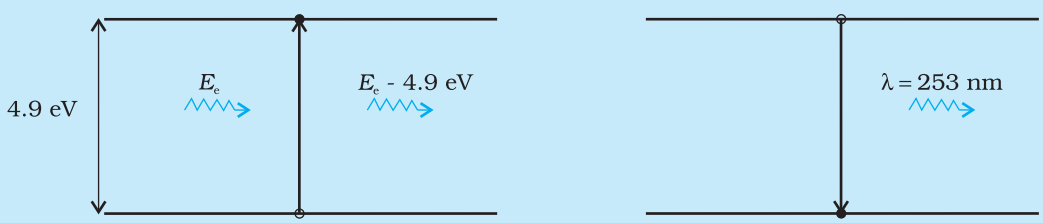
FIGURE 12.8 The energy level diagram for the hydrogen atom. The electron in a hydrogen atom at room temperature spends most of its time in the ground state. To ionise a hydrogen atom an electron from the ground state, 13.6 eV of energy must be supplied. (The horizontal lines specify the presence of allowed energy states.)

* An electron can have any total energy above E = 0 eV. In such situations the electron is free. Thus there is a continuum of energy states above E = 0 eV, as shown in Fig. 12.8.

Fig. 12.8. The principal quantum number n labels the stationary states in the ascending order of energy. In this diagram, the highest energy state corresponds to $n = \infty$ in Eq, (12.19) and has an energy of 0 eV. This is the energy of the atom when the electron is completely removed ($r = \infty$) from the nucleus and is at rest. Observe how the energies of the excited states come closer and closer together as n increases.

FRANCK – HERTZ EXPERIMENT

The existence of discrete energy levels in an atom was directly verified in 1914 by James Franck and Gustav Hertz. They studied the spectrum of mercury vapour when electrons having different kinetic energies passed through the vapour. The electron energy was varied by subjecting the electrons to electric fields of varying strength. The electrons collide with the mercury atoms and can transfer energy to the mercury atoms. This can only happen when the energy of the electron is higher than the energy difference between an energy level of Hg occupied by an electron and a higher unoccupied level (see Figure). For instance, the difference between an occupied energy level of Hg and a higher unoccupied level is 4.9 eV. If an electron of having an energy of 4.9 eV or more passes through mercury, an electron in mercury atom can absorb energy from the bombarding electron and get excited to the higher level [Fig (a)]. The colliding electron's kinetic energy would reduce by this amount.



(a) (b)

The excited electron would subsequently fall back to the ground state by emission of radiation [Fig. (b)]. The wavelength of emitted radiation is:

$$\lambda = \frac{hc}{E} = \frac{6.625\times10^{-34}\times3\times10^{8}}{4.9\times1.6\times10^{-19}} = 253 \text{ nm}$$

By direct measurement, Franck and Hertz found that the emission spectrum of mercury has a line corresponding to this wavelength. For this experimental verification of Bohr's basic ideas of discrete energy levels in atoms and the process of photon emission, Frank and Hertz were awarded the Nobel prize in 1925.

12.5 THE LINE SPECTRA OF THE HYDROGEN ATOM

According to the third postulate of Bohr's model, when an atom makes a transition from the higher energy state with quantum number n_i to the lower energy state with quantum number n_f ($n_f < n_i$), the difference of energy is carried away by a photon of frequency ν_{if} such that

$$h\nu_{if} = E_{n_i} - E_{n_f} \tag{12.20}$$

Using Eq. (12.16), for E_{n_f} and E_{n_i}, we get

$$h\nu_{if} = \frac{me^4}{8\varepsilon_0^2 h^2}\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right) \tag{12.21}$$

$$\text{or } \nu_{if} = \frac{me^4}{8\varepsilon_0^2 h^2}\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right) \tag{12.22}$$

Equation (12.21) is the Rydberg formula, for the spectrum of the hydrogen atom. In this relation, if we take $n_f = 2$ and $n_i = 3, 4, 5...$, it reduces to a form similar to Eq. (12.10) for the Balmer series. The Rydberg constant R is readily identified to be

$$R = \frac{me^4}{8\varepsilon_0^2 h^3 c} \tag{12.23}$$

If we insert the values of various constants in Eq. (12.23), we get

$R = 1.03 \times 10^7\ \text{m}^{-1}$

This is a value very close to the value ($1.097 \times 10^7\ \text{m}^{-1}$) obtained from the empirical Balmer formula. This agreement between the theoretical and experimental values of the Rydberg constant provided a direct and striking confirmation of the Bohr's model.

Since both n_f and n_i are integers, this immediately shows that in transitions between different atomic levels, light is radiated in various discrete frequencies. For hydrogen spectrum, the Balmer formula corresponds to $n_f = 2$ and $n_i = 3, 4, 5$, etc. The results of the Bohr's model suggested the presence of other series spectra for hydrogen atom–those corresponding to transitions resulting from $n_f = 1$ and $n_i = 2, 3$, etc.; $n_f = 3$ and $n_i = 4, 5$, etc., and so on. Such series were identified in the course of spectroscopic investigations and are known as the Lyman, Balmer, Paschen, Brackett, and Pfund series. The electronic transitions corresponding to these series are shown in Fig. 12.9.

The various lines in the atomic spectra are produced when electrons jump from higher energy state to a lower energy state and photons are emitted. These spectral lines are called emission lines. But when an atom absorbs a photon that has precisely

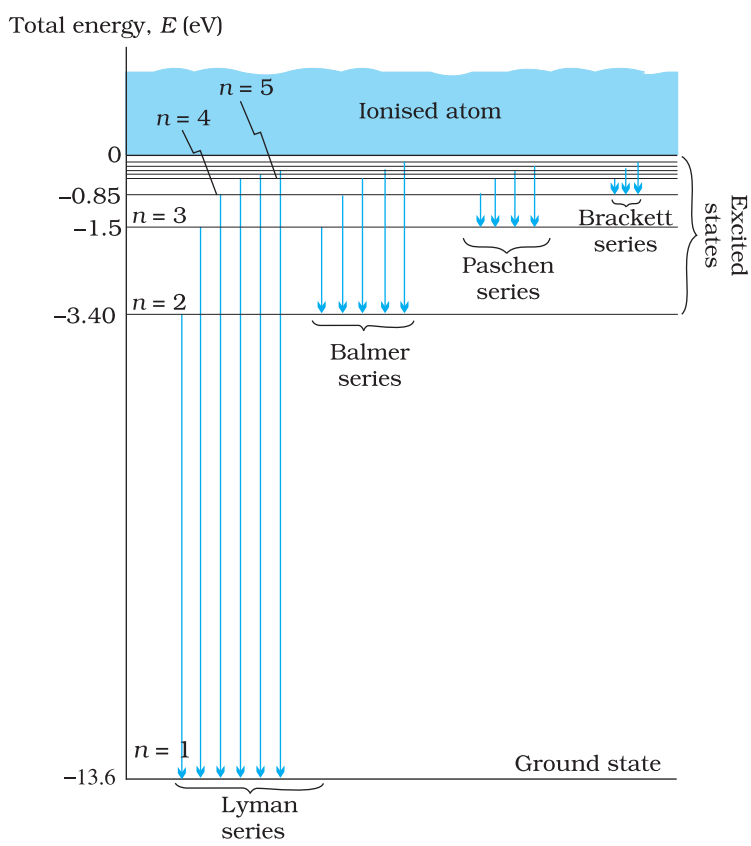


FIGURE 12.9 Line spectra originate in transitions between energy levels.

the same energy needed by the electron in a lower energy state to make transitions to a higher energy state, the process is called absorption. Thus if photons with a continuous range of frequencies pass through a rarefied gas and then are analysed with a spectrometer, a series of dark spectral absorption lines appear in the continuous spectrum. The dark lines indicate the frequencies that have been absorbed by the atoms of the gas.

The explanation of the hydrogen atom spectrum provided by Bohr's model was a brilliant achievement, which greatly stimulated progress towards the modern quantum theory. In 1922, Bohr was awarded Nobel Prize in Physics.

EXAMPLE 12.6

Example 12.6 Using the Rydberg formula, calculate the wavelengths of the first four spectral lines in the Lyman series of the hydrogen spectrum.

Solution The Rydberg formula is

$$hc/\lambda_{if} = \frac{me^4}{8\varepsilon_0^2 h^2}\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right)$$

The wavelengths of the first four lines in the Lyman series correspond to transitions from n_i = 2,3,4,5 to n_f = 1. We know that

$$\frac{me^4}{8\varepsilon_0^2 h^2} = 13.6 \text{ eV} = 21.76 \times 10^{-19} \text{ J}$$

Therefore,

$$\lambda_{i1} = \frac{hc}{21.76 \times 10^{-19}\left(\frac{1}{1} - \frac{1}{n_i^2}\right)} \text{ m}$$

$$= \frac{6.625\times10^{-34} \times 3\times10^{8} \times n_i^2}{21.76\times10^{-19} \times (n_i^2 - 1)} \text{ m} = \frac{0.9134\ n_i^2}{(n_i^2 - 1)} \times 10^{-7} \text{ m}$$

$$= 913.4\ n_i^2/(n_i^2 - 1) \text{ Å}$$

Substituting n_i = 2,3,4,5, we get λ_{21} = 1218 Å, λ_{31} = 1028 Å, λ_{41} = 974.3 Å, and λ_{51} = 951.4 Å.

12.6 De Broglie's Explanation of Bohr's Second Postulate of Quantisation

Of all the postulates, Bohr made in his model of the atom, perhaps the most puzzling is his second postulate. It states that the angular momentum of the electron orbiting around the nucleus is quantised (that is, $L_n = nh/2\pi$; n = 1, 2, 3 ...). Why should the angular momentum have only those values that are integral multiples of $h/2\pi$? The French physicist Louis de Broglie explained this puzzle in 1923, ten years after Bohr proposed his model.

We studied, in Chapter 11, about the de Broglie's hypothesis that material particles, such as electrons, also have a wave nature. C. J. Davisson and L. H. Germer later experimentally verified the wave nature of electrons

in 1927. Louis de Broglie argued that the electron in its circular orbit, as proposed by Bohr, must be seen as a particle wave. In analogy to waves travelling on a string, particle waves too can lead to standing waves under resonant conditions. From Chapter 15 of Class XI Physics textbook, we know that when a string is plucked, a vast number of wavelengths are excited. However only those wavelengths survive which have nodes at the ends and form the standing wave in the string. It means that in a string, standing waves are formed when the total distance travelled by a wave down the string and back is one wavelength, two wavelengths, or any integral number of wavelengths. Waves with other wavelengths interfere with themselves upon reflection and their amplitudes quickly drop to zero. For an electron moving in n^{th} circular orbit of radius r_n, the total distance is the circumference of the orbit, $2\pi r_n$. Thus

$$2\pi\, r_n = n\lambda, \qquad n = 1, 2, 3... \tag{12.24}$$

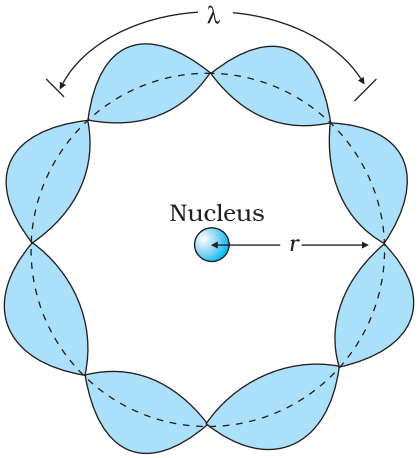


FIGURE 12.10 A standing wave is shown on a circular orbit where four de Broglie wavelengths fit into the circumference of the orbit.

Figure 12.10 illustrates a standing particle wave on a circular orbit for $n = 4$, i.e., $2\pi r_n = 4\lambda$, where λ is the de Broglie wavelength of the electron moving in n^{th} orbit. From Chapter 11, we have $\lambda = h/p$, where p is the magnitude of the electron's momentum. If the speed of the electron is much less than the speed of light, the momentum is mv_n. Thus, $\lambda = h/mv_n$. From Eq. (12.24), we have

$$2\pi\, r_n = n\, h/mv_n \qquad \text{or} \qquad m\, v_n\, r_n = nh/2\pi$$

This is the quantum condition proposed by Bohr for the angular momentum of the electron [Eq. (12.13)]. In Section 12.5, we saw that this equation is the basis of explaining the discrete orbits and energy levels in hydrogen atom. Thus de Broglie hypothesis provided an explanation for Bohr's second postulate for the quantisation of angular momentum of the orbiting electron. The quantised electron orbits and energy states are due to the wave nature of the electron and only resonant standing waves can persist.

Bohr's model, involving classical trajectory picture (planet-like electron orbiting the nucleus), correctly predicts the gross features of the hydrogenic atoms*, in particular, the frequencies of the radiation emitted or selectively absorbed. This model however has many limitations. Some are:

(i) The Bohr model is applicable to hydrogenic atoms. It cannot be extended even to mere two electron atoms such as helium. The analysis of atoms with more than one electron was attempted on the lines of Bohr's model for hydrogenic atoms but did not meet with any success. Difficulty lies in the fact that each electron interacts not only with the positively charged nucleus but also with all other electrons.

* Hydrogenic atoms are the atoms consisting of a nucleus with positive charge +Ze and a single electron, where Z is the proton number. Examples are hydrogen atom, singly ionised helium, doubly ionised lithium, and so forth. In these atoms more complex electron-electron interactions are nonexistent.

The formulation of Bohr model involves electrical force between positively charged nucleus and electron. It does not include the electrical forces between electrons which necessarily appear in multi-electron atoms.

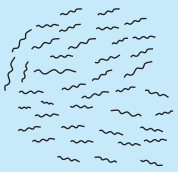
(ii) While the Bohr's model correctly predicts the frequencies of the light emitted by hydrogenic atoms, the model is unable to explain the relative intensities of the frequencies in the spectrum. In emission spectrum of hydrogen, some of the visible frequencies have weak intensity, others strong. Why? Experimental observations depict that some transitions are more favoured than others. Bohr's model is unable to account for the intensity variations.

Bohr's model presents an elegant picture of an atom and cannot be generalised to complex atoms. For complex atoms we have to use a new and radical theory based on Quantum Mechanics, which provides a more complete picture of the atomic structure.

LASER LIGHT

Imagine a crowded market place or a railway platform with people entering a gate and going towards all directions. Their footsteps are random and there is no phase correlation between them. On the other hand, think of a large number of soldiers in a regulated march. Their footsteps are very well correlated. See figure here.

This is similar to the difference between light emitted by an ordinary source like a candle or a bulb and that emitted by a laser. The acronym LASER stands for Light Amplification by Stimulated Emission of Radiation. Since its development in 1960, it has entered into all areas of science and technology. It has found applications in physics, chemistry, biology, medicine, surgery, engineering, etc. There are low power lasers, with a power of 0.5 mW, called pencil lasers, which serve as pointers. There are also lasers of different power, suitable for delicate surgery of eye or glands in the stomach. Finally, there are lasers which can cut or weld steel.



(a) Light from a bulb

(b) Laser light

Light is emitted from a source in the form of packets of waves. Light coming out from an ordinary source contains a mixture of many wavelengths. There is also no phase relation between the various waves. Therefore, such light, even if it is passed through an aperture, spreads very fast and the beam size increases rapidly with distance. In the case of laser light, the wavelength of each packet is almost the same. Also the average length of the packet of waves is much larger. This means that there is better phase correlation over a longer duration of time. This results in reducing the divergence of a laser beam substantially.

If there are N atoms in a source, each emitting light with intensity I, then the total intensity produced by an ordinary source is proportional to NI, whereas in a laser source, it is proportional to N^2I. Considering that N is very large, we see that the light from a laser can be much stronger than that from an ordinary source.

When astronauts of the Apollo missions visited the moon, they placed a mirror on its surface, facing the earth. Then scientists on the earth sent a strong laser beam, which was reflected by the mirror on the moon and received back on the earth. The size of the reflected laser beam and the time taken for the round trip were measured. This allowed a very accurate determination of (a) the extremely small divergence of a laser beam and (b) the distance of the moon from the earth.

SUMMARY

1. Atom, as a whole, is electrically neutral and therefore contains equal amount of positive and negative charges.
2. In *Thomson's model*, an atom is a spherical cloud of positive charges with electrons embedded in it.
3. In *Rutherford's model*, most of the mass of the atom and all its positive charge are concentrated in a tiny nucleus (typically one by ten thousand the size of an atom), and the electrons revolve around it.
4. Rutherford nuclear model has two main difficulties in explaining the structure of atom: (a) It predicts that atoms are unstable because the accelerated electrons revolving around the nucleus must spiral into the nucleus. This contradicts the stability of matter. (b) It cannot explain the characteristic line spectra of atoms of different elements.
5. Atoms of each element are stable and emit characteristic spectrum. The spectrum consists of a set of isolated parallel lines termed as line spectrum. It provides useful information about the atomic structure.
6. The atomic hydrogen emits a line spectrum consisting of various series. The frequency of any line in a series can be expressed as a difference of two terms;

   Lyman series: $\nu = Rc\left(\frac{1}{1^2} - \frac{1}{n^2}\right)$; $n = 2, 3, 4, \ldots$

   Balmer series: $\nu = Rc\left(\frac{1}{2^2} - \frac{1}{n^2}\right)$; $n = 3, 4, 5, \ldots$

   Paschen series: $\nu = Rc\left(\frac{1}{3^2} - \frac{1}{n^2}\right)$; $n = 4, 5, 6, \ldots$

   Brackett series: $\nu = Rc\left(\frac{1}{4^2} - \frac{1}{n^2}\right)$; $n = 5, 6, 7, \ldots$

   Pfund series: $\nu = Rc\left(\frac{1}{5^2} - \frac{1}{n^2}\right)$; $n = 6, 7, 8, \ldots$
7. To explain the line spectra emitted by atoms, as well as the stability of atoms, Niel's Bohr proposed a model for hydrogenic (single elctron) atoms. He introduced three postulates and laid the foundations of quantum mechanics:
   (a) In a hydrogen atom, an electron revolves in certain stable orbits (called stationary orbits) without the emission of radiant energy.
   (b) The stationary orbits are those for which the angular momentum is some integral multiple of $h/2\pi$. (Bohr's quantisation condition.) That is $L = nh/2\pi$, where n is an integer called a quantum number.
   (c) The third postulate states that an electron might make a transition from one of its specified non-radiating orbits to another of lower energy. When it does so, a photon is emitted having energy equal to the energy difference between the initial and final states. The frequency (ν) of the emitted photon is then given by

   $h\nu = E_i - E_f$

   An atom absorbs radiation of the same frequency the atom emits, in which case the electron is transferred to an orbit with a higher value of n.

   $E_i + h\nu = E_f$

8. As a result of the quantisation condition of angular momentum, the electron orbits the nucleus at only specific radii. For a hydrogen atom it is given by

$$r_n = \left(\frac{n^2}{m}\right)\left(\frac{h}{2\pi}\right)^2 \frac{4\pi\varepsilon_0}{e^2}$$

The total energy is also quantised:

$$E_n = -\frac{me^4}{8n^2\varepsilon_0^2 h^2}$$

$$= -13.6 \text{ eV}/n^2$$

The n = 1 state is called ground state. In hydrogen atom the ground state energy is –13.6 eV. Higher values of n correspond to excited states (n > 1). Atoms are excited to these higher states by collisions with other atoms or electrons or by absorption of a photon of right frequency.

9. de Broglie's hypothesis that electrons have a wavelength $\lambda = h/mv$ gave an explanation for Bohr's quantised orbits by bringing in the wave-particle duality. The orbits correspond to circular standing waves in which the circumference of the orbit equals a whole number of wavelengths.

10. Bohr's model is applicable only to hydrogenic (single electron) atoms. It cannot be extended to even two electron atoms such as helium. This model is also unable to explain for the relative intensities of the frequencies emitted even by hydrogenic atoms.

POINTS TO PONDER

1. Both the Thomson's as well as the Rutherford's models constitute an unstable system. Thomson's model is unstable electrostatically, while Rutherford's model is unstable because of electromagnetic radiation of orbiting electrons.
2. What made Bohr quantise angular momentum (second postulate) and not some other quantity? Note, h has dimensions of angular momentum, and for circular orbits, angular momentum is a very relevant quantity. The second postulate is then so natural!
3. The orbital picture in Bohr's model of the hydrogen atom was inconsistent with the uncertainty principle. It was replaced by modern quantum mechanics in which Bohr's orbits are regions where the electron may be found with large probability.
4. Unlike the situation in the solar system, where planet-planet gravitational forces are very small as compared to the gravitational force of the sun on each planet (because the mass of the sun is so much greater than the mass of any of the planets), the electron-electron electric force interaction is comparable in magnitude to the electron-nucleus electrical force, because the charges and distances are of the same order of magnitude. This is the reason why the Bohr's model with its planet-like electron is not applicable to many electron atoms.
5. Bohr laid the foundation of the quantum theory by postulating specific orbits in which electrons do not radiate. Bohr's model include only

one quantum number n. The new theory called quantum mechanics supportes Bohr's postulate. However in quantum mechanics (more generally accepted), a given energy level may not correspond to just one quantum state. For example, a state is characterised by four quantum numbers (n, l, m, and s), but for a pure Coulomb potential (as in hydrogen atom) the energy depends only on n.

6. In Bohr model, contrary to ordinary classical expectation, the frequency of revolution of an electron in its orbit is not connected to the frequency of spectral line. The later is the difference between two orbital energies divided by h. For transitions between large quantum numbers (n to $n-1$, n very large), however, the two coincide as expected.
7. Bohr's semiclassical model based on some aspects of classical physics and some aspects of modern physics also does not provide a true picture of the simplest hydrogenic atoms. The true picture is quantum mechanical affair which differs from Bohr model in a number of fundamental ways. But then if the Bohr model is not strictly correct, why do we bother about it? The reasons which make Bohr's model still useful are:
   (i) The model is based on just three postulates but accounts for almost all the general features of the hydrogen spectrum.
   (ii) The model incorporates many of the concepts we have learnt in classical physics.
   (iii) The model demonstrates how a theoretical physicist occasionally must quite literally ignore certain problems of approach in hopes of being able to make some predictions. If the predictions of the theory or model agree with experiment, a theoretician then must somehow hope to explain away or rationalise the problems that were ignored along the way.

EXERCISES

12.1 Choose the correct alternative from the clues given at the end of the each statement:

(a) The size of the atom in Thomson's model is ..... the atomic size in Rutherford's model. (much greater than/no different from/much less than.)

(b) In the ground state of ..... electrons are in stable equilibrium, while in ..... electrons always experience a net force. (Thomson's model/ Rutherford's model.)

(c) A *classical* atom based on ..... is doomed to collapse. (Thomson's model/ Rutherford's model.)

(d) An atom has a nearly continuous mass distribution in a ..... but has a highly non-uniform mass distribution in ..... (Thomson's model/ Rutherford's model.)

(e) The positively charged part of the atom possesses most of the mass in ..... (Rutherford's model/both the models.)

12.2 Suppose you are given a chance to repeat the alpha-particle scattering experiment using a thin sheet of solid hydrogen in place of the gold foil. (Hydrogen is a solid at temperatures below 14 K.) What results do you expect?

12.3 What is the shortest wavelength present in the Paschen series of spectral lines?

12.4 A difference of 2.3 eV separates two energy levels in an atom. What is the frequency of radiation emitted when the atom make a transition from the upper level to the lower level?

12.5 The ground state energy of hydrogen atom is –13.6 eV. What are the kinetic and potential energies of the electron in this state?

12.6 A hydrogen atom initially in the ground level absorbs a photon, which excites it to the n = 4 level. Determine the wavelength and frequency of photon.

12.7 (a) Using the Bohr's model calculate the speed of the electron in a hydrogen atom in the n = 1, 2, and 3 levels. (b) Calculate the orbital period in each of these levels.

12.8 The radius of the innermost electron orbit of a hydrogen atom is 5.3×10^{-11} m. What are the radii of the n = 2 and n =3 orbits?

12.9 A 12.5 eV electron beam is used to bombard gaseous hydrogen at room temperature. What series of wavelengths will be emitted?

12.10 In accordance with the Bohr's model, find the quantum number that characterises the earth's revolution around the sun in an orbit of radius 1.5×10^{11} m with orbital speed 3×10^{4} m/s. (Mass of earth = 6.0×10^{24} kg.)

ADDITIONAL EXERCISES

12.11 Answer the following questions, which help you understand the difference between Thomson's model and Rutherford's model better.

(a) Is the average angle of deflection of α-particles by a thin gold foil predicted by Thomson's model much less, about the same, or much greater than that predicted by Rutherford's model?

(b) Is the probability of backward scattering (i.e., scattering of α-particles at angles greater than 90°) predicted by Thomson's model much less, about the same, or much greater than that predicted by Rutherford's model?

(c) Keeping other factors fixed, it is found experimentally that for small thickness t, the number of α-particles scattered at moderate angles is proportional to t. What clue does this linear dependence on t provide?

(d) In which model is it completely wrong to ignore multiple scattering for the calculation of average angle of scattering of α-particles by a thin foil?

12.12 The gravitational attraction between electron and proton in a hydrogen atom is weaker than the coulomb attraction by a factor of about 10^{-40}. An alternative way of looking at this fact is to estimate the radius of the first Bohr orbit of a hydrogen atom if the electron and proton were bound by gravitational attraction. You will find the answer interesting.

12.13 Obtain an expression for the frequency of radiation emitted when a hydrogen atom de-excites from level n to level (n–1). For large n, show that this frequency equals the classical frequency of revolution of the electron in the orbit.

12.14 Classically, an electron can be in any orbit around the nucleus of an atom. Then what determines the typical atomic size? Why is an atom not, say, thousand times bigger than its typical size? The question had greatly puzzled Bohr before he arrived at his famous model of the atom that you have learnt in the text. To simulate what he might well have done before his discovery, let us play as follows with the basic constants of nature and see if we can get a quantity with the dimensions of length that is roughly equal to the known size of an atom (~ 10^{-10}m).

(a) Construct a quantity with the dimensions of length from the fundamental constants e, m_e, and c. Determine its numerical value.

(b) You will find that the length obtained in (a) is many orders of magnitude smaller than the atomic dimensions. Further, it involves c. But energies of atoms are mostly in non-relativistic domain where c is not expected to play any role. This is what may have suggested Bohr to discard c and look for 'something else' to get the right atomic size. Now, the Planck's constant h had already made its appearance elsewhere. Bohr's great insight lay in recognising that h, m_e, and e will yield the right atomic size. Construct a quantity with the dimension of length from h, me, and e and confirm that its numerical value has indeed the correct order of magnitude.

12.15 The total energy of an electron in the first excited state of the hydrogen atom is about –3.4 eV.

(a) What is the kinetic energy of the electron in this state?

(b) What is the potential energy of the electron in this state?

(c) Which of the answers above would change if the choice of the zero of potential energy is changed?

12.16 If Bohr's quantisation postulate (angular momentum = $nh/2\pi$) is a basic law of nature, it should be equally valid for the case of planetary motion also. Why then do we never speak of quantisation of orbits of planets around the sun?

12.17 Obtain the first Bohr's radius and the ground state energy of a *muonic hydrogen atom* [i.e., an atom in which a negatively charged muon (μ^-) of mass about $207m_e$ orbits around a proton].

Chapter Thirteen

NUCLEI

13.1 Introduction

In the previous chapter, we have learnt that in every atom, the positive charge and mass are densely concentrated at the centre of the atom forming its nucleus. The overall dimensions of a nucleus are much smaller than those of an atom. Experiments on scattering of α-particles demonstrated that the radius of a nucleus was smaller than the radius of an atom by a factor of about 10^4. This means the volume of a nucleus is about 10^{-12} times the volume of the atom. In other words, an atom is almost empty. If an atom is enlarged to the size of a classroom, the nucleus would be of the size of pinhead. Nevertheless, the nucleus contains most (more than 99.9%) of the mass of an atom.

Does the nucleus have a structure, just as the atom does? If so, what are the constituents of the nucleus? How are these held together? In this chapter, we shall look for answers to such questions. We shall discuss various properties of nuclei such as their size, mass and stability, and also associated nuclear phenomena such as radioactivity, fission and fusion.

13.2 Atomic Masses and Composition of Nucleus

The mass of an atom is very small, compared to a kilogram; for example, the mass of a carbon atom, ^{12}C, is 1.992647×10^{-26} kg. Kilogram is not a very convenient unit to measure such small quantities. Therefore, a

different mass unit is used for expressing atomic masses. This unit is the atomic mass unit (u), defined as $1/12^{th}$ of the mass of the carbon (^{12}C) atom. According to this definition

$$1u = \frac{\text{mass of one } ^{12}C \text{ atom}}{12}$$

$$= \frac{1.992647 \times 10^{-26} \text{ kg}}{12}$$

$$= 1.660539 \times 10^{-27} \text{ kg} \qquad (13.1)$$

The atomic masses of various elements expressed in atomic mass unit (u) are close to being integral multiples of the mass of a hydrogen atom. There are, however, many striking exceptions to this rule. For example, the atomic mass of chlorine atom is 35.46 u.

Accurate measurement of atomic masses is carried out with a mass spectrometer, The measurement of atomic masses reveals the existence of different types of atoms of the same element, which exhibit the same chemical properties, but differ in mass. Such atomic species of the same element differing in mass are called *isotopes*. (In Greek, isotope means the same place, i.e. they occur in the same place in the periodic table of elements.) It was found that practically every element consists of a mixture of several isotopes. The relative abundance of different isotopes differs from element to element. Chlorine, for example, has two isotopes having masses 34.98 u and 36.98 u, which are nearly integral multiples of the mass of a hydrogen atom. The relative abundances of these isotopes are 75.4 and 24.6 per cent, respectively. Thus, the average mass of a chlorine atom is obtained by the weighted average of the masses of the two isotopes, which works out to be

$$= \frac{75.4 \times 34.98 + 24.6 \times 36.98}{100}$$

$$= 35.47 \text{ u}$$

which agrees with the atomic mass of chlorine.

Even the lightest element, hydrogen has three isotopes having masses 1.0078 u, 2.0141 u, and 3.0160 u. The nucleus of the lightest atom of hydrogen, which has a relative abundance of 99.985%, is called the proton. The mass of a proton is

$$m_p = 1.00727 \text{ u} = 1.67262 \times 10^{-27} \text{ kg} \qquad (13.2)$$

This is equal to the mass of the hydrogen atom (= 1.00783u), minus the mass of a single electron (m_e = 0.00055 u). The other two isotopes of hydrogen are called deuterium and tritium. Tritium nuclei, being unstable, do not occur naturally and are produced artificially in laboratories.

The positive charge in the nucleus is that of the protons. A proton carries one unit of fundamental charge and is stable. It was earlier thought that the nucleus may contain electrons, but this was ruled out later using arguments based on quantum theory. All the electrons of an atom are outside the nucleus. We know that the number of these electrons outside the nucleus of the atom is Z, the atomic number. The total charge of the

atomic electrons is thus ($-Ze$), and since the atom is neutral, the charge of the nucleus is ($+Ze$). The number of protons in the nucleus of the atom is, therefore, exactly Z, the atomic number.

Discovery of Neutron

Since the nuclei of deuterium and tritium are isotopes of hydrogen, they must contain only one proton each. But the masses of the nuclei of hydrogen, deuterium and tritium are in the ratio of 1:2:3. Therefore, the nuclei of deuterium and tritium must contain, in addition to a proton, some neutral matter. The amount of neutral matter present in the nuclei of these isotopes, expressed in units of mass of a proton, is approximately equal to one and two, respectively. This fact indicates that the nuclei of atoms contain, in addition to protons, neutral matter in multiples of a basic unit. This hypothesis was verified in 1932 by James Chadwick who observed emission of neutral radiation when beryllium nuclei were bombarded with alpha-particles (α-particles are helium nuclei, to be discussed in a later section). It was found that this neutral radiation could knock out protons from light nuclei such as those of helium, carbon and nitrogen. The only neutral radiation known at that time was photons (electromagnetic radiation). Application of the principles of conservation of energy and momentum showed that if the neutral radiation consisted of photons, the energy of photons would have to be much higher than is available from the bombardment of beryllium nuclei with α-particles. The clue to this puzzle, which Chadwick satisfactorily solved, was to assume that the neutral radiation consists of a new type of neutral particles called *neutrons*. From conservation of energy and momentum, he was able to determine the mass of new particle 'as very nearly the same as mass of proton'.

The mass of a neutron is now known to a high degree of accuracy. It is

$$m_n = 1.00866 \text{ u} = 1.6749\times10^{-27} \text{ kg} \qquad (13.3)$$

Chadwick was awarded the 1935 Nobel Prize in Physics for his discovery of the neutron.

A free neutron, unlike a free proton, is unstable. It decays into a proton, an electron and a antineutrino (another elementary particle), and has a mean life of about 1000s. It is, however, stable inside the nucleus.

The composition of a nucleus can now be described using the following terms and symbols:

Z - *atomic number* = number of protons [13.4(a)]

N - *neutron number* = number of neutrons [13.4(b)]

A - *mass number* = $Z + N$

= total number of protons and neutrons [13.4(c)]

One also uses the term nucleon for a proton or a neutron. Thus the number of nucleons in an atom is its mass number A.

Nuclear species or nuclides are shown by the notation $^{A}_{Z}X$ where X is the chemical symbol of the species. For example, the nucleus of gold is denoted by $^{197}_{79}Au$. It contains 197 nucleons, of which 79 are protons and the rest118 are neutrons.

The composition of isotopes of an element can now be readily explained. The nuclei of isotopes of a given element contain the same number of protons, but differ from each other in their number of neutrons. Deuterium, ${}^{2}_{1}\text{H}$, which is an isotope of hydrogen, contains one proton and one neutron. Its other isotope tritium, ${}^{3}_{1}\text{H}$, contains one proton and two neutrons. The element gold has 32 isotopes, ranging from A =173 to A = 204. We have already mentioned that chemical properties of elements depend on their electronic structure. As the atoms of isotopes have identical electronic structure they have identical chemical behaviour and are placed in the same location in the periodic table.

All nuclides with same mass number A are called *isobars*. For example, the nuclides ${}^{3}_{1}\text{H}$ and ${}^{3}_{2}\text{He}$ are isobars. Nuclides with same neutron number N but different atomic number Z, for example ${}^{198}_{80}\text{Hg}$ and ${}^{197}_{79}\text{Au}$, are called *isotones*.

13.3 Size of the Nucleus

As we have seen in Chapter 12, Rutherford was the pioneer who postulated and established the existence of the atomic nucleus. At Rutherford's suggestion, Geiger and Marsden performed their classic experiment: on the scattering of α-particles from thin gold foils. Their experiments revealed that the distance of closest approach to a gold nucleus of an α-particle of kinetic energy 5.5 MeV is about 4.0×10^{-14} m. The scattering of α-particle by the gold sheet could be understood by Rutherford by assuming that the coulomb repulsive force was solely responsible for scattering. Since the positive charge is confined to the nucleus, the actual size of the nucleus has to be less than 4.0×10^{-14} m.

If we use α-particles of higher energies than 5.5 MeV, the distance of closest approach to the gold nucleus will be smaller and at some point the scattering will begin to be affected by the short range nuclear forces, and differ from Rutherford's calculations. Rutherford's calculations are based on pure coulomb repulsion between the positive charges of the α-particle and the gold nucleus. From the distance at which deviations set in, nuclear sizes can be inferred.

By performing scattering experiments in which fast electrons, instead of α-particles, are projectiles that bombard targets made up of various elements, the sizes of nuclei of various elements have been accurately measured.

It has been found that a nucleus of mass number A has a radius

$$R = R_0 A^{1/3} \tag{13.5}$$

where $R_0 = 1.2 \times 10^{-15}$ m (=1.2 fm; 1 fm = 10^{-15} m). This means the volume of the nucleus, which is proportional to R^3 is proportional to A. Thus the density of nucleus is a constant, independent of A, for all nuclei. Different nuclei are like a drop of liquid of constant density. The density of nuclear matter is approximately 2.3×10^{17} kg m^{-3}. This density is very large compared to ordinary matter, say water, which is 10^3 kg m^{-3}. This is understandable, as we have already seen that most of the atom is empty. Ordinary matter consisting of atoms has a large amount of empty space.

EXAMPLE 13.1

Example 13.1 Given the mass of iron nucleus as 55.85u and A=56, find the nuclear density?

Solution

m_{Fe} = 55.85, u = 9.27×10^{-26} kg

$$\text{Nuclear density} = \frac{\text{mass}}{\text{volume}} = \frac{9.27\times10^{-26}}{(4\pi/3)(1.2\times10^{-15})^3}\times\frac{1}{56}$$

$$= 2.29 \times 10^{17}\ \text{kg m}^{-3}$$

The density of matter in neutron stars (an astrophysical object) is comparable to this density. This shows that matter in these objects has been compressed to such an extent that they resemble a *big nucleus*.

13.4 Mass-Energy and Nuclear Binding Energy

13.4.1 Mass – Energy

Einstein showed from his theory of special relativity that it is necessary to treat mass as another form of energy. Before the advent of this theory of special relativity it was presumed that mass and energy were conserved separately in a reaction. However, Einstein showed that mass is another form of energy and one can convert mass-energy into other forms of energy, say kinetic energy and vice-versa.

Einstein gave the famous mass-energy equivalence relation

$$E = mc^2 \tag{13.6}$$

Here the energy equivalent of mass m is related by the above equation and c is the velocity of light in vacuum and is approximately equal to 3×10^8 m s^{-1}.

EXAMPLE 13.2

Example 13.2 Calculate the energy equivalent of 1 g of substance.

Solution

Energy, $E = 10^{-3} \times (3 \times 10^8)^2$ J

$$E = 10^{-3} \times 9 \times 10^{16} = 9 \times 10^{13}\ \text{J}$$

Thus, if one gram of matter is converted to energy, there is a release of enormous amount of energy.

Experimental verification of the Einstein's mass-energy relation has been achieved in the study of nuclear reactions amongst nucleons, nuclei, electrons and other more recently discovered particles. In a reaction the conservation law of energy states that the initial energy and the final energy are equal provided the energy associated with mass is also included. This concept is important in understanding nuclear masses and the interaction of nuclei with one another. They form the subject matter of the next few sections.

13.4.2 Nuclear binding energy

In Section 13.2 we have seen that the nucleus is made up of neutrons and protons. Therefore it may be expected that the mass of the nucleus is equal to the total mass of its individual protons and neutrons. However,

the nuclear mass M is found to be always less than this. For example, let us consider $^{16}_{8}O$; a nucleus which has 8 neutrons and 8 protons. We have

Mass of 8 neutrons = 8 × 1.00866 u

Mass of 8 protons = 8 × 1.00727 u

Mass of 8 electrons = 8 × 0.00055 u

Therefore the expected mass of $^{16}_{8}O$ nucleus
= 8 × 2.01593 u = 16.12744 u.

The atomic mass of $^{16}_{8}O$ found from mass spectroscopy experiments is seen to be 15.99493 u. Substracting the mass of 8 electrons (8 × 0.00055 u) from this, we get the experimental mass of $^{16}_{8}O$ nucleus to be 15.99053 u.

Thus, we find that the mass of the $^{16}_{8}O$ nucleus is less than the total mass of its constituents by 0.13691u. The difference in mass of a nucleus and its constituents, ΔM, is called the *mass defect*, and is given by

$$\Delta M = [Zm_p + (A - Z)m_n] - M \qquad (13.7)$$

What is the meaning of the mass defect? It is here that Einstein's equivalence of mass and energy plays a role. Since the mass of the oxygen nucleus is less that the sum of the masses of its constituents (8 protons and 8 neutrons, in the unbound state), the equivalent energy of the oxygen nucleus is less than that of the sum of the equivalent energies of its constituents. If one wants to break the oxygen nucleus into 8 protons and 8 neutrons, this extra energy $\Delta M\, c^2$, has to supplied. This energy required E_b is related to the mass defect by

$$E_b = \Delta M\, c^2 \qquad (13.8)$$

Example 13.3 Find the energy equivalent of one atomic mass unit, first in Joules and then in MeV. Using this, express the mass defect of $^{16}_{8}O$ in MeV/c^2.

Solution

1u = 1.6605×10^{-27} kg

To convert it into energy units, we multiply it by c^2 and find that energy equivalent = $1.6605 \times 10^{-27} \times (2.9979 \times 10^8)^2$ kg m^2/s^2

$= 1.4924 \times 10^{-10}$ J

$= \dfrac{1.4924 \times 10^{-10}}{1.602 \times 10^{-19}}$ eV

$= 0.9315 \times 10^9$ eV

= 931.5 MeV

or, 1u = 931.5 MeV/c^2

For $^{16}_{8}O$, ΔM = 0.13691 u = 0.13691×931.5 MeV/c^2

= 127.5 MeV/c^2

The energy needed to separate $^{16}_{8}O$ into its constituents is thus 127.5 MeV/c^2.

EXAMPLE 13.3

If a certain number of neutrons and protons are brought together to form a nucleus of a certain charge and mass, an energy E_b will be released

in the process. The energy E_b is called the *binding energy* of the nucleus. If we separate a nucleus into its nucleons, we would have to supply a total energy equal to E_b, to those particles. Although we cannot tear apart a nucleus in this way, the nuclear binding energy is still a convenient measure of how well a nucleus is held together. A more useful measure of the binding between the constituents of the nucleus is the *binding energy per nucleon*, E_{bn}, which is the ratio of the binding energy E_b of a nucleus to the number of the nucleons, A, in that nucleus:

$$E_{bn} = E_b / A \tag{13.9}$$

We can think of binding energy per nucleon as the average energy per nucleon needed to separate a nucleus into its individual nucleons.

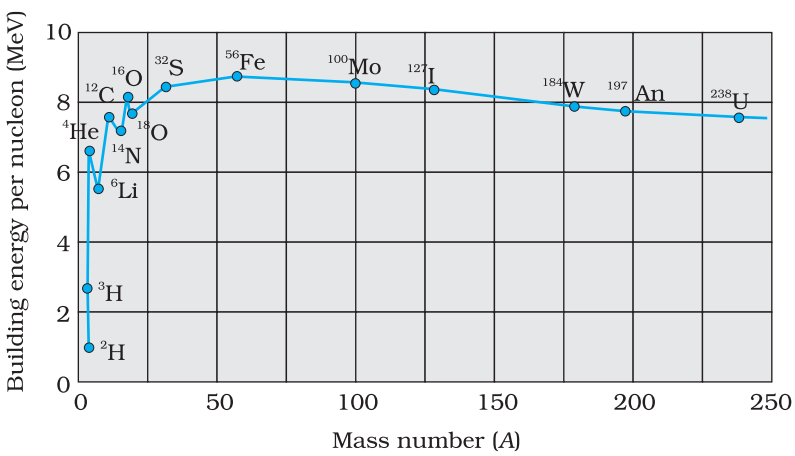


FIGURE 13.1 The binding energy per nucleon as a function of mass number.

Figure 13.1 is a plot of the binding energy per nucleon E_{bn} versus the mass number A for a large number of nuclei. We notice the following main features of the plot:

(i) the binding energy per nucleon, E_{bn}, is practically constant, i.e. practically independent of the atomic number for nuclei of middle mass number (30 < A < 170). The curve has a maximum of about 8.75 MeV for A = 56 and has a value of 7.6 MeV for A = 238.

(ii) E_{bn} is lower for both light nuclei (A<30) and heavy nuclei (A>170).

We can draw some conclusions from these two observations:

(i) The force is attractive and sufficiently strong to produce a binding energy of a few MeV per nucleon.

(ii) The constancy of the binding energy in the range 30 < A < 170 is a consequence of the fact that the nuclear force is short-ranged. Consider a particular nucleon inside a sufficiently large nucleus. It will be under the influence of only some of its neighbours, which come within the range of the nuclear force. If any other nucleon is at a distance more than the range of the nuclear force from the particular nucleon it will have no influence on the binding energy of the nucleon under consideration. If a nucleon can have a maximum of p neighbours within the range of nuclear force, its binding energy would be proportional to p. Let the binding energy of the nucleus be pk, where k is a constant having the dimensions of energy. If we increase A by adding nucleons they will not change the binding energy of a nucleon inside. Since most of the nucleons in a large nucleus reside inside it and not on the surface, the change in binding energy per nucleon would be small. The binding energy per nucleon is a constant and is approximately equal to pk. The property that a given nucleon

influences only nucleons close to it is also referred to as saturation property of the nuclear force.

(iii) A very heavy nucleus, say A = 240, has lower binding energy per nucleon compared to that of a nucleus with A = 120. Thus if a nucleus A = 240 breaks into two A = 120 nuclei, nucleons get more tightly bound. This implies energy would be released in the process. It has very important implications for energy production through *fission*, to be discussed later in Section 13.7.1.

(iv) Consider two very light nuclei ($A \leq 10$) joining to form a heavier nucleus. The binding energy per nucleon of the fused heavier nuclei is more than the binding energy per nucleon of the lighter nuclei. This means that the final system is more tightly bound than the initial system. Again energy would be released in such a process of *fusion*. This is the energy source of sun, to be discussed later in Section 13.7.3.

13.5 Nuclear Force

The force that determines the motion of atomic electrons is the familiar Coulomb force. In Section 13.4, we have seen that for average mass nuclei the binding energy per nucleon is approximately 8 MeV, which is much larger than the binding energy in atoms. Therefore, to bind a nucleus together there must be a strong attractive force of a totally different kind. It must be strong enough to overcome the repulsion between the (positively charged) protons and to bind both protons and neutrons into the tiny nuclear volume. We have already seen that the constancy of binding energy per nucleon can be understood in terms of its short-range. Many features of the nuclear binding force are summarised below. These are obtained from a variety of experiments carried out during 1930 to 1950.

(i) The nuclear force is much stronger than the Coulomb force acting between charges or the gravitational forces between masses. The nuclear binding force has to dominate over the Coulomb repulsive force between protons inside the nucleus. This happens only because the nuclear force is much stronger than the coulomb force. The gravitational force is much weaker than even Coulomb force.

(ii) The nuclear force between two nucleons falls rapidly to zero as their distance is more than a few femtometres. This leads to *saturation of forces* in a medium or a large-sized nucleus, which is the reason for the constancy of the binding energy per nucleon.

A rough plot of the potential energy between two nucleons as a function of distance is shown in the Fig. 13.2. The potential energy is a minimum at a distance r_0 of about 0.8 fm. This means that the force is attractive for distances larger than 0.8 fm and repulsive if they are separated by distances less than 0.8 fm.

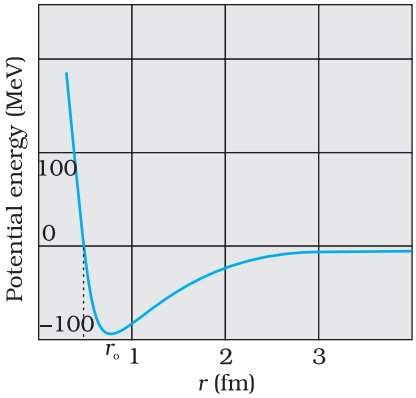


FIGURE 13.2 Potential energy of a pair of nucleons as a function of their separation. For a separation greater than r_0, the force is attractive and for separations less than r_0, the force is strongly repulsive.

(iii) The nuclear force between neutron-neutron, proton-neutron and proton-proton is approximately the same. The nuclear force does not depend on the electric charge.

Unlike Coulomb's law or the Newton's law of gravitation there is no simple mathematical form of the nuclear force.

13.6 Radioactivity

A. H. Becquerel discovered radioactivity in 1896 purely by accident. While studying the fluorescence and phosphorescence of compounds irradiated with visible light, Becquerel observed an interesting phenomenon. After illuminating some pieces of uranium-potassium sulphate with visible light, he wrapped them in black paper and separated the package from a photographic plate by a piece of silver. When, after several hours of exposure, the photographic plate was developed, it showed blackening due to something that must have been emitted by the compound and was able to penetrate both black paper and the silver.

Experiments performed subsequently showed that radioactivity was a nuclear phenomenon in which an unstable nucleus undergoes a decay. This is referred to as *radioactive decay*. Three types of radioactive decay occur in nature :

(i) α-decay in which a helium nucleus ${}^{4}_{2}He$ is emitted;

(ii) β-decay in which electrons or positrons (particles with the same mass as electrons, but with a charge exactly opposite to that of electron) are emitted;

(iii) γ-decay in which high energy (hundreds of keV or more) photons are emitted.

Each of these decay will be considered in subsequent sub-sections.

13.6.1 Law of radioactive decay

In any radioactive sample, which undergoes α, β or γ-decay, it is found that the number of nuclei undergoing the decay per unit time is proportional to the total number of nuclei in the sample. If N is the number of nuclei in the sample and ΔN undergo decay in time Δt then

$$\frac{\Delta N}{\Delta t} \propto N$$

or, $\Delta N/\Delta t = \lambda N$, (13.10)

where λ is called the radioactive *decay constant* or *disintegration constant*.

The change in the number of nuclei in the sample* is $dN = -\Delta N$ in time Δt. Thus the rate of change of N is (in the limit $\Delta t \to 0$)

$$\frac{dN}{dt} = -\lambda N$$

* ΔN is the number of nuclei that decay, and hence is always positive. dN is the change in N, which may have either sign. Here it is negative, because out of original N nuclei, ΔN have decayed, leaving $(N-\Delta N)$ nuclei.

or, $\frac{dN}{N} = -\lambda dt$

Now, integrating both sides of the above equation,we get,

$$\int_{N_0}^{N} \frac{dN}{N} = -\lambda \int_{t_0}^{t} dt \qquad (13.11)$$

or, $\ln N - \ln N_0 = -\lambda (t - t_0)$ (13.12)

Here N_0 is the number of radioactive nuclei in the sample at some arbitrary time t_0 and N is the number of radioactive nuclei at any subsequent time t. Setting $t_0 = 0$ and rearranging Eq. (13.12) gives us

$$\ln \frac{N}{N_0} = -\lambda t \qquad (13.13)$$

which gives

$$N(t) = N_0 \, e^{-\lambda t} \qquad (13.14)$$

Note, for example, the light bulbs follow no such exponential decay law. If we test 1000 bulbs for their life (time span before they burn out or fuse), we expect that they will 'decay' (that is, burn out) at more or less the same time. The decay of radionuclides follows quite a different law, the *law of radioactive decay* represented by Eq. (13.14).

The total decay rate R of a sample is the number of nuclei disintegrating per unit time. Suppose in a time interval dt, the decay count measured is ΔN. Then $dN = -\Delta N$.

The positive quantity R is then defined as

$$R = -\frac{dN}{dt}$$

Differentiating Eq. (13.14), we get

$$R = \lambda N_0 \, e^{-\lambda t}$$

or, $R = R_0 \, e^{-\lambda t}$ (13.15)

This is equivalant to the law of radioactivity decay, since you can integrate Eq. (13.15) to get back Eq. (13.14). Clearly, $R_0 = \lambda N_0$ is the decay rate at $t = 0$. The decay rate R at a certain time t and the number of undecayed nuclei N at the same time are related by

$$R = \lambda N \qquad (13.16)$$

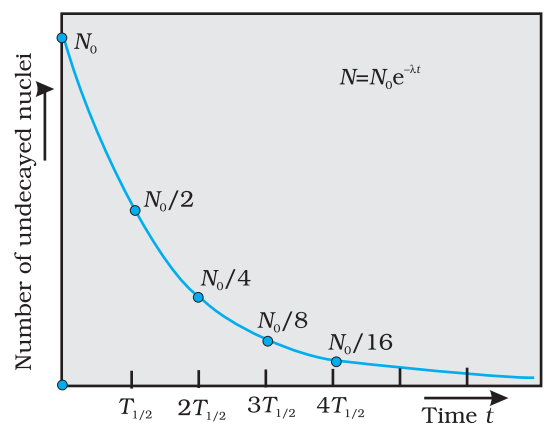


FIGURE 13.3 Exponential decay of a radioactive species. After a lapse of $T_{1/2}$, population of the given species drops by a factor of 2.

The decay rate of a sample, rather than the number of radioactive nuclei, is a more direct experimentally measurable quantity and is given a specific name: *activity*. The SI unit for activity is becquerel, named after the discoverer of radioactivity, Henry Becquerel.

1 becquerel is simply equal to 1 disintegration or decay per second. There is also another unit named "curie" that is widely used and is related to the SI unit as:

$$1 \text{ curie} = 1 \text{ Ci} = 3.7 \times 10^{10} \text{ decays per second}$$
$$= 3.7 \times 10^{10} \text{ Bq}$$

Different radionuclides differ greatly in their rate of decay. A common way to characterize this feature is through the notion of *half-life*. Half-life of a radionuclide (denoted by $T_{1/2}$) is the time it takes for a sample that has initially, say N_0 radionuclei to reduce to $N_0/2$. Putting $N = N_0/2$ and $t = T_{1/2}$ in Eq. (13.14), we get

$$T_{1/2} = \frac{\ln 2}{\lambda} = \frac{0.693}{\lambda} \qquad (13.17)$$

Clearly if N_0 reduces to half its value in time $T_{1/2}$, R_0 will also reduce to half its value in the same time according to Eq. (13.16).

Another related measure is the *average* or *mean life* τ. This again can be obtained from Eq. (13.14). The number of nuclei which decay in the time interval t to $t + \Delta t$ is $R(t)\Delta t$ $(= \lambda N_0 e^{-\lambda t}\Delta t)$. Each of them has lived for time t. Thus the total life of all these nuclei would be $t\lambda N_0 e^{-\lambda t}\Delta t$. It is clear that some nuclei may live for a short time while others may live longer. Therefore to obtain the mean life, we have to sum (or integrate) this expression over all times from 0 to ∞, and divide by the total number N_0 of nuclei at $t = 0$. Thus,

$$\tau = \frac{\lambda N_0 \int_0^\infty t e^{-\lambda t} dt}{N_0} = \lambda \int_0^\infty t e^{-\lambda t} dt$$

One can show by performing this integral that

$$\tau = 1/\lambda$$

We summarise these results with the following:

$$T_{1/2} = \frac{\ln 2}{\lambda} = \tau \ln 2 \qquad (13.18)$$

MARIE SKLODOWSKA CURIE (1867-1934)

Marie Sklodowska Curie (1867-1934) Born in Poland. She is recognised both as a physicist and as a chemist. The discovery of radioactivity by Henri Becquerel in 1896 inspired Marie and her husband Pierre Curie in their researches and analyses which led to the isolation of radium and polonium elements. She was the first person to be awarded two Nobel Prizes- for Physics in 1903 and for Chemistry in 1911.

Radioactive elements (e.g., tritium, plutonium) which are short-lived i.e., have half-lives much less than the age of the universe (~ 15 billion years) have obviously decayed long ago and are not found in nature. They can, however, be produced artificially in nuclear reactions.

EXAMPLE 13.4

Example 13.4 The half-life of $^{238}_{92}U$ undergoing α-decay is 4.5×10^9 years. What is the activity of 1g sample of $^{238}_{92}U$?

Solution

$T_{1/2} = 4.5 \times 10^9$ y
$= 4.5 \times 10^9$ y x 3.16×10^7 s/y
$= 1.42 \times 10^{17}$s

One k mol of any isotope contains Avogadro's number of atoms, and so 1g of $^{238}_{92}U$ contains

$$\frac{1}{238\times10^{-3}}\text{kmol} \times 6.025 \times 10^{26} \text{ atoms/kmol}$$

$= 25.3 \times 10^{20}$ atoms.

The decay rate R is

$R = \lambda N$

$$= \frac{0.693}{T_{1/2}}N = \frac{0.693\times25.3\times10^{20}}{1.42\times10^{17}}\text{s}^{-1}$$

$= 1.23 \times 10^{4}\ \text{s}^{-1}$

$= 1.23 \times 10^{4}$ Bq

EXAMPLE 13.4

Example 13.5 Tritium has a half-life of 12.5 y undergoing beta decay. What fraction of a sample of pure tritium will remain undecayed after 25 y.

Solution

By definition of half-life, half of the initial sample will remain undecayed after 12.5 y. In the next 12.5 y, one-half of these nuclei would have decayed. Hence, one fourth of the sample of the initial pure tritium will remain undecayed.

EXAMPLE 13.5

13.6.2 Alpha decay

A well-known example of alpha decay is the decay of uranium $^{238}_{92}U$ to thorium $^{234}_{90}Th$ with the emission of a helium nucleus $^{4}_{2}He$

$$^{238}_{92}U \rightarrow {}^{234}_{90}Th + {}^{4}_{2}He \qquad (\alpha\text{-decay}) \qquad (13.19)$$

In α-decay, the mass number of the product nucleus (daughter nucleus) is four less than that of the decaying nucleus (parent nucleus), while the atomic number decreases by two. In general, α-decay of a parent nucleus $^{A}_{Z}X$ results in a daughter nucleus $^{A-4}_{Z-2}Y$

$$^{A}_{Z}X \rightarrow {}^{A-4}_{Z-2}Y + {}^{4}_{2}He \qquad (13.20)$$

From Einstein's mass-energy equivalance relation [Eq. (13.6)] and energy conservation, it is clear that this spontaneous decay is possible only when the total mass of the decay products is less than the mass of the initial nucleus. This difference in mass appears as kinetic energy of the products. By referring to a table of nuclear masses, one can check that the total mass of $^{234}_{90}Th$ and $^{4}_{2}He$ is indeed less than that of $^{238}_{92}U$.

The disintegration energy or the Q-value of a nuclear reaction is the difference between the initial mass energy and the total mass energy of the decay products. For α-decay

$$Q = (m_X - m_Y - m_{He})\, c^2 \qquad (13.21)$$

Q is also the net kinetic energy gained in the process or, if the initial nucleus X is at rest, the kinetic energy of the products. Clearly, $Q > 0$ for exothermic processes such as α-decay.

Example 13.6 We are given the following atomic masses:

$^{238}_{92}U$ = 238.05079 u $^{4}_{2}He$ = 4.00260 u

$^{234}_{90}Th$ = 234.04363 u $^{1}_{1}H$ = 1.00783 u

$^{237}_{91}Pa$ = 237.05121 u

Here the symbol Pa is for the element protactinium (Z = 91).

(a) Calculate the energy released during the alpha decay of $^{238}_{92}U$.

(b) Show that $^{238}_{92}U$ can not spontaneously emit a proton.

Solution

(a) The alpha decay of $^{238}_{92}U$ is given by Eq. (13.20). The energy released in this process is given by

$Q = (M_U - M_{Th} - M_{He})\, c^2$

Substituting the atomic masses as given in the data, we find

$Q = (238.05079 - 234.04363 - 4.00260)\text{u} \times c^2$

$= (0.00456\ \text{u})\, c^2$

$= (0.00456\ \text{u})\ (931.5\ \text{MeV/u})$

$= 4.25$ MeV.

(b) If $^{238}_{92}U$ spontaneously emits a proton, the decay process would be

$^{238}_{92}U \rightarrow {}^{237}_{91}Pa + {}^{1}_{1}H$

The Q for this process to happen is

$= (M_U - M_{Pa} - M_H)\, c^2$

$= (238.05079 - 237.05121 - 1.00783)\ \text{u} \times c^2$

$= (-\ 0.00825\ \text{u})\, c^2$

$= -\ (0.00825\ \text{u})(931.5\ \text{MeV/u})$

$= -\ 7.68$ MeV

Thus, the Q of the process is negative and therefore it cannot proceed spontaneously. We will have to supply an energy of 7.68 MeV to a $^{238}_{92}U$ nucleus to make it emit a proton.

Example 13.6

13.6.3 Beta decay

In beta decay, a nucleus spontaneously emits an electron (β^- decay) or a positron (β^+ decay). A common example of β^- decay is

$$^{32}_{15}P \rightarrow {}^{32}_{16}S + e^- + \bar{\nu} \qquad (13.22)$$

and that of β^+ decay is

$$^{22}_{11}Na \rightarrow {}^{22}_{10}Ne + e^+ + \nu \qquad (13.23)$$

The decays are governed by the Eqs. (13.14) and (13.15), so that one can never predict *which* nucleus will undergo decay, but one can characterize the decay by a half-life $T_{1/2}$. For example, $T_{1/2}$ for the decays above is respectively 14.3 d and 2.6y. The emission of electron in β^- decay is accompanied by the emission of an antineutrino ($\bar{\nu}$); in β^+ decay, instead, a neutrino (ν) is generated. Neutrinos are neutral particles with very small (possiblly, even zero) mass compared to electrons. They have only weak interaction with other particles. They are, therefore, very difficult to detect, since they can penetrate large quantity of matter (even earth) without any interaction.

In both β^- and β^+ decay, the mass number A remains unchanged. In β^- decay, the atomic number Z of the nucleus goes up by 1, while in β^+ decay Z goes down by 1. The basic nuclear process underlying β^- decay is the conversion of neutron to proton

$$n \rightarrow p + e^- + \bar{\nu} \tag{13.24}$$

while for β^+ decay, it is the conversion of proton into neutron

$$p \rightarrow n + e^+ + \nu \tag{13.25}$$

Note that while a free neutron decays to proton, the decay of proton to neutron [Eq. (13.25)] is possible only inside the nucleus, since proton has smaller mass than neutron.

13.6.4 Gamma decay

Like an atom, a nucleus also has discrete energy levels - the ground state and excited states. The scale of energy is, however, very different. Atomic energy level spacings are of the order of eV, while the difference in nuclear energy levels is of the order of MeV. When a nucleus in an excited state spontaneously decays to its ground state (or to a lower energy state), a photon is emitted with energy equal to the difference in the two energy levels of the nucleus. This is the so-called *gamma decay*. The energy (MeV) corresponds to radiation of extremely short wavelength, shorter than the hard X-ray region.

Typically, a gamma ray is emitted when a α or β decay results in a daughter nucleus in an excited state. This then returns to the ground state by a single photon transition or successive transitions involving more than one photon. A familiar example is the successive emmission of gamma rays of energies 1.17 MeV and 1.33 MeV from the deexcitation of $^{60}_{28}$Ni nuclei formed from β^- decay of $^{60}_{27}$Co.

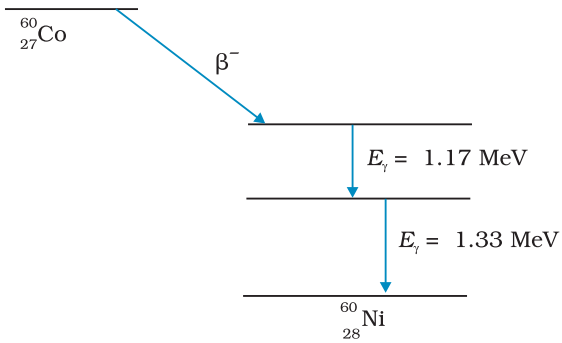


FIGURE 13.4 β-decay of $^{60}_{28}$Ni nucleus followed by emission of two γ rays from deexcitation of the daughter nucleus $^{60}_{28}$Ni .

13.7 Nuclear Energy

The curve of binding energy per nucleon E_{bn}, given in Fig. 13.1, has a long flat middle region between A = 30 and A = 170. In this region the binding energy per nucleon is nearly constant (8.0 MeV). For the lighter nuclei region, A < 30, and for the heavier nuclei region, A > 170, the binding energy per nucleon is less than 8.0 MeV, as we have noted earlier. Now, the greater the binding energy, the less is the total mass of a bound system, such as a nucleus. Consequently, if nuclei with less total binding energy transform to nuclei with greater binding energy, there will be a net energy release. This is what happens when a heavy nucleus decays into two or more intermediate mass fragments (*fission*) or when light nuclei fuse into a havier nucleus (*fusion*.)

Exothermic chemical reactions underlie conventional energy sources such as coal or petroleum. Here the energies involved are in the range of

electron volts. On the other hand, in a nuclear reaction, the energy release is of the order of MeV. Thus for the same quantity of matter, nuclear sources produce a million times more energy than a chemical source. Fission of 1 kg of uranium, for example, generates 10^{14} J of energy; compare it with burning of 1 kg of coal that gives 10^{7} J.

13.7.1 Fission

New possibilities emerge when we go beyond natural radioactive decays and study nuclear reactions by bombarding nuclei with other nuclear particles such as proton, neutron, α-particle, etc.

A most important neutron-induced nuclear reaction is fission. An example of fission is when a uranium isotope ${}^{235}_{92}\mathrm{U}$ bombarded with a neutron breaks into two intermediate mass nuclear fragments

$$ {}^{1}_{0}\mathrm{n} + {}^{235}_{92}\mathrm{U} \rightarrow {}^{236}_{92}\mathrm{U} \rightarrow {}^{144}_{56}\mathrm{Ba} + {}^{89}_{36}\mathrm{Kr} + 3\,{}^{1}_{0}\mathrm{n} \qquad (13.26) $$

The same reaction can produce other pairs of intermediate mass fragments

$$ {}^{1}_{0}\mathrm{n} + {}^{235}_{92}\mathrm{U} \rightarrow {}^{236}_{92}\mathrm{U} \rightarrow {}^{133}_{51}\mathrm{Sb} + {}^{99}_{41}\mathrm{Nb} + 4\,{}^{1}_{0}\mathrm{n} \qquad (13.27) $$

Or, as another example,

$$ {}^{1}_{0}\mathrm{n} + {}^{235}_{92}\mathrm{U} \rightarrow {}^{140}_{54}\mathrm{Xe} + {}^{94}_{38}\mathrm{Sr} + 2\,{}^{1}_{0}\mathrm{n} \qquad (13.28) $$

The fragment products are radioactive nuclei; they emit β particles in succession to achieve stable end products.

The energy released (the Q value) in the fission reaction of nuclei like uranium is of the order of 200 MeV per fissioning nucleus. This is estimated as follows:

Let us take a nucleus with A = 240 breaking into two fragments each of A = 120. Then

E_{bn} for A = 240 nucleus is about 7.6 MeV,

E_{bn} for the two A = 120 fragment nuclei is about 8.5 MeV.

∴ Gain in binding energy for nucleon is about 0.9 MeV.

Hence the total gain in binding energy is 240×0.9 or 216 MeV.

The disintegration energy in fission events first appears as the kinetic energy of the fragments and neutrons. Eventually it is transferred to the surrounding matter appearing as heat. The source of energy in nuclear reactors, which produce electricity, is nuclear fission. The enormous energy released in an atom bomb comes from uncontrolled nuclear fission. We discuss some details in the next section how a nuclear reactor functions.

13.7.2 Nuclear reactor

Notice one fact of great importance in the fission reactions given in Eqs. (13.26) to (13.28). There is a release of *extra* neutron (s) in the fission process. Averagely, 2½ neutrons are released per fission of uranium nucleus. It is a fraction since in some fission events 2 neutrons are

India's atomic energy programme

The atomic energy programme in India was launched around the time of independence under the leadership of Homi J. Bhabha (1909-1966). An early historic achievement was the design and construction of the first nuclear reactor in India (named Apsara) which went critical on August 4, 1956. It used enriched uranium as fuel and water as moderator. Following this was another notable landmark: the construction of CIRUS (Canada India Research U.S.) reactor in 1960. This 40 MW reactor used natural uranium as fuel and heavy water as moderator. Apsara and CIRUS spurred research in a wide range of areas of basic and applied nuclear science. An important milestone in the first two decades of the programme was the indigenous design and construction of the plutonium plant at Trombay, which ushered in the technology of fuel reprocessing (separating useful fissile and fertile nuclear materials from the spent fuel of a reactor) in India. Research reactors that have been subsequently commissioned include ZERLINA, PURNIMA (I, II and III), DHRUVA and KAMINI. KAMINI is the country's first large research reactor that uses U-233 as fuel. As the name suggests, the primary objective of a research reactor is not generation of power but to provide a facility for research on different aspects of nuclear science and technology. Research reactors are also an excellent source for production of a variety of radioactive isotopes that find application in diverse fields: industry, medicine and agriculture.

The main objectives of the Indian Atomic Energy programme are to provide safe and reliable electric power for the country's social and economic progress and to be self-reliant in all aspects of nuclear technology. Exploration of atomic minerals in India undertaken since the early fifties has indicated that India has limited reserves of uranium, but fairly abundant reserves of thorium. Accordingly, our country has adopted a three-stage strategy of nuclear power generation. The first stage involves the use of natural uranium as a fuel, with heavy water as moderator. The Plutonium-239 obtained from reprocessing of the discharged fuel from the reactors then serves as a fuel for the second stage — the fast breeder reactors. They are so called because they use fast neutrons for sustaining the chain reaction (hence no moderator is needed) and, besides generating power, also breed more fissile species (plutonium) than they consume. The third stage, most significant in the long term, involves using fast breeder reactors to produce fissile Uranium-233 from Thorium-232 and to build power reactors based on them.

India is currently well into the second stage of the programme and considerable work has also been done on the third — the thorium utilisation — stage. The country has mastered the complex technologies of mineral exploration and mining, fuel fabrication, heavy water production, reactor design, construction and operation, fuel reprocessing, etc. Pressurised Heavy Water Reactors (PHWRs) built at different sites in the country mark the accomplishment of the first stage of the programme. India is now more than self-sufficient in heavy water production. Elaborate safety measures both in the design and operation of reactors, as also adhering to stringent standards of radiological protection are the hallmark of the Indian Atomic Energy Programme.

produced, in some 3, etc. The extra neutrons in turn can initiate fission processes, producing still more neutrons, and so on. This leads to the possibility of a chain reaction, as was first suggested by Enrico Fermi. If the chain reaction is controlled suitably, we can get a steady energy

output. This is what happens in a nuclear reactor. If the chain reaction is uncontrolled, it leads to explosive energy output, as in a nuclear bomb.

There is, however, a hurdle in sustaining a chain reaction, as described here. It is known experimentally that slow neutrons (thermal neutrons) are much more likely to cause fission in $^{235}_{92}U$ than fast neutrons. Also fast neutrons liberated in fission would escape instead of causing another fission reaction.

The average energy of a neutron produced in fission of $^{235}_{92}U$ is 2 MeV. These neutrons unless slowed down will escape from the reactor without interacting with the uranium nuclei, unless a very large amount of fissionable material is used for sustaining the chain reaction. What one needs to do is to slow down the fast neutrons by elastic scattering with light nuclei. In fact, Chadwick's experiments showed that in an elastic collision with hydrogen the neutron almost comes to rest and proton carries away the energy. This is the same situation as when a marble hits head-on an identical marble at rest. Therefore, in reactors, light nuclei called *moderators* are provided along with the fissionable nuclei for slowing down fast neutrons. The moderators commonly used are water, heavy water (D_2O) and graphite. The Apsara reactor at the Bhabha Atomic Research Centre (BARC), Mumbai, uses water as moderator. The other Indian reactors, which are used for power production, use heavy water as moderator.

Because of the use of moderator, it is possible that the ratio, K, of number of fission produced by a given generation of neutrons to the number of fission of the preceeding generation may be greater than one. This ratio is called the *multiplication factor;* it is the measure of the growth rate of the neutrons in the reactor. For $K = 1$, the operation of the reactor is said to be *critical*, which is what we wish it to be for steady power operation. If K becomes greater than one, the reaction rate and the reactor power increases exponentially. Unless the factor K is brought down very close to unity, the reactor will become supercritical and can even explode. The explosion of the Chernobyl reactor in Ukraine in 1986 is a sad reminder that accidents in a nuclear reactor can be catastrophic.

The reaction rate is controlled through control-rods made out of neutron-absorbing material such as cadmium. In addition to control rods, reactors are provided with *safety rods* which, when required, can be inserted into the reactor and K can be reduced rapidly to less than unity.

The more abundant isotope $^{238}_{92}U$ in naturally occurring uranium is non-fissionable. When it captures a neutron, it produces the highly radioactive plutonium through these reactions

$$^{238}_{92}U + n \rightarrow {}^{239}_{92}U \rightarrow {}^{239}_{93}Np + e^- + \bar{\nu}$$

$$^{239}_{93}Np \rightarrow {}^{239}_{94}Pu + e^- + \bar{\nu} \qquad (13.29)$$

Plutonium undergoes fission with slow neutrons.

Figure 13.5 shows the schematic diagram of a nuclear reactor based on thermal neutron fission. The *core* of the reactor is the site of nuclear

Nuclear power plants in India
http://www.npcil.nic.in/main/AllProjectOperationDisplay.aspx
PHYSICS

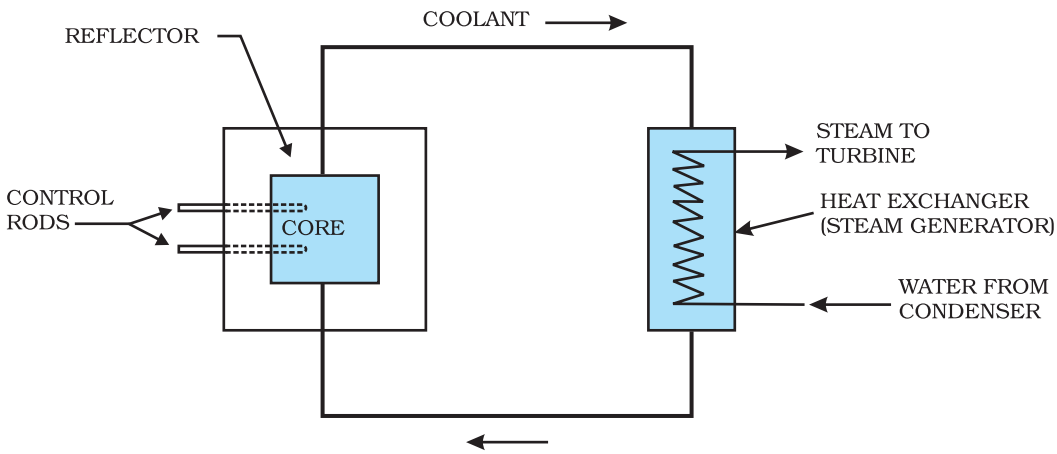


FIGURE 13.5 Schematic diagram of a nuclear reactor based on thermal neutron fission.

PHYSICS

A simplified online simulation of a nuclear reactor
http://esa21.kennesaw.edu/activities/nukeenergy/nuke.htm

fission. It contains the fuel elements in suitably fabricated form. The fuel may be say enriched uranium (i.e., one that has greater abundance of $^{235}_{92}U$ than naturally occurring uranium). The core contains a moderator to slow down the neutrons. The core is surrounded by a *reflector* to reduce leakage. The energy (heat) released in fission is continuously removed by a suitable *coolant*. A containment vessel prevents the escape of radioactive fission products. The whole assembly is shielded to check harmful radiation from coming out. The reactor can be shut down by means of rods (made of, for example, cadmium) that have high absorption of neutrons. The coolant transfers heat to a working fluid which in turn may produce stream. The steam drives turbines and generates electricity.

Like any power reactor, nuclear reactors generate considerable waste products. But nuclear wastes need special care for treatment since they are radioactive and hazardous. Elaborate safety measures, both for reactor operation as well as handling and reprocessing the spent fuel, are required. These safety measures are a distinguishing feature of the Indian Atomic Energy programme. An appropriate plan is being evolved to study the possibility of converting radioactive waste into less active and short-lived material.

13.7.3 Nuclear fusion – energy generation in stars

When two light nuclei fuse to form a larger nucleus, energy is released, since the larger nucleus is more tightly bound, as seen from the binding energy curve in Fig.13.1. Some examples of such energy liberating nuclear fusion reactions are :

$$^{1}_{1}H + ^{1}_{1}H \rightarrow ^{2}_{1}H + e^{+} + \nu + 0.42 \text{ MeV} \qquad [13.29(a)]$$

$$^{2}_{1}H + ^{2}_{1}H \rightarrow ^{3}_{2}He + n + 3.27 \text{ MeV} \qquad [13.29(b)]$$

$$^{2}_{1}H + ^{2}_{1}H \rightarrow ^{3}_{1}H + ^{1}_{1}H + 4.03 \text{ MeV} \qquad [13.29(c)]$$

In the first reaction, two protons combine to form a deuteron and a positron with a release of 0.42 MeV energy. In reaction [13.29(b)], two deuterons combine to form the light isotope of helium. In reaction (13.29c), two deuterons combine to form a triton and a proton. For fusion to take place, the two nuclei must come close enough so that attractive short-range nuclear force is able to affect them. However, since they are both positively charged particles, they experience coulomb repulsion. They, therefore, must have enough energy to overcome this coulomb barrier. The height of the barrier depends on the charges and radii of the two interacting nuclei. It can be shown, for example, that the barrier height for two protons is ~ 400 keV, and is higher for nuclei with higher charges. We can estimate the temperature at which two protons in a proton gas would (averagely) have enough energy to overcome the coulomb barrier:

$(3/2)k\,T = K \simeq 400$ keV, which gives $T \sim 3 \times 10^9$ K.

When fusion is achieved by raising the temperature of the system so that particles have enough kinetic energy to overcome the coulomb repulsive behaviour, it is called *thermonuclear fusion*.

Thermonuclear fusion is the source of energy output in the interior of stars. The interior of the sun has a temperature of 1.5×10^7 K, which is considerably less than the estimated temperature required for fusion of particles of average energy. Clearly, fusion in the sun involves protons whose energies are much above the average energy.

The fusion reaction in the sun is a multi-step process in which the hydrogen is burned into helium. Thus, the fuel in the sun is the hydrogen in its core. The *proton-proton* (p, p) *cycle* by which this occurs is represented by the following sets of reactions:

$$^1_1\text{H} + {}^1_1\text{H} \rightarrow {}^2_1\text{H} + e^+ + \nu + 0.42 \text{ MeV} \qquad \text{(i)}$$

$$e^+ + e^- \rightarrow \gamma + \gamma + 1.02 \text{ MeV} \qquad \text{(ii)}$$

$$^2_1\text{H} + {}^1_1\text{H} \rightarrow {}^3_2\text{He} + \gamma + 5.49 \text{ MeV} \qquad \text{(iii)}$$

$$^3_2\text{He} + {}^3_2\text{He} \rightarrow {}^4_2\text{He} + {}^1_1\text{H} + {}^1_1\text{H} + 12.86 \text{ MeV} \quad \text{(iv)} \qquad (13.30)$$

For the fourth reaction to occur, the first three reactions must occur twice, in which case two light helium nuclei unite to form ordinary helium nucleus. If we consider the combination 2(i) + 2(ii) + 2(iii) +(iv), the net effect is

$$4\,{}^1_1\text{H} + 2e^- \rightarrow {}^4_2\text{He} + 2\nu + 6\gamma + 26.7 \text{ MeV}$$

$$\text{or } (4\,{}^1_1\text{H} + 4e^-) \rightarrow ({}^4_2\text{He} + 2e^-) + 2\nu + 6\gamma + 26.7 \text{ MeV} \qquad (13.31)$$

Thus, four hydrogen atoms combine to form an $^4_2$He atom with a release of 26.7 MeV of energy.

Helium is not the only element that can be synthesized in the interior of a star. As the hydrogen in the core gets depleted and becomes helium, the core starts to cool. The star begins to collapse under its own gravity

which increases the temperature of the core. If this temperature increases to about 10^8 K, fusion takes place again, this time of helium nuclei into carbon. This kind of process can generate through fusion higher and higher mass number elements. But elements more massive than those near the peak of the binding energy curve in Fig. 13.1 cannot be so produced.

The age of the sun is about 5×10^9 y and it is estimated that there is enough hydrogen in the sun to keep it going for another 5 billion years. After that, the hydrogen burning will stop and the sun will begin to cool and will start to collapse under gravity, which will raise the core temperature. The outer envelope of the sun will expand, turning it into the so called *red giant*.

NUCLEAR HOLOCAUST

In a single uranium fission about 0.9×235 MeV (≈200 MeV) of energy is liberated. If each nucleus of about 50 kg of ^{235}U undergoes fission the amount of energy involved is about 4×10^{15}J. This energy is equivalent to about 20,000 tons of TNT, enough for a superexplosion. Uncontrolled release of large nuclear energy is called an atomic explosion. On August 6, 1945 an atomic device was used in warfare for the first time. The US dropped an atom bomb on Hiroshima, Japan. The explosion was equivalent to 20,000 tons of TNT. Instantly the radioactive products devastated 10 sq km of the city which had 3,43,000 inhabitants. Of this number 66,000 were killed and 69,000 were injured; more than 67% of the city's structures were destroyed.

High temperature conditions for fusion reactions can be created by exploding a fission bomb. Super-explosions equivalent to 10 megatons of explosive power of TNT were tested in 1954. Such bombs which involve fusion of isotopes of hydrogen, deuterium and tritium are called hydrogen bombs. It is estimated that a nuclear arsenal sufficient to destroy every form of life on this planet several times over is in position to be triggered by the press of a button. Such a nuclear holocaust will not only destroy the life that exists now but its radioactive fallout will make this planet unfit for life for all times. Scenarios based on theoretical calculations predict a long *nuclear winter*, as the radioactive waste will hang like a cloud in the earth's atmosphere and will absorb the sun's radiation.

13.7.4 Controlled thermonuclear fusion

The natural thermonuclear fusion process in a star is replicated in a thermonuclear fusion device. In controlled fusion reactors, the aim is to generate steady power by heating the nuclear fuel to a temperature in the range of 10^8 K. At these temperatures, the fuel is a mixture of positive ions and electrons (plasma). The challenge is to confine this plasma, since no container can stand such a high temperature. Several countries around the world including India are developing techniques in this connection. If successful, fusion reactors will hopefully supply almost unlimited power to humanity.

Example 13.7 Answer the following questions:

(a) Are the equations of nuclear reactions (such as those given in Section 13.7) 'balanced' in the sense a chemical equation (e.g., $2H_2 + O_2 \rightarrow 2\ H_2O$) is? If not, in what sense are they balanced on both sides?

(b) If both the number of protons and the number of neutrons are conserved in each nuclear reaction, in what way is mass converted into energy (or vice-versa) in a nuclear reaction?

(c) A general impression exists that mass-energy interconversion takes place only in nuclear reaction and never in chemical reaction. This is strictly speaking, incorrect. Explain.

Solution

(a) A chemical equation is balanced in the sense that the number of atoms of each element is the same on both sides of the equation. A chemical reaction merely alters the original combinations of atoms. In a nuclear reaction, elements may be transmuted. Thus, the number of atoms of each element is not necessarily conserved in a nuclear reaction. However, the number of protons and the number of neutrons are both separately conserved in a nuclear reaction. [Actually, even this is not strictly true in the realm of very high energies – what is strictly conserved is the total charge and total 'baryon number'. We need not pursue this matter here.]

In nuclear reactions (e.g., Eq. 13.26), the number of protons and the number of neutrons are the same on the two sides of the equation.

(b) We know that the binding energy of a nucleus gives a negative contribution to the mass of the nucleus (mass defect). Now, since proton number and neutron number are conserved in a nuclear reaction, the total rest mass of neutrons and protons is the same on either side of a reaction. But the total binding energy of nuclei on the left side need not be the same as that on the right hand side. The difference in these binding energies appears as energy released or absorbed in a nuclear reaction. Since binding energy contributes to mass, we say that the difference in the total mass of nuclei on the two sides get converted into energy or vice-versa. It is in these sense that a nuclear reaction is an example of mass-energy interconversion.

(c) From the point of view of mass-energy interconversion, a chemical reaction is similar to a nuclear reaction *in principle.* The energy released or absorbed in a chemical reaction can be traced to the difference in chemical (not nuclear) binding energies of atoms and molecules on the two sides of a reaction. Since, strictly speaking, chemical binding energy also gives a negative contribution (mass defect) to the total mass of an atom or molecule, we can equally well say that the difference in the total mass of atoms or molecules, on the two sides of the chemical reaction gets converted into energy or vice-versa. However, the mass defects involved in a chemical reaction are almost a million times smaller than those in a nuclear reaction.This is the reason for the general impression, (which is *incorrect*) that mass-energy interconversion does not take place in a chemical reaction.

EXAMPLE 13.7

SUMMARY

1. An atom has a nucleus. The nucleus is positively charged. The radius of the nucleus is smaller than the radius of an atom by a factor of 10^4. More than 99.9% mass of the atom is concentrated in the nucleus.
2. On the atomic scale, mass is measured in atomic mass units (u). By definition, 1 atomic mass unit (1u) is $1/12^{th}$ mass of one atom of ^{12}C; $1u = 1.660563 \times 10^{-27}$ kg.
3. A nucleus contains a neutral particle called neutron. Its mass is almost the same as that of proton
4. The atomic number Z is the number of protons in the atomic nucleus of an element. The mass number A is the total number of protons and neutrons in the atomic nucleus; $A = Z+N$; Here N denotes the number of neutrons in the nucleus.

   A nuclear species or a nuclide is represented as $^{A}_{Z}X$, where X is the chemical symbol of the species.

   Nuclides with the same atomic number Z, but different neutron number N are called *isotopes*. Nuclides with the same A are *isobars* and those with the same N are *isotones*.

   Most elements are mixtures of two or more isotopes. The atomic mass of an element is a weighted average of the masses of its isotopes. The masses are the relative abundances of the isotopes.
5. A nucleus can be considered to be spherical in shape and assigned a radius. Electron scattering experiments allow determination of the nuclear radius; it is found that radii of nuclei fit the formula

   $R = R_0 A^{1/3}$,

   where R_0 = a constant = 1.2 fm. This implies that the nuclear density is independent of A. It is of the order of 10^{17} kg/m^3.
6. Neutrons and protons are bound in a nucleus by the short-range strong nuclear force. The nuclear force does not distinguish between neutron and proton.
7. The nuclear mass M is always less than the total mass, Σm, of its constituents. The difference in mass of a nucleus and its constituents is called the *mass defect*,

   $\Delta M = (Z\, m_p + (A - Z)m_n) - M$

   Using Einstein's mass energy relation, we express this mass difference in terms of energy as

   $\Delta E_b = \Delta M\, c^2$

   The energy ΔE_b represents the *binding energy* of the nucleus. In the mass number range A = 30 to 170, the binding energy per nucleon is nearly constant, about 8 MeV/nucleon.
8. Energies associated with nuclear processes are about a million times larger than chemical process.
9. The Q-value of a nuclear process is

   Q = final kinetic energy – initial kinetic energy.

   Due to conservation of mass-energy, this is also,

   Q = (sum of initial masses – sum of final masses)c^2
10. Radioactivity is the phenomenon in which nuclei of a given species transform by giving out α or β or γ rays; α-rays are helium nuclei;

β-rays are electrons. γ-rays are electromagnetic radiation of wavelengths shorter than X-rays;

11. Law of radioactive decay : $N(t) = N(0)\, e^{-\lambda t}$

    where λ is the decay constant or disintegration constant.

    The half-life $T_{1/2}$ of a radionuclide is the time in which N has been reduced to one-half of its initial value. The mean life τ is the time at which N has been reduced to e^{-1} of its initial value

    $$T_{1/2} = \frac{\ln 2}{\lambda} = \tau \ln 2$$

12. Energy is released when less tightly bound nuclei are transmuted into more tightly bound nuclei. In fission, a heavy nucleus like ${}^{235}_{92}\mathrm{U}$ breaks into two smaller fragments, e.g., ${}^{235}_{92}\mathrm{U} + {}^{1}_{0}\mathrm{n} \rightarrow {}^{133}_{51}\mathrm{Sb} + {}^{99}_{41}\mathrm{Nb} + 4\,{}^{1}_{0}\mathrm{n}$

13. The fact that more neutrons are produced in fission than are consumed gives the possibility of a chain reaction with each neutron that is produced triggering another fission. The chain reaction is uncontrolled and rapid in a nuclear bomb explosion. It is controlled and steady in a nuclear reactor. In a reactor, the value of the neutron multiplication factor k is maintained at 1.

14. In fusion, lighter nuclei combine to form a larger nucleus. Fusion of hydrogen nuclei into helium nuclei is the source of energy of all stars including our sun.

Physical Quantity	Symbol	Dimensions	Units	Remarks
Atomic mass unit		[M]	u	Unit of mass for expressing atomic or nuclear masses. One atomic mass unit equals $1/12^{th}$ of the mass of ${}^{12}\mathrm{C}$ atom.
Disintegration or decay constant	λ	$[T^{-1}]$	s^{-1}	
Half-life	$T_{1/2}$	[T]	s	Time taken for the decay of one-half of the initial number of nuclei present in a radioactive sample.
Mean life	τ	[T]	s	Time at which number of nuclei has been reduced to e^{-1} of its initial value
Activity of a radio-active sample	R	$[T^{-1}]$	Bq	Measure of the activity of a radioactive source.

POINTS TO PONDER

1. The density of nuclear matter is independent of the size of the nucleus. The mass density of the atom does not follow this rule.
2. The radius of a nucleus determined by electron scattering is found to be slightly different from that determined by alpha-particle scattering.

This is because electron scattering senses the charge distribution of the nucleus, whereas alpha and similar particles sense the nuclear matter.

3. After Einstein showed the equivalence of mass and energy, $E = mc^2$, we cannot any longer speak of separate laws of conservation of mass and conservation of energy, but we have to speak of a unified law of conservation of mass and energy. The most convincing evidence that this principle operates in nature comes from nuclear physics. It is central to our understanding of nuclear energy and harnessing it as a source of power. Using the principle, Q of a nuclear process (decay or reaction) can be expressed also in terms of initial and final masses.
4. The nature of the binding energy (per nucleon) curve shows that exothermic nuclear reactions are possible, when two light nuclei fuse or when a heavy nucleus undergoes fission into nuclei with intermediate mass.
5. For fusion, the light nuclei must have sufficient initial energy to overcome the coulomb potential barrier. That is why fusion requires very high temperatures.
6. Although the binding energy (per nucleon) curve is smooth and slowly varying, it shows peaks at nuclides like ^{4}He, ^{16}O etc. This is considered as evidence of atom-like shell structure in nuclei.
7. Electrons and positron are a particle-antiparticle pair. They are identical in mass; their charges are equal in magnitude and opposite. (It is found that when an electron and a positron come together, they annihilate each other giving energy in the form of gamma-ray photons.)
8. In â$^-$-decay (electron emission), the particle emitted along with electron is anti-neutrino ($\bar{\nu}$). On the other hand, the particle emitted in β^+-decay (positron emission) is neutrino (ν). Neutrino and anti-neutrino are a particle-antiparticle pair. There are anti particles associated with every particle. What should be antiproton which is the anti particle of the proton?
9. A free neutron is unstable ($\mathrm{n} \rightarrow \mathrm{p} + e^- + \bar{\nu}$). But a similar free proton decay is not possible, since a proton is (slightly) lighter than a neutron.
10. Gamma emission usually follows alpha or beta emission. A nucleus in an excited (higher) state goes to a lower state by emitting a gamma photon. A nucleus may be left in an excited state after alpha or beta emission. Successive emission of gamma rays from the same nucleus (as in case of ^{60}Ni, Fig. 13.4) is a clear proof that nuclei also have discrete energy levels as do the atoms.
11. Radioactivity is an indication of the instability of nuclei. Stability requires the ratio of neutron to proton to be around 1:1 for light nuclei. This ratio increases to about 3:2 for heavy nuclei. (More neutrons are required to overcome the effect of repulsion among the protons.) Nuclei which are away from the stability ratio, i.e., nuclei which have an excess of neutrons or protons are unstable. In fact, only about 10% of knon isotopes (of all elements), are stable. Others have been either artificially produced in the laboratory by bombarding α, p, d, n or other particles on targets of stable nuclear species or identified in astronomical observations of matter in the universe.

EXERCISES

You may find the following data useful in solving the exercises:

$e = 1.6\times10^{-19}$C $\quad N = 6.023\times10^{23}$ per mole

$1/(4\pi\varepsilon_0) = 9 \times 10^9$ N m^2/C^2 $\quad k = 1.381\times10^{-23}$J ^{0}K^{-1}

1 MeV $= 1.6\times10^{-13}$J $\quad$ 1 u = 931.5 MeV/c^2

1 year $= 3.154\times10^7$ s

$m_H = 1.007825$ u $\quad m_n = 1.008665$ u

$m(^4_2\text{He}) = 4.002603$ u $\quad m_e = 0.000548$ u

13.1 (a) Two stable isotopes of lithium $^6_3$Li and $^7_3$Li have respective abundances of 7.5% and 92.5%. These isotopes have masses 6.01512 u and 7.01600 u, respectively. Find the atomic mass of lithium.

(b) Boron has two stable isotopes, $^{10}_5$B and $^{11}_5$B. Their respective masses are 10.01294 u and 11.00931 u, and the atomic mass of boron is 10.811 u. Find the abundances of $^{10}_5$B and $^{11}_5$B.

13.2 The three stable isotopes of neon: $^{20}_{10}$Ne, $^{21}_{10}$Ne and $^{22}_{10}$Ne have respective abundances of 90.51%, 0.27% and 9.22%. The atomic masses of the three isotopes are 19.99 u, 20.99 u and 21.99 u, respectively. Obtain the average atomic mass of neon.

13.3 Obtain the binding energy (in MeV) of a nitrogen nucleus $\left(^{14}_7\text{N}\right)$, given $m\left(^{14}_7\text{N}\right)$=14.00307 u

13.4 Obtain the binding energy of the nuclei $^{56}_{26}$Fe and $^{209}_{83}$Bi in units of MeV from the following data:

$m\left(^{56}_{26}\text{Fe}\right) = 55.934939$ u $\quad m\left(^{209}_{83}\text{Bi}\right) = 208.980388$ u

13.5 A given coin has a mass of 3.0 g. Calculate the nuclear energy that would be required to separate all the neutrons and protons from each other. For simplicity assume that the coin is entirely made of $^{63}_{29}$Cu atoms (of mass 62.92960 u).

13.6 Write nuclear reaction equations for

(i) α-decay of $^{226}_{88}$Ra $\quad$ (ii) α-decay of $^{242}_{94}$Pu

(iii) β^--decay of $^{32}_{15}$P $\quad$ (iv) β^--decay of $^{210}_{83}$Bi

(v) β^+-decay of $^{11}_6$C $\quad$ (vi) β^+-decay of $^{97}_{43}$Tc

(vii) Electron capture of $^{120}_{54}$Xe

13.7 A radioactive isotope has a half-life of T years. How long will it take the activity to reduce to a) 3.125%, b) 1% of its original value?

13.8 The normal activity of living carbon-containing matter is found to be about 15 decays per minute for every gram of carbon. This activity arises from the small proportion of radioactive $^{14}_6$C present with the stable carbon isotope $^{12}_6$C. When the organism is dead, its interaction with the atmosphere (which maintains the above equilibrium activity) ceases and its activity begins to drop. From the known half-life (5730 years) of $^{14}_6$C, and the measured activity, the age of the specimen can be approximately estimated. This is the principle of $^{14}_6$C dating

used in archaeology. Suppose a specimen from Mohenjodaro gives an activity of 9 decays per minute per gram of carbon. Estimate the approximate age of the Indus-Valley civilisation.

13.9 Obtain the amount of $^{60}_{27}\text{Co}$ necessary to provide a radioactive source of 8.0 mCi strength. The half-life of $^{60}_{27}\text{Co}$ is 5.3 years.

13.10 The half-life of $^{90}_{38}\text{Sr}$ is 28 years. What is the disintegration rate of 15 mg of this isotope?

13.11 Obtain approximately the ratio of the nuclear radii of the gold isotope $^{197}_{79}\text{Au}$ and the silver isotope $^{107}_{47}\text{Ag}$.

13.12 Find the Q-value and the kinetic energy of the emitted α-particle in the α-decay of (a) $^{226}_{88}\text{Ra}$ and (b) $^{220}_{86}\text{Rn}$.

Given $m\,(^{226}_{88}\text{Ra}) = 226.02540$ u, $m\,(^{222}_{86}\text{Rn}) = 222.01750$ u,

$m\,(^{222}_{86}\text{Rn}) = 220.01137$ u, $m\,(^{216}_{84}\text{Po}) = 216.00189$ u.

13.13 The radionuclide ^{11}C decays according to

$$^{11}_{6}\text{C} \rightarrow ^{11}_{5}\text{B} + e^{+} + \nu: \quad T_{1/2} = 20.3 \text{ min}$$

The maximum energy of the emitted positron is 0.960 MeV.

Given the mass values:

$m\,(^{11}_{6}\text{C}) = 11.011434$ u and $m\,(^{11}_{6}\text{B}) = 11.009305$ u,

calculate Q and compare it with the maximum energy of the positron emitted.

13.14 The nucleus $^{23}_{10}\text{Ne}$ decays by β^- emission. Write down the β-decay equation and determine the maximum kinetic energy of the electrons emitted. Given that:

$m\,(^{23}_{10}\text{Ne}) = 22.994466$ u

$m\,(^{23}_{11}\text{Na}) = 22.089770$ u.

13.15 The Q value of a nuclear reaction $A + b \rightarrow C + d$ is defined by

$Q = [\,m_A + m_b - m_C - m_d]c^2$

where the masses refer to the respective nuclei. Determine from the given data the Q-value of the following reactions and state whether the reactions are exothermic or endothermic.

(i) $^{1}_{1}\text{H} + ^{3}_{1}\text{H} \rightarrow ^{2}_{1}\text{H} + ^{2}_{1}\text{H}$

(ii) $^{12}_{6}\text{C} + ^{12}_{6}\text{C} \rightarrow ^{20}_{10}\text{Ne} + ^{4}_{2}\text{He}$

Atomic masses are given to be

$m\,(^{2}_{1}\text{H}) = 2.014102$ u

$m\,(^{3}_{1}\text{H}) = 3.016049$ u

$m\,(^{12}_{6}\text{C}) = 12.000000$ u

$m\,(^{20}_{10}\text{Ne}) = 19.992439$ u

13.16 Suppose, we think of fission of a $^{56}_{26}\text{Fe}$ nucleus into two equal fragments, $^{28}_{13}\text{Al}$. Is the fission energetically possible? Argue by working out Q of the process. Given $m\,(^{56}_{26}\text{Fe}) = 55.93494$ u and $m\,(^{28}_{13}\text{Al}) = 27.98191$ u.

13.17 The fission properties of $^{239}_{94}\text{Pu}$ are very similar to those of $^{235}_{92}\text{U}$. The average energy released per fission is 180 MeV. How much energy,

in MeV, is released if all the atoms in 1 kg of pure $^{239}_{94}Pu$ undergo fission?

13.18 A 1000 MW fission reactor consumes half of its fuel in 5.00 y. How much $^{235}_{92}U$ did it contain initially? Assume that the reactor operates 80% of the time, that all the energy generated arises from the fission of $^{235}_{92}U$ and that this nuclide is consumed only by the fission process.

13.19 How long can an electric lamp of 100W be kept glowing by fusion of 2.0 kg of deuterium? Take the fusion reaction as

$$^{2}_{1}H + ^{2}_{1}H \rightarrow ^{3}_{2}He + n + 3.27 \text{ MeV}$$

13.20 Calculate the height of the potential barrier for a head on collision of two deuterons. (Hint: The height of the potential barrier is given by the Coulomb repulsion between the two deuterons when they just touch each other. Assume that they can be taken as hard spheres of radius 2.0 fm.)

13.21 From the relation $R = R_0 A^{1/3}$, where R_0 is a constant and A is the mass number of a nucleus, show that the nuclear matter density is nearly constant (i.e. independent of A).

13.22 For the β^+ (positron) emission from a nucleus, there is another competing process known as electron capture (electron from an inner orbit, say, the K–shell, is captured by the nucleus and a neutrino is emitted).

$$e^+ + ^{A}_{Z}X \rightarrow ^{A}_{Z-1}Y + \nu$$

Show that if β^+ emission is energetically allowed, electron capture is necessarily allowed but not vice–versa.

ADDITIONAL EXERCISES

13.23 In a periodic table the average atomic mass of magnesium is given as 24.312 u. The average value is based on their relative natural abundance on earth. The three isotopes and their masses are $^{24}_{12}Mg$ (23.98504u), $^{25}_{12}Mg$ (24.98584u) and $^{26}_{12}Mg$ (25.98259u). The natural abundance of $^{24}_{12}Mg$ is 78.99% by mass. Calculate the abundances of other two isotopes.

13.24 The neutron separation energy is defined as the energy required to remove a neutron from the nucleus. Obtain the neutron separation energies of the nuclei $^{41}_{20}Ca$ and $^{27}_{13}Al$ from the following data:

$m(^{40}_{20}Ca) = 39.962591$ u

$m(^{41}_{20}Ca) = 40.962278$ u

$m(^{26}_{13}Al) = 25.986895$ u

$m(^{27}_{13}Al) = 26.981541$ u

13.25 A source contains two phosphorous radio nuclides $^{32}_{15}P$ ($T_{1/2}$ = 14.3d) and $^{33}_{15}P$ ($T_{1/2}$ = 25.3d). Initially, 10% of the decays come from $^{33}_{15}P$. How long one must wait until 90% do so?

13.26 Under certain circumstances, a nucleus can decay by emitting a particle more massive than an α-particle. Consider the following decay processes:

$$^{223}_{88}Ra \rightarrow ^{209}_{82}Pb + ^{14}_{6}C$$

$^{223}_{88}\text{Ra} \rightarrow {}^{219}_{86}\text{Rn} + {}^{4}_{2}\text{He}$

Calculate the Q-values for these decays and determine that both are energetically allowed.

13.27 Consider the fission of $^{238}_{92}\text{U}$ by fast neutrons. In one fission event, no neutrons are emitted and the final end products, after the beta decay of the primary fragments, are $^{140}_{58}\text{Ce}$ and $^{99}_{44}\text{Ru}$. Calculate Q for this fission process. The relevant atomic and particle masses are

$m(^{238}_{92}\text{U}) = 238.05079$ u

$m(^{140}_{58}\text{Ce}) = 139.90543$ u

$m(^{99}_{44}\text{Ru}) = 98.90594$ u

13.28 Consider the D–T reaction (deuterium–tritium fusion)

$^{2}_{1}\text{H} + {}^{3}_{1}\text{H} \rightarrow {}^{4}_{2}\text{He} + \text{n}$

(a) Calculate the energy released in MeV in this reaction from the data:

$m(^{2}_{1}\text{H}) = 2.014102$ u

$m(^{3}_{1}\text{H}) = 3.016049$ u

(b) Consider the radius of both deuterium and tritium to be approximately 2.0 fm. What is the kinetic energy needed to overcome the coulomb repulsion between the two nuclei? To what temperature must the gas be heated to initiate the reaction?

(Hint: Kinetic energy required for one fusion event =average thermal kinetic energy available with the interacting particles = $2(3kT/2)$; k = Boltzman's constant, T = absolute temperature.)

13.29 Obtain the maximum kinetic energy of β-particles, and the radiation frequencies of γ decays in the decay scheme shown in Fig. 13.6. You are given that

$m(^{198}\text{Au}) = 197.968233$ u

$m(^{198}\text{Hg}) = 197.966760$ u

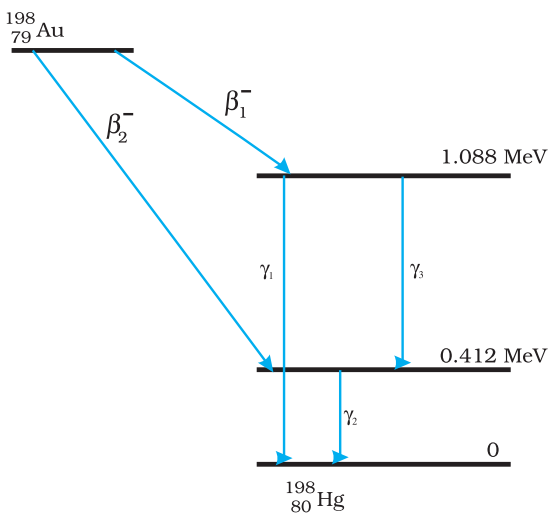


FIGURE 13.6

13.30 Calculate and compare the energy released by a) fusion of 1.0 kg of hydrogen deep within Sun and b) the fission of 1.0 kg of ^{235}U in a fission reactor.

13.31 Suppose India had a target of producing by 2020 AD, 200,000 MW of electric power, ten percent of which was to be obtained from nuclear power plants. Suppose we are given that, on an average, the efficiency of utilization (i.e. conversion to electric energy) of thermal energy produced in a reactor was 25%. How much amount of fissionable uranium would our country need per year by 2020? Take the heat energy per fission of ^{235}U to be about 200MeV.

Chapter Fourteen

SEMICONDUCTOR ELECTRONICS: MATERIALS, DEVICES AND SIMPLE CIRCUITS

14.1 Introduction

Devices in which a controlled flow of electrons can be obtained are the basic *building blocks* of all the electronic circuits. Before the discovery of transistor in 1948, such devices were mostly vacuum tubes (also called valves) like the vacuum diode which has two electrodes, viz., anode (often called plate) and cathode; triode which has three electrodes – cathode, plate and grid; tetrode and pentode (respectively with 4 and 5 electrodes). In a vacuum tube, the electrons are supplied by a heated cathode and the controlled flow of these electrons *in vacuum* is obtained by varying the voltage between its different electrodes. Vacuum is required in the inter-electrode space; otherwise the moving electrons may lose their energy on collision with the air molecules in their path. In these devices the electrons can flow only from the cathode to the anode (i.e., only in one direction). Therefore, such devices are generally referred to as *valves*. These vacuum tube devices are bulky, consume high power, operate generally at high voltages (~100 V) and have limited life and low reliability. The seed of the development of modern *solid-state semiconductor electronics* goes back to 1930's when it was realised that some solid-state semiconductors and their junctions offer the possibility of controlling the number and the direction of flow of charge carriers through them. Simple excitations like light, heat or small applied voltage can change the number of mobile charges in a semiconductor. Note that the supply

and flow of charge carriers in the semiconductor devices are *within the solid itself*, while in the earlier vacuum tubes/valves, the mobile electrons were obtained from a heated cathode and they were made to flow in an *evacuated* space or vacuum. No external heating or large evacuated space is required by the semiconductor devices. They are small in size, consume low power, operate at low voltages and have long life and high reliability. Even the Cathode Ray Tubes (CRT) used in television and computer monitors which work on the principle of vacuum tubes are being replaced by Liquid Crystal Display (LCD) monitors with supporting solid state electronics. Much before the full implications of the semiconductor devices was formally understood, a naturally occurring crystal of *galena* (Lead sulphide, PbS) with a metal point contact attached to it was used as *detector* of radio waves.

In the following sections, we will introduce the basic concepts of semiconductor physics and discuss some semiconductor devices like junction diodes (a 2-electrode device) and bipolar junction transistor (a 3-electrode device). A few circuits illustrating their applications will also be described.

14.2 Classification of Metals, Conductors and Semiconductors

On the basis of conductivity

On the basis of the relative values of electrical conductivity (σ) or resistivity ($\rho = 1/\sigma$), the solids are broadly classified as:

(i) ***Metals:*** They possess very low resistivity (or high conductivity).
$\rho \sim 10^{-2} - 10^{-8}\ \Omega$ m
$\sigma \sim 10^{2} - 10^{8}$ S m^{-1}

(ii) ***Semiconductors:*** They have resistivity or conductivity intermediate to metals and insulators.
$\rho \sim 10^{-5} - 10^{6}\ \Omega$ m
$\sigma \sim 10^{5} - 10^{-6}$ S m^{-1}

(iii) ***Insulators:*** They have high resistivity (or low conductivity).
$\rho \sim 10^{11} - 10^{19}\ \Omega$ m
$\sigma \sim 10^{-11} - 10^{-19}$ S m^{-1}

The values of ρ and σ given above are indicative of magnitude and could well go outside the ranges as well. Relative values of the resistivity are not the only criteria for distinguishing metals, insulators and semiconductors from each other. There are some other differences, which will become clear as we go along in this chapter.

Our interest in this chapter is in the study of semiconductors which could be:

(i) *Elemental semiconductors*: Si and Ge

(ii) *Compound semiconductors*: Examples are:
- Inorganic: CdS, GaAs, CdSe, InP, etc.
- Organic: anthracene, doped pthalocyanines, etc.
- Organic polymers: polypyrrole, polyaniline, polythiophene, etc.

Most of the currently available semiconductor devices are based on elemental semiconductors Si or Ge and compound *inorganic*

semiconductors. However, after 1990, a few semiconductor devices using organic semiconductors and semiconducting polymers have been developed signalling the birth of a futuristic technology of polymer-electronics and molecular-electronics. In this chapter, we will restrict ourselves to the study of inorganic semiconductors, particularly elemental semiconductors Si and Ge. The general concepts introduced here for discussing the elemental semiconductors, by-and-large, apply to most of the compound semiconductors as well.

On the basis of energy bands

According to the Bohr atomic model, in an *isolated atom* the energy of any of its electrons is decided by the orbit in which it revolves. But when the atoms come together to form a solid they are close to each other. So the outer orbits of electrons from neighbouring atoms would come very close or could even overlap. This would make the nature of electron motion in a solid very different from that in an isolated atom.

Inside the crystal each electron has a unique position and no two electrons see exactly the same pattern of surrounding charges. Because of this, each electron will have a different *energy level*. These different energy levels with continuous energy variation form what are called *energy bands*. The energy band which includes the energy levels of the valence electrons is called the *valence band*. The energy band above the valence band is called the *conduction band*. With no external energy, all the valence electrons will reside in the valence band. If the lowest level in the conduction band happens to be lower than the highest level of the valence band, the electrons from the valence band can easily move into the conduction band. Normally the conduction band is empty. But when it overlaps on the valence band electrons can move freely into it. This is the case with metallic conductors.

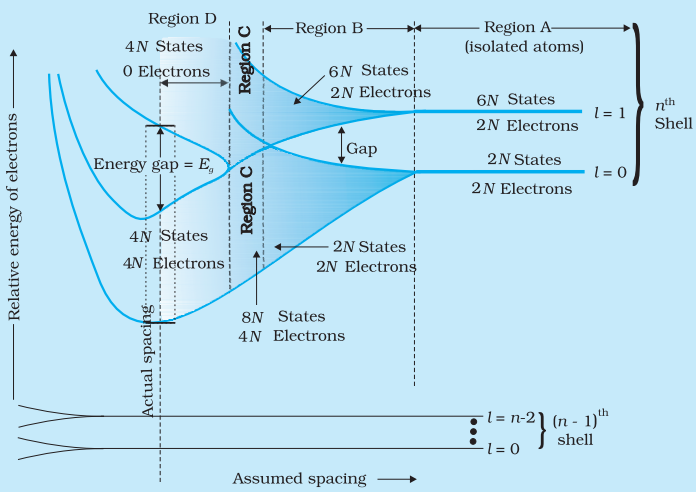
If there is some gap between the conduction band and the valence band, electrons in the valence band all remain bound and no free electrons are available in the conduction band. This makes the material an insulator. But some of the electrons from the valence band may gain external energy to cross the gap between the conduction band and the valence band. Then these electrons will move into the conduction band. At the same time they will create vacant energy levels in the valence band where other valence electrons can move. Thus the process creates the possibility of conduction due to electrons in conduction band as well as due to vacancies in the valence band.

Let us consider what happens in the case of Si or Ge crystal containing N atoms. For Si, the outermost orbit is the third orbit ($n = 3$), while for Ge it is the fourth orbit ($n = 4$). The number of electrons in the outermost orbit is 4 ($2s$ and $2p$ electrons). Hence, the total number of outer electrons in the crystal is $4N$. The maximum possible number of electrons in the outer orbit is 8 ($2s + 6p$ electrons). So, for the $4N$ valence electrons there are $8N$ available energy states. These $8N$ discrete energy levels can either form a continuous band or they may be grouped in different bands depending upon the distance between the atoms in the crystal (see box on Band Theory of Solids).

At the distance between the atoms in the crystal lattices of Si and Ge, the energy band of these $8N$ states is split apart into two which are separated by an *energy gap* E_g (Fig. 14.1). The lower band which is

completely occupied by the $4N$ valence electrons at temperature of absolute zero is the *valence band.* The other band consisting of $4N$ energy states, called the *conduction band*, is completely empty at absolute zero.

Band theory of solids



Consider that the Si or Ge crystal contains N atoms. Electrons of each atom will have discrete energies in different orbits. The electron energy will be same if all the atoms are *isolated,* i.e., separated from each other by a large distance. However, in a crystal, the atoms are close to each other (2 to 3 Å) and therefore the electrons interact with each other and also with the neighbouring atomic cores. The overlap (or interaction) will be more felt by the electrons in the outermost orbit while the inner orbit or core electron energies may remain unaffected. Therefore, for understanding electron energies in Si or Ge crystal, we need to consider the changes in the energies of the electrons in the outermost orbit only. For Si, the outermost orbit is the third orbit (n = 3), while for Ge it is the fourth orbit (n = 4). The number of electrons in the outermost orbit is 4 ($2s$ and $2p$ electrons). Hence, the total number of outer electrons in the crystal is $4N$. The maximum possible number of outer electrons in the orbit is 8 ($2s + 6p$ electrons). So, out of the $4N$ electrons, $2N$ electrons are in the $2N$ *s-states* (orbital quantum number $l = 0$) and 2N electrons are in the available $6N$ *p-states*. Obviously, some p-electron states are empty as shown in the extreme right of Figure. This is the case of well separated or isolated atoms [region A of Figure].

Suppose these atoms start coming nearer to each other to form a solid. The energies of these electrons in the outermost orbit may change (both increase and decrease) due to the interaction between the electrons of different atoms. The $6N$ states for $l = 1$, which originally had identical energies in the isolated atoms, spread out and form an *energy band* [region B in Figure]. Similarly, the $2N$ states for $l = 0$, having identical energies in the isolated atoms, split into a second band (carefully see the region B of Figure) separated from the first one by an *energy gap*.

At still smaller spacing, however, there comes a region in which the bands merge with each other. The lowest energy state that is a split from the upper atomic level appears to drop below the upper state that has come from the lower atomic level. In this region (region C in Figure), *no energy gap exists where the upper and lower energy states get mixed.*

Finally, if the distance between the atoms further decreases, the energy bands again split apart and are separated by an *energy gap* E_g (region D in Figure). The total number of available energy states $8N$ has been *re-apportioned* between the two bands ($4N$ states each in the lower and upper energy bands). Here the significant point is that there are exactly as many states in the lower band ($4N$) as there are available valence electrons from the atoms ($4N$).

Therefore, this band (called the *valence band*) is completely filled while the upper band is completely empty. The upper band is called the *conduction band.*

The lowest energy level in the conduction band is shown as E_C and highest energy level in the valence band is shown as E_V. Above E_C and below E_V there are a large number of closely spaced energy levels, as shown in Fig. 14.1.

The gap between the top of the valence band and bottom of the conduction band is called the *energy band gap* (Energy gap E_g). It may be large, small, or zero, depending upon the material. These different situations, are depicted in Fig. 14.2 and discussed below:

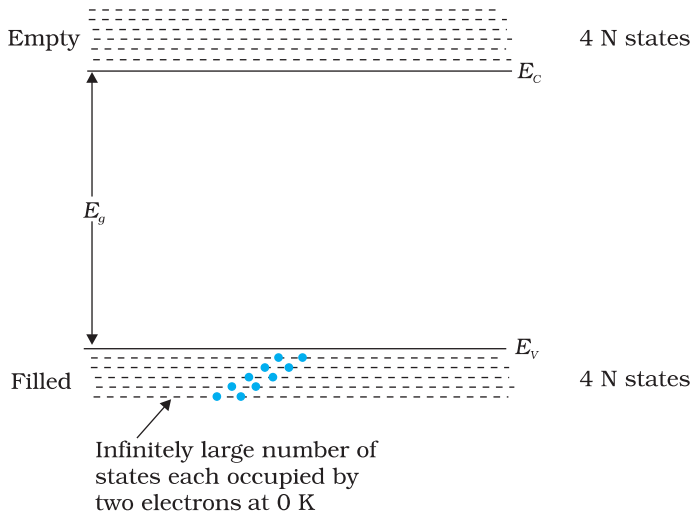


FIGURE 14.1 The energy band positions in a semiconductor at 0 K. The upper band, called the conduction band, consists of infinitely large number of closely spaced energy states. The lower band, called the valence band, consists of closely spaced completely filled energy states.

Case I: This refers to a situation, as shown in Fig. 14.2(a). One can have a metal either when the conduction band is partially filled and the balanced band is partially empty or when the conduction and valance bands overlap. When there is overlap electrons from valence band can easily move into the conduction band. This situation makes a large number of electrons available for electrical conduction. When the valence band is partially empty, electrons from its lower level can move to higher level making conduction possible. Therefore, the resistance of such materials is low or the conductivity is high.

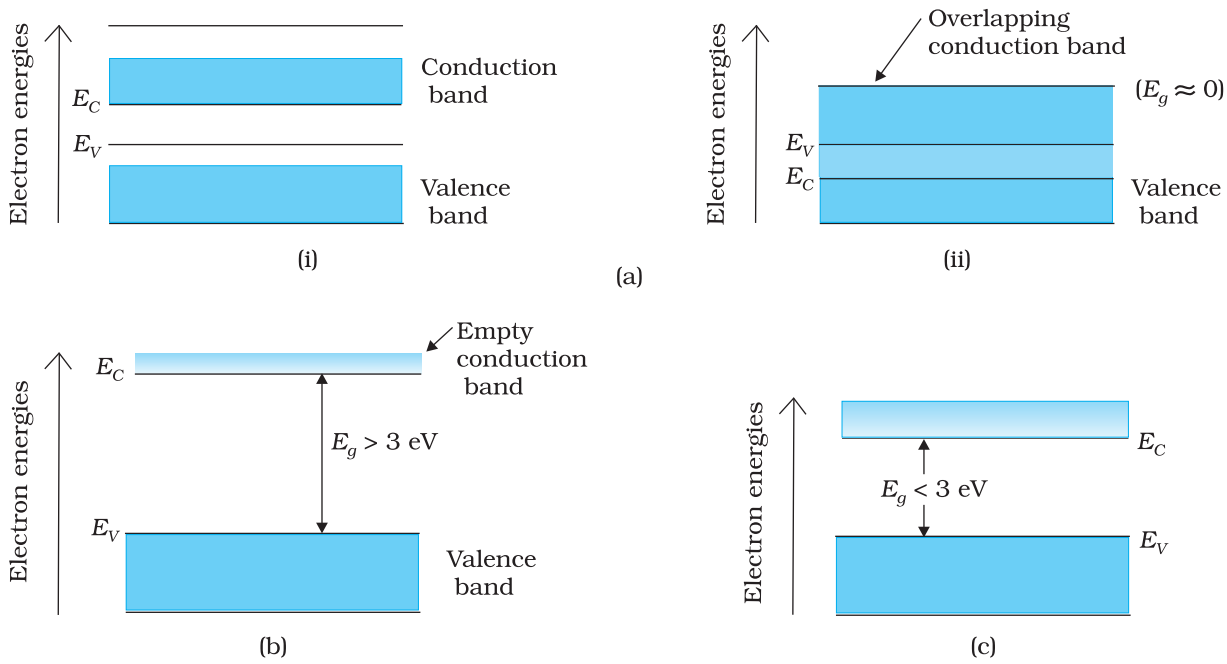


FIGURE 14.2 Difference between energy bands of (a) metals, (b) insulators and (c) semiconductors.

Case II: In this case, as shown in Fig. 14.2(b), a large band gap E_g exists ($E_g > 3$ eV). There are no electrons in the conduction band, and therefore no electrical conduction is possible. Note that the energy gap is so large that electrons cannot be excited from the valence band to the conduction band by thermal excitation. This is the case of *insulators*.

Case III: This situation is shown in Fig. 14.2(c). Here a finite but small band gap ($E_g < 3$ eV) exists. Because of the small band gap, at room temperature some electrons from valence band can acquire enough energy to cross the energy gap and enter the *conduction band*. These electrons (though small in numbers) can move in the conduction band. Hence, the resistance of *semiconductors* is not as high as that of the insulators.

In this section we have made a broad classification of metals, conductors and semiconductors. In the section which follows you will learn the conduction process in semiconductors.

14.3 Intrinsic Semiconductor

We shall take the most common case of Ge and Si whose lattice structure is shown in Fig. 14.3. These structures are called the diamond-like structures. Each atom is surrounded by four nearest neighbours. We know that Si and Ge have four valence electrons. In its crystalline structure, every Si or Ge atom tends *to share* one of its four valence electrons with each of its four nearest neighbour atoms, and also *to take share* of one electron from each such neighbour. These shared electron pairs are referred to as forming a *covalent bond* or simply a *valence bond*. The two shared electrons can be assumed to shuttle back-and-forth between the associated atoms holding them together strongly. Figure 14.4 schematically shows the 2-dimensional representation of Si or Ge structure shown in Fig. 14.3 which overemphasises the covalent bond. It shows an idealised picture in which no bonds are broken (all bonds are intact). Such a situation arises at low temperatures. As the temperature increases, more thermal energy becomes available to these electrons and some of these electrons may break–away (becoming *free* electrons contributing to conduction). The thermal energy effectively ionises only a few atoms in the crystalline lattice and creates a *vacancy* in the bond as shown in Fig. 14.5(a). The neighbourhood, from which the free electron (with charge $-q$) has come out leaves a vacancy with an effective charge ($+q$). This *vacancy* with the effective positive electronic charge is called a *hole*. The hole behaves as an *apparent free particle* with effective positive charge.

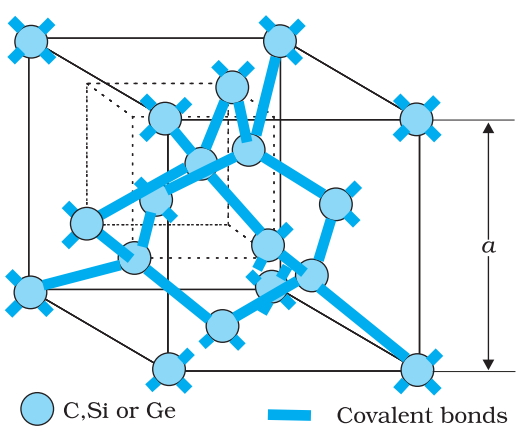


FIGURE 14.3 Three-dimensional diamond-like crystal structure for Carbon, Silicon or Germanium with respective lattice spacing a equal to 3.56, 5.43 and 5.66 Å.

In intrinsic semiconductors, the number of free electrons, n_e is equal to the number of holes, n_h. That is

$$n_e = n_h = n_i \qquad (14.1)$$

where n_i is called intrinsic carrier concentration.

Semiconductors posses the unique property in which, apart from electrons, the holes also move.

Suppose there is a hole at site 1 as shown in Fig. 14.5(a). The movement of holes can be visualised as shown in Fig. 14.5(b). An electron from the covalent bond at site 2 may jump to the vacant site 1 (hole). Thus, after such a jump, the hole is at site 2 and the site 1 has now an electron. Therefore, apparently, the hole has moved from site 1 to site 2. Note that the electron originally set free [Fig. 14.5(a)] is not involved in this process of hole motion. The free electron moves completely independently as conduction electron and gives rise to an electron current, I_e under an applied electric field. Remember that the motion of hole is only a convenient way of describing the actual motion of *bound* electrons, whenever there is an empty bond anywhere in the crystal. Under the action of an electric field, these holes move towards negative potential giving the hole current, I_h. The total current, I is thus the sum of the electron current I_e and the hole current I_h:

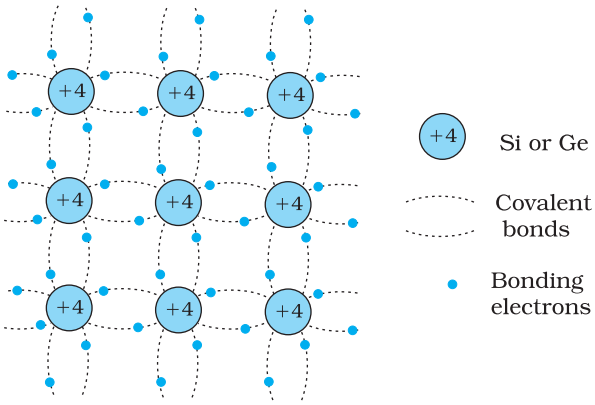


FIGURE 14.4 Schematic two-dimensional representation of Si or Ge structure showing covalent bonds at low temperature (all bonds intact). +4 symbol indicates inner cores of Si or Ge.

$$I = I_e + I_h \tag{14.2}$$

It may be noted that apart from the *process of generation* of conduction electrons and holes, a simultaneous *process of recombination* occurs in which the electrons *recombine* with the holes. At equilibrium, the rate of generation is equal to the rate of recombination of charge carriers. The recombination occurs due to an electron colliding with a hole.

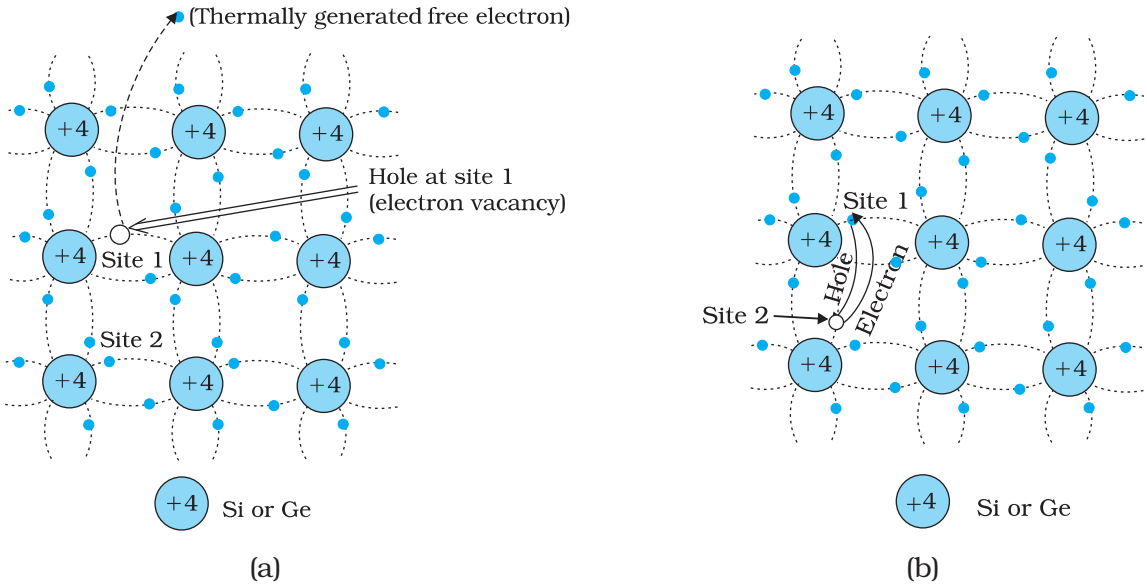


FIGURE 14.5 (a) Schematic model of generation of hole at site 1 and conduction electron due to thermal energy at moderate temperatures. (b) Simplified representation of possible thermal motion of a hole. The electron from the lower left hand covalent bond (site 2) goes to the earlier hole site1, leaving a hole at its site indicating an apparent movement of the hole from site 1 to site 2.

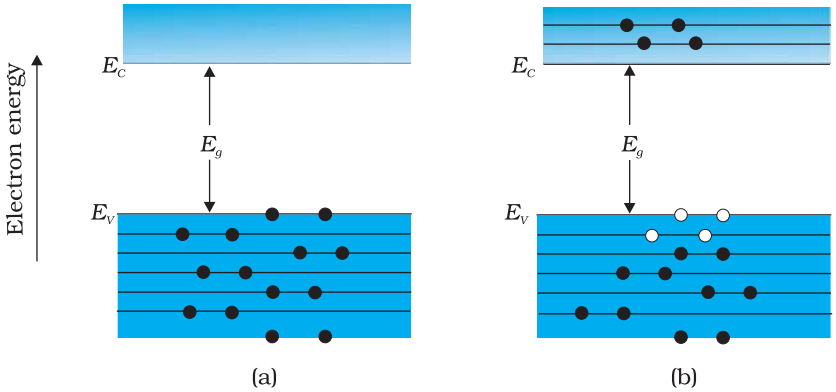


FIGURE 14.6 (a) An intrinsic semiconductor at $T = 0$ K behaves like insulator. (b) At $T > 0$ K, four thermally generated electron-hole pairs. The filled circles (•) represent electrons and empty circles (○) represent holes.

An intrinsic semiconductor will behave like an insulator at $T = 0$ K as shown in Fig. 14.6(a). It is the thermal energy at higher temperatures ($T > 0$K), which excites some electrons from the valence band to the conduction band. These thermally excited electrons at $T > 0$ K, partially occupy the conduction band. Therefore, the energy-band diagram of an intrinsic semiconductor will be as shown in Fig. 14.6(b). Here, some electrons are shown in the conduction band. These have come from the valence band leaving equal number of holes there.

EXAMPLE 14.1

Example 14.1 C, Si and Ge have same lattice structure. Why is C insulator while Si and Ge intrinsic semiconductors?

Solution The 4 bonding electrons of C, Si or Ge lie, respectively, in the second, third and fourth orbit. Hence, energy required to take out an electron from these atoms (i.e., ionisation energy E_g) will be least for Ge, followed by Si and highest for C. Hence, number of free electrons for conduction in Ge and Si are significant but negligibly small for C.

14.4 Extrinsic Semiconductor

The conductivity of an intrinsic semiconductor depends on its temperature, but at room temperature its conductivity is very low. As such, no important electronic devices can be developed using these semiconductors. Hence there is a necessity of improving their conductivity. This can be done by making use of impurities.

When a small amount, say, a few parts per million (ppm), of a suitable impurity is added to the pure semiconductor, the conductivity of the semiconductor is increased manifold. Such materials are known as *extrinsic semiconductors* or *impurity semiconductors*. The deliberate addition of a desirable impurity is called *doping* and the impurity atoms are called *dopants*. Such a material is also called a *doped semiconductor*. The dopant has to be such that it does not distort the original pure semiconductor lattice. It occupies only a very few of the original semiconductor atom sites in the crystal. A necessary condition to attain this is that the sizes of the dopant and the semiconductor atoms should be nearly the same.

There are two types of dopants used in doping the tetravalent Si or Ge:

(i) Pentavalent (valency 5); like Arsenic (As), Antimony (Sb), Phosphorous (P), etc.

(ii) Trivalent (valency 3); like Indium (In), Boron (B), Aluminium (Al), etc.

We shall now discuss how the doping changes the number of charge carriers (and hence the conductivity) of semiconductors. Si or Ge belongs to the fourth group in the Periodic table and, therefore, we choose the dopant element from nearby fifth or third group, expecting and taking care that the size of the dopant atom is nearly the same as that of Si or Ge. Interestingly, the pentavalent and trivalent dopants in Si or Ge give two entirely different types of semiconductors as discussed below.

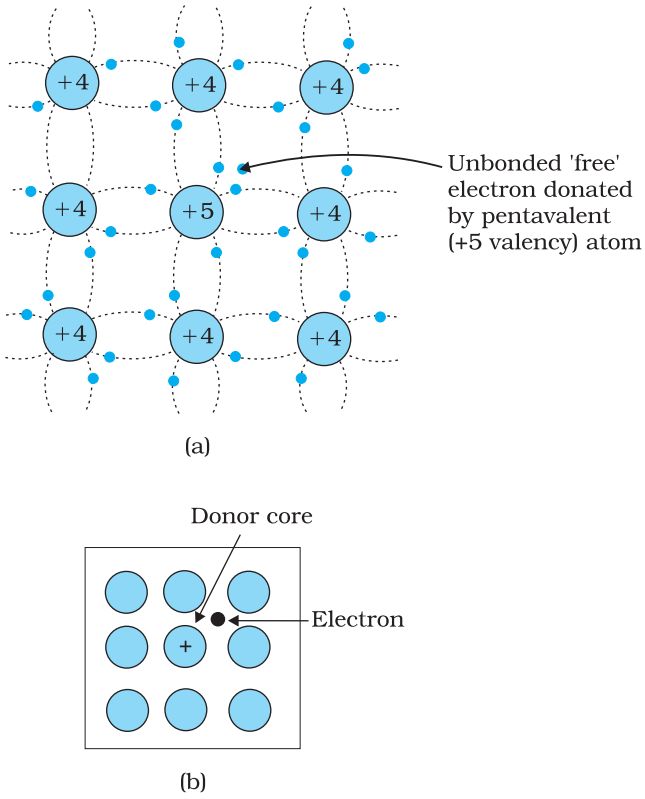


FIGURE 14.7 (a) Pentavalent donor atom (As, Sb, P, etc.) doped for tetravalent Si or Ge giving n-type semiconductor, and (b) Commonly used schematic representation of n-type material which shows only the fixed cores of the substituent donors with one additional effective positive charge and its associated extra electron.

(i) n-type semiconductor

Suppose we dope Si or Ge with a pentavalent element as shown in Fig. 14.7. When an atom of +5 valency element occupies the position of an atom in the crystal lattice of Si, four of its electrons bond with the four silicon neighbours while the fifth remains very weakly bound to its parent atom. This is because the four electrons participating in bonding are seen as part of the effective core of the atom by the fifth electron. As a result the ionisation energy required to set this electron free is very small and even at room temperature it will be free to move in the lattice of the semiconductor. For example, the energy required is ~ 0.01 eV for germanium, and 0.05 eV for silicon, to separate this electron from its atom. This is in contrast to the energy required to jump the forbidden band (about 0.72 eV for germanium and about 1.1 eV for silicon) at room temperature in the intrinsic semiconductor. Thus, the pentavalent dopant is donating one extra electron for conduction and hence is known as *donor* impurity. The number of electrons made available for conduction by dopant atoms depends strongly upon the doping level and is independent of any increase in ambient temperature. On the other hand, the number of free electrons (with an equal number of holes) generated by Si atoms, increases weakly with temperature.

In a doped semiconductor the total number of conduction electrons n_e is due to the electrons contributed by donors and those generated intrinsically, while the total number of holes n_h is only due to the holes from the intrinsic source. But the rate of recombination of holes would increase due to the increase in the number of electrons. As a result, the number of holes would get reduced further.

Thus, with proper level of doping the number of conduction electrons can be made much larger than the number of holes. Hence in an extrinsic

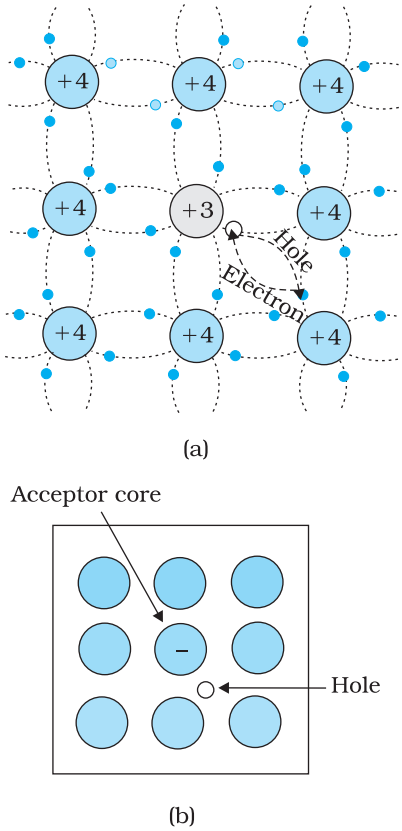


FIGURE 14.8 (a) Trivalent acceptor atom (In, Al, B etc.) doped in tetravalent Si or Ge lattice giving p-type semiconductor. (b) Commonly used schematic representation of p-type material which shows only the fixed core of the substituent acceptor with one effective additional negative charge and its associated hole.

semiconductor doped with pentavalent impurity, electrons become the *majority carriers* and holes the *minority carriers*. These semiconductors are, therefore, known as n-*type semiconductors*. For n-type semiconductors, we have,

$$n_e >> n_h \qquad (14.3)$$

(ii) p-type semiconductor

This is obtained when Si or Ge is doped with a trivalent impurity like Al, B, In, etc. The dopant has one valence electron less than Si or Ge and, therefore, this atom can form covalent bonds with neighbouring three Si atoms but does not have any electron to offer to the fourth Si atom. So the bond between the fourth neighbour and the trivalent atom has a vacancy or hole as shown in Fig. 14.8. Since the neighbouring Si atom in the lattice wants an electron in place of a hole, an electron in the outer orbit of an atom in the neighbourhood may jump to fill this vacancy, leaving a vacancy or hole at its own site. Thus the *hole* is available for conduction. Note that the trivalent foreign atom becomes effectively negatively charged when it shares fourth electron with neighbouring Si atom. Therefore, the dopant atom of p-type material can be treated as *core of one negative charge* along with its associated hole as shown in Fig. 14.8(b). It is obvious that one *acceptor* atom gives one *hole*. These holes are in addition to the intrinsically generated holes while the source of conduction electrons is only intrinsic generation. Thus, for such a material, the holes are the majority carriers and electrons are minority carriers. Therefore, extrinsic semiconductors doped with trivalent impurity are called p-*type semiconductors*. For p-type semiconductors, the recombination process will reduce the number (n_i)of intrinsically generated electrons to n_e. We have, for p-type semiconductors

$$n_h >> n_e \qquad (14.4)$$

Note that *the crystal maintains an overall charge neutrality as the charge of additional charge carriers is just equal and opposite to that of the ionised cores in the lattice.*

In extrinsic semiconductors, because of the abundance of majority current carriers, the minority carriers produced thermally have more chance of meeting majority carriers and thus getting destroyed. Hence, the dopant, by adding a large number of current carriers of one type, which become the majority carriers, indirectly helps to reduce the intrinsic concentration of minority carriers.

The semiconductor's energy band structure is affected by doping. In the case of extrinsic semiconductors, additional energy states due to donor impurities (E_D) and acceptor impurities (E_A) also exist. In the energy band diagram of n-type Si semiconductor, the donor energy level E_D is slightly below the bottom E_C of the conduction band and electrons from this level move into the conduction band with very small supply of energy. At room temperature, most of the donor atoms get ionised but very few ($\sim 10^{-12}$) atoms of Si get ionised. So the conduction band will have most electrons coming from the donor impurities, as shown in Fig. 14.9(a). Similarly,

for p-type semiconductor, the acceptor energy level E_A is slightly above the top E_V of the valence band as shown in Fig. 14.9(b). With very small supply of energy an electron from the valence band can jump to the level E_A and ionise the acceptor negatively. (Alternately, we can also say that with very small supply of energy the hole from level E_A sinks down into the valence band. Electrons rise up and holes fall down when they gain external energy.) At room temperature, most of the acceptor atoms get ionised leaving holes in the valence band. Thus at room temperature the density of holes in the valence band is predominantly due to impurity in the extrinsic semiconductor. The electron and hole concentration in a semiconductor *in thermal equilibrium* is given by

$$n_e n_h = n_i^2 \qquad (14.5)$$

Though the above description is grossly approximate and hypothetical, it helps in understanding the difference between metals, insulators and semiconductors (extrinsic and intrinsic) in a simple manner. The difference in the resistivity of C, Si and Ge depends upon the energy gap between their conduction and valence bands. For C (diamond), Si and Ge, the energy gaps are 5.4 eV, 1.1 eV and 0.7 eV, respectively. Sn also is a group IV element but it is a metal because the energy gap in its case is 0 eV.

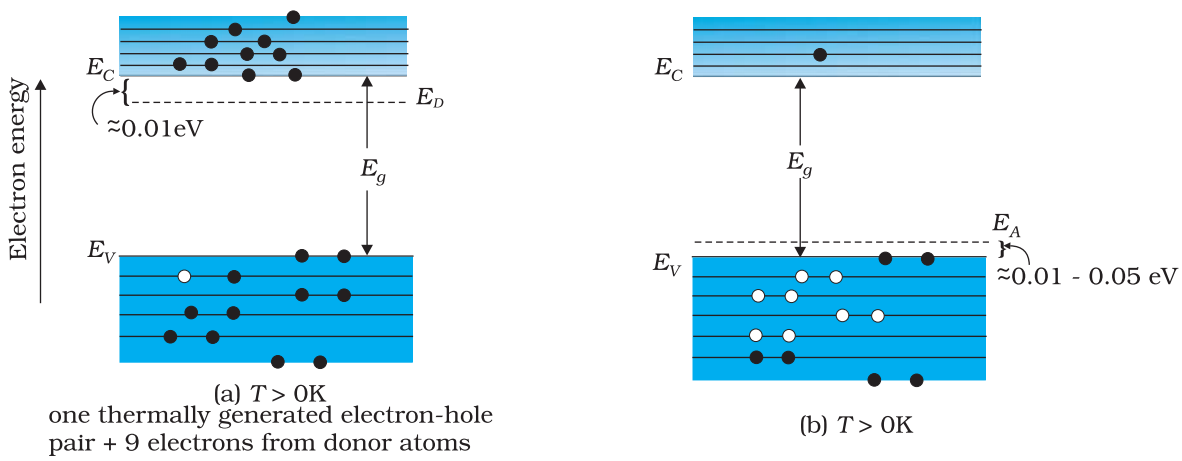


FIGURE 14.9 Energy bands of (a) n-type semiconductor at $T > 0$K, (b) p-type semiconductor at $T > 0$K.

Example 14.2 Suppose a pure Si crystal has 5×10^{28} atoms m^{-3}. It is doped by 1 ppm concentration of pentavalent As. Calculate the number of electrons and holes. Given that $n_i = 1.5 \times 10^{16}$ m^{-3}.

Solution Note that thermally generated electrons ($n_i \sim 10^{16}$ m^{-3}) are negligibly small as compared to those produced by doping.
Therefore, $n_e \approx N_D$.
Since $n_e n_h = n_i^2$, The number of holes
$n_h = (2.25 \times 10^{32})/(5 \times 10^{22})$

$\sim 4.5 \times 10^9$ m^{-3}

EXAMPLE 14.2

14.5 p-n Junction

A p-n junction is the basic building block of many semiconductor devices like diodes, transistor, etc. A clear understanding of the junction behaviour is important to analyse the working of other semiconductor devices. We will now try to understand how a junction is formed and how the junction behaves under the influence of external applied voltage (also called *bias*).

Formation and working of p-n junction diode
http://hyperphysics.phy-astr.gsu.edu/hbase/solids/pnjun.html

14.5.1 p-n junction formation

Consider a thin p-type silicon (p-Si) semiconductor wafer. By adding precisely a small quantity of pentavelent impurity, part of the p-Si wafer can be converted into n-Si. There are several processes by which a semiconductor can be formed. The wafer now contains p-region and n-region and a metallurgical junction between p-, and n- region.

Two important processes occur during the formation of a p-n junction: *diffusion* and *drift*. We know that in an n-type semiconductor, the concentration of electrons (number of electrons per unit volume) is more compared to the concentration of holes. Similarly, in a p-type semiconductor, the concentration of holes is more than the concentration of electrons. During the formation of p-n junction, and due to the concentration gradient across p-, and n- sides, holes diffuse from p-side to n-side (p → n) and electrons diffuse from n-side to p-side (n → p). This motion of charge carries gives rise to diffusion current across the junction.

When an electron diffuses from n → p, it leaves behind an ionised donor on n-side. This ionised donor (positive charge) is immobile as it is bonded to the surrounding atoms. As the electrons continue to diffuse from n → p, a layer of positive charge (or positive space-charge region) on n-side of the junction is developed.

Similarly, when a hole diffuses from p → n due to the concentration gradient, it leaves behind an ionised acceptor (negative charge) which is immobile. As the holes continue to diffuse, a layer of negative charge (or negative space-charge region) on the p-side of the junction is developed. This space-charge region on either side of the junction together is known as *depletion region* as the electrons and holes taking part in the initial movement across the junction *depleted* the region of its free charges (Fig. 14.10). The thickness of depletion region is of the order of one-tenth of a micrometre. Due to the positive space-charge region on n-side of the junction and negative space charge region on p-side of the junction, an electric field directed from positive charge towards negative charge develops. Due to this field, an electron on p-side of the junction moves to n-side and a hole on n-side of the junction moves to p-side. The motion of charge carriers due to the electric field is called drift. Thus a drift current, which is opposite in direction to the diffusion current (Fig. 14.10) starts.

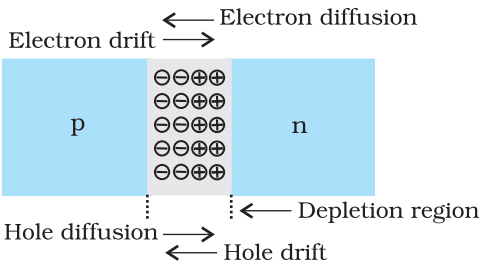


FIGURE 14.10 p-n junction formation process.

Initially, diffusion current is large and drift current is small. As the diffusion process continues, the space-charge regions on either side of the junction extend, thus increasing the electric field strength and hence drift current. This process continues until the diffusion current equals the drift current. Thus a p-n junction is formed. In a p-n junction under equilibrium there is *no net* current.

The loss of electrons from the n-region and the gain of electron by the p-region causes a difference of potential across the junction of the two regions. The polarity of this potential is such as to oppose further flow of carriers so that a condition of equilibrium exists. Figure 14.11 shows the p-n junction at equilibrium and the potential across the junction. The n-material has lost electrons, and p material has acquired electrons. The n material is thus positive relative to the p material. Since this potential tends to prevent the movement of electron from the n region into the p region, it is often called a *barrier potential*.

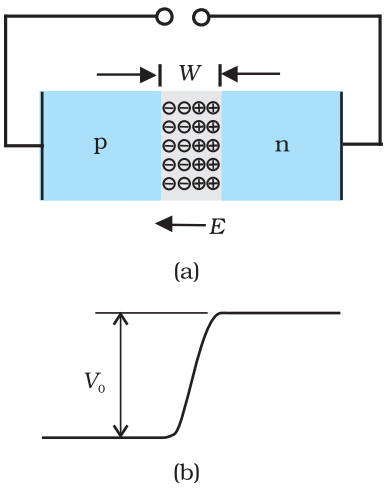


FIGURE 14.11 (a) Diode under equilibrium (V = 0), (b) Barrier potential under no bias.

Example 14.3 Can we take one slab of p-type semiconductor and physically join it to another n-type semiconductor to get p-n junction?

Solution No! Any slab, howsoever flat, will have roughness much larger than the inter-atomic crystal spacing (~2 to 3 Å) and hence *continuous contact* at the atomic level will not be possible. The junction will behave as a *discontinuity* for the flowing charge carriers.

EXAMPLE 14.3

14.6 SEMICONDUCTOR DIODE

A semiconductor diode [Fig. 14.12(a)] is basically a p-n junction with metallic contacts provided at the ends for the application of an external voltage. It is a two terminal device. A p-n junction diode is symbolically represented as shown in Fig. 14.12(b).

The direction of arrow indicates the conventional direction of current (when the diode is under forward bias). The equilibrium barrier potential can be altered by applying an external voltage *V* across the diode. The situation of p-n junction diode under equilibrium (without bias) is shown in Fig. 14.11(a) and (b).

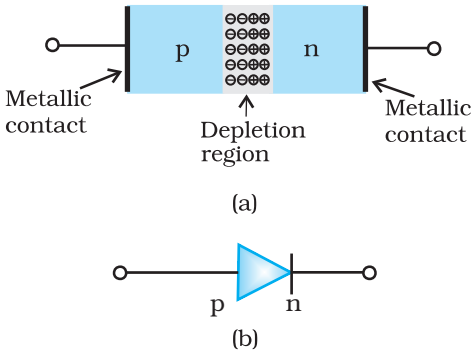


FIGURE 14.12 (a) Semiconductor diode, (b) Symbol for p-n junction diode.

14.6.1 p-n junction diode under forward bias

When an external voltage *V* is applied across a semiconductor diode such that p-side is connected to the positive terminal of the battery and n-side to the negative terminal [Fig. 14.13(a)], it is said to be *forward biased*.

The applied voltage mostly drops across the depletion region and the voltage drop across the p-side and n-side of the junction is negligible. (This is because the resistance of the depletion region – a region where there are no charges – is very high compared to the resistance of n-side and p-side.) The direction of the applied voltage (*V*) is opposite to the

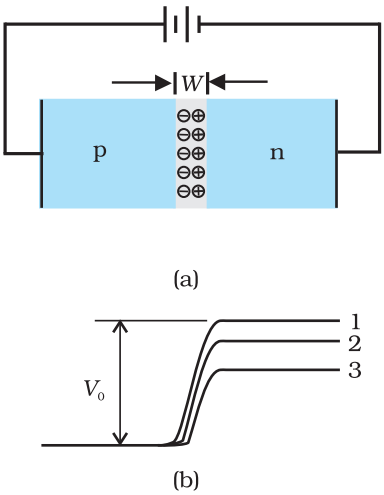


FIGURE 14.13 (a) p-n junction diode under forward bias, (b) Barrier potential (1) without battery, (2) Low battery voltage, and (3) High voltage battery.

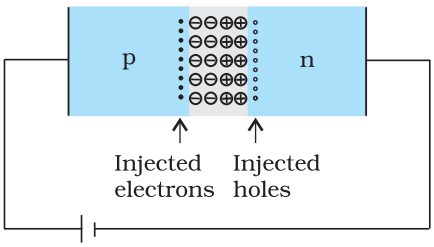


FIGURE 14.14 Forward bias minority carrier injection.

built-in potential V_0. As a result, the depletion layer width decreases and the barrier height is reduced [Fig. 14.13(b)]. The effective barrier height under forward bias is $(V_0 - V)$.

If the applied voltage is small, the barrier potential will be reduced only slightly below the equilibrium value, and only a small number of carriers in the material—those that happen to be in the uppermost energy levels—will possess enough energy to cross the junction. So the current will be small. If we increase the applied voltage significantly, the barrier height will be reduced and more number of carriers will have the required energy. Thus the current increases.

Due to the applied voltage, electrons from n-side cross the depletion region and reach p-side (where they are minority carries). Similarly, holes from p-side cross the junction and reach the n-side (where they are minority carries). This process under forward bias is known as minority carrier injection. At the junction boundary, on each side, the minority carrier concentration increases significantly compared to the locations far from the junction.

Due to this concentration gradient, the injected electrons on p-side diffuse from the junction edge of p-side to the other end of p-side. Likewise, the injected holes on n-side diffuse from the junction edge of n-side to the other end of n-side (Fig. 14.14). This motion of charged carriers on either side gives rise to current. The total diode forward current is sum of hole diffusion current and conventional current due to electron diffusion. The magnitude of this current is usually in mA.

14.6.2 p-n junction diode under reverse bias

When an external voltage (V) is applied across the diode such that n-side is positive and p-side is negative, it is said to be *reverse biased* [Fig.14.15(a)]. The applied voltage mostly drops across the depletion region. The direction of applied voltage is same as the direction of barrier potential. As a result, the barrier height increases and the depletion region widens due to the change in the electric field. The effective barrier height under reverse bias is $(V_0 + V)$, [Fig. 14.15(b)]. This suppresses the flow of electrons from n $\rightarrow$ p and holes from p $\rightarrow$ n. Thus, diffusion current, decreases enormously compared to the diode under forward bias.

The electric field direction of the junction is such that if electrons on p-side or holes on n-side in their random motion come close to the junction, they will be swept to its majority zone. This drift of carriers gives rise to current. The drift current is of the order of a few μA. This is quite low because it is due to the motion of carriers from their minority side to their majority side across the junction. The drift current is also there under forward bias but it is negligible (μA) when compared with current due to injected carriers which is usually in mA.

The diode reverse current is not very much dependent on the applied voltage. Even a small voltage is sufficient to sweep the minority carriers from one side of the junction to the other side of the junction. The current

is not limited by the magnitude of the applied voltage but is limited due to the concentration of the minority carrier on either side of the junction.

The current under reverse bias is essentially voltage independent upto a critical reverse bias voltage, known as breakdown voltage (V_{br}). When $V = V_{br}$, the diode reverse current increases sharply. Even a slight increase in the bias voltage causes large change in the current. If the reverse current is not limited by an external circuit below the rated value (specified by the manufacturer) the p-n junction will get destroyed. Once it exceeds the rated value, the diode gets destroyed due to overheating. This can happen even for the diode under forward bias, if the forward current exceeds the rated value.

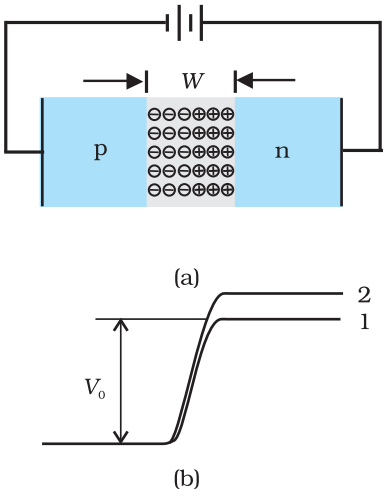


FIGURE 14.15 (a) Diode under reverse bias, (b) Barrier potential under reverse bias.

The circuit arrangement for studying the *V-I* characteristics of a diode, (i.e., the variation of current as a function of applied voltage) are shown in Fig. 14.16(a) and (b). The battery is connected to the diode through a potentiometer (or reheostat) so that the applied voltage to the diode can be changed. For different values of voltages, the value of the current is noted. A graph between *V* and *I* is obtained as in Fig. 14.16(c). Note that in forward bias measurement, we use a milliammeter since the expected current is large (as explained in the earlier section) while a micrometer is used in reverse bias to measure the current. You can see in Fig. 14.16(c) that in forward

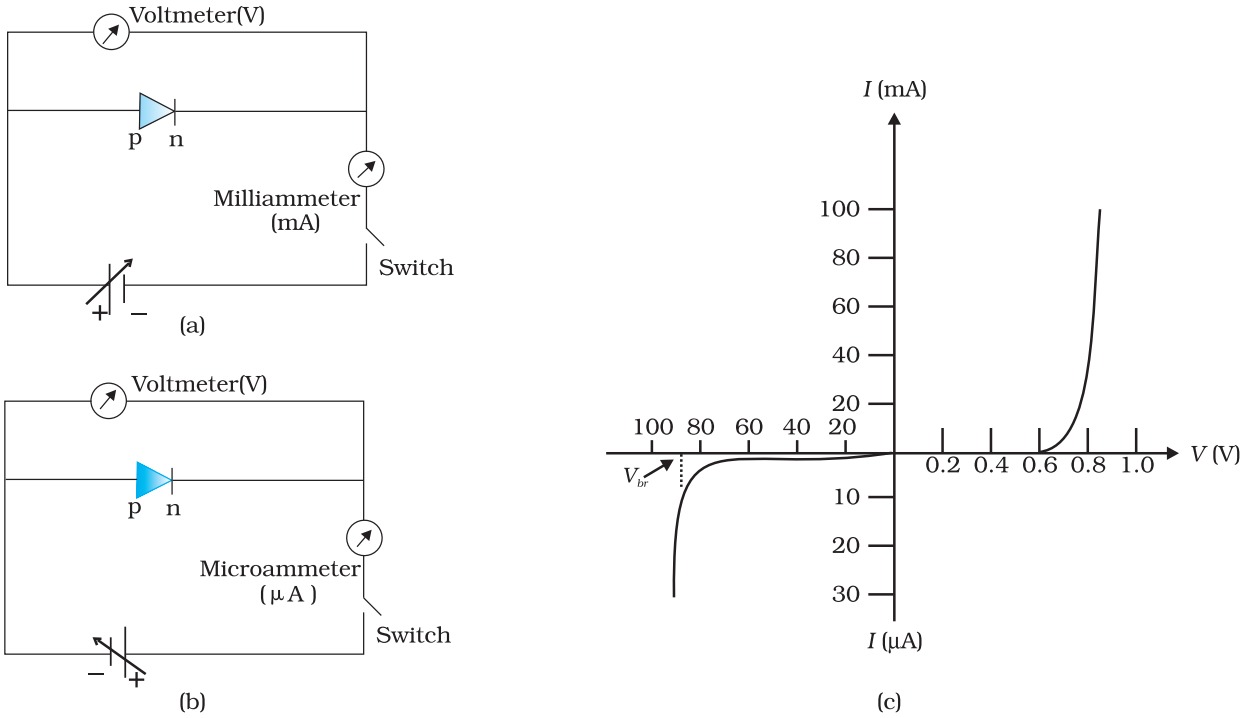


FIGURE 14.16 Experimental circuit arrangement for studying *V-I* characteristics of a p-n junction diode (a) in forward bias, (b) in reverse bias. (c) Typical *V-I* characteristics of a silicon diode.

bias, the current first increases very slowly, almost negligibly, till the voltage across the diode crosses a certain value. After the characteristic voltage, the diode current increases significantly (exponentially), even for a very small increase in the diode bias voltage. This voltage is called the *threshold voltage* or cut-in voltage (~0.2V for germanium diode and ~0.7 V for silicon diode).

For the diode in reverse bias, the current is very small (~μA) and almost remains constant with change in bias. It is called *reverse saturation current*. However, for special cases, at very high reverse bias (break down voltage), the current suddenly increases. This special action of the diode is discussed later in Section 14.8. The general purpose diode are not used beyond the reverse saturation current region.

The above discussion shows that the p-n junction diode primerly allows the flow of current only in one direction (forward bias). The forward bias resistance is low as compared to the reverse bias resistance. This property is used for rectification of ac voltages as discussed in the next section. For diodes, we define a quantity called *dynamic resistance* as the ratio of small change in voltage ΔV to a small change in current ΔI:

$$r_d = \frac{\Delta V}{\Delta I} \tag{14.6}$$

Example 14.4 The *V-I* characteristic of a silicon diode is shown in the Fig. 14.17. Calculate the resistance of the diode at (a) I_D = 15 mA and (b) V_D = –10 V.

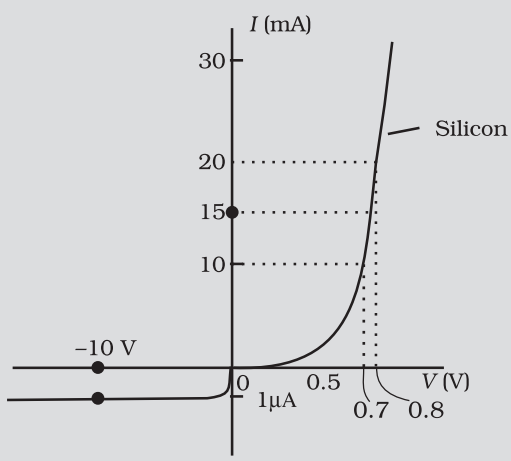


FIGURE 14.17

Solution Considering the diode characteristics as a straight line between I = 10 mA to I = 20 mA passing through the origin, we can calculate the resistance using Ohm's law.

(a) From the curve, at I = 20 mA, V = 0.8 V; I = 10 mA, V = 0.7 V
$r_{fb} = \Delta V/\Delta I$ = 0.1V/10 mA = 10 Ω

(b) From the curve at V = –10 V, I = –1 μA,
Therefore,
r_{rb} = 10 V/1μA= 1.0×10^7 Ω

EXAMPLE 14.4

14.7 Application of Junction Diode as a Rectifier

From the *V-I* characteristic of a junction diode we see that it allows current to pass only when it is forward biased. So if an alternating voltage is applied across a diode the current flows only in that part of the cycle when the diode is forward biased. This property is used *to rectify* alternating voltages and the circuit used for this purpose is called a *rectifier*.

If an alternating voltage is applied across a diode in series with a load, a pulsating voltage will appear across the load only during the half cycles of the ac input during which the diode is forward biased. Such rectifier circuit, as shown in Fig. 14.18, is called a *half-wave rectifier*. The secondary of a transformer supplies the desired ac voltage across terminals A and B. When the voltage at A is positive, the diode is forward biased and it conducts. When A is negative, the diode is reverse-biased and it does not conduct. The reverse saturation current of a diode is negligible and can be considered equal to zero for practical purposes. (The reverse breakdown voltage of the diode must be sufficiently higher than the peak ac voltage at the secondary of the transformer to protect the diode from reverse breakdown.)

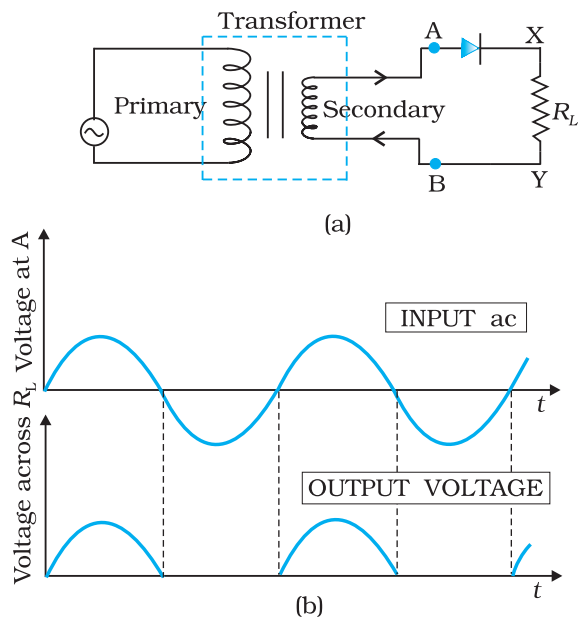


FIGURE 14.18 (a) Half-wave rectifier circuit, (b) Input ac voltage and output voltage waveforms from the rectifier circuit.

Therefore, in the positive *half-cycle* of ac there is a current through the load resistor R_L and we get an output voltage, as shown in Fig. 14.18(b), whereas there is no current in the negative half-cycle. In the next positive half-cycle, again we get the output voltage. Thus, the output voltage, though still varying, is restricted to *only one direction* and is said to be *rectified*. Since the rectified output of this circuit is only for half of the input ac wave it is called as *half-wave rectifier*.

The circuit using two diodes, shown in Fig. 14.19(a), gives output rectified voltage corresponding to both the positive as well as negative half of the ac cycle. Hence, it is known as *full-wave rectifier*. Here the p-side of the two diodes are connected to the ends of the secondary of the transformer. The n-side of the diodes are connected together and the output is taken between this common point of diodes and the midpoint of the secondary of the transformer. So for a full-wave rectifier the secondary of the transformer is provided with a centre tapping and so it is called *centre-tap transformer*. As can be seen from Fig.14.19(c) the voltage rectified by each diode is only half the total secondary voltage. Each diode rectifies only for half the cycle, but the two do so for alternate cycles. Thus, the output between their common terminals and the centre-tap of the transformer becomes a full-wave rectifier output. (Note that there is another circuit of full wave rectifier which does not need a centre-tap transformer but needs four diodes.) Suppose the input voltage to A

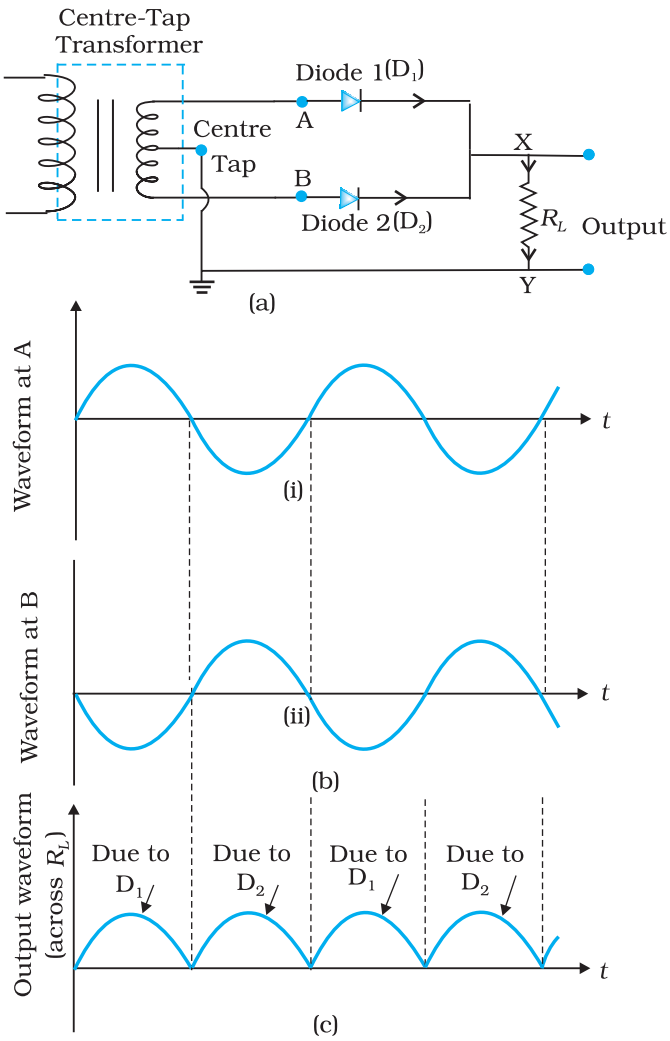


FIGURE 14.19 (a) A Full-wave rectifier circuit; (b) Input wave forms given to the diode D_1 at A and to the diode D_2 at B; (c) Output waveform across the load R_L connected in the full-wave rectifier circuit.

with respect to the centre tap at any instant is positive. It is clear that, at that instant, voltage at B being out of phase will be negative as shown in Fig.14.19(b). So, diode D_1 gets forward biased and conducts (while D_2 being reverse biased is not conducting). Hence, during this positive half cycle we get an output current (and a output voltage across the load resistor R_L) as shown in Fig.14.19(c). In the course of the ac cycle when the voltage at A becomes negative with respect to centre tap, the voltage at B would be positive. In this part of the cycle diode D_1 would not conduct but diode D_2 would, giving an output current and output voltage (across R_L) during the negative half cycle of the input ac. Thus, we get output voltage during both the positive as well as the negative half of the cycle. Obviously, this is a more efficient circuit for getting rectified voltage or current than the half-wave rectifier.

The rectified voltage is in the form of pulses of the shape of half sinusoids. Though it is unidirectional it does not have a steady value. To get steady dc output from the pulsating voltage normally a capacitor is connected across the output terminals (parallel to the load R_L). One can also use an inductor in series with R_L for the same purpose. Since these additional circuits appear to *filter* out the *ac ripple* and give a *pure dc* voltage, so they are called filters.

Now we shall discuss the role of capacitor in filtering. When the voltage across the capacitor is rising, it gets charged. If there is no external load, it remains charged to the peak voltage of the rectified output. When there is a load, it gets discharged through the load and the voltage across it begins to fall. In the next half-cycle of rectified output it again gets charged to the peak value (Fig. 14.20). The rate of fall of the voltage across the capacitor depends inversely upon the product of capacitance C and the effective resistance R_L used in the circuit and is called the *time constant*. To make the time constant large value of C should be large. So capacitor input filters use large capacitors. The *output voltage* obtained by using capacitor input filter is nearer to the *peak voltage* of the rectified voltage. This type of filter is most widely used in power supplies.

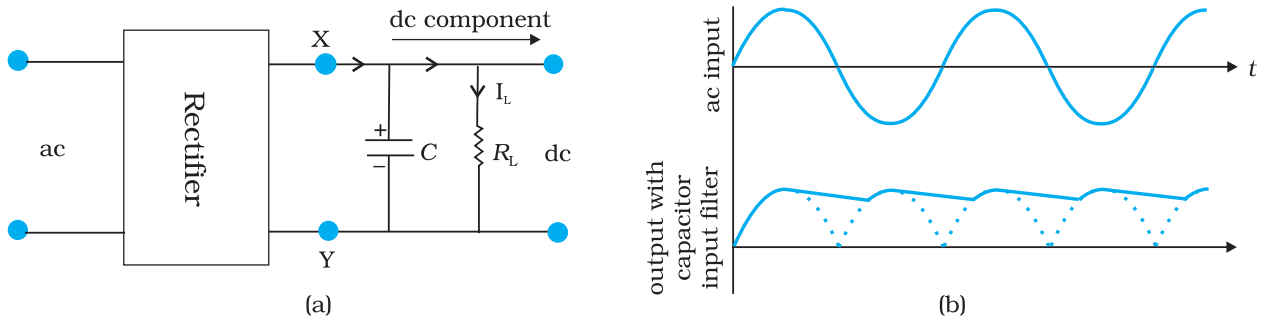


FIGURE 14.20 (a) A full-wave rectifier with capacitor filter, (b) Input and output voltage of rectifier in (a).

14.8 Special Purpose p-n Junction Diodes

In the section, we shall discuss some devices which are basically junction diodes but are developed for different applications.

14.8.1 Zener diode

It is a special purpose semiconductor diode, named after its inventor C. Zener. It is designed to operate under reverse bias in the breakdown region and used as a voltage regulator. The symbol for Zener diode is shown in Fig. 14.21(a).

Zener diode is fabricated by heavily doping both p-, and n- sides of the junction. Due to this, depletion region formed is very thin ($<10^{-6}$ m) and the electric field of the junction is extremely high ($\sim 5\times10^{6}$ V/m) even for a small reverse bias voltage of about 5V. The I-V characteristics of a Zener diode is shown in Fig. 14.21(b). It is seen that when the applied reverse bias voltage(V) reaches the breakdown voltage (V_z) of the Zener diode, there is a large change in the current. Note that after the breakdown voltage V_z, a large change in the current can be produced by almost insignificant change in the reverse bias voltage. In other words, Zener voltage remains constant, even though current through the Zener diode varies over a wide range. This property of the Zener diode is used for regulating supply voltages so that they are constant.

Let us understand how reverse current suddenly increases at the breakdown voltage. We know that reverse current is due to the flow of electrons (minority carriers) from p $\rightarrow$ n and holes from n $\rightarrow$ p. As the reverse bias voltage is increased, the electric field at the junction becomes significant. When the reverse bias voltage V = V_z, then the electric field strength is high enough to pull valence electrons from the host atoms on the p-side which are accelerated to n-side. These electrons account for high current observed at the breakdown. The emission of electrons from the host atoms due to the high electric field is known as internal field emission or field ionisation. The electric field required for field ionisation is of the order of 10^6 V/m.

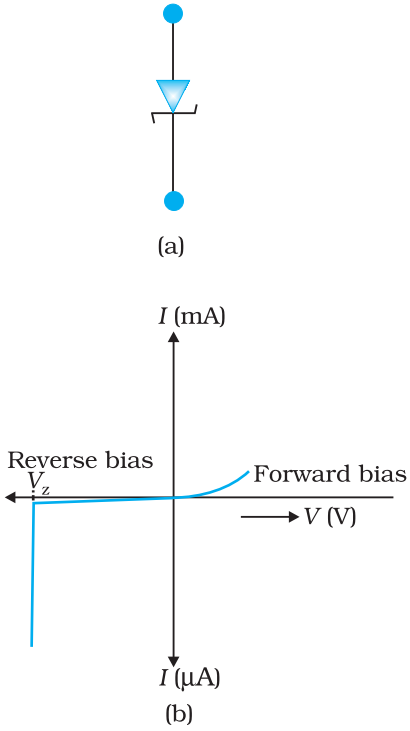


FIGURE 14.21 Zener diode, (a) symbol, (b) *I-V* characteristics.

Zener diode as a voltage regulator

We know that when the ac input voltage of a rectifier fluctuates, its rectified output also fluctuates. To get a constant dc voltage from the dc unregulated output of a rectifier, we use a Zener diode. The circuit diagram of a voltage regulator using a Zener diode is shown in Fig. 14.22.

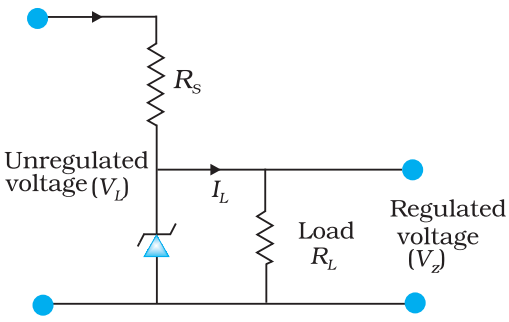


FIGURE 14.22 Zener diode as DC voltage regulator

The unregulated dc voltage (filtered output of a rectifier) is connected to the Zener diode through a series resistance R_s such that the Zener diode is reverse biased. If the input voltage increases, the current through R_s and Zener diode also increases. This increases the voltage drop across R_s without any change in the voltage across the Zener diode. This is because in the breakdown region, Zener voltage remains constant even though the current through the Zener diode changes. Similarly, if the input voltage decreases, the current through R_s and Zener diode also decreases. The voltage drop across R_s decreases without any change in the voltage across the Zener diode. Thus any increase/decrease in the input voltage results in, increase/decrease of the voltage drop across R_s without any change in voltage across the Zener diode. Thus the Zener diode acts as a voltage regulator. We have to select the Zener diode according to the required output voltage and accordingly the series resistance R_s.

EXAMPLE 14.5

Example 14.5 In a Zener regulated power supply a Zener diode with V_Z = 6.0 V is used for regulation. The load current is to be 4.0 mA and the unregulated input is 10.0 V. What should be the value of series resistor R_S?

Solution

The value of R_S should be such that the current through the Zener diode is much larger than the load current. This is to have good load regulation. Choose Zener current as five times the load current, i.e., I_Z = 20 mA. The total current through R_S is, therefore, 24 mA. The voltage drop across R_S is 10.0 – 6.0 = 4.0 V. This gives $R_S = 4.0\text{V}/(24 \times 10^{-3})\ \text{A} = 167\ \Omega$. The nearest value of carbon resistor is 150 Ω. So, a series resistor of 150 Ω is appropriate. Note that slight variation in the value of the resistor does not matter, what is important is that the current I_Z should be sufficiently larger than I_L.

14.8.2 Optoelectronic junction devices

We have seen so far, how a semiconductor diode behaves under applied electrical inputs. In this section, we learn about semiconductor diodes in which carriers are generated by photons (photo-excitation). All these devices are called *optoelectronic devices*. We shall study the functioning of the following optoelectronic devices:

(i) *Photodiodes* used for detecting optical signal (photodetectors).

(ii) *Light emitting diodes* (LED) which convert electrical energy into light.

(iii) *Photovoltaic devices* which convert optical radiation into electricity (*solar cells*).

(i) *Photodiode*

A Photodiode is again a special purpose p-n junction diode fabricated with a transparent window to allow light to fall on the diode. It is operated under reverse bias. When the photodiode is illuminated with light (photons) with energy ($h\nu$) greater than the energy gap (E_g) of the semiconductor, then electron-hole pairs are generated due to the absorption of photons. The diode is fabricated such that the generation of *e-h* pairs takes place in or near the depletion region of the diode. Due to electric field of the junction, electrons and holes are separated before they recombine. The direction of the electric field is such that electrons reach n-side and holes reach p-side. Electrons are collected on n-side and holes are collected on p-side giving rise to an emf. When an external load is connected, current flows. The magnitude of the photocurrent depends on the intensity of incident light (photocurrent is proportional to incident light intensity).

It is easier to observe the change in the current with change in the light intensity, if a reverse bias is applied. Thus photodiode can be used as a photodetector to detect optical signals. The circuit diagram used for the measurement of *I-V* characteristics of a photodiode is shown in Fig. 14.23(a) and a typical *I-V* characteristics in Fig. 14.23(b).

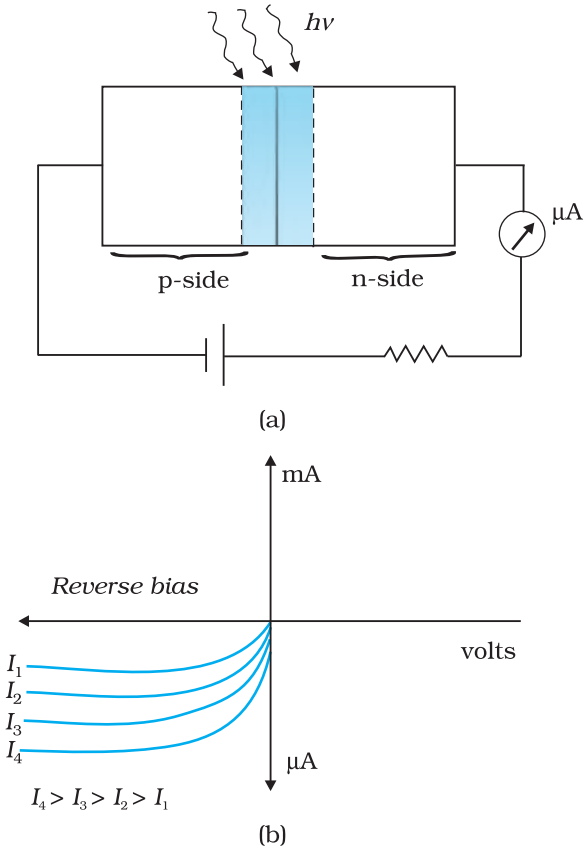


FIGURE 14.23 (a) An illuminated photodiode under reverse bias , (b) *I-V* characteristics of a photodiode for different illumination intensity $I_4 > I_3 > I_2 > I_1$.

Example 14.6 The current in the forward bias is known to be more (~mA) than the current in the reverse bias (~μA). What is the reason then to operate the photodiodes in reverse bias?

Solution Consider the case of an n-type semiconductor. Obviously, the majority carrier density (n) is considerably larger than the minority hole density p (i.e., $n >> p$). On illumination, let the excess electrons and holes generated be Δn and Δp, respectively:

$$n' = n + \Delta n$$

$$p' = p + \Delta p$$

Here n' and p' are the electron and hole concentrations* at any particular illumination and n and p are carriers concentration when there is no illumination. Remember $\Delta n = \Delta p$ and $n >> p$. Hence, the

EXAMPLE 14.6

* Note that, to create an e-h pair, we spend some energy (photoexcitation, thermal excitation, etc.). Therefore when an electron and hole recombine the energy is released in the form of light (radiative recombination) or heat (non-radiative recombination). It depends on semiconductor and the method of fabrication of the p-n junction. For the fabrication of LEDs, semiconductors like GaAs, GaAs-GaP are used in which radiative recombination dominates.

Example 14.6

fractional change in the majority carriers (i.e., $\Delta n/n$) would be much less than that in the minority carriers (i.e., $\Delta p/p$). In general, we can state that the fractional change due to the photo-effects on the *minority carrier dominated reverse bias current* is more easily measurable than the fractional change in the forward bias current. Hence, photodiodes are preferably used in the reverse bias condition for measuring light intensity.

(ii) *Light emitting diode*

It is a heavily doped p-n junction which under forward bias emits spontaneous radiation. The diode is encapsulated with a transparent cover so that emitted light can come out.

When the diode is forward biased, electrons are sent from n $\rightarrow$ p (where they are minority carriers) and holes are sent from p $\rightarrow$ n (where they are minority carriers). At the junction boundary the concentration of minority carriers increases compared to the equilibrium concentration (i.e., when there is no bias). Thus at the junction boundary on either side of the junction, excess minority carriers are there which recombine with majority carriers near the junction. On recombination, the energy is released in the form of photons. Photons with energy equal to or slightly less than the band gap are emitted. When the forward current of the diode is small, the intensity of light emitted is small. As the forward current increases, intensity of light increases and reaches a maximum. Further increase in the forward current results in decrease of light intensity. LEDs are biased such that the light emitting efficiency is maximum.

The *V-I* characteristics of a LED is similar to that of a Si junction diode. But the threshold voltages are much higher and slightly different for each colour. The reverse breakdown voltages of LEDs are very low, typically around 5V. So care should be taken that high reverse voltages do not appear across them.

LEDs that can emit red, yellow, orange, green and blue light are commercially available. The semiconductor used for fabrication of visible LEDs must at least have a band gap of 1.8 eV (spectral range of visible light is from about 0.4 μm to 0.7 μm, i.e., from about 3 eV to 1.8 eV). The compound semiconductor Gallium Arsenide – Phosphide ($GaAs_{1-x}P_x$) is used for making LEDs of different colours. $GaAs_{0.6}\,P_{0.4}$ (E_g ~ 1.9 eV) is used for red LED. GaAs (E_g ~ 1.4 eV) is used for making infrared LED. These LEDs find extensive use in remote controls, burglar alarm systems, optical communication, etc. Extensive research is being done for developing white LEDs which can replace incandescent lamps.

LEDs have the following advantages over conventional incandescent low power lamps:

(i) Low operational voltage and less power.
(ii) Fast action and no warm-up time required.
(iii) The bandwidth of emitted light is 100 Å to 500 Å or in other words it is nearly (but not exactly) monochromatic.
(iv) Long life and ruggedness.
(v) Fast on-off switching capability.

(iii) *Solar cell*

A solar cell is basically a p-n junction which generates emf when solar radiation falls on the p-n junction. It works on the same principle (photovoltaic effect) as the photodiode, except that no external bias is applied and the junction area is kept much larger for solar radiation to be incident because we are interested in more power.

A simple p-n junction solar cell is shown in Fig. 14.24.

A p-Si wafer of about 300 μm is taken over which a thin layer (~0.3 μm) of n-Si is grown on one-side by diffusion process. The other side of p-Si is coated with a metal (back contact). On the top of n-Si layer, metal finger electrode (or metallic grid) is deposited. This acts as a front contact. The metallic grid occupies only a very small fraction of the cell area (<15%) so that light can be incident on the cell from the top.

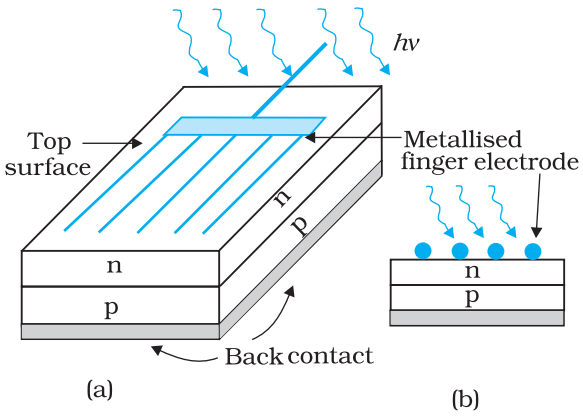


FIGURE 14.24 (a) Typical p-n junction solar cell; (b) Cross-sectional view.

The generation of emf by a solar cell, when light falls on, it is due to the following three basic processes: generation, separation and collection—(i) generation of e-h pairs due to light (with $h\nu > E_g$) close to the junction; (ii) separation of electrons and holes due to electric field of the depletion region. Electrons are swept to n-side and holes to p-side; (iii) the electrons reaching the n-side are collected by the front contact and holes reaching p-side are collected by the back contact. Thus p-side becomes positive and n-side becomes negative giving rise to *photovoltage*.

When an external load is connected as shown in the Fig. 14.25(a) a photocurrent I_L flows through the load. A typical *I-V* characteristics of a solar cell is shown in the Fig. 14.25(b).

Note that the $I - V$ characteristics of solar cell is drawn in the fourth quadrant of the coordinate axes. This is because a solar cell does not draw current but supplies the same to the load.

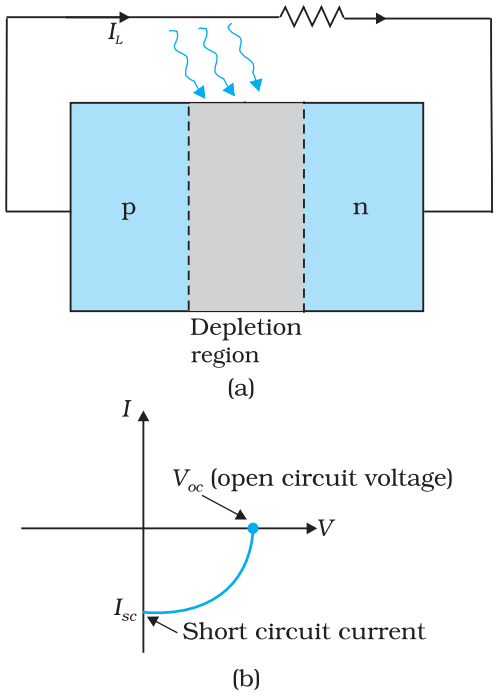


FIGURE 14.25 (a) A typical illuminated p-n junction solar cell; (b) *I-V* characteristics of a solar cell.

Semiconductors with band gap close to 1.5 eV are ideal materials for solar cell fabrication. Solar cells are made with semiconductors like Si (E_g = 1.1 eV), GaAs (E_g = 1.43 eV), CdTe (E_g = 1.45 eV), $CuInSe_2$ (E_g = 1.04 eV), etc. The important criteria for the selection of a material for solar cell fabrication are (i) band gap (~1.0 to 1.8 eV), (ii) high optical absorption (~10^4 cm^{-1}), (iii) electrical conductivity, (iv) availability of the raw material, and (v) cost. Note that sunlight is not always required for a solar cell. *Any* light with photon energies greater than the bandgap will do. Solar cells are used to power electronic devices in satellites and space vehicles and also as power supply to some calculators. Production of low-cost photovoltaic cells for large-scale solar energy is a topic for research.

Example 14.7 Why are Si and GaAs are preferred materials for solar cells?

Solution The solar radiation spectrum received by us is shown in Fig. 14.26.

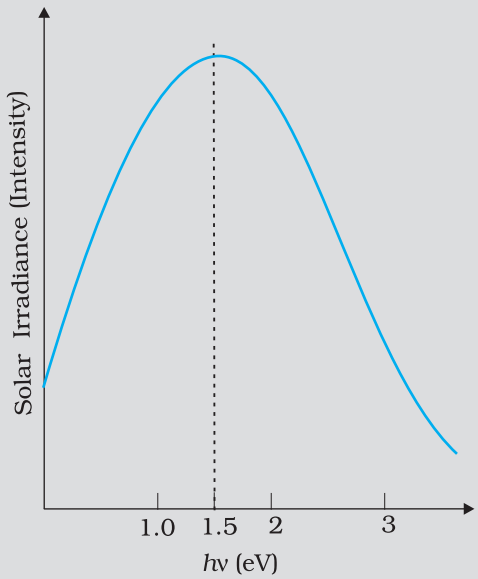


FIGURE 14.26

The maxima is near 1.5 eV. For photo-excitation, $h\nu > E_g$. Hence, semiconductor with band gap ~1.5 eV or lower is likely to give better solar conversion efficiency. Silicon has E_g ~ 1.1 eV while for GaAs it is ~1.53 eV. In fact, GaAs is better (in spite of its higher band gap) than Si because of its relatively higher absorption coefficient. If we choose materials like CdS or CdSe (E_g ~ 2.4 eV), we can use only the high energy component of the solar energy for photo-conversion and a significant part of energy will be of no use.

The question arises: why we do not use material like PbS (E_g ~ 0.4 eV) which satisfy the condition $h\nu > E_g$ for ν maxima corresponding to the solar radiation spectra? If we do so, most of the solar radiation will be absorbed on the *top-layer* of solar cell and will not reach in or near the depletion region. For effective electron-hole separation, due to the junction field, we want the photo-generation to occur in the junction region only.

EXAMPLE 14.7

14.9 Junction Transistor

The credit of inventing the transistor in the year 1947 goes to J. Bardeen and W.H. Brattain of Bell Telephone Laboratories, U.S.A. That transistor was a point-contact transistor. The first junction transistor consisting of two back-to-back p-n junctions was invented by William Schockley in 1951.

As long as only the junction transistor was known, it was known simply as transistor. But over the years new types of transistors were invented and to differentiate it from the new ones it is now called the Bipolar Junction Transistor (BJT). Even now, often the word transistor

is used to mean BJT when there is no confusion. Since our study is limited to only BJT, we shall use the word transistor for BJT without any ambiguity.

14.9.1 Transistor: structure and action

A transistor has three doped regions forming two p-n junctions between them. Obviously, there are two types of transistors, as shown in Fig. 14.27.

(i) n-p-n transistor: Here two segments of n-type semiconductor (emitter and collector) are separated by a segment of p-type semiconductor (base).

(ii) p-n-p transistor: Here two segments of p-type semiconductor (termed as emitter and collector) are separated by a segment of n-type semiconductor (termed as base).

The schematic representations of an n-p-n and a p-n-p configuration are shown in Fig. 14.27(a). All the three segments of a transistor have different thickness and their doping levels are also different. In the schematic symbols used for representing p-n-p and n-p-n transistors [Fig. 14.27(b)] the arrowhead shows the direction of conventional current in the transistor. A brief description of the three segments of a transistor is given below:

- *Emitter*: This is the segment on one side of the transistor shown in Fig. 14.27(a). It is of *moderate size* and *heavily doped*. It supplies a large number of majority carriers for the current flow through the transistor.
- *Base*: This is the central segment. *It is very thin* and *lightly doped*.
- *Collector*: This segment collects a *major* portion of the majority carriers supplied by the emitter. The collector side is *moderately doped* and *larger* in size as compared to the *emitter*.

We have seen earlier in the case of a p-n junction, that there is a formation of depletion region acorss the junction. In case of a transistor depletion regions are formed at the emitter base-junction and the base-collector junction. For understanding the action of a transistor, we have to consider the nature of depletion regions formed at these junctions. The charge carriers move across different regions of the transistor when proper voltages are applied across its terminals.

The biasing of the transistor is done differently for different uses. The transistor can be used in two distinct ways. Basically, it was invented to function as an amplifier, a device which produces a enlarged copy of a signal. But later its use as a switch acquired equal importance. We shall study both these functions and the ways the transistor is biased to achieve these mutually exclusive functions.

First we shall see what gives the transistor its amplifying capabilities. The transistor works as an amplifier, with its emitter-base junction forward biased and the base-collector junction reverse biased. This situation is shown in Fig. 14.28, where V_{CC} and V_{EE} are used for creating the respective biasing. When the transistor is biased in this way it is said to be in *active* state.We represent the voltage between emitter and base as V_{EB} and that between the collector and the base as V_{CB}. In

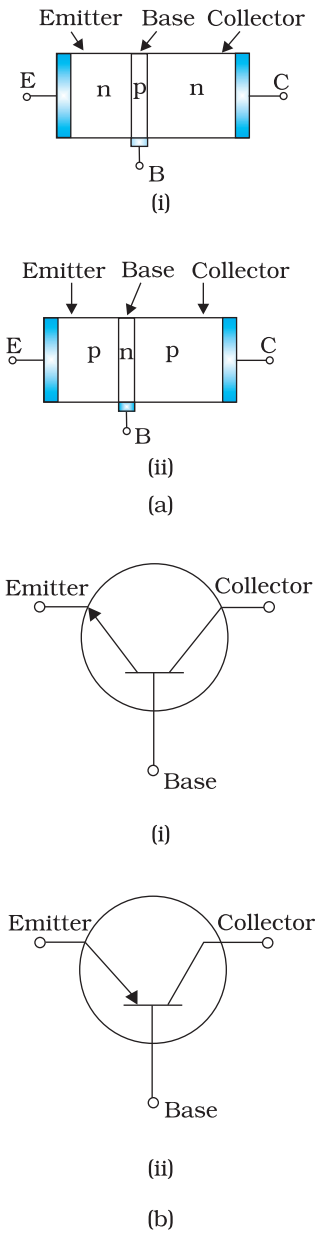


FIGURE 14.27 (a) Schematic representations of a n-p-n transistor and p-n-p transistor, and (b) Symbols for n-p-n and p-n-p transistors.

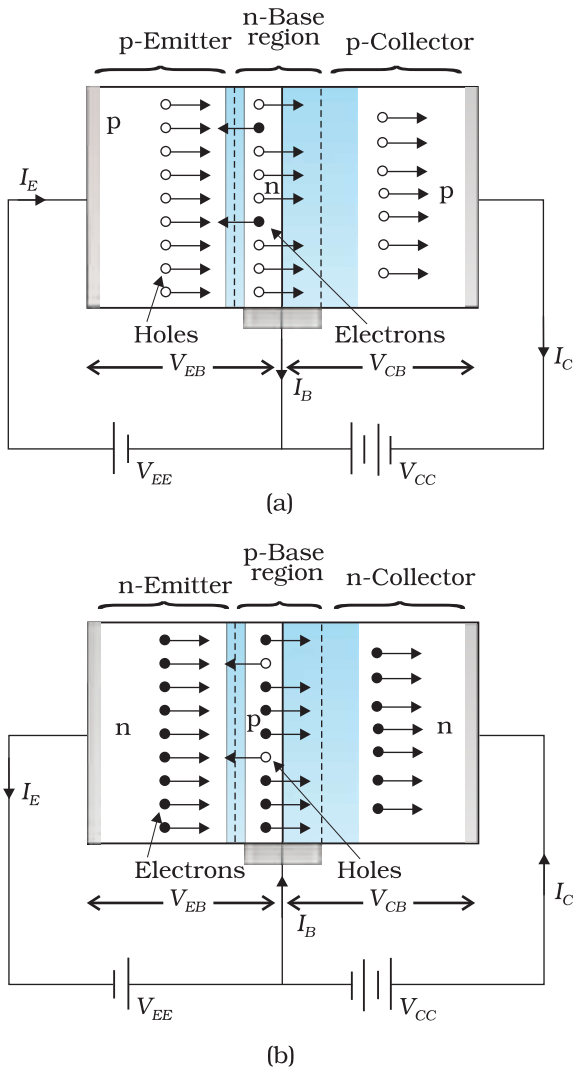


FIGURE 14.28 Bias Voltage applied on: (a) p-n-p transistor and (b) n-p-n transistor.

Fig. 14.28, base is a common terminal for the two power supplies whose other terminals are connected to emitter and collector, respectively. So the two power supplies are represented as V_{EE}, and V_{CC}, respectively. In circuits, where emitter is the common terminal, the power supply between the base and the emitter is represented as V_{BB} and that between collector and emitter as V_{CC}.

Let us see now the paths of current carriers in the transistor with emitter-base junction forward biased and base-collector junction reverse biased. The heavily doped emitter has a high concentration of majority carriers, which will be holes in a p-n-p transistor and electrons in an n-p-n transistor. These majority carriers enter the base region in large numbers. The base is thin and lightly doped. So the majority carriers there would be few. In a p-n-p transistor the majority carriers in the base are electrons since base is of n-type semiconductor. The large number of holes entering the base from the emitter swamps the small number of electrons there. As the base collector-junction is reverse-biased, these holes, which appear as minority carriers at the junction, can easily cross the junction and enter the collector. The holes in the base could move either towards the base terminal to combine with the electrons entering from outside or cross the junction to enter into the collector and reach the collector terminal. The base is made thin so that most of the holes find themselves near the reverse-biased base-collector junction and so cross the junction instead of moving to the base terminal.

It is interesting to note that due to forward bias a large current enters the emitter-base junction, but most of it is diverted to adjacent reverse-biased base-collector junction and the current coming out of the base becomes a very small fraction of the current that entered the junction. If we represent the hole current and the electron current crossing the forward biased junction by I_h and I_e respectively then the total current in a forward biased diode is the sum $I_h + I_e$. We see that the emitter current $I_E = I_h + I_e$ but the base current $I_B << I_h + I_e$, because a major part of I_E goes to collector instead of coming out of the base terminal. The base current is thus a small fraction of the emitter current.

The current entering into the emitter from outside is equal to the emitter current I_E. Similarly the current emerging from the base terminal is I_B and that from collector terminal is I_C. It is obvious from the above description and also from a straight forward application of Kirchhoff's law to Fig. 14.28(a) that the emitter current is the sum of collector current and base current:

$$I_E = I_C + I_B \quad (14.7)$$

We also see that $I_C \approx I_E$.

Our description of the direction of motion of the holes is identical with the direction of the conventional current. But the direction of motion of electrons is just opposite to that of the current. Thus in a p-n-p transistor the current enters from emitter into base whereas in a n-p-n transistor it enters from the base into the emitter. The arrowhead in the emitter shows the direction of the conventional current.

The description about the paths followed by the majority and minority carriers in a n-p-n is exactly the same as that for the p-n-p transistor. But the current paths are exactly opposite, as shown in Fig. 14.28. In Fig. 14.28(b) the electrons are the majority carriers supplied by the n-type emitter region. They cross the thin p-base region and are able to reach the collector to give the collector current, I_C. From the above description we can conclude that in the active state of the transistor the emitter-base junction acts as a low resistance while the base collector acts as a high resistance.

14.9.2 Basic transistor circuit configurations and transistor characteristics

In a transistor, only three terminals are available, viz., *Emitter* (E), *Base* (B) and *Collector* (C). Therefore, in a circuit the input/output connections have to be such that one of these (E, B or C) is common to both the input and the output. Accordingly, the transistor can be *connected* in either of the following three configurations:

Common Emitter (CE), Common Base (CB), Common Collector (CC)

The transistor is most widely used in the CE configuration and we shall restrict our discussion to only this configuration. Since more commonly used transistors are n-p-n Si transistors, we shall confine our discussion to such transistors only. With p-n-p transistors the polarities of the external power supplies are to be inverted.

Common emitter transistor characteristics

When a transistor is used in CE configuration, the input is between the base and the emitter and the output is between the collector and the emitter. The variation of the base current I_B with the base-emitter voltage V_{BE} is called the *input characteristic*. Similarly, the variation of the collector current I_C with the collector-emitter voltage V_{CE} is called the *output characteristic*. You will see that the output characteristics are controlled by the input characteristics. This implies that the collector current changes with the base current.

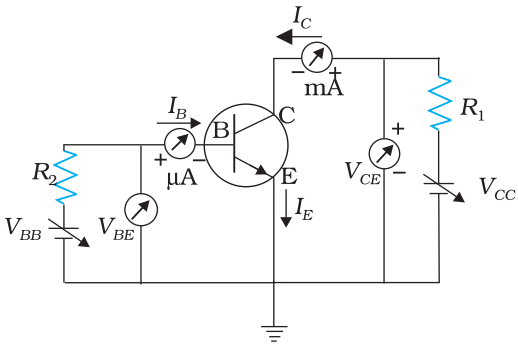


FIGURE 14.29 Circuit arrangement for studying the input and output characteristics of n-p-n transistor in CE configuration.

The input and the output characteristics of an n-p-n transistors can be studied by using the circuit shown in Fig. 14.29.

To study the input characteristics of the transistor in C_E configuration, a curve is plotted between the base current I_B against the base-emitter voltage V_{BE}. The

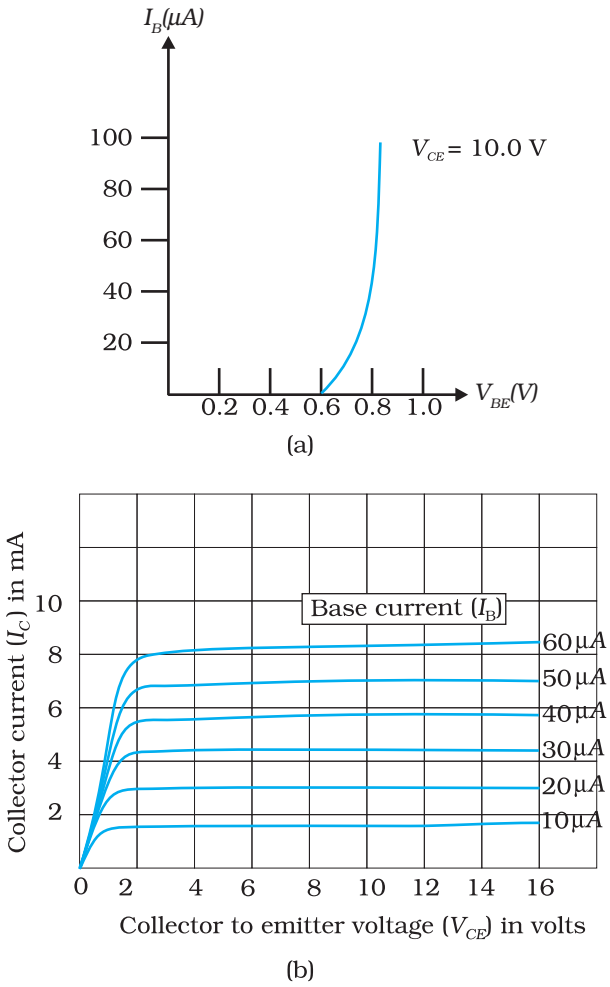


FIGURE 14.30 (a) Typical input characteristics, and (b) Typical output characteristics.

collector-emitter voltage V_{CE} is kept fixed while studying the dependence of I_B on V_{BE}. We are interested to obtain the input characteristic when the transistor is in active state. So the collector-emitter voltage V_{CE} is kept large enough to make the base collector junction reverse biased. Since $V_{CE} = V_{CB} + V_{BE}$ and for Si transistor V_{BE} is 0.6 to 0.7 V, V_{CE} must be sufficiently larger than 0.7 V. Since the transistor is operated as an amplifier over large range of V_{CE}, the reverse bias across the base-collector junction is high most of the time. Therefore, the input characteristics may be obtained for V_{CE} somewhere in the range of 3 V to 20 V. Since *the increase in V_{CE} appears as increase in V_{CB}, its effect on I_B is negligible.* As a consequence, input characteristics for various values of V_{CE} will give almost identical curves. Hence, *it is enough to determine only one input characteristics*. The input characteristics of a transistor is as shown in Fig. 14.30(a).

The output characteristic is obtained by observing the variation of I_C as V_{CE} is varied keeping I_B constant. It is obvious that if V_{BE} is increased by a small amount, both hole current from the emitter region and the electron current from the base region will increase. As a consequence both I_B and I_C will increase proportionately. This shows that when I_B increases I_C also increases. The plot of I_C versus V_{CE} for different fixed values of I_B gives one output characteristic. So there will be different output characteristics corresponding to different values of I_B as shown in Fig. 14.30(b).

The linear segments of both the input and output characteristics can be used to calculate some important ac parameters of transistors as shown below.

(i) ***Input resistance*** **(r_i):** This is defined as the ratio of change in base-emitter voltage (ΔV_{BE}) to the resulting change in base current (ΔI_B) at constant collector-emitter voltage (V_{CE}). This is dynamic (ac resistance) and as can be seen from the input characteristic, its value varies with the operating current in the transistor:

$$r_i = \left(\frac{\Delta V_{BE}}{\Delta I_B}\right)_{V_{CE}} \tag{14.8}$$

The value of r_i can be anything from a few hundreds to a few thousand ohms.

(ii) *Output resistance (r_o):* This is defined as the ratio of change in collector-emitter voltage (ΔV_{CE}) to the change in collector current (ΔI_C) at a constant base current I_B.

$$r_o = \left(\frac{\Delta V_{CE}}{\Delta I_C}\right)_{I_B} \tag{14.9}$$

The output characteristics show that initially for very small values of V_{CE}, I_C increases almost linearly. This happens because the base-collector junction is not reverse biased and the transistor is not in active state. In fact, the transistor is in the saturation state and the current is controlled by the supply voltage V_{CC} (=V_{CE}) in this part of the characteristic. When V_{CE} is more than that required to reverse bias the base-collector junction, I_C increases very little with V_{CE}. The reciprocal of the slope of the linear part of the output characteristic gives the values of *r*o. The output resistance of the transistor is mainly controlled by the bias of the base-collector junction. The high magnitude of the output resistance (of the order of 100 kΩ) is due to the reverse-biased state of this diode. This also explains why the resistance at the initial part of the characteristic, when the transistor is in saturation state, is very low.

(iii) *Current amplification factor (β):* This is defined as the ratio of the change in collector current to the change in base current at a constant collector-emitter voltage (V_{CE}) when the transistor is in active state.

$$\beta_{ac} = \left(\frac{\Delta I_C}{\Delta I_B}\right)_{V_{CE}} \tag{14.10}$$

This is also known as *small signal current gain* and its value is very large.

If we simply find the ratio of I_C and I_B we get what is called dc β of the transistor. Hence,

$$\beta_{dc} = \frac{I_C}{I_B} \tag{14.11}$$

Since I_C increases with I_B almost linearly and $I_C = 0$ when $I_B = 0$, the values of both β_{dc} and β_{ac} are nearly equal. So, for most calculations β_{dc} can be used. Both β_{ac} and β_{dc} vary with V_{CE} and I_B (or I_C) slightly.

Example 14.8 From the output characteristics shown in Fig. 14.30(b), calculate the values of β_{ac} and β_{dc} of the transistor when V_{CE} is 10 V and I_C = 4.0 mA.

Solution

$$\beta_{ac} = \left(\frac{\Delta I_C}{\Delta I_B}\right)_{V_{CE}}, \quad \beta_{dc} = \frac{I_C}{I_B}$$

For determining β_{ac} and β_{dc} at the stated values of V_{CE} and I_C one can proceed as follows. Consider any two characteristics for two values of I_B which lie above and below the given value of I_C. Here I_C = 4.0 mA. (Choose characteristics for I_B= 30 and 20 μA.) At V_{CE} = 10 V we read the two values of I_C from the graph. Then

EXAMPLE 14.8

EXAMPLE 14.8

ΔI_B = (30 – 20) μA = 10 μA, ΔI_C = (4.5 – 3.0) mA = 1.5 mA

Therefore, β_{ac} = 1.5 mA/ 10 μA = 150

For determining $\beta_{dc,}$ either estimate the value of I_B corresponding to I_C = 4.0 mA at V_{CE} = 10 V or calculate the two values of β_{dc} for the two characteristics chosen and find their mean.

Therefore, for I_C = 4.5 mA and I_B = 30 μA,

β_{dc} = 4.5 mA/ 30 μA = 150

and for I_C = 3.0 mA and I_B = 20 μA

β_{dc} =3.0 mA / 20 μA = 150

Hence, β_{dc} =(150 + 150) /2 = 150

14.9.3 Transistor as a device

The transistor can be used as a device application depending on the configuration used (namely CB, CC and CE), the biasing of the E-B and B-C junction and the operation region namely cutoff, active region and saturation. As mentioned earlier we have confined only to the CE configuration and will be concentrating on the biasing and the operation region to understand the working of a device.

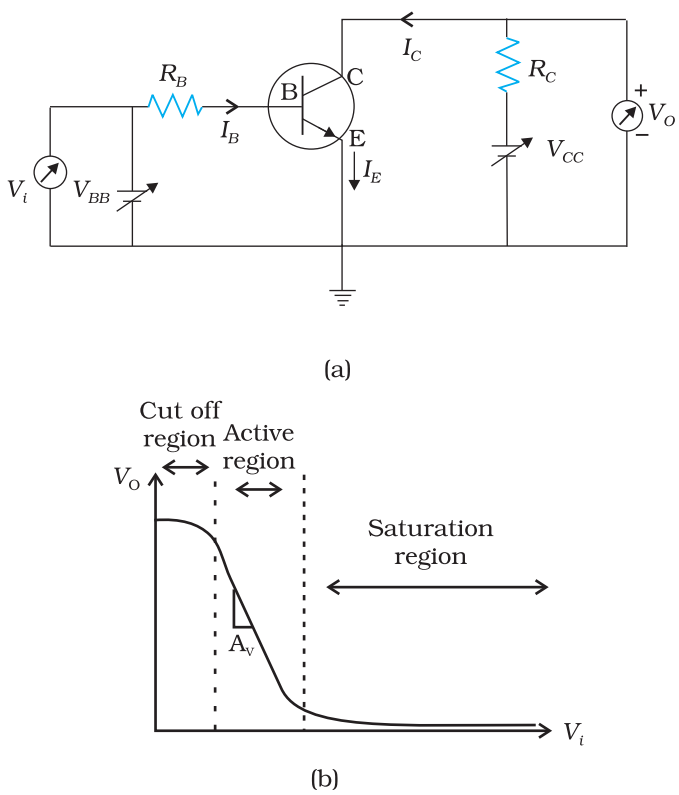


FIGURE 14.31 (a) Base-biased transistor in CE configuration, (b) Transfer characteristic.

When the transistor is used in the cutoff or saturation state it acts as a *switch*. On the other hand for using the transistor as an *amplifier*, it has to operate in the active region.

(i) Transistor as a switch

We shall try to understand the operation of the transistor as a switch by analysing the behaviour of the base-biased transistor in CE configuration as shown in Fig. 14.31(a).

Applying Kirchhoff's voltage rule to the input and output sides of this circuit, we get

$$V_{BB} = I_B R_B + V_{BE} \quad (14.12)$$

and

$$V_{CE} = V_{CC} - I_C R_C. \quad (14.13)$$

We shall treat V_{BB} as the dc input voltage V_i and V_{CE} as the dc output voltage V_O. So, we have

$V_i = I_B R_B + V_{BE}$ and

$V_o = V_{CC} - I_C R_C$.

Let us see how V_o changes as V_i increases from zero onwards. In the case of Si transistor, as long as input V_i is less than 0.6 V, the transistor will be in cut off state and current I_C will be zero.

Hence $V_o = V_{CC}$

When V_i becomes greater than 0.6 V the transistor is in active state with some current I_C in the output path and the output V_o decrease as the

term $I_C R_C$ increases. With increase of V_i, I_C increases almost linearly and so V_o decreases linearly till its value becomes less than about 1.0 V.

Beyond this, the change becomes non linear and transistor goes into saturation state. With further increase in V_i the output voltage is found to decrease further towards zero though it may never become zero. If we plot the V_o vs V_i curve, [also called the transfer characteristics of the base-biased transistor (Fig. 14.31(b)], we see that between cut off state and active state and also between active state and saturation state there are regions of non-linearity showing that the transition from cutoff state to active state and from active state to saturation state are not sharply defined.

Let us see now how the transistor is operated as a switch. As long as V_i is *low* and unable to forward-bias the transistor, V_o is *high* (at V_{CC}). If V_i is *high* enough to drive the transistor into saturation, then V_o is *low*, very near to zero. When the transistor is not conducting it is said to be *switched off* and when it is driven into saturation it is said to be *switched on*. This shows that if we define low and high states as below and above certain voltage levels corresponding to cutoff and saturation of the transistor, then we can say that a *low* input switches the transistor off and a *high* input switches it on. Alternatively, we can say that a *low* input to the transistor gives a *high* output and a *high* input gives a *low* output. The switching circuits are designed in such a way that the transistor does not remain in active state.

(ii) Transistor as an amplifier

For using the transistor as an amplifier we will use the active region of the V_o versus V_i curve. The slope of the linear part of the curve represents the rate of change of the output with the input. It is negative because the output is $V_{CC} - I_C R_C$ and not $I_C R_C$. That is why as input voltage of the CE amplifier increases its output voltage decreases and the output is said to be out of phase with the input. If we consider ΔV_o and ΔV_i as small changes in the output and input voltages then $\Delta V_o / \Delta V_i$ is called the small signal voltage gain A_V of the amplifier.

If the V_{BB} voltage has a fixed value corresponding to the mid point of the active region, the circuit will behave as a CE amplifier with voltage gain $\Delta V_o / \Delta V_i$. We can express the voltage gain A_V in terms of the resistors in the circuit and the current gain of the transistor as follows.

We have, $V_o = V_{CC} - I_C R_C$

Therefore, $\Delta V_o = 0 - R_C \Delta I_C$

Similarly, from $V_i = I_B R_B + V_{BE}$

$\Delta V_i = R_B \Delta I_B + \Delta V_{BE}$

But ΔV_{BE} is negligibly small in comparison to $\Delta I_B R_B$ in this circuit.

So, the *voltage gain* of this *CE amplifier* (Fig. 14.32) is given by

$$A_V = -R_C \Delta I_C / R_B \Delta I_B$$

$$= -\beta_{ac}(R_C / R_B) \qquad (14.14)$$

where β_{ac} is equal to $\Delta I_C / \Delta I_B$ from Eq. (14.10). Thus the linear portion of the active region of the transistor can be exploited for the use in amplifiers. Transistor as an amplifier (CE configuration) is discussed in detail in the next section.

14.9.4 Transistor as an Amplifier (CE-Configuration)

To operate the transistor as an amplifier it is necessary to fix its operating point somewhere in the middle of its active region. If we fix the value of V_{BB} corresponding to a point in the middle of the linear part of the transfer curve then the dc base current I_B would be constant and corresponding collector current I_C will also be constant. The dc voltage $V_{CE} = V_{CC} - I_C R_C$ would also remain constant. The operating values of V_{CE} and I_B determine the operating point, of the amplifier.

If a small sinusoidal voltage with amplitude v_s is superposed on the dc base bias by connecting the source of that signal in series with the V_{BB} supply, then the base current will have sinusoidal variations superimposed on the value of I_B. As a consequence the collector current also will have sinusoidal variations superimposed on the value of I_C, producing in turn corresponding change in the value of V_O. We can measure the ac variations across the input and output terminals by blocking the dc voltages by large capacitors.

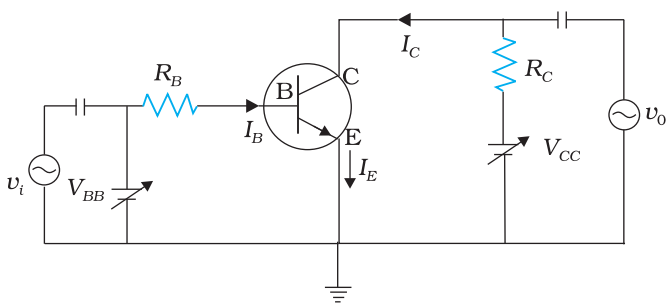


FIGURE 14.32 A simple circuit of a CE-transistor amplifier.

In the discription of the amplifier given above we have not considered any ac signal. In general, amplifiers are used to amplify alternating signals. Now let us superimpose an ac input signal v_i (to be amplified) on the bias V_{BB} (dc) as shown in Fig. 14.32. The output is taken between the collector and the ground.

The working of an amplifier can be easily understood, if we first assume that $v_i = 0$. Then applying Kirchhoff's law to the output loop, we get

$$V_{cc} = V_{CE} + I_c R_L \tag{14.15}$$

Likewise, the input loop gives

$$V_{BB} = V_{BE} + I_B R_B \tag{14.16}$$

When v_i is not zero, we get

$$V_{BE} + v_i = V_{BE} + I_B R_B + \Delta I_B (R_B + r_i)$$

The change in V_{BE} can be related to the input resistance r_i [see Eq. (14.8)] and the change in I_B. Hence

$$v_i = \Delta I_B (R_B + r_i)$$

$$= r \Delta I_B$$

The change in I_B causes a change in I_c. We define a parameter β_{ac}, which is similar to the β_{dc} defined in Eq. (14.11), as

$$\beta_{ac} = \frac{\Delta I_c}{\Delta I_B} = \frac{i_c}{i_b} \tag{14.17}$$

which is also known as the *ac current gain* A_i. Usually β_{ac} is close to β_{dc} in the linear region of the output characteristics.

The change in I_c due to a change in I_B causes a change in V_{CE} and the voltage drop across the resistor R_L because V_{CC} is fixed.

These changes can be given by Eq. (14.15) as

$\Delta V_{CC} = \Delta V_{CE} + R_L \Delta I_C = 0$

or $\Delta V_{CE} = -R_L \Delta I_C$

The change in V_{CE} is the output voltage υ_0. From Eq. (14.10), we get

$\upsilon_0 = \Delta V_{CE} = -\beta_{ac} R_L \Delta I_B$

The voltage gain of the amplifier is

$$A_\upsilon = \frac{\upsilon_0}{\upsilon_i} = \frac{\Delta V_{CE}}{r \Delta I_B}$$

$$= -\frac{\beta_{ac} R_L}{r} \qquad (14.18)$$

The negative sign represents that output voltage is opposite with phase with the input voltage.

From the discussion of the transistor characteristics you have seen that there is a current gain β_{ac} in the CE configuration. Here we have also seen the voltage gain A_υ. Therefore the power gain A_p can be expressed as the product of the current gain and voltage gain. Mathematically

$$A_p = \beta_{ac} \times A_\upsilon \qquad (14.19)$$

Since β_{ac} and A_υ are greater than 1, we get ac power gain. However it should be realised that transistor is not a power generating device. The energy for the higher ac power at the output is supplied by the battery.

Example 14.9 In Fig. 14.31(a), the V_{BB} supply can be varied from 0V to 5.0 V. The Si transistor has β_{dc} = 250 and R_B = 100 kΩ, R_C = 1 KΩ, V_{CC} = 5.0V. Assume that when the transistor is saturated, V_{CE} = 0V and V_{BE} = 0.8V. Calculate (a) the minimum base current, for which the transistor will reach saturation. Hence, (b) determine V_1 when the transistor is 'switched on'. (c) find the ranges of V_1 for which the transistor is 'switched off' and 'switched on'.

Solution

Given at saturation V_{CE} = 0V, V_{BE} = 0.8V

$V_{CE} = V_{CC} - I_C R_C$

$I_C = V_{CC}/R_C$ = 5.0V/1.0kΩ = 5.0 mA

Therefore $I_B = I_C/\beta$ = 5.0 mA/250 = 20μA

The input voltage at which the transistor will go into saturation is given by

$V_{IH} = V_{BB} = I_B R_B + V_{BE}$

= 20μA × 100 kΩ + 0.8V = 2.8V

The value of input voltage below which the transistor remains cutoff is given by

V_{IL} = 0.6V, V_{IH} = 2.8V

Between 0.0V and 0.6V, the transistor will be in the 'switched off' state. Between 2.8V and 5.0V, it will be in 'switched on' state.

Note that the transistor is in active state when I_B varies from 0.0mA to 20mA. In this range, $I_C = \beta I_B$ is valid. In the saturation range, $I_C \leq \beta I_B$.

EXAMPLE 14.9

EXAMPLE 14.10

Example 14.10 For a CE transistor amplifier, the audio signal voltage across the collector resistance of 2.0 kΩ is 2.0 V. Suppose the current amplification factor of the transistor is 100, What should be the value of R_B in series with V_{BB} supply of 2.0 V if the dc base current has to be 10 times the signal current. Also calculate the dc drop across the collector resistance. (Refer to Fig. 14.33).

Solution The output ac voltage is 2.0 V. So, the ac collector current i_C = 2.0/2000 = 1.0 mA. The signal current through the base is, therefore given by $i_B = i_C / \beta$ = 1.0 mA/100 = 0.010 mA. The dc base current has to be 10× 0.010 = 0.10 mA.
From Eq.14.16, $R_B = (V_{BB} - V_{BE}) / I_B$. Assuming V_{BE} = 0.6 V,
R_B = (2.0 – 0.6)/0.10 = 14 kΩ.
The dc collector current I_C = 100×0.10 = 10 mA.

14.9.5 Feedback amplifier and transistor oscillator

In an amplifier, we have seen that a sinusoidal input is given which appears as an amplified signal in the output. This means that an *external input is necessary to sustain ac signal in the output for an amplifier.* In an oscillator, we get ac output without any external input signal. In other words, the output in an oscillator is *self-sustained*. To attain this, an amplifier is taken. A portion of the output power is returned back (feedback) to the input *in phase* with the starting power (this process is termed *positive feedback)* as shown in Fig. 14.33(a). The feedback can be achieved by inductive coupling (through mutual inductance) or *LC* or *RC* networks. Different types of oscillators essentially use different methods of coupling the output to the input (feedback network), apart from the resonant circuit for obtaining oscillation at a particular frequency. For understanding the oscillator action, we consider the circuit shown in Fig. 14.33(b) in which the feedback is accomplished by *inductive coupling* from one coil winding (T_1) to another coil winding (T_2). Note that the coils T_2 and T_1 are wound on the same core and hence are inductively coupled through their mutual inductance. As in an amplifier, the base-emitter junction is forward biased while the base-collector junction is reverse biased. Detailed biasing circuits actually used have been omitted for simplicity.

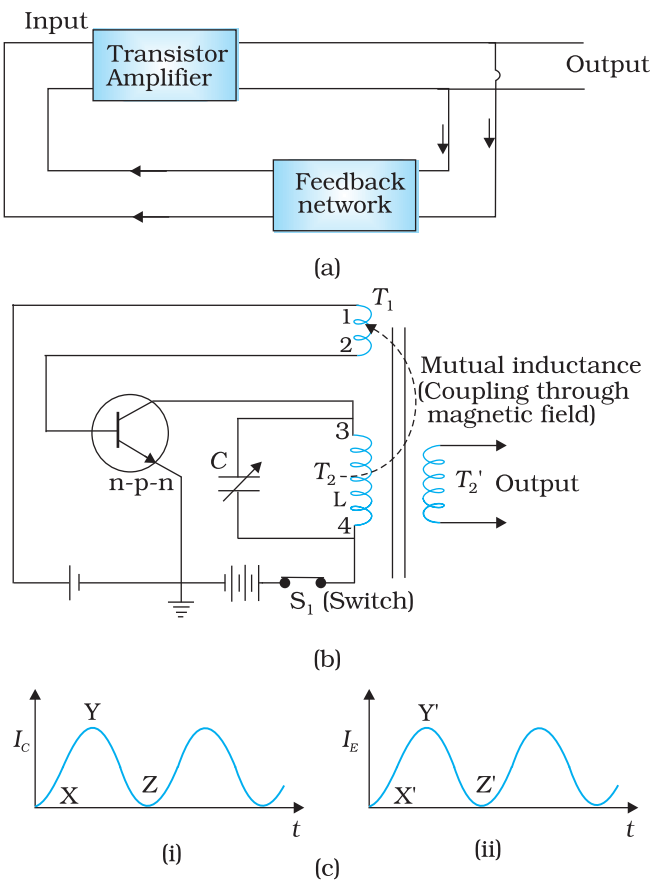


FIGURE 14.33 (a) Principle of a transistor amplifier with positive feedback working as an oscillator and (b) Tuned collector oscillator, (c) Rise and fall (or built up) of current I_c and I_e due to the inductive coupling.

Let us try to understand how oscillations are built. Suppose switch S_1 is put *on* to

apply proper bias for the first time. Obviously, a *surge* of collector current flows in the transistor. This current flows through the coil T_2 where terminals are numbered 3 and 4 [Fig. 14.33(b)]. This current does not reach full amplitude instantaneously but increases from X to Y, as shown in Fig. [14.33(c)(i)]. The inductive coupling between coil T_2 and coil T_1 now causes a current to flow in the emitter circuit (note that this actually is the 'feedback' from input to output). As a result of this positive feedback, this current (in T_1; emitter current) also increases from X′ to Y′ [Fig. 14.33(c)(ii)]. The current in T_2 (collector current) connected in the collector circuit acquires the value Y when the transistor becomes *saturated*. This means that maximum collector current is flowing and can increase no further. Since there is no further change in collector current, the magnetic field around T_2 ceases to grow. As soon as the field becomes static, there will be no further feedback from T_2 to T_1. Without continued feedback, the emitter current begins to fall. Consequently, collector current decreases from Y towards Z [Fig. 14.33(c)(i)]. However, a decrease of collector current causes the magnetic field to decay around the coil T_2. Thus, T_1 is now seeing a decaying field in T_2 (opposite from what it saw when the field was growing at the initial *start* operation). This causes a further decrease in the emitter current till it reaches Z′ when the transistor is *cut-off*. This means that both I_E and I_C cease to flow. Therefore, the transistor has reverted back to its original state (when the power was first switched on). The whole process now repeats itself. That is, the transistor is driven to saturation, then to cut-off, and then back to saturation. The time for change from saturation to cut-off and back is determined by the constants of the tank circuit or tuned circuit (inductance L of coil T_2 and C connected in parallel to it). The resonance frequency (ν) of this tuned circuit determines the frequency at which the oscillator will oscillate.

$$\nu = \left(\frac{1}{2\pi\sqrt{LC}}\right) \tag{14.20}$$

In the circuit of Fig. 14.33(b), the tank or tuned circuit is connected in the collector side. Hence, it is known as *tuned collector oscillator*. If the tuned circuit is on the base side, it will be known as *tuned base oscillator*. There are many other types of tank circuits (say *RC*) or feedback circuits giving different types of oscillators like Colpitt's oscillator, Hartley oscillator, *RC*-oscillator.

14.10 Digital Electronics and Logic Gates

In electronics circuits like amplifiers, oscillators, introduced to you in earlier sections, the signal (current or voltage) has been in the form of continuous, time-varying voltage or current. Such signals are called continuous or *analog signals*. A typical analog signal is shown in Figure. 14.34(a). Fig. 14.34(b) shows a *pulse waveform* in which only discrete values of voltages are possible. It is convenient to use binary numbers to represent such signals. A binary number has only two digits '0' (say, 0V) and '1' (say, 5V). In digital electronics we use only these two levels of voltage as shown in Fig. 14.34(b). Such signals are called *Digital Signals*. In digital circuits only two values (represented by 0 or 1) of the input and output voltage are permissible.

This section is intended to provide the first step in our understanding of digital electronics. We shall restrict our study to some basic building blocks of digital electronics (called *Logic Gates*) which process the digital signals in a specific manner. Logic gates are used in calculators, digital watches, computers, robots, industrial control systems, and in telecommunications.

A light switch in your house can be used as an example of a digital circuit. The light is either ON or OFF depending on the switch position. When the light is ON, the output value is '1'. When the light is OFF the output value is '0'. The inputs are the position of the light switch. The switch is placed either in the ON or OFF position to activate the light.

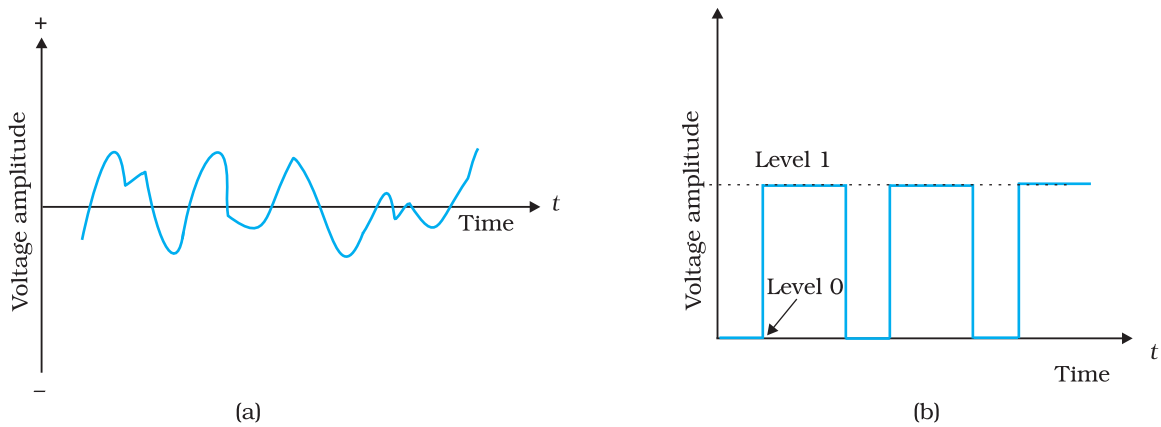
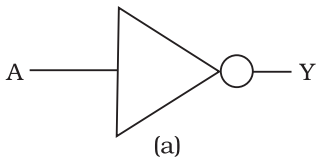


FIGURE 14.34 (a) Analog signal, (b) Digital signal.



Input	Output
A	Y
0	1
1	0

(b)

FIGURE 14.35 (a) Logic symbol, (b) Truth table of NOT gate.

14.10.1 Logic gates

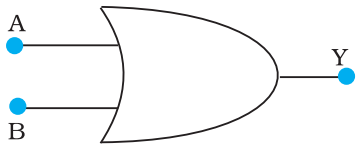
A gate is a digital circuit that follows curtain *logical* relationship between the input and output voltages. Therefore, they are generally known as *logic gates* — gates because they control the flow of information. The five common logic gates used are NOT, AND, OR, NAND, NOR. Each logic gate is indicated by a symbol and its function is defined by a *truth table* that shows all the possible input logic level combinations with their respective output logic levels. Truth tables help understand the behaviour of logic gates. These logic gates can be realised using semiconductor devices.

(i) NOT gate

This is the most basic gate, with one input and one output. It produces a '1' output if the input is '0' and vice-versa. That is, it produces an inverted version of the input at its output. This is why it is also known as an *inverter*. The commonly used symbol together with the truth table for this gate is given in Fig. 14.35.

(ii) OR Gate

An *OR* gate has two or more inputs with one output. The logic symbol and truth table are shown in Fig. 14.36. The output Y is 1 when either input A *or* input B *or* both are 1s, that is, if any of the input is high, the output is high.



(a)

Input		Output
A	B	Y
0	0	0
0	1	1
1	0	1
1	1	1

(b)

FIGURE 14.36 (a) Logic symbol (b) Truth table of OR gate.

Apart from carrying out the above mathematical logic operation, this *gate* can be used for modifying the pulse waveform as explained in the following example.

Example 14.11 Justify the output waveform (Y) of the OR gate for the following inputs A and B given in Fig. 14.37.

Solution Note the following:

- At $t < t_1$; A = 0, B = 0; Hence Y = 0
- For t_1 to t_2; A = 1, B = 0; Hence Y = 1
- For t_2 to t_3; A = 1, B = 1; Hence Y = 1
- For t_3 to t_4; A = 0, B = 1; Hence Y = 1
- For t_4 to t_5; A = 0, B = 0; Hence Y = 0
- For t_5 to t_6; A = 1, B = 0; Hence Y = 1
- For $t > t_6$; A = 0, B = 1; Hence Y = 1

Therefore the waveform Y will be as shown in the Fig. 14.37.

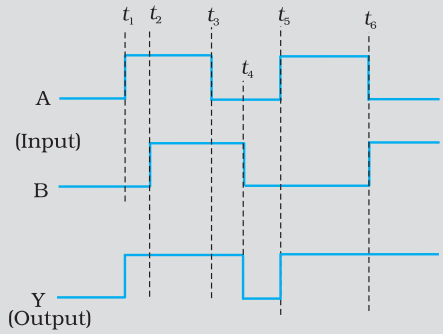
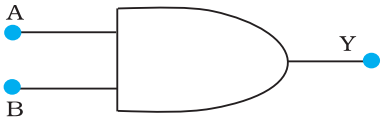


FIGURE 14.37

EXAMPLE 14.11

(iii) AND Gate

An *AND* gate has two or more inputs and one output. The output Y of AND gate is 1 only when input A *and* input B are both 1. The logic symbol and truth table for this gate are given in Fig. 14.38



(a)

Input		Output
A	B	Y
0	0	0
0	1	0
1	0	0
1	1	1

(b)

FIGURE 14.38 (a) Logic symbol, (b) Truth table of AND gate.

EXAMPLE 14.12

Example 14.12 Take A and B input waveforms similar to that in Example 14.11. Sketch the output waveform obtained from AND gate.

Solution

- For $t \le t_1$; A = 0, B = 0; Hence Y = 0
- For t_1 to t_2; A = 1, B = 0; Hence Y = 0
- For t_2 to t_3; A = 1, B = 1; Hence Y = 1
- For t_3 to t_4; A = 0, B = 1; Hence Y = 0
- For t_4 to t_5; A = 0, B = 0; Hence Y = 0
- For t_5 to t_6; A = 1, B = 0; Hence Y = 0
- For $t > t_6$; A = 0, B = 1; Hence Y = 0

Based on the above, the output waveform for AND gate can be drawn as given below.

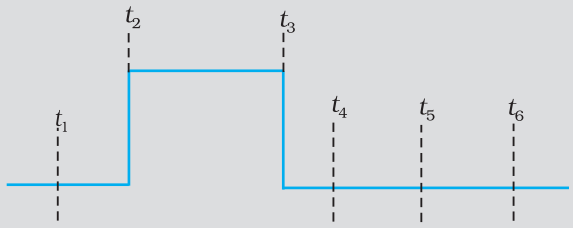
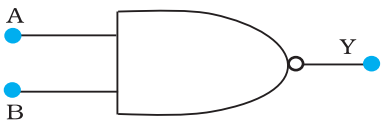


FIGURE 14.39

(iv) NAND Gate

This is an AND gate followed by a NOT gate. If inputs A *and* B are both '1', the output Y is *not* '1'. The gate gets its name from this NOT AND behaviour. Figure 14.40 shows the symbol and truth table of NAND gate.

NAND gates are also called *Universal Gates* since by using these gates you can realise other basic gates like OR, AND and NOT (Exercises 14.16 and 14.17).



(a)

Input		Output
A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

(b)

FIGURE 14.40 (a) Logic symbol, (b) Truth table of NAND gate.

EXAMPLE 14.13

Example 14.13 Sketch the output Y from a NAND gate having inputs A and B given below:

Solution

- For $t < t_1$; A = 1, B = 1; Hence Y = 0
- For t_1 to t_2; A = 0, B = 0; Hence Y = 1
- For t_2 to t_3; A = 0, B = 1; Hence Y = 1
- For t_3 to t_4; A = 1, B = 0; Hence Y = 1

- For t_4 to t_5; A = 1, B = 1; Hence Y = 0
- For t_5 to t_6; A = 0, B = 0; Hence Y = 1
- For $t > t_6$; A = 0, B = 1; Hence Y = 1

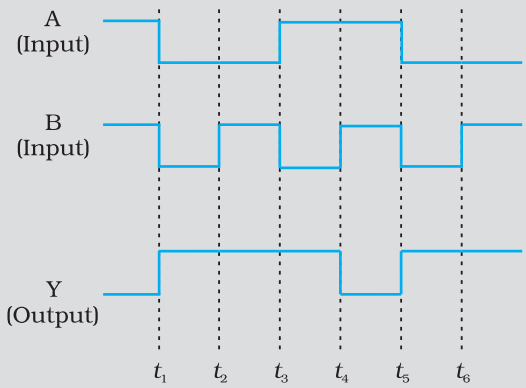
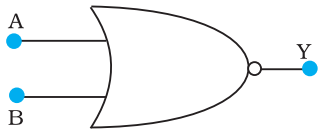


FIGURE 14.41

EXAMPLE 14.13

(v) NOR Gate

It has two or more inputs and one output. A NOT- operation applied *after* OR gate gives a NOT-OR gate (or simply NOR gate). Its output Y is '1' only when both inputs A and B are '0', i.e., neither one input *nor* the other is '1'. The symbol and truth table for NOR gate is given in Fig. 14.42.



(a)

Input		Output
A	B	Y
0	0	1
0	1	0
1	0	0
1	1	0

(b)

FIGURE 14.42 (a) Logic symbol, (b) Truth table of NOR gate.

NOR gates are considered as *universal* gates because you can obtain all the gates like AND, OR, NOT by using only NOR gates (Exercises 14.18 and 14.19).

14.11 INTEGRATED CIRCUITS

The conventional method of making circuits is to choose components like diodes, transistor, *R*, *L*, *C* etc., and connect them by soldering wires in the desired manner. Inspite of the miniaturisation introduced by the discovery of transistors, such circuits were still bulky. Apart from this, such circuits were less reliable and less shock proof. The concept of fabricating *an entire circuit* (consisting of many passive components like *R* and *C* and active devices like diode and transistor) on a small single block (or chip) of a semiconductor has revolutionised the electronics technology. Such a circuit is known as *Integrated Circuit* (IC). The most widely used technology is the *Monolithic Integrated Circuit.* The word

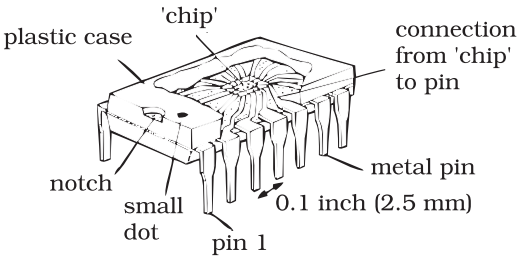


FIGURE 14.43 The casing and connection of a 'chip'.

monolithic is a combination of two greek words, *monos* means single and *lithos* means stone. This, in effect, means that the entire circuit is formed on a single silicon crystal (or *chip*). The *chip* dimensions are as small as 1mm × 1mm or it could even be smaller. Figure 14.43 shows a chip in its protective plastic case, partly removed to reveal the connections coming out from the 'chip' to the pins that enable it to make external connections.

Depending on nature of input signals, IC's can be grouped in two categories: (a) *linear* or *analogue IC's* and (b) digital IC's. The linear IC's process analogue signals which change smoothly and continuously over a range of values between a maximum and a minimum. The output is more or less directly proportional to the input, i.e., it varies *linearly* with the input. One of the most useful linear IC's is the operational amplifier.

The digital IC's process signals that have only two values. They contain circuits such as logic gates. Depending upon the level of integration (i.e., the number of circuit components or logic gates), the ICs are termed as Small Scale Integration, SSI (logic gates $\leq$ 10); Medium Scale Integration, MSI (logic gates $\leq$ 100); Large Scale Integration, LSI (logic gates $\leq$ 1000); and Very Large Scale Integration, VLSI (logic gates > 1000). The technology of fabrication is very involved but large scale industrial production has made them very inexpensive.

FASTER AND SMALLER: THE FUTURE OF COMPUTER TECHNOLOGY

The *Integrated Chip* (IC) is at the heart of all computer systems. In fact ICs are found in almost all electrical devices like cars, televisions, CD players, cell phones etc. The miniaturisation that made the modern personal computer possible could never have happened without the IC. ICs are electronic devices that contain many transistors, resistors, capacitors, connecting wires – all in one package. You must have heard of the *microprocessor*. The microprocessor is an IC that processes all information in a computer, like keeping track of what keys are pressed, running programmes, games etc. The IC was first invented by Jack Kilky at Texas Instruments in 1958 and he was awarded Nobel Prize for this in 2000. ICs are produced on a piece of semiconductor crystal (or chip) by a process called *photolithography*. Thus, the entire Information Technology (IT) industry hinges on semiconductors. Over the years, the complexity of ICs has increased while the size of its features continued to shrink. In the past five decades, a dramatic miniaturisation in computer technology has made modern day computers *faster and smaller*. In the 1970s, Gordon Moore, co-founder of INTEL, pointed out that the memory capacity of a chip (IC) approximately doubled every one and a half years. This is popularly known as *Moore's law*. The number of transistors per chip has risen exponentially and each year computers are becoming more powerful, yet cheaper than the year before. It is intimated from current trends that the computers available in 2020 will operate at 40 GHz (40,000 MHz) and would be much smaller, more efficient and less expensive than present day computers. The explosive growth in the semiconductor industry and computer technology is best expressed by a famous quote from Gordon Moore: "If the auto industry advanced as rapidly as the semiconductor industry, a Rolls Royce would get half a million miles per gallon, and it would be cheaper to throw it away than to park it".

SUMMARY

1. Semiconductors are the basic materials used in the present solid state electronic devices like diode, transistor, ICs, etc.
2. Lattice structure and the atomic structure of constituent elements decide whether a particular material will be insulator, metal or semiconductor.
3. Metals have low resistivity (10^{-2} to 10^{-8} Ωm), insulators have very high resistivity (>10^{8} $\Omega\,m^{-1}$), while semiconductors have intermediate values of resistivity.
4. Semiconductors are elemental (Si, Ge) as well as compound (GaAs, CdS, etc.).
5. Pure semiconductors are called 'intrinsic semiconductors'. The presence of charge carriers (electrons and holes) is an 'intrinsic' property of the material and these are obtained as a result of thermal excitation. The number of electrons (n_e) is equal to the number of holes (n_h) in intrinsic conductors. Holes are essentially electron vacancies with an effective positive charge.
6. The number of charge carriers can be changed by 'doping' of a suitable impurity in pure semiconductors. Such semiconductors are known as extrinsic semiconductors. These are of two types (n-type and p-type).
7. In n-type semiconductors, $n_e >> n_h$ while in p-type semiconductors $n_h >> n_e$.
8. n-type semiconducting Si or Ge is obtained by doping with pentavalent atoms (donors) like As, Sb, P, etc., while p-type Si or Ge can be obtained by doping with trivalent atom (acceptors) like B, Al, In etc.
9. $n_e n_h = n_i^2$ in all cases. Further, the material possesses an *overall charge neutrality*.
10. There are two distinct band of energies (called valence band and conduction band) in which the electrons in a material lie. Valence band energies are low as compared to conduction band energies. All energy levels in the valence band are filled while energy levels in the conduction band may be fully empty or partially filled. The electrons in the conduction band are free to move in a solid and are responsible for the conductivity. The extent of conductivity depends upon the energy gap (E_g) between the top of valence band (E_V) and the bottom of the conduction band E_C. The electrons from valence band can be excited by heat, light or electrical energy to the conduction band and thus, produce a change in the current flowing in a semiconductor.
11. For insulators E_g> 3 eV, for semiconductors E_g is 0.2 eV to 3 eV, while for metals $E_g \approx 0$.
12. p-n junction is the 'key' to all semiconductor devices. When such a junction is made, a 'depletion layer' is formed consisting of immobile ion-cores devoid of their electrons or holes. This is responsible for a junction potential barrier.
13. By changing the external applied voltage, junction barriers can be changed. In forward bias (n-side is connected to negative terminal of the battery and p-side is connected to the positive), the barrier is decreased while the barrier increases in reverse bias. Hence, forward bias current is more (mA) while it is very small (μA) in a p-n junction diode.
14. Diodes can be used for rectifying an ac voltage (restricting the ac voltage to one direction). With the help of a capacitor or a suitable filter, a dc voltage can be obtained.
15. There are some special purpose diodes.

16. Zener diode is one such special purpose diode. In reverse bias, after a certain voltage, the current suddenly increases (breakdown voltage) in a Zener diode. This property has been used to obtain *voltage regulation*.
17. p-n junctions have also been used to obtain many photonic or optoelectronic devices where one of the participating entity is 'photon': (a) Photodiodes in which photon excitation results in a change of reverse saturation current which helps us to measure light intensity; (b) Solar cells which convert photon energy into electricity; (c) Light Emitting Diode and Diode Laser in which electron excitation by a bias voltage results in the generation of light.
18. Transistor is an n-p-n or p-n-p junction device. The central block (thin and lightly doped) is called 'Base' while the other electrodes are 'Emitter' and 'Collectors'. The emitter-base junction is forward biased while collector-base junction is reverse biased.
19. The transistors can be connected in such a manner that either C or E or B is common to both the input and output. This gives the three configurations in which a transistor is used: Common Emitter (CE), Common Collector (CC) and Common Base (CB). The plot between I_C and V_{CE} for fixed I_B is called output characteristics while the plot between I_B and V_{BE} with fixed V_{CE} is called input characteristics. The important transistor parameters for CE-configuration are:

    input resistance, $r_i = \left(\frac{\Delta V_{BE}}{\Delta I_B}\right)_{V_{CE}}$

    output resistance, $r_o = \left(\frac{\Delta V_{CE}}{\Delta I_C}\right)_{I_B}$

    current amplification factor, $\beta = \left(\frac{\Delta I_C}{\Delta I_B}\right)_{V_{CE}}$
20. Transistor can be used as an amplifier and oscillator. In fact, an oscillator can also be considered as a self-sustained amplifier in which a part of output is fed-back to the input in the same phase (positive feed back). The voltage gain of a transistor amplifier in common emitter configuration is: $A_v = \left(\frac{v_o}{v_i}\right) = \beta\frac{R_C}{R_B}$, where R_C and R_B are respectively the resistances in collector and base sides of the circuit.
21. When the transistor is used in the cutoff or saturation state, it acts as a switch.
22. There are some special circuits which handle the digital data consisting of 0 and 1 levels. This forms the subject of Digital Electronics.
23. The important digital circuits performing special logic operations are called logic gates. These are: OR, AND, NOT, NAND, and NOR gates.
24. In modern day circuit, many logical gates or circuits are integrated in one single 'Chip'. These are known as Intgrated circuits (IC).

POINTS TO PONDER

1. The energy bands (E_C or E_V) in the semiconductors are space delocalised which means that these are not located in any specific place inside the solid. The energies are the overall averages. When you see a picture in which E_C or E_V are drawn as straight lines, then they should be respectively taken simply as the *bottom* of conduction band energy levels and *top* of valence band energy levels.

2. In elemental semiconductors (Si or Ge), the n-type or p-type semiconductors are obtained by introducing 'dopants' as defects. In compound semiconductors, the change in relative stoichiometric ratio can also change the type of semiconductor. For example, in ideal GaAs the ratio of Ga:As is 1:1 but in Ga-rich or As-rich GaAs it could respectively be $Ga_{1.1}\ As_{0.9}$ or $Ga_{0.9}\ As_{1.1}$. In general, the presence of defects control the properties of semiconductors in many ways.
3. In transistors, the base region is both narrow and lightly doped, otherwise the electrons or holes coming from the input side (say, emitter in CE-configuration) will not be able to reach the collector.
4. We have described an oscillator as a positive feedback amplifier. For stable oscillations, the voltage feedback (V_{fb}) from the output voltage (V_o) should be such that after amplification (A) it should again become V_o. If a fraction β' is feedback, then $V_{fb} = V_o.\ \beta'$ and after amplification its value $A(v_o.\beta')$ should be equal to V_o. This means that the criteria for stable oscillations to be sustained is $A\ \beta' = 1$. This is known as Barkhausen's Criteria.
5. In an oscillator, the feedback is in the same phase (positive feedback). If the feedback voltage is in opposite phase (negative feedback), the gain is less than 1 and it can never work as oscillator. It will be an amplifier with reduced gain. However, the negative feedback also reduces noise and distortion in an amplifier which is an advantageous feature.

EXERCISES

14.1 In an n-type silicon, which of the following statement is true:
(a) Electrons are majority carriers and trivalent atoms are the dopants.
(b) Electrons are minority carriers and pentavalent atoms are the dopants.
(c) Holes are minority carriers and pentavalent atoms are the dopants.
(d) Holes are majority carriers and trivalent atoms are the dopants.

14.2 Which of the statements given in Exercise 14.1 is true for p-type semiconductos.

14.3 Carbon, silicon and germanium have four valence electrons each. These are characterised by valence and conduction bands separated by energy band gap respectively equal to $(E_g)_C$, $(E_g)_{Si}$ and $(E_g)_{Ge}$. Which of the following statements is true?
(a) $(E_g)_{Si} < (E_g)_{Ge} < (E_g)_C$
(b) $(E_g)_C < (E_g)_{Ge} > (E_g)_{Si}$
(c) $(E_g)_C > (E_g)_{Si} > (E_g)_{Ge}$
(d) $(E_g)_C = (E_g)_{Si} = (E_g)_{Ge}$

14.4 In an unbiased p-n junction, holes diffuse from the p-region to n-region because
(a) free electrons in the n-region attract them.
(b) they move across the junction by the potential difference.
(c) hole concentration in p-region is more as compared to n-region.
(d) All the above.

14.5 When a forward bias is applied to a p-n junction, it
(a) raises the potential barrier.
(b) reduces the majority carrier current to zero.
(c) lowers the potential barrier.
(d) None of the above.

14.6 For transistor action, which of the following statements are correct:
(a) Base, emitter and collector regions should have similar size and doping concentrations.
(b) The base region must be very thin and lightly doped.
(c) The emitter junction is forward biased and collector junction is reverse biased.
(d) Both the emitter junction as well as the collector junction are forward biased.

14.7 For a transistor amplifier, the voltage gain
(a) remains constant for all frequencies.
(b) is high at high and low frequencies and constant in the middle frequency range.
(c) is low at high and low frequencies and constant at mid frequencies.
(d) None of the above.

14.8 In half-wave rectification, what is the output frequency if the input frequency is 50 Hz. What is the output frequency of a full-wave rectifier for the same input frequency.

14.9 For a CE-transistor amplifier, the audio signal voltage across the collected resistance of 2 kΩ is 2 V. Suppose the current amplification factor of the transistor is 100, find the input signal voltage and base current, if the base resistance is 1 kΩ.

14.10 A p-n photodiode is fabricated from a semiconductor with band gap of 2.8 eV. Can it detect a wavelength of 6000 nm?

ADDITIONAL EXERCISES

14.11 The number of silicon atoms per m^3 is 5×10^{28}. This is doped simultaneously with 5×10^{22} atoms per m^3 of Arsenic and 5×10^{20} per m^3 atoms of Indium. Calculate the number of electrons and holes. Given that $n_i = 1.5 \times 10^{16}\ m^{-3}$. Is the material n-type or p-type?

14.12 In an intrinsic semiconductor the energy gap E_g is 1.2eV. Its hole mobility is much smaller than electron mobility and independent of temperature. What is the ratio between conductivity at 600K and that at 300K? Assume that the temperature dependence of intrinsic carrier concentration n_i is given by

$$n_i = n_0 \exp\left(-\frac{E_g}{2k_B T}\right)$$

where n_0 is a constant.

14.13 In a p-n junction diode, the current I can be expressed as

$$I = I_0 \exp\left(\frac{eV}{2k_B T} - 1\right)$$

where I_0 is called the reverse saturation current, V is the voltage across the diode and is positive for forward bias and negative for reverse bias, and I is the current through the diode, k_B is the Boltzmann constant (8.6×10^{-5} eV/K) and T is the absolute temperature. If for a given diode $I_0 = 5 \times 10^{-12}$ A and T = 300 K, then

(a) What will be the forward current at a forward voltage of 0.6 V?

(b) What will be the increase in the current if the voltage across the diode is increased to 0.7 V?

(c) What is the dynamic resistance?

(d) What will be the current if reverse bias voltage changes from 1 V to 2 V?

14.14 You are given the two circuits as shown in Fig. 14.44. Show that circuit (a) acts as OR gate while the circuit (b) acts as AND gate.

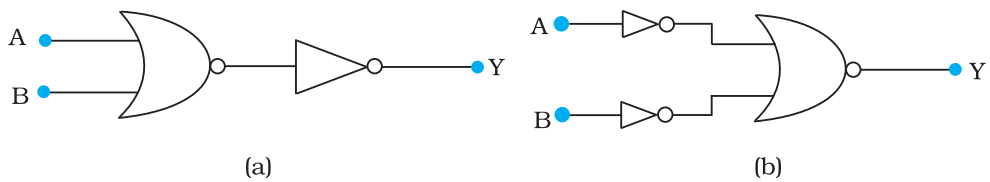


FIGURE 14.44

14.15 Write the truth table for a NAND gate connected as given in Fig. 14.45.

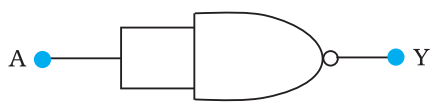


FIGURE 14.45

Hence identify the exact logic operation carried out by this circuit.

14.16 You are given two circuits as shown in Fig. 14.46, which consist of NAND gates. Identify the logic operation carried out by the two circuits.

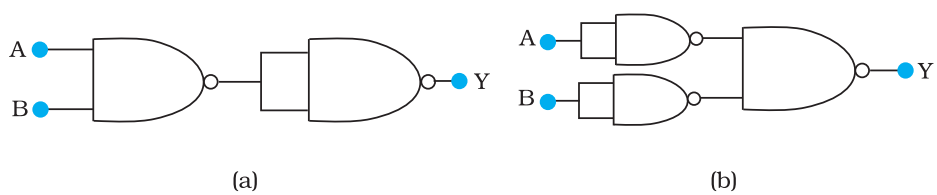


FIGURE 14.46

14.17 Write the truth table for circuit given in Fig. 14.47 below consisting of NOR gates and identify the logic operation (OR, AND, NOT) which this circuit is performing.

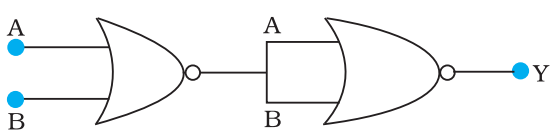


FIGURE 14.47

(Hint: A = 0, B = 1 then A and B inputs of second NOR gate will be 0 and hence Y=1. Similarly work out the values of Y for other combinations of A and B. Compare with the truth table of OR, AND, NOT gates and find the correct one.)

14.18 Write the truth table for the circuits given in Fig. 14.48 consisting of NOR gates only. Identify the logic operations (OR, AND, NOT) performed by the two circuits.

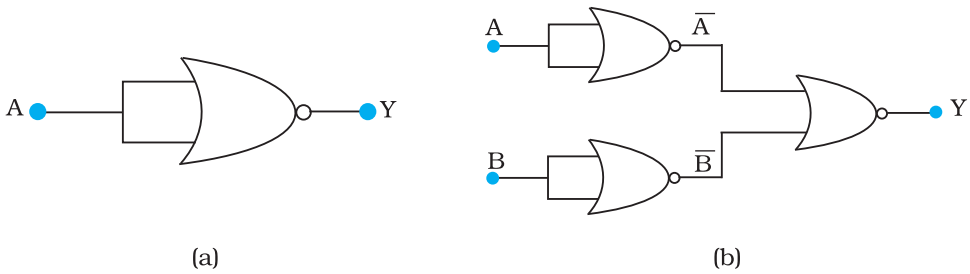


FIGURE 14.48

14.19 Two amplifiers are connected one after the other in series (cascaded). The first amplifier has a voltage gain of 10 and the second has a voltage gain of 20. If the input signal is 0.01 volt, calculate the output ac signal.

Chapter Fifteen

COMMUNICATION SYSTEMS

15.1 Introduction

Communication is the act of transmission of information. Every living creature in the world experiences the need to impart or receive information almost continuously with others in the surrounding world. For communication to be successful, it is essential that the sender and the receiver understand a common *language*. Man has constantly made endeavors to improve the quality of communication with other human beings. Languages and methods used in communication have kept evolving from prehistoric to modern times, to meet the growing demands in terms of speed and complexity of information. It would be worthwhile to look at the major milestones in events that promoted developments in communications, as presented in Table 15.1.

Modern communication has its roots in the 19th and 20th century in the work of scientists like J.C. Bose, F.B. Morse, G. Marconi and Alexander Graham Bell. The pace of development seems to have increased dramatically after the first half of the 20th century. We can hope to see many more accomplishments in the coming decades. The aim of this chapter is to introduce the concepts of communication, namely the mode of communication, the need for modulation, production and detection of amplitude modulation.

15.2 Elements of a Communication System

Communication pervades all stages of life of all living creatures. Irrespective of its nature, every communication system has three essential elements-

TABLE 15.1 SOME MAJOR MILESTONES IN THE HISTORY OF COMMUNICATION

Year	Event	Remarks
Around 1565 A.D.	The reporting of the delivery of a child by queen using drum beats from a distant place to King Akbar.	It is believed that minister Birbal experimented with the arrangement to decide the number of drummers posted between the place where the queen stayed and the place where the king stayed.
1835	Invention of telegraph by Samuel F.B. Morse and Sir Charles Wheatstone	It resulted in tremendous growth of messages through post offices and reduced physical travel of messengers considerably.
1876	Telephone invented by Alexander Graham Bell and Antonio Meucci	Perhaps the most widely used means of communication in the history of mankind.
1895	Jagadis Chandra Bose and Guglielmo Marconi demonstrated wireless telegraphy.	It meant a giant leap – from an era of communication using wires to communicating without using wires. (wireless)
1936	Television broadcast(John Logi Baird)	First television broadcast by BBC
1955	First radio FAX transmitted across continent.(Alexander Bain)	The idea of FAX transmission was patented by Alexander Bain in 1843.
1968	ARPANET- the first internet came into existence(J.C.R. Licklider)	ARPANET was a project undertaken by the U.S. defence department. It allowed file transfer from one computer to another connected to the network.
1975	Fiber optics developed at Bell Laboratories	Fiber optical systems are superior and more economical compared to traditional communication systems.
1989-91	Tim Berners-Lee invented the **World Wide Web**.	WWW may be regarded as the mammoth encyclopedia of knowledge accessible to everyone round the clock throughout the year.

transmitter, medium/channel and receiver. The block diagram shown in Fig. 15.1 depicts the general form of a communication system.

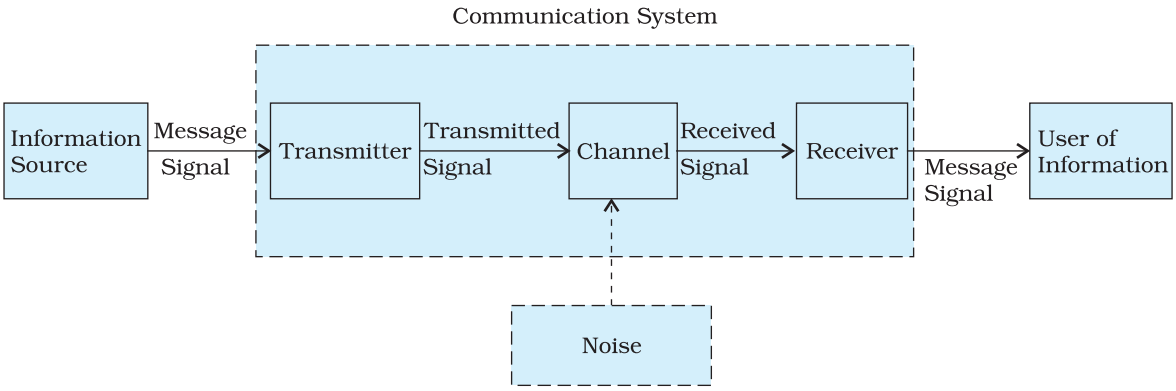


FIGURE 15.1 Block diagram of a generalised communication system.

In a communication system, the transmitter is located at one place, the receiver is located at some other place (far or near) separate from the transmitter and the channel is the physical medium that connects them. Depending upon the type of communication system, a channel may be in the form of wires or cables connecting the transmitter and the receiver or it may be wireless. The purpose of the transmitter is to convert the message signal produced by the source of information into a form suitable for transmission through the channel. If the output of the information source is a non-electrical signal like a voice signal, a transducer converts it to electrical form before giving it as an input to the transmitter. When a transmitted signal propagates along the channel it may get distorted due to channel imperfection. Moreover, noise adds to the transmitted signal and the receiver receives a corrupted version of the transmitted signal. The receiver has the task of operating on the received signal. It reconstructs a recognisable form of the original message signal for delivering it to the user of information.

There are two basic modes of communication: *point-to-point* and *broadcast*.

In point-to-point communication mode, communication takes place over a link between a single transmitter and a receiver. Telephony is an example of such a mode of communication. In contrast, in the broadcast mode, there are a large number of receivers corresponding to a single transmitter. Radio and television are examples of broadcast mode of communication.

15.3 Basic Terminology Used in Electronic Communication Systems

By now, we have become familiar with some terms like information source, transmitter, receiver, channel, noise, etc. It would be easy to understand the principles underlying any communication, if we get ourselves acquainted with the following basic terminology.

JAGADIS CHANDRA BOSE (1858 – 1937)

Jagadis Chandra Bose (1858 – 1937) He developed an apparatus for generating ultrashort electro-magnetic waves and studied their quasi-optical properties. He was said to be the first to employ a semiconductor like galena as a self-recovering detector of electromagnetic waves. Bose published three papers in the British magazine, 'The Electrician' of 27 Dec. 1895. His invention was published in the 'Proceedings of The Royal Society' on 27 April 1899 over two years before Marconi's first wireless communication on 13 December 1901. Bose also invented highly sensitive instruments for the detection of minute responses by living organisms to external stimulii and established parallelism between animal and plant tissues.

(i) *Transducer:* Any device that converts one form of energy into another can be termed as a transducer. In electronic communication systems, we usually come across devices that have either their inputs or outputs in the electrical form. An electrical transducer may be defined as a device that converts some physical variable (pressure, displacement, force, temperature, etc.) into corresponding variations in the electrical signal at its output.

(ii) *Signal:* Information converted in electrical form and suitable for transmission is called a signal. Signals can be either *analog or digital*. Analog signals are continuous variations of voltage or current. *They are essentially single-valued functions of time*. Sine wave is a fundamental analog signal. All other analog signals can be fully understood in terms of their sine wave components. Sound and picture signals in TV are analog in nature. Digital signals are those which can take only discrete stepwise values. Binary system that is extensively used in digital electronics employs just two levels of a signal. '0' corresponds to a low level and '1' corresponds to a high level of voltage/current. There are several coding schemes useful for digital communication. They employ suitable combinations of number systems such as the binary coded decimal (BCD)*. American Standard Code for Information Interchange (ASCII)** is a universally popular digital code to represent numbers, letters and certain characters.
(Nowadays, optical signals are also in use.)

(iii) *Noise:* Noise refers to the unwanted signals that tend to disturb the transmission and processing of message signals in a communication system. The source generating the noise may be located inside or outside the system.

(iv) *Transmitter:* A transmitter processes the incoming message signal so as to make it suitable for transmission through a channel and subsequent reception.

(v) *Receiver:* A receiver extracts the desired message signals from the received signals at the channel output.

(vi) *Attenuation:* The loss of strength of a signal while propagating through a medium is known as attenuation.

* In BCD, a digit is usually represented by four binary (0 or 1) bits. For example the numbers 0, 1, 2, 3, 4 in the decimal system are written as 0000, 0001, 0010, 0011 and 0100. 1000 would represent eight.

** It is a character encoding in terms of numbers based on English alphabet since the computer can only understand numbers.

(vii) *Amplification:* It is the process of *increasing the amplitude* (and consequently the strength) of a signal using an electronic circuit called the amplifier (reference Chapter 14). Amplification is necessary to compensate for the attenuation of the signal in communication systems. The energy needed for additional signal strength is obtained from a DC power source. Amplification is done at a place between the source and the destination wherever signal strength becomes weaker than the required strength.

(viii) *Range:* It is the largest distance between a source and a destination up to which the signal is received with sufficient strength.

(ix) *Bandwidth:* Bandwidth refers to the frequency range over which an equipment operates or the portion of the spectrum occupied by the signal.

(x) *Modulation:* The original low frequency message/information signal cannot be transmitted to long distances because of reasons given in Section 15.7. Therefore, at the transmitter, information contained in the low frequency message signal is superimposed on a high frequency wave, which acts as a carrier of the information. This process is known as modulation. As will be explained later, there are several types of modulation, abbreviated as AM, FM and PM.

(xi) *Demodulation:* The process of retrieval of information from the carrier wave at the receiver is termed demodulation. This is the reverse process of modulation.

(xii) *Repeater:* A repeater is a combination of a receiver and a transmitter. A repeater, picks up the signal from the transmitter, amplifies and retransmits it to the receiver sometimes with a change in carrier frequency. Repeaters are used to extend the range of a communication system as shown in Fig. 15.2. A communication satellite is essentially a repeater station in space.

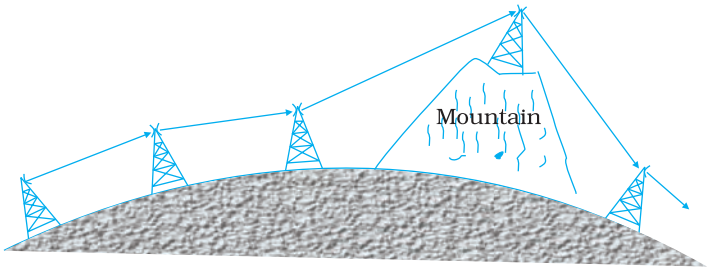


FIGURE 15.2 Use of repeater station to increase the range of communication.

15.4 Bandwidth of Signals

In a communication system, the message signal can be voice, music, picture or computer data. Each of these signals has different ranges of frequencies. The type of communication system needed for a given signal depends on the band of frequencies which is considered essential for the communication process.

For speech signals, frequency range 300 Hz to 3100 Hz is considered adequate. Therefore speech signal requires a bandwidth of 2800 Hz (3100 Hz – 300 Hz) for commercial telephonic communication. To transmit music,

an approximate bandwidth of 20 kHz is required because of the high frequencies produced by the musical instruments. The audible range of frequencies extends from 20 Hz to 20 kHz.

Video signals for transmission of pictures require about 4.2 MHz of bandwidth. A TV signal contains both voice and picture and is usually allocated 6 MHz of bandwidth for transmission.

In the preceeding paragraph, we have considered only analog signals. Digital signals are in the form of rectangular waves as shown in Fig. 15.3. One can show that this rectangular wave can be decomposed into a superposition of sinusoidal waves of frequencies ν_0, $2\nu_0$, $3\nu_0$, $4\nu_0$ ... $n\nu_0$ where n is an integer extending to infinity and $\nu_0 = 1/T_0$. The fundamental (ν_0), fundamental (ν_0) + second harmonic ($2\nu_0$), and fundamental (ν_0) + second harmonic ($2\nu_0$) + third harmonic ($3\nu_0$), are shown in the same figure to illustrate this fact. It is clear that to reproduce the rectangular wave shape exactly we need to superimpose all the harmonics ν_0, $2\nu_0$, $3\nu_0$, $4\nu_0$..., which implies an infinite bandwidth. However, for practical purposes, the contribution from higher harmonics can be neglected, thus limiting the bandwidth. As a result, received waves are a distorted version of the transmitted one. If the bandwidth is large enough to accommodate a few harmonics, the information is not lost and the rectangular signal is more or less recovered. This is so because the higher the harmonic, less is its contribution to the wave form.

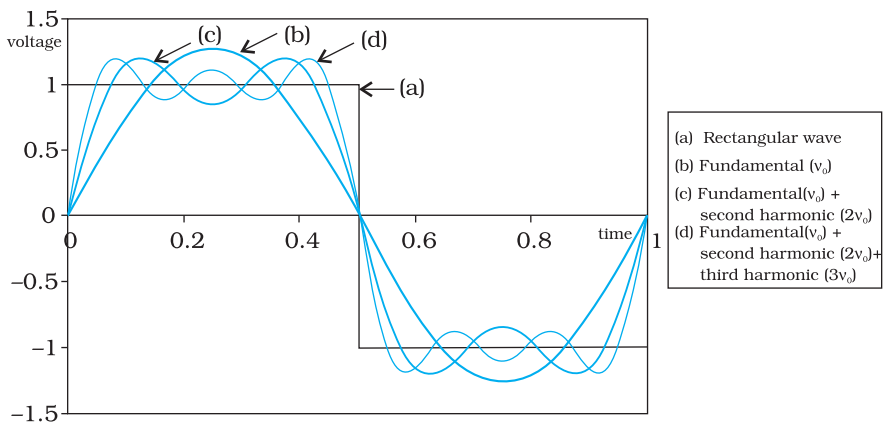


FIGURE 15.3 Approximation of a rectangular wave in terms of a fundamental sine wave and its harmonics.

15.5 Bandwidth of Transmission Medium

Similar to message signals, different types of transmission media offer different bandwidths. The commonly used transmission media are wire, free space and fiber optic cable. Coaxial cable is a widely used wire medium, which offers a bandwidth of approximately 750 MHz. Such cables are normally operated below 18 GHz. Communication through free space using radio waves takes place over a very wide range of frequencies: from a few hundreds of kHz to a few GHz. This range of frequencies is further subdivided and allocated for various services as indicated in Table 15.2. Optical communication using fibers is performed in the frequency range of 1 THz to 1000 THz (microwaves to ultraviolet). An optical fiber can offer a transmission bandwidth in excess of 100 GHz.

Spectrum allocations are arrived at by an international agreement. The International Telecommunication Union (ITU) administers the present system of frequency allocations.

TABLE 15.2 SOME IMPORTANT WIRELESS COMMUNICATION FREQUENCY BANDS

Service	Frequency bands	Comments
Standard AM broadcast	540-1600 kHz	
FM broadcast	88-108 MHz	
Television	54-72 MHz	VHF (very high frequencies)
	76-88 MHz	TV
	174-216 MHz	UHF (ultra high frequencies)
	420-890 MHz	TV
Cellular Mobile Radio	896-901 MHz	Mobile to base station
	840-935 MHz	Base station to mobile
Satellite Communication	5.925-6.425 GHz	Uplink
	3.7-4.2 GHz	Downlink

15.6 PROPAGATION OF ELECTROMAGNETIC WAVES

In communication using radio waves, an antenna at the transmitter radiates the Electromagnetic waves (em waves), which travel through the space and reach the receiving antenna at the other end. As the em wave travels away from the transmitter, the strength of the wave keeps on decreasing. Several factors influence the propagation of em waves and the path they follow. At this point, it is also important to understand the composition of the earth's atmosphere as it plays a vital role in the propagation of em waves. A brief discussion on some useful layers of the atmosphere is given in Table 15.3.

15.6.1 Ground wave

To radiate signals with high efficiency, the antennas should have a size comparable to the wavelength λ of the signal (at least ~ $\lambda/4$). At longer wavelengths (i.e., at lower frequencies), the antennas have large physical size and they are located on or very near to the ground. In standard AM broadcast, ground based vertical towers are generally used as transmitting antennas. For such antennas, ground has a strong influence on the propagation of the signal. The mode of propagation is called surface wave propagation and the wave glides over the surface of the earth. A wave induces current in the ground over which it passes and it is attenuated as a result of absorption of energy by the earth. The attenuation of surface waves increases very rapidly with increase in frequency. The maximum range of coverage depends on the transmitted power and frequency (less than a few MHz).

TABLE 15.3 DIFFERENT LAYERS OF ATMOSPHERE AND THEIR INTERACTION WITH THE PROPAGATING ELECTROMAGNETIC WAVES

Name of the stratum (layer)		Approximate height over earth's surface	Exists during	Frequencies most affected
Troposphere		10 km	Day and night	VHF (up to several GHz)
D (part of stratosphere)	PARTS OF IONOSPHERE	65-75 km	Day only	Reflects LF, absorbs MF and HF to some degree
E (part of Stratosphere)		100 km	Day only	Helps surface waves, reflects HF
F_1 (Part of Mesosphere)		170-190 km	Daytime, merges with F_2 at night	Partially absorbs HF waves yet allowing them to reach F_2
F_2 (Thermosphere)		300 km at night, 250-400 km during daytime	Day and night	Efficiently reflects HF waves, particularly at night

15.6.2 Sky waves

In the frequency range from a few MHz up to 30 to 40 MHz, long distance communication can be achieved by ionospheric reflection of radio waves back towards the earth. This mode of propagation is called *sky wave propagation* and is used by short wave broadcast services. The ionosphere is so called because of the presence of a large number of ions or charged particles. It extends from a height of ~ 65 Km to about 400 Km above the earth's surface. Ionisation occurs due to the absorption of the ultraviolet and other high-energy radiation coming from the sun by air molecules. The ionosphere is further subdivided into several layers, the details of which are given in Table 15.3. The degree of ionisation varies with the height. The density of atmosphere decreases with height. At great heights the solar radiation is intense but there are few molecules to be ionised. Close to the earth, even though the molecular concentration is very high, the radiation intensity is low so that the ionisation is again low. However, at some intermediate heights, there occurs a peak of ionisation density. The ionospheric layer acts as a reflector for a certain range of frequencies (3 to 30 MHz). Electromagnetic waves of frequencies higher than 30 MHz penetrate the ionosphere and escape. These phenomena are shown in the Fig. 15.4. The phenomenon of bending of em waves so that they are diverted towards the earth is similar to total internal reflection in optics*.

* Compare this with the phenomenon of mirage.

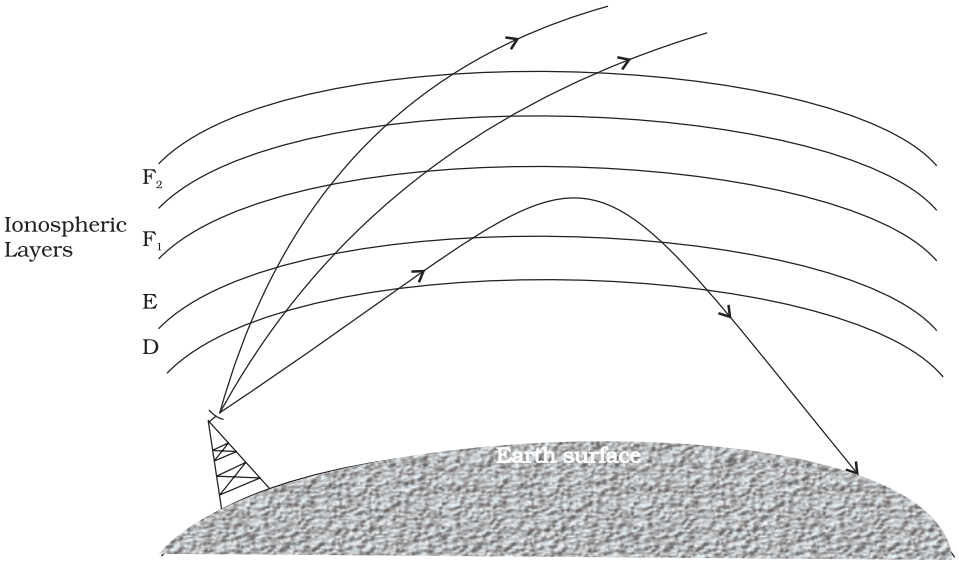


FIGURE 15.4 Sky wave propagation. The layer nomenclature is given in Table 15.3.

15.6.3 Space wave

Another mode of radio wave propagation is by *space waves*. A space wave travels in a straight line from transmitting antenna to the receiving antenna. Space waves are used for line-of-sight (LOS) communication as well as satellite communication. At frequencies above 40 MHz, communication is essentially limited to line-of-sight paths. At these frequencies, the antennas are relatively smaller and can be placed at heights of many wavelengths above the ground. Because of line-of-sight nature of propagation, direct waves get blocked at some point by the curvature of the earth as illustrated in Fig. 15.5. If the signal is to be received beyond the horizon then the receiving antenna must be high enough to intercept the line-of-sight waves.

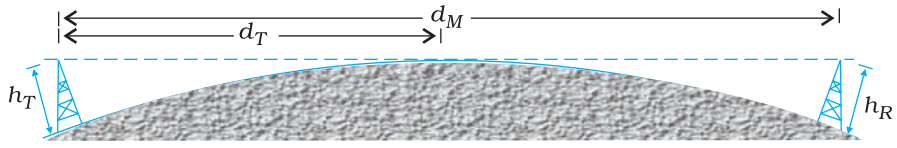


FIGURE 15.5 Line of sight communication by space waves.

If the transmitting antenna is at a height h_T, then you can show that the distance to the horizon d_T is given as $d_T = \sqrt{2Rh_T}$, where R is the radius of the earth (approximately 6400 km). d_T is also called the radio horizon of the transmitting antenna. With reference to Fig. 15.5 the maximum line-of-sight distance d_M between the two antennas having heights h_T and h_R above the earth is given by

$$d_M = \sqrt{2Rh_T} + \sqrt{2Rh_R} \qquad (15.1)$$

where h_R is the height of receiving antenna.

Television broadcast, microwave links and satellite communication are some examples of communication systems that use space wave mode of propagation. Figure 15.6 summarises the various modes of wave propagation discussed so far.

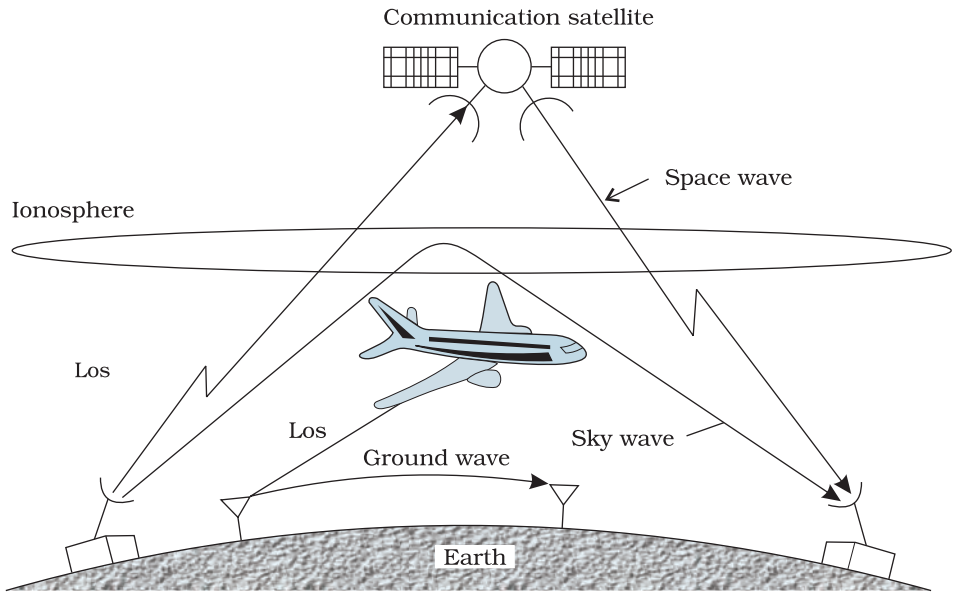


FIGURE 15.6 Various propagation modes for em waves.

Example 15.1 A transmitting antenna at the top of a tower has a height 32 m and the height of the receiving antenna is 50 m. What is the maximum distance between them for satisfactory communication in LOS mode? Given radius of earth 6.4×10^6 m.

Solution

$$d_m = \sqrt{2\times 64\times 10^5 \times 32} + \sqrt{2\times 64\times 10^5 \times 50}\ \text{m}$$

$$= 64\times 10^2 \times \sqrt{10} + 8\times 10^3 \times \sqrt{10}\ \text{m}$$

$$= 144\times 10^2 \times \sqrt{10}\ \text{m} = 45.5\ \text{km}$$

EXAMPLE 15.1

15.7 Modulation and its Necessity

As already mentioned, the purpose of a communication system is to transmit information or message signals. Message signals are also called *baseband signals*, which essentially designate the band of frequencies representing the original signal, as delivered by the source of information. No signal, in general, is a single frequency sinusoid, but it spreads over a range of frequencies called the signal *bandwidth*. Suppose we wish to transmit an electronic signal in the audio frequency (AF) range (baseband signal frequency less than 20 kHz) over a long distance directly. Let us find what factors prevent us from doing so and how we overcome these factors.

15.7.1 Size of the antenna or aerial

For transmitting a signal, we need an antenna or an aerial. This antenna should have a size comparable to the wavelength of the signal (at least $\lambda/4$ in dimension) so that the antenna properly senses the time variation of the signal. For an electromagnetic wave of frequency 20 kHz, the wavelength λ is 15 km. Obviously, such a long antenna is not possible to construct and operate. Hence direct transmission of such baseband signals is not practical. We can obtain transmission with reasonable antenna lengths if transmission frequency is high (for example, if v is 1 MHz, then λ is 300 m). Therefore, there is a need of *translating the information contained in our original low frequency baseband signal into high or radio frequencies before transmission.*

15.7.2 Effective power radiated by an antenna

A theoretical study of radiation from a linear antenna (length l) shows that the power radiated is proportional to $(l/\lambda)^2$. This implies that for the same antenna length, the power radiated increases with decreasing λ, i.e., increasing frequency. Hence, the effective power radiated by a long wavelength baseband signal would be small. For a good transmission, we need high powers and hence this also points out to *the need* of using high frequency transmission.

15.7.3 Mixing up of signals from different transmitters

Another important argument against transmitting baseband signals directly is more *practical* in nature. Suppose many people are talking at the same time or many transmitters are transmitting baseband information signals simultaneously. All these signals will get mixed up and there is no simple way to distinguish between them. This points out towards a possible solution by using communication at high frequencies and allotting a *band* of frequencies to each message signal for its transmission.

The above arguments suggest that there is a *need for translating the original low frequency baseband message or information signal into high frequency wave before transmission such that the translated signal continues to possess the information contained in the original signal.* In doing so, we take the help of a high frequency signal, known as the *carrier wave*, and a process known as modulation which attaches information to it. The carrier wave may be continuous (sinusoidal) or in the form of pulses as shown in Fig. 15.7.

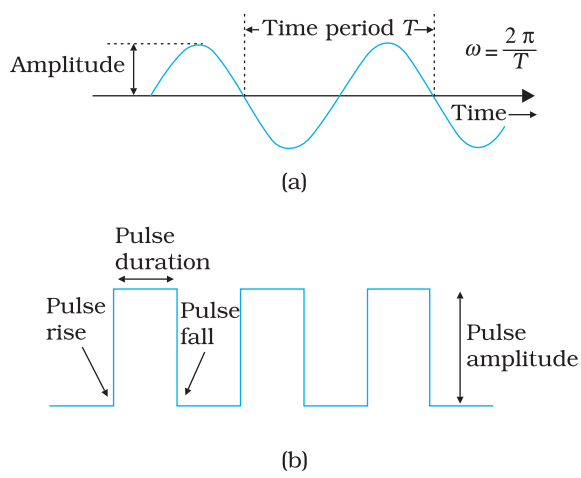


FIGURE 15.7 (a) Sinusoidal, and (b) pulse shaped signals.

A sinusoidal carrier wave can be represented as

$$c(t) = A_c \sin(\omega_c t + \phi) \qquad (15.2)$$

where $c(t)$ is the signal strength (voltage or current), A_c is the amplitude, ω_c ($= 2\pi v_c$) is the angular frequency and ϕ is the initial phase of the carrier wave. During the process of modulation, any of the three parameters, *viz* A_c, ω_c and ϕ, of the carrier wave can be controlled by the message or

information signal. This results in three types of modulation: (i) Amplitude modulation (AM), (ii) Frequency modulation (FM) and (iii) Phase modulation (PM), as shown in Fig. 15.8.

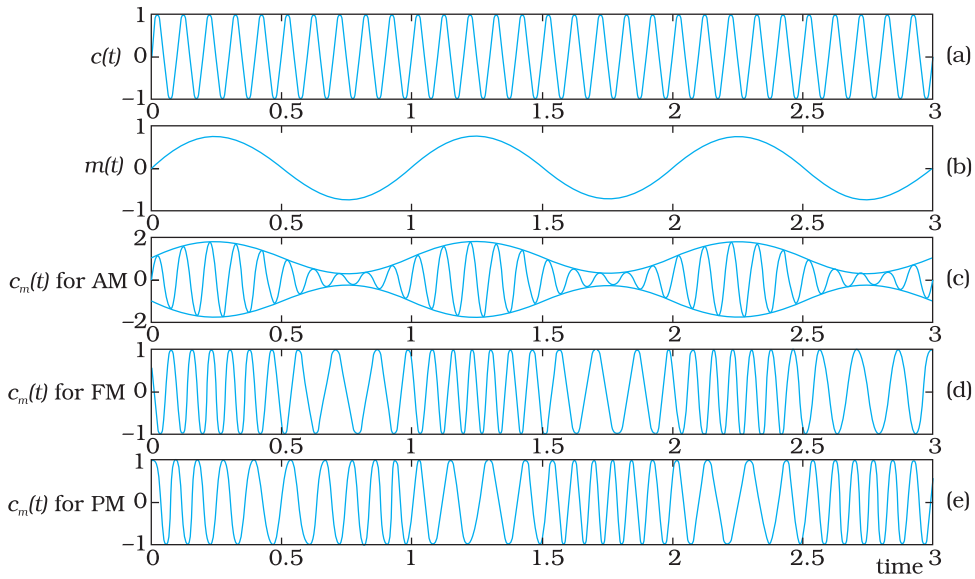


FIGURE 15.8 Modulation of a carrier wave: (a) a sinusoidal carrier wave; (b) a modulating signal; (c) amplitude modulation; (d) frequency modulation; and (e) phase modulation.

Similarly, the significant characteristics of a pulse are: pulse amplitude, pulse duration or pulse Width, and pulse position (denoting the time of *rise* or *fall* of the pulse amplitude) as shown in Fig. 15.7(b). Hence, different types of pulse modulation are: (a) pulse amplitude modulation (PAM), (b) pulse duration modulation (PDM) or pulse width modulation (PWM), and (c) pulse position modulation (PPM). In this chapter, we shall confine to amplitude modulation only.

15.8 Amplitude Modulation

In amplitude modulation the amplitude of the carrier is varied in accordance with the information signal. Here we explain amplitude modulation process using a sinusoidal signal as the modulating signal.

Let $c(t) = A_c \sin \omega_c t$ represent carrier wave and $m(t) = A_m \sin \omega_m t$ represent the message or the modulating signal where $\omega_m = 2\pi f_m$ is the angular frequency of the message signal. The modulated signal $c_m(t)$ can be written as

$$c_m(t) = (A_c + A_m \sin \omega_m t) \sin \omega_c t$$

$$= A_c \left(1 + \frac{A_m}{A_c} \sin\omega_m t\right) \sin\omega_c t \qquad (15.3)$$

Note that the modulated signal now contains the message signal. This can also be seen from Fig. 15.8(c). From Eq. (15.3), we can write,

$$c_m(t) = A_c \sin\omega_c t + \mu A_c \sin\omega_m t \sin\omega_c t \qquad (15.4)$$

Modulation and Demodulation
http://iitg.vlab.co.in/?sub=59&brch=163&sim=259&cnt=358
PHYSICS

Here $\mu = A_m/A_c$ is the *modulation index;* in practice, μ is kept ≤ 1 to avoid distortion.

Using the trignomatric relation $\sin A \sin B = ½ (\cos(A - B) - \cos (A + B)$, we can write $c_m(t)$ of Eq. (15.4) as

$$c_m(t) = A_c \sin \omega_c t + \frac{\mu A_c}{2} \cos(\omega_c - \omega_m)\, t - \frac{\mu A_c}{2} \cos(\omega_c + \omega_m)t \qquad (15.5)$$

Here $\omega_c - \omega_m$ and $\omega_c + \omega_m$ are respectively called the lower side and upper side frequencies. The modulated signal now consists of the carrier wave of frequency ω_c plus two sinusoidal waves each with a frequency slightly different from, known as side bands. The frequency spectrum of the amplitude modulated signal is shown in Fig. 15.9.

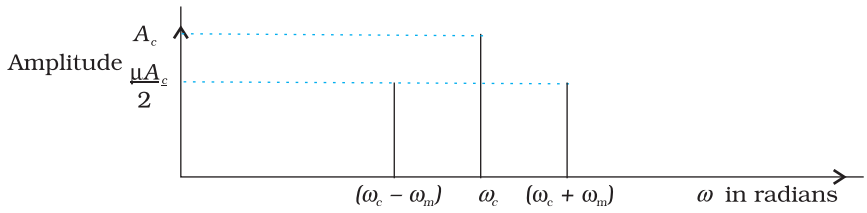


FIGURE 15.9 A plot of amplitude versus ω for an amplitude modulated signal.

As long as the broadcast frequencies (carrier waves) are sufficiently spaced out so that sidebands do not overlap, different stations can operate without interfering with each other.

Example 15.2 A message signal of frequency 10 kHz and peak voltage of 10 volts is used to modulate a carrier of frequency 1 MHz and peak voltage of 20 volts. Determine (a) modulation index, (b) the side bands produced.

Solution

(a) Modulation index =10/20 = 0.5

(b) The side bands are at (1000+10 kHz)=1010 kHz and (1000 –10 kHz) = 990 kHz.

EXAMPLE 15.2

15.9 Production of Amplitude Modulated Wave

Amplitude modulation can be produced by a variety of methods. A conceptually simple method is shown in the block diagram of Fig. 15.10.

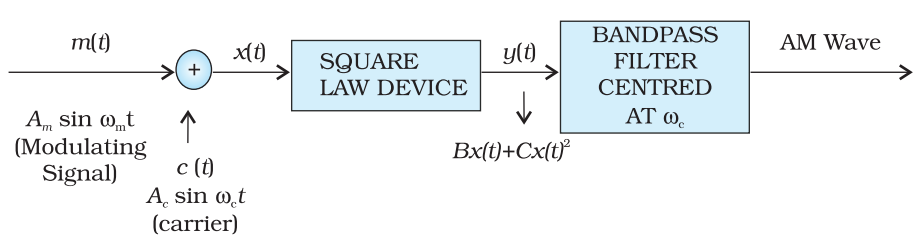


FIGURE 15.10 Block diagram of a simple modulator for obtaining an AM signal.

Here the modulating signal $A_m \sin \omega_m t$ is added to the carrier signal $A_c \sin \omega_c t$ to produce the signal $x(t)$. This signal $x(t) = A_m \sin\omega_m t + A_c \sin\omega_c t$ is passed through a square law device which is a non-linear device which produces an output

$$y(t) = B\,x(t) + Cx^2(t) \tag{15.6}$$

where B and C are constants. Thus,

$$y(t) = BA_m \sin \omega_m t + BA_c \sin \omega_c t$$

$$+C\ A_m^2 \sin^2 \omega_m t + A_c^2 \sin^2 \omega_c t + 2A_m A_c \sin \omega_m t \sin \omega_c t \tag{15.7}$$

$$= BA_m \sin \omega_m t + BA_c \sin \omega_c t$$

$$+\frac{C\,A_m^2}{2} + \frac{C\,A_c^2}{2} - \frac{C\,A_m^2}{2}\cos 2\omega_m t - \frac{C\,A_c^2}{2}\cos 2\omega_c t$$

$$+ CA_m A_c \cos(\omega_c - \omega_m)\,t - CA_m A_c \cos(\omega_c + \omega_m)\,t \tag{15.8}$$

where the trigonometric relations $\sin^2 A = (1 - \cos 2A)/2$ and the relation for $\sin A \sin B$ mentioned earlier are used.

In Eq. (15.8), there is a dc term $C/2\left(A_m^2 + A_c^2\right)$ and sinusoids of frequencies ω_m, $2\omega_m$, ω_c, $2\omega_c$, $\omega_c - \omega_m$ and $\omega_c + \omega_m$. As shown in Fig. 15.10 this signal is passed through a band pass filter* which rejects dc and the sinusoids of frequencies ω_m, $2\omega_m$ and $2\,\omega_c$ and retains the frequencies ω_c, $\omega_c - \omega_m$ and $\omega_c + \omega_m$. The output of the band pass filter therefore is of the same form as Eq. (15.5) and is therefore an AM wave.

It is to be mentioned that the modulated signal cannot be transmitted as such. The modulator is to be followed by a power amplifier which provides the necessary power and then the modulated signal is fed to an antenna of appropriate size for radiation as shown in Fig. 15.11.

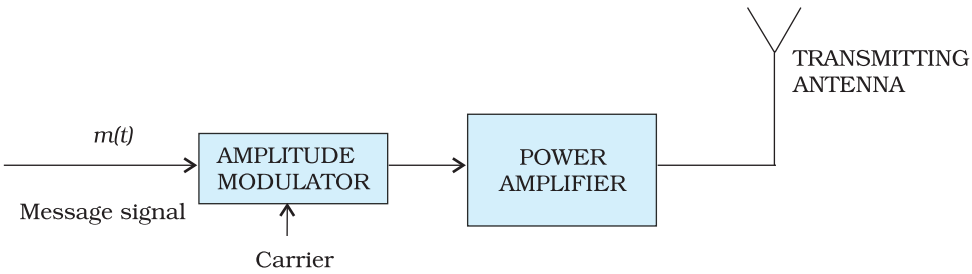


FIGURE 15.11 Block diagram of a transmitter.

15.10 Detection of Amplitude Modulated Wave

The transmitted message gets attenuated in propagating through the channel. The receiving antenna is therefore to be followed by an amplifier and a detector. In addition, to facilitate further processing, the carrier frequency is usually changed to a lower frequency by what is called an *intermediate frequency (IF) stage* preceding the detection. The detected signal may not be strong enough to be made use of and hence is required

* A band pass filter rejects low and high frequencies and allows a band of frequencies to pass through.

to be amplified. A block diagram of a typical receiver is shown in Fig. 15.12

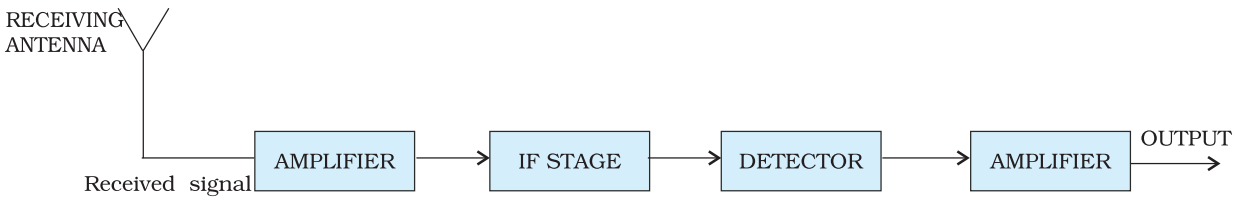


FIGURE 15.12 Block diagram of a receiver.

Detection is the process of recovering the modulating signal from the modulated carrier wave. We just saw that the modulated carrier wave contains the frequencies ω_c and $\omega_c \pm \omega_m$. In order to obtain the original message signal $m(t)$ of angular frequency ω_m, a simple method is shown in the form of a block diagram in Fig. 15.13.

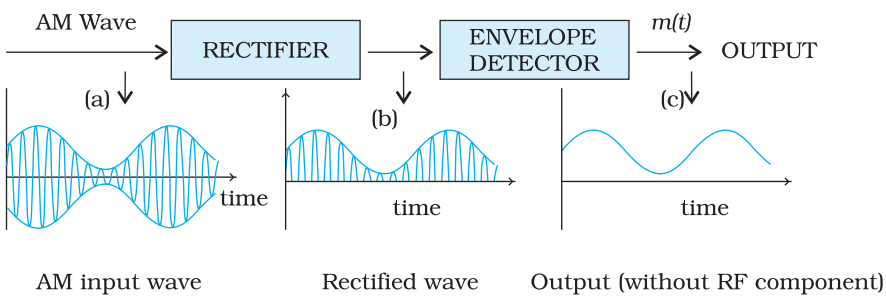


FIGURE 15.13 Block diagram of a detector for AM signal. The quantity on y-axis can be current or voltage.

The modulated signal of the form given in (a) of fig. 15.13 is passed through a rectifier to produce the output shown in (b). This envelope of signal (b) is the message signal. In order to retrieve $m(t)$, the signal is passed through an envelope detector (which may consist of a simple RC circuit).

In the present chapter we have discussed some basic concepts of communication and communication systems. We have also discussed one specific type of analog modulation namely Amplitude Modulation (AM). Other forms of modulation and digital communication systems play an important role in modern communication. These and other exciting developments are taking place everyday.

So far we have restricted our discussion to some basic communication systems. Before we conclude this chapter, it is worth taking a glance at some of the communication systems (see the box) that in recent times have brought major changes in the way we exchange information even in our day-to-day life:

Additional Information

The Internet

It is a system with billions of users worldwide. It permits communication and sharing of all types of information between any two or more computers connected through a large and complex network. It was started in 1960's and opened for public use in 1990's. With the passage of time it has witnessed tremendous growth and it is still expanding its reach. Its applications include

(i) *E mail* – It permits exchange of text/graphic material using email software. We can write a letter and send it to the recipient through ISP's (Internet Service Providers) who work like the dispatching and receiving post offices.

(ii) *File transfer* – A FTP (File Transfer Programmes) allows transfer of files/software from one computer to another connected to the Internet.

(iii) *World Wide Web (WWW)* – Computers that store specific information for sharing with others provide *websites* either directly or through web service providers. Government departments, companies, NGO's (Non-Government Organisations) and individuals can post information about their activities for restricted or free use on their websites. This information becomes accessible to the users. Several search engines like Google, Yahoo! etc., help us in finding information by listing the related websites. *Hypertext* is a powerful feature of the web that automatically links relevant information from one page on the web to another using *HTML (hypertext markup language)*.

(iv) *E-commerce* – Use of the Internet to promote business using electronic means such as using credit cards is called E-commerce. Customers view images and receive all the information about various products or services of companies through their websites. They can do *on-line shopping* from home/office. Goods are dispatched or services are provided by the company through mail/courier.

(v) *Chat* – Real time conversation among people with common interests through typed messages is called chat. Everyone belonging to the *chat group* gets the message instantaneously and can respond rapidly.

Facsimile (FAX)

It scans the contents of a document (as an image, not text) to create electronic signals. These signals are then sent to the destination (another FAX machine) in an orderly manner using telephone lines. At the destination, the signals are reconverted into a replica of the original document. Note that FAX provides image of a static document unlike the image provided by television of objects that might be dynamic.

Mobile telephony

The concept of mobile telephony was developed first in 1970's and it was fully implemented in the following decade. The central concept of this system is to divide the service area into a suitable number of *cells* centred on an office called *MTSO (Mobile Telephone Switching Office)*. Each cell contains a low-power transmitter called a *base station* and caters to a large number of mobile receivers (popularly called cell phones). Each cell could have a service area of a few square kilometers or even less depending upon the number of customers. When a mobile receiver crosses the coverage area of one base station, it is necessary for the mobile user to be transferred to another base station. This procedure is called *handover* or *handoff*. This process is carried out very rapidly, to the extent that the consumer does not even notice it. Mobile telephones operate typically in the UHF range of frequencies (about 800-950 MHz).

SUMMARY

1. Electronic communication refers to the faithful transfer of information or message (available in the form of electrical voltage and current) from one point to another point.
2. Transmitter, transmission channel and receiver are three basic units of a communication system.
3. Two important forms of communication system are: Analog and Digital. The information to be transmitted is generally in continuous waveform for the former while for the latter it has only discrete or quantised levels.
4. Every message signal occupies a range of frequencies. The bandwidth of a message signal refers to the band of frequencies, which are necessary for satisfactory transmission of the information contained in the signal. Similarly, any practical communication system permits transmission of a range of frequencies only, which is referred to as the bandwidth of the system.
5. Low frequencies cannot be transmitted to long distances. Therefore, they are superimposed on a high frequency carrier signal by a process known as modulation.
6. In modulation, some characteristic of the carrier signal like amplitude, frequency or phase varies in accordance with the modulating or message signal. Correspondingly, they are called Amplitude Modulated (AM), Frequency Modulated (FM) or Phase Modulated (PM) waves.
7. Pulse modulation could be classified as: Pulse Amplitude Modulation (PAM), Pulse Duration Modulation (PDM) or Pulse Width Modulation (PWM) and Pulse Position Modulation (PPM).
8. For transmission over long distances, signals are radiated into space using devices called antennas. The radiated signals propagate as electromagnetic waves and the mode of propagation is influenced by the presence of the earth and its atmosphere. Near the surface of the earth, electromagnetic waves propagate as surface waves. Surface wave propagation is useful up to a few MHz frequencies.
9. Long distance communication between two points on the earth is achieved through reflection of electromagnetic waves by ionosphere. Such waves are called sky waves. Sky wave propagation takes place up to frequency of about 30 MHz. Above this frequency, electromagnetic waves essentially propagate as space waves. Space waves are used for line-of-sight communication and satellite communication.
10. If an antenna radiates electromagnetic waves from a height h_T, then the range d_T is given by $\sqrt{2Rh_T}$ where R is the radius of the earth.
11. Amplitude modulated signal contains frequencies $(\omega_c - \omega_m)$, ω_c and $(\omega_c + \omega_m)$.
12. Amplitude modulated waves can be produced by application of the message signal and the carrier wave to a non-linear device, followed by a band pass filter.
13. AM detection, which is the process of recovering the modulating signal from an AM waveform, is carried out using a rectifier and an envelope detector.

POINTS TO PONDER

1. In the process of transmission of message/ information signal, noise gets added to the signal anywhere between the information source and the receiving end. Can you think of some sources of noise?
2. In the process of modulation, new frequencies called sidebands are generated on either side (higher and lower than the carrier frequency) of the carrier by an amount equal to the highest modulating frequency. Is it possible to retrieve the message by transmitting (a) only the side bands, (b) only one side band?
3. In amplitude modulation, modulation index $\mu \leq 1$ is used. What will happen if $\mu > 1$?

EXERCISES

15.1 Which of the following frequencies will be suitable for beyond-the-horizon communication using sky waves?
(a) 10 kHz
(b) 10 MHz
(c) 1 GHz
(d) 1000 GHz

15.2 Frequencies in the UHF range normally propagate by means of:
(a) Ground waves.
(b) Sky waves.
(c) Surface waves.
(d) Space waves.

15.3 Digital signals
(i) do not provide a continuous set of values,
(ii) represent values as discrete steps,
(iii) can utilize binary system, and
(iv) can utilize decimal as well as binary systems.
Which of the above statements are true?
(a) (i) and (ii) only
(b) (ii) and (iii) only
(c) (i), (ii) and (iii) but not (iv)
(d) All of (i), (ii), (iii) and (iv).

15.4 Is it necessary for a transmitting antenna to be at the same height as that of the receiving antenna for line-of-sight communication? A TV transmitting antenna is 81m tall. How much service area can it cover if the receiving antenna is at the ground level?

15.5 A carrier wave of peak voltage 12V is used to transmit a message signal. What should be the peak voltage of the modulating signal in order to have a modulation index of 75%?

15.6 For an amplitude modulated wave, the maximum amplitude is found to be 10V while the minimum amplitude is found to be 2V. Determine the modulation index, μ.
What would be the value of μ if the minimum amplitude is zero volt?

15.7 Due to economic reasons, only the upper sideband of an AM wave is transmitted, but at the receiving station, there is a facility for generating the carrier. Show that if a device is available which can multiply two signals, then it is possible to recover the modulating signal at the receiver station.

15.8 A modulating signal is a square wave, as shown in Fig. 15.14.

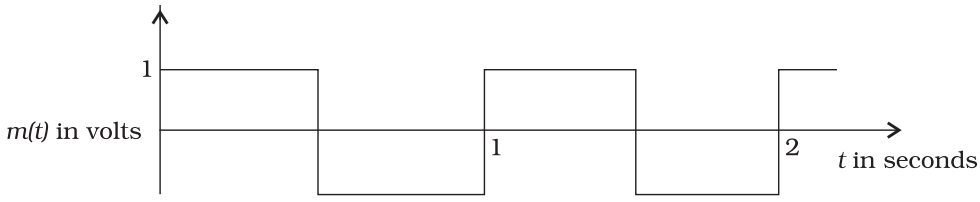


FIGURE 15.14

The carrier wave is given by $c(t) = 2\sin(8\pi t)$ volts.

(i) Sketch the amplitude modulated waveform

(ii) What is the modulation index?