Physics for Computer Science Students

Narciso Garcia
Arthur Damask

Physics for Computer Science Students

With Emphasis on Atomic and Semiconductor Physics

Springer-Verlag
New York Berlin Heidelberg London Paris
Tokyo Hong Kong Barcelona Budapest

Narciso Garcia
Arthur Damask
Physics Department
Queen's College
of the City University of New York
Flushing, NY 11367
USA

Library of Congress Cataloging in Publication Data
Garcia, Narciso
Physics for computer science students.
Includes index.
1. Physics 2. Computers. I. Damask, A. C. II. Title.
QC21.2.G32 530 91-31527

Printed on acid-free paper.

© 1991 Springer-Verlag New York Inc.
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

9 8 7 6 5 4 3 2 1

ISBN-13: 978-0-387-97656-3 e-ISBN-13: 978-1-4684-0421-0
DOI: 10.1007/978-1-4684-0421-0

Preface

This text is the product of several years' effort to develop a course to fill a specific educational gap. It is our belief that computer science students should know how a computer works, particularly in light of rapidly changing technologies. The text was designed for computer science students who have a calculus background but have not necessarily taken prior physics courses. However, it is clearly not limited to these students. Anyone who has had first-year physics can start with Chapter 17. This includes all science and engineering students who would like a survey course of the ideas, theories, and experiments that made our modern electronics age possible.

This textbook is meant to be used in a two-semester sequence. Chapters 1 through 16 can be covered during the first semester, and Chapters 17 through 28 in the second semester. At Queens College, where preliminary drafts have been used, the material is presented in three lecture periods (50 minutes each) and one recitation period per week, 15 weeks per semester. The lecture and recitation are complemented by a two-hour laboratory period per week for the first semester and a two-hour laboratory period biweekly for the second semester.

Computer Science students are usually required to take one year of physics; therefore, we examined conventional first-year physics courses with care so as to delete topics not relevant to the physics of semiconductors and logic circuits (such as Bernoulli's equation or nuclear decay rates). To achieve this contraction with some precision, we developed the second semester, Chapters 17–28, and taught these chapters for two years before writing Chapters 1–16 for the first semester.

The first-semester chapters include the First Principles of physics. The student normally develops the concepts of sliding and colliding blocks because these can be readily visualized and the principles understood from common experience. These concepts are later applied to electrons and holes, which are not readily visualized; hence the early foundation permits the understanding of more complex conceptual models.

The second part of this textbook begins with a presentation of some of the phenomena that led to the breakdown of classical physics and the advent of the quantum in blackbody radiation, photoelectric effect, electromagnetic spectrum of atoms, and so forth (Chapters 17 and 18). After laying down the fundamental principles of wave mechanics, de Broglie's hypothesis, and the Uncertainty Principle, we introduce the student to the Schrödinger theory of quantum physics. This theory is applied to the simple example of the infinite potential well (Chapters 19 and 20). An outline of the solution of the Schrödinger equation (i.e., the three quantum numbers n, l, m_l) and some of the features of the wavefunctions are presented in Chapter 21. After investigating

the need for the electron spin, we use these results, together with Pauli's exclusion principle, to determine the electron configurations and some of the properties of multi-electron atoms.

The student is then ready to study the electrical properties of solids. A brief discussion of the crystal structure and bonding mechanisms (Chapter 22) precedes the presentation of the classical and quantum free electron theories and their successes and shortcomings (Chapter 23). In order to explain the large differences in the electrical properties of solids as well as the peculiar properties of semiconductors, the existence of allowed and forbidden energy bands is investigated (Chapter 24). In this chapter, we introduce the concepts of the electron effective mass and of holes. Intrinsic and doped semiconductors, their electron and hole densities, and their electrical properties are discussed in Chapter 25.

It is now a rather simple matter for the student to understand the behavior and the characteristics of semiconductor devices: diodes, bipolar transistors, field effect transistors, etc. Semiconductor devices are the subject of Chapter 26. The text concludes with two chapters unique to this physics textbook. In Chapter 27, we show how diodes and transistors can be used to construct the logic circuits (gates) that constitute the fundamental building blocks of the computer. Chapter 28 is a layman's introduction to some of the techniques used in the fabrication of integrated circuits.

The laboratory experiments for the first semester are standard in any physics department, and thus we do not feel that is is necessary to include them in this book. The seven experiments in the second semester have been designed to illustrate important measurements; they are not standard in most physics courses. The equipment needed for these experiments is not generally available in quantity in physics departments but can be ordered from standard suppliers or constructed in school shops. For this reason the seven laboratory experiments are described in the instructor's manual that complements this textbook.

Although some of the topics and the level of treatment of Chapters 20–27 may appear to be potentially formidable for a first course in physics, in practice they have proved not to be so. We have tried to soften the impact of mathematical complexities with intuitive physical arguments. For example, the existence of energy bands in solids is first introduced by outlining the solution of the Schrödinger equation for an electron moving in the periodic potential of the ions. The student is shown that the imposition of the boundary conditions leads to a transcendental equation for the dispersion relation whose numerical solution leads in turn to the existence of forbidden and allowed energy bands. (Students are encouraged to use their computer programming skills to solve this equation.) This presentation is followed by an alternative approach that relies on simple intuitive arguments; here, the mathematics is kept to a minimum. The student is shown how, when two atoms are brought together to form a molecule, two separate energy levels are formed from each level of the individual atom. This is then extended to a situation where a large number of atoms come together to form a solid. It is thus relatively simple

to show that each atomic level becomes a band of very closely spaced energy levels.

By blending mathematical and physical arguments, we believe we have achieved a thorough presentation of the concepts of solid state physics while staying within the reach of students with an intermediate background in mathematics.

Narciso Garcia

Arthur C. Damask

Acknowledgments

The course for which this text was developed was conceived by Professor Joseph Klarfeld of the Physics Department and Professor Jacob Rootenberg, Chairman of the Computer Science Department of Queens College. Many useful suggestions on style, pace, and content were made by Professor Arthur Paskin of the Physics Department, who taught the course for several trial semesters. Topics requiring fuller explanation and wording requiring greater clarity were pointed out by Jay Damask, a high school junior with a knowledge of elementary calculus.

We appreciate the thorough work and many suggestions of the reviewers of this book: Professor Joseph I. Budnick-University of Connecticut, CT, Professor William Faissler-Northeastern University, MA, Professor David B. Fenner-University of Santa Clara, CA, Professor H. L. Helfer-The University of Rochester, NY, Professor Marvin D. Kemple-Purdue University, IN, and Professor Lawrence A. Mink-Arkansas State University, AR. The manuscript was typed and prepared for the publisher by Mrs. Marion Gaffga of Spring Hill, Florida. Mrs. Shirley Allen and Mrs. Susan Wasserman of Queens College contributed to the typing.

N. G.
A. C. D.

Contents

CHAPTER 1 ***Physical Quantities*** **1**

1.1 Introduction 2
1.2 Quantities and Units 3
1.3 Powers of 10 5
1.4 Accuracy of Numbers 6

CHAPTER 2 ***Vectors*** **9**

2.1 Introduction 10
2.2 Vector Components 10
2.3 Unit Vectors 14
2.4 Dot Product 15
2.5 Cross Product 16

CHAPTER 3 ***Uniformly Accelerated Motion*** **21**

3.1 Introduction 22
3.2 Speed and Velocity 22
3.3 Acceleration 24
3.4 Linear Motion 25
3.5 Projectile Motion 30

CHAPTER 4 ***Newton's Laws*** **37**

4.1 Introduction 38
4.2 Newton's Laws 38
4.3 Mass 41
4.4 Weight 42
4.5 Applications of Newton's Laws 43
4.6 Friction 48

CHAPTER 5 ***Work, Energy and Power*** **53**

5.1 Introduction 54
5.2 Work 54
5.3 Potential Energy 56
5.4 Work Done by a Variable Force 57
5.5 Kinetic Energy 58
5.6 Energy Conservation 59
5.7 Power 62

CHAPTER 6 ***Momentum and Collisions*** **67**
6.1 Introduction 68
6.2 Center of Mass 68
6.3 Motion of the Center of Mass 71
6.4 Momentum and its Conservation 72
6.5 Collisions 74

CHAPTER 7 ***Rotational Motion*** **81**
7.1 Introduction 82
7.2 Measurement of Rotation 82
7.3 Rotational Motion 83
7.4 Equations of Rotational Motion 86
7.5 Radial Acceleration 88
7.6 Centripetal Force 89
7.7 Orbital Motion and Gravitation 91

CHAPTER 8 ***Rotational Dynamics*** **97**
8.1 Introduction 98
8.2 Moment of Inertia and Torque 98
8.3 Rotational Kinetic Energy 101
8.4 Power 105
8.5 Angular Momentum 106
8.6 Conservation of Angular Momentum 107

CHAPTER 9 ***Kinetic Theory of Gases and the Concept of Temperature*** **111**
9.1 Introduction 112
9.2 Molecular Weight 112
9.3 Thermometers 113
9.4 Ideal Gas Law and Absolute Temperature 114
9.5 Kinetic Theory of Gas Pressure 117
9.6 Kinetic Theory of Temperature 119
9.7 Measurement of Heat 121
9.8 Specific Heats of Gases 122
9.9 Work Done by a Gas 123
9.10 First Law of Thermodynamics 124
Supplement 9.1 Maxwell-Boltzmann Statistical Distribution 126

CHAPTER 10 ***Oscillatory Motion*** **131**
10.1 Introduction 132
10.2 Characterization of Springs 132
10.3 Frequency and Period 133
10.4 Amplitude and Phase Angle 134
10.5 Oscillation of a Spring 134
10.6 Energy of Oscillation 140

CHAPTER 11 ***Wave Motion*** **145**

11.1 Introduction 146
11.2 Wavelength, Velocity, Frequency, and Amplitude 146
11.3 Traveling Waves in a String 147
11.4 Energy Transfer of a Wave 151

CHAPTER 12 ***Interference of Waves*** **157**

12.1 Introduction 158
12.2 The Superposition Principle 158
12.3 Interference from Two Sources 159
12.4 Double Slit Interference of Light 162
12.5 Single Slit Diffraction 166
12.6 Resolving Power 168
12.7 X-Ray Diffraction by Crystals: Bragg Scattering 170
12.8 Standing Waves 173

CHAPTER 13 ***Electrostatics*** **177**

13.1 Introduction 178
13.2 Attraction and Repulsion of Charges 178
13.3 Coulomb's Law 179
13.4 Charge of an Electron 182
13.5 Superposition Principle 183

CHAPTER 14 ***The Electric Field and the Electric Potential*** **187**

14.1 Introduction 188
14.2 The Electric Field 188
14.3 Electrical Potential Energy 191
14.4 Electric Potential 194
14.5 The Electron Volt 197
14.6 Electromotive Force 197
14.7 Capacitance 198

CHAPTER 15 ***Electric Current*** **203**

15.1 Introduction 204
15.2 Motion of Charges in an Electric Field 204
15.3 Electric Current 205
15.4 Resistance and Resistivity 208
15.5 Resistances in Series and Parallel 210
15.6 Kirchhoff's Rules 214
15.7 Ammeters and Voltmeters 219
15.8 Power Dissipation by Resistors 221
15.9 Charging a Capacitor—*RC* Circuits 222

CHAPTER 16 *Magnetic Fields and Electromagnetic Waves* 227
16.1 Introduction 228
16.2 Magnetic Fields 228
16.3 Force on Current-Carrying Wires 229
16.4 Torque on a Current Loop 230
16.5 Magnetic Dipole Moment 232
16.6 Force on a Moving Charge 234
16.7 The Hall Effect 235
16.8 Electromagnetic Waves: The Nature of Light 237

CHAPTER 17 *The Beginning of the Quantum Story* 243
17.1 Introduction 244
17.2 Blackbody Radiation 244
17.3 The Photoelectric Effect 247
17.4 Further Evidence for the Photon Theory 253
Supplement 17.1 Momentum of the Photon 260

CHAPTER 18 *Atomic Models* 263
18.1 Introduction 264
18.2 The Rutherford Model 264
18.3 The Spectrum of Hydrogen 267
18.4 The Bohr Atom 269
18.5 The Franck-Hertz Experiment 274

CHAPTER 19 *Fundamental Principles of Quantum Mechanics* 279
19.1 Introduction 280
19.2 De Broglie's Hypothesis and Its Experimental Verification 280
19.3 Nature of the Wave 283
19.4 The Uncertainty Principle 285
19.5 Physical Origin of the Uncertainty Principle 288
19.6 Matter Waves and the Uncertainty Principle 289
19.7 Velocity of the Wave Packet: Group Velocity 292
19.8 The Principle of Complementarity 293

CHAPTER 20 *An Introduction to the Methods of Quantum Mechanics* 297
20.1 Introduction 298
20.2 The Schrödinger Theory of Quantum Mechanics 298
20.3 Application of the Schrödinger Theory 311

CHAPTER 21 *Quantum Mechanics of Atoms* 321
21.1 Introduction 322
21.2 Outline of the Solution of the Schrödinger Equation for the H Atom 323
21.3 Physical Significance of the Results 325
21.4 Space Quantization: The Experiments 326
21.5 The Spin 331
21.6 Some Features of the Atomic Wavefunctions 333
21.7 The Periodic Table 336

CHAPTER 22 *Crystal Structures and Bonding in Solids* 347
22.1 Introduction 348
22.2 Crystal Structures 348
22.3 Crystal Bonding 352

CHAPTER 23 *Free Electron Theories of Solids* 361
23.1 Introduction 362
23.2 Classical Free Electron Model 363
23.3 Quantum-Mechanical Free Electron Model 370
Supplement 23.1 The Wiedemann-Franz Law 387
Supplement 23.2 Fermi-Dirac Statistics 390

CHAPTER 24 *Band Theory of Solids* 395
24.1 Introduction 396
24.2 Bloch's Theorem 397
24.3 The Kronig-Penney Model 398
24.4 Tight-Binding Approximation 407
24.5 Conductors, Insulators, and Semiconductors 415
24.6 Effective Mass 418
24.7 Holes 422

CHAPTER 25 *Semiconductors* 429
25.1 Introduction 430
25.2 Intrinsic Semiconductors 430
25.3 Extrinsic or Impurity Semiconductors 438
25.4 Electrical Conductivity of Semiconductors 446
25.5 Photoconductivity 448

CHAPTER 26 *Semiconductor Devices* 453
26.1 Introduction 454
26.2 Metal-Metal Junction: The Contact Potential 454
26.3 The Semiconductor Diode 456
26.4 The Bipolar Junction Transistor (BJT) 465
26.5 Field-Effect Transistors (FET) 473

CHAPTER 27 *Some Basic Logic Circuits of Computers* 481

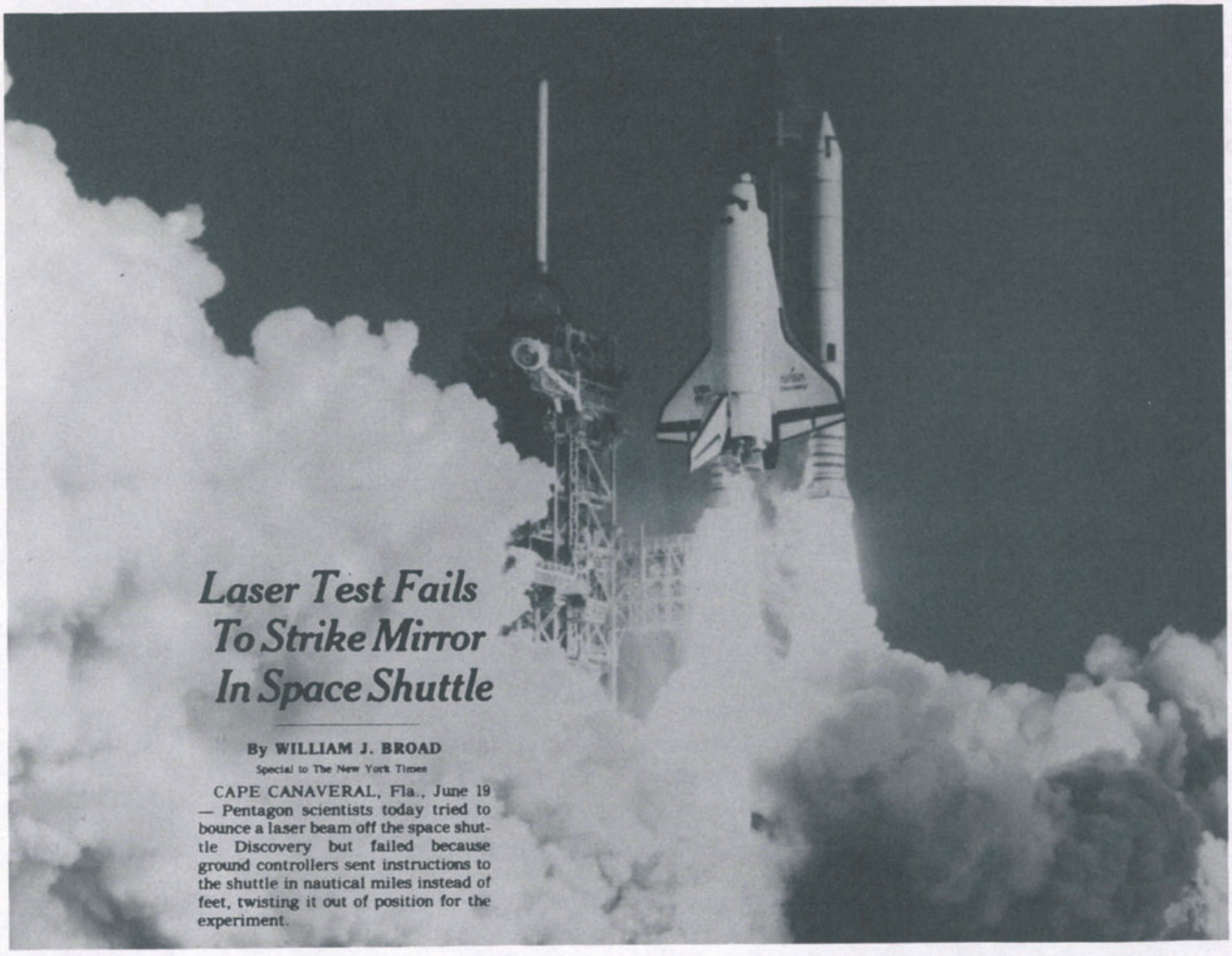
27.1 Introduction 482
27.2 Rudiments of Boolean Algebra 482
27.3 Electronic Logic Circuits 485
27.4 Semiconductor Gates 488
27.5 NAND and NOR Gates 493
27.6 Other Gates, RTL and TTL 494
27.7 Memory Circuits 497
27.8 Clock Circuits 498

CHAPTER 28 *The Technology of Manufacturing Integrated Circuits* 503

28.1 Introduction 504
28.2 Semiconductor Purification: Zone Refining 504
28.3 Single-Crystal Growth 505
28.4 The Processes of IC Production 508
28.5 Electronic Component Fabrication on a Chip 512
28.6 Conclusion 516

Photo Credits 519

Index 523



Laser Test Fails To Strike Mirror In Space Shuttle

By WILLIAM J. BROAD
Special to The New York Times

CAPE CANAVERAL, Fla., June 19 — Pentagon scientists today tried to bounce a laser beam off the space shuttle Discovery but failed because ground controllers sent instructions to the shuttle in nautical miles instead of feet, twisting it out of position for the experiment.

CHAPTER 1

Physical Quantities

1.1 INTRODUCTION

People have always observed natural phenomena and then verbalized their observations for discussion with others. With the development of *physics,* the words in the verbal descriptions were given symbols, which could then be manipulated according to the rules of mathematics.

In physics two fundamental processes are involved. The first is the description of natural phenomena based on experiments, which control variables. Theories are not accepted by physicists until verified by experiment. The second is mathematical manipulation or theorizing, which is a predictive process. If the observed quantities have been described properly and given the proper symbols, then the subsequent mathematical manipulations will result in new relationships that must be correct on testing, else their formulation will be rejected as incorrect or inadequate. Furthermore, the results of the relationships must stand the test of time. In this sense, time means enough time for many experiments to be performed to test the relationships. The laws presented in this book have met these requirements.

Although human beings have observed Nature from their first existence, it was not until the time of Galileo (1564–1642) that these observations began to be expressed in modern mathematical terms. Subsequent studies, measurements, and critical evaluations developed what we now call the *First Principles* of physics, which have truly stood the test of time. The first six chapters of this book discuss these principles and show how they are used in their simplest form. *Other areas of physics must satisfy these First Principles.* Chapters 7 through 16 illustrate the use of these principles in rotational motion, the behavior of gases, and electric and magnetic phenomena. Chapters 1 through 16 constitute what is usually known as *classical physics.* In the remaining chapters, we introduce a different way of describing the behavior of small physical particles. For example, although we may continue to consider the electron as the smallest negatively charged particle, experiments have shown that its behavior can also be described as a wave instead of a particle. The mathematics of waves, instead of that of particles, must be used to explain the electron's behavior in certain situations, whereas the mathematics of particles still applies in other situations. This revolution in thought, begun in the early part of this century, has led to the method, or science, of *wave mechanics,* which is more generally called *quantum mechanics.*

When the behavior of an electron within a solid is sought, very little can be learned by the particle treatment, but a vast amount of understanding can be achieved by the wave approach. How does this fit in with our mention of the test of time and observation? Although the actual length of time of this modern model (about 80 yr) is short compared with the time since Galileo, the number of experiments that have been performed is far greater. It has been said that of all the physicists who have ever lived, 95% are still alive.

The observation requirement is somewhat more subtle. We cannot observe fundamental particles such as an electron in the same way that we observe macroscopic objects; they are too small. In fact, we will later discuss

the *Uncertainty Principle*, which will show that the mere act of observing will change some state of the particle. Because of this principle, experiments that would completely characterize a small particle are too difficult to perform. What we do observe is the statistical behavior of a vast number of particles, and we infer the average behavior of a member of the statistical ensemble. Therefore, although we may never know the physical parameters of the individual particle, there are many physical experiments that can tell us if the statistical model is satisfactory.

1.2 QUANTITIES AND UNITS

If a physical phenomenon is to be quantified, there must be suitable, agreed-on measuring devices. Many measuring systems have been created in the past, none perfect. It is desirable to have a measurement system that has the least number of fundamental parameters. In classical physics, these are length, mass, and time. All other quantities can be defined in terms of these three. For example, speed is the ratio of a length to a time. We then choose a standard for each of the fundamental parameters. Most scientific measurements use the metric system. There are two versions of the metric system in use, the *cgs* (centimeter, gram, second) and the *mks* (meter, kilogram, second). Although the cgs system is still often used in the biological sciences, most measurements by physicists now use the mks system. This is the mechanical part of the more general SI (Système International) that covers all physical measurements. The English system of units (foot, pound, second) is used largely by engineers. We will mostly use SI units in this book.

The unit of a quantity is as important as the magnitude, as indicated in the quote from *The New York Times* at the beginning of this chapter. It is meaningless to say "the distance between two points is 10," because the 10 may be meters, miles, or inches. The units are an integral part of the measurement and must be treated algebraically. One may substitute for them or convert them to a different system, but they cannot be gotten rid of except by an algebraic process. For example, π is dimensionless, but it is defined on the basis of two measured quantities, the circumference of a circle divided by its diameter. It is independent of units because they cancel: circumference (meters)/diameter (meters), and it is seen that meters cancel algebraically.

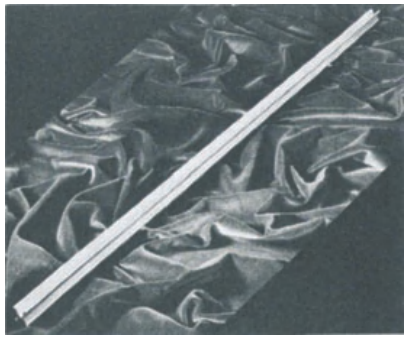
Conversion of units must be done with care, and, in order to convert, a relationship between units of two different systems must be known. Let us illustrate this with a trivial example. How many feet are there in 5 mi?

$$5 \text{ mi} = ? \text{ ft} \tag{1.1}$$

We know the relation

$$5280 \text{ ft} = 1 \text{ mi} \tag{1.2}$$

How do we substitute this with care? The safest rule is always to multiply by one (unity), because we know that in algebra, multiplication of anything



United States copy of the original platinum-iridium bar which for many years was the standard of length: the meter.

by unity leaves it unchanged. Our conversion relation, Eq. 1.2, can be made into two different forms of unity depending on whether we divide both sides by miles or by feet:

$$\frac{5280 \text{ ft}}{1 \text{ mi}} = 1 \qquad \frac{1 \text{ mi}}{5280 \text{ ft}} = 1$$

We can now multiply the left side of Eq. 1.1 by the first form of unity, which algebraically results in the unit

$$\frac{\text{mi}\cdot\text{ft}}{\text{mi}} = \text{ft}$$

Multiplication by this gives

$$5 \text{ mi} = 5 \cancel{\text{mi}} \times \frac{5280 \text{ ft}}{1 \cancel{\text{mi}}} = 5 \times 5280 \text{ ft} = 26{,}400 \text{ ft}$$

We see that miles in both the numerator and the denominator have cancelled algebraically, leaving feet as the unit.

We can do more than one algebraic step at a time. For example, how many seconds are in 1 day?

$$1 \text{ day} = ? \text{ sec}$$

We know three conversion relations

$$1 \text{ day} = 24 \text{ h} \qquad 1 \text{ h} = 60 \text{ min} \qquad 1 \text{ min} = 60 \text{ sec}$$

We select our choices of unity to give successive algebraic cancellations

$$1 \text{ day} = 1 \cancel{\text{day}} \left(\frac{24 \cancel{\text{h}}}{1 \cancel{\text{day}}}\right)\left(\frac{60 \cancel{\text{min}}}{1 \cancel{\text{h}}}\right)\left(\frac{60 \text{ sec}}{1 \cancel{\text{min}}}\right)$$

$$= 24 \times 60 \times 60 \text{ sec} = 86{,}400 \text{ sec}$$

We may convert two units simultaneously in a single equation to save a lot of writing. For example, a car traveling at 60 mi/h travels how many feet per second?

$$60 \frac{\text{mi}}{\text{h}} = 60 \frac{\cancel{\text{mi}}}{\cancel{\text{h}}}\left(\frac{5280 \text{ ft}}{1 \cancel{\text{mi}}}\right)\left(\frac{1 \cancel{\text{h}}}{60 \cancel{\text{min}}}\right)\left(\frac{1 \cancel{\text{min}}}{60 \text{ sec}}\right) = \frac{60 \times 5280}{60 \times 60} \text{ ft/sec} = 88 \text{ ft/sec}$$

Remember that in square or cubic units, all measurements must be in the same units. It makes no sense to calculate the area of a room if its length is measured in feet and its width is measured in meters. Both measurements should be in the same system. Also remember that when converting square or cubic units they must be squared or cubed just as algebraic quantities. For example, how many cubic centimeters (1 m = 100 cm) are there in a volume of 1 m^3?

$$1 \text{ m}^3 = 1 \cancel{\text{m}}^3 \left(\frac{100 \text{ cm}}{1 \cancel{\text{m}}}\right)\left(\frac{100 \text{ cm}}{1 \cancel{\text{m}}}\right)\left(\frac{100 \text{ cm}}{1 \cancel{\text{m}}}\right) = 1{,}000{,}000 \text{ cm}^3$$

1.3 POWERS OF 10

Often, very large and very small numbers arise in physics. In 1 cm^3 of a solid there are a vast number of atoms, about 1 followed by 21 zeros. Measurements have shown that the range of atomic diameters in meters is between 0.0000000001 and 0.0000000003 m. Because of the difficulty of reading and writing numbers with many zeros, we use powers of 10 notation. Recall the algebraic postulate that any quantity to the zeroth power is, identically, unity. A brief table of some powers of 10 follows.

$$\begin{aligned} 10^0 &= 1 & 10^{-1} &= \tfrac{1}{10} = 0.1 \\ 10^1 &= 10 & 10^{-2} &= \tfrac{1}{100} = 0.01 \\ 10^2 &= 100 & 10^{-3} &= \tfrac{1}{1000} = 0.001 \\ 10^3 &= 1000 & 10^{-4} &= \tfrac{1}{10000} = 0.0001 \\ 10^4 &= 10{,}000 \end{aligned}$$

Some of these are used as prefixes; for example, 10^3 = kilo, 10^{-3} = milli, 10^{-6} = micro, 10^{-9} = nano, 10^{-12} = pico.

Let us review the algebra of adding and multiplying powers with the letter a representing 10. (These rules apply for a equal to any value other than zero.)

Example 1-1

$$(b \times a^n)(c \times a^m) = bca^n a^m = bca^{n+m}$$

Substitute arbitrary numbers for the letters; for example, $b = 2$, $c = 3$, $n = 4$, and $m = 2$.

$$(2 \times 10^4)(3 \times 10^2) = 2 \times 3 \times 10^4 \times 10^2 = 6 \times 10^6$$

Example 1-2

$$\frac{e \times a^n}{f \times a^m} = \frac{e}{f} a^{n-m}$$

If $e = 6$, $f = 2$, $n = 4$, and $m = 2$,

$$\frac{6 \times 10^4}{2 \times 10^2} = \frac{6}{2} 10^{4-2} = 3 \times 10^2$$

Example 1-3

$$(b \times a^n) + (c \times a^m)$$

This form can be simplified if $n = m$, then

$$(b \times a^n) + (c \times a^n) = (b + c)\, a^n$$

If $n \neq m$, they can be made equal. For example, let

$$b = 2 \text{ and } c = 3, \quad n = 4 \text{ and } m = 5$$

substituting in the algebraic relation, we have

$$2 \times 10^4 + 3 \times 10^5$$

But $10^5 = 10^1 \times 10^4$ and therefore

$$2 \times 10^4 + 3 \times 10^5 = 2 \times 10^4 + 3 \times 10^1 \times 10^4$$
$$= (2 + 30)10^4 = 32 \times 10^4 = 3.2 \times 10^5$$

This sum could also be done by converting the first term

$$2 \times \frac{10^1}{10^1} \times 10^4 + 3 \times 10^5 = 0.2 \times 10^5 + 3 \times 10^5 = 3.2 \times 10^5$$

In scientific notation, usually only one digit is placed in front of the decimal point.

1.4 ACCURACY OF NUMBERS

Suppose we wish to find the area of a rectangular surface. We know that we multiply the length l by the width w. Suppose we take a metric ruler to measure l and w. The metric ruler's smallest division is the millimeter, 10^{-3} m. Figure 1-1 illustrates the use of a metric ruler to measure the length of the rectangle. We can see that the measure of length lies between 47.6 and 47.7 cm; from the position of the edge of the rectangle we can estimate the second decimal as being less than 0.5 mm but not less than 0.3 mm. We can therefore express our measurement as 47.64 ± 0.01 cm, or 0.4764 ± 0.0001 m. Thus, the last digit is always uncertain and the ± value is the magnitude of the uncertainty. Suppose we have measured the width as 0.6343 ± 0.0001 m; what is the accuracy of the calculation of the area? We examine the two extremes. The largest area is

$$0.4765 \text{ m} \times 0.6344 \text{ m} = 0.3023 \text{ m}^2$$

and the smallest is

$$0.4763 \text{ m} \times 0.6342 \text{ m} = 0.3021 \text{ m}^2$$

We can write the answer as the average between the two values with ± the uncertainty, or 0.3022 ± 0.0001 m^2. Therefore, the accuracy of the product cannot exceed the accuracy of any of the components in the product.

This type of analysis must be extended to all data-handling, for example, sums, differences and quotients. Suppose that we want to sum two numbers known with different degrees of accuracy. What is the accuracy of the sum? As an example, consider three successive points A, B, and C on a straight line. If the distance between A and B is 15.75 m and the distance between B and C is 2.432 m, what is the distance between A and C? The answer is obtained by summing the two numbers. If we use a calculator to evaluate the sum, it will tell us that the answer is 18.182 m. However, the only reliable

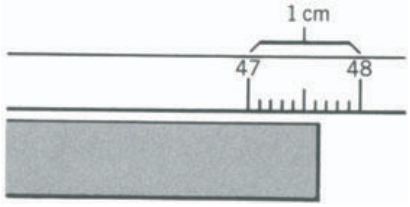


FIGURE 1-1

answer is 18.18 m, because we do not know what the third digit after the decimal point of the first number (15.75 m) is. No matter how accurately a given parameter is measured, when it is combined arithmetically with another measurement the result is only as accurate as the least-accurate measurement. The number of accurate figures in a measurement is called the number of *significant figures*. Computer science students are often misled because computer or calculator answers may have 8 to 10 figures. In general, most of these figures have no significance and the answer should be rounded off to the lowest number of significant figures of the quantities used in the calculation. It is *incorrect* to give an answer with a greater number of figures.

PROBLEMS

1.1 Express your height in meters, using the relation 1 in. = 2.54 cm.

1.2 Use the relations 1 mi = 5280 ft and 1 m = 3.28 ft to express the speed 60 mi/h in meters per second (m/sec).

1.3 How many kilometers are in 1 mi?

1.4 Express the following in scientific notation (a single digit to the left of the decimal):

0.038, 0.000042, 5280, 62.356,
$(4 \times 10^3 + 3 \times 10^2)6 \times 10^{-3}$,
$3.2 \times 10^4/(6.1 \times 10^{-2} + 9.2 \times 10^{-3})$

1.5 Express the following operation in scientific notation

$$\frac{4 \times 10^2 + 6 \times 10^3}{2 \times 10^{-4}}$$

1.6 Light travels at 186,000 mi/sec. Assume an average of 365 days in a year. (a) How many years does it take the light to reach us from the sun, which is 9.3×10^7 mi from the earth? (b) How many years does it take the light to reach us from the nearest star, other than the sun, which is 1.8×10^{13} mi from us?

1.7 Astronomers measure large distances in a unit called the *light-year*. This is the distance that light traveling at approximately 186,000 mi/sec will travel in 1 yr. How many miles in 1 light-year?

1.8 Assume that the average lecture period is 1 microcentury (10^{-6} centuries); how long is the lecture period in minutes?

(*Answer*: 52.6 min.)

1.9 A *light-fermi* is a unit of time proposed by science-fiction writer Isaac Asimov. It is defined as the time taken by light to travel the distance of 1 fermi (10^{-15} m), which is the approximate size of the proton. How long is a light-fermi in seconds? Light travels at 3×10^8 m/sec.

1.10 There are approximately 8×10^{28} copper atoms in 1 m^3 of copper. (a) What is the volume occupied by a copper atom? (b) What is the radius of a sphere having that volume? (volume of a sphere = $4\pi r^3/3$)

1.11 Assume that atoms have spherical shape with average radius 4×10^{-10} m. How many atoms are there in the earth? Neglect the volume lost in packing the spheres and take the average radius of the earth to be 6.37×10^6 m.

(*Answer*: 4.04×10^{48}.)

1.12 In the Old Testament the Lord commanded Noah to build an ark 300 cubits long, 50 cubits wide, and 30 cubits high. A cubit is the length from a man's elbow to the tip of his extended middle finger. We do not know Noah's height, so measure a cubit from both a short person and a tall person. Assume the ark was a parallelepiped with right angles. (a) What

is the maximum and the minimum values of its volume? (b) Assuming that the average animal required a space of $2 \times 4 \times 6$ ft^3, and that one half the volume of the ark was for food and passengers, what is the possible variation in the number of animals that could be accommodated?

1.13 Density is defined as the mass per unit volume. Take the average density of the earth to be 5.5 g/cm^3 and assume that the earth is a sphere of radius 6.37×10^3 km. Calculate the mass of the earth.

(*Answer*: 5.96×10^{24} kg.)

1.14 A neutron is one of the constituent particles of the nucleus. The mass of the neutron is 1.67×10^{-27} kg. Assuming that the neutron is a sphere of radius 1 F (10^{-15} m), what is the density of the neutron in g/cm^3? Compare your answer with the average density of the earth (see problem 1.13).

1.15 The radius of a carbon atom is about 2.5×10^{-8} cm. (a) How many could fit in a row 1-cm long? (b) How many could fit in a layer one atom deep and area 1 cm^2? (c) How many could fit in a cube 1 cm on each size? (d) If a crystal of carbon atoms (diamond) had this form, what is the minimum number of impurity atoms that could block the light coming through the faces of a 1-cm^3 cube? Express your answer in both percent and in parts per million. (*Hint:* Assume as an approximation that a layer of impurity atoms on each of three faces of the cube at right angles to each other could block all the light.)

(*Answer*: (a) 2×10^7 cm^{-1}, (b) 4×10^{14} cm^{-2}, (c) 8×10^{21} cm^{-3}, (d) 1.2×10^{15}, 1.5×10^{-5} %, 0.15 ppm.)

1.16 The distance x of an object from a certain origin is found to vary with time t as $x = a_1 + a_2 t + a_3 t^2$, where x is in meters, t is in seconds, and a_1, a_2, and a_3 are constants. What are the units of a_1, a_2, a_3?

1.17 A student is trying to find what parameters determine the period (time for a full swing) of a pendulum. After some experimentation, he concludes that the period T is given by $T = 2\pi\frac{g}{l}$ where $g = 9.8$ m/s^2 is the acceleration of free-falling bodies near the surface of the earth and l is the length of the pendulum. (a) Show by the units of the terms that the student's conclusion is incorrect. (b) Assuming that the period depends only on g and l, what is the proper functional dependence of T on these two quantities?

CHAPTER 2

Vectors

2.1 INTRODUCTION

We will be dealing with two types of quantities in this book. Some quantities are fully specified only by a number and a unit, such as a quart of milk or a pound of potatoes. Such a quantity consisting only of magnitude is called a *scalar* quantity. Other measurements have meaning only if direction is specified along with the magnitude. For example, telling a stranger that a gas station is 1 mi away will not help him unless you specify the direction also. A quantity that has both magnitude and direction and obeys certain algebraic laws is called a *vector* quantity and will be indicated in this book by boldface type.

2.2 VECTOR COMPONENTS

A vector direction must be specified in relation to a given coordinate system. Given a coordinate system, any vector can be expressed in terms of its components.

Let us examine the concept of vector components first by simply using the compass points north, south, east, and west as the directions.

Suppose you walk 5.0 mi east and then 4.0 mi north. How far are you and in what direction from the starting point?

We draw Fig. 2-1. We see that we have right-angle geometry and, because R is the hypotenuse, using the pythagorean theorem, we have

$$R = \sqrt{(4.0 \text{ mi})^2 + (5.0 \text{ mi})^2} = 6.4 \text{ mi}$$

We also know from trigonometry that

$$\tan \theta = \frac{\text{opposite side}}{\text{adjacent side}} = \frac{4.0 \cancel{\text{mi}}}{5.0 \cancel{\text{mi}}} = 0.8$$

or

$$\theta = \arctan 0.8 = 39°$$

The distance R together with its orientation θ is called the *vector sum* of the two vectors 5.0 mi east and 4.0 mi north and is given the name *resultant*. The 5.0 mi east vector and the 4.0 mile north vector are called the *components* of R.

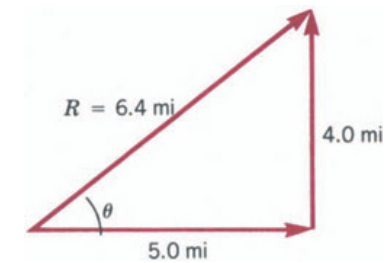


FIGURE 2-1

Suppose, instead, you walk 5.0 mi east and 4.0 mi northeast, namely, 45° north of east as in Fig. 2-2. What is the resultant?

This is a little more complicated and, although you could use the law of cosines and the law of sines to solve the problem, there is a simpler way that will be used throughout the book. This method is particularly useful when you deal with problems that involve more than two vectors. Let us reduce the problem to two questions.

1. How far are you to the east of the starting point?
2. How far are you north of the starting point?

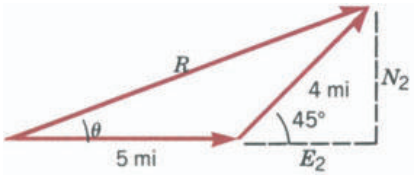


FIGURE 2-2

Note that the additional dashed lines in Fig. 2-2 labeled E_2 and N_2 represent the respective distances east and north traveled in the second leg. In this second part of the walk we see again a right triangle with E_2 and N_2 as the legs and with 4.0 mi as the hypotenuse. Recall from trigonometry that sin 45° = N_2/4.0 mi and cos 45° = E_2/4.0 mi. Therefore, N_2 = (4.0 mi) sin 45° and E_2 = (4.0 mi) cos 45°. Now make a table of the data

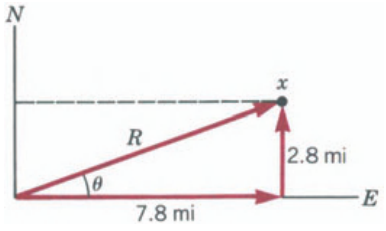


FIGURE 2-3

		East		*North*
Walk 1		5 mi		0 mi
Walk 2	(4.0 mi) cos 45° =	2.8 mi	(4.0 mi) sin 45° =	2.8 mi
	Total	7.8 mi		2.8 mi

Take a fresh piece of graph paper and plot these distances from a starting point, which will be at the origin of a compass coordinate system, as in Fig. 2-3. The point marked x is your location from the starting point, R is the distance, and θ is the angle. We now have right-triangle geometry again and may write as before

$$R = \sqrt{(7.8 \text{ mi})^2 + (2.8 \text{ mi})^2} = 8.3 \text{ mi}$$

$$\theta = \arctan\left(\frac{2.8 \cancel{\text{mi}}}{7.8 \cancel{\text{mi}}}\right) = 19.7°$$

Consider the more complicated walk of Fig. 2-4. What is the resultant R of the four displacements shown?

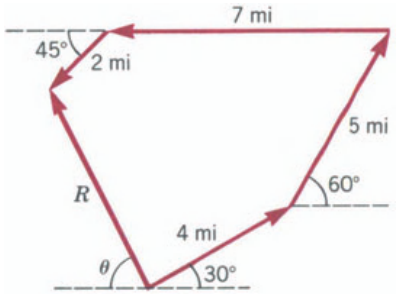


FIGURE 2-4

To find R and θ, make a table of east-west and north-south displacements. Note here that the table will be in terms of east and north, so that a displacement to the west will be a negative east displacement and one to the south will be negative north. The components of the 4.0 mi, 5.0 mi, 7.0 mi, and 2.0 mi displacements can be found by putting pieces of graph paper with the origins at the starting points of these legs and finding the components by right-triangle geometry, as in Fig. 2-5.

Construct a table as in the previous example.

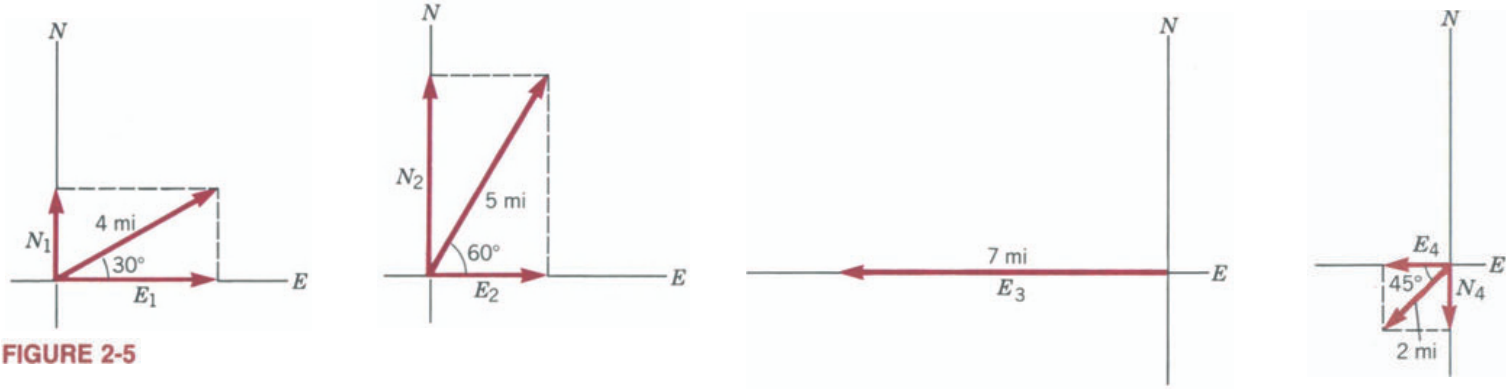


FIGURE 2-5

		East		North
Walk 1	(4.0 mi) cos 30° =	3.5 mi	(4.0 mi) sin 30° =	2.0 mi
Walk 2	(5.0 mi) cos 60° =	2.5 mi	(5.0 mi) sin 60° =	4.3 mi
Walk 3	−7.0 mi =	−7.0 mi	0 mi =	0 mi
Walk 4	−(2.0 mi) cos 45° =	−1.4 mi	−(2.0 mi) sin 45° =	−1.4 mi
	Total	−2.4 mi		4.9 mi

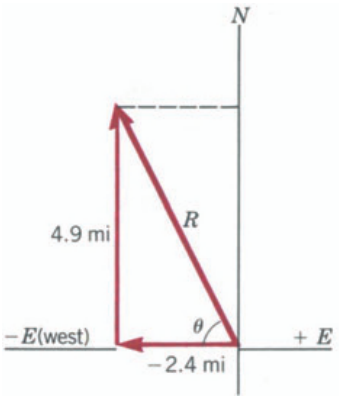


FIGURE 2-6

Now take a piece of graph paper and plot the total east and north displacements as in Fig. 2-6.

$$R = \sqrt{(-2.4 \text{ mi})^2 + (4.9 \text{ mi})^2} = 5.5 \text{ mi}$$

$$\theta = \arctan\left(\frac{4.9 \cancel{\text{mi}}}{|-2.4 \cancel{\text{mi}}|}\right) = 63.9° \text{ north of the west direction}$$

What we have done in Fig. 2-6 is to define an angle θ as less than 90°. This makes both the calculation and the spatial location much simpler. To work with $\theta < 90°$ we had to ignore the sign of the coordinate and locate the resulting angle on the graph. Had we kept the sign of the coordinate in the calculation of the angle, it would not have helped much because the tangent is negative in two of the quadrants. We would still have to rely on some construction to locate the angle. The method shown in the preceding example, which first uses a graph to show where you are, leaves no question as to the meaning of the angle θ. How does one specify the angle? It is equally correct to say it is 64° north of west or, using the 360° scale with east as 0°, the angle would be 180° − 64° = 116°. One other point should be noted. In Fig. 2-5 we symbolically used four pieces of graph paper to obtain the components of the vectors. We could have equally used a single piece by putting the beginning of each of the walks at the origin of the graph paper as in Fig. 2-7; henceforth, we will conserve paper by this method.

Now that we have related coordinate systems to navigation, we can apply the same techniques to cartesian coordinates *x*-*y* instead of compass directions. To show how the vector component method is used in more complicated

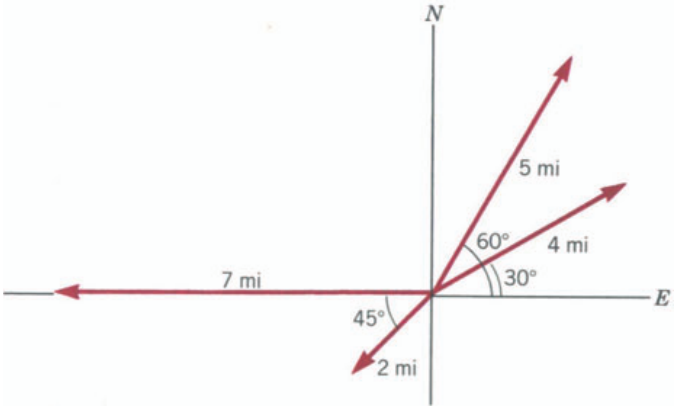


FIGURE 2-7

situations, we will abandon our walks and consider forces. The concept of forces will be developed more fully in Chapter 4. For now, we can simply rely on our experience that a force is that which when exerted in some direction against an object may or may not cause it to move. It is a vector quantity.

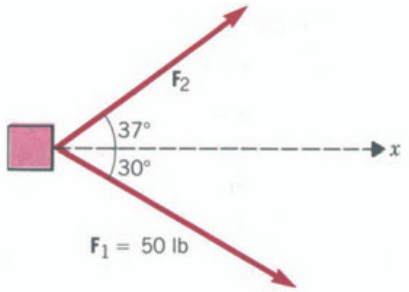


FIGURE 2-8
Example 2-1.

Example 2-1

A box is pulled by two persons exerting the forces $\mathbf{F}_1$ and $\mathbf{F}_2$ shown in Fig. 2-8, where $\mathbf{F}_1$ is given as 50 lb. Two questions may now be asked. 1. What force $\mathbf{F}_2$ must be applied so that the box moves only in the x direction? 2. What single force could replace $\mathbf{F}_1$ and $\mathbf{F}_2$ so that the box moves only in the x direction?

Solution We obtain answers to these two questions by first constructing a vector diagram of the forces, as in Fig. 2-9, and tabulating the components of the forces.

Force	*x components*	*y components*
$\mathbf{F}_1$	$(50 \text{ lb}) \cos 30° = 43.3 \text{ lb}$	$-(50 \text{ lb}) \sin 30° = -25.0 \text{ lb}$
$\mathbf{F}_2$	$F_2 \cos 37° = 0.8\,F_2$	$F_2 \sin 37° = 0.6\,F_2$
Total	$43.3 \text{ lb} + 0.8\,F_2$	$-25.0 \text{ lb} + 0.6\,F_2$

If the object is going to move in the x direction, the resultant force must be in the x direction only, with no component in the y direction. If there is to be no net force in the y direction that could cause the box to move in that direction, then the sum of the positive and negative y forces must be zero. We can express this as

$$\Sigma F_y = 0$$

$$-25 \text{ lb} + 0.6\,F_2 = 0$$

or

$$F_2 = \frac{25 \text{ lb}}{0.6} = 41.7 \text{ lb}$$

Question 2 can be answered from the sum of forces in the x direction.

$$\Sigma F_x = 43.3 \text{ lb} + 0.8\,F_2$$

$$= 43.3 \text{ lb} + 0.8 \times 41.7 \text{ lb}$$

$$= 76.7 \text{ lb}$$

Therefore, we conclude that instead of the two forces $\mathbf{F}_1$ and $\mathbf{F}_2$ acting on the box, the same result can be obtained by a single force of 76.7 lb pulling in the x direction.

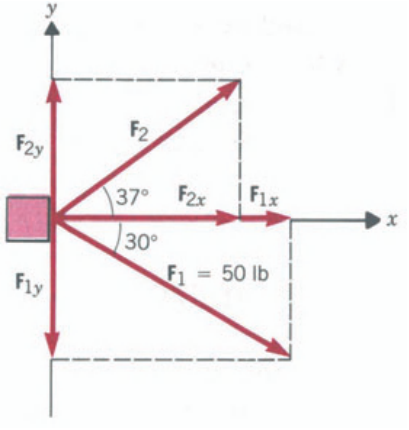


FIGURE 2-9
Example 2.1.

We can now summarize the preceding discussion concerning the addition of vectors by the method of components.

1. Establish a coordinate system. Choose an x-y cartesian coordinate system that is convenient for calculation. For example, we will see in Chapter 4 that when an object slides on an incline plane, it is convenient to choose the plane as the x axis.
2. Construct each vector with its tail at the origin of this coordinate system.
3. Drop construction lines from the head of each vector perpendicular to the x and y axes, and note from the laws of trigonometry that the x component of a vector is equal to the magnitude of the vector multiplied by the cosine of the angle that the vector makes with the x axis. Similarly, the y component of a vector equals the magnitude of the vector times the sine of the same angle.
4. Add algebraically the individual x components and y components of all the vectors to find the x and y components, respectively, of the resultant vector.
5. The square of the resultant equals the sum of the squares of the x and y components of the resultant.
6. To determine the angle of the resultant, it is best to make a sketch of the resultant, showing its x and y components, which will indicate in what quadrant the resultant is. The angle between R and the x axis (either positive or negative direction) is equal to the tangent of the absolute value of the y component of the resultant divided by the absolute value of the x component.

We conclude this section by defining certain rules of vector algebra.

1. If we have a vector **A**, we define a vector $-\mathbf{A}$ as one whose magnitude is the same as that of **A** but its direction is opposite to that of **A**.
2. A vector $2\mathbf{A}$ is one whose magnitude is twice that of **A** and whose direction is the same as that of **A**. More generally, when a vector is multiplied or divided by a scalar quantity, we obtain a vector of different magnitude but of the same direction as the initial vector.
3. When several vectors are added, they can be added in any order and thus the distributive law holds; for example, $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$.

2.3 UNIT VECTORS

It is lengthy to write and say x component, y component, and z component. A shorthand notation is used. Unit vectors **i**, **j**, and **k** are introduced. These have the respective directions x, y, and z and a magnitude of unity, so they give direction without changing the magnitude. For example, the vector $\mathbf{F}_1$

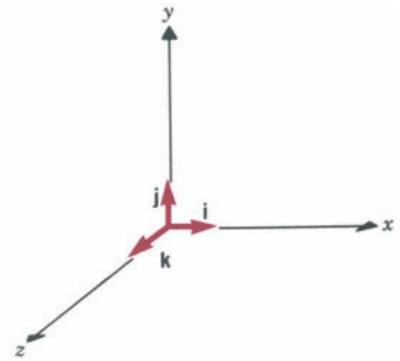


FIGURE 2-10
Unit vectors **i**, **j**, **k** on the three coordinate axes.

in Fig. 2-9 that has components 43.3 lb in the x direction and -25 lb in the y direction would simply be written as

$$\mathbf{F}_1 = (43.3\,\mathbf{i} - 25\,\mathbf{j})\text{lb}$$

The conventional diagram for three dimensions is shown in Fig. 2-10 with the corresponding unit vectors.

The vector **F** in Fig. 2-11 has the components shown, which are obtained by the right-triangle method of extending perpendiculars to the axes. This vector would be written as

$$\mathbf{F} = (4\,\mathbf{i} + 8\,\mathbf{j} + 5\,\mathbf{k})\text{ lb}$$

and its magnitude is obtained from the three-dimensional pythagorean theorem

$$F = \sqrt{(4\text{ lb})^2 + (8\text{ lb})^2 + (5\text{ lb})^2}$$
$$= 10.2\text{ lb}$$

When desired, the direction can be obtained by standard methods of analytic geometry to obtain its location in space.

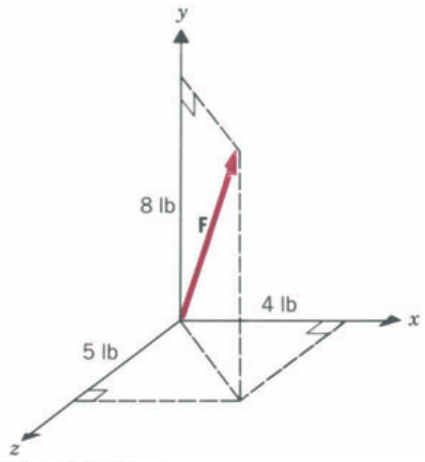


FIGURE 2-11
Components of vector **F** on the three coordinate axes.

2.4 DOT PRODUCT

Very often in physics we have two vectors with an angle θ between them, and we wish to find the product of their components that lie in the direction of one or the other vector.

Consider Fig. 2-12. If we, for instance, select the **A** direction, then the component of vector **B** in that direction is given by dropping a perpendicular (Fig. 2-12 *a*) and noting from the resulting right triangle that the component of **B** in the **A** direction is B cos θ and the product of this component and vector **A** is

$$AB\cos\theta$$

If, instead, we had selected the **B** direction we could equally have dropped a perpendicular from vector **A** to the line of vector **B** (Fig. 2-12*b*) and obtained the identical result. Because there is no specified direction for the resulting product, we define such a product as a scalar. We use the shorthand notation of a dot ($\cdot$) to represent this type of product, which is referred to as the *dot product*

$$\mathbf{A}\cdot\mathbf{B} = AB\cos\theta \qquad (2.1)$$

where on the right side A and B are simply the magnitude of each of the vectors. We will use this dot product in Chapter 5 on work and energy, both of which arise from vector relationships although neither in itself has direction.

$$\mathbf{A}\cdot\mathbf{B} = AB\cos\theta$$

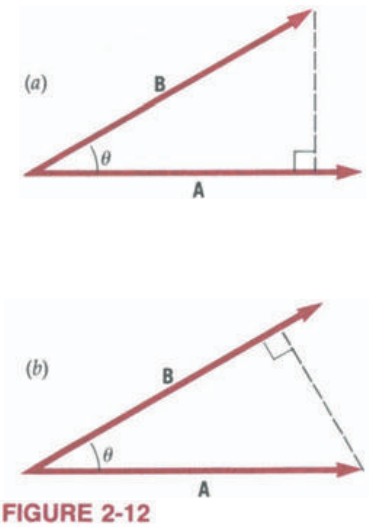


FIGURE 2-12
Geometric representation of two ways of forming a dot product of vectors **A** and **B**.

Let us apply our definition of the dot product to the unit vectors **i**, **j**, and **k**.

$$\mathbf{i} \cdot \mathbf{j} = (1)(1) \cos 90° = 0$$
$$\mathbf{i} \cdot \mathbf{k} = (1)(1) \cos 90° = 0$$
$$\mathbf{j} \cdot \mathbf{k} = (1)(1) \cos 90° = 0$$
$$\mathbf{i} \cdot \mathbf{i} = (1)(1) \cos 0° = 1$$
$$\mathbf{j} \cdot \mathbf{j} = (1)(1) \cos 0° = 1$$
$$\mathbf{k} \cdot \mathbf{k} = (1)(1) \cos 0° = 1$$

We see that when a unit vector is dotted with a different unit vector the result is zero, whereas when a unit vector is dotted with itself the result is unity.

Example 2-2

Find $\mathbf{A} \cdot \mathbf{B}$ if $\mathbf{A} = 3\mathbf{i} + 2\mathbf{j}$ and $\mathbf{B} = -\mathbf{i} + 3\mathbf{j}$.

Solution

$$\begin{aligned}\mathbf{A} \cdot \mathbf{B} &= (3\mathbf{i} + 2\mathbf{j}) \cdot (-\mathbf{i} + 3\mathbf{j}) \\ &= 3\mathbf{i} \cdot (-\mathbf{i}) + 3\mathbf{i} \cdot 3\mathbf{j} + 2\mathbf{j} \cdot (-\mathbf{i}) + 2\mathbf{j} \cdot 3\mathbf{j} \\ &= -3 + 0 + 0 + 6 \\ &= 3\end{aligned}$$

You can verify that $\mathbf{B} \cdot \mathbf{A}$ gives the same answer; therefore the commutative law holds for the dot product of two vectors.

2.5 CROSS PRODUCT

In some topics of physics we often need to define a vector **C**, whose magnitude is equal to the magnitude of one vector **A** times the component of a second vector **B** in the direction perpendicular to **A** (see Fig. 2-13). Moreover, we want the direction of **C** to be perpendicular to **A** and **B**.

We thus introduce a new type of product called the *cross product* of **A** and **B**. If **C** is the cross product of **A** and **B**, we write

$$\mathbf{C} = \mathbf{A} \times \mathbf{B} \tag{2.2}$$

By this definition, it is seen in Fig. 2-13 that the magnitude of the vector **C** is

$$C = AB \sin \theta$$

The direction of **C** is perpendicular to both **A** and **B** and consequently perpendicular to the plane containing **A** and **B**. There are obviously two possible directions for a vector perpendicular to a plane. This ambiguity can be re-

$$|\mathbf{A} \times \mathbf{B}| = AB \sin \theta$$

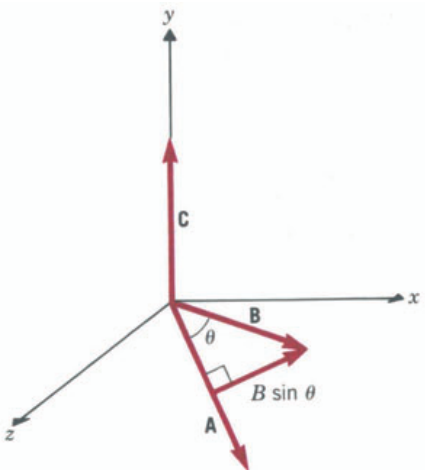


FIGURE 2-13
Geometric representation of the cross product $\mathbf{A} \times \mathbf{B} = \mathbf{C}$, where $C = AB \sin \theta$.

moved by using a *right-hand rule*. A simple mnenomic is the following. Consider the two vectors to be two sticks connected by a hinge at the apex of the angle. Mentally place the palm of your right hand against the outside of the first stick to be crossed (in our case **A**) as if to push the two sticks together (see Fig. 2-14). Do this with the thumb extended and the thumb will point in the direction of the vector cross product. The vector $\mathbf{C}' = \mathbf{B} \times \mathbf{A}$ will, from the definition, have the same magnitude as **C**. However, it is clear from Fig. 2-15 that the direction of $\mathbf{C}'$ is opposite to that of **C**; namely,

$$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}$$

Note that this is in contrast to the dot product where $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$.

We can examine the resulting directions of cross products by operating on the unit vectors of Fig. 2-10 with the right-hand rule. We find that

$$\mathbf{i} \times \mathbf{j} = \mathbf{k}$$
$$\mathbf{j} \times \mathbf{k} = \mathbf{i}$$
$$\mathbf{k} \times \mathbf{i} = \mathbf{j}$$
$$\mathbf{j} \times \mathbf{i} = -\mathbf{k}$$
$$\mathbf{k} \times \mathbf{j} = -\mathbf{i}$$
$$\mathbf{i} \times \mathbf{k} = -\mathbf{j}$$

It should be noted in the definition that $\mathbf{i} \times \mathbf{i} = \mathbf{j} \times \mathbf{j} = \mathbf{k} \times \mathbf{k} = 0$ because the angle between a vector and itself is zero and $\sin 0° = 0$.

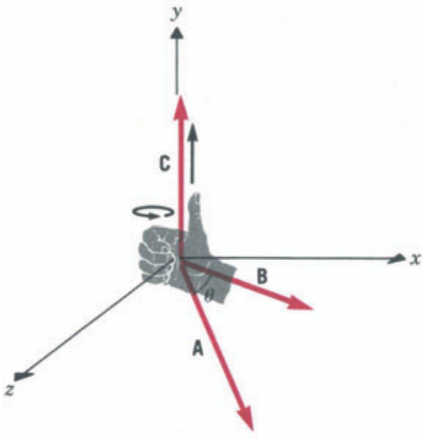


FIGURE 2-14
Right-hand rule for the vector cross product. For $\mathbf{A} \times \mathbf{B}$, curl the fingers of the right hand in a direction such that the fingers seem to push vector **A** toward vector **B**. The direction of the thumb points in the direction of the vector $\mathbf{C} = \mathbf{A} \times \mathbf{B}$.

Example 2-3

Find $\mathbf{A} \times \mathbf{B}$ if $\mathbf{A} = 3\mathbf{i} + 2\mathbf{j}$ and $\mathbf{B} = -\mathbf{i} + 3\mathbf{j}$

Solution

$$\begin{aligned}\mathbf{A} \times \mathbf{B} &= (3\mathbf{i} + 2\mathbf{j}) \times (-\mathbf{i} + 3\mathbf{j}) \\ &= 3\mathbf{i} \times (-\mathbf{i}) + 3\mathbf{i} \times 3\mathbf{j} + 2\mathbf{j} \times (-\mathbf{i}) + 2\mathbf{j} \times 3\mathbf{j} \\ &= 0 + 9\mathbf{k} + 2\mathbf{k} + 0 \\ &= 11\mathbf{k}\end{aligned}$$

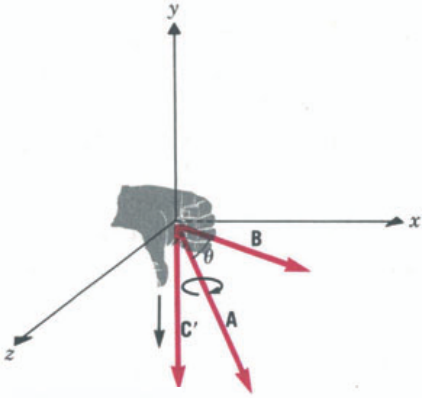


FIGURE 2-15
Right-hand rule for the vector cross product of $\mathbf{B} \times \mathbf{A} = -\mathbf{C}$.

PROBLEMS

2.1 What are the x and y components of the following vector displacements? (a) 2 m at 20°? (b) 3 m at 120°? (c) 4 m at 240°? (d) 2.5 m at 325°? All angles are with respect to the positive x axis.

2.2 A sailboat follows a series of racing buoys. On the first lap it goes 8 mi at 20°, then 10 mi at 40°, then 6 mi at 130°. What distance is it from its starting point, and in what direction must it sail to return? (*Answer*: 17.8 mi, 230.6°.)

2.3 A sailboat sails 7 mi in the direction 37° north of east, then 4 mi in the direction 53° west of north. What is the magnitude and the direction of the final leg that will bring it to the starting point?

2.4 A vector displacement **A** in the x-y plane has an x component of 10 m. The angle between the y axis and the vector **A** is 37°. What is the magnitude of the vector **A**?

2.5 A force of 100 lb acts on an object at an angle of 20° with respect to the x axis, and a force of 300 lb acts at an angle of 60° with respect to the x axis. What single force must be applied at what angle to be the equivalent of these two forces?

(*Answer*: 382 lb, 50.3°.)

2.6 In problem 2.5 an additional 200 lb acts at 215°. What single force at what angle will be the equivalent of these three forces?

2.7 Consider the case discussed in Example 2-1, except that $\mathbf{F}_1 = \mathbf{F}_2 = 50$ lb. The two men exerting these forces on the box ask a small boy to push on the box while they pull it so that it moves only in the x direction with a net force in that direction of 90 lb. With how large a force and in what direction does the boy push on the box?

(*Answer*: 8.4 lb, −37°.)

2.8 Express the vectors of problem 2.1 in **i**, **j**, **k** notation.

2.9 Express the vectors of problem 2.2 in **i**, **j**, **k** notation and perform the summation in that notation.

(*Answer*: $11.3\mathbf{i} + 13.8\mathbf{j}$.)

2.10 What is the magnitude and the angle of the resultant of vectors **A**, **B** and **C**, where $\mathbf{A} = 2\mathbf{i} + 3\mathbf{j}$, $\mathbf{B} = 4\mathbf{i} - 2\mathbf{j}$, and $\mathbf{C} = -\mathbf{i} + \mathbf{j}$?

2.11 The sum of three vectors **A**, **B**, and **C** is equal to vector **R**. If $\mathbf{A} = 2\mathbf{i} - 3\mathbf{j}$, $\mathbf{B} = -\mathbf{i} + 2\mathbf{j}$, and $\mathbf{R} = -2\mathbf{i} + 3\mathbf{j}$, what are the components of vector **C**? Make a sketch of vector **C** on a cartesian system, find its magnitude and the angle it makes with the x axis.

2.12 The resultant of vectors **A**, **B**, and **C** is $2\mathbf{i} + \mathbf{j}$. If $\mathbf{A} = 6\mathbf{i} - 3\mathbf{j}$ and $\mathbf{B} = 2\mathbf{i} + 5\mathbf{j}$, find the components, the magnitude, and the angle of vector **C**.

2.13 Vectors **A** and **B** have magnitudes of 3 m and 4 m, respectively, and are 30° apart. Find $\mathbf{A} \cdot \mathbf{B}$ and $\mathbf{A} \times \mathbf{B}$.

(*Answer*: 10.4 m², 6.0 m².)

2.14 If the vectors **A** and **B** of problem 2.13 are 150° apart, find $\mathbf{A} \cdot \mathbf{B}$ and $\mathbf{A} \times \mathbf{B}$.

2.15 Find the dot and cross products of vectors **A** and **B** of problem 2.13 if they are 0° apart. If they are 180° apart.

2.16 Find the vector $\mathbf{A} \times \mathbf{B}$ if $\mathbf{A} = \mathbf{i} - 3\mathbf{j} + 2\mathbf{k}$ and $\mathbf{B} = -2\mathbf{i} - \mathbf{j} + 3\mathbf{k}$.

2.17 Find the dot product $\mathbf{A} \cdot \mathbf{B}$ if $\mathbf{A} = 3\mathbf{i} + 4\mathbf{j} - \mathbf{k}$ and $\mathbf{B} = -3\mathbf{j} - 12\mathbf{k}$. What are the magnitudes of vector **A** and vector **B**?

2.18 Use the dot product to find the angle between vectors **A** and **B** in problem 2.17.

2.19 Find the angle between the vectors $\mathbf{A} = 4\mathbf{i} + 3\mathbf{j}$ and $\mathbf{B} = 6\mathbf{i} - 3\mathbf{j}$.

(*Answer*: 63.4°.)

2.20 What is the angle between the vector $\mathbf{A} = 3\mathbf{i} - 7\mathbf{j}$ and the x-axis?

2.21 Find the angles between the vector $\mathbf{A} = 2\mathbf{i} - 3\mathbf{j} + 5\mathbf{k}$ and the x, y, and z axes, respectively.

(*Answer*: 71.1°, 119.1°, 35.8°.)

2.22 Consider a vector $\mathbf{A} = 4\mathbf{i} - 9\mathbf{j}$. Find a vector in the x-y plane that is perpendicular to **A**.

(*Answer*: a $(2.25\mathbf{i} + \mathbf{j})$, where a is an arbitrary constant.)

2.23 A vector $\mathbf{R} = 9\mathbf{i} - 12\mathbf{j}$ can be expressed as a linear combination of two vectors **A** and **B**; namely, $\mathbf{R} = C_1\,\mathbf{A} + C_2\,\mathbf{B}$, where C_1 and C_2 are two scalar constants. If $\mathbf{A} = 5\mathbf{i} - 3\mathbf{j}$ and $\mathbf{B} = -\mathbf{i} + 12\mathbf{j}$, what are C_1 and C_2?

(*Answer*: $C_1 = 1.68$, $C_2 = -0.58$.)

2.24 The resultant **R** of two vectors **A** and **B** has half the magnitude of **A** and is perpendicular to **B**.

If the magnitude of **A** is 5, what is the magnitude of **B** and the angle between **A** and **B**?

2.25 Many theorems in geometry can be readily proved by vector algebra. Consider the triangle *OAB* in Fig. 2-16. Show that the line joining the midpoint of side *OA* to the midpoint of side *AB* is parallel to *OB* and its length is half that of *OB*. (*Hint:* Make vectors out of *OA*, *OB*, *AB*, *CA*, *AD*, and *CD* and find relations between these vectors)

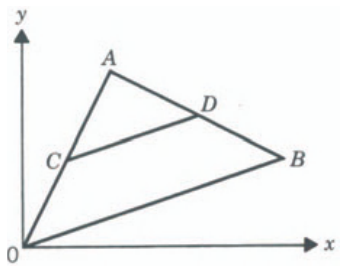


FIGURE 2-16
Problem 2.25.

2.26 Consider the line obtained by joining the origin and the point $x = 5$, $y = 3$ (see Fig. 2-17). Find the perpendicular distance h from a point P with coordinates $x = 1$, $y = 7$ to that line.

(*Answer*: 5.49.)

2.27 The magnitude of vectors **A** and **B** are 4 and 10, respectively. The magnitude of the resultant **R** is 12. What is the angle between **A** and **B**?

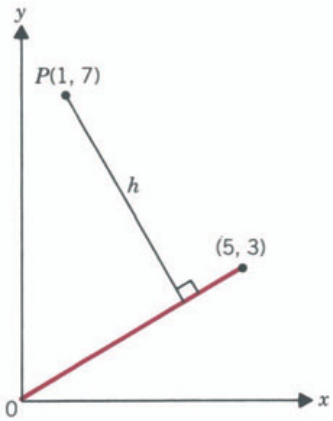


FIGURE 2-17
Problem 2.27.

2.28 What is the area of the triangle formed by joining the following three points: $x = 0$ m, $y = 0$ m; $x = 3$ m, $y = 4$ m; $x = 7$ m, $y = 2$ m. Recall that the area of a triangle is Area = 1/2 base $\times$ height. (*Hint*: Make vectors out of two of the sides of the triangle and consider the magnitude of the cross product of those two vectors.)

(*Answer*: 11 m^2.)

CHAPTER 3

Uniformly Accelerated Motion

3.1 INTRODUCTION

In this chapter we introduce certain vector quantities—position, displacement, velocity and acceleration—used to describe the motion of a body. We define these quantities and discuss the mathematical relations between them. We then derive specific functional relations between them and time (a scalar quantity) for the case where the object moves in a straight line with constant acceleration. The chapter concludes with a discussion of projectile motion, one of the simplest types of two-dimensional motion.

3.2 SPEED AND VELOCITY

Two words in English, "speed" and "velocity," are used interchangeably to indicate how fast a body is moving. In physics we make a distinction between them. The word "speed" is defined as a scalar quantity and "velocity" is a vector quantity. Thus, the average speed (where average will be represented by a bar on top of the quantity involved) is the distance traveled in any direction, Δs, divided by the time Δt, or

$$\overline{\text{speed}} = \frac{\Delta s}{\Delta t} \tag{3.1}$$

where

$$\Delta(\text{anything}) = \text{final value} - \text{initial value}$$

Velocity is defined differently. Consider a particle moving in space. Let the particle be at point P in Fig. 3-1 at some initial time t_0 and at point P' some later time t_f. The initial position of the particle can be specified by a *position vector* $\mathbf{r}_0$ obtained by drawing an arrow from the origin of the coordinate system to point P. Similarly, the position at the later time is specified by a second position vector $\mathbf{r}_f$ that results when an arrow is drawn from the origin to point P'. The position at any other point in the motion is specified by a corresponding position vector $\mathbf{r}$. We can now define the *displacement vector* $\Delta\mathbf{r}$ as the vector difference between the final and the initial position vectors, namely, $\Delta\mathbf{r} = \mathbf{r}_f - \mathbf{r}_0$ (see Fig. 3-1). Correspondingly, we define the *average velocity* $\overline{\mathbf{v}}$ as the ratio of the displacement vector to the time taken for the displacement to occur, namely,

$$\overline{\mathbf{v}} = \frac{\mathbf{r}_f - \mathbf{r}_0}{t_f - t_0} = \frac{\Delta\mathbf{r}}{\Delta t} \tag{3.2}$$

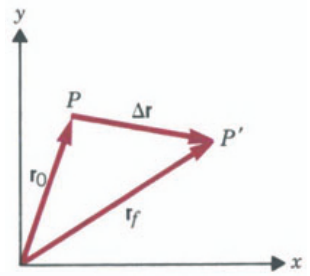


FIGURE 3-1
The displacement vector $\Delta\mathbf{r}$ is obtained by drawing an arrow from the initial position vector $\mathbf{r}_o$ to the final position vector $\mathbf{r}_f$.

The distinction between speed and velocity is difficult to grasp at first, but it is extremely important. Consider the walk taken in Fig. 2-1. Suppose it took 1 h. Then, by definition, Eq 3.1, the average speed of walking was $\Delta s/\Delta t =$

(4.0 mi + 5.0 mi)/1 h = 9 mi/h, whereas the average velocity was $\Delta\mathbf{r}/\Delta t$ = 6.4 mi/1 h = 6.4 mi/h in the direction 39° north of east. Consider an even more extreme example. Suppose a race car is traveling around a circular track of 1-mi diameter and its speedometer reads 100 mi/h. This is the speed. The time taken to reach any point is, from Eq. 3.1, $\Delta t = \Delta s/\overline{\text{speed}}$. Because the track length is $\pi \times$ diameter = 3.14 mi, the time to complete one circuit is

$$\Delta t = \frac{3.14 \text{ mi}}{100 \text{ mi/h}} = 3.14 \times 10^{-2} \text{ h}$$

and the time to go halfway around is 1.57×10^{-2} h. However, the car's average velocity by the definition of Eq. 3.2 depends on its position. When the car has gone halfway around, the vector displacement from the starting point is the diameter or 1 mi. Hence, its average velocity to that point is

$$\bar{\mathbf{v}} = \frac{1 \text{ mi}}{1.57 \times 10^{-2} \text{ h}} = 63.7 \text{ mi/h}$$

In one complete circuit, as the car passes the starting point its vector displacement is zero and hence

$$\bar{\mathbf{v}} = \frac{0 \text{ mi}}{3.14 \times 10^{-2} \text{ h}} = 0 \text{ mi/h}$$

This seeming contradiction has occurred because we have taken large displacements for $\Delta\mathbf{r}$. If we shrink the displacement to a minute amount by taking the limit

$$\mathbf{v} = \lim_{\Delta t \to 0} \frac{\Delta \mathbf{r}}{\Delta t} = \frac{d\mathbf{r}}{dt} \tag{3.3}$$

$$\mathbf{v} = \frac{d\mathbf{r}}{dt}$$

then the magnitude of the velocity, which is now called the *instantaneous velocity*, at any point on the track will equal the speed. This can be seen in Fig. 3-2, where we notice that as Δs becomes smaller, the difference between Δs and the corresponding $\Delta\mathbf{r}$ decreases. We should also note that in the limit

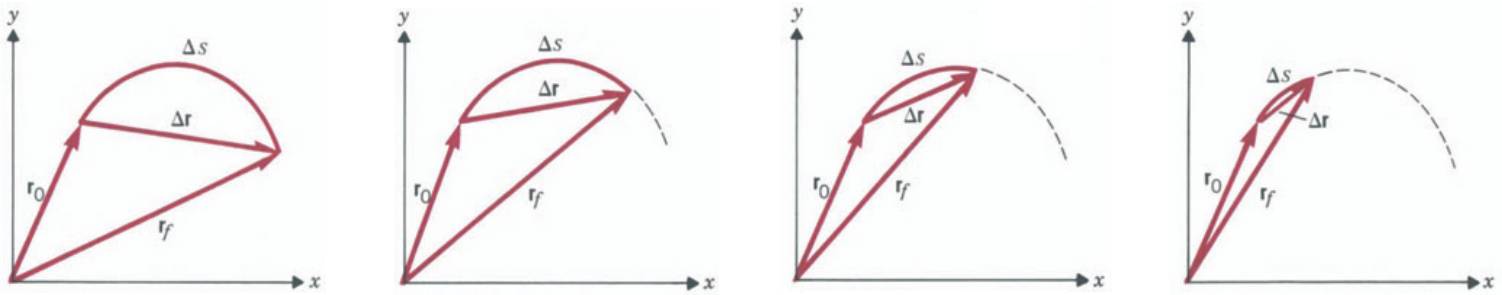


FIGURE 3-2

A curved path of a car traveling clockwise. Δs is the distance traveled by the car, and $\Delta\mathbf{r}$ is the displacement vector between the position of the car $\mathbf{r}_f$ at some instant and the position $\mathbf{r}_o$ at the initial time. As the distance Δs becomes smaller, its value approaches $\Delta\mathbf{r}$.

where $\Delta\mathbf{r}$ becomes infinitesimally small, it becomes tangential to the path, and therefore the direction of the instantaneous velocity is the tangent to the path. Thus, while the magnitude of the instantaneous velocity remains equal to the speed, the direction part of the instantaneous velocity is changing. Velocity is a vector because it is equal to a vector displacement divided by time, which is scalar, and the division of a vector by a scalar does not remove the vector property. We should note that by definition, Eq. 3.3, the instantaneous velocity $\mathbf{v}$ is the first derivative of the position vector with respect to time. It should also be pointed out that because Eq. 3.3 is a vector equation, it holds for each of the cartesian components of the vectors $\mathbf{v}$ and $\mathbf{r}$, namely,

$$v_x = \frac{dx}{dt}, \quad v_y = \frac{dy}{dt}, \quad \text{and} \quad v_z = \frac{dz}{dt}$$

where v_x, v_y, and v_z are the cartesian components of $\mathbf{v}$ and x, y, and z are those of $\mathbf{r}$.

3.3 ACCELERATION

In the preceding section we introduced a convention in which Δ is a measurable change between a final value and an initial value and d is used for an infinitesimally small change.

If there is a velocity change $\Delta\mathbf{v}$ in a certain time Δt, we define the average acceleration as

$$\bar{\mathbf{a}} = \frac{\Delta\mathbf{v}}{\Delta t} \tag{3.4}$$

or

$$\bar{\mathbf{a}} = \frac{\mathbf{v}_f - \mathbf{v}_0}{t_f - t_0}$$

where the subscripts f and o represent final and initial values, respectively. Usually in a problem we start our stopwatch at $t_0 = 0$, so the elapsed time is simply t_f and we drop the subscript f. We may define an instantaneous acceleration as

$$\mathbf{a} = \lim_{\Delta t \to 0} \frac{\Delta\mathbf{v}}{\Delta t} = \frac{d\mathbf{v}}{dt} \tag{3.5}$$

$\mathbf{a} = \dfrac{d\mathbf{v}}{dt}$

which is the first derivative of $\mathbf{v}$ with respect to time. Substituting Eq. 3.3 for $\mathbf{v}$, we write

$$\mathbf{a} = \frac{d}{dt}\left(\frac{d\mathbf{r}}{dt}\right) = \frac{d^2\mathbf{r}}{dt^2} \tag{3.6}$$

which is the second derivative of $\mathbf{r}$ with respect to time.

Example 3-1

The position of a body on the x axis varies as a function of time according to the following equation

$$x(\text{meters}) = (3t + 2t^2)\text{m}$$

Find its velocity and acceleration when $t = 3$ sec.

Solution Because the body moves in a straight line, $\mathbf{r} = x$. From Eq. 3.3

$$v = \frac{dx}{dt} = \frac{d}{dt}(3t + 2t^2) = (3 + 4t) \text{ m/sec}$$

The velocity of the body at $t = 3$ sec is therefore

$$v\ (t = 3 \text{ sec}) = 3 + 4 \times 3 = 15 \text{ m/sec}$$

From Eq. 3.5,

$$a = \frac{dv}{dt} = \frac{d}{dt}(3 + 4t) = 4 \text{ m/sec}^2$$

Notice that a is a constant, and therefore $a(t = 3 \text{ sec}) = 4 \text{ m/sec}^2$.

3.4 LINEAR MOTION

Because displacement, velocity, and acceleration are vectors, we may treat them by the method of cartesian components introduced in Chapter 2. First, let us consider motion only in the direction of a single component, for example, the x direction, that is, motion in a straight line.

If we start timing an object moving in the x direction when it starts from or is passing the $x = 0$ point, we may write Eq. 3.2 as

$$\bar{v}_x = \frac{x - 0}{t - 0}$$

$x = \bar{v}_x t$

or

$$x = \bar{v}_x t \tag{3.7}$$

Because in this section we will be talking about motion in one direction, we will drop the subscript x from the velocity.

Equation 3.7 results from the definition of average velocity; thus it holds in all cases whether or not the acceleration is constant. In the remainder of this chapter, we will consider only *constant acceleration*.

The derivative of a variable, for example, the velocity v, with respect to a second variable, for example, time t, represents the instantaneous *rate of change* of the first variable (v) with respect to the second (t). Thus, the acceleration as defined in Eq. 3.5 is the rate of change of the velocity with time.

If the acceleration is constant, the change in the velocity during the first, second, third, and all succeeding seconds of the motion will be the same and equal to the acceleration a. Thus, if the motion lasts t seconds, the change in the velocity $\Delta v = v - v_0 = at$, where v is the final velocity and v_0 is the initial velocity. We can rewrite this result as

$$v = v_0 + at \tag{3.8}$$

If velocity v is plotted against time t, it is seen that Eq. 3.8 is a straight line, as indicated in Fig. 3-3. The slope of this line is the constant acceleration a.

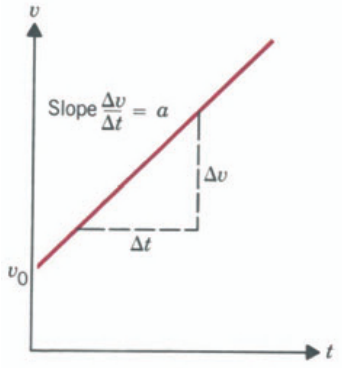


FIGURE 3-3

Plot of v versus t for constant acceleration.

With a little bit of thought, we can obtain another important relation. When the velocity increases at a constant rate as in Eq. 3.8 and Fig. 3-3, the average velocity is one half the sum of the initial velocity v_0 and the final velocity v, namely,

$$\bar{v} = \frac{v + v_0}{2} \tag{3.9}$$

$$x = \frac{v + v_0}{2} t$$

and Eq. 3.7 becomes

$$x = \frac{v + v_0}{2} t \tag{3.10}$$

$$x = \frac{v + v_0}{2} t$$

The three equations (3.7), (3.8), and (3.9) define linear motion for constant acceleration. Often, however, at least two of these, and sometimes all three, must be used to solve a problem. It is convenient, therefore, to combine these three equations to obtain two auxiliary ones and have all five available (thereby avoiding the necessity of solving simultaneous equations). We obtain the auxiliary equations in the following way.

From Eq. 3.8 write

$$t = \frac{v - v_0}{a}$$

and substituting it into Eq. 3.10 we obtain

$$x = \frac{(v + v_0)}{2} \frac{(v - v_0)}{a}$$

and

$$v^2 - v_0^2 = 2ax \tag{3.11}$$

If we substitute Eq. 3.8 for v in Eq. 3.10, we obtain

$$x = \frac{v_0 + at + v_0}{2} t$$

and

$$x = v_0 t + \frac{1}{2} a t^2 \tag{3.12}$$

We may derive these equations more formally by integration. By definition

$$a = \frac{dv}{dt} \tag{3.5}$$

Rearranging terms and integrating, we write

$$\int_{v_0}^{v} dv = \int_0^t a \, dt$$

This equation holds in general whether or not acceleration is a constant. In the present case, acceleration is taken as constant, so a can be taken out of the integral and we write

$$\int_{v_0}^{v} dv = a \int_0^t dt$$

This integrates to

$$v - v_0 = at$$

and

$$v = v_0 + at \tag{3.8}$$

$v = v_0 + at$

From the definition

$$v = \frac{dx}{dt} \tag{3.3}$$

$$\int_{x_0}^{x} dx = \int_0^t v \, dt$$

Substitute Eq. 3.8 for v

$$\int_{x_0}^{x} dx = \int_0^t (v_0 + at) \, dt = v_0 \int_0^t dt + a \int_0^t t \, dt$$

$$x - x_0 = v_0 t + \frac{1}{2} a t^2 \tag{3.12}$$

$x - x_0 = v_0 t + \frac{1}{2} a t^2$

Note that in this formulation of Eq. 3.12 we have not required that $x = 0$ at $t = 0$ as in the previous algebraic derivations.

We may use the chain rule to write

$$a = \frac{dv}{dt} = \frac{dv}{dx}\frac{dx}{dt} = v\frac{dv}{dx} \tag{3.13}$$

or

$$\int_{v_0}^{v} v\,dv = a\int_{x_0}^{x} dx$$

$$\frac{v^2}{2} - \frac{v_0^2}{2} = a(x - x_0)$$

$$v^2 - v_0^2 = 2a(x - x_0) \qquad (3.11)$$

$v^2 - v_0^2 = 2a(x - x_0)$

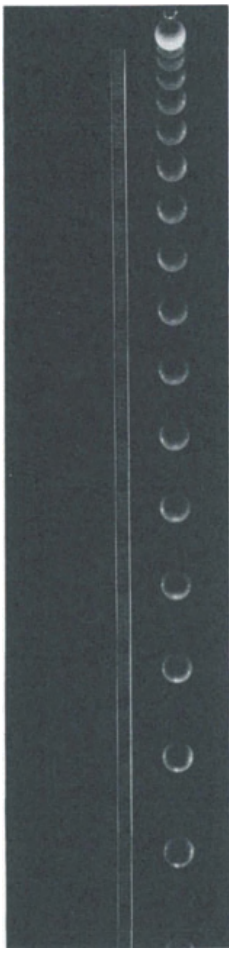
Equations 3.7 through 3.12 have been derived for motion in the x direction. Similar equations can simply be written for motion in the y and z directions when the components of the acceleration in these directions are also constant.

There is one important thing to be noted here. In the solution of motion problems we must assign vector directions. Suppose we observe a boy throwing a ball as we look through a transparent piece of graph paper and draw lines of motion and displacement. We could lie on our side or stand on our head and draw the lines without having any effect on the boy and his ball. Therefore, choosing a particular coordinate system is a matter of personal convenience. The upward direction could be the positive y direction or the negative y direction, or even a direction at an angle on the graph paper, although we will always try to choose a system that will minimize the calculational steps. It is important to note that once we choose a coordinate system, all parameters have their vector direction controlled by it. If we choose the positive y direction as up and the boy throws the ball straight up, then the vector displacement from the ground to its highest position is positive. During its upward travel, because velocity is the displacement divided by the scalar time, it too is positive. The only motion is in the y direction, so we therefore use y, v_y, and a_y in the equations previously derived. Thus, Eq. 3.8 is

$$v_y = v_{0y} + a_y t \qquad (3.8')$$

We observe that in throwing the ball upward the largest value for the magnitude of the y velocity occurs as it leaves the boy's hand; then the ball begins to slow down until the upward velocity is zero. The only way that this can occur is if a_y in Eq. 3.8′ is negative. Now a_y is the acceleration caused by the force of gravity acting on the ball, and we will see in the next chapter that because the force of gravity is downward then the corresponding acceleration must also be downward. *The acceleration caused by gravity is usually written as the symbol* g *and has the approximate sea-level value* $g = 9.8\ m/sec^2$.

When solving problems, the best approach is to tabulate what is known and what is to be found and select the appropriate equation.



Multiflash photograph of a falling ball. Note the increase in the distance traveled by the ball between flashes as it falls. This reflects an increase in the velocity of the ball caused by the downward-directed gravitational acceleration.

Example 3-2

A boy throws a ball upward with an initial velocity of 12 m/sec. How high does it go?

Solution We choose the starting point as the origin and the upward direction as positive. Because velocity is a vector displacement divided by time, upward velocity is also positive. The force of gravity is in the negative y direction, so the sign of the acceleration is therefore negative. First list what is known and what is to be found

$$v_{0y} = 12 \text{ m/sec}, \quad v_y = 0 \text{ (at its highest point)}, \quad a_y = g = -9.8 \text{ m/sec}^2$$

$$y = ?$$

We select the y equivalent of Eq. 3.11 because all the quantities in that equation are known except y, the quantity that we want to find

$$v_y^2 - v_{0y}^2 = 2a_y y$$

Solving for y, we write

$$y = \frac{v_y^2 - v_{0y}^2}{2\,a_y}$$

Substituting the numerical values for the quantities in the equation,

$$y = \frac{0 - (12 \text{ m/sec})^2}{2\,(-9.8 \text{ m/sec}^2)}$$

$$= 7.3 \text{ m}$$

Example 3-3

A boy throws a ball upward with an initial velocity of 12 m/sec and catches it when it returns. How long was it in the air?

Solution As in the previous example, we choose the starting point as the origin and the upward direction as positive.

$$v_{0y} = 12 \text{ m/sec}, \quad a_y = -9.8 \text{ m/sec}^2, \quad y = 0 \text{ (vector displacement is zero because it returns to his hand)}, \quad t = ?$$

Select Eq. 3.12

$$y = v_{0y}t + \frac{1}{2}a_y t^2$$

Using the fact that $y = 0$, Eq. 3.12 becomes

$$0 = v_{0y}t + \frac{1}{2}a_y t^2$$

We see immediately that if we divide both sides of the equation by t, we obtain

$$0 = v_{0y} + \frac{1}{2}a_y t$$

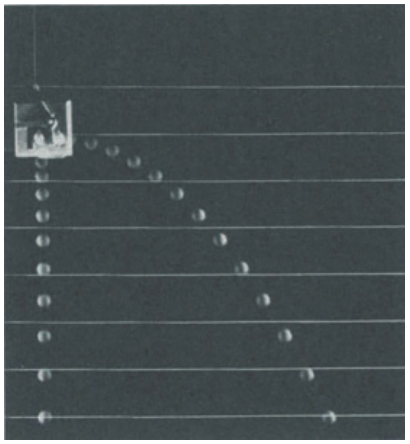
and

$$t = -\frac{2\,v_{0y}}{a_y}$$

$$= -\frac{2 \times 12 \text{ m/sec}}{-9.8 \text{ m/sec}^2}$$

$$= 2.45 \text{ sec}$$

In this example, the ball returned to its starting point, so the displacement y was zero. This simplified Eq. 3.12, because dividing both sides by t linearized the equation. If the ball had landed on a roof, then the left side of Eq. 3.12 would not be zero and the equation to be solved would be quadratic.



Multiflash photograph of two falling balls: One released from rest and the other launched with an initial horizontal velocity. The vertical position of the two balls at any given time (time of the flashes) is the same for both, indicating that the vertical motion is unaffected by the initial horizontal velocity given to the second ball.

3.5 PROJECTILE MOTION

We have treated motion in one dimension in the preceding section. Suppose we have a smooth, frictionless wall in the x-y plane. If we set an object in motion along the wall, we find experimentally that the object is accelerated downward in the y-vector direction but that there is no acceleration in the x-vector (horizontal) direction. That is, the object moves in the x direction with its constant initial x velocity but its y velocity is increasing downward owing to the acceleration of gravity. If we now perform the same experiment with an imaginary wall, we have what is called *projectile motion*. The characteristic in the coordinate system in which we are working is that the x and y motions and velocities are at right angles to each other and that there is an acceleration only in the y direction, a_y; there is no acceleration in the x direction. The equations of motion in the x and y directions are therefore

$$x = v_{0x}t$$

$$y = v_{0y}t + \frac{1}{2}a_y t^2$$

In view of the foregoing, projectile problems are treated as two separate linear motion problems, one in the x direction and another in the y direction, with only time as the common element.

Example 3-4

A ball moving at 2 m/sec rolls off of a 1-m-high table, Fig. 3-4. How far horizontally from the edge of the table does it land?

Solution The ball will continue moving in the x direction for as long as it is in the air. We can use Eq. 3.7 to determine the x coordinate of the ball as it lands

$$x_f = \bar{v}_x t_f$$

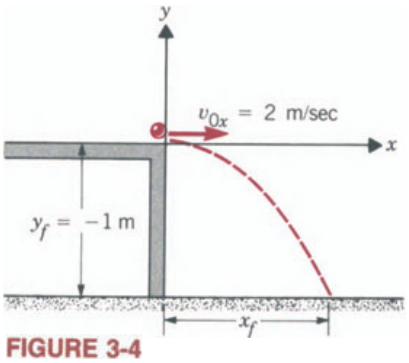


FIGURE 3-4

Example 3-4. The edge of the table is chosen as the origin of the $x-y$ coordinate system.

where t_f is the time that the ball is in the air. Note here that, because the velocity in the x direction is unchanged, $\bar{v}_x = v_{0x}$ and therefore

$$x_f = 2 \text{ m/sec } t_f$$

t_f is the time when the y coordinate of the ball becomes -1 m. We have

$$y_f = -1 \text{ m}, \quad v_{0y} = 0, \quad a_y = -9.8 \text{ m/sec}^2, \quad t_f = ?$$

Using Eq. 3.12,

$$y = v_{0y}t + \frac{1}{2}a_y t^2$$

Because $v_{0y} = 0$, this equation becomes

$$y = \frac{1}{2}a_y t^2$$

and

$$t = \pm\sqrt{\frac{2y}{a_y}}$$

$$t_f = \pm\sqrt{\frac{2(-1 \text{ m})}{-9.8 \text{ m/sec}^2}}$$

$$= \pm 0.45 \text{ sec}$$

and because in deriving Eqs. 3.9 through 3.12 we chose $t = 0$ as the initial time, only the positive root is acceptable, therefore,

$$x_f = 2 \text{ m/sec} \times 0.45 \text{ sec}$$

$$= 0.9 \text{ m}$$

Let us examine a specific case of projectile motion along level ground. We will find the general formula for the distance that a person can throw a ball or that a gun can fire a projectile. The variables are shown in Fig. 3-5*a* and the initial velocity components in Fig. 3-5*b*. The distance x that the projectile travels just before it strikes the ground is

$$x_f = \bar{v}_x t_f = v_{0x} t_f \quad (\text{because } \bar{v}_x = v_{0x})$$

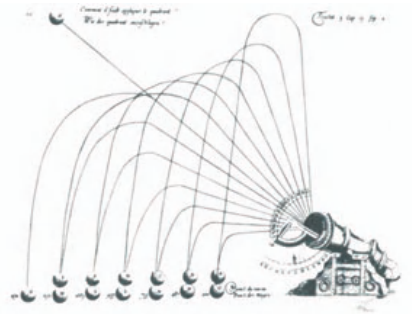
From Fig. 3-5*b*

$$v_{0x} = v_0 \cos\theta$$

and therefore

$$x_f = v_0 \cos\theta \, t_f$$

We determine the time in the air from the y-direction problem when the projectile is thrown upward with an initial velocity of $v_{0y} = v_0 \sin\theta$ and is acted on by the acceleration of gravity, $a_y = g = -9.8$ m/sec^2. At the end of



Cannonball trajectories for various launching angles as conceived by Diego Ufano in 1621.

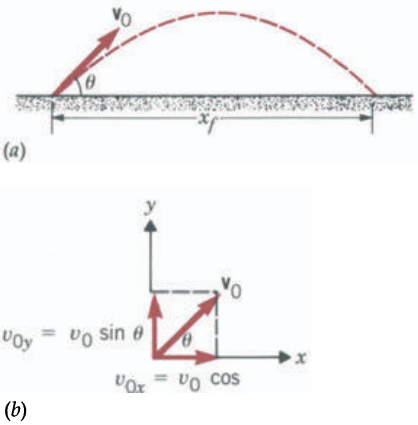


FIGURE 3-5
(*a*) Projectile motion on level ground with $\mathbf{v}_o$ and θ the initial velocity and angle, respectively. (*b*) The x and y components of the initial velocity $\mathbf{v}_o$.

its trajectory the vector displacement is $y = 0$, so we may write

$$y_f = 0, \quad v_{0y} = v_0 \sin \theta, \quad a_y = -9.8 \text{ m/sec}^2, \quad t_f = ?$$

using

$$y = v_{0y}t + \frac{1}{2}a_y t^2 \tag{3.14}$$

$$0 = v_0 \sin \theta \, t_f - \frac{1}{2} g t_f^2$$

$$t_f = \frac{2}{g} v_0 \sin \theta$$

Substitute this for t_f in the equation for x-direction motion

$$x_f = \frac{v_0^2}{g} 2 \sin \theta \cos \theta$$

Substitute the trigonometric relation $2 \sin \theta \cos \theta = \sin 2\theta$ and obtain

$$x_f = \frac{v_0^2}{g} \sin 2\theta \tag{3.15}$$

We can readily find from Eq. 3.15 the angle at which the projectile should be thrown (or fired) to achieve a maximum distance in the x direction for a fixed value of v_0. The only variable is the angle in the term $\sin 2\theta$, and the sine has a maximum value of unity when the argument is 90°. Therefore, x_f of Eq. 3.15 is maximum when $2\theta = 90°$ or $\theta = 45°$.

Note that Eq. 3.15 is valid only when the projectile returns to the starting level, because we set $y_f = 0$ in Eq. 3.14. If it returns to some other level, then the quadratic equation in time must be solved (see problems 3.16 and 3.17).

Example 3-5

A boy stands on the edge of a roof 10 m above the ground and throws a ball with a velocity of 15 m/sec at an angle of 37° above the horizontal. How far from the building does it land? See Fig. 3-6.

Solution Let us choose the edge of the roof as the origin of the coordinate system. There is no acceleration in the x direction, so we may simply write the following for x distance

$$x_f = \bar{v}_x t_f = v_{0x} t_f$$

$$v_{0x} = v_0 \cos 37° = 15 \text{ m/sec} \times 0.8 = 12 \text{ m/sec}$$

$$x_f = 12 \text{ m/sec} \, t_f$$

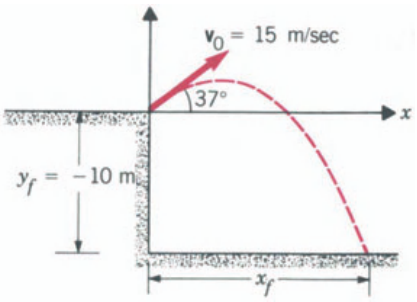


FIGURE 3-6
Example 3-5.

We now need to find the time in the air, which is a motion problem in the y direction only.

$$y_f = -10 \text{ m}, \quad v_{0y} = 15 \text{ m/sec} \sin 37° = 15 \text{ m/sec} \times 0.6 = 9 \text{ m/sec}$$

$$a_y = -9.8 \text{ m/sec}^2, \quad t_f = ?$$

We use Eq. 3.12

$$y = v_{0y}t + \frac{1}{2}a_y t^2$$

If $t = t_f$ when $y = y_f = -10$ m, this equation can be written as

$$\frac{1}{2}a_y t_f^2 + v_{0y} t_f - y_f = 0$$

Solving this quadratic equation for t_f

$$t_f = \frac{-v_{0y} \pm \sqrt{v_{0y}^2 - (4)\left(\frac{1}{2}a_y\right)(-y_f)}}{(2)\left(\frac{1}{2}a_y\right)}$$

$$= \frac{-v_{0y} \pm \sqrt{v_{0y}^2 + 2\,a_y\,y_f}}{a_y}$$

Substituting the numerical values for y_f, v_{0y}, and a_y

$$t_f = \frac{-(9 \text{ m/sec}) \pm \sqrt{(9 \text{ m/sec})^2 + (2)(-9.8 \text{ m/sec}^2)(-10 \text{ m})}}{-9.8 \text{ m/sec}^2}$$

$$t_f = 2.6 \text{ sec}, \quad -0.78 \text{ sec}$$

Because we have started timing when the ball is thrown, the negative time solution is rejected because it has no physical meaning to this problem. Substitute the positive time of 2.6 sec into the equation of motion in the x direction and obtain

$$x_f = 12 \text{ m/sec} \times 2.6 \text{ sec} = 31.2 \text{ m}$$

It will be instructive to find the magnitude of the velocity and the angle at which the ball strikes the ground. This is obtained from the components of the velocity just before it hits, as shown in Fig. 3-7. We see from the vector component method that the ball's vector velocity just before striking the ground is given by the final value of its components, v_{fx} and v_{fy}. Because we have a right triangle, we may use the pythagorean theorem

$$v_f^2 = v_{fx}^2 + v_{fy}^2$$

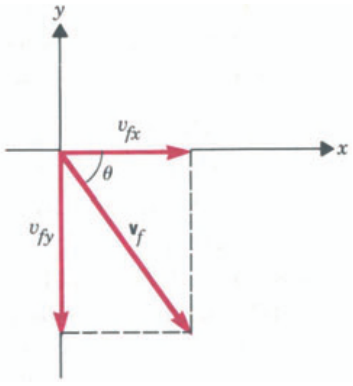


FIGURE 3-7
Example 3-5. Diagram to calculate the angle at which the projectile strikes the ground.

and the angle θ at which it strikes the ground is

$$\theta = \arctan \frac{v_{fy}}{v_{fx}}$$

$v_{fx} = 12$ m/sec because it is unchanged during the ball's flight. We must therefore find v_{fy}. We obtain this from Eq. 3.8 in the y direction.

$$v_{0y} = 9 \text{ m/sec}, \quad t_f = 2.6 \text{ sec}, \quad a_y = -9.8 \text{ m/sec}^2, \quad v_{fy} = ?$$

$$v_{fy} = v_{0y} + a_y t_f$$

$$v_{fy} = 9 \text{ m/sec} - 9.8 \text{ m/sec}^2 \times 2.6 \text{ sec}$$

$$v_{fy} = -16.5 \text{ m/sec}$$

Then

$$v_f = \sqrt{(12 \text{ m/sec})^2 + (-16.5 \text{ m/sec})^2} = 20.4 \text{ m/sec}$$

$$\theta = \arctan \frac{|-16.5 \text{ m/sec}|}{12 \text{ m/sec}} = 54°$$

where θ is the angle indicated in Fig. 3-7.

PROBLEMS

3.1 A student drives to college 15 km away from home in half an hour. After classes, he returns home in 20 min. Find (a) the average speed on his way to college, (b) the average speed for the round trip, (c) his average velocity for the entire trip.

3.2 The position of a particle moving along the x axis is given by $x = 3 + 17t - 5t^2$, where x is in meters and t is in seconds. (a) What is the position of the particle at $t = 1$, 2, and 3 sec? (b) At what time does the particle return to the origin? (c) What is the instantaneous velocity at $t = 1$, 2, and 3 sec? (d) At what time is the instantaneous velocity of the particle zero? (e) What is the velocity of the particle as it passes through the origin? (f) What is the acceleration of the particle as it passes the origin?

3.3 The position of a particle moving in a straight line is given by $x = 5 + 2t + 4t^2 - t^3$. (a) Find an expression for the instantaneous velocity as a function of time. (b) Find the position of the particle at $t = 0$, 1, 0.1, and 0.01 sec. (c) What is the average velocity between $t = 0$ sec and $t = 1$ sec, between $t = 0$ sec and $t = 0.1$ sec, and between $t = 0$ sec and $t = 0.01$ sec? (d) What is the instantaneous velocity at $t = 0$ sec? (e) What conclusion do you draw from the answers in (c) and (d)?

3.4 A car is driving east at 60 km/h, it then makes a turn and travels north at 50 km/h. If it takes 2 sec to make the turn, what is the average acceleration of the car as a result of the turn?

(*Answer*: 10.85 m/sec^2, directed 39.8° north of west.)

3.5 Consider the particle of problem 3.3. (a) Find an expression for the acceleration of the particle as a function of time. (b) What is the instantaneous velocity of the particle at $t = 0$, 1, 0.1, and 0.01 sec. (c) What is the average acceleration between $t = 0$ sec and $t = 1$ sec, between $t = 0$ sec and $t = 0.1$ sec, between $t = 0$ sec and $t = 0.01$ sec? (d) What is the instantaneous acceleration at $t = 0$ sec. (e) What conclusion can you draw from the answers in (c) and (d)?

3.6 A car starts from rest and accelerates uniformly to a speed of 25 m/sec in 8 sec. (a) What is the acceleration? (b) How far did it travel in the 8 sec?

(*Answer*: (a) 3.13 m/sec^2, (b) 100 m.)

3.7 A rocket starting from rest rises to a height of 20,000 m in 60 sec. (a) What was the average velocity of the rocket? (b) Assuming that the acceleration was constant, what was the acceleration of the rocket? (c) What was the velocity and the height of the rocket after 30 sec?

3.8 A boy stands on the edge of a building 10 m above the ground and throws a ball upward with an initial velocity of 12 m/sec. It misses the roof on the way down and falls to the ground. Find how long the ball was in the air and its velocity just before it strikes the ground. (*Hint*: take $y = 0$ at $t = 0$ and y final as -10 m).

(*Answer*: 3.11 sec, -18.44 m/sec.)

3.9 A car moving at 25 m/sec strikes a tree, and the tree is seen to dent the front by 0.5 m. Assume that the deceleration of the car was constant. Find the deceleration and time it took the car to stop.

3.10 A car moving with constant acceleration covers a distance of 50 m between two points in 5 sec. Its velocity as it passes the second point is 16 m/sec. (a) What is its acceleration? (b) What was its velocity as it passed the first point?

(*Answer*: (a) 2.4 m/sec^2, (b) 4.0 m/sec.)

A ball is dropped from the roof of a building. It is observed to take 0.2 sec to pass by a window 2 m high. How far is the top of the window from the roof?

(*Answer*: 4.15 m.)

A ball is dropped from a bridge 60 m above the surface of the water. One second later, a second ball is thrown down with an initial velocity v_0. Both balls strike the water at the same time. (a) How long were the balls in the air? (b) What was the initial velocity of the second ball? (c) What were the velocities of the balls as they struck the water?

(*Answer*: (a) 3.50 sec, 2.50 sec, (b) -11.76 m/sec, (c) -34.30 m/sec, -36.25 m/sec.)

A motorcycle is waiting at an intersection. As the light turns green it starts with an acceleration of 20 m/sec^2. At that same moment a car, moving with constant velocity of 120 m/sec overtakes and passes the motorcycle. (a) How far from the traffic light will the motorcycle overtake the car? (b) What is the velocity of the motorcycle at that point?

3.14 A girl drops a flowerpot from a window 50 m above the ground. At the same instant a boy directly under the flowerpot throws a stone with an upward velocity of 30 m/sec. (a) How far above the ground will the stone hit the pot? (b) How long after the flowerpot was dropped does the hit take place? (c) What is the minimum velocity with which the stone must be thrown for the hit to occur?

(*Answer*: (a) 36.4 m, (b) 1.67 sec, (c) 15.65 m/sec.)

3.15 An electron is set in motion horizontally with a velocity $v_x = 4 \times 10^6$ m/sec. How far will it fall while traveling a horizontal distance of 10 m?

3.16 A boy standing on the ground throws a ball at an angle of 37° above the horizontal with a velocity of 15 m/sec. It lands on the edge of a flat roof of a building 3 m high. How far horizontally from the boy does it strike the roof?

(*Answer*: 16.86 m.)

3.17 A boy standing on the ground throws a ball at an angle of 37° above the horizontal with a velocity of 15 m/sec. It strikes the wall of a building 16 m away. How high above the ground is the point at which the ball strikes the building?

(*Answer*: 3.32 m.)

3.18 An artillery gunner wishes to have a projectile land at a point on level ground 20,000 m away from

the gun. If the muzzle velocity is 500 m/sec and the muzzle is assumed to be at ground level, at what angle above the horizontal should the gun be aimed?

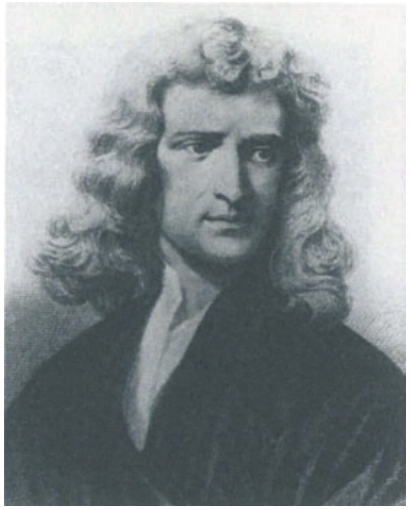
3.19 If the gun of problem 3.18 is on a hill 30 m high and the same angle of elevation is used, how far beyond the target will the projectile land?

3.20 A batter at home plate hits a baseball 1 m above the ground. The ball leaves the bat in the direction of an outfielder with a velocity of 30 m/sec at an angle of 30° above the horizontal. Half a second after the ball is hit, the outfielder 100 m away from home plate runs to catch the ball. How fast must he run to catch the ball just before it hits the ground?
(*Answer*: 7.16 m/sec.)

CHAPTER 4

Newton's Laws

4.1 INTRODUCTION



Isaac Newton (1642-1727).

In this chapter we will consider Newton's three laws of motion. Although when first propounded they were postulates, they have since been verified by experiment in so many ways that they are now considered Laws of Nature. A careful observer will note that they are not quite independent laws—that is, one implies another, although they are usually listed separately as if they were independent. There is one consistent word in these three laws and that is "body." We sometimes speak of this as the *newtonian body*. Notice that body is singular. In a given physical situation we must first define the newtonian body, which may often be a mathematical point. If the situation has two bodies, then Newton's laws must be applied separately to each. Often we solve complicated systems of solids on a computer that can remember and vary 10,000 atoms in a solid. We require an equal number of applications of Newton's laws, one for each, although the complexity of the solution of that number of simultaneous equations is often reducible by symmetry of behavior.

In addition to using the basic terms of Chapter 1—length, mass, and time—we will discuss the term *force*, which we have used a bit freely. A precautionary word might be said of these.

Length, at least on earth, can be measured by adopting a standard such as the length of the king's foot, or the length of a standard bar of metal, and comparing other lengths with it.

Time can be measured by divisions of the motion of the earth either in rotation about its own axis or in revolution about the sun. Again, a standard has been established by our environment. An early recorded question about what time really means is found in a discussion by St. Augustine in his *Confessions* around 400 A.D. "For so it is O Lord, my God, I measure it, but what it is I measure I do not know."

Mass is an even more obscure property. Newton first referred to it as *inertial* mass, that property of a body which resists being set in motion if a force is applied. But what is *force*? It is that entity which under certain conditions, to be discussed shortly, can change the state of motion of a mass when it acts on it. And thus we find ourselves in a circular argument. We may define mass if force is known, or we may define force if mass is known. The customary approach is to start with mass and define force through the motion it causes on mass. This enables us to use a combination of dimensions—length, mass, and time for force. So if we choose an arbitrary object, and agree that it be the standard of mass (the kilogram was selected), we will be able to evaluate the mass of any other body by means of Newton's second law. We will shortly see how this is done.

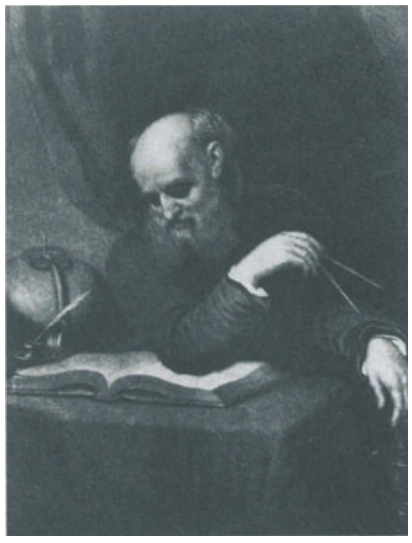
4.2 NEWTON'S LAWS

We will not state the three laws exactly as Newton did; instead we will use modern English so that we may discuss them with no misunderstanding.

First Law: Every body of matter continues in a state of rest or moves with constant velocity in a straight line unless compelled by a force to change that state.

Second Law: When net unbalanced forces act on a body, they will produce a change in the *momentum* (this concept will be defined shortly) of that body proportional to the vector sum of the forces. The direction of the change in momentum is that of the line of action of the resultant force.

Third Law: Forces, arising from the interaction of particles, act in such a way that the force exerted by one particle on the second is equal and opposite to the force exerted by the second on the first and both are directed along the line joining the two particles. (Or, as usually expressed, *action* and *reaction* are equal and opposite.)



Galileo Galilei (1564-1642).

In the first law Newton implies what we call *frames of reference*. He has implied that a body at rest with respect to an observer can be analyzed in the same way as one moving past the observer at a constant velocity. That is, the same force applied to either body will change its state of motion in the same way.

The third law of action and reaction implies that there can be no single force in isolation. A force must act on a body and, when it does, the body acts on the source of the force. Consider a bat striking a ball. The bat is a mass in motion with a certain velocity. If it strikes the ball, it exerts a force on it and changes the ball's state of motion. At the instant of striking, the ball exerts a force on the bat in the opposite direction, thereby changing the bat's state of motion by slowing it down. If the bat misses the ball, it swings through the air with no appreciable change in motion because it has not exerted force on any body of substantial mass and therefore nothing has changed its state of motion. Strictly speaking, the bat is striking air molecules during its motion, thereby changing their state of motion and slightly reducing its speed. There is another force acting to change the direction of the bat's motion that causes it to move in an arc. If it were not for the force of the batter's hands the bat's motion would continue in a straight line. That is, if the batter lets go of the bat, it will fly off in a straight line; we will consider this phenomenon later.

The first and third laws set the stage and the conditions for the second law. To appreciate Newton's approach, let us briefly look at it in historical perspective. It is common knowledge that Galileo (1564–1642) was tried for disobeying a directive from Catholic Church authorities not to state or publish his views of motion of the solar system. The next great natural philosopher to take a keen interest in the laws of mechanics was the Frenchman René Descartes (or in Latin, Renatus Cartesius) (1596–1650). Our cartesian coordinate system bears his Latin name. He decided from his studies that the most important property of a body in a mechanical system was what he called its *momentum*, the mass times the velocity, *mv*. We will deal with this in Chapter 6. Being a careful observer, he could not help but notice what had happened to Galileo, so he declined to publish his thoughts on mechanical

René Descartes (1596-1650).

systems. These were embodied in his manuscript *Le Monde*, which was first published in Amsterdam in 1662, 12 yr after his death. The Reformation had by then reached Holland, so the fear of recrimination by ecclesiastical authorities was of little concern.

Meanwhile, back in England, the Great Plague was raging and everyone who had a relative in the country escaped from the city of London. Among these was Newton, who, according to legend, was letting apples fall on his head. A copy of *Le Monde* was given to him, and he was able to make the creative step shortly thereafter, although his approach was somewhat indirect. The concept of momentum, $m\mathbf{v}$, involves constant or zero velocity, which is embodied in his first law. If a force is applied against a body, the body resists with what he called an inertial force, namely, resistance to having its state of motion (momentum) changed. This is the principle embodied in the third law. If, however, a force is applied to a body for a given length of time, Δt, the momentum will be changed by $\Delta m\mathbf{v}$. He called the product of force and time an *impulse*, and he wrote the basic principle of the second law that the application of an impulse to a body caused a change in its momentum or

$$\mathbf{F}\Delta t = \Delta m\mathbf{v} \tag{4.1}$$

He recognized, however, that direction was equally important. That is, if an impulse was applied in the x direction, the momentum of the body would be changed in only the x direction. Thus Newton introduced the requirement of vectors in calculations.

If we divide Eq. 4.1 by Δt and consider the mass of the body to be constant, we may write

$$\mathbf{F} = m\frac{\Delta \mathbf{v}}{\Delta t} \tag{4.2}$$

where we must recognize that in most physically realizable situations $\mathbf{F}$ is not a constant but rather an average force; for example, the force of a bat against a baseball. In Chapter 3 we defined a Δ as a measurable quantity. Newton realized that he wanted to have a form of Eq. 4.2 for very small, or instantaneous, values of time. Because only tentative beginnings of calculus existed at that time, he proceeded to improve the methods. (G. Leibniz, a contemporary German mathematician, also refined the calculus, independently.) Combining the result of Eq. 3.5, $\mathbf{a} = d\mathbf{v}/dt$, with Eq. 4.2, we may write

$$\mathbf{F} = m\mathbf{a} \tag{4.3}$$

$\mathbf{F} = m\mathbf{a}$

To use Newton's great second law properly, we must include formally the two additional concepts: (1) This is a vector equation, and (2) the force, which is now instantaneous so the average is not required, is actually the net, or unbalanced, force in a given direction. We write this as the algebraic sum in each direction.

The forms of Newton's law that we will use are therefore

$$\Sigma F_x = ma_x, \quad \Sigma F_y = ma_y, \quad \Sigma F_z = ma_z \tag{4.4}$$

$\Sigma F_x = ma_x$
$\Sigma F_y = ma_y$
$\Sigma F_z = ma_z$

The concept of summation of forces can be understood from elementary examples. Consider a tug-of-war with equal numbers of people of equal strength pulling on a rope in opposite directions. If we consider a point in the center of the rope as the newtonian body, we conclude that the sum of forces acting on that body is zero and we will observe no acceleration. If a small child joins one of the sides and pulls, then the sum of forces is no longer zero, but instead there is a net force in the direction that the child pulls with magnitude equal to the force the child exerts. Hence, the rope will be accelerated in his direction.

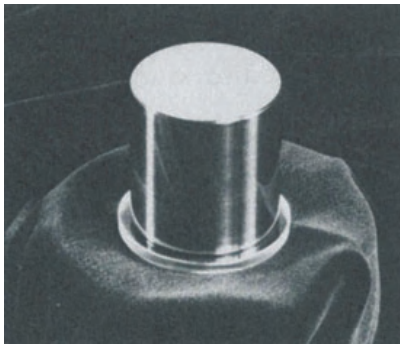
It is convenient at this point to introduce the word *tension*, which is used to convey the transmission of a force through a rope. In this example, none of the contestant's hands are on the newtonian body at the center of the rope; yet, if the center is not being accelerated in either direction, the sum of forces at that point must be zero. If we insert a spring scale in the rope on either side of the center, we note that the same force is present, an example of *action* and *reaction*. Why does the spring scale not read zero, since we have just said that the sum of the force is zero? The answer lies with the third law of action and reaction. If you pull on one end of a spring, the other attached end is being pulled equally and in the opposite direction. Thus, although the sum of the forces is zero and the scale does not accelerate, the spring is nevertheless stretched, which causes the scale to read the value of whatever force is being applied to either end. We may perform the same measurement at any other point along the rope and the scale will read the same. The measured force in the rope at any point is called the *tension* in the rope. We could lower a curtain at the middle of the rope and tie one end of the rope to a wall and send that team home. The other team would not be aware of it, and a measurement in the rope would indicate the same tension. Suppose we went behind the curtain and cut the rope: How would Newton's law apply?

4.3 MASS

Let us now reconsider our dilemma of defining mass and force. At this point we know from Chapter 3 how to measure acceleration, and we have stated that everyone has agreed to accept a certain block of material as having a mass of 1 kg. We do not yet know how to measure force, but we can devise a system to reproduce a given force, such as a pull on a rope with a spring scale to measure the same tension. If we exert this force on the standard kilogram, m_0, with no other forces such as friction or gravity to interfere, we can measure an acceleration $\mathbf{a}_0$. If we apply this same force to a different mass m_1, we measure a different acceleration $\mathbf{a}_1$. From Eq. 4.3 we may write for each experiment

$$\mathbf{F} = m_0 \mathbf{a}_0$$

$$\mathbf{F} = m_1 \mathbf{a}_1$$



The standard kilogram, a platinum-iridium cylinder kept at the International Bureau of Weights and Measures in Sèvre, France.

and, equating the two because the forces are equal, we have

$$\frac{m_1}{m_0} = \frac{\mathbf{a}_0}{\mathbf{a}_1} \tag{4.5}$$

a relation independent of the value of the force. We thus have a method of measuring the mass of any other body in relation to a standard mass.

The unit of force can be defined in terms of mass, length, and time using Eq. 4.3.

$$\mathbf{F} = m\mathbf{a}$$

$$= [M]\left[\frac{L}{T^2}\right]$$

where brackets contain the dimensionality of the quantities involved and M, L, and T stand for dimensions of mass, length, and time, respectively. Because the equation must balance dimensionally, force has units of mass $\times$ length/time^2 or kilogram-meter per second^2

$$F\left(\frac{ML}{T^2}\right)$$

In the SI system of units, this combination of units is called newton (N) for simplicity. *A force of 1* N *is that force which causes a mass of 1 kg to be accelerated at a rate of 1 m/sec*^{2} (or 2 kg accelerated at 0.5 m/sec^2, and so on).

4.4 WEIGHT

A simple way to determine mass is to weigh it on a balance scale. In this method, a balance consists of a rod pivoted in the center so that the weighing pan on each side is equidistant from the center. (In Chapter 8 we will see how a balance scale may be constructed with arms of unequal length.) The unknown mass is placed on one side, and multiples or fractions of a standard kilogram are placed on the other side until a balance is achieved. In this way the magnitude of the unknown mass can be determined because both the unknown and known masses are being acted on by the same force, that of gravity.

The force of gravity can be expressed in terms of Eq. 4.3 by measuring g, the acceleration resulting from gravity. This can be done by noting that the rate of free fall of all objects in a vacuum (to eliminate air resistance) at a given point on earth is the same. The downward acceleration at sea level is approximately the same at all locations, or $g = 9.8$ m/sec^2. So the force on an object of mass m resulting from gravity is, from Newton's second law

$$F = mg$$

and in the English language we call this force the *weight* of an object or

$$\text{Weight} = mg \tag{4.6}$$

Weight = mg

Thus, for 1 kg

$$\text{Weight} = 1\ (\text{kg})g(\text{m/sec}^2) = g(\text{kg m/sec}^2)$$
$$= g \text{ newtons}$$

and 1 kg weighs 9.8 N. On the surface of the moon the acceleration of gravity is about one-sixth that of earth, so the weight of 1 kg will be one-sixth its weight on earth. In outer space at great distance from all other objects, the gravitational force will be near zero, and the kilogram will have almost zero weight. But its mass is unchanged. It takes the same force to produce a given acceleration in space as it does on the moon or on the earth.

In the English system we use units of *pounds* to express the weight of an object. Therefore, the pound is a force. Acceleration is ft/sec^2 and mass has units of

$$\frac{\text{Weight(pound)}}{g(\text{ft/sec}^2)} = m\left(\frac{\text{pound-sec}^2}{\text{ft}}\right)$$

The unit of mass in the English system is called the *slug* (from sluggish). At the surface of the earth the acceleration of gravity is 32.2 ft/sec^2 and 1 lb weight has a mass of 1/32.2 slugs.

4.5 APPLICATIONS OF NEWTON'S LAWS

4.5a Zero Acceleration

It is seen from Eq. 4.4 that when $a = 0$ in a given direction the sum of forces, or net force, in that direction is zero. This fact can be used to gain information about the forces acting on an object when the object is not accelerated. When several forces act on an object but their effects cancel so that the object is not accelerated, the object is said to be in *equilibrium*.

Example 4-1

A block rests on a table. What are the forces acting on the block? See Fig. 4-1.

Solution Take the upward direction ($+y$ direction) as positive. We know that there is a force downward equal to the weight of the block. But because the block is not accelerating there must be an equal force upward, which we will call N for *normal* force, so that the sum of forces in the y direction is zero. Note here that the word *normal* is used in the mathematical sense of the direction perpendicular to a plane. We write Newton's law as

$$\Sigma F_y = 0$$
$$-mg + N = 0$$
$$N = mg$$

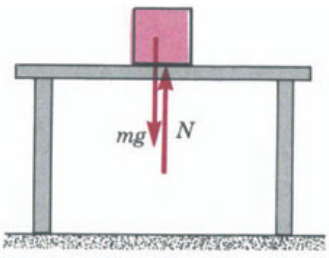


FIGURE 4-1
Example 4-1.

So the table exerts a force equal and opposite to the weight of the block; the table exerts a force on the floor equal to the sum of its weight and that of the block, the floor exerts an equal and opposite force on the table legs, and so forth.

FIGURE 4-2
Example 4-2.

Example 4-2

A child pulls a toy boat through the water at constant velocity by a string parallel to the surface of the water on which he exerts a force of 1 N. What is the force of resistance of the water to the motion of the boat? See Fig. 4-2.

Solution Let F be the force parallel to the water of the string and f be the force of resistance of the water. Let us take the direction of F as the positive x direction. Because constant velocity means zero acceleration,

$$\Sigma F_x = 0$$

$$F - f = 0$$

$$f = F = 1 \text{ N}$$

Example 4-3

Two ropes attached to a ceiling at the angles shown in Fig. 4-3 support a block of weight 50 N. What are the tensions T_1 and T_2 in the ropes?

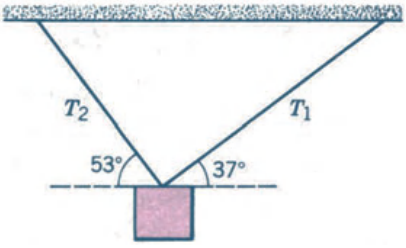


FIGURE 4-3
Example 4-3.

Solution Note here that the ropes exert forces both on the block and on the ceiling. The newtonian body of our concern is one through which all of the forces pass, namely the block. We therefore use the tensions acting on the block. We first draw a vector diagram of the forces (tensions) as in Fig. 4-4. If we examine the newtonian body, we see that it is not accelerating in either the x or y directions. We may therefore write

$$\Sigma F_x = 0, \quad \Sigma F_y = 0$$

By the component method of Chapter 2, we find the x and y components of the forces and substitute them into the equations.

$$\Sigma F_x = 0$$

$$T_1 \cos 37° - T_2 \cos 53° = 0$$

$$0.8\, T_1 - 0.6\, T_2 = 0$$

$$\Sigma F_y = 0$$

$$T_1 \sin 37° + T_2 \sin 53° - 50 \text{ N} = 0$$

$$0.6\, T_1 + 0.8\, T_2 - 50 \text{ N} = 0$$

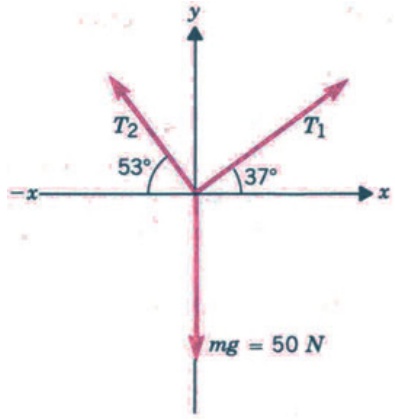


FIGURE 4-4
Example 4-3.

We solve these simultaneously by substituting either T_1 or T_2 from one equation into the other. For example, from the $\Sigma F_x = 0$ equation we get

$$T_1 = \frac{0.6T_2}{0.8} = \frac{3}{4}T_2$$

Substituting into the second equation

$$0.6\left(\frac{3}{4}T_2\right) + 0.8\,T_2 - 50\text{ N} = 0$$

$$1.25\,T_2 = 50\text{ N}$$

$$T_2 = 40\text{ N} \quad \text{and} \quad T_1 = \frac{3}{4}T_2 = 30\text{ N}$$

4.5b Constant Acceleration

In a constant acceleration situation we must examine the motion of the newtonian body in all of the cartesian directions. It may be accelerating in some directions but not in others. The direction or directions in which it is not accelerating may give additional information about the forces acting on the body.

Example 4-4

A child pulls on a string attached to a 1-kg toy boat at an angle of 45° with a constant force of 2 N (Fig. 4-5). The boat goes from rest to a velocity of 0.2 m/sec in 0.5 sec. Assuming constant acceleration, what is the force of resistance of the water?

Solution

$$\Sigma F_x = ma_x$$

From Fig. 4-5, this becomes

$$2\text{ N}\cos 45° - f = ma_x$$

$$(2\text{ N})(0.71) - f = (1\text{ kg})a_x$$

where we used the component of force in the direction of motion and f is the force of resistance of the water. We find a_x by the method of Chapter 3.

$$v_{0x} = 0, \quad v_{fx} = 0.2\text{ m/sec}, \quad t_f = 0.5\text{ sec}, \quad a_x = ?$$

$$v_{fx} = v_{0x} + a_x t_f$$

$$a_x = \frac{v_{fx} - v_{0x}}{t_f} = \frac{0.2\text{ m/sec} - 0\text{ m/sec}}{0.5\text{ sec}} = 0.40\text{ m/sec}^2$$

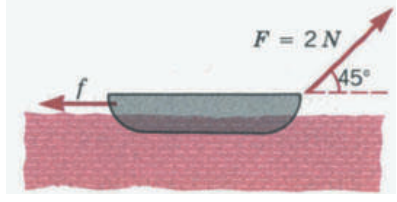


FIGURE 4-5
Example 4-4.

Substituting this result for a_x gives

$$1.42 \text{ N} - f = 0.40 \text{ N}$$

$$f = 1.02 \text{ N}$$

Example 4-5

A block of mass 8 kg is released from rest on a frictionless incline that is at an angle of 37° with the horizontal (Fig. 4-6*a*). What is its acceleration down the incline?

Solution In this situation it is convenient to tilt our graph paper so that the x axis is along the incline, for that is the direction in which the acceleration is to be determined. The y axis will be perpendicular to the incline. The vector force diagram is shown in Fig. 4-6*b*. The only forces exerted on the block are mg downward and the normal force N on the block exerted by the plane that, as we indicated in Example 4-1, is perpendicular to the surface. The component of mg along the x axis, F_x, is determined by dropping a perpendicular from the end of the mg-force vector to the x axis. The angle between mg and this perpendicular is $\theta = 37°$ by the geometric rule that two angles are equal if their sides are mutually perpendicular: A is perpendicular to D, and B is perpendicular to C. Therefore,

$$\sin 37° = \frac{F_x}{mg}$$

$$F_x = mg \sin 37°$$

From Newton's second law, Eq. 4.4,

$$F_x = ma_x$$

$$a_x = \frac{F_x}{m}$$

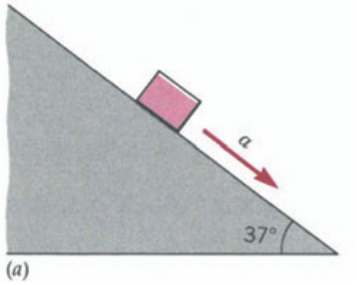
$$= \frac{mg \sin 37°}{m}$$

$$= g \sin 37° = 9.8 \text{ m/sec}^2 \times 0.6$$

$$= 5.9 \text{ m/sec}^2$$

Two important points can be seen in this simple problem:

1. Because the acceleration is independent of the mass, all masses starting from rest at the same height on the same plane will have the same acceleration and, therefore, reach the bottom at the same time.
2. The acceleration is less than the acceleration of gravity because only a component of the force of gravity on the body is directed down the plane.



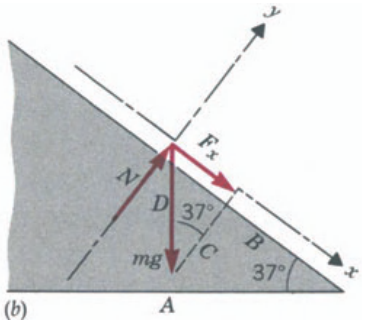


FIGURE 4-6
Example 4-5. (*a*) Diagram of the problem. (*b*) Force diagram.

Example 4-6

Masses of 2 kg and 4 kg connected by a cord are suspended over a frictionless pulley (Fig. 4-7*a*). What is their acceleration when released?

Solution Before solving this problem, we note three important facts. First, because the pulley is frictionless, the tension in the rope is the same on both sides. If it were not, the cord would slide over the pulley until the tensions were the same. Second, the tensions are not the same as in a static situation; that is, we *cannot* equate $T = mg$ because $\Sigma F_y \neq 0$. Third, there are two newtonian bodies and we must write an equation for each, but note that while m_1 moves upward with a positive acceleration, m_2 moves with an acceleration having the same magnitude but directed downward. The force diagrams for the two bodies are given in Fig. 4-7*b*.

For body m_1 we write

$$\Sigma F_y = m_1 a$$

$$T - m_1 g = m_1 a$$

$$T = m_1(g + a)$$

For body m_2

$$\Sigma F_y = m_2 a$$

and, noting that because upward motion was chosen as positive for body 1, the downward acceleration of body 2 must be negative, we write

$$T - m_2 g = m_2(-a)$$

$$T = m_2(g - a)$$

Substituting the T from the body 1 equation, into the equation for body 2, as the tensions are the same, we obtain

$$m_1(g + a) = m_2(g - a)$$

Rearranging terms,

$$a\,(m_1 + m_2) = g\,(m_2 - m_1)$$

$$a = g\,\frac{m_2 - m_1}{m_1 + m_2}$$

$$= 9.8 \text{ m/sec}^2 \times \frac{4 \text{ kg} - 2 \text{ kg}}{2 \text{ kg} + 4 \text{ kg}} = 3.3 \text{ m/sec}^2 \quad \text{for body 1}$$

and

$$a = -3.3 \text{ m/sec}^2 \quad \text{for body 2}$$

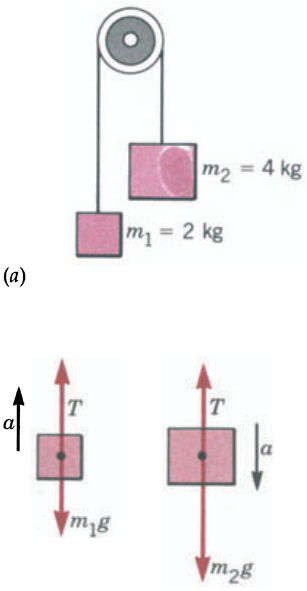


FIGURE 4-7
Example 4-6. (*a*) Diagram of the problem. (*b*) Force diagram for each of the masses.

4.6 FRICTION

We have to exert a steady force to drag an object at constant velocity across the floor. Because the velocity is constant, the acceleration is zero and the sum of forces on the object is correspondingly zero. This means that there is a force equal and opposite to the force that we exert that resists the motion of the object. However, our tug-of-war example does not apply here, for if we stop pulling the object the resistive force does not start to pull it in the opposite direction. Nearly all surfaces have a certain amount of roughness, visible under a microscope, and it is the breakage of these rough protrusions or the rising over them that causes the resistance to motion. This resistive force is called the *force of friction*. The earlier example, 4-2, of the boy with the boat is another type of friction, that of water resisting the motion of an object moving through it. But why consider friction at all in a book about semiconductors and their circuits? Because we commonly experience friction of the types we are discussing here and thus they are easier to comprehend. We will later consider the motion of electrons through a solid under the influence of electric forces. The motion of the electrons is impeded by their banging into the atoms in the solid and losing energy in the process. This is another type of friction.

Returning now to the behavior of an object with an opposing frictional force, we must leave First Principles temporarily and rely on experimental data. There are two types of friction, *static* and *kinetic*. The starting friction is called *static*. The friction of motion is called *kinetic*. We observe from experience that it is harder to start an object moving across a floor than it is to maintain its motion; static friction is larger than kinetic friction. We will only consider kinetic friction. If we wish to measure the force of kinetic friction, we have only to measure the force required to keep an object in motion at constant velocity on a level surface. If we add a weight equal to that of the object on top of it, we find it takes twice the force to keep it moving at constant velocity; with the object weighing three times as much, then three times the force is required. We would correctly conclude that the force of friction is proportional to the weight of the object. But, as can be seen from Fig. 4-8, it is equivalent to say that the force of friction is proportional to the normal force because $mg = N$. Either way we say that the force of friction is proportional to the force pushing the surfaces together. Therefore

$$f \propto N \tag{4.7}$$

Now we know that it is easier to pull or push an object across ice than across a floor. Therefore, we may transform the proportionality of Eq. 4.7 to an equality by introducing a constant that characterizes the surface. We customarily use the Greek letter μ(mu) for this, and μ is called the *coefficient of friction*. Thus

$$f = \mu N \tag{4.8}$$

$f = \mu N$

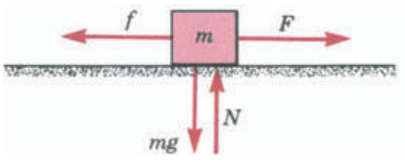


FIGURE 4-8
Forces on a mass that give rise to a force of friction f.

Example 4-7

A force of 10 N is required to keep a box of mass 20 kg moving at a constant velocity across a level floor (Fig. 4-9). What is the coefficient of friction?

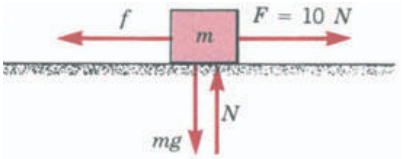


FIGURE 4-9
Example 4-7.

Solution Because the velocity is constant, $a_x = 0$ and $a_y = 0$, and

$$\Sigma F_x = 0$$

$$F - f = 0$$

$$f = 10 \text{ N}$$

and

$$\Sigma F_y = 0$$

$$N - mg = 0$$

$$N = mg$$

But

$$f = \mu N$$

$$f = \mu mg$$

or

$$\mu = \frac{f}{mg}$$

$$= \frac{10 \text{ N}}{20 \text{ kg} \times 9.8 \text{ m/sec}^2}$$

$$\mu = 0.05$$

Suppose that the surface is not level but is inclined by an angle θ with respect to the horizontal, as in Fig. 4-10. We see that the component of the weight mg along the axis perpendicular to the plane is $-mg \cos \theta$ and, by Newton's second law, because $a_y = 0$, the normal force N exerted by the plane on the block is $N = mg \cos \theta$. Thus, Eq. 4.8 can be written in a more general form as

$$f = \mu mg \cos \theta \qquad (4.9)$$

$f = \mu mg \cos \theta$

when $\theta = 0$, $f = \mu mg$ and is at a maximum. When $\theta = 90°$, the incline is standing vertically and there is no force pushing the surfaces together and $f = 0$.

Example 4-8

A block is placed on a plane inclined to the horizontal at 37°. The coefficient of friction between the plane and the block is $\mu = 0.4$. When the block is released, what is its acceleration down the plane?

Solution The forces along the plane are the force of friction f upward and the component of the force of gravity F_D downward (see Fig. 4-10). Choose the downward direction as positive and write Newton's second law.

$$\Sigma F_{\text{plane}} = ma_{\text{plane}}$$

$$F_D - f = ma_{\text{plane}}$$

We have seen in Example 4-5 that

$$F_D = mg \sin \theta$$

and Eq. 4.9 gives the expression for f

$$mg \sin \theta - \mu mg \cos \theta = ma_{\text{plane}}$$

Solving for the acceleration, we obtain

$$a_{\text{plane}} = \frac{mg \sin \theta - \mu mg \cos \theta}{m}$$

$$= g \sin \theta - \mu g \cos \theta$$

Substituting the known quantities

$$a_{\text{plane}} = 9.8 \text{ m/sec}^2 \times 0.6 - 0.4 \times 9.8 \text{ m/sec}^2 \times 0.8$$

$$= 2.74 \text{ m/sec}^2$$

We see that the mass cancels; that is, all blocks with the same coefficient of friction will have the same acceleration down the plane.

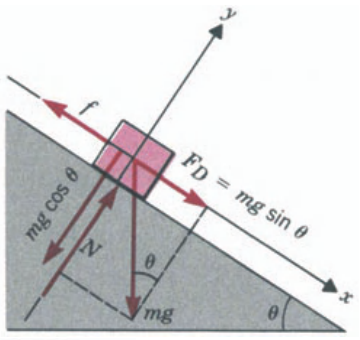


FIGURE 4-10
A block in an inclined plane. The normal force is equal to the component of the weight perpendicular to the plane $mg \cos \theta$ and the force of friction is equal to $-\mu mg \cos \theta$.

PROBLEMS

4.1 A 50-N weight is suspended by a rope from the ceiling. A horizontal force pulls it sideways, causing the rope to make an angle with the ceiling of 53°. When the weight is in equilibrium, what is the force?

4.2 A 40-N weight is suspended by a rope from the ceiling. Another rope pulls horizontally on it sideways so that the suspending rope makes an angle of 60° with the ceiling. What are the tensions in the ropes?

4.3 A 50-N weight is suspended by a rope from the ceiling. A horizontal force of 40 N pulls on the weight in the x direction. (a) What is the angle that the rope makes with the ceiling? (b) What is the tension on the rope?

4.4 A 100-N weight is suspended by ropes as shown in Fig. 4-11. Find the tension on each rope.

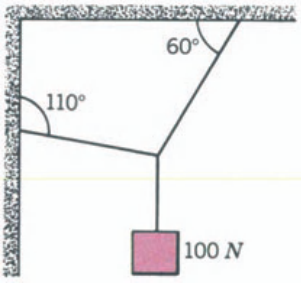


FIGURE 4-11
Problem 4.4.

4.5 A force of 50 N acting at 37° above the horizontal pulls a block along the floor with constant velocity. If the coefficient of friction between the

block and the floor is 0.2, what is the mass of the block?

(*Answer*: 23.4 kg.)

4.6 A 500-kg box is to be lowered down a ramp at constant velocity. The ramp makes an angle of 30° with the ground. The coefficient of friction between the ramp and the box is 0.7. (a) What force applied parallel to the ramp is needed? Must the box be pushed down or held back? (b) Repeat (a) if the coefficient of friction is 0.2.

4.7 A constant horizontal force of 50 N acts on a body that is resting on a smooth, frictionless horizontal plane. The body is observed to go from rest to $v = 5$ m/sec in 10 sec. What is the mass of the body?

(*Answer*: 100 kg.)

4.8 A body of mass 5 kg rests on a horizontal frictionless plane. A force of 10 N is applied at an angle of 37° above the plane for 5 sec. How far has the body moved in that time?

4.9 An electron of mass 9×10^{-31} kg leaves the heated filament of a vacuum tube with $v_0 = 0$ m/sec and travels in a straight line toward a plate 1 cm away. It arrives there with a velocity of 7×10^6 m/sec. Find the magnitude of the acceleration and the accelerating force.

4.10 A 5-kg mass is attached to the end of a string with a breaking strength of 100 N. What is the maximum acceleration that the mass can be given by pulling the string in the upward direction?

(*Answer*: 10.2 m/sec^2.)

4.11 Three blocks of mass—$m_1 = 3$ kg, $m_2 = 4$ kg, $m_3 = 6$ kg—resting on a frictionless table and connected by strings with tensions T_1 and T_2 are being pulled to the right by a force of 6 N (Fig. 4-12). (a) What is the acceleration of the blocks? (b) What are the tensions in the strings?

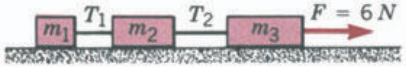


FIGURE 4-12
Problem 4.11.

4.12 What horizontal force is required to drag a 5-kg block along a horizontal surface, with a coefficient of friction of 0.5, at a constant acceleration of 1 m/sec^2?

4.13 A block of 8 kg is held on an incline at 37°. The coefficient of friction between the block and the incline is $\mu = 0.1$. When the block is released, what will be its acceleration?

4.14 Consider an 8-kg block on a frictionless plane inclined at 37° to the horizontal, as in Fig. 4-6*a*. Suppose a force of 40 N is applied to the block upward along the plane. (a) What will be the acceleration of the block? (b) If the upward force applied is 60 N, what will be its acceleration? (c) What force is required along the plane to hold the block motionless?

(*Answer*: (a) 0.90 m/sec^2 downward, (b) 1.60 m/sec^2 upward, (c) 47.18 N.)

4.15 (a) What constant force acting parallel to a 37° plane is required to push a 10-kg block up the plane at constant speed if the coefficient of friction is 0.5? (b) What force is required to push it up with an acceleration of 2 m/sec^2?

4.16 Block A rests on a frictionless plane and the connecting cord passes over a frictionless pulley with block B attached to it (Fig. 4-13). What is the acceleration of block A along the plane when it is released?

(*Answer*: 0.67 m/sec^2 down the plane.)

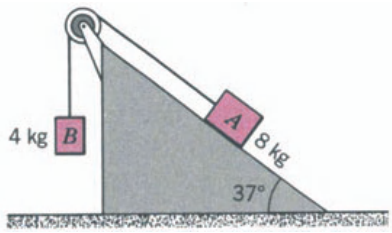


FIGURE 4-13
Problem 4.16.

4.17 An object is hung by a rope to the ceiling of an elevator. When the elevator rises at constant speed, the tension in the rope is 50 N. (a) What is the tension when the elevator is accelerating upward at 3 m/sec^2? (b) What is the acceleration of the elevator if the tension is 30 N?

4.18 An object slides down a 37° incline with constant velocity. After reaching the bottom, it is launched

up the incline with an initial velocity of 5 m/sec. How far up the incline will it move before it stops?

(*Answer*: 1.06 m.)

4.19 In Fig. 4-14, the masses of blocks A, B, and C are 5 kg, 20 kg, and 10 kg, respectively. The blocks are observed to move with constant velocity. What will be the acceleration of blocks A and B when block C is removed?

(*Answer*: 0.65 m/sec^2.)

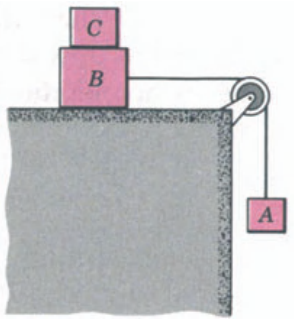


FIGURE 4-14
Problem 4.19

4.20 A 40-N block is connected to a second block by a light rope passing over a frictionless pulley, as in Fig. 4-15. The coefficient of friction between the blocks and the inclines is 0.25. If the 40-N block moves up the plane at constant velocity: (a) What is the weight of the second block? (b) What is the tension in the rope? (c) Suppose that the 40-N block is replaced by a 100-N block and the coefficient of friction remains unchanged, what will be the acceleration of the blocks?

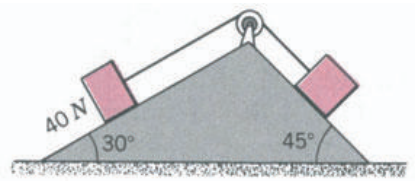


FIGURE 4-15
Problem 4.20.

4.21 A rope of breaking strength 800 N is to be used to drag a box at constant velocity on a horizontal surface. The rope pulls the box at some angle θ above the horizontal. If the coefficient of friction is 0.3, what is the maximum weight of the box that can be moved without the rope breaking?

(*Answer*: 2784 N.)

4.22 A 10-kg ball is hung by a rope from the ceiling of a car. The maximum tension that the rope can withstand is 500 N. (a) What is the maximum horizontal acceleration that the car can reach without the rope breaking? (b) What is the angle between the rope and the vertical for that acceleration?

(*Answer*: (a) 49.03 m/sec^2, (b) 78.7°.)

4.23 A block is held against the front vertical wall of a railroad car, as in Fig. 4-16. The coefficient of friction between the block and the wall is 0.4. When the train begins to accelerate, the block is released and begins to slide down the wall with an acceleration of 9.0 m/sec^2. What is the horizontal acceleration of the train?

(*Answer*: 2 m/sec^2.)

FIGURE 4-16
Problem 4.23.

4.24 In Fig. 4-17, block B of mass 5 kg is being pulled by a force $F = 100$ N. The mass of block A is 2 kg and the coefficient of friction between all surfaces is 0.2. The pulley is frictionless. Find the acceleration of the blocks and the tension in the rope.

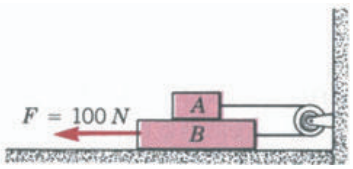


FIGURE 4-17
Problem 4.24.

CHAPTER 5

Work, Energy, and Power

5.1 INTRODUCTION

The terms work, energy, and power are common words in English. In physics we require precise definitions so that these terms can be formulated mathematically. Readers will find that definitions in physics do not always match the usage of the words. For example, although a student may do homework, by the physics definition of work none is being done, although at the biological cellular level, chemical work is being done. We will not consider chemical work in this book. It is important that we consider *mechanical* work, energy, and power, for it is the treatment of these terms from First Principles that will be applied directly to electrical circuits. It is therefore essential that the physics definitions of these terms be learned carefully.

5.2 WORK

In this first treatment of work we will restrict our consideration to that done by a constant force. Our definition of an amount of work ΔW done by a constant force **F** acting on a body is: The product of the distance the body is moved in a given direction by the component of the force in that direction,

$$\Delta W = F_s \, \Delta s \tag{5.1}$$

where F_s represents the component of force in the direction Δs. If F_s is 1 N and Δs is 1 m, then the work ΔW done on the body by the force **F** is 1 newton-meter (N-m). We define a new unit, 1 N-m = 1 joule (pronounced in America as jewel) with symbol J.

Example 5-1

A box is pushed 3 m at constant velocity across a floor by a force **F** of 5 N parallel to the floor. (a) How much work was done on the box by the force **F**, which clearly opposes friction (see Fig. 5-1). (b) How much work is done on the box by the force of friction?

Solution

(a) $W = 5 \text{ N} \times 3 \text{ m} = 15 \text{ J}$

(b) Because $a = 0$

$$\Sigma F_x = 0$$

$$F - f = 0$$

$$f = 5 \text{ N}$$

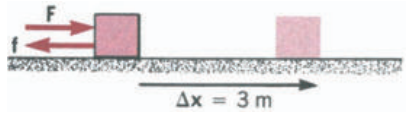


FIGURE 5-1
Example 5-1.

The component of **f** in the direction of the displacement vector $\Delta\mathbf{x}$ is $-f$ (because frictional forces always act against the motion); therefore

$$\Delta W = (-f)\,\Delta x$$
$$= -15\ \text{J}$$

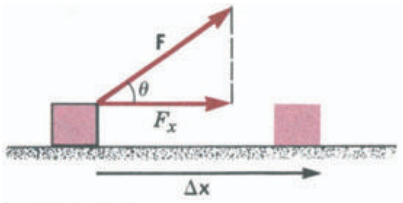


FIGURE 5-2
Force and motion are not in the same direction.

Although work may be positive or negative, it has no direction and is therefore a *scalar* quantity. We will now show that work meets the condition of the vector dot product of Chapter 2, which was defined as a scalar. Suppose the force pulling a box across a floor is not in the direction of motion but is in the direction shown in Fig. 5-2. Following the definition of work, we must take the component of the force in the direction of motion

$$\Delta W = F_x\,\Delta x$$
$$= F\cos\theta\,\Delta x$$
$$\Delta W = F\,\Delta x\cos\theta \tag{5.2}$$

which, by Eq. 2.1, is

$$\Delta W = \mathbf{F}\cdot\Delta\mathbf{x} \tag{5.3}$$

$\Delta W = \mathbf{F}\cdot\Delta\mathbf{x}$

Because Eqs. 5.2 and 5.3 are expressions of the general relation of Eq. 2.1, the F and Δx in Eq. 5.2 are the magnitudes, and we thus have a scalar. Eq. 5.2 has an important conceptual implication. Suppose we carry a weight mg across the room. If it is initially placed in our hands and we carry it slowly, without appreciable acceleration, the only force we exert is in the upward direction.

$$\Sigma F_y = 0$$
$$-mg + F_y = 0$$
$$F_y = mg$$

If we move a distance Δx, from our definition of work

$$\Delta W = F_y\,\Delta x\cos\theta$$

the angle θ between the force direction and the motion direction is 90°. Therefore we do no work.

We may use the definition of Eq. 5.2 to treat formally the work of friction of Example 5-1.

$$\Delta W = f\,\Delta x\cos\theta$$

but $\theta = 180°$ and $\cos 180° = -1$, so we could have simply taken the magnitude of friction times the distance and the $\cos\theta$ would have yielded the correct sign.

5.3 POTENTIAL ENERGY

Consider now that we lift a weight mg at constant velocity a distance y from the floor. Because gravity exerts a downward force mg, we must exert an upward force of equal magnitude over a distance y and therefore have done an amount of work mgy on the weight. If we lower it back to the floor, we still exert an upward force mg, but now the motion is in a direction opposite to the force, or $\theta = 180°$, and the work done by us is $-mgy$ and therefore no net work has been done in the round trip. Suppose we wish to move a weight mg from the floor to a shelf a distance x across the room. There are many paths that we may take, some of which are shown in Fig. 5-3. An examination of these paths shows that each move in the y direction contributes positive or negative work amounts whose sum must be mgy, whereas motion in the x direction contributes no work. We are thus able to draw a very important conclusion. *Work done against the gravitational force is independent of the choice of path between any two fixed endpoints.*

Suppose an object is placed at a height y in a gravitational field, as in Fig. 5-4. If it descended from y, the gravitational force mg on the object would be capable of doing work equal to the force times the displacement mgy. Therefore, because the gravitational force is potentially able to do work on the object, we say that it has a *potential energy* E_p equal to mgy. (Note that the unit of potential energy is the same as that of work, that is, the joule)

$$E_p = mgy \tag{5.4}$$

That is, if we have lifted it by doing work mgy on it, then it has gained a potential energy of mgy, where we use the positive value because the work was put into it. Note that there is a direct relation between the work done on an object in a gravitational field and its gain in potential energy. The measure of y must be considered with care. Suppose an object is lifted above a table to a height y_1. It has $E_p = mgy_1$ with respect to the table. But if the table is at a height y_3 above the floor, the object has $E_p = mg\,(y_1 + y_3)$ with respect to the floor (see Fig. 5-5).

Thus, for potential energy, a reference level must always be specified. If we lift an object two different distances above a table, then we may state the difference in potential energy between the two positions. If we included the distance above the floor for each, then when we take the difference in their E_p, the distance above the floor will cancel,

$$\Delta E_p \text{ (with the table as the reference level)}$$
$$= mgy_2 - mgy_1 = mg(y_2 - y_1)$$
$$\Delta E_p \text{ (with the floor as the reference level)}$$
$$= mg\,(y_2 + y_3) - mg(y_1 + y_3)$$
$$= mgy_2 - mgy_1 = mg(y_2 - y_1)$$

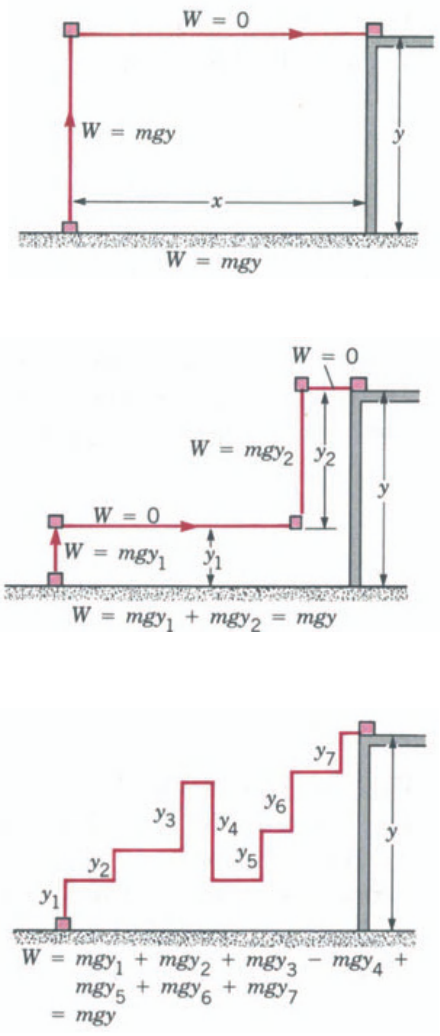


FIGURE 5-3
Examples of different paths used in raising a mass to a height y.

We see in this answer that only the difference in heights needs to be specified to give the relative difference in potential energy. We will use this concept in electricity to specify the relative difference in potential energy of a charged particle in two different positions in an electric field. To apply the concept of potential energy in later chapters we will need to expand our consideration of work to that done by a variable force.

FIGURE 5-4

5.4 WORK DONE BY A VARIABLE FORCE

We have seen that when the force acting through a distance Δx is constant, then the work done may be written as

$$\Delta W = \mathbf{F} \cdot \Delta \mathbf{x} = F_x \, \Delta x \tag{5.3}$$

Suppose that at each small displacement of the motion the force has a different value. Then we would write Eq. 5.3 as

$$W = F_{x1} \, \Delta x_1 + F_{x2} \, \Delta x_2 + F_{x3} \, \Delta x_3 + \cdots F_{xN} \, \Delta x_N$$

or

$$W = \sum_{i=1}^{N} F_{xi} \, \Delta x_i \tag{5.5}$$

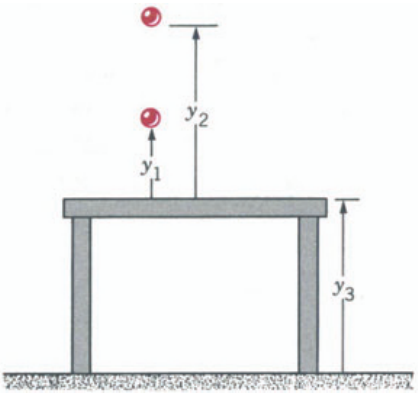


FIGURE 5-5

A sketch of this summation is shown in Fig. 5-6 for the work done in moving a body from $x = a$ to $x = b$. We see that each vertical segment of Δx_i width has associated with it an average F_{xi}. The total work is the sum of the areas of these segments. If we make the width of each segment very small so that $\Delta x \to 0$, the number of segments required to obtain the area under the curve approaches infinity and the sum of these infinitesimal areas becomes the precise area under the curve of Fig. 5-6, or the total work. This is the definition of an integral. That is, Eq. 5.5 as $\Delta x \to 0$ becomes

$$W = \int_a^b F_x \, dx \tag{5.6}$$

Furthermore, by the definition of an integral, and with reference to Fig. 5-6, we see that work is the area under the F_x versus the x curve. In the derivation of Eq. 5.6 we have assumed that **F**, although not constant in magnitude, is always in the direction of the displacements Δx's. In the more general case where **F** and the general displacement $\Delta \mathbf{s}$'s are not in the same direction, the expression for the work becomes

$$dW = \mathbf{F} \cdot d\mathbf{s} \tag{5.7}$$

or the integral form

$$W = \int_a^b \mathbf{F} \cdot d\mathbf{s} \tag{5.7'}$$

$$W = \int_a^b \mathbf{F} \cdot d\mathbf{s}$$

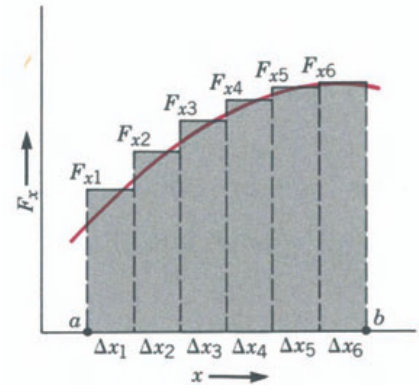


FIGURE 5-6
Area under the curve obtained by summation of the area of rectangles.

where the dot product takes care of changes in orientation between the vectors **F** and $d\mathbf{s}$.

5.5 KINETIC ENERGY

The word "kinetic" is used frequently in all branches of science and is from the Greek word for motion. If work is done on a body that changes its state of motion, namely, its velocity, we say that the work has caused the body to gain or lose *kinetic energy*, E_k. We will show two derivations of kinetic energy, the first for the restricted condition of a constant force and the second for a variable force. In both cases we will consider motion in the x direction alone.

If the force is constant and the initial position is $x = 0$, we may write the definition of work as

$$W = F_x x$$

which from Newton's second law (Eq. 4.4) can be written as

$$W = ma_x x$$

or, if we take a only in the x direction, we may drop the subscript and write

$$W = max \tag{5.8}$$

Because the force is constant and we are considering a single body of constant mass, the acceleration is constant and we may use Eq. 3.11

$$v^2 - v_0^2 = 2\,ax \tag{3.11}$$

where v_0 is the velocity at $x = 0$ and v is the velocity at x. Substituting this equation into Eq. 5.8 we obtain

$$W = m\left\{\frac{v^2 - v_0^2}{2}\right\}$$

$$W = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2 \tag{5.9}$$

Thus the work done on a body that changes its velocity actually changes the quantity $\frac{1}{2}mv^2$, which is called the *kinetic energy* E_k.

$$E_k = \frac{1}{2}mv^2 \tag{5.10}$$

$E_k = \frac{1}{2}mv^2$

Because change in kinetic energy is equal to work and work is a scalar quantity, kinetic energy is also a scalar quantity and the unit of kinetic energy is the same as that of work, that is, the joule.

We obtain the same result if the applied force is not constant but is variable. Let us consider the x axis motion again so that

$$W = \int_{x_0}^{x} F\,dx$$

In this case F must remain inside the integral, but we can substitute Newton's second law for it, $F = ma$

$$W = m\int_{x_0}^{x} a\,dx \tag{5.11}$$

And we may substitute for a from Eq. 3.13

$$a = v\frac{dv}{dx} \tag{3.13}$$

Thus Eq. 5.11 becomes

$$W = m\int_{x_0}^{x} v\frac{dv}{dx}\,dx$$

$$W = m\int_{v_0}^{v} v\,dv$$

where we have changed the limits of integration because the velocity is v_0 at x_0 and v at x. This integrates to

$$W = m\frac{v^2}{2}\Bigg]_{v_0}^{v}$$

$$W = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2 \tag{5.9}$$

$$W = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2$$

which is the same result as found before for a constant force. Note that in both derivations we have used $F = ma$ where F is the net, or resultant, force. If $a = 0$, then $F = 0$ and there is no net, or resultant, force and hence there can be no change in the kinetic energy. Eq. 5.9 is known as the *work-energy theorem*, which states that *the work done by the resultant force acting on a particle is equal to the change in kinetic energy of the particle.*

5.6 ENERGY CONSERVATION

We define a *mechanically conservative* system as one in which *no* energy enters or leaves the system (as heat, radiation, or such). Therefore, the system's initial energy is unchanged, which is the same as saying it is conserved. This fact simplifies the solution of many types of problems because we do not

have to calculate the acceleration. For conservative systems we know that the total energy in state 1 is the same as that in state 2. Because the total energy is the sum of the potential and kinetic energies, we write

$$(E_k + E_p)_{\text{initial}} = (E_k + E_p)_{\text{final}} \tag{5.12}$$

We can verify explicitly the conservation of the total energy with the simple example shown in Fig. 5-7.

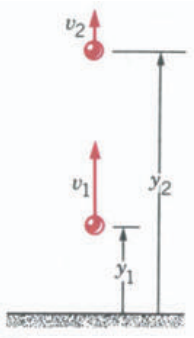


FIGURE 5-7

Let us launch an object of mass m from a point y_1 above the floor with an initial velocity v_1. Owing to the gravitational force that acts on the object, its velocity decreases as it rises. Sometime later, the velocity of the object will be v_2 and its position y_2. We can relate this new velocity and position to the initial velocity and position by means of Eq. 3.11.

$$v_2^2 - v_1^2 = 2(-g)(y_2 - y_1)$$

Rearranging terms,

$$v_2^2 + 2gy_2 = v_1^2 + 2gy_1$$

If we divide both sides by 2 and multiply by the mass of the object m, we get

$$\frac{1}{2}mv_2^2 + mgy_2 = \frac{1}{2}mv_1^2 + mgy_1 \tag{5.13}$$

$\frac{1}{2}mv_2^2 + mgy_2 = \frac{1}{2}mv_1^2 + mgy_1$

Thus, although the velocity of the object (and therefore its E_k) decreases as it rises, this decrease in E_k is compensated by an increase in E_p in such a way that the sum of E_k and E_p remains constant and equal to the sum of the initial kinetic and potential energies. That is, total energy is conserved.

A falling tree illustrates the conversion of potential energy into kinetic energy.

Example 5-2

Suppose a ball is dropped from a height $h = 10$ m. What is its velocity just before it strikes the ground?

Solution

$$E_{ki} + E_{pi} = E_{kf} + E_{pf}$$

$$0 + mgh = \frac{1}{2}mv^2 + 0$$

$$v = \pm\sqrt{2gh} = \pm\sqrt{2 \times 9.8\ \text{m/sec}^2 \times 10\ \text{m}} = -14\ \text{m/sec}$$

where the negative sign is chosen because the motion is downward.

The pendulum is another simple example of the conservation of energy. Let us assume an idealized pendulum that swings in a vacuum so that there is no energy lost to air friction and that there is no frictional loss at the pivot

(see Fig. 5-8). If we start the pendulum by pulling it to one side and releasing it with no initial velocity, it has an initial potential energy of mgh_0, which is also the total energy. If there is no energy loss during its subsequent motion, it must always have this amount of total energy. When it is released, it begins to fall and potential energy is lost. But as it falls it picks up speed and thereby gains kinetic energy. At the bottom of its path, $h = 0$ and all its energy is kinetic. As it starts to rise again, the kinetic energy is converted to potential energy. Thus, if h_0 is its initial height, mgh_0 is the energy of the pendulum and the sum of the potential and kinetic energies at all other positions must equal this value. We may write this as

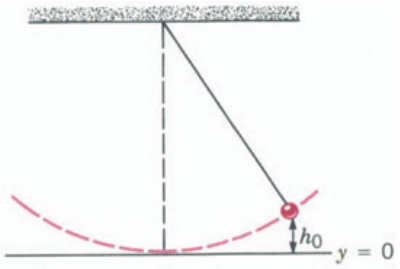


FIGURE 5-8

$$E_{p0} + E_{k0} = E_{p2} + E_{k2}$$

$$mgh_0 + 0 = mgh_2 + \frac{1}{2}mv_2^2$$

The string does no work on the pendulum because of the definition (Eq. 5.7)

$$dW = \mathbf{F} \cdot d\mathbf{s} = F \cos \theta \, ds$$

The angle θ is that between the string direction and ds, the instantaneous direction of motion. This angle is always 90°, so that dW due to the string is always zero.

If the system loses energy or energy is put into the system, it is no longer a mechanically conservative system. Later, when we understand other types of energy and can enlarge the system to include all sources of input and output, we will again develop the concept of conservation of energy. For now, however, let us say that all energy must be accounted for and use the term *accountability of energy*. We become accountants and keep books of assets and liabilities. All initial energy plus any energy put in, E_{in}, is on the left side of the ledger as an asset. All energy converted to another form or escaping from the system, E_{out}, may go on the right side. Thus, Eq. 5.12 is written as

$$E_{ki} + E_{pi} + E_{\text{in}} = E_{kf} + E_{pf} + E_{\text{out}} \tag{5.14}$$

Example 5-3

A skier is on a 37° slope of length $s = 100$ m (Fig. 5-9). The coefficient of friction between his skis and the snow is 0.2. If he starts from rest, what is his velocity at the bottom of the slope?

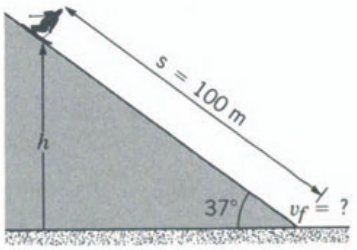


FIGURE 5-9
Example 5-3.

Solution

$$E_{ki} + E_{pi} + E_{\text{in}} = E_{kf} + E_{pf} + E_{\text{out}}$$

No energy is put in, but there is energy lost to work against friction, or $E_{\text{out}} = |W_{\text{friction}}| = |\mathbf{f} \cdot \mathbf{s}|$

As before, let us tilt our coordinate axis so that the slope becomes the x axis and the normal becomes the y axis.

$$\Sigma F_y = 0$$

$$N - mg \cos 37° = 0$$

$$N = mg \cos 37°$$

$$f = \mu N = \mu mg \cos 37°$$

$$E_{\text{out}} = |\mathbf{f} \cdot \mathbf{s}| = \mu mg \cos 37° \, (s)$$

$$0 + mgh + 0 = \frac{1}{2} mv_f^2 + 0 + \mu mg \cos 37° \, (s)$$

The mass cancels, and we solve for v_f

$$\begin{aligned} v_f &= [2(gh - \mu gs \cos 37°)]^{1/2} \\ &= [2(9.8 \text{ m/sec}^2 \times 100 \text{ m} \times \sin 37° \\ &\quad - 0.2 \times 9.8 \text{ m/sec}^2 \times 100 \text{ m} \times \cos 37°)]^{1/2} \\ &= 29.4 \text{ m/sec} \end{aligned}$$

5.7 POWER

Different persons or different machines may take different amounts of time to do the same amount of work. The term used to describe this rate of performance of work is *power*.

$$\text{Power} = \frac{\text{work done}}{\text{time taken}}$$

$$P = \frac{W}{t} \tag{5.15}$$

$P = \frac{W}{t}$

Work is measured in joules, time in seconds, therefore the unit of power is joules per second (J/sec). We introduce a new unit: 1 J/sec = 1 watt. Conversely, work (or energy) is equal to power × time,

$$W(\text{joules}) = P(\text{watts}) \, t(\text{sec})$$

The symbol used for watt is W. A 100-W light bulb uses 100 J of electrical energy each second. Your electric light bill is in kilowatt-hours. A kilowatt-hour is the energy dissipated by a device that uses 10^3 W for a period of 1 h, that is, 1 kwh = 10^3 J/sec × 3600 sec = 3.6×10^6 J.

We may develop another expression for power because work is defined by Eq. 5.3 as the dot product of force **F** and displacement **Δs**, where **Δs** can

be in the x direction. From Eq. 5.15

$$P = \frac{\mathbf{F} \cdot \Delta \mathbf{x}}{\Delta t}$$

or for infinitesimally small displacements $\mathbf{dx}$

$$P = \mathbf{F} \cdot \frac{\mathbf{dx}}{dt}$$

but

$$\frac{\mathbf{dx}}{dt} = \mathbf{v}, \text{ the velocity}$$

and therefore an alternative equation for power is

$$P = \mathbf{F} \cdot \mathbf{v} \qquad (5.16)$$

$P = \mathbf{F} \cdot \mathbf{v}$

There is an engineering unit of power that we might introduce in passing: horsepower. This is defined as

$$1 \text{ hp} = 746 \text{ W}$$

Example 5-4

A tractor can exert a force of 3×10^4 N while moving at a constant speed of 5 m/sec. What is its horsepower?

Solution

$$\begin{aligned} P &= \mathbf{F} \cdot \mathbf{v} \\ &= 3 \times 10^4 \text{ N} \times 5 \text{ m/sec} \\ &= 1.5 \times 10^5 \text{W} \left(\frac{1 \text{ hp}}{746 \text{ W}} \right) = 200 \text{ hp} \end{aligned}$$

PROBLEMS

5.1 A car is pulling a trailer with a force of 900 N on a level road. The car is moving at 70 km/h. How much work is done by the car on the trailer in 15 min?

5.2 A 50-kg box is being pushed on a horizontal surface, at constant velocity, by a 90-N force acting at an angle of 15° below the horizontal. (a) How much work is done by the 90-N force in moving the box a distance of 20 m? (b) How much work is done by friction over the same distance? (c) What is the force of friction?

5.3 A force of 20 N parallel to a 37° plane pulls a 2-kg block 5 m up the plane at a constant speed. (a) How much work has been done by the 20-N force? (b) How much work has been done by friction? (c) How much work has been done by the gravitational

force acting on the block? (d) What can you say about the total work done?

(*Answer*: (a) 100 J, (b) −41 J, (c) −59 J, (d) zero.)

5.4 In Fig. 5-10 the force acting on a body for $x < 10$ m is $F_x = 0.2\,x$ N and the force for $x > 10$ m is constant at $F_x = 2$ N. What is the work done in going from $x = 0$ to $x = 15$ m?

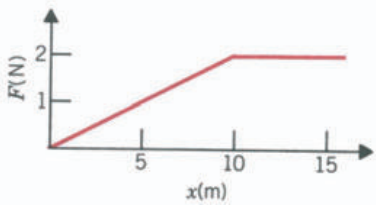


FIGURE 5-10
Problem 5.4.

5.5 A boy pulls a 15-kg sled at constant speed a distance of 30 m along rough level snow that has a coefficient of friction of 0.1. How much work did he do?

5.6 A 40-kg box is to be pushed at constant speed a distance of 5 m up a ramp by a force parallel to the ramp. The coefficient of friction between the box and the ramp is 0.25. The ramp makes an angle of 37° with the horizontal. How much work must be done?

(*Answer*: 1571 J.)

5.7 A car traveling at 30 m/sec suddenly brakes. If the coefficient of friction between the tires and the road is 0.7, what is the minimum stopping distance? Solve by energy methods.

5.8 A bead having an initial speed at point A of 2 m/sec slides down a frictionless wire (see Fig. 5-11). What are its speeds at points B and C?

(*Answer*: $v_B = 4.21$ m/sec, $v_C = 3.44$ m/sec.)

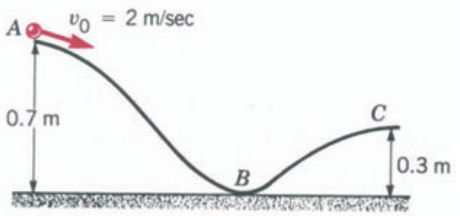


FIGURE 5-11
Problem 5.8.

5.9 A horizontal force of 20 N pulls a 10-kg block on a level frictionless surface a distance of 5 m. (a) How much work is done, and what becomes of this work? (b) Show, using the methods of Chapters 3 and 4, that the change in the kinetic energy is equal to the work done by the force.

5.10 An automobile is moving with a velocity of 90 km/h. From what height would it have to fall to acquire that velocity?

5.11 Using the conservation of energy principle, find the maximum height reached by a projectile launched with a velocity of 80 m/sec at an angle of 37° with the horizontal?

5.12 A pendulum consists of a mass at the end of a string 1.5 m long. The mass is pulled sideways until the string makes an angle of 30° with the vertical; then it is released. What is the speed of the mass as it passes through its lowest point?

(*Answer*: 1.98 m/sec.)

5.13 A light rope passing over a frictionless pulley connects two blocks of mass $m_1 = 3$ kg and $m_2 = 5$ kg (see Fig. 5.12). (a) If the blocks are released from rest in the position shown in Fig. 5-12, what will be the velocity of m_2 as it hits the ground? (b) What is the final height reached by m_1?

(*Answer*: (a) 5.42 m/sec, (b) 7.5 m.)

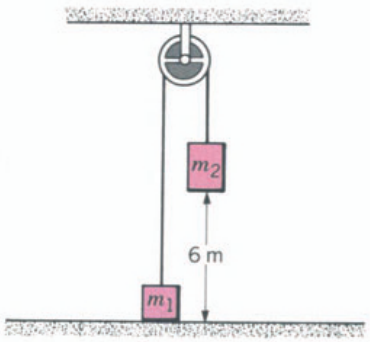


FIGURE 5-12
Problem 5.13.

5.14 A 40-kg box slides 5 m down a ramp inclined at 37° to the horizontal. If there is no friction, what is its speed at the bottom? If the coefficient of friction is 0.2, what is its speed at the bottom?

5.15 A 5-kg block rests at the top of a rough plank inclined at 25° with respect to the horizontal and 4 m long. It is given an initial speed downward of 2 m/sec, and it just reaches the bottom before it stops. What is the coefficient of friction? Solve by energy methods.

(*Answer*: 0.52.)

5.16 A constant force of 60 N parallel to an inclined plane of 30° above the horizontal pushes a 6-kg block 10 m up the incline. The coefficient of friction between the block and the incline is 0.25. (a) What is the velocity of the block at the 10-m point if the block starts from rest? (b) If the force is removed at that point, how much farther up the incline will the block go? (c) At the uppermost point the block starts sliding down. What is its speed when it reaches the bottom? Use the energy methods.

(*Answer*: (a) 7.72 m/sec, (b) 4.24 m, (c) 8.90 m/sec.)

5.17 A rigid rod of negligible weight and length $l = 2$ m has a mass of 5 kg attached to one end. The other end is pivoted about a point 0, as shown in Fig. 5-13. The mass is released from the position shown with some initial speed v_0. As the mass swings around it experiences an average frictional force of 12 N and just reaches the top of the circle, point P. (a) What is the initial speed v_0 of the mass? (b) What is the speed of the mass as it passes through the lowest point P'?

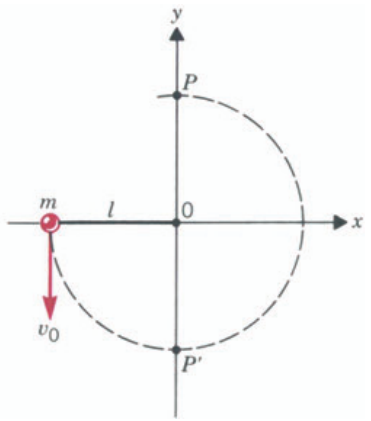


FIGURE 5-13
Problem 5.17.

5.18 A mass $m_1 = 3$ kg resting on a long table is connected by a light string passing over a frictionless pulley to a second mass $m_2 = 5$ kg hanging 2 m above the floor (see Fig. 5-14). The coefficient of friction between the table and m_1 is 0.3. The blocks are released from rest. (a) What is the velocity of m_2 as it hits the floor? (b) What is the total distance traveled by m_1 before it stops?

(*Answer*: (a) 4.48 m/sec, (b) 5.42 m.)

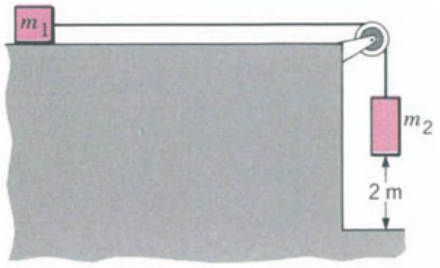


FIGURE 5-14
Problem 5.18.

5.19 An 80-kg man ascends a 4 m high staircase in 12 sec. What is his horsepower?

5.20 An elevator of mass 800 kg is raised 10 m in 5 sec. How much power is required? Express your answer in watts and in horsepower.

5.21 A cable car is operated on a slope 1000 m long making an angle ot 20° with the horizontal. The cable car moves up the slope with a speed of 3 m/sec and carries 20 persons of average weight 600 N. What power is needed?

(*Answer*: 1.23×10^4 W.)

5.22 The piston of a steam engine is driven 120 times per minute. The length of the stroke is 0.5 m. If the engine develops 150 kW of power, what is the average force exerted by the steam on the piston?

5.23 A pump is needed to lift 100 kg of water per minute from a well 30 m deep. The water is ejected with a speed of 5 m/sec. What must be the power output of the pump?

(*Answer*: 510.8 W.)

5.24 A 2500-kg automobile develops 30 kW of power to drive with a constant velocity of 90 km/h on a level road. What power must it develop to drive up a 15° hill with the same velocity?

(*Answer*: 1.88×10^5 W.)

CHAPTER 6

Momentum and Collisions

6.1 INTRODUCTION

Momentum is the product of the mass of a body and its velocity. It is therefore a vector. We first mentioned momentum in Chapter 4 as the seed whose germination led to modern thoughts on the mechanical behavior of newtonian bodies. In our description of this development, however, we did not consider with sufficient care what was meant by a body. In many cases it is an assembly of particles. In this chapter we will first show how such an assembly can be mathematically represented by a point mass, called the *center of mass*. We will then show that the motion of the center of mass is that predicted by Newton's second law for a particle whose mass is the sum of the masses of the individual particles and is acted on by the resultant of the forces acting on the body. Having established these facts, we will turn our attention to the momentum changes of colliding bodies with confidence, knowing that the treatment of bodies is as mathematically sound as if they were very small masses. Collision theory is very important in our later analysis of conduction electrons in solids.

6.2 CENTER OF MASS

If you have a stick, whether uniform or not, you can find a point along its length that we call the "balance point." If you place your finger there, you can support the stick. Clearly, from Newton's law, the sum of the forces in the y direction is zero at that point, with your finger supplying the upward force. The weight of the stick supplies the downward force, and it appears to be located at that point, although we know that every segment of the stick has weight. We call this point the *center of gravity* of the stick.

If, while balancing the stick on your finger, you suddenly exert an impulse on it by moving your finger rapidly upward, the stick will fly upward without rotating. That is, all parts of the stick will move upward uniformly and the stick will therefore retain the configuration that it had on your finger. If, when it is in motion, you catch it at the center of gravity, the entire stick will stop. These experiments show that the center of gravity of the stick behaves as a point mass in Newton's second law, $\mathbf{F} = m\mathbf{a}$, and that the center of gravity may also be considered as the *center of mass*. That is, if you perform these same experiments in distant space where the force of gravity is negligible, you will obtain the same results. Similarly, if you have a piece of cardboard, you can find a point on the surface at which you can place your finger and support it. All the weight of the cardboard can be considered to be located at that point. Although it is a more difficult experiment to perform on a three-dimensional object, it may be shown mathematically that it too has a center of mass. A better understanding of what the center of mass is will be obtained when we introduce the concept of torque in Chapter 8.

Suppose that we have two masses, m_1 and m_2, on a weightless stick as in Fig. 6-1*a*. If we perform a measurement, we will find that the balance point (center of mass), is the point at which the products of the masses and their

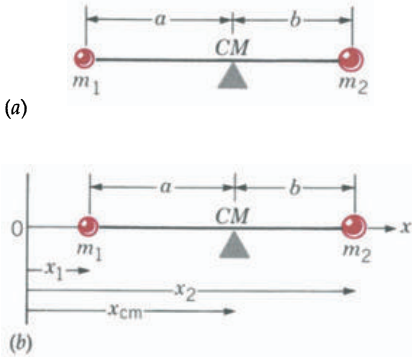


FIGURE 6-1

respective distances from the balance point are equal. From Fig. 6-1*a*, the center of mass is the point such that

$$m_1 a = m_2 b \tag{6.1}$$

Let us put this figure on an x axis, as in Fig. 6-1*b*. We may express Eq. 6.1 in terms of x distances by noting that

$$a = x_{cm} - x_1 \quad \text{and} \quad b = x_2 - x_{cm}$$

Equation 6.1 becomes

$$m_1(x_{cm} - x_1) = m_2(x_2 - x_{cm})$$

Rearranging gives

$$(m_1 + m_2)x_{cm} = m_1 x_1 + m_2 x_2$$

or

$$x_{cm} = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2}$$

This relation holds true regardless of the number of masses placed on the balance, so we may write

$$x_{cm} = \frac{\sum_{i=1}^{n} m_i x_i}{\sum_{i=1}^{n} m_i} \tag{6.2}$$

But $\sum_{i=1}^{n} m_i = M$, where M is the total mass. We can therefore express Eq. 6.2 as

$$x_{cm} = \frac{1}{M} \sum_{i=1}^{n} m_i x_i \tag{6.3}$$

Example 6-1

Find the center of mass of the configuration in Fig. 6-2 when $m_1 = 1$ kg, $m_2 = 2$ kg, and $m_3 = 3$ kg.

Solution

$$x_{cm} = \frac{1}{M} \Sigma m_i x_i$$

If we take the position of m_1 as the origin, we write this equation as

$$x_{cm} = \frac{1}{1 \text{ kg} + 2 \text{ kg} + 3 \text{ kg}} (1 \text{ kg} \times 0 \text{ m} + 2 \text{ kg} \times 0.5 \text{ m} + 3 \text{ kg} \times 1.3 \text{ m})$$

$$= 0.82 \text{ m}$$

FIGURE 6-2
Example 6-1.

If we had chosen any other point as the origin, the position of the center of mass relative to the individual masses would have been the same, although the numerical value of x_{cm} would have been different.

In the general case, when the masses do not lie on one of the axis, we define the center of mass as the point whose cartesian coordinates are

$$x_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i x_i \tag{6.3}$$

$$y_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i y_i \tag{6.4}$$

$$z_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i z_i \tag{6.5}$$

$$x_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i x_i$$

$$y_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i y_i$$

$$z_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i z_i$$

where x_i, y_i, z_i are the coordinates of the ith particle, all measured from the same arbitrary origin. The reason for defining the center of mass in this manner will become obvious in the next section.

Example 6-2

Find the x and y coordinates of the center of mass of the system shown in Fig. 6-3, where $m_1 = 2$ kg, $m_2 = 3$ kg, $m_3 = 4$ kg, and $m_4 = 1$ kg, and the coordinates are (3,4)m, (4,6)m, (5,5)m, and (6,8)m, respectively.

Solution We note that each mass has both an x and y coordinate and therefore each contributes to the x_{cm} and the y_{cm}.

For x_{cm} we write

$$x_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i x_i$$

$$= \frac{1}{10\text{ kg}}(2\text{ kg} \times 3\text{ m} + 3\text{ kg} \times 4\text{ m} + 4\text{ kg} \times 5\text{ m} + 1\text{ kg} \times 6\text{ m})$$

$$= 4.4\text{ m}$$

For y_{cm} we write

$$y_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i y_i$$

$$= \frac{1}{10\text{ kg}}(2\text{ kg} \times 4\text{ m} + 3\text{ kg} \times 6\text{ m} + 4\text{ kg} \times 5\text{ m} + 1\text{ kg} \times 8\text{ m})$$

$$= 5.4\text{ m}$$

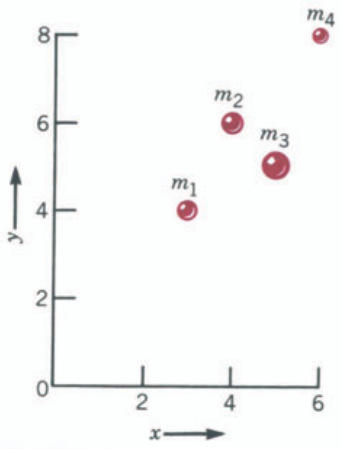


FIGURE 6-3
Example 6-2.

6.3 MOTION OF THE CENTER OF MASS

Let us rewrite the expression for the x coordinate of the center of mass of an array of masses, Eq. 6.3, in the following way.

$$Mx_{cm} = m_1x_1 + m_2x_2 + \cdots m_nx_n \tag{6.6}$$

Differentiating this with respect to time, we obtain

$$M\frac{dx_{cm}}{dt} = m_1\frac{dx_1}{dt} + m_2\frac{dx_2}{dt} + \cdots m_n\frac{dx_n}{dt}$$

or

$$Mv_{xcm} = m_1v_{x1} + m_2v_{x2} + \cdots m_nv_{xn} \tag{6.7}$$

Similar expressions can be readily obtained for Mv_{ycm} and Mv_{zcm}.

Equation 6.7 and the equivalent equations for the y and z motion show that the total momentum of all the particles is equal to the momentum of a single particle whose mass is equal to the sum of the masses of the particles and moves with the velocity of the center of mass. Let us now differentiate Eq. 6.7 with respect to time.

$$M\frac{dv_{xcm}}{dt} = m_1\frac{dv_{x1}}{dt} + m_2\frac{dv_{x2}}{dt} + \cdots m_n\frac{dv_{xn}}{dt}$$

or

$$Ma_{xcm} = m_1a_{x1} + m_2a_{x2} + \cdots m_na_{xn} \tag{6.8}$$

We can apply Newton's second law ($\mathbf{F} = m\mathbf{a}$) to each individual particle; that is, $F_{x1} = m_1a_{x1}$, $F_{x2} = m_2a_{x2}$, ... Substituting this in Eq. 6.8, we have

$$Ma_{xcm} = F_{x1} + F_{x2} + \cdots F_{xn} \tag{6.9}$$

where F_{xi} is the x component of the resultant (i.e., the sum) of the forces acting on the ith particle. A system of masses may be connected by internal forces such as the binding forces of a solid. There may be additional external forces acting on the solid. By Newton's third law of action and reaction, each force exerted internally by a particle on another has an equal and opposite force exerted internally on it. Therefore, the sum of internal forces on the right side of Eq. 6.9 must be zero. We conclude that the sum of the forces in Eq. 6.9 includes only the *external* forces acting on the system of particles. We can rewrite Eq. 6.9 simply as

$$\sum_{i=1}^{n} F_{xi} = Ma_{xcm}$$

The same equation can be derived for the y and the z components of the external forces, and we have the vector equation of Chapter 4

$$\mathbf{F} = M\mathbf{a}_{cm} \tag{6.10}$$

$\mathbf{F} = M\mathbf{a}_{cm}$

where **F** is the resultant of the external forces acting on all the particles. This result shows that the center of mass moves as if it were a point whose mass is equal to the total mass of the system and all the external forces were acting on it. And this is why the point defined by Eqs. 6.3, 6.4, and 6.5 is called the center of mass.

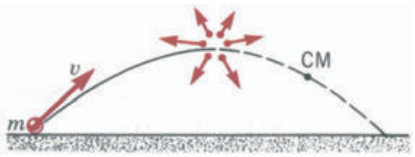


FIGURE 6-4
A grenade explodes while in a trajectory of projectile motion. Because there has been no external force involved in the explosion, the motion of the center of mass is unchanged.

Example 6-3

Suppose a grenade is thrown that has the trajectory shown in Fig. 6-4. If it explodes in midair, only internal forces have acted on the fragments and therefore from Eq. 6.10 the acceleration of the center of mass of the fragments, regardless of their subsequent dispersal, is unchanged by the explosion, and thus follows the original trajectory.

6.4 MOMENTUM AND ITS CONSERVATION

Recall Newton's approach to mechanics from Chapter 4. He said that an impulse applied to a body will change its state of momentum (Eq. 4.1).

$$\mathbf{F}\,\Delta t = \Delta\, m\mathbf{v} \qquad \mathbf{F}\,\Delta t = m\mathbf{v}_f - m\mathbf{v}_0 \tag{4.1}$$

where $\mathbf{v}_0$ is the velocity of the body before the force begins to act on it and $\mathbf{v}_f$ is the velocity when the force stops acting on the body.

Momentum is often represented by the letter **p**,

$$\mathbf{F}\,\Delta t = \mathbf{p}_f - \mathbf{p}_0 \tag{4.1'}$$

If there is no external force exerted on a mass, the left side of Eq. 4.1′ is zero and we may write

$$\mathbf{p}_0 = \mathbf{p}_f \tag{6.11}$$

This simple equation is called the law of *conservation of momentum*. It is important to recognize that momentum is a vector and that Eq. 6.11 must be satisfied in all three cartesian coordinates.

We can easily extend this law to a system of particles using the results developed in the preceding section.

If the resultant of the external forces acting on all the particles is zero, then from Eq. 6.10 the acceleration of the center of mass is $a_{cm} = 0$. Therefore, the velocity of the center of mass will be constant. We can then conclude, from Eq. 6.7 and the equivalent equations for the y and z directions, that the total momentum of all the particles will not change, or

$$\left(\sum_{i=1}^{n} \mathbf{p}_i\right)_{\text{before}} = \left(\sum_{i=1}^{n} \mathbf{p}_i\right)_{\text{after}} \tag{6.12}$$

If $\sum \mathbf{F}_{\text{ext}} = 0$

$$\left(\sum_{i=1}^{n} \mathbf{p}_i\right)_{\text{before}} = \left(\sum_{i=1}^{n} \mathbf{p}_i\right)_{\text{after}}$$

We should note that Eq. 6.12 does not imply that the momenta of the individual particles remains constant. The individual momenta can change as a

result of internal forces such as in Example 6-3, but the total momentum remains unchanged.

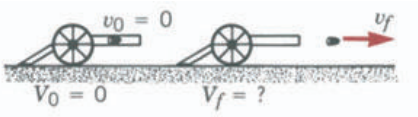


FIGURE 6-5
Example 6-4.

Example 6-4

A cannon of mass 1000 kg fires a 100-kg projectile with a muzzle velocity of 400 m/sec (see Fig. 6-5). With what speed and in what direction does the cannon move?

Solution Let M be the mass of the cannon, $\mathbf{V}_0$ its initial velocity, and $\mathbf{V}_f$ its final velocity. Let m be the mass of the projectile and $\mathbf{v}_0$ and $\mathbf{v}_f$ its initial and final velocities, respectively. If we consider the cannon and projectile as our system of particles, no external force is involved in the firing of the projectile and we conclude that

$$\mathbf{p}_0 = \mathbf{p}_f$$

Substitute the terms on each side of this equation

$$m\mathbf{v}_0 + M\mathbf{V}_0 = m\mathbf{v}_f + M\mathbf{V}_f$$

If we choose the direction of motion of the projectile as the positive direction, and noting that $\mathbf{V}_0$ and $\mathbf{v}_0$ are zero, we get

$$0 + 0 = m\mathbf{v}_f + M\mathbf{V}_f$$

or

$$\mathbf{V}_f = -\frac{m\mathbf{v}_f}{M} = -\frac{100 \text{ kg} \times 400 \text{ m/sec}}{1000 \text{ kg}}$$

$$\mathbf{V}_f = -40 \text{ m/sec}$$

Note that although from experience we know that the cannon will recoil (i.e., move in a direction opposite to that of the projectile), we are not told this as one of the facts of the problem. So we put $\mathbf{V}_f$ in as positive; the vector aspect of the formulation shows in the result (i.e., the fact that $\mathbf{V}_f$ is negative) that the direction of recoil is opposite to that of the projectile.

In the next section we will consider problems in collisions whose solutions involve both momentum conservation and energy accountability. However, as the following example illustrates, some questions about collisions can be answered by momentum conservation alone.

Example 6-5

A 10,000-kg truck traveling east at 20 m/sec collides with a 2000-kg car traveling north at 30 m/sec. After the collision, they are locked together. With what velocity and at what angle do the locked vehicles move immediately after the collision? (See the schematic diagram, Fig. 6-6.)

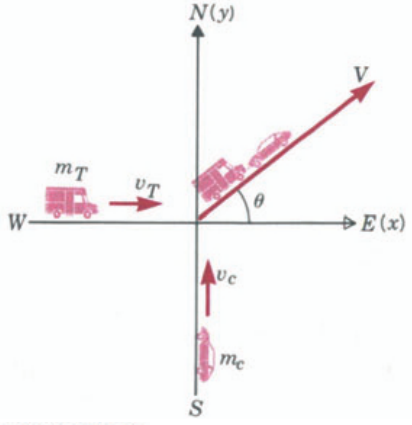


FIGURE 6-6
Example 6-5.

Solution Because no external force is involved in the collision, momentum is conserved. In the x direction

$$p_{x0} = p_{xf}$$

$$m_T v_T = (m_T + m_c)V \cos \theta$$

Rearranging terms,

$$V \cos \theta = \frac{m_T v_T}{m_T + m_c} = \frac{10{,}000 \text{ kg} \times 20 \text{ m/sec}}{10{,}000 \text{ kg} + 2000 \text{ kg}}$$

$$= 16.7 \text{ m/sec}$$

In the y direction,

$$p_{y0} = p_{yf}$$

$$m_c v_c = (m_T + m_c)V \sin \theta$$

Solving for $V \sin \theta$,

$$V \sin \theta = \frac{m_c v_c}{m_T + m_c} = \frac{2000 \text{ kg} \times 30 \text{ m/sec}}{10{,}000 \text{ kg} + 2000 \text{ kg}}$$

$$= 5.0 \text{ m/sec}$$

First find the angle by dividing the two velocity components

$$\frac{\cancel{V} \sin \theta}{\cancel{V} \cos \theta} = \tan \theta = \frac{5.0 \text{ m/sec}}{16.7 \text{ m/sec}} = 0.30$$

$$\theta = \arctan 0.30 = 16.7°$$

Find V by substituting the angle into either the x or y momentum solutions

$$V \sin 16.7° = 5 \text{ m/sec}$$

$$V = 17.4 \text{ m/sec}$$

or

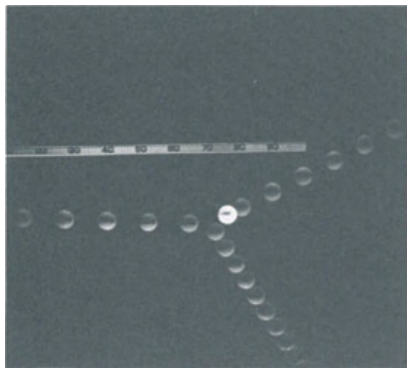
$$V \cos 16.7° = 16.7 \text{ m/sec}$$

$$V = 17.4 \text{ m/sec}$$

6.5 COLLISIONS

One of the most important applications of the conservation of momentum law occurs in the theory of collisions. We will deal only with collisions between two bodies, as it is exceedingly difficult to obtain any but approximate solutions for three-body collisions. There are two types of collisions, to which we give the names *elastic* and *inelastic*. In an elastic collision kinetic energy is

conserved (i.e., no energy is lost from the system). This type of collision can occur only between atomic particles, although in physics problems we often assume elastic collisions between colliding bodies. In actuality, there are no elastic collisions, but in some the energy loss is very small and it may be considered negligible. An inelastic collision is one in which kinetic energy is not conserved (e.g., some energy is lost to friction, crumpled fenders, or such).



Multiflash photograph of a collision between a moving ball coming in from the left and a stationary ball.

6-5a Elastic Collisions

Example 6-6

A neutron with a mass of $m = 1$ u (atomic mass unit) strikes a larger atom at rest and rebounds elastically along its original path with 0.9 of its initial forward velocity. What is the mass M, in atomic mass units, of the atom it struck?

Solution Let v_0 be the initial velocity of the neutron and $v_f = -0.9\ v_0$ its final velocity. Note that the problem tells us that it rebounds; therefore the direction of the final velocity is opposite to its initial velocity. Let M be the mass of the atom, V_0 its initial velocity, and V_f its velocity after collision. Both momentum and kinetic energy are conserved. From the conservation of momentum

$$mv_0 + MV_0 = mv_f + MV_f$$

and on rearranging and using the fact that $V_0 = 0$

$$V_f = \frac{m(v_0 - v_f)}{M}$$

$$= \frac{1\text{ u}(v_0 + 0.9\ v_0)}{M} = \frac{(1\text{ u})(1.9\ v_0)}{M}$$

From the conservation of kinetic energy

$$\frac{1}{2}mv_0^2 + \frac{1}{2}MV_0^2 = \frac{1}{2}mv_f^2 + \frac{1}{2}MV_f^2$$

Solving for M, noting that $V_0 = 0$, we obtain

$$M = \frac{m(v_0^2 - v_f^2)}{V_f^2} = \frac{(1\text{ u})(0.19\ v_0^2)}{V_f^2}$$

If we substitute for V_f from the momentum equation into the energy equation, we obtain

$$M = \frac{(1\text{ u})(0.19\ v_0^2)\ M^2}{(1\text{ u}^2)(3.61\ v_0^2)}$$

which simplifies to

$$M = \frac{(1\ \text{u}^2)(3.61)}{(1\ \text{u})\ (0.19)}$$

$$= 19\ \text{u}$$

Example 6-7

An important type of elastic collision at the atomic level, whose results we will use later, is the collision between a very small mass particle, such as an electron, with another particle of comparatively large mass, such as an atom. The mass of a copper atom, for example, is about 10^5 times that of an electron. In this type of collision one is often interested in finding the velocity of the electron after the collision with the copper atom. To solve this type of collision, we follow the usual procedure of conserving momentum and kinetic energy. We will assume a one-dimensional collision.

Solution Let m, v_0, and v_f be the mass and the initial and the final velocity of the electron and M, V_0, and V_f those of the atom. From the conservation of momentum law

$$mv_0 + MV_0 = mv_f + MV_f$$

and, on rearranging,

$$M(V_0 - V_f) = m(v_f - v_0) \tag{6.13}$$

Conserving kinetic energy yields

$$\frac{1}{2}mv_0^2 + \frac{1}{2}MV_0^2 = \frac{1}{2}mv_f^2 + \frac{1}{2}MV_f^2$$

or

$$M(V_0^2 - V_f^2) = m(v_f^2 - v_0^2)$$

On factoring,

$$M(V_0 + V_f)(V_0 - V_f) = m(v_f + v_0)(v_f - v_0) \tag{6.14}$$

Dividing Eq. 6.14 by Eq. 6.13 obtains

$$V_0 + V_f = v_0 + v_f \tag{6.15}$$

For simplicity, let us have the atom initially at rest, $V_0 = 0$, and substitute V_f of Eq. 6.15 into Eq. 6.13,

$$-M\ (v_0 + v_f) = m\ (v_f - v_0)$$

$$-Mv_0 - Mv_f = mv_f - mv_0$$

Rearranging,

$$-v_f(M + m) = v_0(M - m)$$

$$v_f = -v_0 \frac{(M - m)}{(M + m)} \tag{6.16}$$

For our case $m << M$; therefore Eq. 6.16 reduces to

$$v_f \approx -v_0 \tag{6.17}$$

The electron rebounds (recoils) with the same magnitude of velocity; thus it does not lose any kinetic energy; therefore the atom does not gain any and is not set in motion. This is, of course, an approximation. If the mass of the electron is taken as 1 unit and the mass of the atom as 10^5 units and these numbers are substituted into Eq. 6.16, then $v_f = -0.99998\, v_0$. Because kinetic energy is proportional to v^2, the remaining energy of the recoiling electron is 0.99996 of its initial kinetic energy. Therefore, in the collision 0.00004 of the initial kinetic energy of the electron has been transferred to the atom. This concept will be important later when we develop the loss of electron energy to atoms in an electrical conductor in which electrons flow as a current. The increase in energy of the atoms from electron collisions manifests itself as an increase in temperature of the conductor.

6.5b Inelastic Collisions

In all the preceding examples, the energy of the system was unchanged by the collision. We now give an example of a collision in which the energy is changed by the collision.

Example 6-8

A ballistic pendulum is used to measure the velocity of a bullet. The bullet is shot into a wooden block suspended by strings. It lodges in the block, losing energy in its penetration, and the increase in the height of the swinging block and bullet is measured (see Fig. 6-7). If the bullet has a mass of 0.01 kg, the block has a mass of 0.5 kg, and the swing rises 0.1 m, what was the velocity of the incident bullet and what fraction of its energy was lost during penetration?

Solution We first conserve momentum between situation (*a*) and (*b*) in Fig. 6-7.

$$\mathbf{p}_0 = \mathbf{p}_f$$

$$mv + 0 = (m + M)V$$

$$0.01 \text{ kg } v = 0.51 \text{ kg } V$$

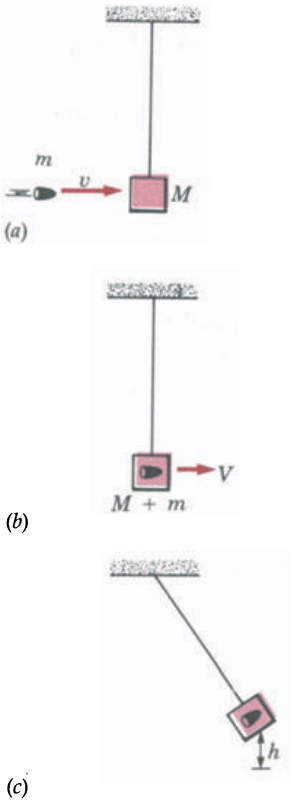


FIGURE 6-7
Example 6-8.

We saw in Section 5.6 in the discussion of the pendulum that the string does no work on the block. We may therefore conserve mechanical energy between situations (*b*) and (*c*) in Fig. 6-7.

$$(E_k)_b = (E_p)_c$$

$$\frac{1}{2}(m + M)V^2 = (m + M)gh$$

$$V = \sqrt{2gh} = 1.4 \text{ m/sec}$$

Substitute this value into the momentum equation and obtain

$$v = \frac{0.51 \text{ kg}}{0.01 \text{ kg}} \times 1.4 \text{ m/sec} = 71.4 \text{ m/sec}$$

We find the fraction of the bullet's initial energy lost in penetration by calculating the energy of the system before (situation *a*) and after the collision (situation *b*).

$$(E_k)_a = \frac{1}{2}mv^2 = \frac{1}{2}(0.01 \text{ kg})(71.4 \text{ m/sec})^2 = 25.5 \text{ J}$$

$$(E_k)_b = \frac{1}{2}(m + M)V^2 = \frac{1}{2}(0.51 \text{ kg})(1.4 \text{ m/sec})^2 = 0.5 \text{ J}$$

The fraction remaining is

$$\text{Fraction remaining} = \frac{(E_k)_b}{(E_k)_a} = \frac{0.5 \text{ J}}{25.5 \text{ J}} = 0.02$$

and the fraction lost is

$$\text{Fraction lost} = 1 - \text{fraction remaining} = 0.98$$

When colliding objects stick together we find that the kinetic energy is not conserved. We should note that energy can be lost, and therefore the collision is inelastic, in certain cases where objects do not stick together.

PROBLEMS

6.1 The equilibrium separation between the centers of the sodium ($m = 23$ u) and chlorine ($m = 35$ u) ions in the sodium chloride molecule is 2.4×10^{-10} m. Where is the center of mass of the molecule?

6.2 In the Bohr model of the hydrogen atom, the electron ($m = 9.1 \times 10^{-31}$ kg) revolves around a proton ($m = 1.67 \times 10^{-27}$ kg) in a circular orbit of radius $r = 0.5 \times 10^{-10}$ m. Where is the center of mass of the hydrogen atom?

6.3 What are the x and y coordinates of the center of mass of the system of particles shown in Fig. 6-8?

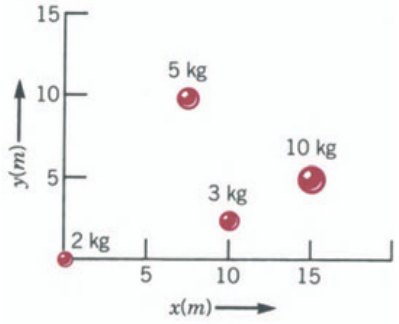


FIGURE 6-8
Problem 6.3.

6.4 A thorium nucleus ($m = 232$ u), at rest at the origin of a coordinate system, decays into a radium nucleus ($m = 228$ u) and an alpha particle ($m = 4$ u). Sometime later, the alpha particle passes the point $x = 3$ m, $y = 2$ m with a velocity $v = 2 \times 10^6$ m/sec. What is the position and the velocity of the radium nucleus at that moment?

(*Answer*: $x = 5.3 \times 10^{-2}$ m, $y = 3.5 \times 10^{-2}$ m, $v = 3.5 \times 10^4$ m/sec.)

6.5 Two particles of mass $m_1 = 1$ kg and $m_2 = 99$ kg are held 2 m apart. The particles attract each other with a constant force directed along the line joining the two particles. (a) When the particles are released, where will the collision occur? (b) Does the answer to (a) depend on the actual value of the force?

(*Answer*: (a) 1.98 m from m_1, (b) no.)

6.6 A 0.25-kg baseball has an initial velocity toward a bat of 15 m/sec. The batter strikes the ball and it goes out in the reverse direction at 30 m/sec. (a) What is the change in the momentum of the ball? (b) What is its change in kinetic energy?

6.7 A swimmer in a pool makes a racing turn at the end by suddenly straightening his legs while his feet are pressed against the end of the pool. If his mass is 80 kg and he exerts an average force of 120 N for 0.8 sec, what is his initial velocity on leaving the pool end?

6.8 A car crashes into a tree. If the car has a mass of 1200 kg and its speed is reduced from 30 m/sec to zero in 0.2 sec, what is the average force exerted by the tree on the car?

6.9 A fire hose delivers water at the rate of 20 kg/sec with a speed of 25 m/sec. A riot police officer uses the hose to control an unruly crowd. The water from the hose strikes a person horizontally and then falls down to the ground. What is the average force experienced by that person?

(*Answer*: 500 N.)

6.10 A gun fires a 0.01-kg bullet with a velocity of 250 m/sec at a 0.5-kg melon resting on a post. The bullet penetrates the melon and leaves the back of it with a velocity of 100 m/sec. With what velocity and in what direction does the melon leave the post?

6.11 A radium atom at rest with a mass of 226 u suddenly emits an alpha particle of mass 4 u with a speed of 2×10^7 m/sec. With what speed and in what direction does the resulting radon atom of mass 222 u move?

6.12 A truck of mass 5×10^3 kg moving at 20 m/sec collides head-on with a car of mass 1×10^3 kg moving at 25 m/sec in the opposite direction. If they stick together after the collision, in what direction and with what speed do they move immediately after the collision?

6.13 An atom of mass 10 u strikes a stationary atom of mass M and rebounds elastically with one half its original velocity. What is the mass of the atom it struck?

(*Answer*: 30 u.)

6.14 A sled of mass 10 kg slides on level, frictionless ice with a velocity of 12 m/sec. It collides elastically with another sled of different mass pointed in the same direction but at rest. After the collision, the first sled continues in the same direction but with a velocity of 4 m/sec. What is the mass of the second sled and its velocity after the collision?

(*Answer*: 5 kg, 16 m/sec.)

6.15 A 9000-N open-top railroad car is coasting with a velocity $v = 10$ km/h on a frictionless horizontal track. A 1200-kg meteorite falls vertically into the car with a velocity of 200 km/h. (a) What is the velocity of the railroad car after the meteorite lands on it? (b) What is the magnitude and the direction of the impulse of the meteorite on the car?

6.16 An object at rest in space explodes into three equal parts. The velocities of two of them are, re-

spectively, $2i$ and $-4j$. Find the resulting velocity of the third part.

6.17 A bullet of mass 80 g is moving east with a velocity v_0. The bullet strikes a 200-g wooden block moving south with a velocity of 2 m/sec. The bullet remains embedded in the block, which then moves in the direction 37° south of east. (a) What is the velocity of the block after the collision? (b) What was the initial velocity v_0 of the bullet? (c) What is the fractional change in the energy of the system?

(*Answer*: (a) 2.37 m/sec, (b) 6.65 m/sec, (c) 64% decrease.)

6.18 Two particles of mass $m_1 = 5$ kg and $m_2 = 2$ kg move toward each other as shown in Fig. 6-9. After the collision, they stick together. (a) What is the speed of the particles after they collide? (b) What is the direction of motion of the particles after the collision? (c) What is the change in the total kinetic energy of the particles?

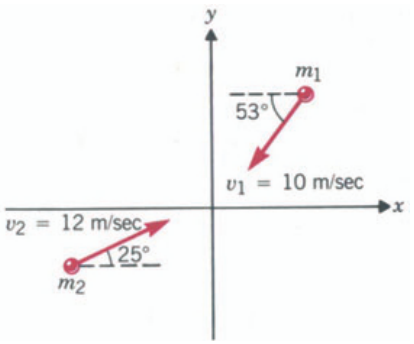


FIGURE 6-9
Problem 6.18.

6.19 A puck sliding on a frictionless table with a velocity $v = 2$ m/sec strikes a second puck of equal mass initially at rest. The collision is elastic, and it is found that after the collision both pucks move with the same speed. (a) What is the speed of the pucks after the collision? (b) What is the angle between the directions of motion of the pucks?

(*Answer*: (a) 1.41 m/sec, (b) 90°.)

6.20 Three boys stand on a 10-kg wagon resting on a frictionless horizontal surface. The boys take turns running off the same end of the wagon with a velocity of 1.5 m/sec relative to the wagon. The mass of each of the boys is 40 kg. What is the final velocity of the wagon?

(*Answer*: 7.87 m/sec.)

6.21 A 2-kg block rests on the ground. The coefficient of friction between the block and the ground is 0.4. A man fires a 0.01-kg bullet parallel to the ground. It lodges in the block, and the block and bullet are observed to slide 2 m before coming to rest. What was the velocity of the bullet?

(*Answer*: 796 m/sec.)

6.22 A block of mass 1 kg rests over a hole in a tabletop. A bullet of mass 0.01 kg is fired upward into the block with a velocity of 200 m/sec. If the bullet imbeds itself in the block, how high will the block rise?

6.23 A bullet of mass 100 g is shot into a 3-kg wooden block resting on an incline plane, as shown in Fig. 6-10. The bullet remains embedded in the block, which then slides down the incline plane 2 m before coming to rest. The coefficient of friction between the block and the incline is 0.5. What was the initial velocity of the bullet?

(*Answer*: 91.9 m/sec.)

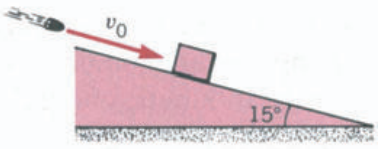


FIGURE 6-10
Problem 6.23.

6.24 A bomb is launched with a velocity $v_0 = 500$ m/sec at an angle of 37° with the horizontal. At the highest point of the trajectory it explodes in two equal pieces. One piece lands 20 sec later directly below the point where the explosion occurred. When and where does the second piece land?

(*Answer*: 77.8 sec, 4.99×10^4 m from launching.)

CHAPTER 7

Rotational Motion

7.1 INTRODUCTION

In the preceding chapters we dealt with translational motion—that is, the change of position in the cartesian coordinate system of the center of mass of a body. However, some systems simply rotate and some rotate while the center of mass translates through space: A roulette wheel simply rotates, whereas a car wheel both rotates and translates. Newton's laws and momentum and energy conservation still apply, but the formulation is somewhat different. In this chapter and the next we will consider rotational motion. We need these properties in order to build the model of the electron rotational motion in atoms.

7.2 MEASUREMENT OF ROTATION

The most common measurement of rotation is a count of the number of revolutions about an axis of rotation. We also use degrees as a measure, where 360° corresponds to one revolution. In physics we mostly use radians for a variety of reasons. One of these reasons is that the formulation affords a quick and easy bridge between linear and rotational motion. Let us examine this.

A measure of an angle in radians is the length of the circular arc subtended by the angle divided by the radius of the circle (see Fig. 7-1). If the length of the arc from a to b is s and r is the radius, then the measure of angle θ in radians is given by

$$\theta \text{ (in radians)} = \frac{s}{r} \tag{7.1}$$

$\theta = \frac{s}{r}$

Because both s and r are in units of length, the units cancel on the right side and θ is dimensionless. Other measures of θ, such as degrees or revolutions, are also dimensionless. However, the numerical magnitudes of these quantities differ, so we must state the system of measure used. Obviously, we must maintain a consistency of angular measure in a given problem.

Returning to Eq. 7.1, we may ask how many radians there are in a whole revolution. The arc length subtended by a revolution is the circumference, or $2\pi r$. Therefore,

$$\frac{s \text{ of 1 revolution}}{\text{radius}} = \frac{2\pi r}{r} = 2\pi \text{ radians (rad)}$$

We may convert between systems of angular measure as shown in Chapter 1 using the identities

$$2\pi \text{ (radians)} = 360 \text{ (degrees)} = 1 \text{ revolution (rev)}$$

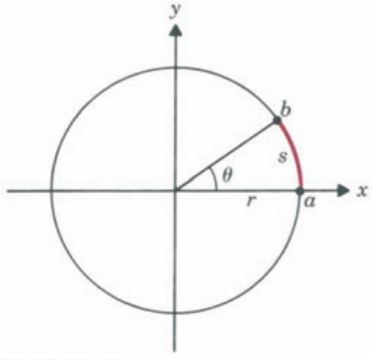


FIGURE 7-1

7.3 ROTATIONAL MOTION

Suppose we have a reference marker a on the x axis of a coordinate system and a wheel whose center coincides with the origin. We also have a mark b on the wheel. We can measure the time during which the wheel marker b moves from the coordinate marker a, a distance Δs (see Fig. 7-2). The $\text{speed}_{a\to b}$ of the marker on the wheel is measured by the time it takes for the marker to move an arc length Δs, or

$$\text{speed}_{a\to b} = \frac{\Delta s}{\Delta t}$$

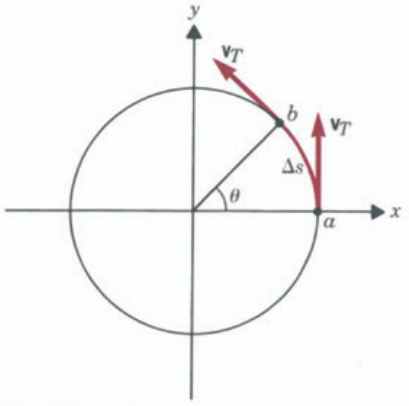


FIGURE 7-2

In the limit $\Delta t \to 0$ the distance Δs becomes a vector and the speed becomes $\mathbf{v}$, the instantaneous velocity (see Section 3.2)

$$\mathbf{v} = \frac{d\mathbf{s}}{dt} \tag{7.2}$$

The direction of the instantaneous velocity of the marker on the rotating wheel is the tangent to the circle of motion and is called the *tangential* velocity; it is sometimes indicated by writing $\mathbf{v}$ with the subscript T. Note that as the marker rotates the direction of $\mathbf{v}_T$ constantly changes even though the marker may rotate at a constant rate. Therefore, the vector $\mathbf{v}_T$ is constantly changing. In the previous chapters we have dealt largely with vectors whose direction remained constant while the magnitude changed, whereas here we have a vector whose magnitude may remain constant while its direction always changes. This has important implications in the development of the centripetal force that we will consider later in this chapter.

In Fig. 7-2 we see that while s is increasing, the angle θ is also increasing as the moving radius vector (line from origin to the marker on the circumference) sweeps out a larger arc. The average rate of change of the angle θ with time is called the average *angular* or *rotational velocity;* we use the small Greek letter ω (omega) for this.

$$\bar{\omega} = \frac{\Delta\theta}{\Delta t}$$

and, as $\Delta t \to 0$, the average angular velocity $\bar{\omega}$ becomes the instantaneous angular velocity ω, namely,

$$\omega = \frac{d\theta}{dt} \tag{7.3}$$

$$\omega = \frac{d\theta}{dt}$$

Because θ is dimensionless, ω has units of reciprocal time, although it must be specified whether ω is radians/second, degrees/second or revolutions/second. We may relate the tangential velocity to the rotational velocity by differentiating Eq. 7.1 with respect to time

$$\frac{1}{r}\frac{ds}{dt} = \frac{d\theta}{dt}$$

Rearranging terms and using the definitions of Eqs. 7.2 and 7.3 we obtain

$$v_T = r\omega \tag{7.4}$$

$v_T = r\omega$

where ω is in radians/second.

Example 7-1

A car is traveling at a constant velocity of 24 m/sec. The radius of its wheels is $r = 0.30$ m. (a) How many revolutions have the wheels turned after the car has gone 120 m? (b) How many revolutions have the wheels turned after 60 sec?

Solution

(*a*) If there is no slipping between the wheels of the car and the road, the arc length moved by a marker on the outermost radius of the wheel is equal to the distance traveled by the car; that is, $s = 120$ m. Using Eq. 7.1

$$\theta = \frac{s}{r} = \frac{120 \text{ m}}{0.30 \text{ m}}$$

$$\theta = 400 \text{ rad} = 400 \text{ rad} \left(\frac{1 \text{ rev}}{2\pi \text{ rad}}\right) = 63.7 \text{ rev}$$

(*b*) Because the car travels at constant velocity, the distance traveled by the car in 60 sec can be found with Eq. 3.12, keeping in mind that the acceleration $a = 0$

$$x = (24 \text{ m/sec})(60 \text{ sec}) = 1440 \text{ m}$$

This, as we have indicated, is also the arc length moved by a marker on the rim of the wheel. We now use Eq. 7.1 to find the angle rotated by the wheel in 60 sec.

$$\theta\ (t = 60 \text{ sec}) = \frac{s}{r}$$

$$\theta = \frac{1440 \text{ m}}{0.30 \text{ m}} = 4800 \text{ rad} = 4800 \text{ rad} \left(\frac{1 \text{ rev}}{2\pi \text{ rad}}\right) = 764 \text{ rev}$$

Suppose the marker on the rotating wheel of Fig. 7-2 is not rotating at a constant rate but is speeding up or slowing down. Then from Eq. 7.4 the marker on the wheel has an average tangential acceleration that from Eq. 3.4 is

$$\bar{\mathbf{a}}_T = \frac{\Delta \mathbf{v}_T}{\Delta t}$$

$$= \frac{\mathbf{v}_{Tf} - \mathbf{v}_{T0}}{\Delta t}$$

In the limit as $\Delta t \to 0$, $\bar{\mathbf{a}}_T$ becomes the instantaneous acceleration $\mathbf{a}_T$, that is,

$$\mathbf{a}_T = \frac{d\mathbf{v}_T}{dt} \tag{7.5}$$

The rate at which the angle θ is being swept out, the angular velocity ω, is also changing. We call the rate of change of ω the average *angular* or *rotational* acceleration and use as the symbol the small Greek letter α (alpha), so that

$$\bar{\alpha} = \frac{\Delta\omega}{\Delta t}$$

To find the instantaneous value we let $\Delta t \to 0$, and

$$\alpha = \frac{d\omega}{dt} \tag{7.6}$$

$\alpha = \frac{d\omega}{dt}$

Angular acceleration has dimensions of $(\text{time})^{-2}$, although, as before, we must specify the measure of the angle. We may connect the angular acceleration with the tangential acceleration of the marker on the wheel by differentiating Eq. 7.4 with respect to time.

$$\frac{dv_T}{dt} = r\frac{d\omega}{dt}$$

$$a_T = r\alpha \tag{7.7}$$

$a_T = r\alpha$

where α is in radians/second squared.

Because the radian is a dimensionless quantity, the units of a_T will be the same as those of r divided by second^2. Thus if r is expressed in meters, a_T will be in m/sec^2.

Example 7-2

A driver of a car traveling at 24 m/sec applies the brakes, decelerates uniformly, and comes to a stop in 100 m. If the wheels have a radius of 0.30 m, what is the angular deceleration of the wheels in rev/sec^2?

Solution There are several ways to solve this, but let us use the most straightforward one, finding first the linear deceleration and then relating it to rotational deceleration

$$v_0 = 24 \text{ m/sec}, \quad v_f = 0, \quad x = 100 \text{ m}, \quad a = ?$$

From Eq. 3.11,

$$v_f^2 - v_0^2 = 2ax$$

$$a = \frac{v_f^2 - v_0^2}{2x} = \frac{0 - (24 \text{ m/sec})^2}{(2)(100 \text{ m})} = -2.88 \text{ m/sec}^2$$

Because a is the acceleration of the car, it is also the tangential acceleration of every point on the rim of its wheels (assuming no slipping between the

wheels and the road). By Eq. 7.7

$$a_T = r\alpha$$

$$\alpha = \frac{a_T}{r} = \frac{-2.88 \text{ m/sec}^2}{0.30 \text{ m}}$$

$$\alpha = -9.6 \frac{\text{rad}}{\text{sec}^2} \frac{1 \text{ rev}}{2\pi \text{ rad}} = -1.5 \text{ rev/sec}^2$$

7.4 EQUATIONS OF ROTATIONAL MOTION

We may derive the equations of rotational motion by the method of Chapter 3. As in Chapter 3, we will limit our discussion to the case of constant angular acceleration.

The basic relations obtained by arguments analogous to those of Eqs. 3.7 and 3.9 are

$$\theta = \bar{\omega} t \tag{7.8}$$

$$\theta = \bar{\omega} t$$

$$\bar{\omega} = \frac{\omega_0 + \omega_f}{2} \tag{7.9}$$

$$\bar{\omega} = \frac{\omega_0 + \omega_f}{2}$$

We may obtain the other three relations analogous to Eqs. 3.8, 3.11, and 3.12 by integration for the condition that α = constant.
From the definition of α, Eq. 7.6,

$$\alpha = \frac{d\omega}{dt}$$

$$\int_{\omega_0}^{\omega} d\omega = \alpha \int_0^t dt$$

$$\omega - \omega_0 = \alpha t$$

$$\omega = \omega_0 + \alpha t \tag{7.10}$$

$$\omega = \omega_0 + \alpha t$$

From the definition

$$\omega = \frac{d\theta}{dt}$$

$$\int_{\theta_0}^{\theta} d\theta = \int_0^t \omega \, dt$$

Substitute Eq. 7.10 for ω

$$\int_{\theta_0}^{\theta} d\theta = \int_0^t (\omega_0 + \alpha t) dt = \omega_0 \int_0^t dt + \alpha \int_0^t t \, dt$$

$$\theta - \theta_0 = \omega_0 t + \frac{1}{2} \alpha t^2 \tag{7.11}$$

$$\theta = \theta_0 + \omega_0 t + \frac{1}{2} \alpha t^2$$

To obtain our final equation, we use the chain rule to write

$$\alpha = \frac{d\omega}{dt} = \frac{d\omega}{d\theta}\frac{d\theta}{dt} = \omega\frac{d\omega}{d\theta}$$

Multiply both sides by $d\theta$

$$\alpha\, d\theta = \omega\, d\omega$$

and the integration is

$$\alpha\int_{\theta_0}^{\theta} d\theta = \int_{\omega_0}^{\omega} \omega\, d\omega$$

$$\alpha(\theta - \theta_0) = \frac{1}{2}(\omega^2 - \omega_0^2)$$

which is usually written in the form

$$\omega^2 - \omega_0^2 = 2\alpha(\theta - \theta_0) \qquad (7.12)$$

$\omega^2 = \omega_0^2 + 2\alpha(\theta - \theta_0)$

Example 7-3

A roulette wheel is given an initial rotational velocity of 2 rev/sec. It is observed to be rotating at 1.5 rev/sec 5 sec after it was set in motion. (*a*) What is the angular deceleration (assumed constant) of the wheel? (*b*) How long will it take to stop? (*c*) How many revolutions will it make from start to finish?

Solution

(a) $\omega_0 = 2.0$ rev/sec $\quad \omega = 1.5$ rev/sec $\quad t = 5$ sec $\quad \alpha = ?$

$$\omega = \omega_0 + \alpha t$$

$$\alpha = \frac{\omega - \omega_0}{t} = \frac{1.5 \text{ rev/sec} - 2.0 \text{ rev/sec}}{5 \text{ sec}}$$

$$\alpha = -0.1 \text{ rev/sec}^2$$

(b) $\omega_0 = 2$ rev/sec $\quad \omega_f = 0 \quad \alpha = -0.1$ rev/sec^2 $\quad t_f = ?$

$$\omega_f = \omega_0 + \alpha t_f$$

$$t_f = \frac{\omega_f - \omega_0}{\alpha} = \frac{0 - 2 \text{ rev/sec}}{-0.1 \text{ rev/sec}^2}$$

$$t_f = 20 \text{ sec}$$

(c) $\omega_0 = 2$ rev/sec $\quad \omega_f = 0 \quad t_f = 20$ sec $\quad \theta = ?$

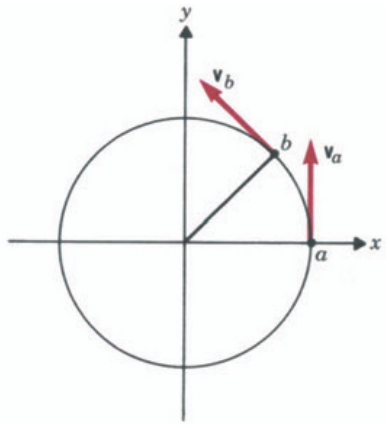
$$\theta = \frac{\omega_0 + \omega_f}{2} t_f = \frac{2 \text{ rev/sec} + 0}{2} \times 20 \text{ sec}$$

$$\theta = 20 \text{ rev}$$

7.5 RADIAL ACCELERATION

Let us consider more carefully the motion of the marker on the wheel in Fig. 7-2 as the wheel rotates at constant speed. In Fig. 7-3*a* we have the same wheel with the velocity vectors indicated at points *a* and *b*. We see that even though the velocity vectors at points *a* and *b* may have the same magnitude, their direction is different. This difference is indicated in Fig. 7-3*b* by the vector $\Delta\mathbf{v}_\perp$. In this figure the vector tails have been put at a common point. Thus, in a time Δt the vector $\mathbf{v}_a$ has changed in value by $\Delta\mathbf{v}_\perp$. This implies an acceleration has taken place. Because a velocity vector of a point on a circle is always tangent to the circle, it is perpendicular to the radius. For infinitesimal changes Δt and thus $\Delta\mathbf{v}_\perp$, there is an acceleration $\mathbf{a}_R$ inward along the radius called a *radial acceleration* $\mathbf{a}_R$ given by

$$\mathbf{a}_R = \lim_{\Delta t \to 0} \frac{\Delta \mathbf{v}_\perp}{\Delta t \to 0}$$



(*a*)

(*b*)

FIGURE 7-3

We will now examine this radial acceleration analytically. We see from Fig. 7-4*a* by the method of vector components of Chapter 2 that the coordinates of the marker at some time *t* are

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned} \tag{7.13}$$

Let the marker rotate about the circle at a constant rotational velocity ω so that $\overline{\omega} = \omega$. Substitute Eq. 7.8 into Eqs. 7.13 and obtain

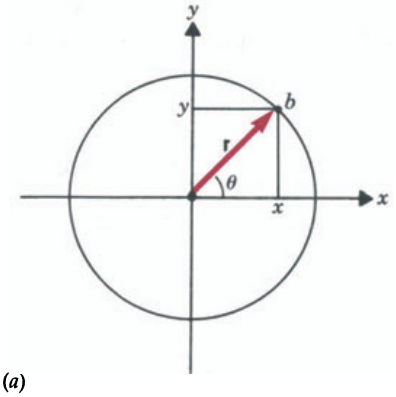
$$\begin{aligned} x &= r\cos\omega t \\ y &= r\sin\omega t \end{aligned} \tag{7.14}$$

The *x* component of the velocity of the marker in Fig. 7-4*b* is $v_x = dx/dt$, and the *y* component is $v_y = dy/dt$. Performing this differentiation of Eqs. 7.14 yields

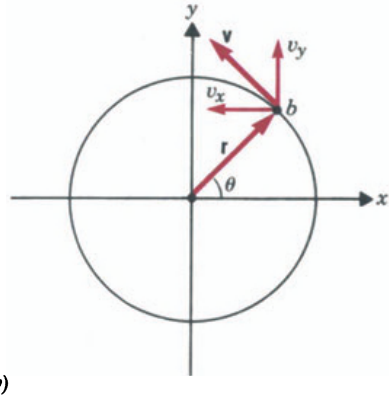
$$\begin{aligned} v_x &= r\frac{d}{dt}(\cos\omega t) = -r\omega\sin\omega t \\ v_y &= r\frac{d}{dt}(\sin\omega t) = r\omega\cos\omega t \end{aligned} \tag{7.15}$$

We see that in Eqs. 7.15 both v_x and v_y are functions of time and therefore the point must be accelerating in both the *x* and *y* directions. We may obtain the components of acceleration by differentiating Eqs. 7.15 with respect to time:

$$\begin{aligned} a_x &= \frac{dv_x}{dt} = -r\omega\frac{d}{dt}(\sin\omega t) = -r\omega^2\cos\omega t \\ a_y &= \frac{dv_y}{dt} = r\omega\frac{d}{dt}(\cos\omega t) = -r\omega^2\sin\omega t \end{aligned} \tag{7.16}$$



(*a*)



(*b*)

FIGURE 7-4

The square of the resultant acceleration a_R^2 is the sum of the squares of the components, or

$$a_R^2 = a_x^2 + a_y^2$$

$$a_R^2 = r^2\omega^4\cos^2\omega t + r^2\omega^4\sin^2\omega t$$

$$= r^2\omega^4(\cos^2\omega t + \sin^2\omega t)$$

Using the trigonometric identity that

$$\sin^2\theta + \cos^2\theta = 1$$

we obtain

$$a_R^2 = r^2\omega^4$$

or

$$a_R = \pm r\omega^2 \qquad (7.17)$$

$a_R = r\omega^2$

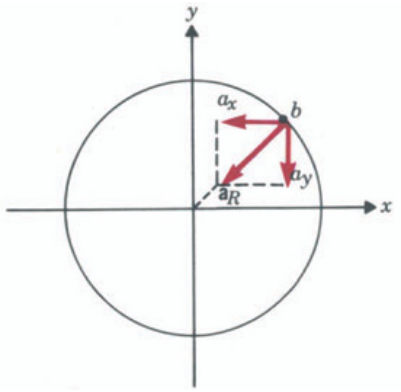


FIGURE 7-5

The direction of $\mathbf{a}_R$ can be found by comparing Eq. 7.16 with Eq. 7.14. It is seen that a_x is ω^2 times the *negative* x coordinate of the radius vector $\mathbf{r}$ and a_y is ω^2 times the *negative* y coordinate of $\mathbf{r}$. This implies that $\mathbf{a}_R = -\omega^2\,\mathbf{r}$, and, consequently, the direction of $\mathbf{a}_R$ is along the radius toward the center, that is, opposite to the vector direction of the radius. This is sketched in Fig. 7-5, where the resultant $\mathbf{a}_R$ is seen to be directed inward along the radius *toward* the center.

7.6 CENTRIPETAL FORCE

Newton's second law, $\mathbf{F} = m\mathbf{a}$, states that if there is a net force on a body there is an associated acceleration. The converse is true; if there is an acceleration there must be a net force. We have shown in the previous section that a particle or a body moving in circular motion at constant speed is being accelerated inward along the radius. Therefore, the particle must be acted on by a force along the radius toward the center. This situation corresponds to the statement of Newton's second law that "If a body in a state of motion is acted on by an external force, it will be accelerated in the direction of the force." The particle *cannot* undergo circular motion unless there is a force along the radius directed inward toward the center. This force is called the *centripetal* (center-seeking) force. A demonstration of this is easily performed by whirling a weight at the end of a string in a circle. You must exert a constant force (tension in the string) to maintain the motion. If you let go of the string, the weight will fly off in a straight line with a velocity whose direction will be the tangent to the circle at the point of release.

We indicate radial (or centripetal) force by F_R. We may use Newton's

second law to write

$$\Sigma F_R = ma_R \tag{7.18}$$

or, using Eq. 7.17

$$\Sigma F_R = mr\omega^2 \tag{7.19}$$

Another convenient form is obtained by substituting Eq. 7.4, $v_T = r\omega$ for ω

$$\Sigma F_R = m\frac{v_T^2}{r} \tag{7.20}$$

$\Sigma F_R = m\frac{v_T^2}{r}$

In the solution of problems involving radial acceleration, two rules must be observed, based on the derivations: (1) a_R has dimensions of m/sec^2 and therefore ω must have dimensions of rad/sec^2; and (2) radial forces directed toward the center of rotation are positive, whereas those directed away from the center are negative. We also note from Section 3-2 that the magnitude of the instantaneous tangential velocity at any point is equal to the speed.

Example 7-4

A person whose weight is 600 N is riding a roller coaster. This person sits on a scale as the roller coaster passes over the top of a rise of radius 80 m. (a) What is the minimum speed of the car if the scale reads zero (the sensation of weightlessness is experienced)? (b) If the car increases its speed to 40 m/sec in descending to a dip with a radius of 80 m, what will the scale read? See Fig. 7-6.

Solution Let us consider the forces on the rider at the rise. The rider's weight, mg, is directed toward the center. The scale exerts a normal force N upward. N is the reading of the scale.

$$\Sigma F_R = m\frac{v_T^2}{r}$$

$$mg - N = m\frac{v_T^2}{r}$$

If the scale reads zero, $N = 0$ and

$$mg = m\frac{v_T^2}{r}$$

$$v_T = \sqrt{gr}$$

$$= \sqrt{9.8 \text{ m/sec}^2 \times 80 \text{ m}} = 28 \text{ m/sec}$$

In the dip $v_T = 40$ m/sec, mg is downward directed away from the center whereas N' now is upward toward the center. Following our sign convention

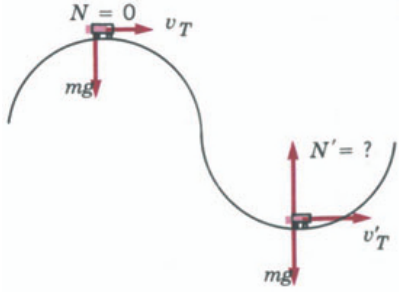


FIGURE 7-6
Example 7-4.

we write

$$-mg + N' = m\frac{v_T'^2}{r}$$

$$N' = mg + m\frac{v_T'^2}{r}$$

$$= 600\text{ N} + \left(\frac{600\text{ N}}{9.8\text{ m/sec}^2}\right)\frac{(40\text{ m/sec})^2}{(80\text{ m})}$$

$$= 600\text{ N} + 1224\text{ N} = 1824\text{ N}$$

Notice that now the scale reads more than three times the person's weight.

Johannes Kepler (1571-1630).

7.7 ORBITAL MOTION AND GRAVITATION

Johannes Kepler (1571–1630), a German astronomer and mathematician, plotted the orbits followed by the planets around the sun. He found three empirical rules for planetary motion, the first two were published in 1609 and the third in 1621. The reason for planetary behavior was not known until Newton found that he could derive Kepler's rules if he postulated a universal gravitational law. Namely, any two bodies are gravitationally attracted to each other by a force proportional to the product of their masses (m_1m_2) and inversely proportional to the square of the distance between them, r^2. If we call the proportionality constant G, the *universal gravitational constant*, we may write

$$F = G\frac{m_1m_2}{r^2} \tag{7.21}$$

$F = G\frac{m_1m_2}{r^2}$

Henry Cavendish (1731-1810).

The value of this constant is $G = 6.67 \times 10^{-11}\text{ Nm}^2/\text{kg}^2$. Such a small number was not measurable in Newton's time, and it was first measured in 1798 by Henry Cavendish (1731–1810). Newton found a method of performing calculations without it. For example, he was able to calculate the acceleration of gravity at the earth's surface, $g = 9.8\text{ m/sec}^2$, which compared favorably with the experimental measurement of the acceleration of falling bodies. He reasoned as follows. Let m_e be the mass of the earth, m_o the mass of an object near the surface of the earth, m_m the mass of the moon, r_e the radius of the earth, and r_{em} the distance from the center of mass of the earth to the center of mass of the moon (assume constant radius of the moon's orbit).

At the surface of the earth the force on an object is its weight m_og. This is equal to the force of gravity between the object and the earth, as given by Eq. 7.21. Considering all the mass of the earth to be at its center of mass, which is the geometrical center of a sphere, then

$$m_og = G\frac{m_om_e}{r_e^2} \tag{7.22}$$

and

$$G = \frac{gr_e^2}{m_e} \tag{7.23}$$

The moon is also attracted to the earth by the gravitational force

$$F = G\frac{m_m m_e}{r_{em}^2} \tag{7.24}$$

This force is the centripetal force, which from Eq. 7.20 is

$$F_R = m_m \frac{v_T^2}{r_{em}}$$

Substituting Eq. 7.24 for this force yields

$$G\frac{m_m m_e}{r_{em}^2} = \frac{m_m v_T^2}{r_{em}}$$

from which

$$G = \frac{v_T^2\, r_{em}}{m_e} \tag{7.25}$$

Equate the G's of Eqs. 7.23 and 7.25

$$g\frac{r_e^2}{\not{m}_e} = \frac{v_T^2\, r_{em}}{\not{m}_e}$$

from which

$$g = \frac{v_T^2\, r_{em}}{r_e^2} \tag{7.26}$$

He had the quantities $r_{em} = 3.8 \times 10^8$ m and $r_e = 6.3 \times 10^6$ m measured by astronomers. v_T is the speed of the moon, which is the distance around its orbit $2\pi r_{em}$ divided by the period of rotation of the moon around the earth, 27.3 days (2.36×10^6 sec)

$$v_T = \frac{2\pi \times 3.8 \times 10^8 \text{m}}{2.36 \times 10^6 \text{sec}} = 1.01 \times 10^3 \text{ m/sec}$$

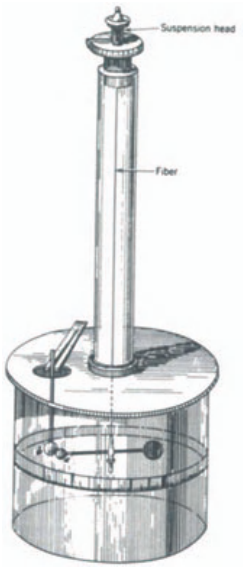
Substituting these numbers into Eq. 7.26 results in

$$g = \frac{(1.01 \times 10^3 \text{ m/sec})^2\,(3.8 \times 10^8 \text{m})}{(6.3 \times 10^6 \text{m})^2}$$

$$= 9.8 \text{ m/sec}^2$$

And thus Newton was able to verify his gravitational law.

Later, when Cavendish measured G, the mass of the earth and the sun could be calculated (see Example 7-5).



Torsion balance used by Cavendish to determine the universal gravitational constant G.

Example 7-5

The radius of the earth is $r_e = 6.3 \times 10^6$m and $G = 6.67 \times 10^{-11}$ Nm²/kg². Find the mass of the earth.

Solution Use Eq. 7.22

$$\cancel{m_o} g = G \frac{\cancel{m_o} m_e}{r_e^2}$$

$$m_e = \frac{r_e^2 g}{G}$$

$$= \frac{(6.3 \times 10^6\text{m})^2(9.8\ \text{m/sec}^2)}{6.67 \times 10^{-11}\ \text{Nm}^2/\text{kg}^2} = 5.8 \times 10^{24}\ \text{kg}$$

The same answer can be obtained if we use Eq. 7.19 instead of Eq. 7.22.

$$F_R = mr\omega^2$$

Applying this to the motion of the moon around the earth, we write

$$G \frac{\cancel{m_m} m_e}{r_{em}^2} = \cancel{m_m} r_{em}\omega^2$$

The rotational velocity must be in rad/sec. One orbit of the moon is 2π rad, which it completes in 27.3 days. Therefore, $\omega = 2\pi/27.3$ day $= 2\pi/2.36 \times 10^6$ sec $= 2.66 \times 10^{-6}$ rad/sec. When this value is used the same answer is obtained for m_e.

$$m_e = \frac{r_{em}^3 \omega^2}{G}$$

$$= \frac{(3.8 \times 10^8\text{m})^3(2.66 \times 10^{-6}\ \text{rad/sec})^2}{6.67 \times 10^{-11}\ \text{Nm}^2/\text{kg}^2}$$

$$= 5.8 \times 10^{24}\ \text{kg}$$

PROBLEMS

7.1 A wheel of radius 0.5 m is rotating at 120 rev/min. (a) What is its rotational speed in rad/sec? (b) What is the tangential velocity of a point of the rim? (c) How many radians does the wheel turn in 10 sec? (d) If the wheel were rolling on the ground, what distance would it travel in 10 sec?

7.2 Calculate the angular velocity of the hour hand, the minute hand, and the second hand of a wristwatch.

7.3 A wheel rotating at 5 rev/sec coasts to rest in 30 sec. (a) What is its deceleration in rev/sec² and in

rad/sec^2? (b) If the radius of the wheel is 0.4 m, what is the tangential acceleration of a point on the rim? (c) Through how many revolutions did the wheel turn in coming to rest?

(*Answer*: (a) -0.167 rev/sec^2, -1.05 rad/sec^2, (b) -0.42 m/sec^2, (c) 75 rev.)

7.4 A bicycle with a wheel radius of 0.34 m is traveling at 10 m/sec. What is the rotational speed of the wheels?

7.5 A wheel rotating at 10 rev/sec makes 1000 rev while coasting to a stop with constant deceleration. How long did it take to stop?

7.6 A wheel of radius 2 m starts rotating with constant angular acceleration $\alpha = 1.5$ rad/sec^2. What are the tangential and radial accelerations of a point on the rim after the wheel has rotated 20π rad?

7.7 A pulley of radius $r_p = 8$ cm is connected to the shaft of an electric motor. A belt couples the pulley to a wheel of radius $r_w = 24$ cm (see Fig. 7-7). The motor shaft begins to rotate with an angular acceleration $\alpha = 25$ rad/sec^2. (a) What is the angular velocity of the wheel after 3 sec? (b) Through what angle has the wheel rotated when the centripetal acceleration of a point on the rim of the wheel is 100 g? ($g = 9.8$ m/sec^2)

(*Answer*: (a) 25 rad/sec, (b) 245 rad.)

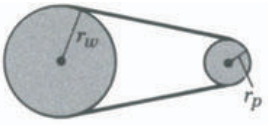


FIGURE 7-7
Problem 7.7.

7.8 A wheel makes 40 rev in 2 sec. The angular velocity at the end of the 2-sec time is 18 rev/sec. (a) What is the angular velocity at the beginning of the 2 sec? (b)What is the angular acceleration (assume it to be constant) of the wheel?

7.9 Assume that the orbit of the earth around the sun is circular and that the period of rotation is 365 days. The earth-sun distance is 1.5×10^{11} m. What is the centripetal acceleration of the earth resulting from its motion around the sun?

7.10 A 0.4-kg object on a string 0.5 m long attached to a pin on a frictionless table is made to rotate. If the breaking strength of the string is 20 N, what is the maximum rotational speed?

(*Answer*: 10 rad/sec.)

An object of mass 0.2 kg on a 0.4-m string is whirled in a vertical circle. (a) If the rotational speed is slowed until the object just completes the top of the circle with no tension in the string, what is its tangential velocity at that point? (b) If the same velocity is maintained at the bottom of the circle, what is the tension in the string at that point? See Fig. 7-8.

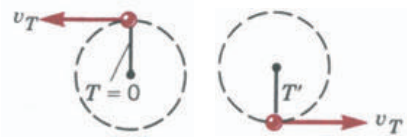


FIGURE 7-8
Problem 7.11.

7.12 A bug sits on a phonograph record 0.18 m from the center. If the record turns at 33 rev/min, what is the radial acceleration of the bug? If it has a mass of 0.5 gm, what is the centripetal force acting on it? See Fig. 7-9.

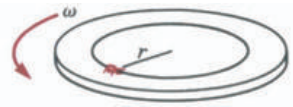


FIGURE 7-9
Problem 7.12.

7.13 The earth-sun distance is 1.5×10^{11} m. If the earth goes around the sun once in 365 days, find the mass of the sun. Assume the earth makes a circular orbit around the sun.

(Answer: 2×10^{30} kg.)

7.14 The force of attraction between oppositely charged particles is given by Coulomb's law, which has the same form as Newton's gravitational law

$$F = K\frac{q_1 q_2}{r^2}$$

where q_1 and q_2 are the charges on the particles in Coulombs (C), r is the distance between them, and K is a constant. In the Bohr model of the hydrogen atom, the electron revolves in a circular orbit around the stationary proton. The magnitude of the charge on the electron is the same as the charge on the proton $q_1 = q_2 = 1.6 \times 10^{-19}$ C and $K = 9 \times 10^9$ Nm^2/C^2. The radius of the smallest electron orbit is 5.3×10^{-11} m, and the mass of the electron is 9.1

$\times 10^{-31}$ kg. Find the number of rev/sec of the electron around the proton, according to the model.

(*Answer*: 6.56×10^{15} rev/sec.)

7.15 What should the duration of a day be in order for a person standing at the equator to have the feeling of weightlessness, namely, for the normal force exerted by the ground on the person to be zero. The radius of the earth is 6.37×10^6 m.

(*Answer*: 1.41 h.)

7.16 (a) What is the centripetal acceleration of a person standing on the earth at a point of latitude 45°? (b) What is the magnitude and the direction of the force exerted by the ground on that person? Express your answer in terms of the weight mg of the person. The radius of the earth is 6.37×10^6 m.

7.17 A 2-kg block is rotating on a frictionless table with angular velocity $\omega = 2$ rev/sec. The block is connected to a 15-kg block by means of a string of total length 2 m that passes through a small hole in the table (see Fig. 7-10). How far below the tabletop does the 15-kg block hang?

(*Answer*: 1.53 m

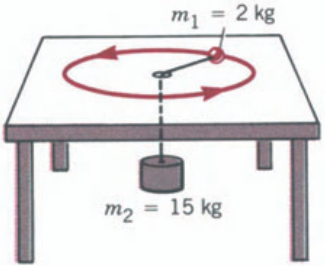


FIGURE 7-10
Problem 7.17.

7.18 A 1.5-kg mass is attached to one end of a rod of length $l = 1$ m and negligible weight. The other end of the rod is pivoted, and the mass rotates in a vertical circle. The tangential velocity of the mass at the top of the circle is 3 m/sec. (a) What is the magnitude and the direction of the force exerted by the rod on the mass at the top of the circle? (b) If friction at the pivot is negligible, what is the tangential velocity at the bottom of the circle? (c) What force does the rod exert on the mass at the bottom?

(*Answer*: (a) 1.20 N upward, (b) 6.94 m/sec, (c) 87 N.)

7.19 A particle of mass $m = 0.7$ kg is released from rest at point A in Fig. 7-11. It slides down and around the frictionless loop. (a) What are the radial and tangential accelerations at points B, C, and D? (b) What is the normal force exerted by the track at those three points? (This is the basis for a popular ride in amusement parks.)

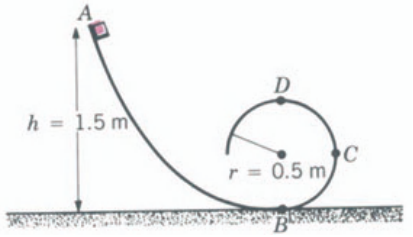


FIGURE 7-11
Problem 7.19.

The length of the string of a conical pendulum is 0.6 m (see Fig. 7-12). The mass of the bob is 1.2 kg. The angular velocity of the bob (which moves in a circle in the horizontal plane) is such that the angle between the string and the vertical is 30° and is constant. (a) What is the tension in the string? (b) What is the angular velocity of the bob?

(*Answer*: (a) 13.6 N, (b) 4.34 rad/sec.)

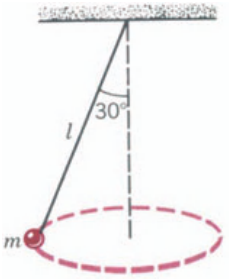


FIGURE 7-12
Problem 7.20.

7.21 A small particle of mass m is placed on top of a stationary, frictionless spherical ball of radius 0.5 m (see Fig. 7-13). It is given a slight kick to start sliding down. (a) Find the tangential velocity of the particle when it looses contact with the sphere. (b) What is the angle θ when contact is lost?

(*Answer*: (a) 1.81 m/sec, (b) 48.2°.)

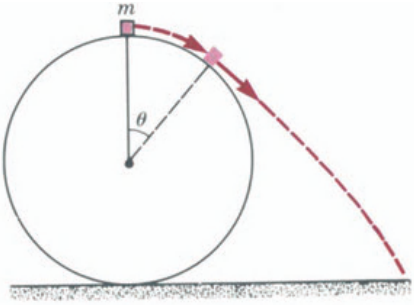


FIGURE 7-13
Problem 7.21.

CHAPTER 8

Rotational Dynamics

8.1 INTRODUCTION

In this chapter we introduce the concepts of rotational dynamics. Previously, we developed the first principles of linear dynamics; now we adapt the principles of linear dynamics to rotating bodies. The same laws apply, but their formulation is different. We will see, however, that Newton's laws, momentum, energy, and power all have equations equivalent to their linear counterparts.

8.2 MOMENT OF INERTIA AND TORQUE

In Newton's second law, mass is the proportionality constant between force and acceleration. Newton called it the *inertial mass*, that is, the resistance of a body to having its state of motion changed. We encounter a similar concept in rotational motion. Independent of friction, it is easier to spin a bicycle wheel on its axle than it is to so spin a car wheel. This resistance to having the state of rotational motion changed is called the *moment of inertia*, with symbol I. To demonstrate it in its simplest form, let us consider the rotation of a point mass m at one end of a rigid massless rod of length r. Let the other end of the rod be fastened to a point of rotation so that the system can rotate in the plane of the paper, as in Fig. 8-1. Suppose a force $\mathbf{F}$ is applied to the mass in the direction shown. We construct cartesian coordinates with the origin at m and x' as an extension of $\mathbf{r}$. The force $\mathbf{F}$ has two components obtained by constructing the indicated lines perpendicular to the x' and y' coordinate axes. The x' component $F_R = F \cos \phi$ is in the direction of $\mathbf{r}$. But because the rod is rigid, there can be no motion of the mass in the x' direction. The component in the y' direction is $F \sin \phi$. It should be observed that by construction this component is tangent to the circle of rotation at the point where m is located, so $F_T = F \sin \phi$. By Newton's second law, this tangential force causes a tangential acceleration

$$F_T = ma_T \tag{8.1}$$

From Eq. 7.7

$$a_T = r\alpha \tag{7.7}$$

Substituting for F_T and a_T in Eq. 8.1 yields

$$F \sin \phi = mr\alpha$$

Now multiply both sides of this equation by r

$$rF \sin \phi = mr^2\alpha \tag{8.2}$$

From Fig. 8-1, $r \sin \phi$ on the left side of Eq. 8.2 is equal to h, the perpendicular distance from the origin of the x-y coordinate system 0 to the line of F. Eq. 8.2 can therefore be rewritten as

$$Fh = mr^2\alpha \tag{8.3}$$

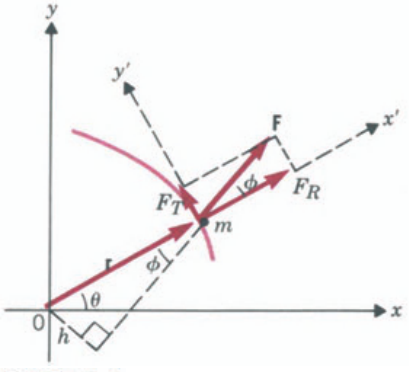


FIGURE 8-1

where

$$Fh = Fr \sin \phi$$

The quantity $Fh = Fr \sin \phi$, *the product of a force times the perpendicular distance from the point of rotation to the line of the force,* is called the *torque* produced by **F**, which is usually represented by τ, the small Greek letter *tau*. The quantity mr^2 on the right side of Eq. 8.2 is called the moment of inertia, I, of a point mass. We may therefore write Eq. 8.3 as

$$\tau = I\alpha \tag{8.4}$$

$\tau = I\alpha$

This is Newton's second law, which governs rotation. When compared with $\mathbf{F} = m\mathbf{a}$, we see that τ corresponds to force, I to mass, and α to linear acceleration. Just as in the linear case, τ must be the net torque and, if it is zero, there is no angular acceleration. Note that from the original definition, Eq. 7.6, the units of α are rad/sec^2.

It should be pointed out that if the force in Fig. 8-1 did not lie in the x-y plane but in some other plane while making the same angle with r, the torque would still have the same magnitude, $rF \sin \phi$, but the ensuing plane of rotation would not be the same. This ambiguity can be removed by assigning a direction to τ. It is conventional to define τ as the cross product of the position vector **r** and the force vector **F**, namely,

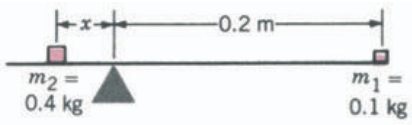
$$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F} \tag{8.5}$$

$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}$

From the definition of the cross product, Eq. 2.2, the magnitude of $\boldsymbol{\tau}$ is $rF \sin \phi$, which is the same value assigned previously. Moreover, the direction of $\boldsymbol{\tau}$ is the perpendicular to the two vectors being crossed, **r** and **F**, according to the right-hand rule discussed in Chapter 2. In the case illustrated by Fig. 8-1, $\boldsymbol{\tau}$ is perpendicular to the plane of the paper outward. Equation 8.5 defines $\boldsymbol{\tau}$ unambiguously.

Example 8-1

A balance scale consisting of a weightless rod has a mass of 0.1 kg on the right side 0.2 m from the pivot point. See Fig. 8-2. (a) How far from the pivot point on the left must 0.4 kg be placed so that a balance is achieved? (b) If the 0.4-kg mass is suddenly removed, what is the instantaneous rotational acceleration of the rod? (c) What is the instantaneous tangential acceleration of the 0.1-kg mass when the 0.4-kg mass is removed?



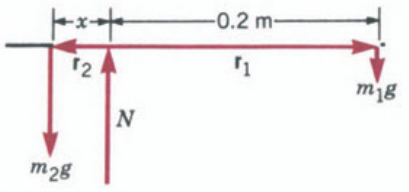


FIGURE 8-2
Example 8-1.

Solution

(a) When a balance is achieved $\alpha = 0$ and therefore

$$\Sigma\tau = 0$$

On the right of the pivot point the force is m_1g downward and the cross product $\mathbf{r} \times \mathbf{F}$ is into the paper or negative. On the left the

force is m_2g downward and the cross product $\mathbf{r} \times \mathbf{F}$ is out of the paper or positive.

$$(m_2g)(x)\sin 90° - (m_1g)(0.2 \text{ m})\sin 90° = 0$$

Solving for x

$$x = \frac{(m_1g)(0.2 \text{ m}) \sin 90°}{(m_2g) \sin 90°}$$

$$= \frac{(0.1 \text{ kg})(9.8 \text{ m/sec}^2)(0.2 \text{ m})}{(0.4 \text{ kg})(9.8 \text{ m/sec}^2)}$$

$$= 0.05 \text{ m}$$

(b) $\tau = I\alpha$

$$\alpha = \frac{\tau}{I} = \frac{(m_1g)(0.2 \text{ m}) \sin 90°}{(m_1)(0.2 \text{ m})^2}$$

$$\alpha = \frac{(0.1 \text{ kg})(9.8 \text{ m/sec}^2)(0.2 \text{ m}) \sin 90°}{(0.1 \text{ kg})(0.2 \text{ m})^2}$$

$$\alpha = 49 \text{ rad/sec}^2 \text{ clockwise}$$

(c) $a_T = r\alpha$

$$= (0.2 \text{ m})(49 \text{ rad/sec}^2)$$

$$= 9.8 \text{ m/sec}^2$$

As expected, the answer to part (c) is the acceleration of a body in free fall. The same answer will be obtained for the left-hand weight if the right-hand one is removed except that the rod will rotate counterclockwise.

In Eq. 8.3, mr^2 is the moment of inertia of a point mass at a distance r from the pivot point. If there are a variety of masses at different distances from the pivot point, the moment of inertia of the assembly is the sum of their individual ones or

$$I = \sum_{i=1}^{n} m_i r_i^2 \tag{8.6}$$

$$I = \sum_{i=1}^{n} m_i r_i^2$$

If all the masses are at the same distance r from the pivot point, $r_i^2 = r^2$ for all the terms in the sum and r^2 can be factored to obtain

$$I = r^2 \sum_{i=1}^{n} m_i \tag{8.7}$$

If we wish to find the moment of inertia of a thin hoop such as a bicycle wheel with essentially massless spokes, then the mass of the wheel M is

simply

$$M = \sum_{i=1}^{n} m_i$$

where m_i is the mass of each infinitesimal element. Therefore, the moment of inertia of a bicycle wheel is approximately

$$I = Mr^2$$

The value of I for spheres, cylinders and such must be either derived by Eq. 8.6 or looked up in tables. We should note that unlike the translational inertia (the mass), the rotational inertia (moment of inertia) of an object depends on the location of the mass relative to the axis of rotation and in general is different for different axes of rotation (see problems 8.5 and 8.6).

8.3 ROTATIONAL KINETIC ENERGY

In Fig. 8-1 the force **F** was divided into two orthogonal components. It is seen that because r is fixed, the component of **F** in the x' direction can do no work.

In an infinitesimally small time interval dt, the tangential component of **F**, $F_T = F \sin \phi$ causes the particle to move an infinitesimal displacement $d\mathbf{s}$, which from Eq. 7.2 is given by

$$d\mathbf{s} = \mathbf{v}_T \, dt \tag{8.8}$$

Because time is a scalar quantity, the direction of $d\mathbf{s}$ is the same as that of $\mathbf{v}_T$, namely, tangent to the path of the particle and therefore in the same direction as $\mathbf{F}_T$ (see Fig. 8-3). The work done by $\mathbf{F}_T$ in this infinitesimal distance is dW, and by the definition of Eq. 5.3 is

$$dW = \mathbf{F}_T \cdot d\mathbf{s} \tag{8.9}$$

But because $\mathbf{F}_T$ and $d\mathbf{s}$ are in the same direction, the dot product in Eq. 8.9 can be deleted.

$$dW = F_T \, ds \tag{8.10}$$

The tangential displacement of the particle is accompanied by an increase in the angle θ (see Fig. 8-3). The two are related by the differential form of Eq. 7.1

$$d\theta = \frac{ds}{r} \tag{7.1'}$$

Substituting $F \sin \phi$ for F_T and $r \, d\theta$ for ds into Eq. 8.10, we obtain

$$dW_\theta = F \sin \phi \, r \, d\theta \tag{8.11}$$

We have added the subscript θ to dW to indicate that the work is associated with an angular displacement $d\theta$.

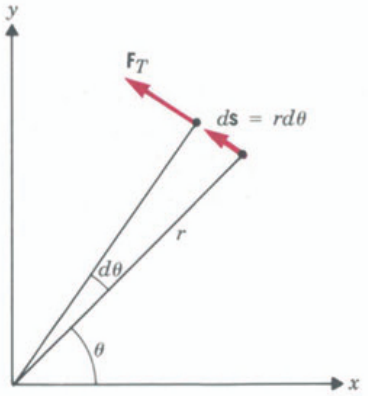


FIGURE 8-3

The product $F \sin \phi \, r$ in the right side of Eq. 8.11 may be recognized as the torque exerted by the force **F** (see Eq. 8.5), therefore

$$dW_\theta = \tau \, d\theta \tag{8.12}$$

To find the work done for a finite rotation we simply integrate Eq. 8.12

$$W_\theta = \int_{\theta_0}^{\theta_f} \tau \, d\theta \tag{8.13}$$

$$W_\theta = \int_{\theta_0}^{\theta_f} \tau \, d\theta$$

where θ_0 and θ_f are the initial and final angles, respectively. Equation 8.13 is equivalent to the expression found in Chapter 5 for the work done in translation

$$W = \int_a^b F \, dx \tag{5.6}$$

In Chapter 5 we indicated that the net work done on a body changes its velocity, or more precisely, its kinetic energy. This was known as the work-energy theorem (Eq. 5.9). We can show that the same occurs in the case of rotation, while at the same time we will find an expression for the kinetic energy of a rotating body in terms of rotational parameters.

Substituting $\tau = I\alpha$ (Eq. 8.4) into Eq. 8.13 yields

$$W_\theta = \int_{\theta_0}^{\theta_f} I\alpha \, d\theta \tag{8.14}$$

But, by definition, $\alpha = \dfrac{d\omega}{dt}$ and $d\theta = \omega \, dt$; therefore,

$$W_\theta = \int_{\theta_0}^{\theta_f} I \frac{d\omega}{dt} \omega \, dt$$

The time dt cancels out in the integral, and we get

$$W_\theta = \int_{\omega_0}^{\omega_f} I\omega \, d\omega \tag{8.15}$$

where we have changed the limits from θ_0 and θ_f to ω_0 (initial angular velocity) and ω_f (final angular velocity), for we now integrate with respect to ω. If the moment of inertia is constant—that is, if the distance of the particle to the point of rotation does not change—then Eq. 8.15 becomes

$$W_\theta = I \int_{\omega_0}^{\omega_f} \omega \, d\omega$$

$$W_\theta = \frac{1}{2} I\omega_f^2 - \frac{1}{2} I\omega_0^2 \tag{8.16}$$

Comparing this with the work-energy theorem of Chapter 5, where we saw that work done is equal to the change in kinetic energy, we may write

$$(\Delta E_k)_{\text{rot}} = \frac{1}{2} I\omega_f^2 - \frac{1}{2} I\omega_0^2 \tag{8.17}$$

where the quantity $\frac{1}{2} I\omega^2$ is called the *rotational kinetic energy*, $(E_k)_{\text{rot}}$.

We may gain insight into the physical significance of $(E_k)_{\text{rot}} = 1/2\ I\omega^2$ by relating it to the linear kinetic energy of the particle. A point on a rotating system has an instantaneous tangential velocity v_T. Its kinetic energy is therefore

$$E_k = \frac{1}{2} mv_T^2$$

But, because $v_T = r\omega$, the kinetic energy may be written as

$$E_k = \frac{1}{2} mr^2\omega^2$$

The moment of inertia of a point mass is $I = mr^2$. Therefore,

$$E_k = \frac{1}{2} I\omega^2$$

Hence, the expression for the rotational kinetic energy in terms of I and ω is simply another form of the kinetic energy of a particle rotating about a fixed axis.

The expression $\frac{1}{2}I\omega^2$ for the rotational kinetic energy can be readily shown to be applicable to the rotation of a rigid body made up of discrete masses m_i. The rotational kinetic energy of the ith particle is

$$(E_k)_{\text{rot}} \text{ of } i\text{th particle} = \frac{1}{2} m_i r_i^2 \omega_i^2 \tag{8.18}$$

and the $(E_k)_{\text{rot}}$ of the body is the sum of $(E_k)_{\text{rot}}$ of the individual masses

$$(E_k)_{\text{rot}} = \frac{1}{2} \sum_{i=1}^{n} m_i r_i^2 \omega_i^2$$

Because the body is rigid, all point masses rotate with the same angular velocity regardless of their distance from the axis, so $\omega_i^2 = \omega^2$ and it can be factored out of the sum. We then have

$$(E_k)_{\text{rot}} = \frac{1}{2} \omega^2 \sum_{i=1}^{n} m_i r_i^2$$

But the quantity in the summation is the definition of the moment of inertia I for a system of particles (see Eq. 8.6) and therefore

$$(E_k)_{\text{rot}} = \frac{1}{2} I\omega^2 \tag{8.19}$$

$(E_k)_{\text{rot}} = \frac{1}{2} I\omega^2$

A body can be rotating as it translates through space; for example, the earth rotates about its axis as its center of mass moves about the sun. Clearly, the earth's rotation gives it more kinetic energy than if it were moving without rotation. Its total kinetic energy is therefore the sum of translational and rotational kinetic energies

$$(E_k)_{\text{total}} = (E_k)_{\text{trans}} + (E_k)_{\text{rot}} = \frac{1}{2} m v_{CM}^2 + \frac{1}{2} I\omega^2$$

$$(E_k)_{\text{total}} = \frac{1}{2} m v_{CM}^2 + \frac{1}{2} I\omega^2$$

where v_{CM} is the translational velocity of the center of mass.

Example 8-2

A large wheel of radius 0.4 m and moment of inertia 1.2 kg-m^2, pivoted at the center, is free to rotate without friction. A rope is wound around it and a 2-kg weight is attached to the rope (see Fig. 8-4). When the weight has descended 1.5 m from its starting position (a) what is its downward velocity? (b) what is the rotational velocity of the wheel?

Solution

(a) We may solve this problem by the conservation of energy, equating the initial potential energy of the weight to its conversion to kinetic energy of the weight and of the wheel.

$$mgh = \frac{1}{2} m v^2 + \frac{1}{2} I\omega^2$$

The downward velocity v of the weight is equal to the tangential velocity at the rim of the wheel v_T; therefore

$$\omega = \frac{v_T}{r} = \frac{v}{r}$$

Substituting for ω

$$mgh = \frac{1}{2} m v^2 + \frac{1}{2} I \frac{v^2}{r^2}$$

We solve for the velocity v

$$v = \left[\frac{mgh}{\frac{1}{2} m + \frac{I}{2r^2}} \right]^{1/2}$$

$$= \left[\frac{(2 \text{ kg})(9.8 \text{ m/sec}^2)(1.5 \text{ m})}{\left(\frac{1}{2}\right)(2 \text{ kg}) + \frac{(1.2 \text{ kg-m}^2)}{(2)(0.4 \text{ m})^2}} \right]^{1/2}$$

$$v = 2.5 \text{ m/sec}$$

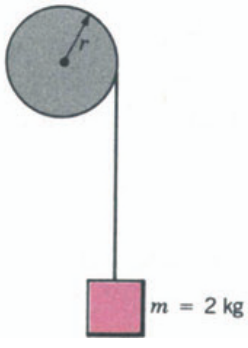


FIGURE 8-4
Example 8-2.

(b) The answer to part (a) shows that any point on the rim of the wheel has a tangential velocity of $v_T = 2.5$ m/sec. We convert this to rotational velocity of the wheel

$$\omega = \frac{v_T}{r} = \frac{2.5 \text{ m/sec}}{0.4 \text{ m}} = 6.2 \text{ rad/sec}$$

8.4 POWER

The definition of *power* is work done per unit time. The incremental amount of work done in moving the mass in Fig. 8-3 a distance $ds = r\,d\theta$ is given in Eq. 8.12.

$$dW_\theta = \tau\,d\theta \tag{8.12}$$

But, from Eq. 5.15

$$\text{Power} = \frac{dW}{dt}$$

Substitute Eq. 8.12 to obtain

$$\text{Power} = \frac{\tau\,d\theta}{dt}$$

or, because $\omega = d\theta/dt$,

$$\text{Power} = \tau\omega \tag{8.20}$$

Power = $\tau\omega$

Example 8-3

A machine shop has a lathe wheel of 40-cm diameter driven by a belt that goes around the rim. If the linear speed of the belt is 2 m/sec and the wheel requires a tangential force of 50 N to turn it, how much power is required to operate the lathe?

Solution Use Eq. 8.20

$$\text{Power} = \tau\omega$$

If there is no slipping between the belt and the wheel, the linear speed of the belt is equal to the tangential velocity at the rim of the wheel and the rotational velocity is therefore

$$\omega = \frac{v_T}{r}$$

$$= \frac{2 \text{ m/sec}}{0.2 \text{ m}}$$

$$= 10 \text{ rad/sec}$$

From Eq. 8.5 the torque is

$$\begin{aligned}\tau &= rF \sin \phi \\ &= (0.2 \text{ m})(50 \text{ N}) \sin 90° \\ &= 10 \text{ Nm}\end{aligned}$$

then

$$\begin{aligned}\text{Power} &= 10 \text{ Nm} \times 10 \text{ rad/sec} \\ &= 100 \text{ W} \left(\frac{1 \text{ hp}}{746 \text{ W}}\right) = 0.13 \text{ hp}\end{aligned}$$

8.5 ANGULAR MOMENTUM

We learned that an important property of a particle or of a system of particles (a body) is its momentum $\mathbf{p} = m\mathbf{v}$. An equivalent property can be associated with a rotating body.

Consider, as shown in Fig. 8-5, a particle of mass m with momentum $\mathbf{p} = m\mathbf{v}$ in the x-y plane. The position vector of m is $\mathbf{r}$, which is not required to be a constant.

From Newton's second law we write

$$\mathbf{F} = \frac{d}{dt}(m\mathbf{v}) \tag{4.2}$$

If we take the cross product of both sides with the position vector $\mathbf{r}$, we obtain

$$\mathbf{r} \times \mathbf{F} = \mathbf{r} \times \frac{d}{dt}(m\mathbf{v}) \tag{8.21}$$

By definition $\mathbf{r} \times \mathbf{F} = \boldsymbol{\tau}$ and therefore

$$\boldsymbol{\tau} = \mathbf{r} \times \frac{dt}{dt}(m\mathbf{v}) \tag{8.22}$$

The right side of Eq. 8.22 can be rewritten as

$$\mathbf{r} \times \frac{d}{dt}(m\mathbf{v}) = \frac{d}{dt}(\mathbf{r} \times m\mathbf{v}) \tag{8.23}$$

The equivalence of the two expressions becomes evident if we differentiate the second term

$$\frac{d}{dt}(\mathbf{r} \times m\mathbf{v}) = \frac{d\mathbf{r}}{dt} \times m\mathbf{v} + \mathbf{r} \times \frac{d}{dt}(m\mathbf{v})$$

But by definition $d\mathbf{r}/dt = \mathbf{v}$, the instantaneous velocity of the particle; therefore

$$\frac{d}{dt}(\mathbf{r} \times m\mathbf{v}) = \mathbf{v} \times m\mathbf{v} + \mathbf{r} \times \frac{d}{dt}(m\mathbf{v})$$

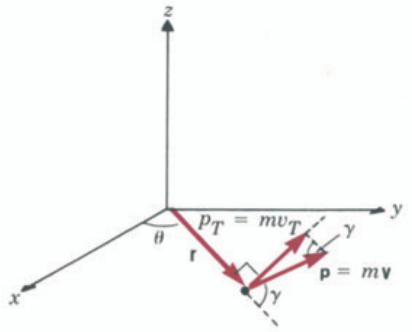


FIGURE 8-5

We saw in Chapter 2 that the cross product of two vectors in the same direction is zero. Therefore

$$\mathbf{v} \times m\mathbf{v} = m(\mathbf{v} \times \mathbf{v}) = 0$$

and, therefore

$$\frac{d}{dt}(\mathbf{r} \times m\mathbf{v}) = \mathbf{r} \times \frac{d}{dt}(m\mathbf{v})$$

Substituting Eq. 8.23 for $\mathbf{r} \times \frac{d}{dt}(m\mathbf{v})$ in Eq. 8.22 yields

$$\boldsymbol{\tau} = \frac{d}{dt}(\mathbf{r} \times m\mathbf{v}) \tag{8.24}$$

$\boldsymbol{\tau} = \frac{d}{dt}(\mathbf{r} \times m\mathbf{v})$

We have already indicated that the torque $\boldsymbol{\tau}$ in rotational motion plays the role of the force $\mathbf{F}$ in translational motion. Thus, if we compare Eq. 8.24 with Eq. 4.2, we are led to the conclusion that the quantity $\mathbf{r} \times m\mathbf{v}$ in rotational motion plays the same role as does the momentum $m\mathbf{v}$ in translation. We therefore call

$$\mathbf{L} = \mathbf{r} \times m\mathbf{v} \tag{8.25}$$

$\mathbf{L} = \mathbf{r} \times m\mathbf{v}$

the *angular momentum* of the particle.

We can find another expression for $\mathbf{L}$ that shows even more clearly its correspondence to the momentum $m\mathbf{v}$.

$$\mathbf{L} = \mathbf{r} \times m\mathbf{v} = rmv \sin \gamma$$

where γ is the angle between the radius vector $\mathbf{r}$ and the linear momentum $m\mathbf{v}$ (see Fig. 8-5). But, from Fig. 8-5, $mv \sin \gamma = mv_T$, and

$$L = rmv_T$$

From Eq. 7.4, $v_T = r\omega$ and therefore

$$L = mr^2\omega \tag{8.26}$$

We have defined mr^2 as the moment of inertia I of a point mass; hence

$$L = I\omega \tag{8.27}$$

$L = I\omega$

The equation of motion for rotation, Eq. 8.24, can be written

$$\tau = \frac{dL}{dt} \quad \text{or} \quad \tau = \frac{d(I\omega)}{dt} \tag{8.28}$$

8.6 CONSERVATION OF ANGULAR MOMENTUM

We will now show that the law of conservation of momentum that was derived in Chapter 6 applies equally to angular momentum. Start with Eq. 8.28

$$\tau = \frac{d(I\omega)}{dt} = \frac{dL}{dt}$$

If we have a situation in which there is no net externally applied torque, then $\tau = 0$. Thus

$$\frac{dL}{dt} = 0$$

and L = constant. Hence, $I\omega$ = constant.

Therefore, with no net external torque

$$(I\omega)_0 = (I\omega)_f \tag{8.29}$$

$(I\omega)_0 = (I\omega)_f$

This is known as the law of *conservation of angular momentum.*

Due to the small torque exerted by the ice, the angular momentum of a spinning skater is almost constant. As a result, when the skater pulls her arms inward, thus reducing its moment of inertia, her angular velocity increases.

Example 8-4

Suppose the body of an ice skater has a moment of inertia $I = 4$ kg-m^2 and her arms have a mass of 5 kg each with the center of mass at 0.4 m from her body. She starts to turn at 0.5 rev/sec on the point of her skate with her arms outstretched. She then pulls her arms inward so that their center of mass is at the axis of her body, $r = 0$. What will be her speed of rotation?

Solution

$$I_0\omega_0 = I_f\omega_f$$

$$(I_{\text{body}} + I_{\text{arms}})\,\omega_0 = I_{\text{body}}\,\omega_f$$

$$(I_{\text{body}} + 2mr^2)\,\omega_0 = I_{\text{body}}\,\omega_f$$

Solving for ω_f

$$\omega_f = \frac{(I_{\text{body}} + 2mr^2)\,\omega_0}{I_{\text{body}}} = \frac{[4 \text{ kg-m}^2 + 2 \times 5 \text{ kg} \times (0.4 \text{ m})^2]\,(0.5 \text{ rev/sec})}{4 \text{ kg-m}^2}$$

$$= 0.7 \text{ rev/sec}$$

PROBLEMS

8.1 A bicycle wheel of mass 2 kg and radius 0.32 m is spinning freely on its axle at 2 rev/sec. When you place your hand against the tire the wheel decelerates uniformly and comes to a stop in 8 sec. What was the torque of your hand against the wheel?

8.2 Two masses, $m_1 = 1$ kg and $m_2 = 5$ kg, are connected by a rigid rod of negligible weight (see Fig. 8-6). The system is pivoted about point 0. The gravitational forces act in the negative z direction. (a) Express the position vectors and the forces on the masses in terms of unit vectors and calculate the torque on the system. (b) What is the angular acceleration of the system at the instant shown in Fig. 8-6?

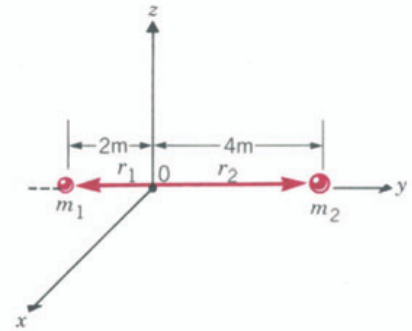


FIGURE 8-6
Problem 8.2.

8.3 A roulette wheel with $I = 0.5$ kg-m^2 rotating initially at 2 rev/sec coasts to a stop from the constant friction torque of the bearing. If the torque is 0.4 N-m, how long does it take to stop?

(*Answer*: 15.7 sec.)

8.4 A grindstone with $I = 240$ kg-m^2 rotates with a speed of 1 rev/sec. A knife blade is pressed against it, and the wheel coasts to a stop with constant deceleration in 12 sec. What torque did the knife exert on the wheel?

8.5 Four identical masses ($m = 2$ kg) are connected by rods of negligible weight to form a rectangle (see Fig. 8-7). The masses are rotated about an axis perpendicular to the plane of the rectangle and passing through its center with an angular acceleration $\alpha = 3$ rev/sec^2. What torque is needed?

(*Answer*: 1282 N-m.)

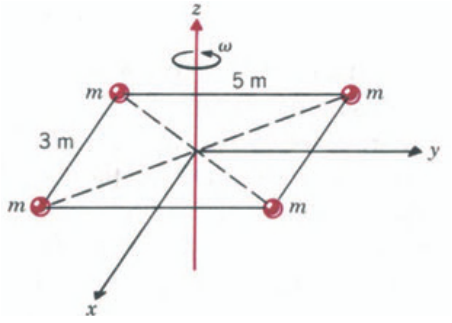


FIGURE 8-7
Problem 8.5.

8.6 Repeat problem 8.5 for a rotation, with the same angular acceleration, about an axis through a corner of the rectangle (see Fig. 8-8).

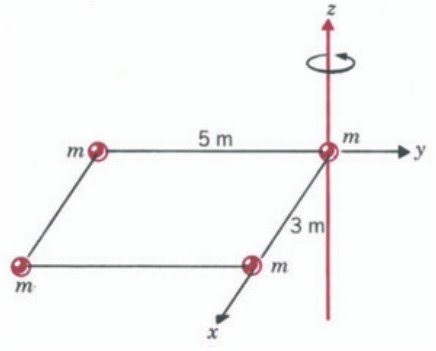


FIGURE 8-8
Problem 8.6.

8.7 A uniform wooden board of mass 20 kg rests on two supports as shown in Fig. 8-9. A 30-kg steel block is placed to the right of support A. How far to the right of A can the steel block be placed without tipping the board?

(*Answer*: 2.0 m.)

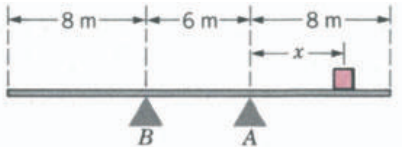


FIGURE 8-9
Problem 8.7.

8.8 A wheel ($I = 30$ kg-m^2) is pivoted about an axis through its center. A 90 N-m torque is applied to the wheel, which then accelerates from rest to an angular velocity of 20 rad/sec in 10 sec. (a) What is the friction torque of the bearings? (b) If the applied torque is removed after 60 sec, how long will it take for the wheel to come to rest?

8.9 Calculate the change in rotational kinetic energy of the roulette wheel and grindstone of problems 8-3 and 8-4.

8.10 A ball of mass 0.3 kg and radius 0.1 m rolls along the ground with a transverse speed of 4 m/sec. It comes to a slope inclined at 30°. How far up the slope does it roll? (I (ball) $= 2/5\ mr^2$).

(*Answer*: 2.29 m.)

8.11 A wheel of moment of inertia 60 kg-m^2 and radius 1.5 m is rotating about an axis through its center with an angular velocity $\omega = 30$ rev/sec. A brake is applied producing a normal force of 450 N against the rim (see Fig. 8-10). The coefficient of friction between the wheel and the brake is 0.5. (a) How long will it take for the wheel to stop? (b) Calculate the work done by friction, and show that this is equal to the change in the kinetic energy of the wheel.

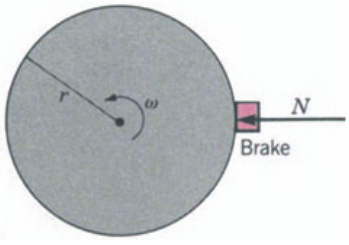


FIGURE 8-10
Problem 8.11.

8.12 Although most people are concerned about the horsepower of a car's motor, the important parameter is the amount of torque that can be given to the rear wheels. The torque of the motor is turned at right angles to the wheels by the differential gear. Assume that in low gear the angular velocity of the rear wheels is 0.1 that of the motor. If the motor has

200 hp and is turning over at a rate of 1400 rev/min, how much torque is delivered to the rear wheels?

(*Answer*: 1.02×10^4 N-m.)

8.13 A 4-kg block is attached to one end of a light rope. The other end of the rope is wrapped around a pulley of moment of inertia $I = 0.5$ kg-m^2 and radius $r = 0.2$ m (see Fig. 8-11). The block is released from rest, and it moves down 9 m in 3 sec. (a) What is the friction torque of the bearings? (b) Use energy principles to calculate the velocity of the block after it has fallen 9 m.

(*Answer*: (a) 1.24 N-m, (b) 6.0 m/sec.)

FIGURE 8-11
Problem 8.13.

8.14 A 2-kg block resting on a frictionless table is connected by a string passing over a pulley to a second block, $m_2 = 5$ kg, hanging over the edge of the table 0.8 m above the floor (see Fig. 8-12). The moment of inertia of the pulley is 0.8 kg-m^2 and the radius is 0.1 m. Neglect the friction of the bearings and assume that there is no slipping between the string and the pulley. Use energy methods to calculate the velocity of m_2 as it hits the floor.

(*Answer*: 0.95 m/sec.)

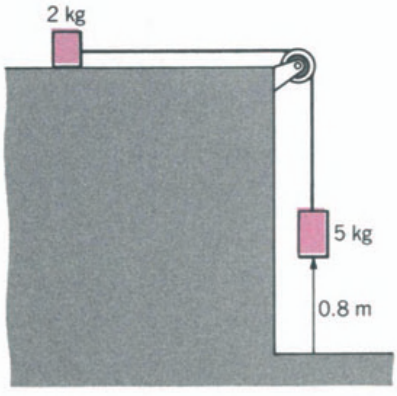


FIGURE 8-12
Problem 8.14.

8.15 Repeat problem 8.14 if the coefficient of friction between the 2-kg block and the table is 0.25.

8.16 Use the data in problem 7.14 to find the angular momentum of an electron in the smallest orbit of the hydrogen atom.

(*Answer*: 1.05×10^{-34} J-sec.)

8.17 A 50-gm mouse falls onto the outer edge of a phonograph turntable of radius 20 cm rotating at 33 rev/min. How much work must it do to walk into the center post? Assume that the angular velocity of the turntable does not change.

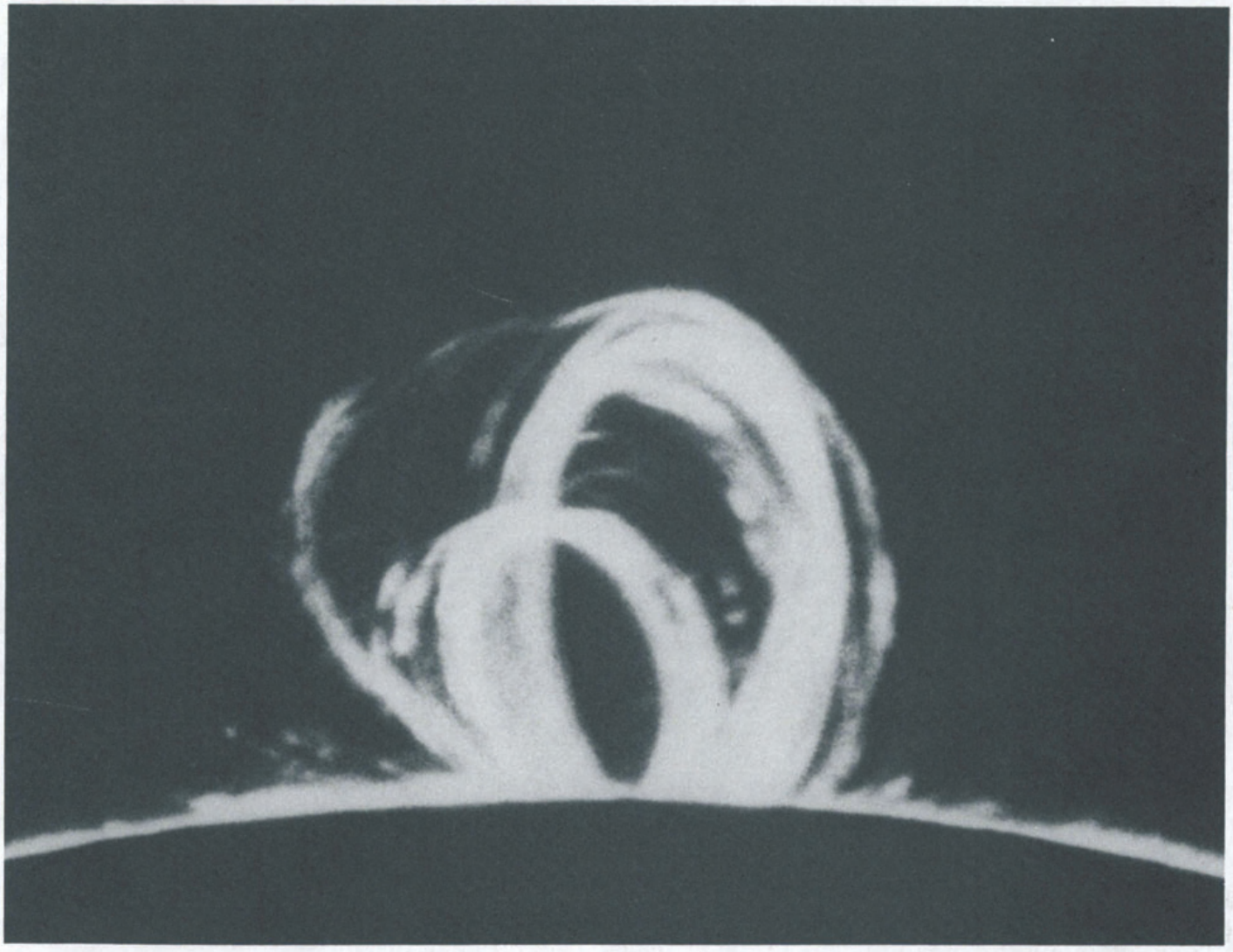
8.18 A children's merry-go-round of radius 4 m and mass 100 kg has an 80-kg man standing at the rim. The merry-go-round coasts on a frictionless bearing at 0.2 rev/sec. The man walks inward 2 m toward the center. What is the new rotational speed of the merry-go-round? What is the source of this energy? (The moment of inertia of a solid disk is $I = 1/2\ mr^2$).

(*Answer*: 0.37 rev/sec.)

8.19 A mass of 0.1 kg on a string is rotating on a frictionless table with $\omega = 1$ rev/sec and $r = 0.2$ m. The string passes through a hole in the table and is held by a hand below (see Fig. 8-13). (a) What is the angular momentum of the mass? (b) What is the kinetic energy of the mass? (c) If the string is pulled down by the hand until $r = 0.1$ m, what is the new rotational speed of the mass? (d) What is the new kinetic energy? (e) How much work did the hand do in pulling the string?

(*Answer*: (a) 2.51×10^{-2} J-sec, (b) 7.9×10^{-2} J, (c) 4 rev/sec, (d) 3.16×10^{-1} J, (e) 2.37×10^{-1} J

FIGURE 8-13
Problem 8.19.



CHAPTER 9

Kinetic Theory of Gases and the Concept of Temperature

9.1 INTRODUCTION

Heat and temperature proved to be very elusive concepts to early scientists. Most historic theories—for examples, the phlogiston and the caloric—assumed that heat was a substance that could flow, much as a gas or a fluid. In fact, the mathematics of heat flow were correctly worked out before scientists learned the true nature of heat and its associated property, temperature. For indeed "heat" does flow; but what is heat?

Our modern understanding of heat, temperature, and the behavior of gases is the result of two and a half centuries of scientific investigation; we now know that heat is a form of energy. We will not tell the entire story, because it is beyond the scope of this book. We will, however, trace the story through the measurement of temperature, the ideal gas law, and then the application of the first principles of mechanics, as developed in the earlier chapters, to the average motion of molecules in gases. The identification of the average kinetic energy of molecules with temperature will then be shown. From that we will be able to write the first law of thermodynamics, which is a broadened statement of the law of conservation of energy.

9.2 MOLECULAR WEIGHT

In real systems there are vast numbers of atoms, all of which obey the first principles of physics, or quantum variations of them, in their motion. The motion of each is different however, so a conclusion about the behavior of a group of atoms is statistical. In this book we will consider only systems composed of identical atoms, or molecules such as oxygen molecules (O_2) or nitrogen molecules (N_2).

Ensembles of different atoms or molecules have different statistical averages of their properties. This is because if objects that have the same kinetic energies, $\frac{1}{2}mv^2$, have different masses, then their velocities are different. It is therefore important to know the mass of the atoms or molecules that make up the ensemble. It is clear that if we know the mass of the ensemble and the weight of each particle of the ensemble, then we can immediately determine the number of particles.

We use a unit called the *mole* (abbr. *mol*) as a measure of the number of particles with the following definition. *A mole of a substance is that quantity which contains the same number of particles as there are atoms in 12 g* (12×10^{-3} *kg) of carbon-12.* The measure in SI units is the *kilomole* (*kmol*), which is the quantity of the substance that contains the same number of particles as there are atoms in 12 kg of carbon-12. All the elements have *isotopes*, that is, atoms with the same chemical properties but with a slightly different mass. Therefore, a single isotope of carbon (carbon-12) is chosen as the reference standard. This standard is said to have a mass of exactly 12 u per atom, where *u* is called an *atomic mass unit* and has the value

$$1 \text{ u} = 1.66057 \times 10^{-27} \text{ kg}$$

Amedo Avogadro (1776–1856).

The mass of an atom (or molecule) in atomic mass units is called the *atomic weight* (or *molecular weight*). Thus, for example, the atomic weight of carbon-12 is 12 u, that of hydrogen-1 is 1.0078 u. The mass, in grams, of a mole of a substance is *numerically* equal to the atomic weight (or molecular weight) of the atoms of that substance, and it is referred to as the *gram atomic weight* (or *gram molecular weight*). The mass of 1 mole of carbon-12 is 12 g/mole, that of hydrogen-1 is 1.0078 g/mole. Often, the word "gram" is deleted from the expressions for the mass of a mole, which may lead to confusion. We can rely on the units to see whether we are dealing with the mass of an atom or that of a mole.

By the use of very careful techniques, chemists and physicists have been able to measure the number of atoms in 1 mole of a substance. This is called *Avogadro's number* and has the value

$$N_A = 6.022 \times 10^{23} \text{ atoms/mol} = 6.022 \times 10^{26} \text{ atoms/kmol}$$

Therefore, if we know the number of moles of substance present in the ensemble, we can calculate the number of atoms or molecules present.

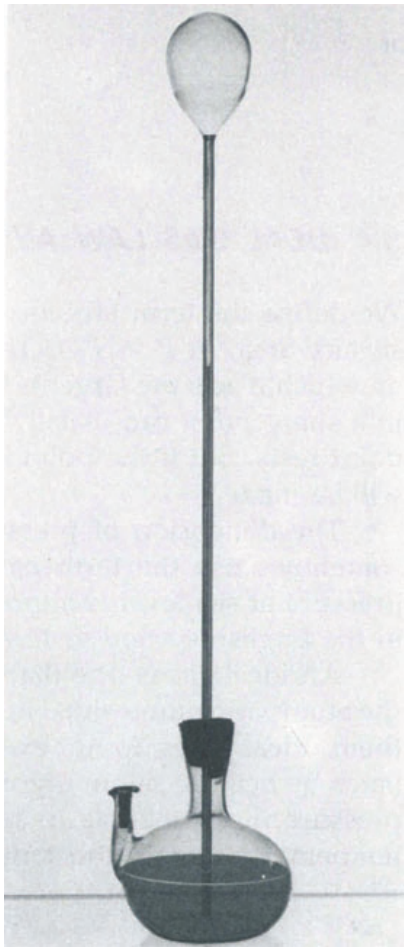
We will use the symbol n to represent the number of moles present, where n may be greater or less than one, or equal to it. Because one usually has less than 1 mole, n has the name *mole fraction*. If we denote M as the gram molecular weight of a substance and m as the mass of the amount present, then $n = m/M$ and, because M has N_A atoms or molecules, the number of atoms or molecules present is nN_A.

9.3 THERMOMETERS

It has been known from ancient times that solids and liquids expand when they are heated. It is not known when the first thermometers were made, but they are believed to have been brought into general use by Duke Ferdinand of Tuscany in 1654. They were generally used shortly thereafter by members of the Academy of Science of Florence (which was founded by him) and were long known as Florentine thermometers. They were much like modern thermometers in that they had a colored liquid, presumably alcohol, hermetically sealed in a tube with a bulb at one end, with little pieces of colored glass to mark even divisions on the scale.

In 1714 Gabriel Fahrenheit proposed that a scale be established in which the temperature of the human body be taken as 100° (which has since been corrected to 98.6°) and 0° be the lowest temperature attainable with a mixture of ice and salt, sodium chloride (NaCl). Using this scale, the melting point of pure ice is 32° and the boiling point of pure water at sea level is 212°. Shortly after Fahrenheit's death in 1736 a different scale, Centigrade or Celsius, came into use; by this scale, the melting point of ice was taken at 0°C and the boiling point of water at 100°C.

The Fahrenheit scale is still popularly used in the United States and Canada, probably because of its finer divisions for meteorologic measure-



One of the earliest thermometers used, was Galileo's thermoscope shown here.

ments, but in scientific laboratories and in most of the world the Celsius scale is used. The conversion between the two has ever since been confusing to the layperson, but, with a little thought, one can eliminate the difficulty. The conversion is easy to see if we construct yet a third scale, which we will call the °F-32 scale, in Fig. 9-1. If we take an arbitrary temperature point on the °C scale and the same point on the °F-32 scale, we may make a ratio between the temperature of °C and °F-32 scales as

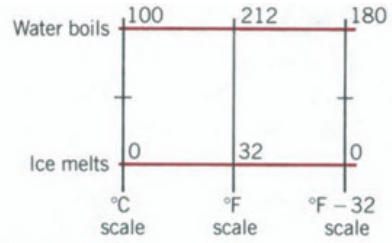


FIGURE 9-1

$$\frac{°C}{°F - 32} = \frac{100}{180}$$

or

$$\frac{°C}{°F - 32} = \frac{5}{9}$$

Thus

$$°C = \frac{5}{9}(°F - 32) \qquad (9.1)$$

$$°C = \frac{5}{9}(°F - 32)$$

or

$$°F = \frac{9}{5}°C + 32 \qquad (9.2)$$

$$°F = \frac{9}{5}°C + 32$$

9.4 IDEAL GAS LAW AND ABSOLUTE TEMPERATURE

We define the term *pressure*, P, as force perpendicular to a surface per unit surface area, or $P = F/A$. Therefore, for a given force, the smaller the area on which it acts the larger is the pressure. For example, a weight terminating in a sharp point can usually make an indentation in a surface on which the point rests. But if the point is changed to a flat face of larger area, no mark will be made.

The dimension of pressure is newton/meter^2 (N/m^2), and in fluids we sometimes use the term pascal (Pa) where 1 Pa = 1 N/m^2. Atmospheric pressure at sea level is approximately 1.01×10^5 N/m^2, which is equivalent in the English system to 14.7 lb/in^2.

An ideal gas is one that has no tendency to condense. This means that the atoms are infinitesimal in size and that there is no attractive force between them. Ideal gases do not exist, but they may be approximated by rare gases (such as helium, neon, argon) at low pressure, or any other gas at *very* low pressure. Robert Boyle in 1662 showed that if the quantity of gas and its temperature remain constant, then the pressure and volume vary inversely.

$$P = \frac{C}{V} \qquad (9.3)$$

where C is a constant. In 1802, Joseph Gay-Lussac showed that if the quantity of gas and its volume remained constant, the pressure is proportional to the

Robert Boyle (1627–1691).

temperature or

$$P = KT \tag{9.4}$$

where K is a constant.

These two laws may be combined into a single law with a single constant:

$$PV = R'T \tag{9.5}$$

When equal volumes of the same gas are taken at the same temperature and pressure, R' remains constant. If, however, equal masses of different gases are taken under the same conditions, R' varies inversely with the molecular weight.

$$PV = nRT$$

In the middle of the last century low temperatures were achieved in Lord Kelvin's laboratory in England. The ideal gas law, Eq. 9.5, was examined over an extended range of temperatures. It seemed desirable to establish R' as a constant and to correct its inverse variation with the mass of the gas used by multiplying by the number of moles, n (because $n = m/M$). With the introduction of this term n the ideal gas law is writen as

$$K = 273.16° + °C$$

$$PV = nRT \tag{9.6}$$

where n is the number of moles (mole fraction) and R is now the same for all gases.

When Kelvin's group examined Eq. 9.6 at constant volume for different amounts of a gas n_1, n_2, n_3, . . ., the data appeared as in Fig. 9-2. It is seen in this graph that the data taken to the lowest achievable temperature T_L all lie on straight lines and, if these lines are extrapolated to $P = 0$, they terminate at a common point, $-273.16°$ C. This was called *absolute zero*, and is the lowest possible temperature. It was therefore logical to establish a new temperature scale with its zero point at $-273.16°$ C. Thus, 0°C = +273.16 K, where K is the symbol for the new scale, called the Kelvin or *absolute* scale. It is related to the Celsius scale as

$$\text{K} = 273.16° + °\text{C} \tag{9.7}$$

With data of the type shown in Fig. 9-2, the gas constant R was evaluated by Kelvin and his associates as

$$R = 8314 \text{ J/kmol-K}$$

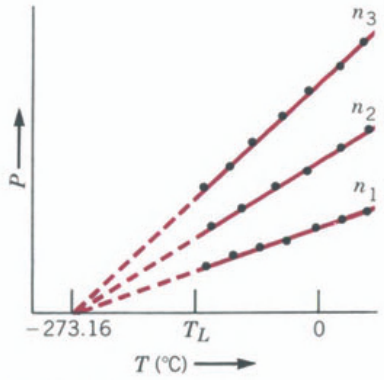


FIGURE 9-2
Results of Lord Kelvin's experiments showing a linear variation of the pressure of a gas with temperature when the volume is kept constant. The three curves correspond to different amounts of gas. When the curves are extrapolated, they terminate at a common point $T = -273.16$°C.

Example 9-1

What is the temperature of absolute zero on the Fahrenheit scale?

Solution Absolute zero = $-273.16°$ C. From Eq. 9.2

$$°\text{F} = \frac{9}{5}°\text{C} + 32$$

$$°\text{F} = \frac{9}{5} \times (-273.16) + 32 = -459.7° \text{ F}$$

Example 9-2

In a typical experiment to determine the value of the gas constant R, 0.152 g of neon gas (atomic weight 20.2 g/mole) is introduced into a 100-cm^3 flask that is closed and attached to a pressure gauge. It is found that when the flask is placed in a constant temperature bath at 50° C the pressure of the gas is 2 atmospheres (atm). What value of R is obtained?

Solution The mole fraction n is the ratio of the number of grams present to the atomic weight in grams.

$$n = \frac{0.152 \text{ g}}{20.2 \text{ g/mole}} = 7.52 \times 10^{-3} \text{ mol} = 7.52 \times 10^{-6} \text{ kmol}$$

The volume is

$$V = 10^2 \text{ cm}^3 \left(\frac{1 \text{ m}}{10^2 \text{ cm}}\right)^3 = 10^{-4} \text{ m}^3$$

The pressure is

$$P = 2 \text{ atm} \left(\frac{1.01 \times 10^5 \text{ N/m}^2}{1 \text{ atm}}\right) = 2.02 \times 10^5 \text{ N/m}^2$$

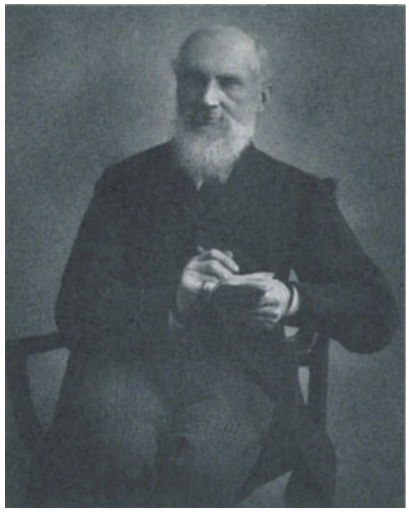
The temperature in K is $T = 273° + 50° = 323$ K.

The ideal gas law $PV = nRT$ is written as

$$R = \frac{PV}{nT}$$

$$= \frac{2.02 \times 10^5 \text{ N/m}^2 \times 10^{-4} \text{ m}^3}{7.52 \times 10^{-6} \text{ kmol} \times 323 \text{ K}}$$

$$= 8316 \text{ J/kmol-K}$$



William Thompson Kelvin (1824–1907).

Example 9-3

In a diesel engine, no spark plug is required because the temperature is raised to the ignition point of the air-fuel mixture by compression. In a typical diesel engine the air intake is at 27° C and at a pressure of 1 atm, and it is compressed to 1/15 of its original volume with its pressure becoming about 50 atm. What is the temperature of the air-fuel mixture in the cylinder in °C?

Solution We note from the ideal gas law that

$$\frac{PV}{T} = nR$$

and, if the quantity of gas is kept constant, then the right side of the equation is a constant. Therefore, if we change any of the quantities on the left side, the other quantities must change to yield the same constant, nR. This means

that the initial conditions of the left side must equal the final conditions because both the initial and the final conditions are equal to the same constant, nR. We write this as

$$\frac{P_0V_0}{T_0} = \frac{P_fV_f}{T_f}$$

where T must be in K.

In this problem

$$T_f = \frac{P_f}{P_0}\frac{V_f}{V_0}T_0$$

$$= \frac{50 \text{ atm}}{1 \text{ atm}}\frac{V_0/15 \text{ m}^3}{V_0\text{m}^3} \times 300 \text{ K}$$

$$= 1000 \text{ K} = 727^\circ \text{ C}$$

9.5 KINETIC THEORY OF GAS PRESSURE

We will now show how the concept of momentum conservation and the definition of pressure can be used to calculate the statistical behavior of a large number of atoms or molecules in a gas. One of the assumptions in this calculation is that all collisions between atoms or molecules are perfectly elastic. This is not strictly true at high temperatures, because in some high-energy collisions electrons are excited or even knocked off atoms. Although this situation can be dealt with theoretically, it will not concern us here.

Because the walls of a container are also made of atoms, then all collisions between the atoms of a gas in a container and the walls are elastic. One other fact must be kept in mind. In an elastic collision of an atom with the container wall, the velocity component normal to the wall is reversed on collision with its magnitude unchanged, and the velocity component in the direction parallel to the surface of the wall is unchanged. This can be seen in the two-dimensional schematic of Fig. 9-3. Viewed from above, it is clear that a v_z velocity component would also be unchanged.

We recall from Chapter 4 that an impulse acting on a body is equal to the change in the momentum of the body. For the situation of Fig. 9-3,

$$\bar{F}_x \, \Delta t = \Delta \, (mv_x)$$

where $\bar{F}_x$ is the average force exerted by the wall of the container on the atom during the time interval Δt, and where m is the mass of the atom. From Newton's third law of action and reaction, the magnitude of the force exerted by the atom on the wall is equal to $\bar{F}_x$ and can be written as

$$\bar{F}_x = \frac{mv_{x \text{ final}} - mv_{x \text{ initial}}}{\Delta t}$$

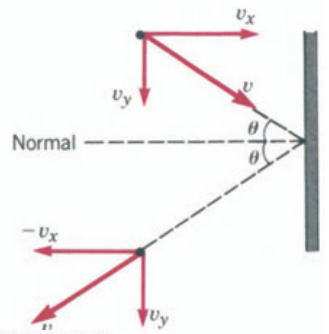


FIGURE 9-3
Two-dimensional representation of an elastic collision of a molecule with the wall of the container.

Because only the direction and not the magnitude of v_x changes on collision

$$\bar{F}_x = \pm\frac{2mv_x}{\Delta t} \tag{9.8}$$

where the choice of sign depends on the assignment of velocity direction sign.

Suppose the atom is moving about in a cubical box of side length l, area of a face $A = l^2$ and a volume $V = l^3$ (see Fig. 9-4). The direction of the force in the impulse of Eq. 9.8 and the initial velocity are reversed if the atom collides with the opposite wall. Thus, it will be convenient to consider only the magnitude of the average force, so we will drop the negative sign in Eq. 9.8. We may approximate Δt of Eq. 9.8 as the time between collisions of the atom against the wall. This is the time for the atom to travel to the opposite wall, bounce off it, and return to the first wall.

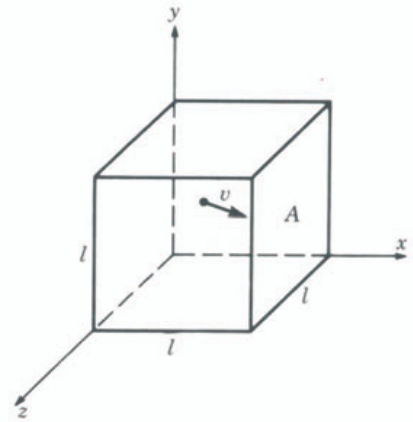


FIGURE 9-4

Because the atom's velocity in the x direction remains constant in magnitude, the time for a round trip between opposite walls is

$$\Delta t = \frac{2l}{v_x} \tag{9.9}$$

Substituting this into Eq. 9.8 obtains the magnitude of the average force on a wall due to the successive striking by one atom

$$\bar{F}_x = \frac{2mv_x}{\dfrac{2l}{v_x}} = \frac{mv_x^2}{l} \tag{9.10}$$

This is the average force on a wall in the y-z plane due to a single atom of the gas. Let us call the force due to this atom $\bar{F}_{x1}$ and the x velocity v_{x1}. Then if there is a second atom with velocity v_{x2}, it would contribute a force $\bar{F}_{x2}$, and so on. The total average force on a wall due to the x motion of N atoms in the box would be the sum of the contribution of each, or

$$\begin{aligned}\bar{F}_x &= \frac{m}{l}v_{x1}^2 + \frac{m}{l}v_{x2}^2 + \cdots\frac{m}{l}v_{xN}^2 \\ &= \frac{m}{l}(v_{x1}^2 + v_{x2}^2 + \cdots v_{xN}^2)\end{aligned} \tag{9.11}$$

If N is the total number of atoms in the box, then by the definition of an average as the sum of the individual amounts divided by the number of items, we may write for $\overline{v_x^2}$, the average of the squared individual x velocities,

$$\overline{v_x^2} = \frac{v_{x1}^2 + v_{x2}^2 + \cdots v_{xN}^2}{N} \tag{9.12}$$

Substitute Eq. 9.12 into Eq. 9.11

$$\bar{F}_x = \frac{mN}{l}\overline{v_x^2} \tag{9.13}$$

By the three-dimensional pythagorean theorem

$$v^2 = v_x^2 + v_y^2 + v_z^2$$

or, expressed in averages,

$$\overline{v^2} = \overline{v_x^2} + \overline{v_y^2} + \overline{v_z^2} \tag{9.14}$$

But in a gas in equilibrium there is no preferred direction of motion; hence

$$\overline{v_x^2} = \overline{v_y^2} = \overline{v_z^2}$$

Therefore, Eq. 9.14 may be written as

$$\overline{v^2} = \overline{3v_x^2} \tag{9.15}$$

Substitute $\overline{v_x^2}$ from Eq. 9.15 into Eq. 9.13, and $\overline{F}_x$ becomes a general force on any wall $\overline{F}$

$$\overline{F} = \frac{mN}{3l}\overline{v^2} \tag{9.16}$$

If we now use the definition of pressure $P = F/A$, we may write Eq. 9.16 as

$$P = \frac{mN}{3Al}\overline{v^2}$$

$$P = \frac{mN}{3V}\overline{v^2} \tag{9.17}$$

where $V = Al$ is the volume of the box. Now multiply and divide Eq. 9.17 by 2 and obtain

$$P = \frac{2}{3}\frac{N}{V}\left(\frac{1}{2}m\,\overline{v^2}\right) \tag{9.18}$$

$$P = \frac{2}{3}\frac{N}{V}\left(\frac{1}{2}m\,\overline{v^2}\right)$$

which shows that the pressure of a gas on the walls of a container is proportional to the average kinetic energy of the atoms or molecules of the gas. One should recognize that atom-atom collisions also take place in a gas. In a more complete calculation these are considered, but the result given by Eq. 9.18 remains unchanged.

9.6 KINETIC THEORY OF TEMPERATURE

We may now show the relation of molecular motion to temperature by using the ideal gas law, Eq. 9.6

$$PV = nRT \tag{9.6}$$

and the equation we just derived for the pressure, Eq. 9.18. Substituting for

the pressure in Eq. 9.6 from Eq. 9.18, we obtain

$$\frac{2}{3} N \left(\frac{1}{2} m \overline{v^2} \right) = nRT \tag{9.19}$$

in which it should be recalled that N is the number of molecules present in the box and n is the number or fraction of moles present. By definition

$$n \text{ (number of moles)} = \frac{N \text{ (number of molecules)}}{N_A \text{ (Avogadro's number)}}$$

Substituting $n = N/N_A$ in Eq. 9.19 gives

$$\frac{2}{3} N \left(\frac{1}{2} m \overline{v^2} \right) = \frac{N}{N_A} RT \tag{9.20}$$

The number of molecules, N, cancels, and we are left with the ratio of two constants R/N_A. This ratio occurs so frequently that it is given the name *Boltzmann's constant*, after the German theorist, with the symbol k_B. Its value is

$$\begin{aligned} k_B &= \frac{R}{N_A} \\ &= \frac{8314 \text{ J/kmol} \cdot \text{K}}{6.02 \times 10^{26} \text{ molecule/kmol}} \\ &= 1.38 \times 10^{-23} \text{ J/K per molecule} \end{aligned}$$

Eq. 9.20 then becomes

$$\frac{1}{2} m \overline{v^2} = \frac{3}{2} k_B T \tag{9.21}$$

$$\frac{1}{2} m \overline{v^2} = \frac{3}{2} k_B T$$

or

$$T = \frac{2}{3} \frac{\overline{E_k}}{k_B} \tag{9.22}$$

and we have shown that temperature is simply proportional to the average kinetic energy E_k of the molecules. Although this calculation has been done for gases, the same result is obtained for liquids and solids.

We may use Eq. 9.21 to find the speed of molecules in a gas. Note, however, that the speed will actually be the square root of average squared velocity. This is called the *root mean square* (RMS) velocity and, although it is not strictly the average speed, its statistical definition is close enough for our purposes. Thus, from Eq. 9.21

$$v_{RMS} = \sqrt{\frac{3k_B T}{m}} \tag{9.23}$$

where T is in K and m is the mass of a single molecule (or atom) in kilograms.

Example 9-4

If we consider air to be made up largely of diatomic nitrogen molecules, N_2, what is their RMS velocity at 27° C? One nitrogen atom has a mass of $14 \times 1.67 \times 10^{-27}$ kg, and the mass of N_2 is twice that.

Solution From Eq. 9.23

$$v_{\text{RMS}} = \sqrt{\frac{3k_BT}{m}} = \sqrt{\frac{3 \times 1.38 \times 10^{-23}\ \text{J/K} \times 300\ \text{K}}{28 \times 1.67 \times 10^{-27}\ \text{kg}}} = 515\ \text{m/sec}$$

9.7 MEASUREMENT OF HEAT

The measure of a quantity of heat ΔQ was established by French scientists as the *calorie*. *One calorie is the quantity of heat required to raise the temperature of 1 g of water by 1° C.* (In the English system the measure is the British thermal unit (BTU), where 1 BTU is the quantity of heat required to raise the temperature of 1 lb of water by 1° F.)

As we discussed in Chapter 5, friction between surfaces causes loss of mechanical energy. However, experience shows us that friction produces heat. Anyone who has used sandpaper on a wooden surface has observed this phenomenon. From the preceding section we recognize that this temperature rise is due to the increased kinetic energy of the molecules. This increase has been produced by the work done on the molecules by the sandpaper. So we see that by our understanding of the nature of temperature we need not restrict the law of energy conservation to frictionless systems; the apparent loss of mechanical energy of the moving system has gone into increased mechanical energy of the molecules.

We may measure how much mechanical energy produces what quantity of heat by a simple experiment (see Fig. 9-5). Suppose we have a paddle wheel driven by a falling weight. The paddle wheel is in a known quantity of water completely insulated from heat flowing in or out. By letting the weight fall with constant velocity and measuring the increased temperature of the water, we may find the mechanical energy equivalent of heat. (*Note*: The weight is allowed to fall at constant speed so that no changes in kinetic energy have to be considered, only changes in potential energy.) The quantity of heat ΔQ is proportional to the temperature rise. This involves the mass of the water m and the *specific heat* of the water c, which is defined as *the amount of heat needed to raise the temperature of 1 g of a substance* (in this case water) 1° C. As mentioned earlier, for water this is 1 cal/g °C. Therefore, the quantity of heat ΔQ required to raise the temperature by ΔT of a mass m of a substance whose specific heat is c, is written as

$$\Delta Q = m\,c\,\Delta T \qquad (9.24)$$

$\Delta Q = m\,c\,\Delta T$

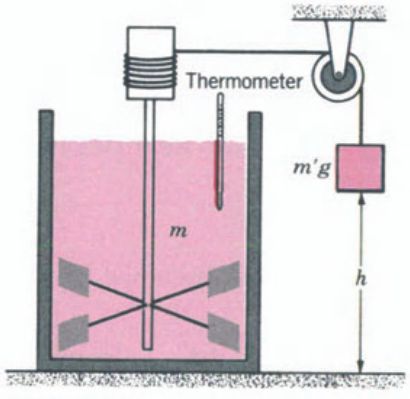


FIGURE 9-5
Diagram of the apparatus for the measurement of the mechanical equivalent of heat.

We can equate this to the loss of potential energy, $\Delta E_p = m'gh$, of the falling weight $m'g$, that is,

$$m'gh = mc\,\Delta T \tag{9.25}$$

For a known amount of ΔE_p and a measured ΔT, the experiment yields the relation

$$4.184\ \text{J} = 1\ \text{cal} \tag{9.26}$$

4.184 J = 1 cal

This relation is called the *mechanical equivalent of heat*.

We may generalize Eq. 9.24 to any substance. On measurement we find, however, that almost all substances have different specific heats. These are not readily calculated but can be determined experimentally and are given by tables in handbooks. The evaluation of the specific heat (or heat capacity) of ideal gases is simpler and is presented next.

9.8 SPECIFIC HEATS OF GASES

If we hold a quantity of gas at a constant volume so that it cannot do work by expanding, then all the heat ΔQ goes into increasing the kinetic energy of the molecules, or

$$\Delta Q = \Delta E_k \tag{9.27}$$

We have discussed c as the specific heat per gram of any substance, but actually we have been talking about solids and liquids. The situation is different for gases because they are compressible. We may define a term C_v as the molar specific heat at constant volume; that is, the specific heat per mole

$$C_v = c_v M \tag{9.28}$$

where M is the mass of a mole of gas and c_v is the specific heat per gram at constant volume. The mass of the gas m is the mass of a mole M multiplied by the number of moles n, that is, $m = Mn$. We may rewrite Eq. 9.24 as ΔQ = (specific heat per mole) × (number of moles) × (ΔT), or

$$\Delta Q = C_v n\,\Delta T \tag{9.29}$$

The resulting increase in the energy of the molecules may be written as

$$\Delta E_k = (\text{number of molecules}) \times \left(\frac{\Delta E_k}{\text{per molecule}}\right)$$

The number of molecules is equal to the number of moles n, multiplied by Avogadro's number N_A, and, from Eq. 9.21, ΔE_k per molecule = 3/2 $k_B\,\Delta T$.

Therefore,

$$\Delta E_k = (nN_A)\left(\frac{3}{2} k_B \Delta T\right) \tag{9.30}$$

Substituting Eq. 9.29 for ΔQ and Eq. 9.30 for ΔE_k in Eq. 9.27, we obtain

$$C_v n \Delta T = \frac{3}{2} nN_A k_B \Delta T$$

or

$$C_v = \frac{3}{2} N_A k_B \tag{9.31}$$

By definition $k_B = \dfrac{R}{N_A}$, and Eq. 9.31 becomes

$$C_v = \frac{3}{2} R \tag{9.32}$$

$C_v = \frac{3}{2} R$

Note that the conversion factor of Eq. 9.26 will reduce R to a value somewhat easier to remember

$$R = 8314 \text{ J/kmol-K} = 8.314 \text{ J/mol-K}\left(\frac{1 \text{ cal}}{4.184 \text{ J}}\right)$$

$$= 1.987 \frac{\text{cal}}{\text{mol K}} \approx 2 \text{ cal/mol-K}$$

Therefore, Eq. 9.32 for ideal gases, which involve only the translational motion of their atoms (no vibration or rotation as in diatomic gases), yields

$$C_v \approx \frac{3}{2} \times 2 \text{ cal/mole-K} \approx 3 \frac{\text{cal}}{\text{mol-K}}$$

This value is expected to hold for all the rare gases that are monatomic, such as helium, neon, and argon. Experiment has proven that the agreement with theory is excellent.

9.9 WORK DONE BY A GAS

Suppose we have a cylinder and a piston with a gas inside, as in Fig. 9-6. Let the cross section of the cylinder be A and the weight of the piston plus a weight resting on it be mg. Suppose the piston was originally at position h_1 and the gas has expanded and pushed it up a distance dx to position h_2.

The definition of work, Eq. 5.3, is force times the distance moved in the direction of the force.

$$dW = F\,dx \tag{5.3}$$

But from the definition of pressure, $P = F/A$, we may substitute for F and obtain

$$dW = PA\,dx$$

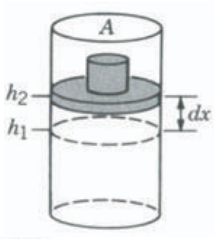


FIGURE 9-6

Since A dx is the change in volume dV, we may write that work done by a gas as

$$dW = P\,dV \tag{9.33}$$

$dW = P\,dV$

Note that by the definition of work, if the gas does work by expanding, the work done is positive, whereas if the gas is compressed by the force on the piston, the work done by the gas is negative.

9.10 FIRST LAW OF THERMODYNAMICS

We now have three factors relating energy and the behavior of a gas:

1. Work done on or by a gas $\Delta W = P\,\Delta V$.
2. The quantity of heat ΔQ that may be added or extracted from the gas.
3. The change in average kinetic energy of the molecules ΔE_k, which we usually call the change in internal energy ΔE.

Because of the interplay of these three terms, it is not meaningful to ask what is the amount of heat in a gas. If heat is added to a gas at constant volume, it all goes into the increase of the internal energy ΔE. If, however, the gas is allowed to expand and do work when heat is added, then the amount of heat available to increase the internal energy depends on the amount of heat that has gone into work. For example, if the gas is allowed to expand and do work while the temperature is held constant ($\Delta T = 0$), then the final internal energy is the same as the initial and $\Delta E = 0$. We may logically write these concepts in the form of the equation

$$\Delta Q = \Delta W + \Delta E \tag{9.34}$$

$\Delta Q = \Delta W + \Delta E$

In Eq. 9.34 ΔQ is taken as positive if heat enters the system (the gas in this case) and as negative if it leaves the system. The work, ΔW, is positive if it is done *by* the system, and negative if done *on* the system. This simple, logical expression bears the ponderous name of the *First Law of Thermodynamics*. It is seen that through the expression for the mechanical equivalent of heat, Eq. 9.26, ΔQ can be expressed as energy, as can ΔW and ΔE. We have then

a full statement of the law of conservation of energy in which the E_{out} (energy out) term of the mechanical law, Eq. 5.14, is now included as well as the energy term E_{in}, which may be heat. Furthermore, because work, ΔW, can give rise to changes both in potential and in kinetic energy of a body or system of bodies, all possible mechanical energy terms have been included. Other forms of energy, for example, radiant energy, which will be discussed in a later chapter, are also included in the first law of thermodynamics.

Example 9-5

Six thousand calories of heat are added to 2 moles of neon gas at 27° C while it does 4100 J of work. (a) How much does the internal energy of the system increase? (b) What is the final temperature of the gas?

Solution

(a) First convert calories to joules

$$\Delta Q = 6000 \text{ cal} \left(\frac{4.184 \text{ J}}{1 \text{ cal}}\right) = 2.51 \times 10^4 \text{ J}$$

Use the first law of thermodynamics

$$\Delta E = \Delta Q - \Delta W$$

$$= 2.51 \times 10^4 \text{ J} - 0.41 \times 10^4 \text{ J} = 2.10 \times 10^4 \text{ J}$$

(b) The relation between the change in the internal energy and the change in temperature is given by Eq. 9.30

$$\Delta E = (nN_A)\left(\frac{3}{2} k_B \Delta T\right)$$

Recalling that $N_A k_B = R$, we write

$$\Delta E = \frac{3}{2} nR \Delta T$$

Therefore

$$\Delta T = \frac{2 \Delta E}{3 nR}$$

$$= \frac{2 \times 2.10 \times 10^4 \text{ J}}{3 \times 2 \text{ mol} \times 8.314 \text{ J/mol-K}}$$

$$= 842 \text{ K or } ^\circ\text{C}$$

$$T_{final} = 27^\circ \text{ C} + 842^\circ \text{ C} = 869^\circ \text{ C}$$

SUPPLEMENT 9-1

Maxwell-Boltzmann Statistical Distribution

Ludwig Boltzmann (1844–1906).

In this chapter we have derived the mean square velocity $\overline{v^2}$ of a large number of atoms or electrons, Eq. 9.21. We did not address the question of how these square velocities are distributed. That is, how many atoms have a square velocity twice the mean or one half the mean. This problem is in the realm of *statistical mechanics*, the science of the application of the first principles of physics to a large number of bodies. The rigorous solution is beyond our interest and is too difficult to present here. However, Professor Richard Feynman has presented a conceptual solution that we will give.

Suppose we have a very tall glass tube with a column of gas going to a great height such as into our upper atmosphere—but, unlike our atmosphere, the temperature is the same at all heights. The problem will be to derive the law of the decrease in density of the atmosphere as the height increases. Let n be the number of moles of gas in volume V at pressure P. From the ideal gas law Eq. 9.6 we write $P = nRT/V$. We recall that $R = k_B N_A$, therefore, $P = (nN_A/V)\,(k_B T)$. The ratio nN_A/V represents the number of molecules per unit volume N. Thus the pressure is proportional to the number of molecules per unit volume because the temperature is constant; that is, $P = Nk_B T$.

The pressure is higher the lower we measure it, for at any point the gas must support all the gas above it. Consider a small cylindrical volume of gas of cross-sectional area $A = 1\ \text{m}^2$ and width dy at a height y (see Fig 9-7). The vertical force from below on this gas is P_y, because $F = P_y A$ and we have taken $A = 1\ m^2$. The vertical force from above at a height $y + dy$ is less than P_y by the weight of the molecules in the section between y and $y + dy$. The total number of molecules in this region is the number per unit volume N times the volume dy (since A was taken as unity). Each molecule has a weight of mg, so the difference in pressure is

$$P_{y\,+\,dy} - P_y = dP = -mg\,N\,dy. \tag{9.35}$$

We have seen that $P = N\,k_B T$. Because T is constant, the pressure depends only on N. Differentiating P with respect to N we obtain

$$dP = k_B T\,dN \tag{9.36}$$

Equating Eq. 9.36 to Eq. 9.35, we have an equation that can be integrated to find N

$$k_B T\,dN = -mg\,N\,dy$$

or

$$\frac{dN}{N} = -\frac{mg}{k_B T}\,dy$$

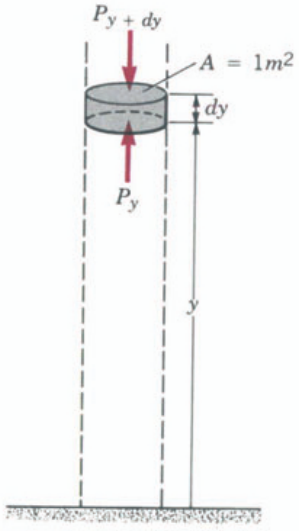


FIGURE 9-7

Integrating

$$\int_{N_0}^{N} \frac{dN}{N} = -\frac{mg}{k_B T} \int_0^y dy$$

which yields

$$N = N_0 \, e^{-mgy/kT} \tag{9.37}$$

where the constant N_0 is the value of N at $y = 0$ and this value of y may be at any predetermined level.

We see that the numerator of the argument of the exponential in Eq. 9.37 contains the potential energy per molecule, and we may write that the density at any point is

$$N = N_0 \, e^{-E_P/k_B T} \tag{9.38}$$

The source of the potential energy need not be gravitational. For example, the potential energies of electrons at different distances from their nuclei have an electrical origin, as we will see in Chapter 14.

In Chapter 5 we saw that as a particle moves in a gravitational field (or any other conservative field) the potential energy of the particle may change. However, the total energy, which is the sum of the potential and kinetic energies, remains constant. Thus it is reasonable to assume that if we start with a given number of molecules having a certain value for the potential energy, the number having that particular value for the total energy, sometime later, will be the same. This means that Eq. 9.38 can be generalized to represent the number of molecules having a certain value for the total energy E; that is,

$$N = N_0 \, e^{-E/k_B T} \tag{9.39}$$

$N = N_0 \, e^{-E/k_B T}$

Eq. 9.39 is known as the *Maxwell-Boltzmann statistical distribution of energy*.

In general, if $E_2 - E_1$ is some energy difference between energy state 1 and a higher energy state 2 of a particle, the ratio of the number occupying the higher energy state to the lower energy state is given by

$$\frac{N_2}{N_1} = e^{-(E_2 - E_1)/k_B T} \tag{9.40}$$

We can use Eq. 9.39 to answer another important question, namely, what fraction of the total number of particles has energy equal to or greater than a certain value E_i? Because Eq. 9.39 represents the number of particles with energy E, the number of particles with energy between E and $E + dE$ will be proportional to $N_0 e^{-E/k_B T} \, dE$. It follows that the fraction of particles with

energies greater than or equal to E_i is

$$\text{Fraction } (E \geq E_i) = \frac{\int_{E_i}^{\infty} N_0\, e^{-E/k_BT}\, dE}{\int_0^{\infty} N_0\, e^{-E/k_BT}\, dE}$$

$$= \frac{\int_{E_i}^{\infty} e^{-E/k_BT} \left(\frac{dE}{k_BT}\right)}{\int_0^{\infty} e^{-E/k_BT} \left(\frac{dE}{k_BT}\right)}$$

$$= \frac{e^{-E/k_BT}\Big|_{E_i}^{\infty}}{e^{-E/k_BT}\Big|_0^{\infty}}$$

$$= e^{-E_i/k_BT} \qquad (9.41)$$

Fraction $(E \geq E_i) = e^{-E_i/k_BT}$

The term of Eq. 9.41 is often referred to as the *Boltzmann factor*.

PROBLEMS

9.1 The molecular weights of sodium and chlorine are 22.99 g/mole and 35.45 g/mole, respectively. How many molecules of sodium chloride (ordinary salt) are there in 100 g of salt?

9.2 The density of copper is 9 g/cm^3, its molecular weight is 64 g/mole. What is the number of copper atoms in 1 m^3?

(*Answer*: 8.47×10^{28}.)

9.3 (a) What is the body temperature of a healthy person on a Celsius clinical thermometer? (b) If the person had a temperature of 102° F what would the thermometer read?

9.4 At what temperature will Fahrenheit and Celsius thermometers read the same value?

(*Answer*: $-40°$.)

9.5 A gas bubble rises from the bottom of a lake to the surface. If the pressure at the bottom is three times atmospheric pressure and the temperature is 4° C while near the surface the temperature is 24° C, what is the ratio of the volume of the bubble just before it reaches the surface to its volume at the bottom of the lake?

9.6 In an ultrahigh vacuum system, the pressure can be lowered to 10^{-10} torr (1 torr = 1/760 atm). How many molecules are there in a vacuum chamber of volume 8×10^6 cm^3 if the pressure is 10^{-10} torr and the temperature is 27° C? 1 atm = 1.01×10^5 N/m^2.

(*Answer*: 2.57×10^{13}.)

9.7 A gas tank contains 10 kg of oxygen at a pressure of 10^7 N/m^2 and a temperature of 27° C. As a result of a leak, the pressure drops to 5×10^6 N/m^2 and the temperature decreases to 7° C. (a) What is the volume of the tank? (b) How much oxygen has leaked out? The molecular weight of oxygen is 32 g/mole.

9.8 The interior of the sun is at a temperature of

about 1.5×10^8 K. The energy is created by the fusion of hydrogen atoms when they collide. In developing the technology for a fusion reactor we simulate this fusion reaction by accelerating protons (hydrogen nuclei) and letting them strike fixed-target hydrogen atoms. What must be the velocity of the protons to simulate 1.5×10^8 K?

9.9 One gram of Ne gas (atomic weight 20.2 g/mole) is in a sealed flask at room temperature, 27° C. If 10 calories of heat are added to the gas, what is the v_{RMS} of the molecules?

(*Answer*: 6.72×10^2 m/sec.)

9.10 The temperature of a gas in a closed container at 27° C is raised to 327° C. By what multiple has v_{RMS} changed?

9.11 The temperature of a room 7 m × 5 m × 3 m is 27° C. (a) How much energy is contained in the air of that room? (b) If that energy could be converted to electrical energy, for how long could a 100-W bulb be lit? Assume the air behaves as an ideal gas.

9.12 A 300-W immersion heater is used to heat a cup of water. If the cup contains 150 g of water at 27° C and 80% of the heater energy is absorbed by the water, how long will it take for the water to begin to boil?

(*Answer*: 191 sec.)

9.13 A 1000-W heater is used to heat the room of problem 9.11. If the molar specific heat of air is 5 cal/mole - K, how long will it take to raise the temperature of the room from 60° F to 70° F? Assume that the quantity, volume, and pressure of air do not change.

9.14 How many calories of heat must be added to 0.5 g of neon gas at constant volume to raise its temperature from 27° C to 127° C? The molecular weight of neon is 20.2 g/mole.

(*Answer*: 7.40 cal.)

9.15 To heat a certain quantity of gas from 27° C to 127° C requires 500 cal when its volume is kept constant. By how much does its internal energy change? How much work could the gas do in cooling back to 27° C?

9.16 A 20-g bullet is shot into a ballistic pendulum with a velocity of 1000 m/sec (see Fig. 9-8). The mass of the wooden block is 2 kg. If the bullet remains embedded in the block and 80% of the energy lost in the collision is absorbed as heat by the bullet, what is the increase in the temperature of the bullet? The specific heat of the bullet is 0.1 cal/g-°C.

(*Answer*: 946° C.)

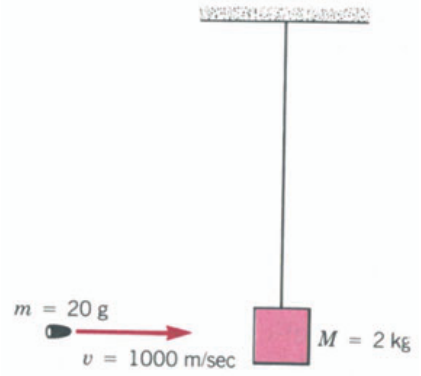


FIGURE 9-8
Problem 9.16.

9.17 Ten moles of an ideal gas at a pressure of 8 atm and with volume 10^{-2} m^3 are allowed to expand isothermally (constant temperature) until the volume doubles. What is the work done by the gas? (1 atm = 1.01×10^5 N/m^2).

(*Answer*: 5.6×10^3 J.)

9.18 Use the ideal gas law, Eq. 9.6, and the first law of thermodynamics, Eq. 9.34, to show that the molar specific heat for a process at constant pressure $C_p = C_v + R$, where C_v is the molar specific heat at constant volume and R is the universal gas constant.

9.19 The initial pressure and volume of 0.1 moles of argon gas are 1 atm (1.01×10^5 N/m^2) and 1 liter (10^{-3} m^3) (see Fig. 9-9). The gas is heated at constant volume until the pressure rises to 4 atm (path A). The gas is then allowed to expand along path B until the pressure drops to 1 atm. The gas is finally cooled down at constant pressure until it returns to its initial state (path C). (a) Find the temperature of the gas at the end of each process (points 1, 2, and 3). (b) Find the internal energy of gas at points 1, 2, and 3. (c) Calculate the work done in each process. (d) Calculate the heat entering or leaving the gas during each process.

(*Answer*: (a) 122 K, 487 K, 609 K, (b) 152 J, 607 J, 760 J, (c) 0 J, 1.01×10^3 J, -4.04×10^2 J, (d) 455 J, 1163 J, -1012 J (leaving the gas).)

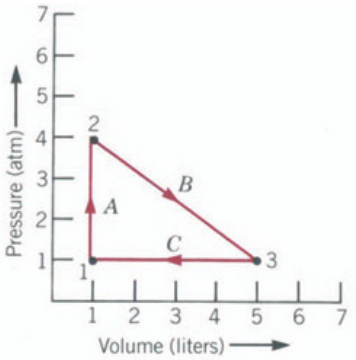


FIGURE 9-9
Problem 9.19.

9.20 A mole of an ideal gas is taken from state A to state C along the path ABC (see Fig. 9-10). (a) If 1000 cal of heat flow into the gas and the gas does 2100 J of work, what is the change in the internal energy of the gas? (b) When the gas is returned from C to A along the path CDA, 700 cal of heat flow out of the gas. How much work is done on the gas? (c) What is the change in the temperature of the gas when it is brought back from C to A? (d) If the pressure of the gas in state A is 2 atm, what is the difference in the volume of the gas between states D and A?

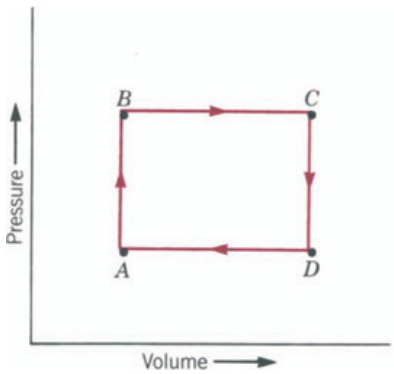


FIGURE 9-10
Problem 9.20.

CHAPTER 10

Oscillatory Motion

10.1 INTRODUCTION

In this chapter we will develop the concepts of oscillatory motion. When a block attached to a spring is set into motion, its position is a periodic function of time. Similarly, in Chapter 7 when we considered the motion of a particle rotating in a circle, we saw that the position coordinates were oscillatory functions of time. Specifically, we showed (Fig. 10-1*a*) that the components of a position vector **r** making an angle θ with the x axis were

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned} \tag{10.1}$$

and that the components of the acceleration, second derivatives with respect to time of these coordinates, were (Fig. 10-1*b*)

$$\begin{aligned} a_x &= \frac{d^2x}{dt^2} = -r\omega^2\cos\theta \\ a_y &= \frac{d^2y}{dt^2} = -r\omega^2\sin\theta \end{aligned} \tag{10.2}$$

As the angle θ goes from 0° to 360° the components of the position vector and of the acceleration vary in value as the sine or cosine functions and go through a reversal of sign; in other words, their values oscillate sinusoidally. In this chapter we will derive some properties of this oscillatory motion with essentially the identical mathematical method that generated Eqs. 10.1 and 10.2.

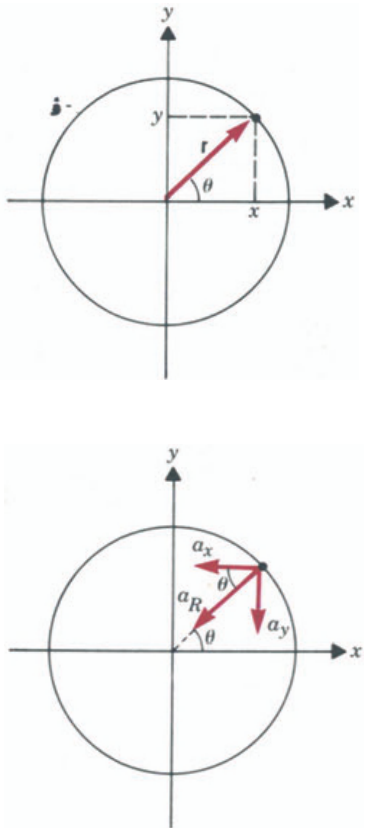


FIGURE 10-1

10.2 CHARACTERIZATION OF SPRINGS

Robert Hooke (1635–1703), an English scientist, was the first to elucidate the behavior of an elastic body such as a spring. He found that the extension or compression x of an elastic body is proportional to the applied force F or

$$F \propto x$$

This simple relationship is now known as Hooke's law. We introduce a proportionality constant to create an equality. This constant has the symbol k and is called the *force constant* or *spring constant*

$$F = kx \tag{10.3}$$

$F = kx$

In the case of a spring, the value of the constant k characterizes the strength (or stiffness) of the spring—a spring with a large k is stronger, or stiffer, than one with a small k. We may readily measure the value of k by simply hanging a weight on the spring and measuring how much it stretches. For the measurement to be valid, the spring must return to its original length when the

weight is removed. The extension of the spring in this case is within its *elastic limit*. If a spring is stretched beyond its elastic limit it will deform permanently and Hooke's law is not obeyed.

10.3 FREQUENCY AND PERIOD

Suppose we have a periodic event, that is, one that occurs regularly with time such as the rising of the sun. We know that it occurs once each day; that is, its *frequency* ν is one event per day, and it has dimensions of $(\text{time})^{-1}$ because event is dimensionless. We use another quantity, the time between periodic events, known as the *period* with symbol T. The period of the sun's rising is 1 day per event, and obviously has dimensions of time. It is seen that period and frequency are reciprocals of each other.

$$\nu = \frac{1}{T} \tag{10.4}$$

$\nu = \frac{1}{T}$

If, for example, the time between periodic events is $T = 0.2$ sec, then the number of events per second, the frequency ν, will be

$$\nu = \frac{1}{0.2 \text{ sec/event}} = 5 \text{ events/sec}$$

As we have indicated, frequency has units of $(\text{time})^{-1}$ or events per second. One event per second is called one hertz, abbreviated Hz.

If a point particle moves on a circle as in Fig. 10-1*a*, its position vector from the center of the circle to the particle has the magnitude of the radius. This is often called the *radius vector* of the point, or simply the radius vector. As the particle moves, the radius vector rotates and if the particle moves with constant speed in the counterclockwise direction, we say that it rotates with a constant positive rotational speed ω. Then θ is a function of time, and from Eq. 7.8

$$\theta = \omega t \tag{10.5}$$

We can use this result to express the coordinates x and y of the rotating particle in Fig. 10-1 as explicit functions of time and the frequency of rotation. Substituting Eq. 10.5 for θ in Eq. 10.1, we obtain

$$\begin{aligned} x &= r \cos \omega t \\ y &= r \sin \omega t \end{aligned} \tag{10.6}$$

The angular speed ω is related to the frequency of rotation ν rather simply. In every rotation θ changes by 2π rad. If the particle performs ν rotations in 1 sec, then θ will change by $2\pi\nu$ rad every second. By definition, ω is the change in θ per unit time (per second). We conclude that

$$\omega = 2\pi\nu \tag{10.7}$$

$\omega = 2\pi\nu$

Substituting Eq. 10.7 for ω in Eq. 10.6 we may write

$$x = r \cos 2\pi\nu t$$
$$y = r \sin 2\pi\nu t \tag{10.8}$$

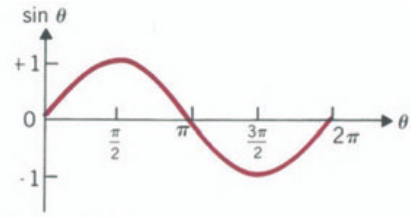


FIGURE 10-2

10.4 AMPLITUDE AND PHASE ANGLE

Fig. 10-2 shows a plot of sin θ versus θ. We see that the value of sin θ oscillates between $+1$ and -1. The maximum value of the magnitude of this oscillation is called the amplitude. In Fig. 10-2, the amplitude is 1. If sin θ were multiplied by a constant A, then A would be the amplitude in the expression A sin θ. Suppose instead of sin θ we plot the function sin$(\theta + \pi/4)$. We see that when $\theta = 0$ the function has the value of sin $\pi/4$ and thereafter attains all values of sin θ at an angle $\pi/4$ earlier, as shown in Fig. 10-3*a*. If, on the other hand, we plot the function sin $(\theta - \pi/4)$ we see in Fig. 10-3*b* that it starts later than the sin θ function. The general form for a function to describe a body undergoing sinusoidal oscillations, such as the one illustrated in Fig. 10-1 is

$$A \sin (\theta + \phi)$$

or, because $\theta = \omega t$, this may be written as

$$A \sin (\omega t + \phi) \tag{10.9}$$

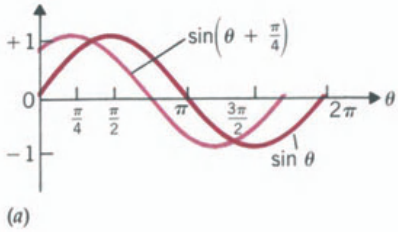
where ϕ is called the *phase angle* and its sign may be positive or negative. Note that if $\phi = \pi/2$, then sin $(\theta + \pi/2) = \cos \theta$, which may be seen by sketching a $+\pi/2$ phase shift on Fig. 10-3*a*. The motion described by Eq. 10.9 is often referred to as *simple harmonic motion*.

10.5 OSCILLATION OF A SPRING

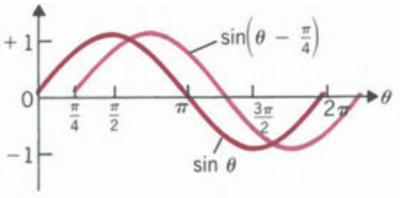
Suppose a body of mass m is connected to a massless spring, with a spring constant k, and the body is free to oscillate on a frictionless surface as in Fig. 10-4. At its rest, or equilibrium position, the position coordinate is $x = 0$, indicated in Fig. 10-4*a*. If the body is pushed to compress the spring a distance x_0, Fig. 10-4*b*, or pulled to stretch it a distance x_0, Fig. 10-4*c*, and then released, the body will then begin to oscillate. We may calculate its subsequent motion from Newton's second law

$$\mathbf{F} = m\mathbf{a}$$

Eq. 10.3 gives an expression of Hooke's law, $F = kx$, for a spring. However, this is the external force needed to compress or to stretch the spring. By Newton's third law of action and reaction, if you pull on a spring with force F it pulls in the opposite direction with force $-F$. Thus, the force that the spring exerts on the body is $-kx$. Because the acceleration is not constant (the force on the body depends on displacement x from equilibrium), we use



(*a*)



(*b*)

FIGURE 10-3

the fundamental definition $a_x = d^2x/dt^2$. Applying Newton's second law to the body, we obtain

$$-kx = m\frac{d^2x}{dt^2} \tag{10.10}$$

This is a second-order differential equation; although there are straightforward mathematical techniques for its solution, we will simply guess a solution and substitute it into Eq. 10.10 to see if an equality is maintained. Such a procedure can verify that the guessed function is a solution but does not prove that it is the only solution. We may try as our guess the function introduced earlier, Eq. 10.9, to describe a body undergoing sinusoidal oscillations. Our guess at a solution will be

$$x = A \sin(\omega t + \phi) \tag{10.9}$$

We may substitute this directly into the left side of Eq. 10.10, but for the right side we need its second derivative

$$\frac{dx}{dt} = A\frac{d}{dt}\sin(\omega t + \phi) = A\omega\cos(\omega t + \phi)$$

$$\frac{d^2x}{dt^2} = A\omega\frac{d}{dt}\cos(\omega t + \phi) = -A\omega^2\sin(\omega t + \phi) \tag{10.11}$$

Substituting Eqs. 10.9 and 10.11 into Eq. 10.10 obtains

$$-kA\sin(\omega t + \phi) = -m\omega^2 A\sin(\omega t + \phi)$$

Cancelling obtains

$$k = m\omega^2$$

and

$$\omega = \sqrt{\frac{k}{m}} \tag{10.12}$$

Therefore, Eq. 10.9 is a solution when the constants have the relation of Eq. 10.12. Using Eq. 10.7, that $\omega = 2\pi\nu$, we immediately obtain the frequency of oscillation

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{10.13}$$

and the period

$$T = \frac{1}{\nu} = 2\pi\sqrt{\frac{m}{k}}$$

To complete the solution of the problem, we must determine the value of the amplitude A and of the phase angle ϕ in the expression for x, Eq. 10.9. This is done by specifying the *boundary conditions*, that is, the behavior of the

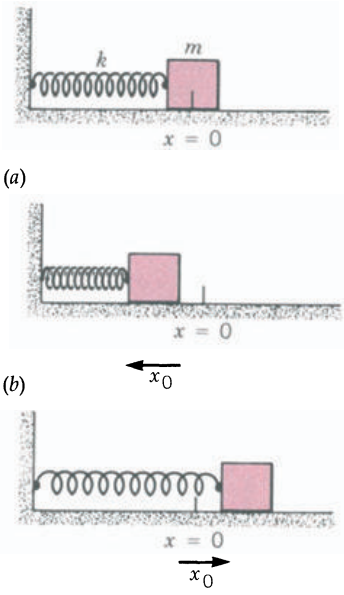


FIGURE 10-4

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

$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$

body at some time such as, $t = 0$. For example, the body in Fig. 10-4 can be set into oscillation by initially stretching the spring a certain distance $x = x_0$ as shown in Fig. 10-4*c* and then releasing it. That is, at $t = 0$.

$$x = x_0 \tag{10.14}$$

$$v_x = 0 \tag{10.15}$$

The first condition, Eq. 10.14, is satisfied by setting $x = x_0$ and $t = 0$ in Eq. 10.9. This yields

$$x_0 = A \sin \phi \tag{10.16}$$

To impose the second condition, we must first determine the velocity of the body as a function of time. This is done as follows:

$$v_x = \frac{dx}{dt} = \frac{d}{dt} A \sin (\omega t + \phi)$$

$$v_x = A\omega \cos (\omega t + \phi) \tag{10.17}$$

The second condition, Eq. 10.15, will be satisfied by setting $v_x = 0$ and $t = 0$ in Eq. 10.17; that is,

$$0 = A\omega \cos \phi \tag{10.18}$$

The amplitude and the phrase angle can now be found by solving simultaneously Eqs. 10.16 and 10.18. If we divide Eq. 10.18 by Eq. 10.16, we obtain

$$\frac{0}{x_0} = \frac{A\omega \cos \phi}{A \sin \phi}$$

or

$$\cot \phi = 0$$

hence

$$\phi = \frac{\pi}{2} \tag{10.19}$$

Substituting Eq. 10.19 for ϕ in Eq. 10.16 yields the result

$$x_0 = A \sin \frac{\pi}{2} = A \tag{10.20}$$

Thus, we see that the amplitude, in this case, is equal to the initial displacement of the body from its equilibrium position, and the phase angle is $\pi/2$ rad. We should note that other boundary conditions will yield different values for A and ϕ.

We can use the facts that $\sin (\theta + \pi/2) = \cos \theta$ and $\cos (\theta + \pi/2) = -\sin \theta$, to eliminate ϕ from the expressions Eq. 10.9 for x and Eq. 10.17 for

v_x, which now become

$$x = A \cos \omega t \tag{10.21}$$

$$v_x = -A\omega \sin \omega t \tag{10.22}$$

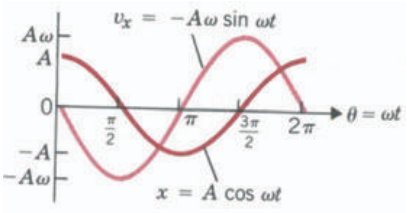


FIGURE 10-5
Plot of x and v as functions of time for ω greater than unity.

We see by Eq. 10.22 that immediately after release from its stretched position to the right, the velocity of the body is toward the left, hence the negative sign. When the argument of the sine, ωt, exceeds π, then the sine function becomes negative and the velocity is positive, or toward the right. We also see from Eq. 10.22 that because the maximum value the sine function may have is ± 1, then the maximum velocity of the block is

$$v_{\max} = \pm A\omega = \pm A\sqrt{\frac{k}{m}} \tag{10.23}$$

$v_{\max} = \pm A\omega$

Furthermore, because $\sin \omega t = 1$ when $\omega t = \pi/2$ and -1 when $\omega t = 3\pi/2$, insertion of these values into Eq. 10.21 shows that the maximum velocity occurs when $x = 0$ or at the midpoint of oscillation. It is instructive to compare a plot of Eq. 10.22 with one of Eq. 10.21 as shown in Fig. 10-5. Note that the amplitude of the displacement A and the maximum value of the velocity $A\omega$ are not the same because ω may be equal to or greater or smaller than unity. Figure 10-5 is a plot with $A\omega \approx 1.2A$. It can be seen that the velocity is maximum when the displacement is zero and zero when the displacement is maximum. The physical significance of this result will become evident when we discuss the energy associated with oscillations in the next section.

We may examine the behavior of the acceleration by taking the time derivative of the velocity in Eq. 10.22.

$$a_x = \frac{dv_x}{dt} = -A\omega \frac{d}{dt} \sin \omega t$$

$$a_x = -A\omega^2 \cos \omega t \tag{10.24}$$

We plot a_x versus θ and compare it with the displacement (x value). Because ω for Fig. 10-5 was taken as 1.2, ω^2 is 1.4, so the maximum value of a_x will be 1.4 times the amplitude of x. The acceleration and displacement curves are plotted in Fig. 10-6. Here we see that the acceleration, although also a cosine curve but with a different amplitude, is a reflection about the θ axis of the displacement. That is, when the displacement is maximum in the positive direction, the acceleration is maximum in the negative direction. Furthermore, when the displacement is zero, so is the acceleration. This relation can be seen physically in Fig. 10-7. When the body is displaced to the right (positive x direction) a distance A and then released (Fig. 10-7*a*), the acceleration is in the negative x direction. The acceleration is also a maximum at this position. This is evident because $F = ma$ and $F = -kx$. Therefore, $-kx = ma$ and a is maximum when x is maximum and x and a have opposite signs. When the body reaches $x = 0$ (Fig. 10-7*b*), the acceleration is 0, but as

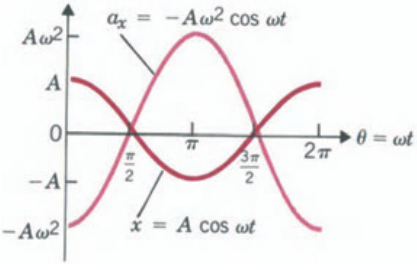


FIGURE 10-6
Plot of x and a as functions of time for ω greater than unity.

it passes to the left the displacement x becomes negative and the acceleration becomes positive or toward the right (Fig. 10-7*c*). This behavior is shown analytically in Fig. 10-6. At $\theta = \pi/2$, which corresponds to zero displacement, the acceleration goes to zero. As the displacement becomes negative ($\pi/2 < \theta < 3\pi/2$) the acceleration becomes positive. We also note from Fig. 10-6 that the two maxima of acceleration occur at $\theta = 0$ or π ($\omega t = 0$ or π). The maximum values of the acceleration at these two positions are, from Eq. 10.24,

$$a_x = -A\omega^2 \cos 0 = -A\omega^2$$

$$a_x = -A\omega^2 \cos \pi = A\omega^2$$

or

$$a_{\max} = \pm A\omega^2 \tag{10.25}$$

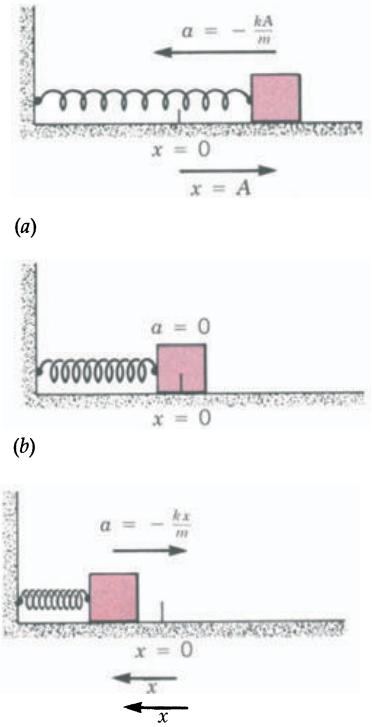


FIGURE 10-7

Example 10-1

Show that $x = A \cos(\omega t + \phi)$ is also a solution of Eq. 10.10.

Solution

$$-kx = m\frac{d^2x}{dt^2} \tag{10.10}$$

Take the second derivative and substitute into the right side of Eq. 10.10.

$$\frac{dx}{dt} = A\frac{d}{dx}\cos(\omega t + \phi) = -A\omega \sin(\omega t + \phi)$$

$$\frac{d^2x}{dt^2} = -A\omega\frac{d}{dx}\sin(\omega t + \phi) = -A\omega^2 \cos(\omega t + \phi)$$

Substituting in Eq. 10.10,

$$-kA \cos(\omega t + \phi) = -mA\omega^2 \cos(\omega t + \phi)$$

After cancelling the cosine term from both sides of this equation, we obtain

$$\omega = \sqrt{\frac{k}{m}}$$

Thus, $A \cos(wt + \phi)$ is a solution of Eq. 10.10 for this value of ω.

$a_{\max} = \pm A\omega^2$

Example 10-2

A given spring stretches 0.1 m when a force of 20 N pulls on it. A 2-kg block attached to it on a frictionless surface as in Fig. 10-4 is pulled to the right 0.2 m and released. (a) What is the frequency of oscillation of the block? (b) What is its velocity at the midpoint? (c) What is its acceleration at either end? (d) What are the velocity and acceleration when $x = 0.12$ m, on the block's first passing this point?

Solution First we must determine the spring constant k.

$$k = \frac{F}{x} = \frac{20 \text{ N}}{0.1 \text{ m}} = 200 \text{ N/m}$$

(a) We may then calculate ω from Eq. 10.12

$$\omega = \sqrt{\frac{k}{m}} = \sqrt{\frac{200 \text{ N/m}}{2 \text{ kg}}} = 10 \text{ rad/sec}$$

Because

$$\omega = 2\pi\nu$$

$$\nu = \frac{\omega}{2\pi} = \frac{10 \text{ rad/sec}}{2\pi} = 1.6 \text{ Hz}$$

(b) The velocity is a maximum when $x = 0$, that is, at the midpoint. Therefore, from Eq. 10.23, (recall, as shown earlier in this section, that when a block is initially displaced a distance x_0 from its equilibrium position and then released, the amplitude of the motion $A = x_0$)

$$v = v_{\max} = \pm A\omega = \pm(0.2\text{m})\ (10 \text{ rad/sec}) = \pm 2 \text{ m/sec}$$

(c) The acceleration is a maximum at the two extremes of the motion. Therefore, from Eq. 10.25

$$a_{\max} = \pm A\omega^2 = \pm(0.2\text{m})\ (10 \text{ rad/sec})^2 = \pm 20 \text{ m/sec}^2$$

(d) To determine the block's velocity and acceleration at some arbitrary value of x, we need to know the angle ωt at that position. In this problem, $x = 0.12$ m. We use the relation

$$x = A \cos \omega t \tag{10.21}$$

$$\omega t = \text{arc cos} \frac{x}{A} = \text{arc cos} \frac{0.12\text{m}}{0.2\text{m}} = 53°$$

Then we may substitute into Eqs. 10.22 and 10.24

$$v = -A\omega \sin \omega t \tag{10.22}$$

$$= -(0.2\text{m})(10 \text{ rad/sec}) \sin 53° = -1.6 \text{ m/sec}$$

(Moving toward the left)

$$a = -A\omega^2 \cos \omega t \tag{10.24}$$

$$= -(0.2\text{m})(10 \text{ rad/sec}^2)^2 \cos 53° = -12 \text{ m/sec}^2$$

(Accelerating toward the left)

10.6 ENERGY OF OSCILLATION

In Chapter 5 we saw that when an object is raised to a height y in the gravitational field on its descent the gravitational force is capable of doing work on the object. Because of this, there is associated with the object at a height y a potential energy $E_p = mgy$. We defined the potential energy as the work done in raising the object to that height. An analogous situation occurs here. When a body attached to a spring is displaced from its equilibrium position ($x = 0$), the spring is potentially capable, on the release of the body, to do work on the body. We can therefore associate with the spring-body system a potential energy E_p. This potential energy will be the work done in stretching or compressing the spring.

When the force **F** and the displacement $d\mathbf{x}$ are in the same direction, work was defined as the product of the magnitudes of the force and the displacement (see Eq. 5.1), that is,

$$dW = F\,dx$$

or

$$W = \int_0^x F\,dx$$

From Hooke's law, Eq. 10.3, the force needed to compress or stretch a spring is $F = kx$; thus

$$W = k\int_0^x x\,dx$$

$$W = \frac{1}{2}kx^2$$

The potential energy of the spring-body system, when the body is displaced a distance x from its equilibrium position, is therefore

$$E_p \text{ (spring)} = \frac{1}{2}kx^2 \tag{10.26}$$

E_p (spring) $= \frac{1}{2}kx^2$

This equation was derived on the assumption that the spring was initially in its equilibrium position, $x = 0$. This assumption is not necessary. If the spring is initially in a position x_1 and is compressed or stretched to position x_2, the work done is as before

$$W = k\int_{x_1}^{x_2} x\,dx$$

$$= \frac{1}{2}kx_2^2 - \frac{1}{2}kx_1^2 \tag{10.27}$$

Note that because the displacement is squared the potential energy of a spring is the same whether it is stretched or compressed an equal distance x from its relaxed position.

If there is no friction we may expect, as was the case with the gravitational force, that the total mechanical energy, kinetic plus potential, will remain constant as the body oscillates. This can be shown rather simply. By the work-energy theorem, Eq. 5.9, the work done by the spring, as the body moves between two arbitrary displacements x_1 and x_2, is equal to the change in the kinetic energy of the body; that is,

$$\int_{x_1}^{x_2} F_{\text{spring}}\, dx = \frac{1}{2} mv_2^2 - \frac{1}{2} mv_1^2 \tag{10.28}$$

where v_1 and v_2 are the velocities of the body at x_1 and x_2, respectively. The force exerted by the spring on the body is $F_{\text{spring}} = -kx$. Substituting this for F_{spring} in Eq. 10.28, and integrating the left side of the equation we obtain

$$-\int_{x_1}^{x_2} kx\, dx = -\left(\frac{1}{2} kx_2^2 - \frac{1}{2} kx_1^2\right)$$

Eq. 10.28 becomes

$$-\left(\frac{1}{2} kx_2^2 - \frac{1}{2} kx_1^2\right) = \frac{1}{2} mv_2^2 - \frac{1}{2} mv_1^2$$

Rearranging terms, we obtain

$$\frac{1}{2} kx_1^2 + \frac{1}{2} mv_1^2 = \frac{1}{2} kx_2^2 + \frac{1}{2} mv_2^2 \tag{10.29}$$

Because x_1 and x_2 are arbitrary points we conclude that the total energy

$$E_{\text{total}} = E_{\text{p}}\left(\frac{1}{2} kx^2\right) + E_k\left(\frac{1}{2} mv^2\right)$$

$$E_{\text{total}} = \frac{1}{2} kx^2 + \frac{1}{2} mv^2$$

remains constant as the body oscillates.

Note that when the spring is stretched, $x = A$. Before the object is released it has no velocity and $1/2\ kA^2$ is the total energy of the system; after it is released the energy remains constant because energy neither enters nor leaves the system.

In Section 10.5 we saw that the velocity was a maximum when the displacement was zero, and it was zero when the displacement was a maximum. This result is intimately tied to the fact that the total mechanical energy of the system remains constant. Therefore, the kinetic energy (and hence the velocity) will be a maximum when the potential energy is a minimum, that is, when x is zero. The kinetic energy will be a minimum (zero) when the potential energy is a maximum, that is, when $x = A$.

Example 10-3

The block of Example 10-2 is released from a position of $x_1 = A = 0.2$ m as before. (a) What is its velocity at $x_2 = 0.1$ m? (b) What is its acceleration at this position?

Solution

(a) The velocity at x_2 can be found with the conservation of energy equation, Eq. 10.29

$$\frac{1}{2}kx_1^2 + \frac{1}{2}mv_1^2 = \frac{1}{2}kx_2^2 + \frac{1}{2}mv_2^2$$

Solving for v_2, noting that $v_1 = 0$, we obtain

$$v_2 = \left[\frac{k(x_1^2 - x_2^2)}{m}\right]^{1/2}$$

$$= \left[\frac{200 \text{ N/m } [(0.2\text{m})^2 - (0.1 \text{ m})^2]}{2 \text{ kg}}\right]^{1/2}$$

$$= 1.73 \text{ m/sec}$$

(b) We may find the acceleration at this position by using Newton's second law

$$F = ma$$

$$-kx = ma$$

$$a = -\frac{kx}{m} = -\frac{(200 \text{ N/m})(0.1 \text{ m})}{2 \text{ kg}} = -10 \text{ m/sec}^2$$

PROBLEMS

10.1 Show that $x = A \sin \omega t + B \cos \omega t$, where A and B are arbitrary constants, is a solution of Eq. 10.10.

10.2 The position of a particle undergoing oscillations is given by $x = 25 \sin (3\pi t + \pi/5)$, where x is in centimeters and t in seconds. Find (a) the frequency of the motion, (b) the amplitude of the motion, (c) the maximum velocity of the particle, (d) the maximum value of the acceleration of the particle, (e) the position, the velocity, and the acceleration of the particle at $t = 0$.

10.3 The same block on the same spring as in Example 10-2 is released after being pulled 0.2 m to the right. Find its position, velocity, and acceleration 0.1 sec after being released.

10.4 A small block attached to a spring is oscillating horizontally on a frictionless surface with an amplitude of 0.12 m. When it is at the position $x = 0.05$ m its velocity is 2 m/sec. (a) What is its frequency of oscillation? (b) What is its position when its velocity is 1 m/sec?

(*Answer*: (a) 2.92 Hz, (b) 0.107 m.)

10.5 An oscillating block of mass 250 g takes 0.15 sec to move between the endpoints of the motion, which are 40 cm apart. (a) What is the frequency of the motion? (b) What is the amplitude of the motion? (c) What is the force constant of the spring?

10.6 When a mass of 0.2 kg is suspended from a spring, it stretches 0.04 m. The mass is pulled down an additional distance 0.1 m from its equilibrium position and released. (a) What is the spring constant? (b) What is the period of oscillation?

(c) What is the frequency of oscillation? (d) What will be the maximum velocity?

(*Answer*: (a) 49 N/m, (b) 0.40 sec, (c) 2.5 Hz, (d) 1.57 m/sec.)

10.7 (a) Write down the equation for the position y (measured from the equilibrium position) of the mass in problem 10-6. (b) What is the equation for the velocity of the mass as a function of time? (c) What is the equation for the acceleration of the mass as a function of time? Take positions below the equilibrium point as positive.

10.8 (a) How long after being released is the position of the mass in problems 10-6 and 10-7 equal to 0.05 m? (b) What is the velocity of the mass when $y = 0.05$ m? (c) What is the acceleration of the mass when $y = 0.05$ m?

(*Answer*: (a) 6.7×10^{-2} sec, (b) -1.36 m/sec, (c) -12.25 m/sec^2.)

10.9 The position of an oscillating particle is given by $x = A \sin(\omega t + \phi)$, Eq. 10.9. A particle of mass $m = 0.5$ kg is connected to a spring of force constant $k = 200$ N/m. The particle is initially at rest on a frictionless table. The particle is given an initial velocity of 1.5 m/sec to start oscillating. What is the amplitude of the motion A and the phase angle ϕ?

(*Answer*: 7.5×10^{-2} m, 0 rad.)

10.10 Two springs with force constants $k_1 = 100$ N/m and $k_2 = 200$ N/m are connected to opposite ends of a block of mass 3 kg (see Fig. 10-8). (a) If the block is displaced 0.1 m to the right, what is the net force exerted by the springs on the block? The block is released from that position. (b) What are the frequency and the period of the motion? (c) What is the amplitude of the motion? (d) Find an expression for the position of the particle as a function of time?

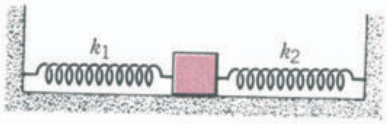


FIGURE 10-8
Problem 10.10.

10.11 A block of mass 2 kg sits on a platform that is oscillating in a vertical plane with an amplitude of 10 cm (see Fig. 10-9). If the frequency of oscillation is 1 Hz, what is the normal force exerted by the platform on the block (a) as they pass the equilibrium point, (b) at the lowest point of the motion, (c) at the highest point of the motion? (d) If the frequency remains constant, at what value of the amplitude will the block and the platform separate?

(*Answer*: (a) 19.6 N, (b) 27.5 N, (c) 11.7 N, (d) 24.8×10^{-2} m.)

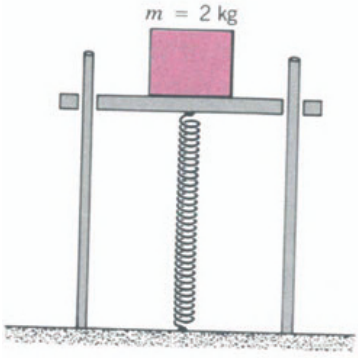


FIGURE 10-9
Problem 10.11.

10.12 A block is oscillating horizontally on a frictionless table with an amplitude of 5 cm. A coin of mass $m = 3$ g is placed on top of the block. The maximum force of friction between the coin and the block is 0.015 N. What is the maximum value of the frequency of oscillation for which the coin will stay on top of the block?

(*Answer*: 1.59 Hz.)

10.13 A block is oscillating with an amplitude of 20 cm. The spring constant is 150 N/m. (a) What is the energy of the system? (b) When the displacement is 5 cm, what is the kinetic energy of the block and the potential energy of the spring?

10.14 A 0.25-kg mass is oscillating with a frequency $\nu = 5$ Hz. What is the amplitude of the motion if the energy of the system is 12 J?

10.15 A mass is oscillating with amplitude A. (a) When the displacement is $x = \frac{1}{2}A$, what fraction of the energy is potential and what fraction is kinetic? (b) For what value of x in terms of A will the energy be half kinetic and half potential?

(*Answer*: (a) $\frac{1}{4}$, $\frac{3}{4}$, (b) 0.707A.)

10.16 A wooden block of mass 0.8 kg rests on a frictionless table connected to a spring ($k = 200$ N/m) as shown in Fig. 10-10. A 20-g bullet moving with

a velocity $v = 500$ m/sec is shot into the block and remains embedded in it. What is the amplitude of the ensuing oscillatory motion?

(*Answer*: 0.78 m.)

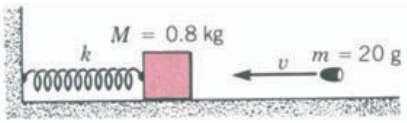


FIGURE 10-10
Problem 10.16.

10.17 A block of mass $m_1 = 3$ kg rests on a frictionless surface connected to a spring ($k = 150$ N/m). A second block of mass $m_2 = 1$ kg is launched toward m_1 with a velocity of 4 m/sec (see Fig. 10-11). After the collision, m_2 bounces back in the opposite direction with a velocity of 1 m/sec. (a) How much will the spring be compressed? (b) What fraction of the energy is lost in the collision?

(*Answer*: (a) 0.24 m, (b) 0.42.)

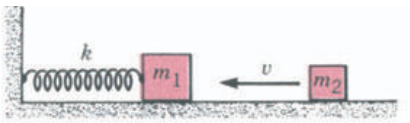


FIGURE 10-11
Problem 10.17.

10.18 A spring ($k = 200$ N/m) is compressed 10 cm between two blocks of mass $m_1 = 1.5$ kg and $m_2 = 4.5$ kg (see Fig. 10-12). The spring is not connected to the blocks, and the table is frictionless. What are the velocities of the blocks after they are released and lose contact with the spring? Assume that the spring falls straight down to the table.

FIGURE 10-12
Problem 10.18.

10.19 A mass of 3 kg is connected to a spring of force constant 250 N/m on a horizontal surface. The coefficient of friction between the block and the surface is 0.1. The block is pulled 20 cm to the right and released. (a) How far to the left of the equilibrium point will the block move? (b) What is the total back and forth distance traveled by the block before it stops?

(*Answer*: (a) 0.176 m, (b) 1.70 m.)

CHAPTER 11

Wave Motion

11.1 INTRODUCTION

Waves are an important concept in physics. We can see water waves and readily demonstrate sound waves with elementary laboratory experiments. In 1801 Thomas Young showed that light can be considered waves by experimental analogies to the behavior of water waves. It will be shown in a later chapter that experiments with fundamental particles, such as electrons, demonstrate that they also have wave characteristics.

If we crack a whip, we produce a brief transverse displacement that can be seen to travel to the end of the whip. If we drop a pebble on the calm surface of a pond, a circular ripple is produced. This ripple travels away from the point where the pebble hit the water with constant speed and, as it reaches a given point of the water surface, it produces a temporary transverse displacement of the water molecules. These are examples of *traveling* waves, which can transmit energy along a medium without any net translation of the particles in the medium through which the wave travels. These readily visible experiments with water or strings led early scientists to conclude that a wave is a disturbance that travels in a medium. When light was shown to have wave characteristics, even though it can travel in the vacuum of space that exists between the earth and the stars, an erroneous conclusion was drawn that there must exist a medium permeating the entire universe. This was called *aether*. As we will see in Chapter 16, because of the nature of the light wave, no medium is necessary for its propagation. Before we enter into the realm of modern physics, we must have a solid understanding of wave motion: the mathematical description of a wave, the parameters that characterize it, the laws that govern its propagation. In the development of these ideas we will consider waves in a visible medium such as a string or a water surface.

11.2 WAVELENGTH, VELOCITY, FREQUENCY, AND AMPLITUDE

Suppose we are sitting in a boat on a body of water in which there is some wave motion, as in Fig. 11-1. If we measure the time between risings on the waves, we have a quantity known as the *period*, with symbol T, which is the time between successive risings. If instead, we ask how many risings did we experience per unit time, such as 1 h, this quantity is called the *frequency*, with symbol ν. Here ν = number of risings/unit time. Just as in the case of oscillatory motion considered in Chapter 10, the period and the frequency are reciprocals of each other.

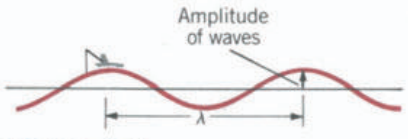


FIGURE 11-1

$$\nu = \frac{1}{T} \tag{11.1}$$

If we ask a friend in another boat to row away from us to a location such that he rises at the same instant that we do, the distance between us is called

$\nu = \frac{1}{T}$

the *wavelength* of the wave with symbol λ. The speed of the wave through the water is the distance between the boats divided by the time it takes for our rising to reach him. This is called the *wave velocity*, with symbol v, and

$$v = \frac{\lambda}{T}$$

or, from Eq. 11.1,

$$v = \lambda\nu \tag{11.2}$$

This is a fundamental equation obeyed by all waves.

Another important parameter that characterizes a wave is its *amplitude*. The amplitude of a wave is the maximum value of the displacement it produces (see Fig. 11-1).

$v = \lambda\nu$

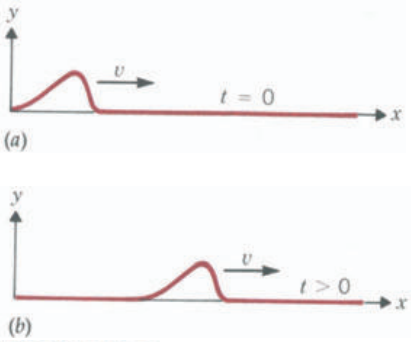


FIGURE 11-2
A transverse displacement in a string traveling in the positive x direction.

11.3 TRAVELING WAVES IN A STRING

Let us have a very long string stretched along the x direction. A pulsed displacement, such as plucking the string, introduced at one end of the string, will cause a transverse displacement of the string in the y or z direction. Let us for now consider only the transverse y direction because it is more easily drawn on a flat sheet of paper.

At $t = 0$, when the pulse is introduced, the string may look as in Fig. 11-2*a*. The wave pulse travels along the string and, consequently, the string will look differently some time later, see Fig. 11-2*b*. This clearly indicates that the transverse displacement of the string, y, varies with x—that is, the point of the string under consideration—and with time t. In mathematical terms we say that the wave pulse, y, is a function of x and t; that is, $y = f(x,t)$. The exact shape of the wave pulse, the precise form of the function $f(x,t)$, will be determined by the source producing the pulse and the nature of the string. One of the most important and most commonly found types of traveling waves is the sinusoidal traveling wave, a wave consisting of a series of consecutive sinusoidal pulses.

Let us attach the end of the string ($x = 0$) to a block connected to a spring hanging from the ceiling, as shown in Fig. 11-3. If we pull on the block, thus stretching the spring, and then release it, we know that the block will begin to oscillate. The y coordinate of the block, and therefore the transverse displacement of that end of the string, will be given by Eq. 10.9

$$y(x = 0,t) = A \sin(\omega t + \phi) \tag{11.3}$$

This transverse displacement, introduced at $x = 0$, moves along the string and, as a result, sometime later other points on the string will begin to oscillate in the transverse y direction. If the velocity of the wave in the string is v, then the time it takes to travel a distance x along the string is x/v. Thus, at time t the displacement produced by the wave at a point x is the same as was the displacement at the origin ($x = 0$) at an earlier time $t - t_0$, where t_0 is

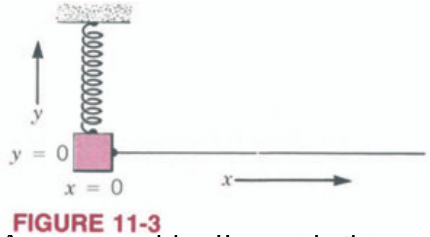


FIGURE 11-3
Arrangement for the production of sinusoidal traveling waves in a long string under tension.

the time that it takes the wave to reach x; that is, $t_0 = x/v$. Putting $t - t_0$ into Eq. 11.3 for t, we obtain

$$y(x,t) = A \sin [\omega(t - t_0) + \phi]$$

and on substituting $t_0 = x/v$ we have

$$y(x,t) = A \sin \left(\omega t - \frac{\omega}{v} x + \phi\right) \tag{11.4}$$

Eq. 11.4 is the general form of the wave equation, but for our purposes we do not require such generality. If we limit ourselves to waves such that $y = 0$ when both $x = 0$ and $t = 0$, then $\phi = 0$ or π. We can then write the commonly used versions of Eq 11.4 as

$$y(x,t) = A \sin (\omega t - kx) \quad \text{when } \phi = 0 \tag{11.5}$$

or

$$y(x,t) = A \sin (kx - \omega t) \quad \text{when } \phi = \pi \tag{11.5'}$$

$y(x,t) = A \sin (kx - \omega t)$

where

$$k = \frac{\omega}{v} \tag{11.6}$$

Either of these versions of the travelling wave equation may be used, and the choice is a writer's preference.

The constant k that we have introduced is called the *propagation constant* (or *wave number*). (Note that this k is a new and different constant from the spring constant k introduced in Chapter 10.) It will be used extensively in the latter part of this book. Its physical significance will soon become apparent.

We can write an alternative representation of a sinusoidal traveling wave that differs from that of Eq. 11.5 in the direction of propagation of the wave that it represents. Equation 11.5 was derived by assuming that the wave traveled toward the right in the positive x direction. For a wave traveling toward the left,

$$y(x,t) = A \sin (kx + \omega t) \tag{11.7}$$

To show that Eq. 11.7 represents a wave traveling in the negative x direction, we look at a particular value of y, the wave displacement, and ask ourselves, as time t increases, what happens to x for that particular y value of the wave? To pick a particular, fixed value of y, the argument of the sine function in Eq. 11.7 must be kept constant, that is,

$$kx + \omega t = \text{constant} \tag{11.8}$$

Obviously, as t increases, x must decrease if the left side of Eq. 11.8 is to remain constant. Therefore, the wave of Eq. 11.7 is moving toward decreasing values of x, that is, in the negative x direction.

To understand a traveling wave in a string produced by an oscillating source such as that of Fig. 11-3, let us examine some stop-action diagrams,

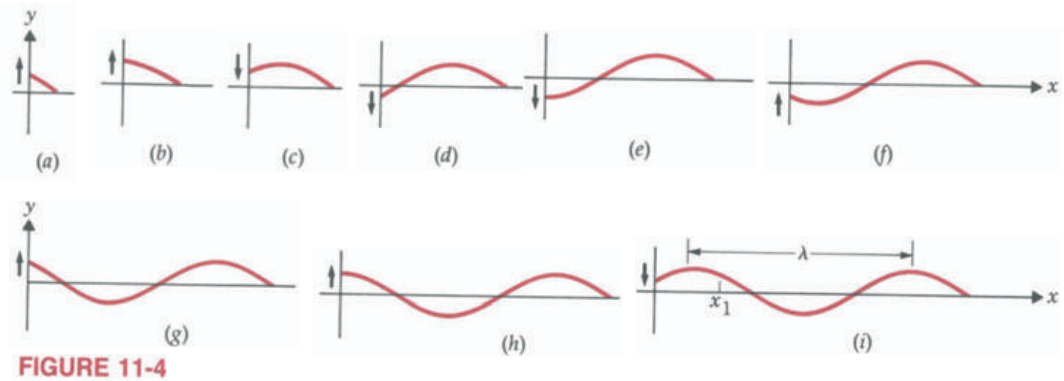


FIGURE 11-4
Stop-action diagrams showing the introduction of a traveling wave in a string by the experimental arrangement of Figure 11-3.

such as those in Fig. 11-4. We see that the y displacement introduced by the oscillation travels to the right in the detailed sketches of Fig. 11-4*a–i*. More than one full wavelength is represented in Fig. 11-4*g–i*.

To gain insight into the physical significance of the wave represented by either Eq. 11.5′ or Eq. 11.7, let us analyze it from two different points of view.

Suppose we take a snapshot of the string as the wave travels through it. What will we see? Taking a snapshot of the string means setting t equal to an instantaneous value t_1 in Eq. 11.5′, which now becomes

$$y(x,t_1) = A \sin (kx - \theta_1) \tag{11.9}$$

where $\theta_1 = \omega t_1$ is a *phase shift* at t_1. This is shown best in Fig. 11-5 in comparison with a sine function plotted as a dashed line. In this figure the solid line is the snapshot at time t_1 and the dashed line is a sine curve with $y = 0$ at the origin. At the time of the photograph the solid line is phase shifted by an angle $\theta_1 = \omega t_1$ from the sine curve of $y = 0$ at $x = 0$. At some other time t_2 the snapshot would show a different phase shift θ_2. Even though there is a phase shift, the string looks like a sine wave. We can use Eq. 11.9 to find the wavelength λ of the wave. In Section 11-2, we defined λ as the separation between two risings in the body of water of Fig. 11-1. Similarly here, λ is the separation between two successive maxima in the transverse displacement of the string; for example, from Fig. 11-5

$$\lambda = x_2 - x_1$$

We identify x_1 and x_2 as two successive values of x for which the sine function

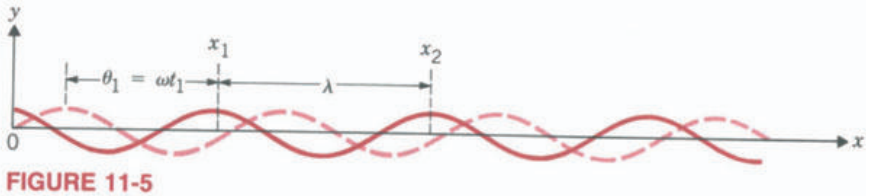


FIGURE 11-5

in Eq. 11.9 equals +1; that is,

$$kx_1 - \theta_1 = \frac{\pi}{2} \text{ rad} \tag{11.10}$$

and

$$kx_2 - \theta_1 = \left(\frac{\pi}{2} + 2\pi\right) \text{ rad} \tag{11.11}$$

Solving Eqs. 11.10 and 11.11 simultaneously, we obtain

$$kx_2 - kx_1 = 2\pi$$

$$x_2 - x_1 = \frac{2\pi}{k}$$

$$\lambda = \frac{2\pi}{k} \tag{11.12}$$

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

We see, therefore, that the propagation constant k in Eq. 11.5′ yields, through the relation of Eq. 11.12, the wavelength of the wave.

An alternative way of looking at Eq. 11.5′ is to consider a particular point in the string of Fig. 11-4*i* and to analyze the motion of that point as a function of time. We can place a little light bulb at that point in the string and follow the motion of the bulb. Suppose we choose the particular point in the string as x_1 in Fig. 11-4*i*. This means we set x equal to a constant value x_1 in Eq. 11.5′, which now becomes

$$y(x_1,t) = A \sin (\theta_1' - \omega t) \tag{11.13}$$

where $\theta_1' = kx_1$ is a constant phase shift that depends on the point chosen in contrast to the previous analysis, which showed that the phase shift depended on the time of the snapshot. We immediately recognize Eq. 11.13 as being similar to Eq. 10.9, which described the oscillatory motion of the body attached to a spring. With the position x fixed, y will vary with sin ωt and the little bulb will undergo simple harmonic motion with amplitude A and frequency

$$\nu = \frac{\omega}{2\pi} \tag{11.14}$$

$$\nu = \frac{\omega}{2\pi}$$

It should be noted that Eq. 11.13, with a change in x and therefore with the corresponding phase shift θ_1', describes the motion of any other point in the string. As the wave moves through the string, all the particles in the string oscillate with the same amplitude and frequency, although *out of phase* with one another, that is, with different phase shifts.

We conclude this section with an alternative demonstration of the relation between frequency and wavelength, Eq. 11.2. Combining Eq. 11.14 and Eq. 11.12, we have the product

$$\lambda\nu = \frac{2\pi}{k}\frac{\omega}{2\pi} = \frac{\omega}{k}$$

By definition $k = \frac{\omega}{v}$ (Eq. 11.6), therefore

$$\lambda\nu = \frac{\omega}{k} = \frac{\omega}{\frac{\omega}{v}} = v \tag{11.2}$$

which is the result found earlier. This result is valid for all waves whether in a medium such as a string, or water or air, or in a vacuum such as light waves.

Example 11-1

A mass of 0.2 kg suspended from a spring of force constant 1000 N/m is attached to a long string as shown in Fig. 11-3. The mass is set into vertical oscillation, and the distance between successive crests of the waves in the string is measured to be 12 cm. What is the velocity of waves in the string?

Solution We use Eq. 10.12 to find the frequency of oscillation

$$\omega = \sqrt{\frac{k}{m}} = \sqrt{\frac{1000 \text{ N/m}}{0.2 \text{ kg}}} = 70.71 \text{ rad/sec}$$

$$\nu = \frac{\omega}{2\pi} = \frac{70.71 \text{ rad/sec}}{6.28} = 11.26 \text{ Hz}$$

The velocity of waves in a medium can be found with Eq. 11.2

$$\begin{aligned} v &= \lambda\nu \\ &= 12 \times 10^{-2}\text{m} \times 11.26 \text{ sec}^{-} \\ &= 1.35 \text{ m/sec} \end{aligned}$$

11.4 ENERGY TRANSFER OF A WAVE

One of the most important aspects of wave motion is that it provides a mechanism for the transfer of energy. A particle in a string before the wave arrives has no mechanical energy. If a sinusoidal wave arrives at the location of the particle, the particle begins to execute simple harmonic motion, and it therefore acquires kinetic as well as potential energy. *The wave has given energy to the particle because the wave carries energy with it.*

In fact if we think carefully, wave motion is one of the two general mechanisms available to transport energy from one point to another. The other occurs when one or more particles move from one point to another and in so doing bring their kinetic energy with them. This kinetic energy can be transferred to other particles in the medium through which they propagate. There are, however, two obvious differences between these two mechanisms; one of them will be crucial in the development of quantum mechanics in a later chapter.

The splashing of the water caused by the sound waves of a tuning fork is a vivid illustration of the transfer of energy by a wave.

1. The first difference is that waves transfer energy without transfer of matter, unlike the motion of particles.
2. The second is that the energy of a beam of particles is localized (it is where the particles are at a given instant). In a wave the energy is distributed over the entire space occupied at a given instant by the wave. (When the ripples in the pond move outward from their source, all the water in the region of the wave is displaced.)

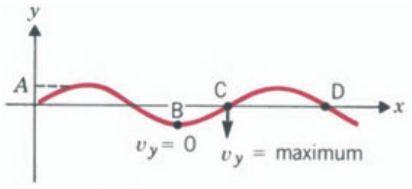


FIGURE 11-6

We will now calculate the rate—that is, energy per unit time—at which energy is transmitted by a wave in a string, noting that a similar calculation, leading to similar results, may be made for any other type of wave. Let us consider the sinusoidal traveling wave represented by Eq. 11.5′.

$$y = A \sin (kx - \omega t) \tag{11.5'}$$

The rate of energy production, consumption, or transmission was defined in Eq. 5.15 as power P. We can obtain P by calculating the energy crossing a given point in a string in 1 sec, for example, point D in Fig. 11-6. This will be equal to the wave energy of the string particles in one wavelength multiplied by the number of wavelengths passing point D in 1 sec, that is, by the frequency ν.

$$P = (\text{energy per wavelength}) \times \nu \tag{11.15}$$

To find the energy in one wavelength we note, as shown in the previous section, that each particle in the string is oscillating with the same amplitude A. Because the total energy of an oscillating particle is proportional to the square of the amplitude of oscillation (see Eq. 10.26 et seq.), we conclude that all the particles in the vibrating string have the same energy. At any given time, the energy of a particular particle may be all kinetic or all potential or a mixture. In Fig. 11-6, the energy of particle C is all kinetic, because C is passing through the equilibrium point. On the other hand, the energy of particle B is all potential, because it is about to reverse the direction of its transverse motion and its velocity is zero.

To obtain the kinetic energy of the particles in the string we need an expression for the transverse velocity v_y. This can be obtained by differentiating y in Eq. 11.5′ with respect to time. Because x is also a variable in the expression for y, we will indicate its omission in the derivative by writing v_y as a partial derivative with respect to t; that is,

$$v_y = \frac{\partial y}{\partial t} = \frac{\partial}{\partial t} A \sin (kx - \omega t)$$

$$v_y = -\omega A \cos (kx - \omega t) \tag{11.16}$$

We can now calculate the energy of particle C of mass Δm.

$$E \text{ (for particle } C) = \frac{1}{2} \Delta m \, v_{y\max}^2 \tag{11.17}$$

But from Eq. 11.16 $v_{y\max} = -\omega A$, because the maximum value of the cosine function is 1. Eq. 11.17 becomes

$$E \text{ (for particle C)} = \frac{1}{2}\Delta m\ \omega^2 A^2 \tag{11.18}$$

Because the energy of all the particles is the same, we can obtain the energy in one wavelength by replacing Δm in Eq. 11.18 with the mass contained in one wavelength. If the amplitude of the wave is small compared with the wavelength (a situation often satisfied), then the mass in one wavelength is $\mu\lambda$, where μ is the mass per unit length of the string. Therefore,

$$\text{Energy per wavelength} = \frac{1}{2}\mu\lambda\omega^2 A^2 \tag{11.19}$$

Substituting Eq. 11.19 for the energy per wavelength in Eq. 11.15, we obtain

$$P = \frac{1}{2}\mu\lambda\nu\omega^2 A^2$$

or

$$P = 2\pi^2\mu v\nu^2 A^2 \tag{11.20}$$

where we have made use of the fact that $\lambda\nu = v$ and $\omega = 2\pi\nu$.

Although Eq. 11.20 has been derived for a wave in a string, two important features hold for any other type of wave: The *power transported by a wave is proportional to the square of the amplitude* and to the velocity of propagation of the wave. We will use these important results later in our discussions of the principles of modern physics.

When considering waves that propagate in three dimensions, such as sound waves or light waves, it is convenient to talk about the energy flowing through a given area of the medium traversed by the wave. The unique term *intensity*, with symbol I, is used for this purpose. The intensity is defined as the power transmitted per unit area perpendicular to the direction of propagation of the wave. Clearly, intensity and power are related by a simple geometric factor. Thus, the intensity of the wave is also proportional to the square of the amplitude. In the SI system, intensity has units of W/m^2.

$I \sim$ (Amplitude of wave)^2

PROBLEMS

11.1 The speed of sound in air is about 330 m/sec, whereas the velocity of light is 3×10^8 m/sec. If you see a flash of lightning and count 8 sec before you hear the thunder, how far away was the lightning? (*Answer*: 2.64×10^3 m.)

11.2 If the principal audio frequency of a thunderclap is the lowest the ear can hear, about 20 Hz, what is the wavelength of the sound wave?

11.3 The speed of all electromagnetic waves in air, both visible and invisible, is 3×10^8 m/sec. The AM radio band ranges in frequency from 550 to 1600 kHz. What is the range of wavelengths? The FM band ranges from 88 to 108 MHz (1 MHz = 10^6 Hz). What is its range of wavelengths?

11.4 The range of sound frequencies detectable by the human ear is 20 to 20,000 Hz. What is the range

of wavelengths? The velocity of sound in air is 330 m/sec.

11.5 A rule of thumb for finding the distance where a flash of lightning occurs is to count the number of seconds from the moment one sees the lightning to the moment one hears the thunder. The distance in kilometers is the number of seconds divided by 3. How accurate is this rule? (See problem 11-1.)

11.6 In an experiment designed to measure the velocity of sound waves in copper, a blow is struck at one end of a copper rod. Detectors at the other end measure the time interval between the arrival of the sound pulse through the rod and the arrival of the sound pulse through the air. If the rod is 3 m long and the sound pulse that traveled through the rod arrives 8.01×10^{-3} sec earlier than the sound pulse that traveled through the air, what is the velocity of sound in copper? The velocity of sound in air is 330 m/sec.

(*Answer*: 2775 m/sec.)

11.7 In the wave configuration shown in Fig. 11-7 the length of the string is 4.2 m, the frequency of the wave is 1.2 Hz, and the amplitude is 0.05 m. What is the speed of the wave? Note that $\lambda >>$ amplitude.

(*Answer*: 2.52 m/sec.)

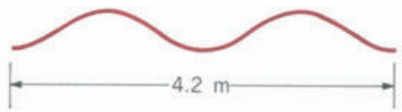


FIGURE 11-7
Problem 11.7.

11.8 The equation of a transverse wave traveling along a very long string is given by

$$y = 6.0 \sin (0.020\pi x + 4.0\pi t)$$

where x and y are expressed in centimeters and t in seconds. Find (a) the amplitude, (b) the wavelength, (c) the frequency, (d) the speed of propagation, (e) the direction of propagation of the wave, and (f) the maximum transverse speed of a particle in the string.

11.10 The equation of a traveling wave in a long stretched string is $y = 10^2 \sin (32t - 4x)$m, where x is in meters and t is in seconds. What is the velocity of the wave in the string?

Write the equation of motion of a traveling wave for the string in problem 11-7 using the same amplitude. Assume that the wave travels in the positive x direction.

11.11 Write the equation for a wave traveling in the negative direction along the x axis and having an amplitude of 0.010 m, a frequency 550 Hz, and a speed 330 m/sec.

(*Answer*: $y = 0.010 \sin (3.33\pi x + 1100\pi t)$ m.)

11.12 Consider the situation illustrated in Fig. 11-3. Let the spring constant $k = 200$ N/m and the mass of the block $m = 2$ kg. The block is given an upward initial kick to start it oscillating. Sketch the shape of the string at $t = 0.15$ sec, $t = 0.3$ sec, $t = 0.45$ sec, $t = 0.6$ sec, and $t = 1.2$ sec.

11.13 A triangular pulse of height 0.5 m and length 2 m moves along the positive x direction on a string with velocity 12 m/sec (see Fig. 11-8). At $t = 0$ the pulse is between $x_1 = 1$ m and $x_2 = 3$ m. Plot the transverse velocity of point x_2 as a function of time.

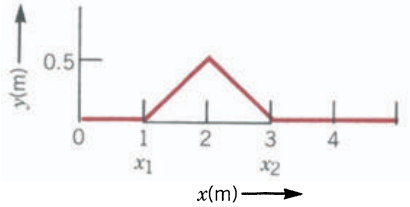


FIGURE 11-8
Problem 11.13.

11.14 A sinusoidal traveling wave on a string has a frequency $\nu = 15$ Hz and a velocity $v = 7.5$ m/sec. (a) How far apart are two points whose transverse displacements are phase-shifted by 30°? (b) At a particular point on the string, what is the phase difference between two displacements occurring 0.05 sec apart?

(*Answer*: (a) 4.17×10^{-2} m, (b) 270°.)

11.15 A block connected to a rigid rod is raised from some initial position $y = 0$ with constant velocity $v_y = 20$ m/sec for 0.4 sec. The block is then suddenly lowered to its initial position and then raised again with the same velocity for the same amount of time. The cycle is repeated indefinitely. A long string under tension is attached to the side of the block (see Fig. 11-9). Let the wave velocity in the string be 5 m/sec. (a) Sketch the shape of the string, using approximate dimensions for the x and y co-

ordinates, at $t = 0.2$ sec, $t = 0.4$ sec and $t = 1.2$ sec. (b) What is the spacing of the pulses on the string?

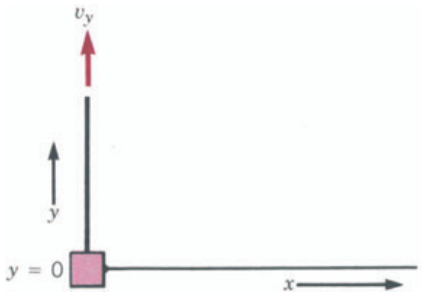


FIGURE 11-9
Problem 11.15.

11.16 A laser produces light pulses of energy 5 J and duration 2×10^{-9} sec. The width of the beam is 1 mm^2. What is the intensity of the laser light?

11.17 A sinusoidal traveling wave of amplitude 2 cm and wavelength 50 cm moves along a string with velocity $v = 6$ m/sec. (a) What is the maximum transverse velocity of the particles in the string? (b) What is the maximum transverse acceleration of the particles in the string?

(*Answer*: (a) 1.51 m/sec, (b) 113.7 m/sec^2.)

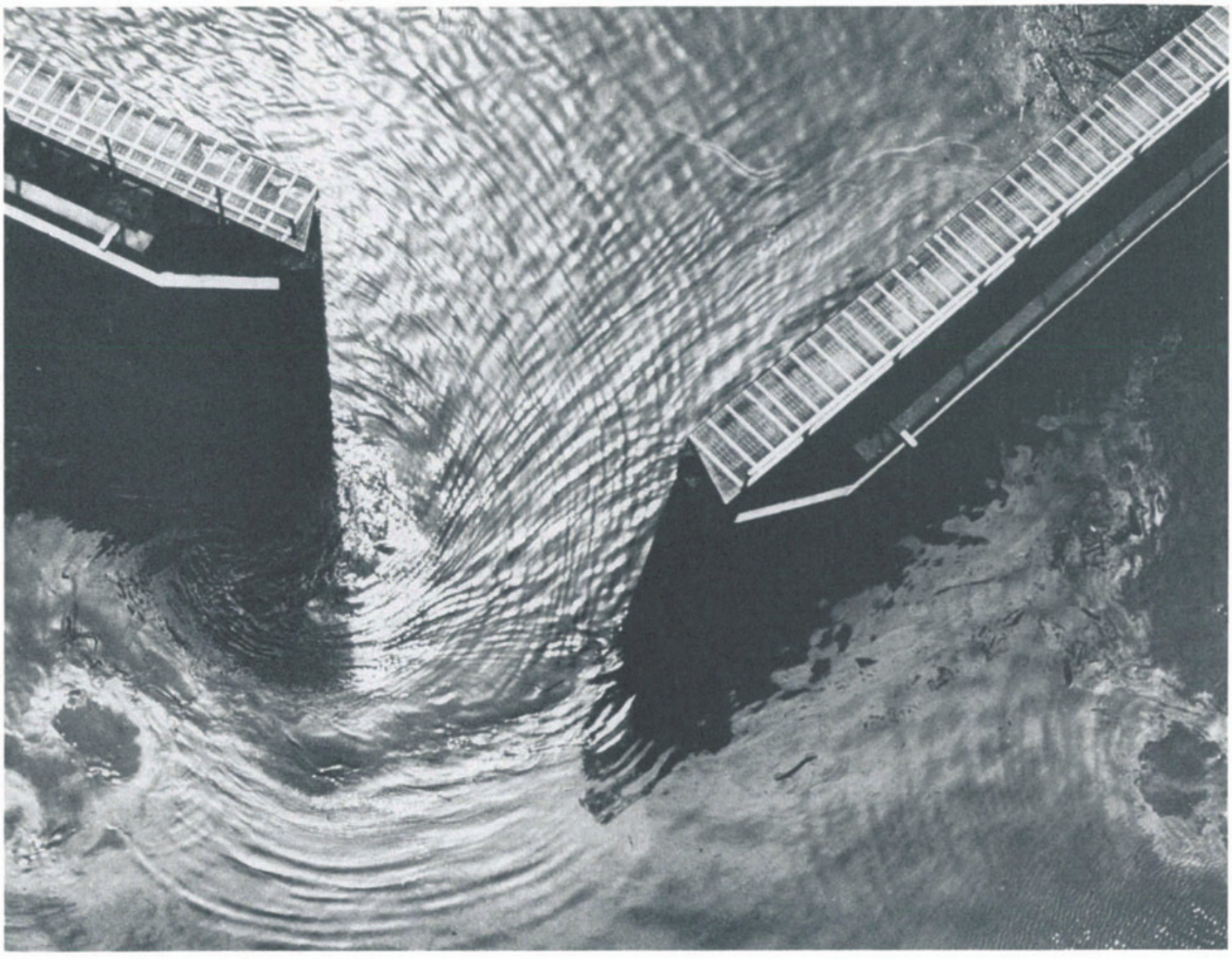
11.18 If the string of problem 11.17 is 30 m long and has a mass of 6 kg, what is the power transmitted by the wave?

11.19 Suppose that during the transmission of waves through a string the frequency is suddenly doubled while maintaining the same amplitude and velocity of propagation, what will happen to the magnitude of the power transmitted? Suppose that the amplitude is doubled while holding the frequency fixed, what will happen to the power?

11.20 Consider a point source emitting waves in all directions. If the medium through which the waves propagate is isotropic, the velocity of propagation will be the same in all directions. As a result, points in the medium where the wave has a certain phase are equidistant from the source and therefore lie on a spherical shell with the source at the center. These waves are called spherical waves. Consider waves emitted from a 5-W source in a nonabsorbing medium. (a) What is the intensity of the waves 1 m away from the source? (b) What should the power of the source be in order that the intensity of the waves 1 m away be the same as that of the laser in problem 11.16?

11.21 A point source emits spherical waves in a nonabsorbing medium (see problem 11.20.) The intensity at some unknown distance from the source is 25 W/m^2. The intensity at some point 10 m farther away from the source is 16 W/m^2. (a) How far is the source from the first point? (b) What is the power output of the source?

(*Answer*: (a) 40 m, (b) 503 kW.)



CHAPTER 12

Interference of Waves

12.1 INTRODUCTION

In the preceding chapter we introduced the concept of waves as a periodic disturbance of a medium. We used familiar concepts such as waves on a string or in water to illustrate the phenomena. The general equation for a traveling wave was derived as was the concept of phase shift. In this chapter we will start with the behavior of two waves when they come together and the effect produced by their relative phase. We will first discuss this phenomenon with waves in water and then extend it to light waves. At this point we will assume, as did early investigators, that air could be the substance in which light-wave motion occurs. However, we will show in a later chapter that light waves do not require some substance or medium to support them.

12.2 THE SUPERPOSITION PRINCIPLE

One of the fundamental principles governing the propagation of waves is called the *superposition principle*. What happens when two different waves meet? Experiments show that waves can move through the same region of space independently and, as a result, when they meet the resultant wave is simply the algebraic sum of the individual waves. (The superposition principle does not hold for waves of very large amplitude in deformable media.) Figure 12-1 shows what happens when a square pulse in the string meets a triangular pulse moving in the opposite direction. In Fig. 12-1*a* two positive waves approach each other. The resulting displacement of the string is the addition of the displacements that each pulse would have produced in the absence of the other (Fig. 12-1*b*). After meeting, the waves move on unaffected (Fig. 12-1*c*). If one wave pulse is positive and the other is negative, as in Fig. 12-1*d–f*, the negatively directed pulse subtracts from the positive pulse. These two examples show that the resultant pulse while the two waves are passing one another is the algebraic sum of the two waves.

The superposition principle leads to a wave phenomenon known as *interference*. Suppose two waves with the same wavelength, velocity, and amplitude, but from different sources, travel together in the same direction. What will be the amplitude of the resulting wave? Figure 12-2*a* shows that if they are in phase, the total amplitude at any point will be the simple sum of the two. If they are out of phase by one-half wavelength, the resulting amplitude will be zero (see Fig. 12-2*b*). The first case is called *constructive interference*, and the second is called *destructive interference*. It is important to note that if either wave is shifted to the right or to the left by a whole wavelength, the situation is unchanged. However, if the shift is by only a half wavelength the situation is reversed; that is, constructive interference becomes destructive and vice versa.

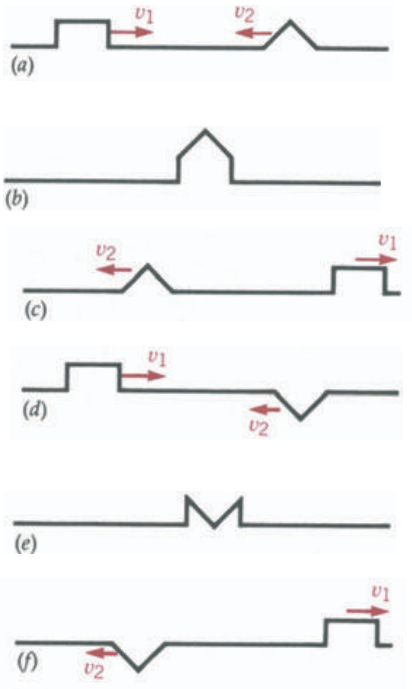


FIGURE 12-1
Superposition principle: (*a*) A square positive pulse and a triangular positive pulse in a string traveling in opposite directions. (*b*) The two pulses meet, and the resulting displacement of the string is the sum of the displacements that each pulse would have produced in the absence of the other. (*c*) After meeting, the pulses move away unaffected. (*d*) A positive square pulse and a negative triangular pulse moving in opposite directions. (*e*) The pulses meet, the resulting displacement is the difference (algebraic sum) of the two pulses. (*f*) After meeting, the two pulses move away unaffected.

12.3 INTERFERENCE FROM TWO SOURCES

If pebbles or water drops fall at regular intervals in still water, a pattern of circular waves, each constantly increasing in radius, will be established. If a similar situation with the same frequency of disturbance occurs nearby, the circular traveling waves will cross one another, producing an interference pattern. Note that the restriction of the previous section that the waves travel together in the same direction has now been removed. At some points the interference will be constructive, and at others it will be destructive. A simple laboratory demonstration of this is shown in Fig. 12-3. This is a photograph of what is called a *ripple tank.* This is a tray of water illuminated from below.

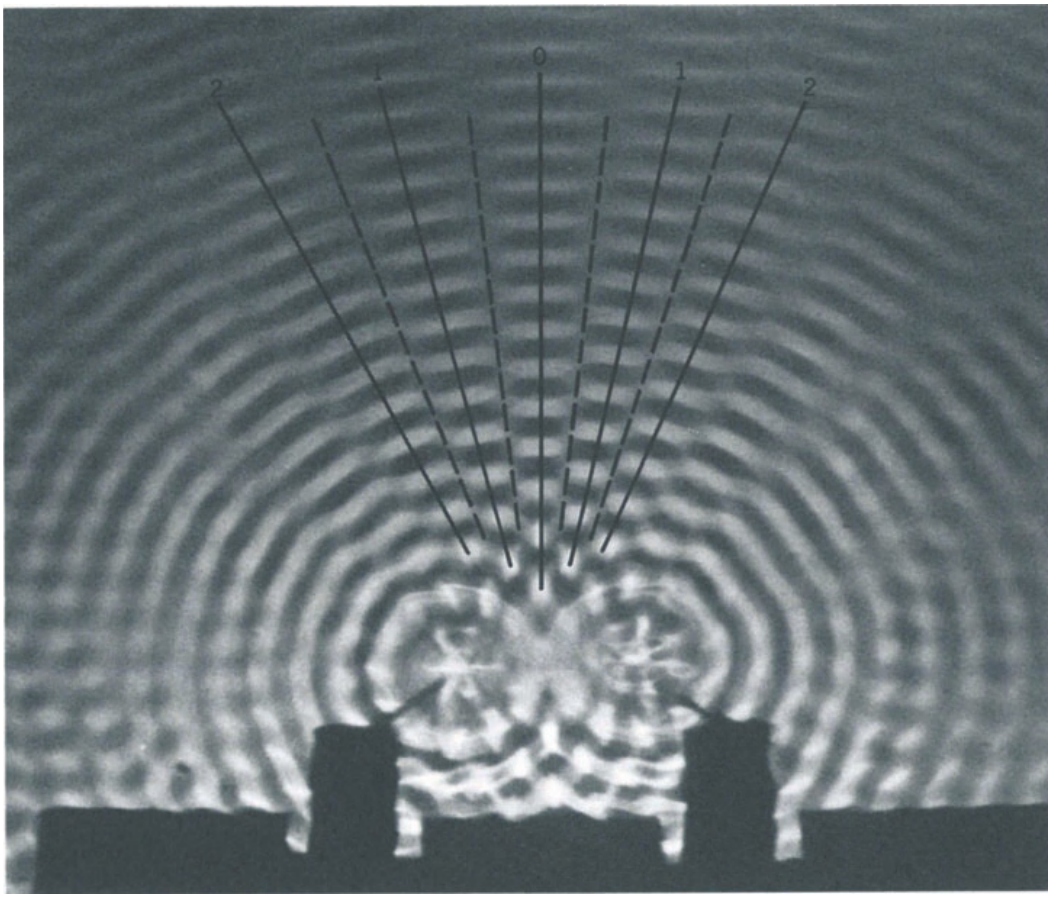


FIGURE 12-3
Ripple-tank demonstration of the phenomenon of interference. Two vibrators striking the water surface in the tank at periodic intervals produce two circular wave patterns. As the two wave patterns cross each other, an interference pattern results. Along the solid lines the waves from the two sources interfere constructively, that is, the displacement of the water is large. Along the dashed lines the interference is destructive, and the displacement of the water is zero. There are additional paths of constructive and destructive interference that have not been marked with lines.

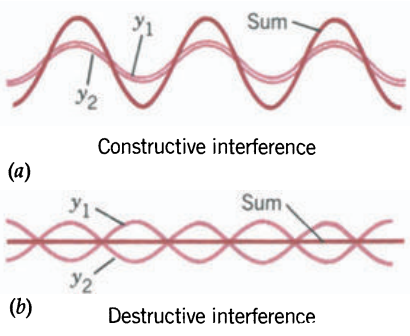


FIGURE 12-2
Interference of two waves: (*a*) *Constructive interference* of two waves traveling together in phase, that is, with the amplitudes coinciding, resulting in a wave with an amplitude that is the sum of the amplitudes of the individual waves. (*b*) *Destructive interference* of two waves of equal amplitude traveling together with a phase difference of one-half wavelength resulting in a wave of zero amplitude.

Instead of having drops of water falling, two vibrators are placed in the tray and the frequency and amplitude of vibration can be controlled more accurately than can that of falling drops of water. The centers of the circles at the bottom are the locations at which the vibrators are located. Some of the paths along which constructive interference occurs are indicated with solid lines. The dashed lines between the paths labeled 0 and 1 are regions of destructive interference.

Figure 12-4 is a schematic representation of the photograph in Fig. 12-3 in which the wave crests, represented by the circular lines, proceed outward from the two sources S_1 and S_2. It is seen in this figure that the phenomenon of Fig. 12-3 is a purely geometric one that can be reproduced on paper with a compass and ruler. The distance between the crests is the wavelength λ. The troughs are located halfway between the crests. As in the photograph, the solid lines labeled 0, 1, 2, . . . represent the paths of constructive interference.

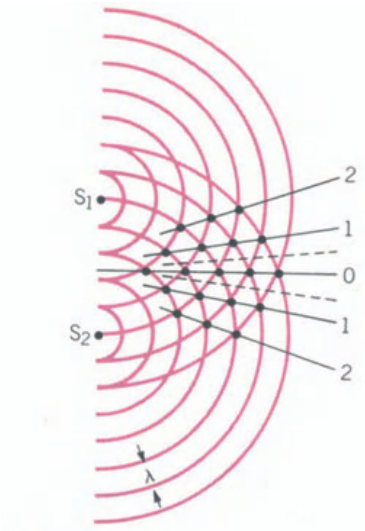


FIGURE 12-4
Geometric representation of the photograph in Fig. 12-3. The crests of the water waves are represented by circular lines whose centers coincide with the location of the vibrators. The distance between adjacent crests is the wavelength. The troughs are halfway between the circular lines. The crests from the two sources coincide along the solid lines labeled 0, 1, 2, where constructive interference occurs. Along the dashed lines, the crests from one source coincide with the troughs from the other, resulting in paths of destructive interference.

We notice in this figure that the paths of constructive interference are symmetric about the line labeled 0. Therefore, we need to consider only one group either above or below the 0 line. We will consider the ones above, knowing that the results will be the same for the ones below. Constructive interference occurs along these paths because the crests from the two sources coincide and add to the disturbance of the water. This is the criterion presented in the previous section for constructive interference. Along the path labeled 0 in Fig. 12-4, the first crest from S_1 coincides with the first crest from S_2, the second crest from S_1 with the second from S_2, and so on. Along the path labeled 1, the first crest from S_2 coincides with the second crest from S_1, the second from S_2 with the third from S_1, and so on. We may view these waves as having traveled for some distance from their sources along their respective paths. When two waves travel in the same medium, the difference in the distances traveled by them from their respective sources to a common point is called the *path difference*. Keeping in mind that the separation between successive crests is the wavelength λ, we can now state the criterion for constructive interference as follows: When waves from two sources are emitted in phase, *constructive interference occurs when the path difference is zero, or one wavelength, or an integral multiple of wavelengths* $n\lambda$. This can be formally shown with the trigonometric relation for the sum of sines,

$$\sin a + \sin b = 2 \sin \frac{1}{2}(a + b) \cos \frac{1}{2}(a - b) \qquad (12.1)$$

Let us consider a point P whose distances from S_1 and S_2 are x_1 and x_2, respectively, as shown in Fig. 12-5. From Eq. 11.5′, the wave y_1 from S_1 and the wave y_2 from S_2 at P are

$$y_1 = A \sin (kx_1 - \omega t)$$

$$y_2 = A \sin (kx_2 - \omega t)$$

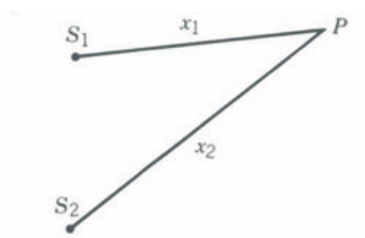


FIGURE 12-5
Arbitrary point P in Fig. 12-3 or 12-4.

From the superposition principle, the resulting wave will be

$$\begin{aligned} y &= y_1 + y_2 \\ &= A\,[\sin(kx_1 - \omega t) + \sin(kx_2 - \omega t)] \end{aligned} \tag{12.2}$$

Now let the path difference be an integral multiple of the wavelength, that is,

$$x_2 - x_1 = n\lambda, \qquad \text{where } n = 0, \pm 1, \pm 2, \ldots$$

and

$$x_2 = x_1 + n\lambda$$

or, because from Eq. 11.12

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

$$x_2 = x_1 + \frac{2\pi n}{k},$$

then Eq. 12.2 becomes

$$y = A\,[\sin(kx_1 - \omega t) + \sin(kx_1 + 2\pi n - \omega t)] \tag{12.3}$$

Using Eq. 12.1 gives

$$y = 2A \sin \frac{1}{2}(kx_1 - \omega t + kx_1 - \omega t + 2\pi n) \cos \frac{1}{2}(2\pi n)$$

$$y = 2A \sin(kx_1 - \omega t) \tag{12.4}$$

In the last step we let $kx_1 - \omega t = \theta$ and used the fact that $\sin(\theta + n\pi) = -\sin\theta$ and $\cos n\pi = -1$ if n is an odd integer, and $\sin(\theta + n\pi) = \sin\theta$ and $\cos n\pi = 1$ if n is an even integer. Equation 12.4 formally verifies that when the path difference is an integral multiple of the wavelength, the resulting wave has an amplitude that is twice that of y_1 or y_2 at that point.

Returning to Fig. 12-4, we note that the dashed lines, representing paths where destructive interference occurs, correspond to points whose distances from S_1 and S_2 differ by $1/2\ \lambda$, or (from Eq. 11.12) $x_2 - x_1 = \lambda/2 = \pi/k$. The resulting wave in this case will be

$$\begin{aligned} y &= y_1 + y_2 \\ &= A[\sin(kx_1 - \omega t) + \sin(kx_1 + \pi - \omega t)] \end{aligned}$$

Using Eq. 12.1, we obtain

$$y = 2A \sin \frac{1}{2}(kx_1 - \omega t + kx_1 - \omega t + \pi) \cos \frac{1}{2}\pi$$

But $\cos \pi/2 = 0$, and therefore $y = y_1 + y_2 = 0$; this by definition is destructive interference. The same result is obtained if $x_2 - x_1 = \frac{3}{2}\lambda$ or $\frac{5}{2}\lambda$ and so on.

12.4 DOUBLE SLIT INTERFERENCE OF LIGHT

If a series of either plane waves or large radius spherical waves strike a barrier with a small opening, circular waves are propagated beyond the opening as if the opening were a point source. The enlarging circumference of these waves is called a *wave front*. Figure 12-6 illustrates this propagation with water

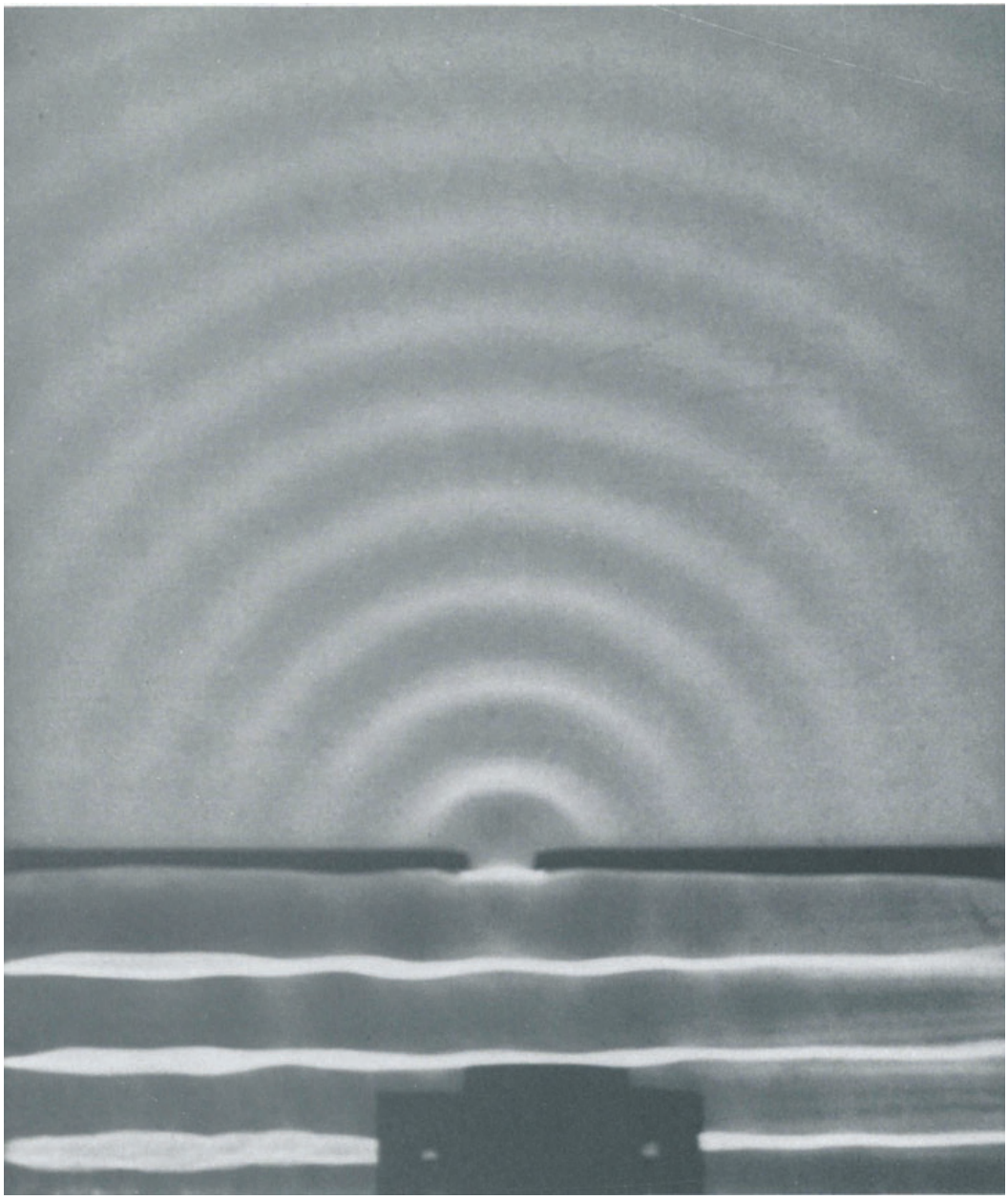


FIGURE 12-6
Huygens' principle. Parallel wave fronts in the ripple tank strike a barrier with a small opening. The opening becomes a source of circular waves.

Christian Huygens (1627–1675)

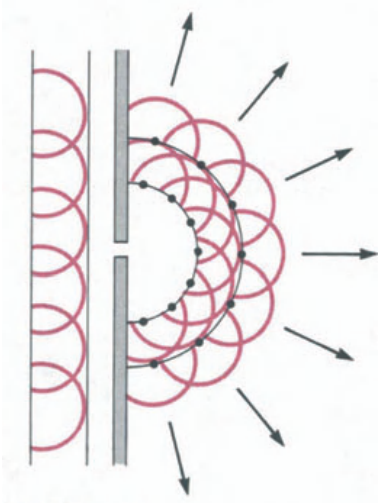


FIGURE 12-7
Geometric representation of the phenomenon in Fig. 12-6. The opening in the barrier produces a circular wave front. Each point in that wave front produces secondary circular wavelets, and the new propagated wave front is the tangent to these secondary wavelets.

waves in a still tank. This phenomenon illustrates what is known as *Huygens' principle* after the Dutch physicist, Christian Huygens (1629–1695). It is more generally stated that every point on a wave front can be considered as a source of secondary wavelets that spread out in all directions with a speed and wavelength equal to those of the propagating wave. The newly propagated wave front is the tangent to these secondary wavelets. Thus, the spherical wavelet passing through the opening in the barrier of Fig. 12-6 produces a secondary wave front that itself produces other wavelets as shown in Fig. 12-7.

Huygens demonstrated that the known facts about the propagation of light could be explained by using his principle. It was many years, however, before light was accepted as a wave phenomenon. This came about when the English physician Thomas Young (1773–1829) performed the first successful experiment that exhibited the interference of light in 1801. The nature of light is discussed in more detail in Chapter 16. For now, we mention that visible light has a wavelength that ranges from about 4×10^{-7} m to 7×10^{-7} m, where the lower values appear to us as violet and the higher values as red. A unit of length often used in specifying the wavelength of light is the Ångstrom, with symbol Å; $1\ \text{Å} = 10^{-8}$ cm $= 10^{-10}$ m, and therefore the wavelength of visible light ranges from about 4000 Å to 7000 Å.

Thomas Young (1773–1829).

Figure 12-8*a* illustrates a schematic arrangement, similar to the one used by Young, to determine the phenomenon of interference with light. A mono-

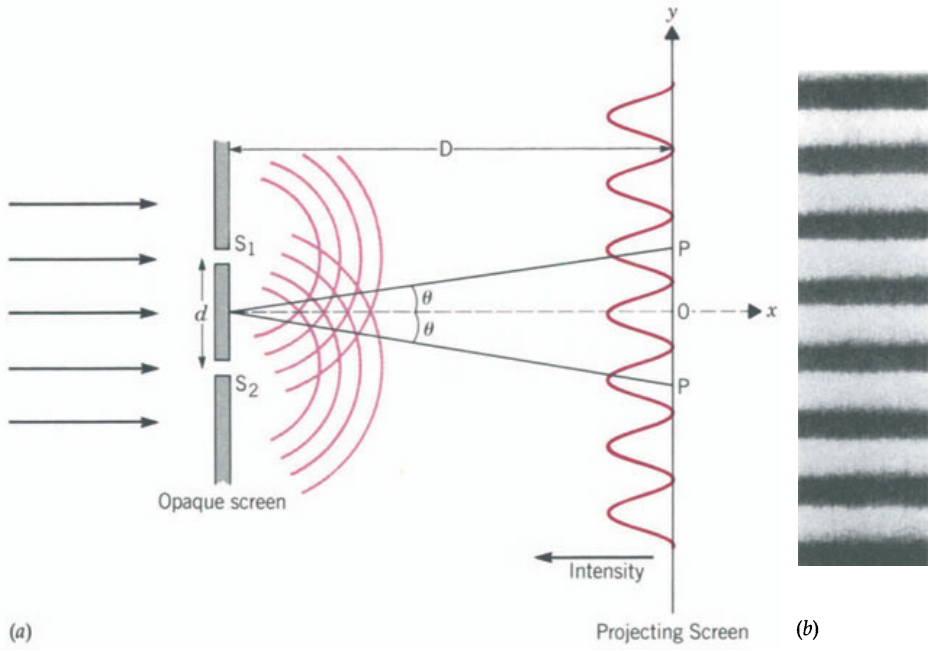


FIGURE 12-8

(*a*) Schematic of Young's double slit interference of light experiment. Light passing through two small slits S_1 and S_2 in an opaque screen produce an interference pattern on a projecting screen to the right of the slits. (*b*) Photograph of the interference pattern on the projecting screen showing a pattern of alternating bright and dark fringes. (*Source:* Cagnet et al., *Atlas of Optical Phenomena*, Springer-Verlag, New York, 1971.)

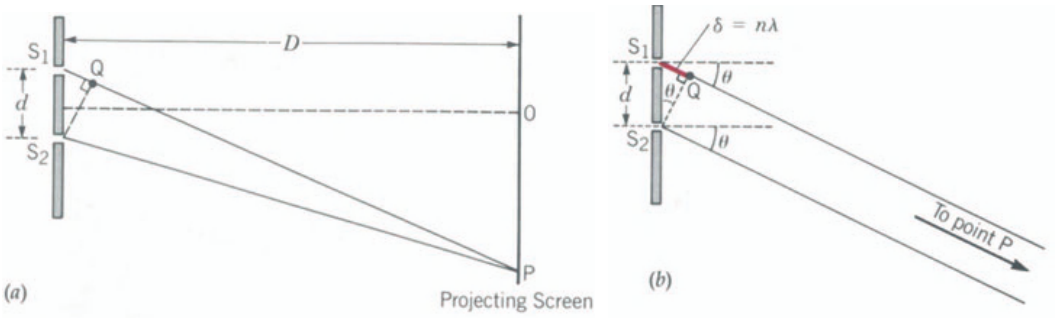


FIGURE 12-9
(*a*) Geometric construct of lines from each of the slits in Fig. 12-8 to a point P on the projecting screen where constructive interference is observed. (*b*) When $D >> d$, the two lines are approximately parallel. The difference between the two paths traveled by the light from S_1 and S_2 is $\delta = n\lambda$.

chromatic light source (a single wavelength) shines on an opaque screen with two narrow slits S_1 and S_2. According to Huygens' principle, these two slits become point sources of light. If we let the light that passes through the slits fall on a screen, we will observe a pattern of bright and dark lines that indicates constructive and destructive interference. A plot of the light intensity on the screen is schematically represented in Fig. 12-8*a*. Figure 12-8*b* is a photograph of the interference pattern observed on the screen.

We can use the principles developed in the previous section to find expressions for the position of the interference maxima and minima. Figure 12-9*a* shows a geometric construct of lines drawn from each of the two slits to a point P of constructive interference. As will become evident soon, for the interference pattern to be easily observable, the separation between the slits, d, cannot be much greater than the wavelength. This implies that in the case of light d might be a few microns (10^{-6} m) whereas the distance, D, between the slits and the screen could be several centimeters, that is, $D >> d$. We therefore conclude that the two lines S_1P and S_2P are almost parallel. The situation of Fig. 12-9*a* can thus be approximated by that shown in Fig. 12-9*b*. In Fig. 12-9*b*, a perpendicular has been dropped from slit S_2 to point Q on the line S_1P. Angles indicated by θ are equal because their sides are mutually perpendicular. The extra distance S_1Q traveled by the wave from slit S_1 is the path difference, with symbol δ, between the two waves when they arrive at P. It is seen from triangle S_1S_2Q that

$$\delta = d \sin \theta \tag{12.5}$$

As we have shown in the preceding section, constructive interference at point P can occur only if this path difference is an integral multiple of wavelengths $n\lambda$. The condition for *constructive interference* in this case becomes

$$d \sin \theta = n\lambda, \quad \text{where } n = 0, 1, 2, 3, \ldots \tag{12.6}$$

Condition for constructive interference

$d \sin \theta = n\lambda$

Similarly, destructive interference will occur when the path difference is 1/2 λ, 3/2 λ, 5/2 λ, or in general $[(2n + 1)/2]\ \lambda$, where $n = 0, 1, 2, 3, \ldots$.

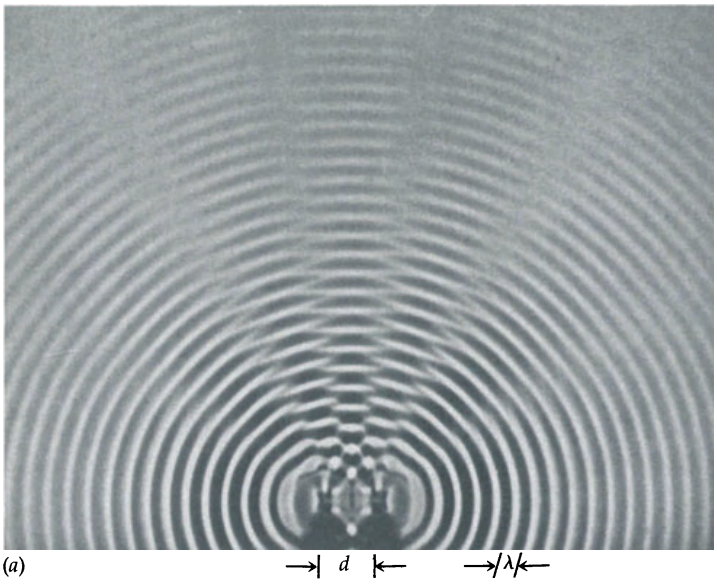


FIGURE 12-10
Interference patterns in the ripple tank of Fig. 12-3. (*a*) and (*b*) If λ is kept constant, the angular separation between the interference maxima decreases when the separation d between the sources is increased. (*a*) and (*c*) Both d and λ are doubled, the angular separation between interference maxima remains the same.

The condition for *destructive interference* can be stated as

$$d \sin \theta = \frac{2n + 1}{2} \lambda, \quad \text{where } n = 0, 1, 2, 3, \ldots \tag{12.7}$$

Condition for destructive interference

$$d \sin \theta = \frac{2n + 1}{2} \lambda$$

We recall that as θ increases, $\sin \theta$ also increases. We conclude from Eq. 12.6 that for a constant λ, the angular separation between successive interference maxima decreases with increasing spacing between the slits, d. A simple laboratory demonstration of this is shown in Fig. 12-10. These are photographs of a ripple tank like the one used in the demonstration of Fig. 12-3. A comparison of Fig. 12-10*a* and 12-10*b* shows that if λ is kept constant, the angular separation between interference maxima decreases as the separation between the sources increases. Similarly a comparison of Fig. 12-10*a* and 12-10*c* indicates that when the separation between the sources is doubled, the wave-

length must also be doubled to have the same angular separation. It should also be clear from Eq. 12.6 that if $d >> \lambda$, a large number of interference maxima will occur within a small angle and, as a result, the pattern will be difficult to observe. To illustrate the point, let us assume that $d = 1000\ \lambda$. If we consider a small angle θ, for example, $\theta = 1°$, we can solve for n in Eq. 12.6, and we will get

$$1000\ \lambda \sin 1° = n\ \lambda$$

therefore

$$n = 1000 \sin 1° = 1000 \times 0.017 = 17$$

Seventeen intensity maxima will be formed within an angle of 1°. It is clear that unless the distance from the slits to the screen is very large, the interference pattern will not be observable.

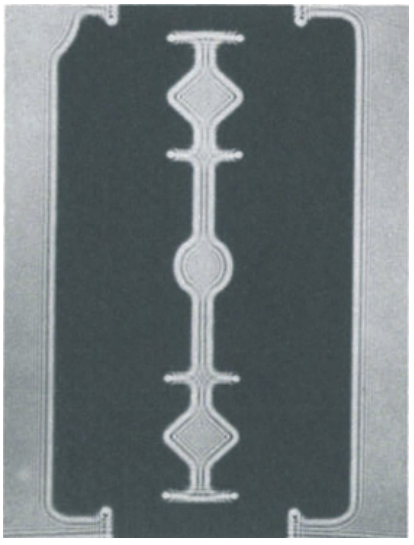
Although we will not derive the relation, it should be mentioned at this point that a derivation similar to that of the two-slits can be made for multiple slits of the same width and the same separation distance: Eqs. 12.6 and 12.7 also apply for multiple slits. An opaque piece of glass with multiple slits is called a *diffraction grating*.

12.5 SINGLE SLIT DIFFRACTION

In the preceding discussion of double slit interference of light, we assumed that the two openings in the opaque screen acted as point sources. We will see that this assumption is correct if the size of the openings is smaller than the wavelength. If the opening is greater than the wavelength but of a size comparable to a few wavelengths, then light waves passing through different portions of a single slit will interfere with each other giving rise to a phenomenon known as *single slit diffraction*.

The method used in Section 12.4 can be employed to analyze the diffraction of light by a single slit. Figure 12.11 illustrates schematically the passage of individual wavelets through a single slit of width d. The pattern seen on a screen to the right of the slit appears with a central bright maximum with alternating bright and dark fringes on either side, with the bright fringes diminishing in intensity as the angle from the normal increases. This is illustrated in Figs. 12-12*a* and *b*.

In our treatment, we will assume that the size of the slit is much smaller than the distance from the slit to the screen. Under this condition, the lines of propagation of the waves emanating from different points in the slit are approximately parallel. In Figs. 12-11*a* and 12-12*a* all forward-directed waves strike the screen in phase, because the path difference is zero. These waves give rise to the central maximum. In Fig. 12-11*b*, wave A has a path difference of $\lambda/2$ from wave B and interferes destructively. This destructive interference effect occurs across the entire slit because any wave slightly below A will interfere destructively with the wave at the same distance below B and so



Photograph illustrating the diffraction of light by the fine edges of a razor blade.

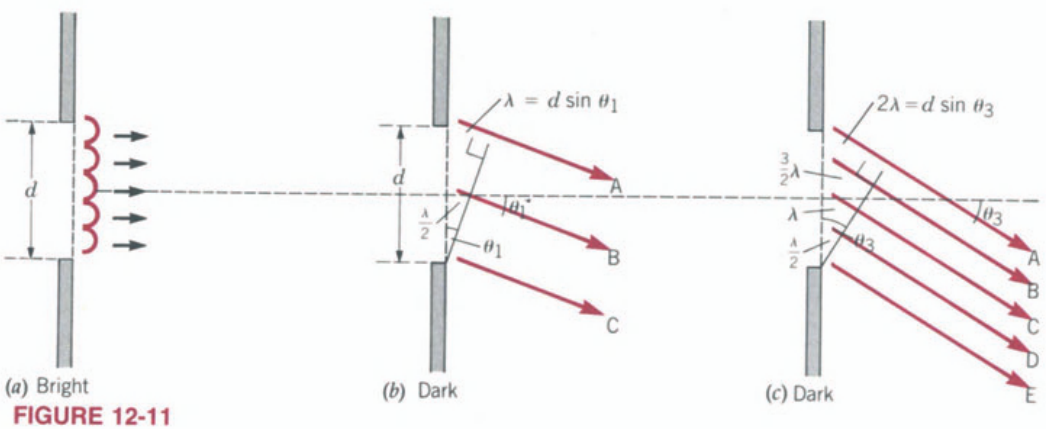


FIGURE 12-11

Single slit diffraction. As light passes through the slit, each point in the slit becomes a secondary source of light. (*a*) All forward-directed waves arrive at the screen in phase, giving rise to the central maximum of Fig. 12-12. (*b*) Waves from the upper half of the slit interfere destructively with those from the lower half, giving rise the first diffraction minimum in Fig. 12-12. (*c*) Waves between A and B interfere destructively with waves between B and C, those between C and D also interfere destructively with those between D and E. This results in the second interference minimum in Fig. 12-12.

on. Therefore, a dark band will appear on the screen for angle θ_1 where

$$\sin \theta_1 = \frac{\lambda}{d} \tag{12.8}$$

At angle θ_3 in Fig. 12-11*c*, the situation is similar to that of Fig. 12-11*b* in that all waves between A and B will destructively interfere with corresponding waves between B and C and all the waves between C and D interfere destructively with those between D and E. Thus, a dark band is observed for the condition $\sin \theta_3 = 2\lambda/d$. Somewhere in between these two minima a

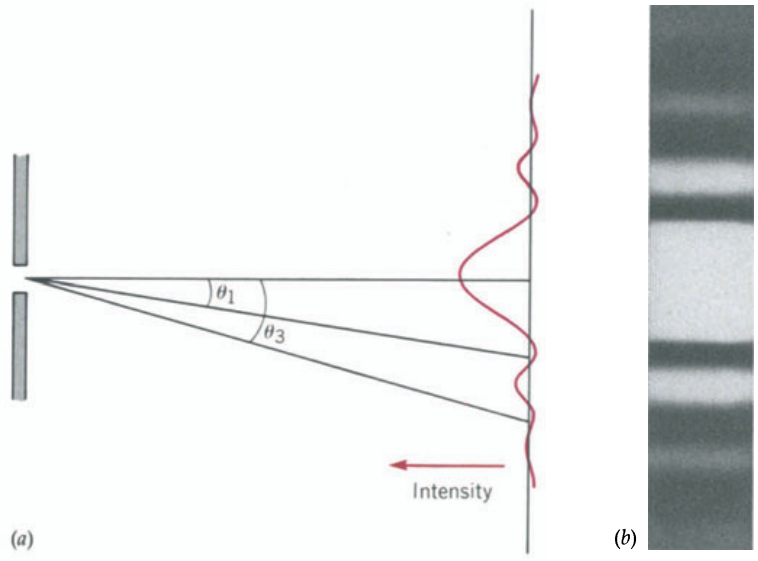
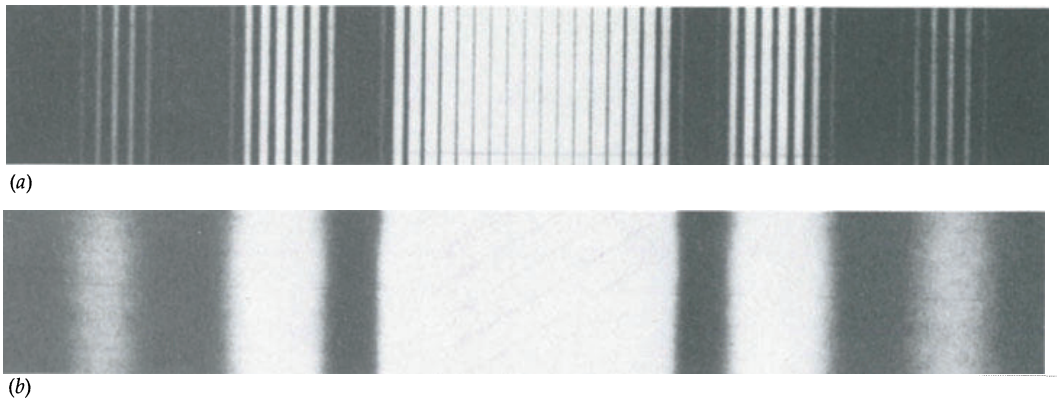


FIGURE 12-12

(*a*) Schematic representation of the interference pattern produced on a screen by light passing through a slit (*b*) Photograph of the interference pattern observed on the screen. (*Source:* Cagnet et al., *Atlas of Optical Phenomena*. New York, Springer-Verlag, 1971.)



(*a*)

(*b*)

FIGURE 12-13

(*a*) Fringe pattern obtained with light shined through a double slit when the size of the slits is greater than but of the same order of magnitude as the wavelength. The pattern is essentially a double slit interference pattern, similar to that of figure 12-8*b*, "modulated" by the diffraction pattern of Fig. 12-12. (*b*) Single slit diffraction pattern obtained when one of the slits is blocked.

maximum will be observed. We can generalize this result by stating that the diffraction minima occur for angles satisfying the relation

$$\sin \theta = n \frac{\lambda}{d}, \quad \text{where } n = 1, 2, 3, \ldots \tag{12.9}$$

$\sin \theta = n \frac{\lambda}{d}$

From Eq. 12.8, we can see that if $d = \lambda$, then $\sin \theta_1 = 1$ and $\theta_1 = 90°$, which means that the central maximum spreads over the entire screen. The same will be true if $d < \lambda$. This is what occurs when the screen is illuminated with a point source. We see, therefore, that an opening can be considered a point source when its size is equal to or smaller than the wavelength. Another important conclusion can be drawn from Eq. 12.9. This equation, which gives the position of the *minima* of single slit diffraction, is identical to Eq. 12.6, which gives the angular position of the interference *maxima* for the double slit case. We found then that if the separation between the slits, d, was much greater than the wavelength, the interference pattern would be difficult to observe. The same conclusion holds here if the size of the slit is much greater than λ.

The patterns of light on a screen from single and double slits are quite distinctive, as shown in Fig. 12-13.

12.6 RESOLVING POWER

We have seen in the preceding section that light effectively bends around corners. That is, when light shines on a slit the edges of the slit are not simply shadowed on the screen; there are also small bright lines within the shadow at angles away from the normal. This has profound implications for our determination of the location of a particle. We will show here that because

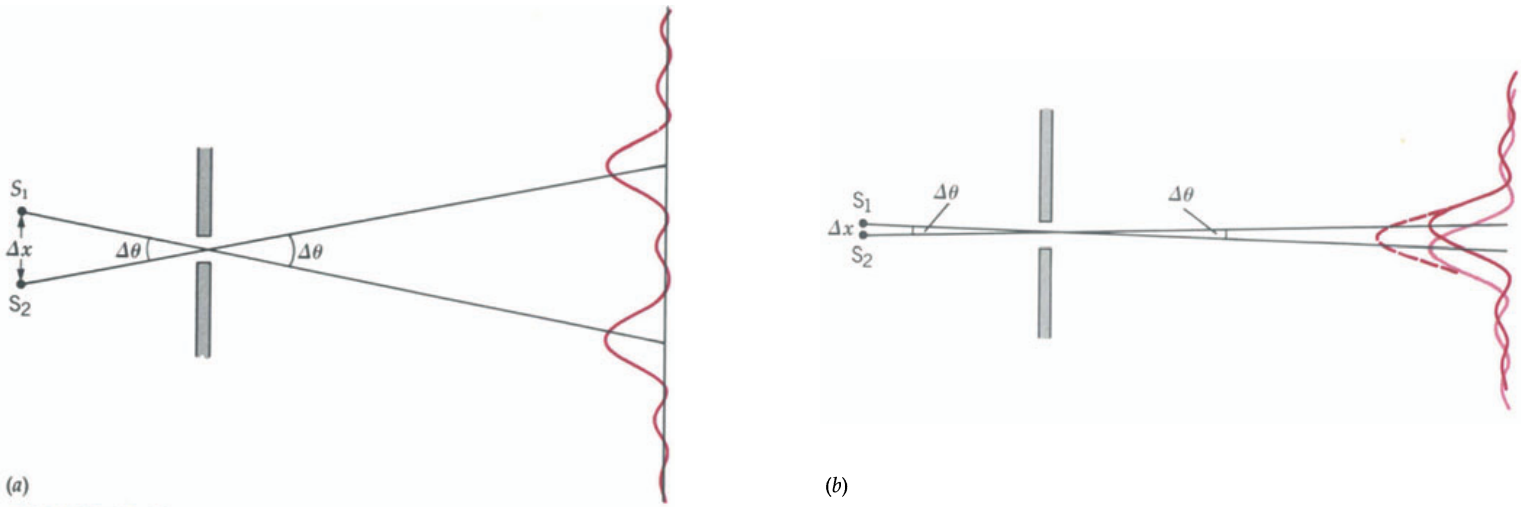


FIGURE 12-14

Two light sources shining on a single slit give rise to two individual diffraction patterns. (*a*) The sources are sufficiently apart and the two central maxima are well resolved. (*b*) The sources are too close together and as a result the two sources cannot be resolved. The two central maxima overlap to the extent that a single intensity maximum (dashed line) appears on the screen.

of diffraction effects the accuracy of the determination of the precise location of a particle depends on the wavelength used to "look" at it; the smaller the wavelength employed, the greater the accuracy. In Chapter 19 we will show that the smaller the wavelength of the light used, the greater the momentum transferred to the particle being examined and, correspondingly, the greater the disturbance of its position in space. Thus, because of these conflicting effects there is a limit to which we may simultaneously know the location and momentum of a particle. This limit is known as the *Uncertainty Principle.* For now, we will show that the determination of location depends on the wavelength; this is called the *resolving power.*

Suppose we have two sources of light, S_1 and S_2, shining on a single slit. If they are sufficiently far apart we will see on a screen two single slit patterns, each of the type of Fig. 12-12. This is illustrated in Fig. 12-14*a*. The light intensity at any point on the screen will be the sum of the contributions from each source. If we know the distance of the sources from the slit and the distance of the screen from the slit, we can calculate the angle $\Delta\theta$ and the distance Δx between the two sources from the separation of the two central maxima. Suppose that the sources are closer together than in Fig. 12-14*a* and we have the arrangement of Fig. 12-14*b*. We see that the bright central maxima of the diffraction pattern from the two sources overlap so that, because it is the sum that is seen, it is difficult to estimate their distance apart and therefore difficult to know Δx. A rather arbitrary, although practical, criterion used to decide when the two sources S_1 and S_2 are just resolvable (i.e., considered as separate sources) is the coincidence of the central maximum of one of them with the first minimum of the other (see Fig. 12-15). This is known as the

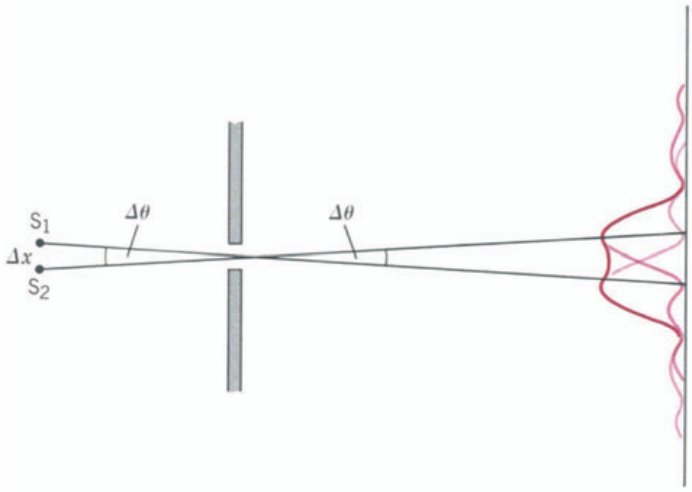


FIGURE 12-15
Rayleigh criterion. The separation between the sources is such that the central maximum produced by one source coincides with the first diffraction minimum from the other. The two sources are barely resolvable.

Rayleigh criterion or the *limit of resolution*. We saw in Eq. 12.8 that the first minimum occurs when $\sin\theta_1 = \lambda/d$. Because this occurs for small angles, $\sin\theta_1 \approx \theta_1$ (in radians) and

$$\Delta\theta = \theta_1 \text{ (in radians)} = \text{limiting angle of resolution} \approx \frac{\lambda}{d} \qquad (12.10)$$

limiting angle of resolution

$$\Delta\theta \approx \frac{\lambda}{d}$$

Thus, for high resolution—that is, small Δx (see Fig. 12-15)—one should have a small λ and a large slit width d.

12.7 X-RAY DIFFRACTION BY CRYSTALS: BRAGG SCATTERING

Waves are scattered or reflected by objects their own size or larger. A wave at the seashore will not be affected by a stick in the water but will be by a large rock or jetty. Atoms have sizes of the order of a few angstroms, 1 to 3×10^{-10} m. Visible light has wavelengths of a few thousand angstroms. Therefore, visible light will not be affected by a single atom. Light is only a small part of the wavelength range of what is called the *electromagnetic spectrum* (Chapter 16). This spectrum ranges between γ-rays and radio waves. The smallest wavelengths that we can conveniently produce are those of X rays, whose wavelengths are about the sizes of atoms. These are produced by bombarding a metal target with high-energy electrons. Their origin will be discussed in Chapter 17.

In 1912, Max von Laue noted that in a crystalline solid the interatomic separation is of the same order of magnitude as the wavelength of X rays. He then showed that the regular, periodic arrangement of atoms in a crystalline solid could be used as a three-dimensional *diffraction grating* (a diffraction grating was defined at the end of Section 12.4). One year later, Sir William Bragg presented a similar but simpler analysis of the problem. We will now present an outline of Bragg's analysis.

Sir William Henry Bragg (1862–1942).

Figure 12-16 is a planar representation of a three-dimensional cubic crystal, that is, a solid in which the atoms are located at the corners of unit cubes. The interatomic separation is d. When X rays strike the crystal at an angle θ with respect to a plane of these atoms, the atoms will scatter them in all directions. We will first limit ourselves to the X rays scattered specularly—that is, X rays for which the angle between the scattered beam and the plane of the crystal is the same as the one between the incident beam and the crystal plane, as shown at points A and B in Fig. 12-16. It is clear from the figure that the path difference between the X rays scattered by atom A and atom B is $2l$ because the ray scattered from atom B must travel that extra distance to rejoin the ray scattered from atom A.

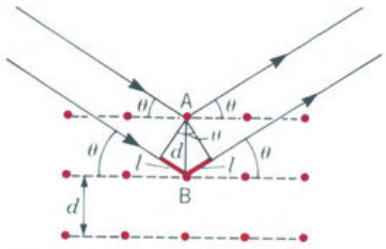


FIGURE 12-16
Planar representation of a three-dimensional cubic crystal of interatomic spacing d. The X rays reflected by the dashed atomic planes will recombine constructively if the Bragg condition is satisfied.

From geometric considerations, $l = d \sin \theta$, and the path difference is $2l = 2d \sin \theta$. By analogy to other interference experiments discussed in previous sections, if this path difference is equal to an integral number of wavelengths, the two beams will add constructively; that is, the radiation reflected by two adjacent layers of atoms will add constructively if

$$2d \sin \theta = n\lambda \quad \text{where } n = 1, 2, 3, \ldots \tag{12.11}$$

Bragg Condition

$$2\,d \sin \theta = n\lambda$$

Equation 12.11 is known as the *Bragg condition*. Obviously, if the waves reflected by the first layer are in phase with those reflected by the second layer, the same will be true for the waves reflected by the second and third, and so on. Thus the condition of Eq. 12.11 guarantees that the radiation reflected by all the atoms in parallel layers of the crystal at the same distance apart will be in phase.

Thus far we have concentrated our attention on the waves that were scattered specularly. Is it possible to have an angle θ', not necessarily equal to the angle of incidence θ, for which the scattered waves recombine constructively? The answer is yes. However, if such an angle exists, it can be shown that a set of atomic planes exists, different from the ones that we have considered, such that with respect to them, the angle of incidence and the angle of scattering are equal. Furthermore, with respect to this new set of atomic planes, the Bragg condition (Eq. 12.11) is satisfied. This situation is illustrated in Fig. 12.17. To have constructive interference for the direction shown in Fig. 12-17, the condition $2d' \sin \phi = n\lambda$ must be satisfied.

We should note that the Bragg equation is very similar to the double slit equation (Eq. 12.6). In the Bragg equation the interatomic spacing plays the

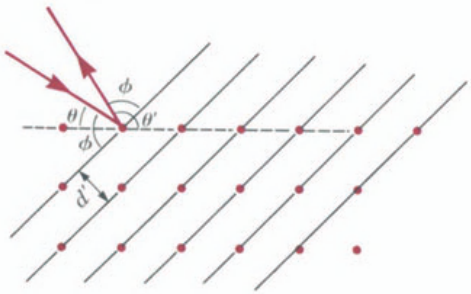


FIGURE 12-17
A different set of atomic planes (indicated by solid lines) in the crystal of Fig. 12-16 can produce constructive reflection if the Bragg condition with respect to those planes is satisfied.

same role as the separation between the slits in the double slit equation. Because the interatomic spacing is of the order of 10^{-10} m, we can see how a crystalline solid can be used as a diffraction grating for radiation of small wavelength, such as X rays or, as will be discussed in Chapter 19, matter waves.

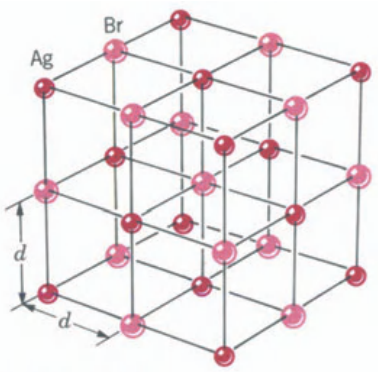


FIGURE 12-18
Example 12-1.

Example 12-1

The crystal structure of silver bromide (AgBr) is represented in Fig. 12-18. The molecular weight and the density[1] of AgBr are 187.80 g/mole and 6.47 g/cm^3, respectively. (a) Calculate the interatomic separation, d, of the atoms in AgBr. (b) If X rays of wavelength $\lambda = 1.50$ Å are incident on a AgBr crystal, at what angle will the first order ($n = 1$) diffraction maxima occur? Assume that the separation between the atomic planes producing the scattering is the interatomic spacing found in part (a).

Solution

(a) Let us consider a cube of AgBr 1 cm a side. In one row of the cube we have 1 cm/d (cm) atoms. Because we have as many rows of atoms in one plane as there are atoms in a row, we conclude that the number of atoms in one plane of the cube is $1/d \times 1/d$ or $1/d^2$. Finally, we have as many planes of atoms as we have atoms in a row; therefore

$$\text{Number of atoms in 1 cm}^3 = \frac{1 \text{ cm}^3}{d^3}$$

where d is expressed in cm.

The actual number of atoms can be found as follows:

$$N \text{ of atoms/cm}^3 = N \text{ of moles/cm}^3 \times N \text{ of atoms/mole}$$

$$= \frac{6.47 \text{ g/cm}^3}{187.80 \text{ g/mole}} \times 2 \times N_A$$

where N_A is Avogadro's number, 6.02×10^{23} molecules/mole and the factor of 2 is included because there are two atoms (one Ag and one Br) per molecule.

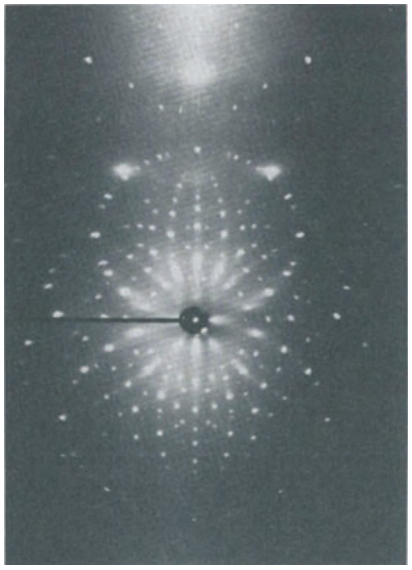
We can now write the following relation

$$\frac{1}{d^3} = \frac{6.47 \text{ g/cm}^3}{187.80 \text{ g/mole}} \times 2 \times 6.02 \times 10^{23} \text{ molecules/mole}$$

$$= 4.15 \times 10^{22} \text{ cm}^{-3}$$

therefore

$$d = 2.89 \times 10^{-8} \text{ cm} = 2.89 \text{ Å}$$



Diffraction pattern produced when X rays are incident on a NaCl crystal. Each dot is produced by a set of atomic planes, satisfying the Bragg condition.

[1]Density is defined as mass per unit volume (see problem 13.1).

(b) We can use the Bragg condition, Eq. 12.11, to find the angle at which the first-order (n = 1) diffraction maximum is observed.

$$2d \sin \theta = \lambda$$

$$\sin \theta = \frac{\lambda}{2d}$$

$$= \frac{1.50 \text{ Å}}{2 \times 2.89 \text{ Å}}$$

$$= 0.26$$

$$\theta = \sin^{-1} 0.26 = 15°$$

12.8 STANDING WAVES

Another interesting phenomenon resulting from the superposition principle is the formation of *standing waves*.

In Chapter 11, when we discussed traveling waves in a string, we implicitly assumed that once the waves were set up at one end, they continue traveling toward the right forever. This is a correct assumption if the string is infinitely long. Consider now that the string is of finite length and the other end is clamped to a rigid support. When the wave disturbances reach the fixed end, they will propagate in the opposite direction. The reflected waves will add to the incident waves according to the superposition principle and, under certain conditions, a standing wave pattern will be formed.

If we assume that the incident waves y_1 travel toward the right in the positive x direction, from Eq. 11.5 we can represent them as

$$y_1 = A \sin (kx - \omega t) \tag{11.5}$$

The reflected waves y_2 will be traveling in the negative x direction and from Eq. 11.7 are given by

$$y_2 = A \sin (kx + \omega t) \tag{11.7}$$

The resulting wave pattern will be

$$y = y_1 + y_2 = A [\sin (kx - \omega t) + \sin (kx + \omega t)]$$

Using the trigonometric relation of Eq. 12.1, we obtain

$$y = 2A \sin kx \cos \omega t \tag{12.12}$$

Equation 12.12 is the equation of a standing wave. We note that, as in the case of a traveling wave, the particles in the string execute simple harmonic motion with the frequency of the wave $\nu = \frac{\omega}{2\pi}$. Unlike the case of the traveling wave, however, the amplitude of oscillation is not the same for all the points (all values of x in Eq. 12.12) in the string. In particular, there are certain points

When this flute player blows on the mouthpiece standing waves are set up inside the flute.

for which the amplitude of oscillation (the coefficient of cos ωt) will be zero. These points, called the *nodes*, are those for which sin $kx = 0$. We can locate these nodes and at the same time find the conditions for standing wave formation by requiring that the value of the wave, y, be zero at the clamped end of the string. If the length of the string is l, then $y(x = l) = 0$. Substituting this in Eq. 12.12, yields

$$0 = 2A \sin kl \cos \omega t \tag{12.13}$$

Because Eq. 12.13 must be satisfied for all times t, we conclude that

$$\sin kl = 0$$

or

$$kl = \pi, 2\pi, 3\pi, \cdots n\pi \tag{12.14}$$

where n is an integer. Note that $kl = -\pi, -2\pi, -3\pi, \ldots$ will also yield a zero value for sin kl. However, these negative values correspond to negative values of k and hence of the wavelength λ and therefore are not physically acceptable. Substituting Eq. 11.12 for k in Eq. 12.14, we obtain

$$\frac{2\pi l}{\lambda} = n\pi$$

or

$$\lambda = \frac{2}{n} l \tag{12.15}$$

This result tells us that the wavelength of the standing wave cannot be arbitrary as was the case with the traveling wave. It can only have the values $2l$, $2/2\,l$, $2/3\,l$, $2/4\,l$, ... A schematic of a photograph of the first few standing wave patterns is shown in Fig. 12-19.

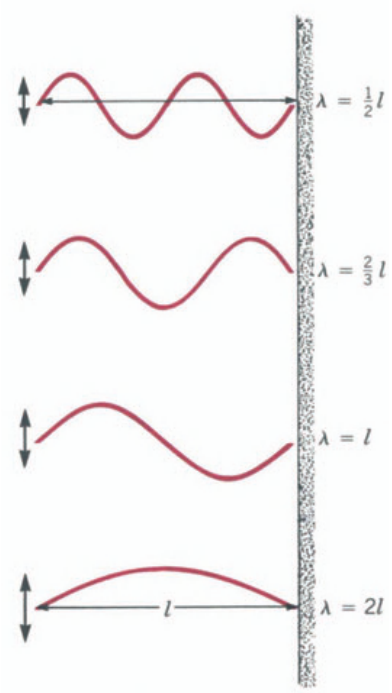


FIGURE 12-19
Configuration of the first four standing waves in a string of length l.

PROBLEMS

12.1 Two sources emit waves of the same frequency, wavelength, and amplitude. What is the amplitude of the resulting wave at a point P at a distance x_1 from source S_1 and a distance x_2 from source S_2 if $x_1 - x_2$ is (a) one wavelength? (b) one-half wavelength?

12.2 Two slits in an opaque screen are separated by a distance $d = 10^{-5}$ m. Light of frequency $\nu = 5 \times 10^{14}$ Hz is shone through the slits. Find the angular position of the first three interference maxima.

12.3 In the double slit of problem 12-2, what is the angular separation between the first interference maxima for two waves of wavelength $\lambda_1 = 6000$ Å and $\lambda_2 = 4000$ Å?

12.4 Two slits separated by a distance $d = 4 \times 10^{-5}$ m, are 1.5 m away from a screen (see Fig. 12-20). What is the separation $y_2 - y_1$ between the first and the second interference maxima if light of wavelength $\lambda = 5000$ Å is sent through the slits?
(*Answer*: 1.88 cm.)

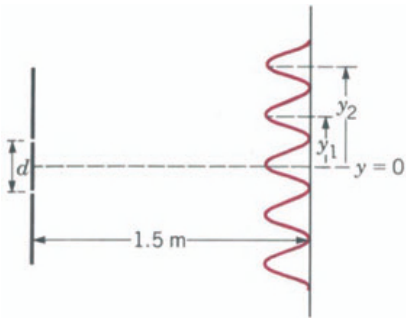


FIGURE 12-20
Problem 12.4.

12.5 Light from a sodium lamp contains waves with wavelength $\lambda_1 = 5880$ Å and waves with $\lambda_2 = 5890$ Å. Find the angular separation and the linear separation of the two wavelengths on a screen 50 cm away from a double slit for the first-order maxima. Do the calculations for a slit separation of 70,000 Å and for a slit separation of 7000 Å.

12.6 In a double slit experiment performed with light of wavelength $\lambda = 5400$ Å, the separation between the tenth interference maximum and the central maximum on a screen 150 cm away is 10 cm. What is the spacing between the slits?

12.7 Two speakers separated by a distance of 3 m emit sound waves of frequency $\nu = 550$ Hz. The velocity of sound is 330 m/sec. Find the position of the points along the line S_1O in Fig. 12-21, at which the intensity of the sound will be a maximum.

(*Answer*: 7.20 m, 3.15 m, 1.60 m, 0.68 m, 0 m from S_1.)

FIGURE 12-21
Problem 12.7.

12.8 A source of waves S and a detector D are located 8 m apart (see Fig. 12-22). A horizontal reflecting surface is placed 3 m above the source and the detector. The direct wave from S to D is found to add constructively with the reflected wave. When the reflecting surface is raised an additional distance of 0.204 m, the direct and the reflected waves add destructively at D. What are the possible values of the wavelength λ?

(*Answer*: 0.500 m, 0.167 m, 0.100 m, 0.071 m, . . .)

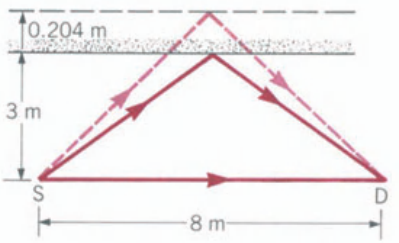


FIGURE 12-22
Problem 12.8.

12.9 Monochromatic (single wavelength) light is directed on a double slit. A light meter is placed to the right of the slits in the position shown in Fig. 12-23. When slit S_2 is closed, the light intensity at the location of the meter is I_1. When slit S_1 is closed the light intensity is I_2. (a) What is the light intensity I_T when both slits are open if $x_1 - x_2 = \lambda$? (b) What is I_T if $x_1 - x_2 = \frac{1}{2}\lambda$? I_1 and I_2 are not necessarily equal. Assume that the size of the slits is smaller than the wavelength so that the slits can be considered to be point sources.

(*Answer*: (a) $I_1 + I_2 + 2\,(I_1I_2)^{1/2}$, (b) $I_1 + I_2 - 2\,(I_1I_2)^{1/2}$.)

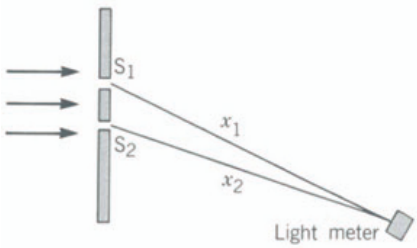


FIGURE 12-23
Problem 12.9.

12.10 Light of wavelength $\lambda = 5000$ Å is incident on a single slit of width 10^{-5} m. What is the angular separation between the central maximum and the eighth-order diffraction minimum?

12.11 Light of wavelength 6000 Å is sent through a single slit. If the angular separation between adjacent diffraction minima is 0.2°, what is the width of the slit? (For small angles, $\sin\theta = \theta$ (in radians)).

(*Answer*: 1.72×10^{-4} m.)

12.12 A screen is placed 2 m to the right of a single slit of unknown width. Light of wavelength 5200 Å is incident on the slit from the left. The separation on the screen between the second-order minima on either side of the central maximum is 5.2 cm. What is the width of the slit?

12.13 Monochromatic X rays of wavelength $\lambda = 1.2$ Å are incident on a crystal. The first-order dif-

fraction maximum is observed when the angle θ between the incident beam and the atomic planes is 12°. (a) What is the separation of the atomic planes responsible for the diffraction? (b) What is the highest order Bragg diffraction produced by those planes that can be observed?

(*Answer*: (a) 2.89 Å, (b) 4th.)

12.14 Sodium chloride (NaCl) has a crystal structure similar to that of silver bromide (AgBr) shown in Fig. 12-18. The atomic weight of NaCl is 58.44 g/mole and its density is 2.16 g/cm^3. (a) Calculate the spacing between the atoms in a NaCl crystal. (b) If X rays of wavelength 1.5 Å are incident on a NaCl crystal, at what angle θ will the first order diffraction maximum be observed?

(*Answer*: (a) 2.82 Å, (b) 15.4°.)

12.15 Potassium chloride (KCl) has the crystal structure of AgBr in Fig. 12-18. The molecular weight and the density of KCl are 74.55 g/mole and 1.98 g/cm^3, respectively. The distance between adjacent atomic planes is 3.14 Å. (a) Calculate Avogadro's number from this data. (b) If the first-order diffraction maximum for X rays incident on these atomic planes is observed when the angle θ between the incident direction and the crystal planes is 30°, what is the wavelength of the X rays?

12.16 The wave velocity in a string 1 m long is 6 m/sec. What are the frequencies of the standing waves in the string?

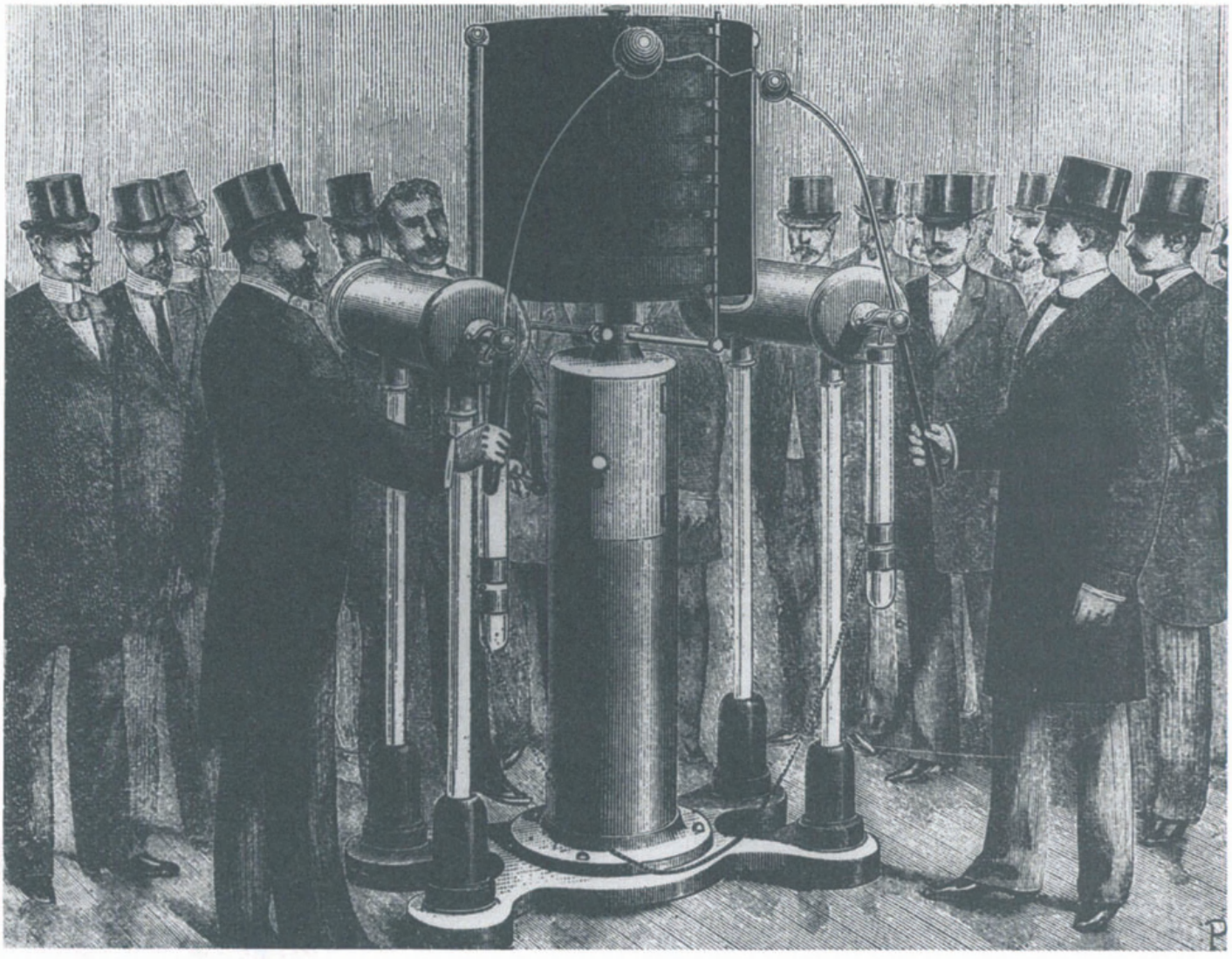
12.17 A standing wave in a string is described by the equation

$$y = (0.7 \text{ m}) \sin (4\pi x) \cos (20\pi t)$$

where x is in meters and t is in seconds. (a) What is the amplitude and the velocity of propagation of the traveling waves that gave rise to such a standing wave? (b) What is the amplitude of vibration of the particles in the string located at $x = 0.45$ m? (c) What is the transverse velocity of the particles at $x = 0.45$ m at $t = 0.25$ sec? (d) What are the locations of the nodes?

(*Answer*: (a) 0.35 m, 5 m/sec, (b) 0.41 m, (c) 0, (d) 0 m, 0.25 m, 0.50 m, 0.75 m.)

12.18 A standing wave of frequency $\nu = 10$ Hz is set up in a string of mass $m = 0.100$ kg and length $l = 2$ m. The maximum amplitude of vibration is 5 cm. What is the total energy of the standing wave? Assume that $\lambda = l$. (*Answer*: 0.25 J.)



CHAPTER 13

Electrostatics

13.1 INTRODUCTION

In this chapter we begin the study of electricity. Although we ultimately wish to understand the flow of electric charges through electrical circuits, we must start with the simple empirical laws of the interaction of charges at rest, called *electrostatics*. Our starting point will be the observed behavior of electric charges at rest and how careful observations by Coulomb led him to postulate laws of the behavior of charges at rest in their interaction with one another. We will also examine the *superposition principle* according to which the behavior of multiple charges on one another is a simple sum of the one-to-one interactions (*pairwise*).

13.2 ATTRACTION AND REPULSION OF CHARGES

Everyone has experienced some of the phenomena of static electricity. When you comb dry, clean hair, it is attracted to the comb. For a short time afterward the comb will attract small particles such as little pieces of paper. These phenomena have been known for thousands of years. The ancient Greeks noticed that if a piece of amber (fossilized tree sap) were rubbed with a piece of cat fur, it would attract pieces of dry leaves. In fact, the Greek word for amber is elektron.

These electrical phenomena fascinated many early investigators, and a useful device called the *electroscope* was invented about 200 years ago to further the studies of these phenomena. An electroscope is simply two very thin gold leaves attached together at the end of a metal rod, usually with a metal knob at the other end. The leaves are enclosed in a protective glass-windowed case (see Fig. 13-1). Gold was used because it is a soft metal that can be beaten very thin. Therefore, the foils have very little mass and not much force is required to push them apart.

If an amber rod (hard rubber or a synthetic polymer will also serve) is rubbed with cat fur and brought close to the metal knob, it is seen that the gold leaves separate into a wide angle as if they are trying to get away from each other. Indeed they are, and their behavior is termed "repelling" one another. If the rod is pulled back from the knob, the leaves collapse again into the downward position. If, however, the rod is touched to the knob before it is removed, the leaves remain apart, in the repelling position. This suggests that something has flowed from the rod to the leaves, and we now know that it is an electric charge. If a glass rod is rubbed with a piece of silk, the same phenomena will occur. However, if the electroscope is first charged with the amber rod and then touched very briefly with the glass rod, it will discharge; that is, the leaves will fall back down. If the glass rod is in contact for a longer period, it is seen that the electroscope will first discharge and then recharge; that is, the leaves will again move apart. This experiment can

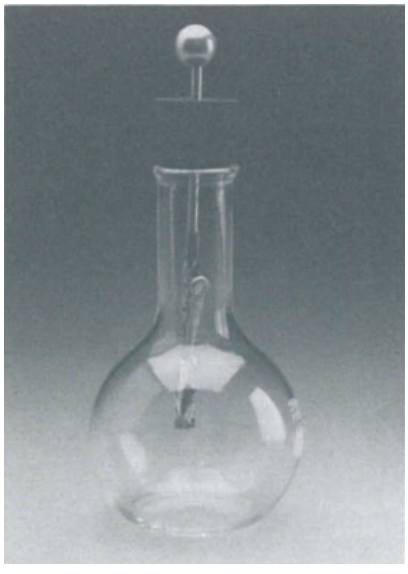


FIGURE 13-1
Electroscope.

be done in the reverse. That is, the electroscope charged by the glass rod can be discharged and recharged with the amber rod. These phenomena were explained by the assumption that there are two different types of charges arbitrarily called *positive* and *negative:* one type was produced on rubbing amber with cat fur and the other by rubbing glass with silk. This is still the model today, and we speak of positive and negative charges. In other words, we now say that the electron has a negative charge, but this is a result of history: its charge could just as easily have been called positive without affecting its behavior. It was also noted in these studies that the rod in the electroscope had to be made of metal. No effect would be observed if it were made of wood, rubber, or some other nonmetallic material. This implies that electrical charges can flow through metal but not through nonmetals, such as the ones just named. Thus metals have been known as electrical *conductors* and nonmetals as *insulators*. In Chapters 24 and 25 we will analyze the difference between these two types of materials and introduce the *semiconductor*, a material with electrical properties that lie between those of a conductor and an insulator and whose characteristics form the basis of the modern computer.

13.3 COULOMB'S LAW

Charles Coulomb (1736–1806) published between 1786 and 1789 the results of a series of experiments that he had performed. Instead of using the electroscope in which the force between the gold leaves was difficult to measure, he used very small lightweight balls on the ends of long threads. The balls were made of the centers of dried reed stems called *pith* and were small so that they could approximate point charges. If two were suspended adjacent to each other and both touched with *either* the amber or glass rod, they would repel each other as did the leaves of the electroscope. He therefore confirmed that *like charges repel*. If he had touched both with the amber rod and achieved repulsion and then touched one with the glass rod they would come together. This experiment could be reversed in that if the repulsion were first achieved by touching both with the glass rod, touching one with the amber rod would cause them to come together. He therefore confirmed that *unlike charges attract*.

Charles Coulomb (1736–1806).

Before continuing with the work of Coulomb, we can use these two findings to interpret the electroscope observations. Accepting that there are two types of charges and that at least one of these can move in a metallic conductor but not in an insulator, we may view the sequence of diagrams in Fig. 13-2. In Fig. 13-2*a*, the negatively charged amber rod is brought near the electroscope and the mobile negative charges in the metal rod, being repelled, go to the ends of the gold leaves, their maximum distance from the amber rod. The gold leaves now both have an excess of negative charge and repel each other. Although at Coulomb's time it was believed that the positive charges were mobile, we now know that this is not generally true and it is

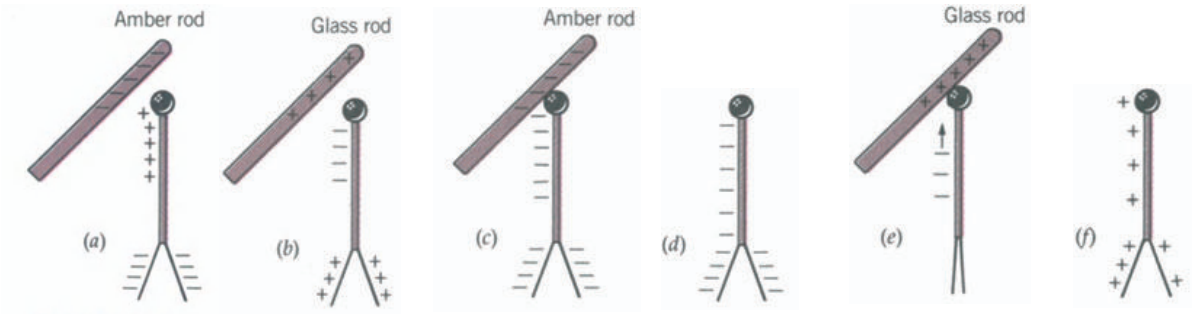


FIGURE 13-2

the negatively charged electrons that move in the metal. The same effect occurs in (*b*) with a positively charged glass rod, for it does not matter if the positive charges in the leaves are there because they have been repelled by the glass rod or because the electrons have been attracted to it, thereby leaving a net excess of positive charge on the leaves. In (*c*), the amber rod is touched to the electroscope. Because the upper end of the metal rod, as shown in (*a*), is positively charged, electrons will flow from the amber to the metal rod. When the amber rod is removed in (*d*), the electroscope has a net excess of negative charges throughout and the leaves therefore repel each other. In (*e*), a positively charged glass rod is touched to the negatively charged electroscope of (*d*) and first the excess electrons leave, causing the leaves to collapse because they now have no excess charge, that is, they are neutral. If this positively charged glass rod has a sufficient amount of charge and it is rubbed further against the metal rod of the electroscope, more electrons will leave the electroscope, resulting in an excess of positive charges in the electroscope. Therefore the leaves repel each other as in (*f*).

We now return to Coulomb's experiments concerning the forces between charges. Let us represent a quantity of charge, a scalar, by the letter q and the distance between charges by the letter r. Coulomb, in his experiments, attached fine threads to the pith balls and passed them over glass rods, which effectively served as frictionless pulleys. To the ends of these threads he fastened various weights (see Fig. 13-3*a*). If the charges q_1 and q_2 placed on the pith balls from the amber and glass rods were of opposite sign, the balls would be attracted to each other by a force F, which would be countered by an appropriate weight Mg. The force diagram for q_2 is shown in Fig. 13-3*b*. Note that the tension T_2 in the horizontal thread is equal to the hanging weight Mg. We have here the equilibrium problem of Chapter 4 where the forces to the left are $T_1 \cos \theta$ plus the electrostatic attraction, and these are equal to Mg ($T_2 = Mg$). By measuring the weights Mg required to hold the balls apart a distance r, Coulomb found that the magnitude of the attractive force was proportional to the reciprocal of the square of the distance between the balls; that is,

$$F \propto \frac{1}{r^2}$$

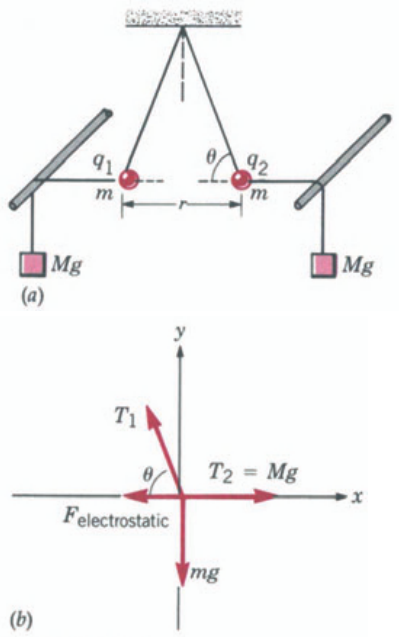


FIGURE 13-3

(*a*) Experimental arrangement for the determination of Coulomb's law. (*b*) Force diagram for q_2.

He then put the same type of charge on each ball and draped the threads over the rods in the opposite direction to measure the force required to hold the balls at various distances r from each other. He found that the force of repulsion also varied inversely as the square of the distance between them. Having established this relation, he was then able to hold them at a fixed distance and vary the charges on them. He could do this accurately by fractionating the charges. He had a set of metal knobs on wooden sticks (insulators). If he touched one of these metal knobs to an amber or glass rod it would acquire a charge of magnitude q. If he then touched this knob to an identical uncharged one they would share the charge equally and each would have a charge $q/2$. Touching either of these to another would reduce the charge to $q/4$, and so on. Touching two knobs in contact with one another would produce a charge of $q/3$. Many combinations of these fractions were used to charge the pith balls at a fixed distance from one another. Coulomb showed that the force of attraction between oppositely charged balls or of repulsion between balls with charges of the same sign was proportional to the product of the magnitude of the two charges, $q_1\, q_2$. Coulomb's law is therefore

$$F \propto \frac{q_1 q_2}{r^2} \tag{13.1}$$

where the sign of q_1 and q_2 may be either plus or minus. This proportionality can be made into an equality by introducing a constant dependent on the system of units used. In the SI system this constant, taken as $1/4\pi\epsilon_0$, has the value

$$\frac{1}{4\pi\epsilon_0} = 9 \times 10^9 \text{ N-m}^2/\text{C}^2$$

The symbol C stands for Coulomb and is the unit of charge. It is important to note at this time that 1 C is *not* the charge of an electron. The charge of the electron in coulombs is $e = -1.6 \times 10^{-19}$ C.

Equation 13.1 can now be written as

$$F = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r^2} \tag{13.2}$$

$$F = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r^2}$$

The direction of the force that q_1 exerts on q_2 is along the line joining the two charges, pointing away from q_1 if the force is repulsive (q_1 and q_2 having the same sign) or toward q_1 if the force is attractive (q_1 and q_2 oppositely charged).

Example 13-1

Two pith balls of mass 0.1 g each are suspended on 50-cm threads. They are given equal charges and assume a position in which each makes an angle of 20° with the vertical, as in Fig. 13-4*a*. What is the charge on each?

Solution The vector diagram of the forces on the right-hand ball is shown in Fig. 13-4*b*, where F is the coulombic force of repulsion between the two

charged pith balls. Because the ball is in equilibrium, we may write

$$\Sigma F_x = 0$$

$$F - T \cos 70° = 0$$

$$F = 0.34\ T$$

$$\Sigma F_y = 0$$

$$T \sin 70° - mg = 0$$

$$T = \frac{mg}{\sin 70°}$$

$$T = \frac{0.1 \times 10^{-3}\ \text{kg} \times 9.8\ \text{m/sec}}{0.94} = 1.04 \times 10^{-3}\ \text{N}$$

Substituting this value of T in the equation for F, we obtain

$$F = 0.34\ T = 0.34 \times 1.04 \times 10^{-3}\ \text{N} = 3.5 \times 10^{-4}\ \text{N}$$

From Fig. 13-4*a*, the distance r between the two balls is

$$r = 2\ l \sin 20°$$

$$r = 2 \times 0.5\ m \times \sin 20°$$

$$r = 0.34\ m$$

Using Coulomb's law,

$$F = 9 \times 10^9 \frac{\text{N-m}^2}{\text{C}^2} \frac{q^2}{r^2}$$

because

$$q_1 = q_2$$

and substituting for F and r we obtain

$$3.5 \times 10^{-4}\ \text{N} = \frac{9 \times 10^9 \dfrac{\text{N-m}}{\text{C}^2}\ q^2}{(0.34\text{m})^2}$$

or

$$q = 6.7 \times 10^{-8}\ \text{C}$$

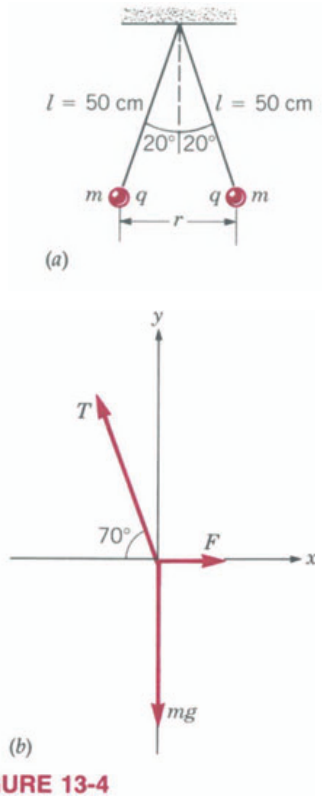


FIGURE 13-4
Example 13-1.

13.4 CHARGE OF AN ELECTRON

In a series of experiments, Robert Millikan (1868–1953) in the years 1909 through 1913 measured the charge on an electron (see Fig. 13-5). With a spray he introduced fine oil drops between two parallel metal plates and observed the motion of a single drop through a telescope. By its rate of fall through

Robert Millikan (1868–1953).

the air he was able to use a formula for the terminal velocity (constant rate of fall through a medium) to estimate its weight. He found that he could arrest its downward motion, that is, hold it stationary by placing a positive charge on the upper plate. (We will see later that he could control the positive charge on the plate by means of the voltage.) Therefore, the balance of forces of mg down and the upward attraction could be used to determine the charge on the drop. First, he found that the drops always acquired a negative charge, never positive. This showed that it is the negative charge that is apparently the more mobile one. His second finding, over hundreds of experiments, was that the smallest charge that was ever acquired by the drop had a magnitude of 1.6×10^{-19} C and that larger charges were always integral multiples of this quantity. He therefore assigned this value to the charge of the electron; it is the smallest negative charge that can be found. Because atoms are neutral and contain equal numbers of electrons and protons, the charge of the heavy and essentially immobile proton also has this magnitude, but it is positive. We will see that the protons are in the nucleus and are relatively massive. It is therefore understandable why it is the negative charge that is the mobile species rather than the positive one. This was not known prior to Millikan's experiment. The great theories of electrical behavior were developed in the nineteenth century when it was assumed that the positive charge was the mobile species. We will see that all electrical definitions are based on the behavior of a unit positive charge. This sometimes leads to the confusion of students; but, as mentioned earlier, the assignment of the words positive and negative are completely arbitrary, and therefore the theories remain valid.

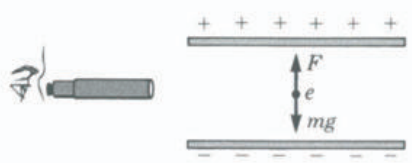


FIGURE 13-5
Experimental arrangement for the determination of the charge of the electron.

13.5 SUPERPOSITION PRINCIPLE

Coulomb's law, defined in Section 13.3, relates to the force between two charges. It is an empirical law derived from experimental measurement. How does one treat a situation in which three or more charges are involved? To answer that requires further experimentation. It is found that the force between any two charges in a group of charges is independent of the presence of the other charges. What this means is that if one selects a given charge in a group and asks for the total force on it, this force would be the resultant of the individual vector forces on it from each of the charges. This is called the *superposition principle* of charges. It makes the calculations straightforward because we can treat them as we did in summing vector forces in Chapter 2. Let us examine the case of three charges in Example 13-2.

Example 13-2

Three charges are arranged in a triangle as shown in Fig. 13-6*a*. What is the direction and the magnitude of the resultant force on the 1×10^{-8} C charge?

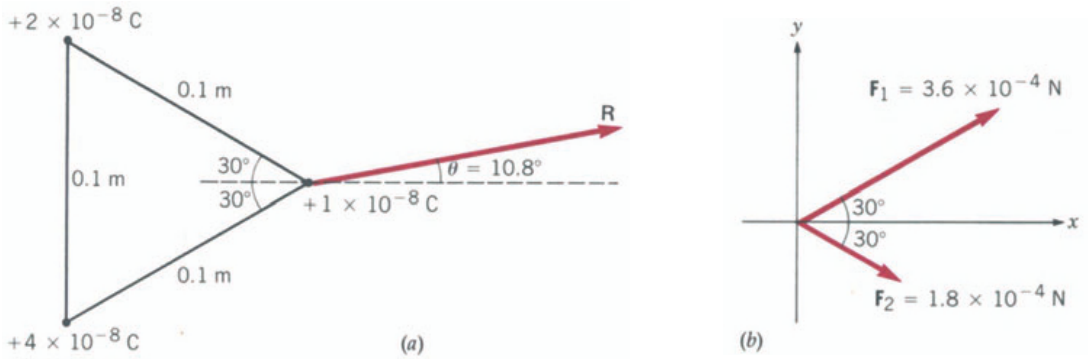


FIGURE 13-6
Example 13-2.

Solution The force resulting from the 4×10^{-8} C charge by Coulomb's law is

$$F_1 = 9 \times 10^9 \frac{\text{N-m}^2}{\text{C}^2} \frac{1 \times 10^{-8}\ \text{C} \times 4 \times 10^{-8}\ \text{C}}{10^{-2}\ \text{m}^2} = 3.6 \times 10^{-4}\ \text{N}$$

at 30° above the positive x axis (see Fig. 13-6*b*).

The force resulting from the 2×10^{-8} C charge is

$$F_2 = 9 \times 10^9 \frac{\text{N-m}^2}{\text{C}^2} \frac{1 \times 10^{-8}\ \text{C} \times 2 \times 10^{-8}\ \text{C}}{10^{-2}\ \text{m}^2} = 1.8 \times 10^{-4}\ \text{N}$$

at 30° below the positive x axis (Fig. 13-6*b*).

We now use the vector diagram of these two forces, Fig. 13-6*b*, and find the resultant by the component method of Chapter 2. We have

$$\mathbf{F}_1 = 3.6 \times 10^{-4}\ \text{N} \cos 30°\ \mathbf{i} + 3.6 \times 10^{-4}\ \text{N} \sin 30°\ \mathbf{j}$$

$$\mathbf{F}_2 = 1.8 \times 10^{-4}\ \text{N} \cos 30°\ \mathbf{i} - 1.8 \times 10^{-4}\ \text{N} \sin 30°\ \mathbf{j}$$

$$\mathbf{R} = 4.7 \times 10^{-4}\ \text{N}\ \mathbf{i} + 0.9 \times 10^{-4}\ \text{N}\ \mathbf{j}$$

$$|R| = \sqrt{(4.7 \times 10^{-4}\ \text{N})^2 + (0.9 \times 10^{-4}\ \text{N})^2} = 4.8 \times 10^{-4}\ \text{N}$$

$$\theta = \arctan \frac{0.9}{4.7} = 10.8° \text{ above the positive } x \text{ axis}$$

The direction of the resultant force vector is indicated by the arrow labeled **R** in Fig. 13-6*a*.

PROBLEMS

13.1 Two particles of charge $q_1 = +2 \times 10^{-9}$ C and $q_2 = +3 \times 10^{-9}$ C are placed 0.04 m from each other. What is the force of repulsion that each experiences?

13.2 A particle of charge $q_3 = -2 \times 10^{-9}$ C is placed midway between the two charged particles of problem 13.1. What is the net force on it and in what direction?

13.3 At what position between particles 1 and 2 of problem 13.1 will particle 3 of problem 13.2 experience no net force?

(*Answer*: 0.018 m from q_1, between q_1 and q_2.)

13.4 Three charges lie on the x axis as in Fig. 13-7. Find the resultant force on the middle charge, q_2.

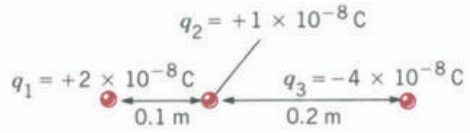


FIGURE 13-7
Problem 13.4.

13.5 An iron atom of mass 9.32×10^{-26} kg has 26 electrons. The density of iron is 7.86 g/cm^3. If two identical iron balls each of volume 1 cm^3 were stripped of all their electrons and placed 1 m apart, (a) What would be the electrostatic force between them? (b) What would be the gravitational force between them? (c) Compare these forces.

13.6 Two particles of mass 5 kg each are given an equal amount of charge. (a) What must the charge be on each particle so that the gravitational attraction exactly balances the electrostatic repulsion? (b) How many electronic charges does that charge correspond to?

13.7 Two threads of length 0.7 m support balls of mass 0.2 gm as in Example 13.1. Equal charges of the same sign are put on the balls and they repel, each making an angle of 30° with the vertical. What is the charge on each ball?

(*Answer*: 2.48×10^{-7} C.)

13.8 A charge q is to be shared by two particles. What must be the charge on each particle so that the force between them, for a fixed separation, is a maximum?

(*Answer*: $\frac{q}{2}$.)

13.9 An *electric dipole* consists of two charges of equal magnitude q but of opposite sign separated by some distance d. The electric dipole moment is defined as $\mu_e = qd$. Consider an electric dipole lying on the y axis as in Fig. 13-8. What is the force exerted by the dipole on a charge q' located on the x axis at a distance x from the origin?

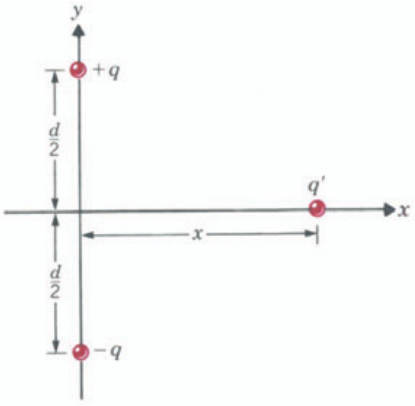


FIGURE 13-8
Problem 13.9.

13.10 Four charges are located at the corners of a square as shown in Fig. 13-9. (a) What is the resultant force on q_2? (b) What should q_1 and q_4 be so that the resultant force on q_2 is zero?

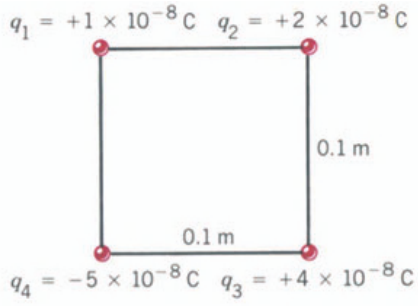


FIGURE 13-9
Problem 13.10.

13.11 Two charges, $q_1 = 3 \times 10^{-6}$ C and $q_2 = -3 \times 10^{-6}$ C, are connected by an insulating rod 10 m long. The rod is pivoted about its midpoint. The rod is kept horizontal 10 cm above the floor. Two identical charges, $q = 5 \times 10^{-6}$ C, are fixed directly below q_1 and q_2, as shown in Fig. 13-10. Where should a 3-kg weight be placed on the rod to keep the rod horizontal?

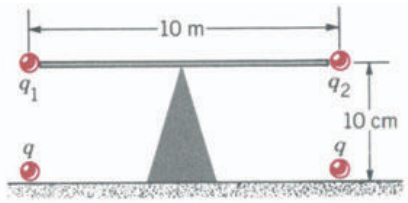


FIGURE 13-10
Problem 13.11.

13.12 A fixed conducting ball has a charge $q_1 = 3 \times 10^{-6}$ C. An identical ball with charge q_2 is held at a distance x away from q_1. The two balls attract each other with a force of 13.5 N. The balls are then connected by a conducting wire. After the wire is removed, the balls repel each other with a force of

0.9 N. (a) What was the charge q_2 of the second ball? (b) What is the separation x between the balls?

(*Answer*: (a) -5×10^{-6} C, (b) 0.10 m.)

13.13 An α particle ($q = 3.2 \times 10^{-19}$ C) is projected from far away directly toward a gold nucleus ($q' = 79 \times 1.6 \times 10^{-19}$ C). The mass of the α particle is 6.7×10^{-27} kg and its initial velocity is 4×10^{6} m/sec. Use the work-energy theorem of Section 5.4 (Eq. 5.9) to calculate the closest distance of approach of the α particle to the nucleus. Assume that the gold nucleus remains stationary. (*Hint:* In this case the force on the α particle is not constant. Therefore, the integral form for work, Eq. 5.7′, must be used to evaluate the work done on the α particle.)

(*Answer:* 6.8×10^{-13} m.)

13.14 Two positive charges q are held fixed and are separated by a distance $2a$. A third positive charge q' of mass m is initially placed halfway between them (Fig. 13-11). q' is then displaced a small distance x ($x << a$) and released. (a) Show that the force on q' is approximately proportional to x and in the opposite direction of the displacement x. (b) The force on q' is therefore similar to the force exerted by a spring on a block connected to it (Chapter 10). What is the period of oscillation of q'?

(*Answer*: $2\pi \left[\dfrac{\pi\epsilon_0 a^3 m}{q\,q'}\right]^{1/2}$.)

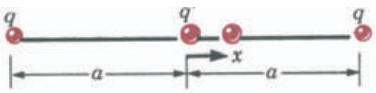


FIGURE 13-11
Problem 13.14.

CHAPTER 14

The Electric Field and the Electric Potential

14.1 INTRODUCTION

We have seen in the preceding chapter how the presence of an electric charge has an effect on another electric charge. This raises the question: What if there is only one electric charge present? The idea of an *electric field* is introduced to describe the effect in all space around a charge so that if another charge is present we can predict the effect on it. If we have multiple charges, such as in Example 13-2, we see that the force of each on a third charge is a vector, and the net effect on the third charge is the resultant of the forces. This resultant will differ with both the position and the charge of the third one. The concept of separating the calculation into the formation of an electric field and the response to the electric field by a given charge placed in it greatly simplifies the calculations.

14.2 THE ELECTRIC FIELD

If, in Example 13-2, we had wished to find the resultant force on a charge of a different magnitude in the same position, we would have to repeat the calculations. However, after a few such repeat calculations we would notice that all one has to do is multiply the first result by the ratio of the magnitude of the new charge to that of the charge used in the first calculation. If the first calculation has been made for a test charge of $+1$ C, the task will be easier, for then the ratio of the magnitude of the new charge to that of the charge of the first calculation is simply the magnitude of the new charge. In other words, let us simply define a new quantity, called the *electric field*, with symbol $\mathscr{E}$, at a point in space as the vector resultant force experienced by a *positive* test charge of magnitude 1 C placed at that point. If an arbitrary test charge q' is placed at that point, the charge will experience a force

$$\mathbf{F} = q'\mathscr{E} \tag{14.1}$$

$\mathbf{F} = q'\mathscr{E}$

Thus, the electric field at a point in space can be calculated by measuring the force experienced by a test charge q' and dividing it by the magnitude of the test charge; that is,

$$\mathscr{E} = \frac{\mathbf{F}}{q'} \tag{14.2}$$

If we consider two charges, q and q', separated by a distance r, from Coulomb's law (Eq. 13.2) the magnitude of the force between them is

$$F = \frac{1}{4\pi\epsilon_0}\frac{qq'}{r^2} \tag{13.2}$$

Let us arbitrarily consider q' as the test charge and q as the charge creating the electric field at the point P where q' is located. On substitution of Eq. 13.2

in Eq. 14.2, the magnitude of the electric field produced by q at P is given by

$$\mathscr{E} = \frac{1}{4\pi\epsilon_0}\frac{qq'}{q'r^2}$$

or

$$\mathscr{E} = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2} \tag{14.3}$$

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

Equation 14.3 is the general expression for the electric field resulting from a charge q at a point located a distance r away from the charge.

The direction of $\mathscr{E}$ can be deduced from its original definition as the force experienced by a unit positive charge. If the charge q producing the field is positive, then a positive test charge of 1 C, when placed at P, will be repelled. We conclude that the electric field produced by a positive charge q at a point P is along the line joining the charge q and the point P and directed away from q. It may be said that the electric field is directed radially away from a point charge q (see Fig. 14-1*a*). On the other hand, a negative point charge q will attract the 1-C positive test charge and consequently the electric field that it produces is directed radially toward it (see Fig. 14-1*b*).

We have indicated (Section 13.5) that Coulomb's law obeys the superposition principle. From the definition it follows that the electric field does too. The field produced by a group of charges is simply the vector sum of the fields produced by the individual charges.

An important point to mention here is that not only is there no electric field when there are no charges, but there is no electric field at a point when the forces from an assembly of charges on a test charge is zero at that point. Also, if an array of equal numbers of positive and negative charges are located in a small region, then at some distant point (distant relative to the distance between the charges) there is no measurable electric field. This is why atoms when they are far away from each other, such as in a dilute gas, experience no measurable electric fields. However, the electrons that make up the atom experience the electric fields of the nucleus and of the other electrons. Furthermore, if a charged particle is shot at an atom, as in a nuclear experiment, it will get close enough to experience the internal electric fields of the atom.

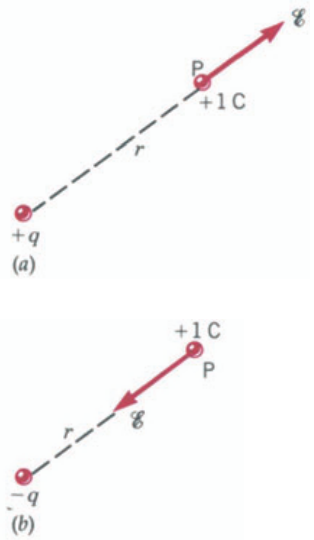


FIGURE 14-1
Direction of the electric field set up at point P: (*a*) by a positive charge q, and (*b*) by a negative charge $-q$.

Example 14-1

A charge $q_1 = 3 \times 10^{-6}$ C is located at the origin of the x axis. A second charge $q_2 = -5 \times 10^{-6}$ C is also on the x axis 4 m from the origin in the positive x direction. (a) Calculate the electric field at the midpoint P of the line joining the two charges. (b) At what point P' on that line is the resultant field zero?

Solution

(a) Because q_1 is positive, its electric field $\mathscr{E}_1$ at P is away from it, that is, in the positive x direction. The electric field $\mathscr{E}_2$ produced at point

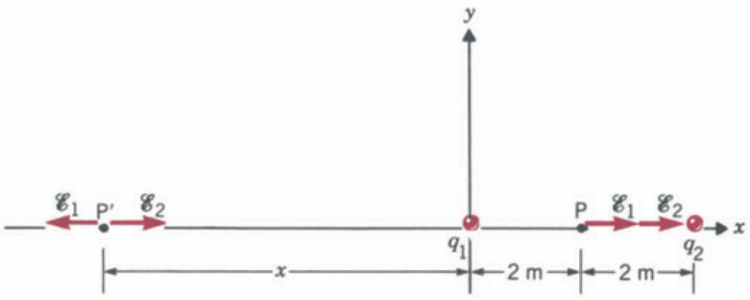


FIGURE 14-2
Example 14-1.

P by q_2 is toward q_2, that is, in the same direction as $\mathscr{E}_1$ (see Fig. 14-2). From Eq. 14.3, the magnitudes of $\mathscr{E}_1$ and $\mathscr{E}_2$ are

$$|\mathcal{E}_1| = 9 \times 10^9 \frac{\text{N-m}^2}{\text{C}^2} \frac{3 \times 10^{-6}\text{C}}{(2\text{m})^2} = 6.75 \times 10^3 \text{ N/C}$$

$$|\mathcal{E}_2| = 9 \times 10^9 \frac{\text{N-m}^2}{\text{C}^2} \frac{5 \times 10^{-6}\text{C}}{(2\text{m})^2} = 11.25 \times 10^3 \text{ N/C}$$

Because it is seen in Fig. 14-2 that both $\mathscr{E}_1$ and $\mathscr{E}_2$ are directed along the positive x direction, the resultant electric field $\mathscr{E}$ at P will be

$$\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_2 = 6.75 \times 10^3 \text{ N/C} + 11.25 \times 10^3 \text{ N/C}$$

$$= 18 \times 10^3 \text{ N/C}$$

(b) From part (a), it is clear that the resultant $\mathscr{E}$ cannot be zero at any point between q_1 and q_2 because both $\mathscr{E}_1$ and $\mathscr{E}_2$ are in the same direction. Similarly $\mathscr{E}$ cannot be zero to the right of q_2 because the magnitude of q_2 is greater than q_1 and the distance r in Eq. 14.3 is smaller for q_2 than q_1. $\mathscr{E}$ can only be zero to the left of q_1 at some point P′ to be found.

$$\mathscr{E} = \mathscr{E}_1 + \mathscr{E}_2 = 0$$

$$\mathscr{E}_1 = -\mathscr{E}_2$$

and

$$|\mathscr{E}_1| = |\mathscr{E}_2|$$

or

$$\frac{1}{4\pi\epsilon_0} \frac{q_1}{x^2} = \frac{1}{4\pi\epsilon_0} \frac{q_2}{(x+4)^2}$$

$$3(x+4)^2 = 5x^2$$

$$2x^2 - 24x - 48 = 0$$

$$x = 13.75 \text{ m}, \quad x = -1.75 \text{ m}$$

The second root of the quadratic equation, $x = -1.75$ m, represents

a point between the charges at which $\mathscr{E}_1 = \mathscr{E}_2$. However, as indicated earlier, between the charges, $\mathscr{E}_1$ and $\mathscr{E}_2$ have the same direction and consequently the resultant field is not zero.

14.3 ELECTRICAL POTENTIAL ENERGY

We now wish to develop an expression for the amount of work required to move a charge in an electric field. First we note that the magnitude of the electric field at a point P resulting from a point charge is independent of the angular position of the point P, because only distance enters into Eq. 14.3. In the preceding section we also showed that the direction of the electric field is radially away from the charge producing the field if the charge is positive or radially toward it if the charge is negative. Thus, the direction of the electric field from a positive point charge may be represented by simply drawing arrows out from it in all directions. Figure 14-3 shows such a schematic in two dimensions only. If we move a positive test charge q' from point A to point B, we have a situation similar to that of Chapter 5 where we showed that the work against a gravitational force is independent of the path. We must recognize that in this case the force is not constant. In Fig. 14-3 work must be done whenever the radial distance between the moving charge, q', and that which creates the electric field, q, is changed. However, when q' moves tangentially, no work is done because the direction of motion is perpendicular to the electric field and therefore to the force acting on q'. Recall that by definition work involves the dot product of the force vector $\mathbf{F}$ and displacement vector $\Delta\mathbf{s}$, that is, $W = \mathbf{F} \cdot \Delta\mathbf{s}$ (Eq. 5.3). In Fig. 14-3 the same amount of work is done in moving a charge from point A to point B either by path 1 (solid line) or by path 2 (dashed line) or by any other path.

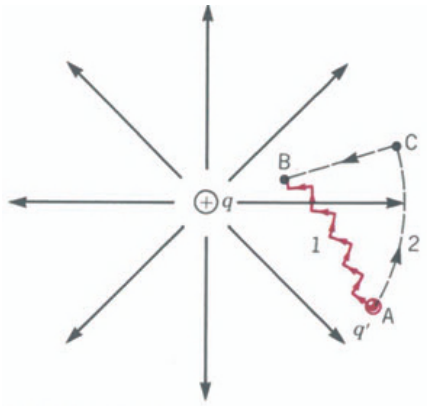


FIGURE 14-3
Two possible paths for bringing a charge q' from point A to point B.

Because the force on the test charge q' is not constant but changes with distance, we use Eq. 5.7′ to evaluate the work done in moving it from point A to point B

$$W_{A\to B} = \int_A^B \mathbf{F} \cdot d\mathbf{s} \qquad (5.7')$$

Just as in the case of gravity, considered in Chapter 5, the force $\mathbf{F}$ needed to move q' at constant velocity must be equal and opposite to the force exerted by the electric field of q, that is, because the force of the electric field on q' is $q'\,\mathscr{E}$, a force $\mathbf{F} = -q'\mathscr{E}$ is needed to move q' with constant velocity, where $\mathscr{E}$ is the electric field produced by q. Substituting this for $\mathbf{F}$ in Eq. 5.7′ we have

$$W_{A\to B} = -q' \int_A^B \mathscr{E} \cdot d\mathbf{s} \qquad (14.4)$$

Because $W_{A\to B}$ is independent of the path followed, we will evaluate it by moving tangentially from point A to C and then radially from point C to point B (see Fig. 14-4). During the first leg of this trip (A to C), the work done is zero because $\mathscr{E}$ is perpendicular to the displacement $d\mathbf{s}$. However, the work

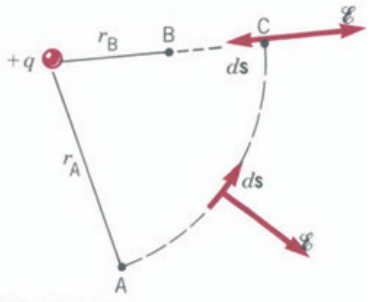


FIGURE 14-4

done from C to B is $\mathscr{E} \cdot d\mathbf{s} = \mathscr{E}\, ds \cos 180° = -\mathscr{E}\, ds$, and Eq. 14.4 becomes

$$W_{A\to B} = q' \int_C^B \mathscr{E}\, ds \tag{14.5}$$

As we move a distance ds toward B from point C, the radius r decreases, which introduces a negative sign, that is, $ds = -dr$. Using this in Eq. 14.5, we obtain

$$W_{A\to B} = -q' \int_C^B \mathscr{E}\, dr \tag{14.6}$$

The electric field for a point charge q is given by Eq. 14.3. Substitution of Eq. 14.3 for $\mathscr{E}$ in Eq. 14.6 yields

$$W_{A\to B} = -\frac{qq'}{4\pi\epsilon_0} \int_{r_A}^{r_B} \frac{dr}{r^2}$$

Note that we have put r_A instead of r_C as the lower limit of the integral because $r_A = r_C$ and we are evaluating the work done in moving q' from A to B. Integrating obtains

$$W_{A\to B} = \frac{qq'}{4\pi\epsilon_0}\left(\frac{1}{r_B} - \frac{1}{r_A}\right) \tag{14.7}$$

By definition (see Section 5.3), the work done in moving an object between two points in a force field is equal to the difference in the potential energy E_p between the two points; that is,

$$E_p(B) - E_p(A) = \frac{qq'}{4\pi\epsilon_0}\left(\frac{1}{r_B} - \frac{1}{r_A}\right) \tag{14.8}$$

Equation 14.8 gives the difference in the potential energy of the two charges when q' is located at two different points. It does not give the potential energy of the charges when q' is at B or at A. For this, as indicated in Section 5.3, we must specify a reference point, that is, a point at which the potential energy is arbitrarily chosen to be zero. In electrostatics, this point is often chosen to be $r = \infty$, that is, when the two charges are separated by an infinite distance. With this assumption, the potential energy of our two charge system q and q' when they are separated by a distance r is simply the work done in bringing one of them (for example q') from infinity to r. Setting $r_A = \infty$ and $r_B = r$ in Eq. 14.8, we have

$$E_p(r) = \frac{1}{4\pi\epsilon_0}\frac{qq'}{r} \tag{14.9}$$

$$E_p(r) = \frac{1}{4\pi\epsilon_0}\frac{qq'}{r}$$

This potential energy is called *electric potential energy* to differentiate it from the gravitational or the elastic potential energies that we encountered earlier. We should note that, because both q and q' are positive, E_p is also positive. To move q' from infinity to r we have to do positive work, we have to overcome

the repulsive force between the two charges, that is, the external force must act in the direction of the displacement.

The same is true if both q and q' are negative. Equation 14.9 holds also in this case because the product of two negative charges will yield a positive value for E_p. If the charges are of unlike sign, they will attract each other and, consequently, to move q' at constant velocity, we will have to hold it back. We will then do negative work and therefore, the potential energy will be negative. It is seen that Eq. 14.9 agrees with this, for if q is positive and q' is negative, or vice versa, E_p will be negative.

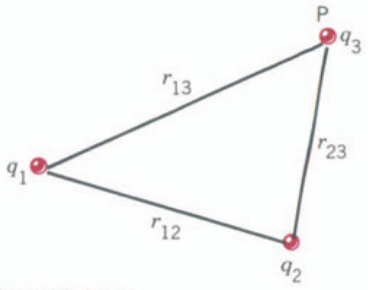


FIGURE 14-5

Let us now have two fixed charges q_1 and q_2 at a distance r_{12} from each other. To achieve this, an amount of work equal to

$$\frac{1}{4\pi\epsilon_0}\frac{q_1q_2}{r_{12}}$$

had to be done; that is, the potential energy of the charges is

$$E_p = \frac{1}{4\pi\epsilon_0}\frac{q_1q_2}{r_{12}} \tag{14.10}$$

Consider now a third charge q_3 that is brought from infinity to point P as shown in Fig. 14-5. How much work must be done? Or equivalently, what is the change ΔE_p in the potential energy of the charges? From Eq. 14.4, setting A $= \infty$, B $=$ P, and $q' = q_3$, we have

$$\Delta E_p = -q_3 \int_\infty^{\mathrm{P}} \mathscr{E} \cdot d\mathbf{s} \tag{14.11}$$

We have already indicated that the electric field obeys the superposition principle. That is, $\mathscr{E} = \mathscr{E}_1 + \mathscr{E}_2$, where $\mathscr{E}_1$ and $\mathscr{E}_2$ are the electric fields produced by q_1 and q_2, respectively. Substituting for $\mathscr{E}$ in Eq. 14.11, we obtain

$$\Delta E_p = -q_3 \int_\infty^{\mathrm{P}} (\mathscr{E}_1 + \mathscr{E}_2) \cdot d\mathbf{s}$$

or

$$\Delta E_p = -q_3 \int_\infty^{\mathrm{P}} \mathscr{E}_1 \cdot d\mathbf{s} - q_3 \int_\infty^{\mathrm{P}} \mathscr{E}_2 \cdot d\mathbf{s} \tag{14.12}$$

Equation 14.12 shows that the total work done in bringing q_3 to point P is simply equal to the sum of the work done against the electric field produced by each charge in the absence of the other. Thus, by using Eq. 14.9 for q_1 and q_3 and for q_2 and q_3 independently, we may write

$$\Delta E_p = \frac{1}{4\pi\epsilon_0}\frac{q_1q_3}{r_{13}} + \frac{1}{4\pi\epsilon_0}\frac{q_2q_3}{r_{23}} \tag{14.13}$$

We should remember that energy is a scalar quantity and, therefore, the sum of the two contributions to the potential energy in Eq. 14.13 is an algebraic sum, not a vector sum as in the case of the electric field. The total energy of

the three-charge system shown in Fig. 14-5 is obtained by combining Eq. 14.13 with Eq. 14.10, that is,

$$E_p = \frac{1}{4\pi\epsilon_0}\left(\frac{q_1q_2}{r_{12}} + \frac{q_1q_3}{r_{13}} + \frac{q_2q_3}{r_{23}}\right) \tag{14.14}$$

Thus, for a system of charges, the procedure to follow is to calculate the potential energy separately for the pairs and then to add these algebraically.

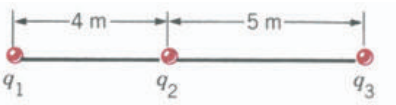


FIGURE 14-6
Example 14-2.

Example 14-2

Three charges—$q_1 = 3 \times 10^{-6}$ C, $q_2 = -5 \times 10^{-6}$ C, and $q_3 = -8 \times 10^{-6}$ C—are positioned on a straight line as shown in Fig. 14-6. Find the potential energy of the charges.

Solution From Eq. 14.14, we may write

$$E_p = 9 \times 10^9 \frac{\text{N-m}^2}{\text{C}^2}\left[\frac{(3 \times 10^{-6}\text{ C})(-5 \times 10^{-6}\text{ C})}{4\text{ m}} + \frac{(3 \times 10^{-6}\text{ C})(-8 \times 10^{-6}\text{ C})}{9\text{ m}} + \frac{(-5 \times 10^{-6}\text{ C})(-8 \times 10^{-6}\text{ C})}{5\text{ m}}\right]$$

$$E_p = 1.43 \times 10^{-2}\text{ J}$$

14.4 ELECTRIC POTENTIAL

In the preceding section we saw that when a test charge q' is moved from point A to point B work is done against the electric field produced by q. The amount of work done (Eq. 14.4) depends on the strength of the field and on the magnitude of the test charge q'. We can introduce a quantity, called the *electric potential*, with symbol V, which is independent of the test charge. The electric potential at a point P is defined as the work done in bringing a unit positive charge from infinity to the point. That is, from Eq. 14.4, setting A = ∞, B = P, and $q' = +1$ C, we have

$$W_{\infty\to P} = V(P) = -\int_{\infty}^{P} \mathscr{E} \cdot d\mathbf{s} \tag{14.15}$$

Because the magnitude of the test charge q' was set equal to unity, the potential at a point depends on the electric field alone and not on the test charge q'. However, if we know the potential at point P, we can immediately conclude, by comparing Eq. 14.4 and Eq. 14.15, that the work done in bringing a charge q' of arbitrary magnitude or sign to P is $W = q'V(P)$. But this work,

Alexander Volta (1745–1827).

by definition, is the potential energy of the charge, and we therefore write

$$E_p = q'V \tag{14.16}$$

The SI unit of potential is known as the *volt* in honor of the Italian scientist Alexandro Volta (1745–1827). From Eq. 14.16 it is seen that *one volt* can be defined as *one joule per Coulomb.*

We can use the results of the preceding section to evaluate the potential resulting from a point charge q at a distance r away from it. If we equate Eqs. 14.9 and 14.16, we conclude that

$$V(r) = \frac{1}{4\pi\epsilon_0}\frac{q}{r} \tag{14.17}$$

In the preceding section, we saw that the work done in moving a charge in the resultant field of several charges could be found by summing the work done against the electric field independently produced by each charge. We therefore conclude that the potential resulting from several point charges is simply equal to the *algebraic* sum (remember that work is a scalar quantity) of the potential resulting from each charge. That is,

$$V = \frac{1}{4\pi\epsilon_0}\frac{q_1}{r_1} + \frac{1}{4\pi\epsilon_0}\frac{q_2}{r_2} + \cdots \tag{14.18}$$

where r_1, r_2, . . . are the distances from q_1 and q_2, respectively, to the point where the potential is being evaluated.

In electricity, like in mechanics, one is often interested in the *difference* in potential between two points rather than in the absolute value of the potential at a point. A potential difference between two points is commonly referred to as a *voltage difference* or simply *voltage.* This difference can be found by applying Eq. 14.18 to the two points in question and finding the difference. Alternatively, the potential difference can be calculated directly from the electric field. From Eq. 14.4 the work done in moving q' from A to B is

$$W_{A\to B} = -q' \int_A^B \mathscr{E} \cdot d\mathbf{s} \tag{14.4}$$

and by definition, this is equal to the difference in potential energy $E_p(B) - E_p(A)$. Using the relation between potential and potential energy, Eq. 14.16, we write

$$\frac{W_{A\to B}}{q'} = \Delta V = V(B) - V(A) = -\int_A^B \mathscr{E} \cdot d\mathbf{s}$$

We can eliminate the minus sign by inverting the limits of integration, that is,

$$\Delta V = V(B) - V(A) = \int_B^A \mathscr{E} \cdot d\mathbf{s} \tag{14.19}$$

Several practical conclusions may be drawn from Eq. 14.19. Consider two plates, B which is positively charged and A which is negatively charged (see

Fig. 14-7). As we saw in Section 14.2, the electric field is directed away from the positive charges and toward the negative charges. Thus in Fig. 14-7, $\mathscr{E}$ is directed from plate B to plate A. A unit of positive charge placed at B will be accelerated toward A. Noting that objects are accelerated when they move from a point to another of lower potential energy (recall the case of gravity), and remembering that by definition, the potential at a point is the potential energy of a unit of positive charge at that point, we conclude that $V(\text{B}) > V(\text{A})$. That is, the positively charged plate is at a higher potential than the negatively charged one. The result illustrated in this example can be generalized by stating that *the electric field is directed from high potential points to low potential points*, and that *positive charges, if free to move, do so from high potential points to low potential points*. For negative charges the opposite is true: A negative charge placed near plate A will be accelerated toward plate B; that is, *negative charges are accelerated from low potential points to high potential points*. In fact, not only can we say in what direction the charge will accelerate but we can calculate the velocity with which it will reach the other plate, if we know the potential difference between the plates. For this we use the conservation of total mechanical energy, which in this case can be written as

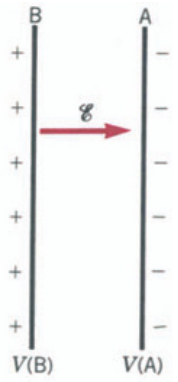


FIGURE 14-7

$$E_k(\text{B}) + qV(\text{B}) = E_k(\text{A}) + qV(\text{A}) \tag{14.20}$$

$E_k(\text{B}) + qV(\text{B}) = E_k(\text{A}) + qV(\text{A})$

That is, the sum of the kinetic and potential energies of a charge q at point B is equal to the sum of these energies at point A.

Example 14-3

A potential difference of 100 V is established between the two plates of Fig. 14-7, B being the high potential plate. A proton of charge $q = 1.6 \times 10^{-19}$ C is released from plate B. What will be the velocity of the proton when it reaches plate A? The mass of the proton is 1.67×10^{-27} kg.

Solution Because the proton is released with no initial velocity, $E_k(\text{B})$ is zero. From Eq. 14.20, we write

$$E_k(\text{A}) = q\,[V(\text{B}) - V(\text{A})] = q\,\Delta V$$

or

$$\frac{1}{2}mv_{\text{A}}^2 = q\,\Delta V$$

Solving for v_{A}

$$v_{\text{A}} = \sqrt{\frac{2q\,\Delta V}{m}}$$

$$= \sqrt{\frac{2 \times 1.6 \times 10^{-19}\ \text{C} \times 100\ \text{V}}{1.67 \times 10^{-27}\ \text{kg}}}$$

$$= 1.38 \times 10^5\ \text{m/sec}$$

14.5 THE ELECTRON VOLT

A useful unit of energy is the electron volt (eV). Because electric potential difference is the work required per coulomb, then $q\,\Delta V$ is the energy required to move a charge q through a voltage difference ΔV. The charge of the electron is $q = e = -1.6 \times 10^{-19}$ C (Section 13.4). If an electron is moved through a potential difference of 1 V (1 J/C) the energy change is

$$|e|\,\Delta V = 1.6 \times 10^{-19}\ \text{C} \times 1\ \text{J/C} = 1.6 \times 10^{-19}\ \text{J}$$

We define 1 electron volt (eV) as 1.6×10^{-19} J. Because the energies of electrons in atoms and solids are of the order of 10^{-19} J, the electron volt is a convenient unit of energy to use in these cases.

Suppose an electron is moved away from a positive charge through a potential difference of 100 V. The electron's potential energy has therefore increased by 100 eV. By energy conservation, if the electron is now released from this point it will acquire a kinetic energy of 100 eV when it arrives back at its starting point.

14.6 ELECTROMOTIVE FORCE

In the electric circuits that we will develop in the next chapter, the symbol ⊣|⊢ will be used and labeled with a voltage magnitude, such as 10 V. This symbol represents a battery that is a source of electrical potential energy. A battery is a contained chemical reaction. There are two types, the wet, or rechargeable type used in an automobile, and the dry type, which is used in flashlights. When the chemical reaction in a wet cell is exhausted, it can be recharged a number of times by sending current through it in the reverse direction. When the chemical reaction in a dry cell is exhausted, the battery is "dead." In both types there are two *electrodes* (plates or rods) whose exposed portions are called terminals, the *anode* and the *cathode*. These are suspended in an ionic solution in the wet cell and an ionic gel in the dry cell. The solutions and gels are called *electrolytes;* in wet cells the electrolyte is usually an acid. The anode is made of a material that strongly attracts the positive charges from the electrolyte whereas the cathode is made of a material that has a strong affinity for negative charges. As the anode and cathode attract their respective ions from the solution, they become electrostatically charged to the extent that they cannot attract further ions from the solution. In most electrolytes this balance of forces on the ions, that is, the attraction of the electrode versus the attraction of the solution occurs at around 1.5 V to 2.0 V. This is the approximate voltage between the terminals of a chemical battery. If a wire is connected between the terminals of a battery, charges can flow between the terminals. As the charges on the electrodes decrease in number, the chemical action inside the battery again takes place and charges again migrate from the electrolyte to the plates. In this way the battery main-

The car battery is an example of a rechargeable wet cell. The standard D, C, or AA batteries used in flashlights, portable radios and many types of toys belong to the dry type.

tains a constant potential difference between the plates. This type of potential difference is called an *electromotive force.* This we now know is a misnomer. There is no "force" between the plates, only an electric potential difference. We simply call it *emf.* Higher emfs are obtained by combining cells in series.

The symbol for battery mentioned earlier, ⊣|⊢, represents a single cell and is generally used for emfs of 1.5 V or *less*. The large line represents the anode as a source of positive charge, and the small line represents the negative side or the cathode. As mentioned before, early scientists assumed that the charges that flow were the positive ones. To this day we indicate charge flow, or current direction as emanating from the anode ⊣|⊢→ . This is called *conventional current*, and its use does not affect the results of common circuit calculations. When a circuit diagram is used in which there is an emf source of several volts the symbol ⊣|||||⊢ is used, which represents many batteries in series. The number of these lines depends on space available or persistence of the draftsman, and one should *not* count the batteries drawn to obtain the magnitude of the emf.

14.7 CAPACITANCE

Suppose we connect the terminals of a battery to two parallel metal plates, as in Fig. 14-8. The plate on the left will quickly attain a negative charge of $-q$ and the one on the right a positive charge of $+q$. The plates are characterized by having a charge q, the magnitude of the charge on either of them. It is evident that if the emf of the battery is small, the charge q on the plate will be small and if the emf is large, the charge q will be large. Experiments show that the charge is proportional to the potential difference, ΔV or emf,

$$q \propto V$$

where V actually means ΔV or voltage difference between the two terminals of the battery. We make this relation into an equality by introducing a constant C so that

$$q = CV \tag{14.21}$$

The arrangement of such a set of plates as in Fig. 14-8 is called a *capacitor*, and the constant C is called the *capacitance.* The constant has units of

$$C = \frac{q}{V} \frac{\text{(coulomb)}}{\text{(volt)}}$$

which is given the special name of *farad* (abb. F) where

$$1 \text{ farad} = 1 \text{ coulomb/volt}$$

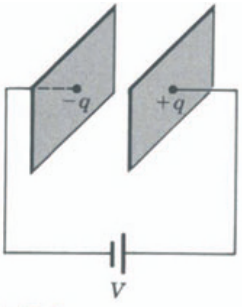


FIGURE 14-8
Two parallel metal plates separated by an insulator, such as air, form a capacitor. When connected to a source of potential difference, the metal plates acquire equal but opposite charges q and $-q$.

A farad is a very large quantity, and the usual capacitor in an electronic circuit is of the order of microfarads (1 μF = 10^{-6} F) or picofarads (1 pF = 10^{-12} F). The symbol for a capacitor in an electric circuit is $\dashv\vdash$.

Equation 14.21 represents the charge on a capacitor in a vacuum, and in air there is very little change. Suppose that some nonconducting material, either liquid or solid, is placed between the plates. It is found experimentally that the capacitor will have a higher charge for the same voltage by a factor κ. The material placed between the plates is called a *dielectric* and the factor κ is called the *dielectric constant*. Therefore, Eq. 14.21 is written as

$$q = \kappa CV \tag{14.22}$$

where κ for air or vacuum is unity (1). Some example of the values of κ are given in Table 14-1.

TABLE 14-1

Material	Dielectric Constant
Vacuum	1
Paper	3.5
Rubber	7

We can also see from Eq. 14.22 that if we wish to maintain a given charge q, less voltage is required with the dielectric present. That is, if it requires V_0 volts to produce a charge q on the capacitor in a vacuum, then when a dielectric is introduced the same charge can be produced by a voltage $V = V_0/\kappa$.

PROBLEMS

14.1 (a) What is the electric field at a point 0.12 m from a point charge of -4×10^{-9} C? (b) What force would an electron experience if it were placed at that point?

14.2 Two equal and opposite charges, $q_1 = 3 \times 10^{-6}$ C and $q_2 = -3 \times 10^{-6}$ C, are held 10 cm apart. (a) What is the electric field at the midpoint of the line joining the two charges? (b) What force would an electron experience if it were placed there?

14.3 The two charges of problem 14.2 are placed on the x axis as shown in Fig. 14-9. What is the electric field at a point on the y axis located at a distance y from the origin?

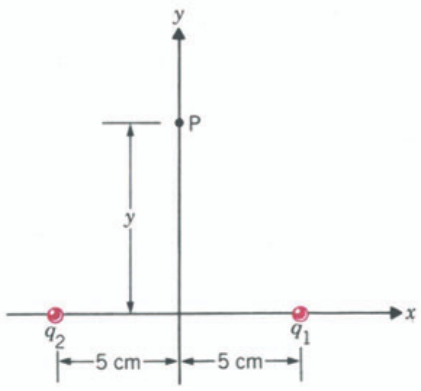


FIGURE 14-9
Problem 14.3.

14.4 Two charges, $q_1 = 7 \times 10^{-8}$ C and $q_2 = -14 \times 10^{-8}$ C, are placed on the x axis as shown in Fig.

14-10. (a) Find the points where the electric field is zero. (b) What is the magnitude and the direction of the electric field at point P with coordinates $x = 0$, $y = 20$ cm?

(*Answer*: (a) 36.2 cm to the left of q_1, (b) 1.70×10^4 N/C, $\theta = -28.8°$.)

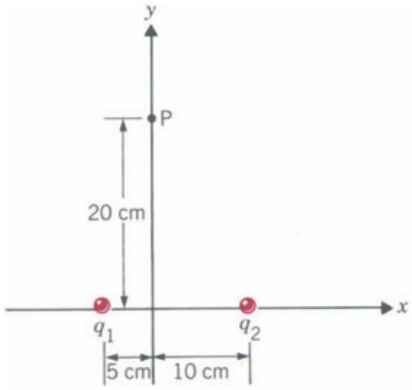


FIGURE 14-10
Problem 14.4.

14.5 Consider the arrangement of charges shown in Fig. 14-11. What is the electric field at point A?

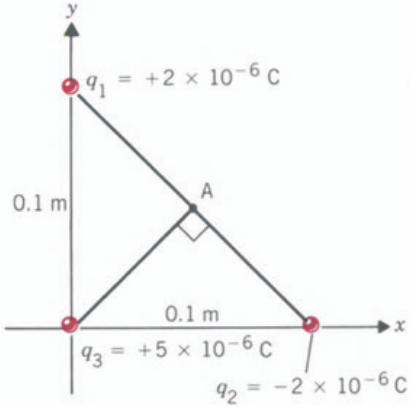


FIGURE 14-11
Problem 14.5.

14.6 Four charges of equal magnitude are placed at the corners of a square as shown in Fig. 14-12. What is the electric field at the center of the square, point O?

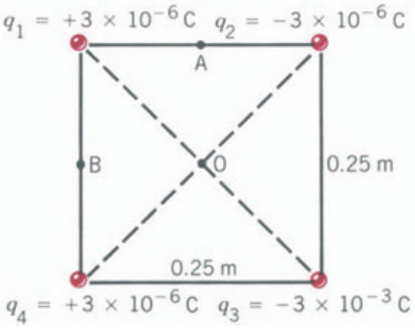


FIGURE 14-12
Problems 14.6 and 14.7.

14.7 Consider the charge configuration of problem 14.6. (a) What is the electric field at point A? (b) What is the electric field at point B?

(*Answer*: (a) 3.77×10^6 N/C directed toward q_2, (b) 6.19×10^5 N/C directed toward point O.)

14.8 Two large parallel plates are separated by a distance of 5 cm. The plates have equal but opposite charges that create an electric field in the region between the plates. An α particle ($q = 3.2 \times 10^{-19}$ C, $m = 6.68 \times 10^{-27}$ kg) is released from the positively charged plate, and it strikes the negatively charged plate 2×10^{-6} sec later. Assuming that the electric field between the plates is uniform and perpendicular to the plates, what is the strength of the electric field?

(*Answer*: 522 N/C.)

14.9 An electron is projected with an initial velocity $v_i = 3 \times 10^6$ m/sec in the x direction in the region between two oppositely charged plates (see Fig. 14-13). By the time the electron leaves the region between the plates, it has undergone a vertical deflection of 2 cm. Assume that the electric field between the plates is uniform and perpendicular to the plates and that the electric field outside the region of the plates is zero. (a) What is the strength of the electric field between the plates? (b) At what point y_f on a screen 1 m away from the plates will the electron land?

(*Answer*: (a) 8.19 N/C, (b) 10 cm.)

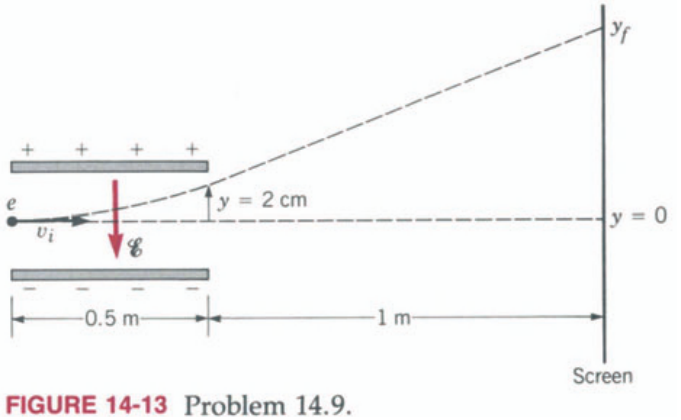


FIGURE 14-13 Problem 14.9.

14.10 A uniform electric field $\mathscr{E} = 500$ N/C exists in the region between two oppositely charged plates (see Fig. 14-14). How much work is done in moving a charge $q = 6 \times 10^{-6}$ C from A to P with constant

velocity (a) along path ABP, (b) along path ADP, (c) along the straight line path AP?

(*Answer*: (a) 1.5×10^{-3} J, (b) same as (a), (c) same as (a).)

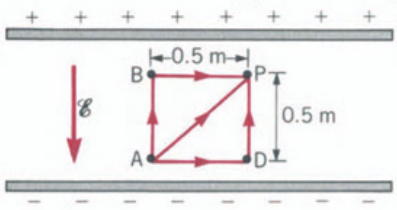


FIGURE 14-14
Problem 14.10.

14.11 The electric field between two parallel plates is uniform and perpendicular to the plates. The potential difference between the plates is 100 V, and the separation between the plates is 1 cm. What is the strength of the electric field between the plates?

(*Answer*: 10^4 N/C.)

14.12 What is the potential difference between the plates in problem 14.8?

(*Answer*: 18.3 V.)

14.13 A charge $q_1 = 3 \times 10^{-6}$ C is brought from infinity to the origin of a set of coordinate axes. A second charge $q_2 = 2 \times 10^{-6}$ C is brought also from infinity to a point with coordinates $x = 5$ cm, $y = 0$ cm. (a) How much work is done in bringing q_1? (b) How much work is done in bringing q_2? (c) What is the potential at $x = 2.5$ cm, $y = 0$ cm? (d) How much work is done in bringing an electron from infinity to the point $x = 2.5$ cm, $y = 0$ cm after q_1 and q_2 have been placed at the locations indicated above?

14.14 (a) What is the potential at point P in Fig. 14.9 of problem 14.3? (b) How much work must be done to move an electron from point P to the origin?

14.15 (a) What is the potential at point O in problem 14.6 (see Fig. 14.13)? (b) What is the potential energy of a charge $q = 1 \times 10^{-6}$ C when it is placed at point O? (c) How much work must be done in bringing q from infinity to point O? (d) How much work must be done to move q from point O to point A?

(*Answer*: (a) 0 V, (b) 0 V, (c) 0 V, (d) 0 V.)

14.16 Charged particles are accelerated through a potential difference of 250 V. What will be the kinetic energy in eV if the particle is (a) an electron, (b) a proton, (c) an α particle ($q = +2e$), (d) a gold nucleus ($q = +79e$). e is the magnitude of the charge of the electron and is 1.6×10^{-19} C.

14.17 In a given vacuum tube, an electron is released from the heated filament with zero velocity. It is attracted by the positive plate and arrives at the plate with a velocity of 4×10^6 m/sec. What is the voltage of the plate with respect to the filament? The mass of the electron is 9.1×10^{-31} kg.

14.18 An α particle ($q = +2e$) is shot directly toward a gold nucleus ($q' = +79e$) with a kinetic energy $E_k = 6$ MeV (6×10^6 eV). How close does the α particle get to the gold nucleus? Assume that the gold nucleus remains stationary.

(*Answer*: 3.79×10^{-14} m.)

14.19 A particle with charge $q_1 = 4 \times 10^{-6}$ C is held fixed at some point in space. A second particle of mass 20 g and charge $q_2 = -5 \times 10^{-6}$ C is placed 3 cm away from the first particle. What velocity must be given to q_2 so that it will reach infinity with zero velocity?

14.20 Two protons are held fixed 10 cm apart. A third proton is projected from far away with some initial velocity v_i as shown in Fig. 14-15. (a) What is the minimum value of v_i that will allow the proton to reach the midpoint of the line joining the two fixed protons? (b) If the initial velocity is half the value found in part (a), how close to point 0 will the proton get before it stops?

(*Answer*: (a) 3.32 m/sec, (b) 1.94×10^{-2} m.)

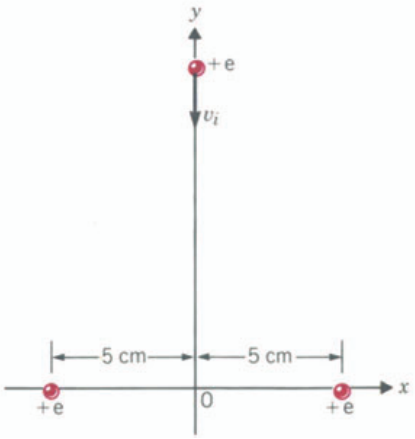


FIGURE 14-15
Problem 14.20.

14.21 An electron is placed midway between two fixed charges, $q_1 = 2.5 \times 10^{-10}$ C and $q_2 = 5 \times 10^{-10}$ C. If the charges are 1 m apart, what is the velocity of the electron when it reaches a point 10 cm from q_2?

14.22 Two particles are placed 1 m apart. Particle 1 has a mass $m_1 = 20$ g and a charge $q_1 = 6 \times 10^{-6}$ C. Particle 2 has a mass $m_2 = 50$ g and a charge $q_2 = -4 \times 10^{-6}$ C. The particles are released from rest simultaneously. (a) What will be the velocities of the particles when they are 0.5 m apart? (b) At what distance from the initial position of particle 1 will the collision occur?

(*Answer*: (a) $v_1 = 3.93$ m/sec, $v_2 = 1.57$ m/sec, (b) 0.714 m.)

14.23 An electric dipole consists of two charged particles of mass $m = 300$ g and charge $q_1 = 3 \times 10^{-5}$ C and $q_2 = -3 \times 10^{-5}$ C connected by a rigid rod of negligible weight and length $d = 20$ cm. The dipole is placed in a region where there is a uniform electric field $\mathscr{E} = 5000$ N/C (see Fig. 14-16). (a) What is the torque exerted by the electric field when $\theta = 30°$? (b) How much work must be done to rotate the dipole from the angular position $\theta = 0°$ to $\theta = 90°$? (c) If the dipole is pivoted about its midpoint and is released from the angular position $\theta = 90°$, what will be the angular velocity of the dipole when it swings back to $\theta = 0°$?

(*Answer*: (a) 1.5×10^{-2} N-m, (b) 3×10^{-2} J, (c) 3.16 rad/sec.)

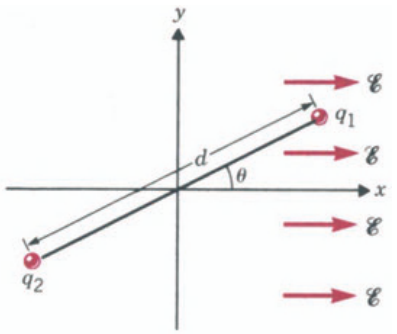
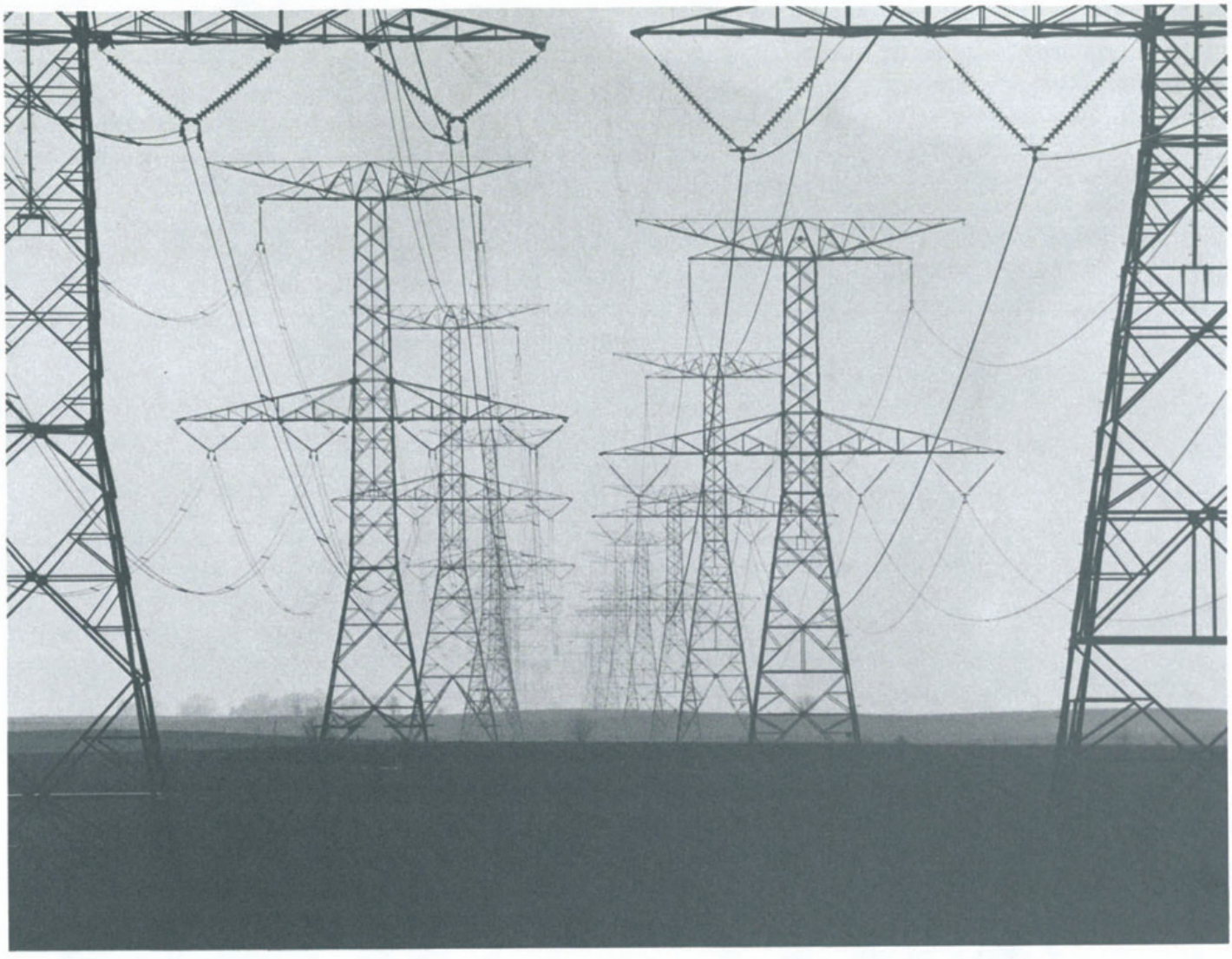


FIGURE 14-16
Problem 14.23.



CHAPTER 15

Electric Current

15.1 INTRODUCTION

In this chapter we consider the motion of electrons in a conductor (a metal) when there is a voltage difference applied between the ends of the conductor. In more common language, how will the electrons in a metal wire behave if the wire is attached to the two opposite terminals of a battery whose potential difference creates an electric field in the wire? After consideration of the basic behavior, we will treat combinations of wires, batteries, and electrical measuring instruments. It will be necessary to understand the rules for charge flow, a *current*, in circuits in order to follow the arrangement of logic circuits in a computer, which will be treated in Chapter 27. We will limit our discussion mostly to *direct currents*, that is, currents whose magnitude and direction do not change with time.

15.2 MOTION OF CHARGES IN AN ELECTRIC FIELD

We have seen in Chapter 14 that the definition of electric field strength $\mathscr{E}$ is the force per unit positive charge

$$\mathscr{E} = \frac{\mathbf{F}}{q}$$

We may substitute Newton's second law $\mathbf{F} = m\mathbf{a}$

$$\mathscr{E} = \frac{m\mathbf{a}}{q}$$

or

$$\mathbf{a} = \frac{q\mathscr{E}}{m} \tag{15.1}$$

Note that in Eq. 15.1 q represents an arbitrary charge. In the case of an electron, $q = e$, where $e = -1.6 \times 10^{-19}$ C and the mass of an electron is $m = 9.1 \times 10^{-31}$ kg. Equation 15.1 will now be written as

$$\mathbf{a} = \frac{-|e|\mathscr{E}}{m} \tag{15.2}$$

in which the negative sign tells us that the direction of acceleration of an electron is opposite to that of the field direction. Unless there is a specific need to know the direction of motion, we may just use the magnitude of the acceleration in Eq. 15.2. For example, suppose a constant electric field is suddenly applied to a metal. What velocity will the electrons have after traveling a distance s, assuming that no scattering (or collisions) occurs over that distance? This is simply a problem from Chapter 3 in which we wish to find a final velocity when the initial velocity is zero and the displacement and the

André Marie Ampère (1775–1836).

constant acceleration are known. From Eq. 3.11, we write

$$v^2 - v_0^2 = 2ax$$

When

$$v_0 = 0, \quad x = s$$

and

$$a = \frac{e\mathscr{E}}{m}$$

the velocity is

$$v = \sqrt{\frac{2e\mathscr{E}s}{m}}$$

Note that in this expression for v, e stands for the magnitude of the charge of the electron.

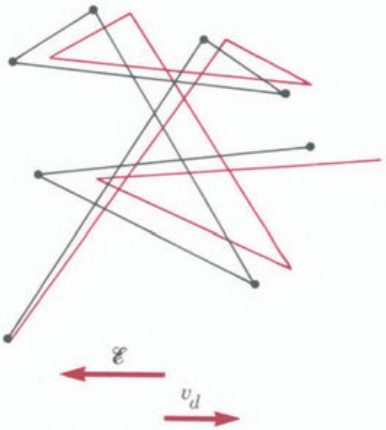


FIGURE 15-1
The random path (black lines) of an electron resulting from collisions with the ions and the effect of an electric field on the path, with a resulting drift velocity (colored lines). [*Source*: David Halliday and Robert Resnick, *Fundamentals of Physics*, 2nd ed. Copyright © by John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.]

15.3 ELECTRIC CURRENT

In the preceding section we saw that an electric field in a conductor can cause charges to undergo accelerated motion. This acceleration is terminated by collisions of the electron with the atoms in the conductor. That is, the motion of an electron in an electric field is a series of short accelerations interrupted by collisions that scatter the electron. It therefore has a random path, although there is a slow net velocity opposite to the field direction, such as illustrated schematically in Fig. 15-1. It is the net velocity of the electrons, called the *drift velocity*, that gives rise to the current, not the brief accelerations.

If we stand at a particular plane perpendicular to a wire and count the charge Δq that flows by in time Δt we define this as electric current i, where

$$i = \frac{\Delta q}{\Delta t} \text{ C/sec} \tag{15.3}$$

The definition of the electric current given by Eq. 15.3 holds only if the rate of charge flow is constant. In the general case where i is not constant, we define it as

$$i = \frac{dq}{dt} \tag{15.3'}$$

$$i = \frac{dq}{dt}$$

The concept of electric current is similar to the concept of measuring the current in a river. One measures the quantity of water, in gallons, cubic feet, or such, which flows past a point in a given time.

Because the charge Δq is a scalar quantity, the current i is also a scalar. In the SI units, current is measured in amperes, or amps, after André Ampère (1775–1836). One ampere (1 A) is equal to one coulomb per second and is a

relatively large quantity. Usually in electronic circuits the current is much smaller than 1 A, and we use the milliampere (1 mA = 10^{-3} A) or the microampere (1 μA = 10^{-6} A)

Consider now a cylindrical conductor with a cross-sectional area A as in Fig. 15-2, and let us assume that there are both positive and negative charges, both of which are mobile in the presence of an electric field $\mathscr{E}$ with a vector direction from left to right. Let us further assume that there are N_p positive charges per unit volume with drift velocity of v_p and N_n negative charges with drift velocity of v_n. In time Δt the positive charges will move from left to right a distance of $v_p \Delta t$. Therefore, in time Δt, all the positively charged particles within the shaded region of the cylinder of cross-section A and length $v_p \Delta t$, and only those particles, will flow out of the shaded region of the cylinder to the right. The volume of the shaded cylinder is $Av_p \Delta t$, and the number of positive particles within is the number per unit volume N_p times the volume or $N_p A v_p \Delta t$; if each has a charge q_p, the charge flowing across the right end of the cylinder is

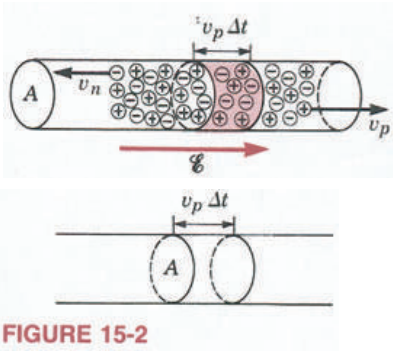


FIGURE 15-2

$$\Delta q_p = q_p N_p A v_p \Delta t$$

Substituting this into Eq. 15.3 gives the current resulting from the positively charged particles as

$$i_p = \frac{\Delta q_p}{\Delta t}$$

$$= \frac{q_p N_p A v_p \Delta t}{\Delta t}$$

$$i_p = q_p N_p A v_p \qquad (15.4)$$

In the same way, the negative particles, each with charge q_n, flow from right to left giving rise to a current

$$i_n = q_n N_n A v_n \qquad (15.5)$$

It is seen in Eq. 15.4 that the current i_p of positive charges is to the right in Fig. 15-2 because both the sign of the charge q_p and their drift velocity are positive and, hence, their product is positive. The current i_n resulting from negative charge motion also results in an effective current of positive charges to the right by subtraction of the negative charges. This is seen in Eq. 15.5 in which both the sign of the charge q_n and the sign of the drift velocity v_n are negative and therefore their product is positive: *A flow of negative charges to the left is equivalent to a flow of positive charges to the right*. When both positive and negative charges move, the total current i is the *sum* of these two currents or

$$i = i_p + i_n$$

$$i = A(q_p N_p v_p + q_n N_n v_n) \qquad (15.6)$$

$i = A(q_p N_p v_p + q_n N_n v_n)$

It is seen by Eq. 15.6 that a measure of electric current does not tell whether the current is being carried by positive or by negative charges or by both.

The vector drift velocity of the positive charge carriers $\mathbf{v}_p$ is in the same vector direction as that of the electric field vector $\mathscr{E}$. The direct current i in a conductor has the same direction as that of the electric field $\mathscr{E}$.

We note that as much charge flows into the cylinder of Fig. 15-2 as flows out. There is no pileup of electric charges in the wire at any point. If there were, then the local electric field would be stronger at that point which would increase the net flow of charge past that point until the charge density at every point in the wire would again be equal. If we connect a wire between the terminals of a battery, it is therefore reasonable to conclude that charge flows at a steady rate throughout the wire.

It is often convenient to introduce a quantity called the current density, with symbol J, which is the current per unit cross-sectional area. This eliminates the area A from Eq. 15.6; thus

$$J = \frac{i}{A} \text{ A/m}^2 \text{ (amp/m}^2\text{)} \qquad (15.7)$$

$$J = \frac{i}{A}$$

Example 15-1

Suppose a copper wire carries 10 A (amps) of current and has a cross-section of 10^{-6} m^2. As will be seen later, each atom of copper contributes one electron that is free to move, so the electron carrier density N_n is about the same as the density of atoms, which is about 7×10^{28} atoms per m^3 (see problem 9.2). The charge on an electron is -1.6×10^{-19} C. (a) What is the drift velocity v_n of the electrons? (b) How long would it take an electron to move from one terminal of a battery to the other if this wire were 1 m long?

Solution

(a) $i = A\, q_n\, N_n\, v_n$

$$v_n = \frac{i}{A\, q_n\, N_n}$$

$$= \frac{10 \text{ A}}{10^{-6} \text{ m}^2 \times 1.6 \times 10^{-19} \text{ C} \times 7 \times 10^{28} \text{ m}^{-3}}$$

$$= 9 \times 10^{-4} \text{ m/sec}$$

(b) $$t = \frac{x}{v_n} = \frac{1 \text{ m}}{9 \times 10^{-4} \text{ m/sec}} = 1.1 \times 10^3 \text{ sec} = 18 \text{ min}$$

So the actual drift velocity of a given electron is very small. However, when we turn on a light switch, the lamp will light almost immediately regardless of the distance from switch to lamp. The reason is that the speed of propagation of the electric field along the wire is that of the speed of light in the

wire. Therefore, all electrons are acted on almost simultaneously by the electric field and begin to drift.

15.4 RESISTANCE AND RESISTIVITY

We have seen that the electric field $\mathscr{E}$ is the force per unit charge and, hence, the field causes the charges to be accelerated. The collisions with atoms scatter the charges, which are then accelerated again. This accelerating-scattering process gives rise to a net drift velocity, resulting in an electrical current in the direction of the field $\mathscr{E}$. Experiment shows that in many cases the electric current i, hence the current density J, are proportional to $\mathscr{E}$.

$$J \propto \mathscr{E}$$

We can change this proportionality to an equality by introducing a quantity ρ, called the *electrical resistivity*

$$\mathscr{E} = \rho J \tag{15.8}$$

$\mathscr{E} = \rho J$

This resistivity is a property of a given material and is independent of its shape. The resistivity was found to be a constant for a given metal at a given temperature by George Ohm (1789–1854); Eq. 15.8 is called Ohm's law. A material obeying Ohm's law is called an *ohmic conductor*, that is, one with a linear relation between current density and electric field. A material with a nonlinear relationship is called a *nonohmic conductor*. For example, in Chapter 26, we will see that a linear dependence does not hold in the case of a circuit element called the diode.

From Eq. 15.8 the units of ρ may be determined

$$\rho = \frac{\mathscr{E}(\text{N/C})}{J(\text{C/sec-m}^2)} = \frac{\mathscr{E}}{J}\left(\frac{\text{N-sec-m}^2}{\text{C}^2}\right)$$

This is such a cumbersome unit that it is shortened to Ω-m (ohm meter).

Some typical values of ρ for conductors and insulators are given in Table 15-1.

TABLE 15-1

Material	ρ (Ω-m) at room temperature
Silver	1.5×10^{-8}
Copper	1.7×10^{-8}
Aluminum	2.7×10^{-8}
Glass	$\sim 10^{+11}$
Teflon	$\sim 10^{+14}$
Dry wood	$\sim 10^{+11}$

It is seen that differences in resistivity between insulators and conductors can be as great as 22 orders of magnitude (powers of 10). One of the early

George Ohm (1787–1854).

successes of the theory of solids was to explain this. We will develop this theory in Chapters 23 and 24.

It is sometimes convenient to use the reciprocal of the resistivity; this is called the *conductivity* and has the symbol σ. Equation 15.8 may be written as

$$J = \sigma \mathscr{E} \tag{15.9}$$

$J = \sigma\mathscr{E}$

where $\sigma = 1/\rho$.

Suppose we have a given metal wire with cross section A, length l, and resistivity ρ with an applied electric field $\mathscr{E}$. The wire is shown schematically in Fig. 15-3. We can use Eq. 14.19 to relate the electric field inside the conductor to the potential difference between the two ends of the conductor, points 1 and 2.

$$\Delta V = V_1 - V_2 = \int_{s_1}^{s_2} \mathscr{E} \cdot d\mathbf{s} \tag{14.19}$$

If the electric field inside the conductor is uniform, the integral becomes $\mathscr{E}l$, and

$$\Delta V = \mathscr{E}l$$

where

$$l = s_2 - s_1$$

or

$$\mathscr{E} = \frac{\Delta V}{l} \tag{15.10}$$

Substitution of Eq. 15.10 for $\mathscr{E}$ and Eq. 15.7 for J in Eq. 15.8 yields

$$\Delta V = i\frac{\rho l}{A}$$

which is written

$$V = iR \tag{15.11}$$

$V = iR$

where V actually means ΔV or voltage difference between the two ends of the wire. This equation (15.11) is also commonly called *Ohm's law*. In this equation $R = \rho l/A$ and is called the *resistance* of the wire and has units of Ω (ohms). It is seen that the longer the wire the more resistance it has to the current, but the larger its cross-sectional area, the less resistance it has. The analogy to water flowing through a pipe is useful. Voltage will be the equivalent of the difference in water pressure and current that of volume per second of flowing water.

We should note an important fact about the direction of the current through a resistance. At the beginning of this chapter, Section 15.3, we indicated that the current has the same direction as that of the electric field. In Chapter 14, Section 14.4, we demonstrated that the electric field is directed

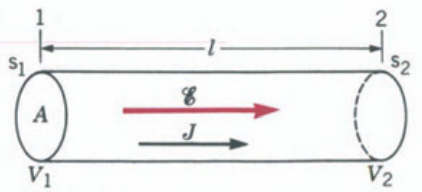


FIGURE 15-3

from high potential points to low potential points. We therefore conclude that *the current in a resistance is from its high potential side to its low potential side.*

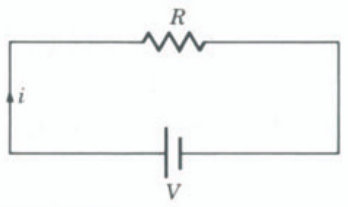


FIGURE 15-4

15.5 RESISTANCES IN SERIES AND PARALLEL

In this and succeeding sections we will study some simple electrical circuits. Resistances will be drawn with the symbol —/\/\/\— and emf sources will be drawn as —|⊢— for a small emf and —|||⊢— for a larger emf. In all cases, the large line represents the positive, or higher potential, side of the emf. A simple direct current (dc) circuit with one emf source and one resistance is drawn as in Fig. 15-4, where the arrow represents the direction of the current.

In Fig. 15-4 and in the following circuit diagrams the emf sources are labeled with the letter V and the voltage magnitude is given. This is the conventional labeling, but it does not conform well to the definitions. If one measures the potential difference between the terminals of a battery, this is ΔV but, as already mentioned, it is customary to refer to a voltage difference simply as the "voltage." We have also seen that the accepted name for such a voltage source is emf. Confusion in terminology often arises for the student when the form of Ohm's law of Eq. 15.11 is used. The V in this equation is the potential, or voltage drop across a resistance R when a current i passes through it. If one takes a meter that measures electric potential difference, called a *voltmeter*, it will read a voltage V when the probes are placed on opposite sides of the resistance. The concept of this voltage difference, commonly called a *voltage drop*, across a resistance should not be confused with the potential difference across the emf source.

Suppose that we replace the single resistance of Fig. 15-4 with three resistances (called *resistors*) of different values R_1, R_2, and R_3, as in Fig. 15-5*a*. We assume in this type of calculation that the connecting wires have zero resistance. Therefore, the electric potential at point A is the same as that at the left side of the battery, and that at point D is the same as the right side of the battery. The same current must pass through each of these resistances as that which passes between points A and D. This combination is therefore called *series* resistances because the current passes through each sequentially. By Ohm's law (Eq. 15.11) we may write the voltage drop across each resistance as

$$V_{AB} = iR_1,\ V_{BC} = iR_2,\ V_{CD} = iR_3$$

and, because the sum of these voltage drops must equal the potential difference between A and D, V_{AD}, which is the emf of the battery, we conclude that

$$\begin{aligned} V &= V_{AB} + V_{BC} + V_{CD} \\ &= iR_1 + iR_2 + iR_3 = i(R_1 + R_2 + R_3) \\ V &= iR_{eq} \end{aligned}$$

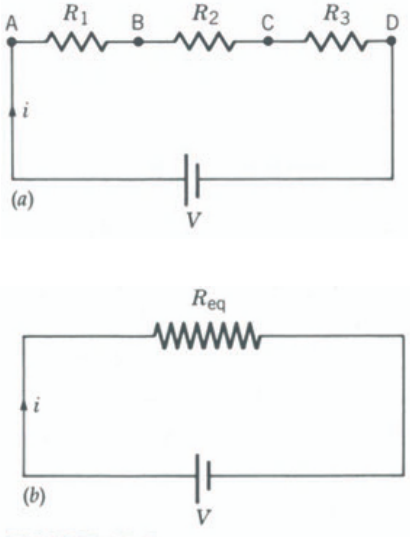


FIGURE 15-5

(*a*) Three resistors in series. (*b*) Equivalent resistor R_{eq}.

where R_{eq} is the equivalent resistance of the three. That is, the same current would flow through the circuit if the three resistances were replaced by a single one (Fig. 15-5*b*) with magnitude

$$R_{eq} = R_1 + R_2 + R_3 \qquad \text{(series)} \tag{15.12}$$

$R_{eq} = R_1 + R_2 + R_3 + \ldots$

It is obvious that the generalization of Eq. 15.12, namely, $R_{eq} = \sum_i R_i$, will be true regardless of the number of resistances in series.

Example 15-2

Suppose in Fig. 15-5 the voltage $V = 1.5$ V and the resistances are $R_1 = 5\ \Omega$, $R_2 = 10\ \Omega$, and $R_3 = 15\ \Omega$. What are the voltages V_{AB}, V_{BC}, and V_{CD}?

Solution First we find the current through the resistors by replacing the individual resistances with a simple equivalent resistance.

$$V = i\,R_{eq} = i\,(R_1 + R_2 + R_3)$$

$$i = \frac{1.5\text{ V}}{(5 + 10 + 15)\ \Omega} = 0.05\text{ A} = 50\text{ mA}$$

Then, applying Ohm's law to each resistance

$$V_{AB} = iR_1 = 0.05\text{ A} \times 5\ \Omega = 0.25\text{ V}$$

$$V_{BC} = iR_2 = 0.05\text{ A} \times 10\ \Omega = 0.50\text{ V}$$

$$V_{CD} = iR_3 = 0.05\text{ A} \times 15\ \Omega = \underline{0.75\text{ V}}$$

$$\text{sum} = V_{AD} = 1.5\text{ V}$$

Suppose we now arrange these resistances in *parallel*, as in Fig. 15-6*a*. As stated previously, we assume the resistance of connecting wires to be negligible and, for clarity, we redraw the circuit of Fig. 15-6*a* in the form of Fig. 15-6*b*. Because all connecting wires are considered to have zero resistance, there can be no voltage drop across them. Therefore, the left side of each resistance is at the same potential and the right side is at the same potential; hence, the same voltage drop V must occur across each. We further note that although the current through each resistance may be different, the sum of the individual currents must equal the current that flows through the wire connecting them to the battery because charge must be conserved. Thus,

$$i = i_1 + i_2 + i_3 \qquad \text{(parallel)} \tag{15.13}$$

Because each resistance has the same voltage drop V across it, we may write Ohm's law for each

$$V = i_1R_1, \quad V = i_2R_2, \quad V = i_3R_3$$

and

$$i_1 = \frac{V}{R_1}, \quad i_2 = \frac{V}{R_2}, \quad i_3 = \frac{V}{R_3}$$

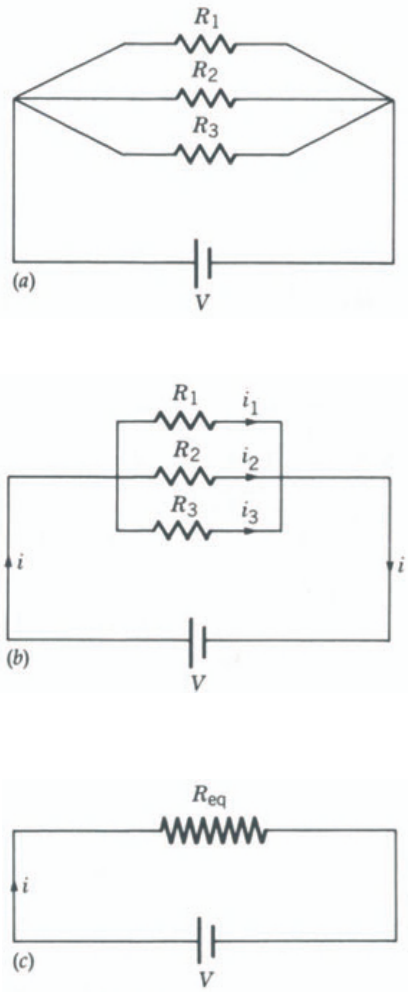


FIGURE 15-6
(*a*) Three resistors in parallel. (*b*) Conventional form of drawing resistors in parallel. (*c*) Equivalent resistor R_{eq}.

Substituting these into Eq. 15.13 gives

$$i = \frac{V}{R_1} + \frac{V}{R_2} + \frac{V}{R_3}$$

$$= V\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right)$$

or

$$V = \frac{i}{\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right)}$$

$$= \frac{i}{\frac{1}{R_{eq}}} = i\,R_{eq}$$

where

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \qquad \text{(parallel)} \qquad (15.14)$$

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$$

We see that any number of parallel resistors can be replaced with an equivalent resistor by a generalization of the relation of Eq. 15.14. The equivalent circuit looks like Fig. 15-6*c*, where R_{eq} is given by Eq. 15.14.

Example 15-3

Suppose two resistors, $R_1 = 5\ \Omega$ and $R_2 = 10\ \Omega$, are connected in parallel to a 1.5-V battery as in Fig. 15-7*a*. (a) What is the current through each? (b) What is the total current in the circuit?

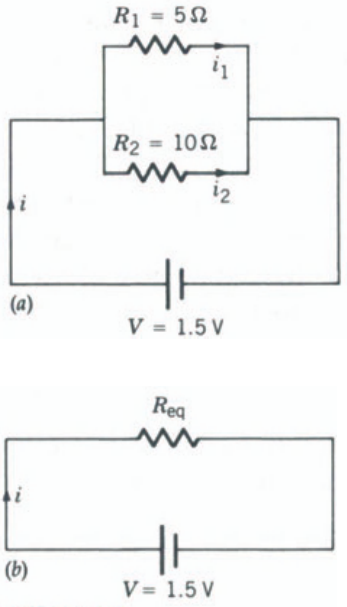


FIGURE 15-7
Example 15-3.

Solution

(a) Using Ohm's law (Eq. 15.11)

$$V = i_1R_1, \quad V = i_2R_2$$

$$i_1 = \frac{V}{R_1} = \frac{1.5\text{ V}}{5\ \Omega} = 0.3\text{ A} = 300\text{ mA},$$

$$i_2 = \frac{V}{R_2} = \frac{1.5\text{ V}}{10\ \Omega} = 0.15\text{ A} = 150\text{ mA}$$

(b) $i = i_1 + i_2 = 300\text{ mA} + 150\text{ mA} = 450\text{ mA}$

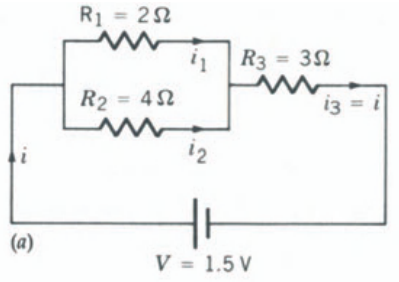
We may check this answer by solving the equivalent circuit, Fig. 15-7*b*.

$$\frac{1}{R_{eq}} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{1}{5\ \Omega} + \frac{1}{10\ \Omega} = 0.2\ \Omega^{-1} + 0.1\ \Omega^{-1} = 0.3\ \Omega^{-1}$$

or

$$R_{eq} = \frac{1}{0.3\ \Omega^{-1}} = 3.33\ \Omega$$

$$i = \frac{V}{R_{eq}} = \frac{1.5\ \text{V}}{3.33\ \Omega} = 0.45\ \text{A} = 450\ \text{mA}$$



Example 15-4

Three resistors are connected in a combination of series and parallel as in Fig. 15-8*a*. What is the current through each?

Solution First we find $R_{eq(p)}$ for the parallel combination

$$\frac{1}{R_{eq(p)}} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{1}{2\ \Omega} + \frac{1}{4\ \Omega} = 0.5\ \Omega^{-1} + 0.25\ \Omega^{-1} = 0.75\ \Omega^{-1}$$

$$R_{eq(p)} = 1.33\ \Omega$$

We then have the equivalent circuit, Fig. 15-8*b*. We now find the equivalent series resistance $R_{eq(s)}$

$$R_{eq(s)} = R_{eq(p)} + R_3 = 1.33\ \Omega + 3\ \Omega = 4.33\ \Omega$$

We now have the simpler equivalent circuit of Fig. 15-8*c*. The current is given by Ohm's law

$$i = \frac{V}{R_{eq(s)}} = \frac{1.5\ \text{V}}{4.33\ \Omega} = 0.35\ \text{A} = 350\ \text{mA}$$

We may return to Fig. 15-8*b* and note that we have already solved part of the problem because all this current flows through R_3, hence, $i_3 = 350$ mA. We may find the voltage drop across the parallel combination by use of Ohm's law

$$V_{(p)} = iR_{eq(p)} = 350\ \text{mA} \times 1.33\ \Omega = 0.47\ \text{V}$$

The current through each of the parallel resistors is then

$$i_1 = \frac{V_{(p)}}{R_1} = \frac{0.47\ \text{V}}{2\ \Omega} = 0.235\ \text{A} = 235\ \text{mA}$$

$$i_2 = \frac{V_{(p)}}{R_2} = \frac{0.47\ \text{V}}{4\ \Omega} = 0.118\ \text{A} = 118\ \text{mA}$$

And, except for the rounding-off error, $i_1 + i_2 = i_3 = i$

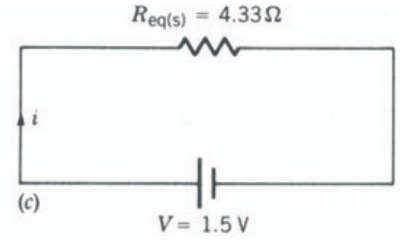


FIGURE 15-8

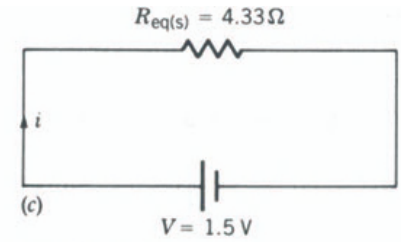


FIGURE 15-8
Example 15-4.

You will have noticed that when there are two resistors in series or in parallel, a method of ratios exists as a quicker way of solution. Consider first a series circuit as in Fig. 15-9. The current through R_1 is the same as that

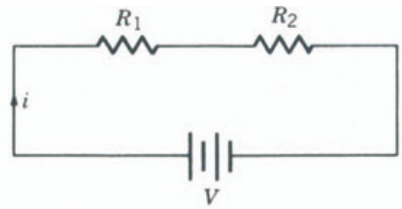


FIGURE 15-9

through R_2, and by Ohm's law

$$i = \frac{V_1}{R_1}, \quad i = \frac{V_2}{R_2}$$

where V_1 and V_2 are the voltage drops across R_1 and R_2, respectively. Equating the i's gives

$$\frac{V_1}{R_1} = \frac{V_2}{R_2}$$

or

$$\frac{V_1}{V_2} = \frac{R_1}{R_2} \qquad \text{(series)} \qquad (15.15)$$

so that *in a series circuit the ratio of the voltage drops is equal to the ratio of the resistances.*

A different ratio can be written for parallel resistors (see Fig. 15-10). In this situation we recall that the voltage across each is the same. From Ohm's law

$$V_1 = i_1 R_1, \quad V_2 = i_2 R_2$$

Equating V_1 and V_2 gives

$$i_1 R_1 = i_2 R_2$$

or

$$\frac{i_1}{i_2} = \frac{R_2}{R_1} \qquad \text{(parallel)} \qquad (15.16)$$

where we see that *in a parallel circuit the ratio of the currents through each resistor is inversely proportional to the resistances.* We might have expected this result from the understanding that resistance impedes the flow of current; hence, the larger the resistance the lower the current.

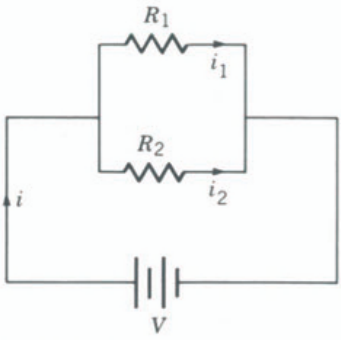


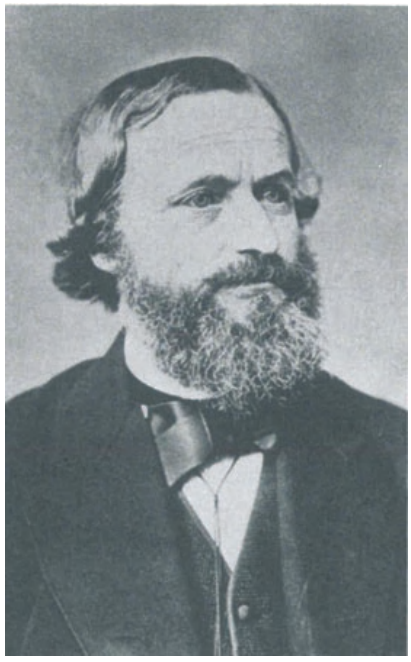
FIGURE 15-10

15.6 KIRCHHOFF'S RULES

Not all electrical circuits can be reduced to simple series or parallel combinations. Two fundamental rules were established by G. R. Kirchhoff (1824–1887) that aid in the solution of electrical networks.

1. The algebraic sum of currents *toward* any branch point is zero.
2. The algebraic sum of all potential changes in a closed loop is zero.

We will consider rule (1) first. As we have stated earlier, charge cannot accumulate (or deplete) in a dc circuit: If it did, there would be a larger (or smaller) electric field at that region which would exert a larger (or smaller)


Gustav R. Kirchhoff (1824–1887).

force and thereby redistribute the charge evenly. Therefore, at any branch point in a circuit, whatever charge flows in must flow out. This is seen in Figs. 15-11*a* and *b*. An analogy would be a series of hoses for water irrigation. At a branch of hoses, whatever water flows in must flow out.

The current equation for Fig. 15-11*a*, $i_1 = i_2 + i_3$, is an obvious application of the first rule. The rule can also be used for the circuit of Fig. 15-11*b*. Consider branch point A. The current equation would be $i_2 = i_3 + i_4$ *if* the currents have the directions indicated by the arrows. We do know that current i_2 has the direction indicated by the arrow. But we do not know if the arrow is the correct direction for i_3 since we do not know if the potential at point A is higher or lower than that at point B. We therefore *assume* a direction and maintain that assumption in formulating other equations. When we finally solve the circuit equations, if i_3 is positive our assumption is correct; if it is negative, then the current i_3 is in the direction opposite to our assumption. To illustrate the consistency of following the original assumption, rule (1) applied to branch point B would be $i_1 + i_3 = i_5$.

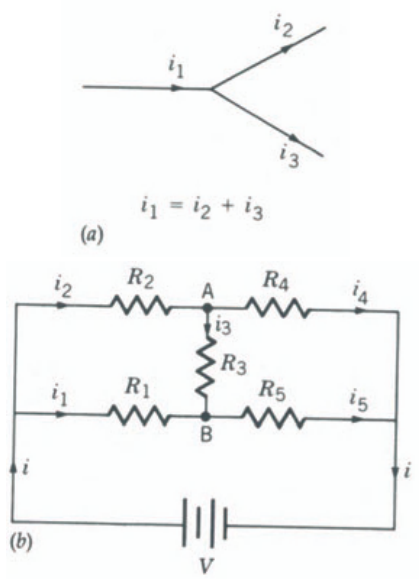


FIGURE 15-11
Kirchhoff's rule for current at a branch point.

Let us consider rule (2). This rule is a statement of the conservation of energy. In a circuit there may be potential differences associated with emf sources present as well as voltage drops associated with resistors. If we mentally start at a point in the circuit, go around any closed loop in either direction adding algebraically all the changes in potential and then return to the starting point, the potential of that point must be the same as when we started; that is, the sum of all the potential changes (increases and decreases) considered in our mental trip must add up to zero.

In applying rule (2), it is useful to follow certain guidelines that will prevent errors in the signs of the potential changes.

(a) As indicated in connection with rule (1), we first assume a direction for the current through each branch of the circuit.

(b) We then choose any closed loop in the circuit and designate the direction (clockwise or counterclockwise) in which we wish to mentally traverse it.

(c) We now go around the loop in the chosen direction adding algebraically all the potential changes and setting the sum equal to zero.

When we meet an emf source, its voltage V is taken as positive if we cross the source from the negative (low potential) side to the positive (high potential) side. The reason for taking V as positive is that in going from the negative side of the source to the positive one, the change in potential represents an *increase* in electric potential. If the source is crossed from positive to negative, its voltage is taken as negative because the electric potential has decreased. Let us now consider what to do when we meet a resistor. Earlier, we indicated that in a resistor the current goes from the high potential side of the resistor to the low potential one and that the potential drop between the two sides is iR. Thus, if in our mental trip around the circuit loop we cross a resistor in the same direction as the current, we must take the iR drop as negative because we are going from high to low potential—a decrease. The iR drop is

taken as positive if the resistor is traversed in the direction opposite to that of the current.

Now we can apply rule (2) to some simple circuits. Consider the circuit of Fig. 15-12. We choose the current in the clockwise direction (we may just as well choose the opposite direction, although the correctness of our assumption is obvious by inspection in this simple circuit). If we now traverse the loop in the clockwise direction starting at point A, we apply rule (2) and write

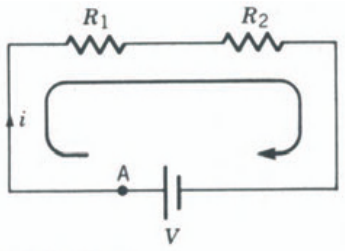


FIGURE 15-12

$$-iR_1 - iR_2 + V = 0$$

Note that both iR drops are written as negative because both resistors were crossed in the direction of the assumed current. V was taken as positive because the emf source was crossed from the negative to the positive side. We can rewrite the result as

$$V = iR_1 + iR_2$$

which is the result obtained when we discussed resistors in series where rule (2) was implicitly used.

Let us now consider the slightly more complicated circuit of Fig. 15-13*a*. We may again choose to go around the loop in a clockwise direction starting at point A. From rule (2) we obtain

$$-iR_1 - iR_2 + V_2 + V_1 = 0$$

or

$$V_2 + V_1 = iR_1 + iR_2$$

We see that because the batteries are pointing in the same direction (relative to the direction of the current) the effective voltage of the two emf sources is the sum of the individual voltages, that is, an emf source of voltage $V = V_1 + V_2$ would give rise to the same current. Suppose we reverse the direction of one of the sources, as in Fig. 15-13*b*. We will still *assume* that the current is in the clockwise direction. With such an assumption, and if we traverse the loop in the clockwise direction, we write

$$-iR_1 - iR_2 + V_2 - V_1 = 0$$

Note that V_1 is now taken as negative because in our mental trip the emf V_1 was crossed from the positive to the negative side. The result can be solved for i

$$i = \frac{V_2 - V_1}{R_1 + R_2}$$

It is clear that if $V_1 > V_2$, i would be negative, indicating that we have assumed the wrong direction for i, that is, i would be counterclockwise.

Let us now use Kirchhoff's rules to solve a circuit with certain similarities to a transistor circuit that we will encounter in a later chapter.

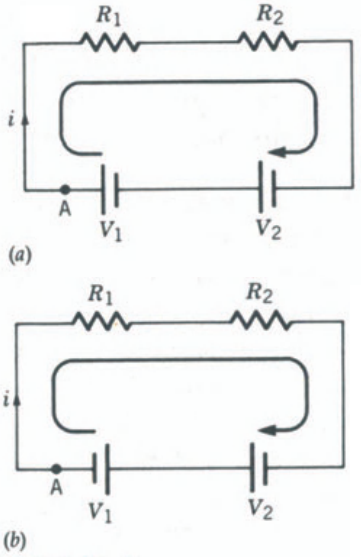


FIGURE 15-13

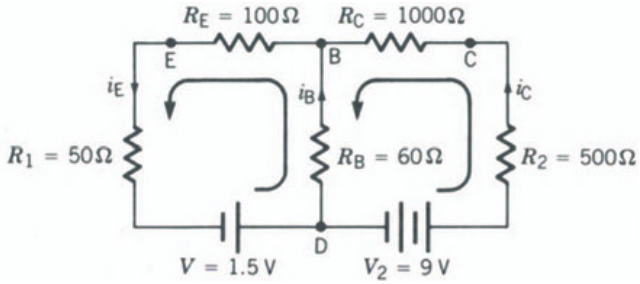


FIGURE 15-14
Example 15-5.

Example 15-5

In the circuit of Fig. 15-14, (a) Find the currents i_C, i_E, and i_B and the voltage drop across resistors R_1 and R_2. (b) Find the voltage difference between points C and D and between D and E.

Solution

(a) From the first rule at branch point B

$$i_C + i_B = i_E$$

We then write the second rule for the two loops. For the right-hand loop, if we traverse it in the counterclockwise direction starting at point D, we write

$$V_2 - i_C R_2 - i_C R_C + i_B R_B = 0$$

$$9\text{ V} - i_C\,500\,\Omega - i_C\,1000\,\Omega + i_B\,60\,\Omega = 0$$

$$9\text{ V} - i_C\,1500\,\Omega + i_B\,60\,\Omega = 0$$

For the left-hand loop, traversing it counterclockwise, we write

$$V_1 - i_B R_B - i_E R_E - i_E R_1 = 0$$

$$1.5\text{ V} - i_B\,60\,\Omega - i_E\,100\,\Omega - i_E\,50\,\Omega = 0$$

$$1.5\text{ V} - i_B\,60\,\Omega - i_E\,150\,\Omega = 0$$

We now have three equations to be solved simultaneously for i_C, i_E, and i_B. They are

$$i_C + i_B = i_E$$

$$9\text{ V} - i_C\,1500\,\Omega + i_B\,60\,\Omega = 0$$

$$1.5\text{ V} - i_B\,60\,\Omega - i_E\,150\,\Omega = 0$$

We can use the first equation to eliminate i_B from the last two. $i_B = i_E - i_C$, therefore

$$9\text{ V} - i_C\,1500\,\Omega + (i_E - i_C)\,60\,\Omega = 0$$

or

$$9\text{ V} - i_C\,1560\,\Omega + i_E\,60\,\Omega = 0 \qquad (15.17)$$

and

$$1.5\ \text{V} - (i_\text{E} - i_\text{C})\ 60\ \Omega - i_\text{E}\ 150\ \Omega = 0$$

or

$$1.5\ \text{V} - i_\text{E}\ 210\ \Omega + i_\text{C}\ 60\ \Omega = 0 \qquad (15.18)$$

Equations 15.17 and 15.18 can be solved for i_E and i_C. Multiplying Eq. 15.17 by 7 and Eq. 15.18 by 2 and adding them together eliminates i_E, that is,

$$7 \times (9\ \text{V} - i_\text{C}\ 1560\ \Omega + i_\text{E}\ 60\ \Omega) = 0$$

$$2 \times (1.5\ \text{V} - i_\text{E}\ 210\ \Omega + i_\text{C}\ 60\ \Omega) = 0$$

$$63\ \text{V} - i_\text{C}\ 10{,}920\ \Omega + 3\ \text{V} + i_\text{C}\ 120\ \Omega = 0$$

$$i_\text{C} = \frac{66\ \text{V}}{10{,}800\ \Omega} = 6.1 \times 10^{-3}\ \text{A} = 6.1\ \text{mA}$$

We can now solve for i_E using either Eq. 15.17 or 15.18

$$1.5\ \text{V} - i_\text{E}\ 210\ \Omega + i_\text{C}\ 60\ \Omega = 0$$

$$i_\text{E} = \frac{1.5\ \text{V} + i_\text{C}\ 60\ \Omega}{210\ \Omega}$$

$$= \frac{1.5\ \text{V} + (6.1 \times 10^{-3}\ \text{A})(60\ \Omega)}{210\ \Omega}$$

$$= 8.9 \times 10^{-3}\ \text{A} = 8.9\ \text{mA}$$

Finally, we can use the result of the first rule, $i_\text{C} + i_\text{B} = i_\text{E}$, to obtain i_B

$$i_\text{B} = i_\text{E} - i_\text{C} = 8.9\ \text{mA} - 6.1\ \text{mA} = 2.8\ \text{mA}$$

The voltage drop across R_1 is

$$V = i_\text{E}R_1 = 8.9 \times 10^{-3}\ \text{A} \times 50\ \Omega = 0.45\ \text{V}$$

and the voltage drop across R_2 is

$$V = i_\text{C}R_2 = 6.1 \times 10^{-3}\ A \times 500\ \Omega = 3.1\ \text{V}$$

(b) To find the voltage difference between points D and C, let us follow the current i_C through the circuit elements in the right-hand loop. Start with the potential V_D.

$$V_\text{D} + V_2 - i_\text{C}R_2 = V_\text{C}$$

$$V_\text{C} - V_\text{D} = V_2 - i_\text{C}R_2$$

$$= 9\ \text{V} - 6.1 \times 10^{-3}\ \text{A} \times 500\ \Omega$$

$$= 6\ \text{V}$$

We may do the same with the left-hand loop to find the potential difference between points D and E.

$$\begin{aligned} V_D - i_B R_B - i_E R_E &= V_E \\ V_E - V_D &= -i_B R_B - i_E R_E \\ &= -2.8 \times 10^{-3}\ \text{A} \times 60\ \Omega \\ &\quad - 8.9 \times 10^{-3}\ \text{A} \times 100\ \Omega \\ &= -1.1\ \text{V} \end{aligned}$$

The result shows that D is at a higher potential than E.

15.7 AMMETERS AND VOLTMETERS

We will see in the next chapter that electric current passing through a wire produces a magnetic field. If a loop of wire is used then, on the passage of current, one end of the loop becomes the north pole of a magnet and the other end becomes the south pole, as in Fig. 15-15*a*. This will be discussed in more detail in Chapter 16. Many loops can be used (see Fig. 15-15*b*), and each one contributes to the forming of a magnet: The larger the number of loops, the stronger the magnet for a given current. A series of loops forms a coil, and if a tightly wound coil is placed between the poles of a permanent horseshoe magnet as in Fig. 15-15*c* and current passes through it, the induced north pole of the coil will be repelled by the north pole of the permanent magnet. If the coil is suspended by a flexible metal strip (see Fig. 15-15*c*), the twisting (torsion) force of this metal strip acts as a spring and will oppose the rotation of the coil, causing it to return to its initial position at zero current. Thus, depending on the strength of this metal strip and the other design parameters of the mechanism, a full-scale deflection of the instrument needle can be established for a given amount of current through the coil. This instrument is called a *galvanometer*. The current for full-scale deflection is called the *current rating* of a meter. Because these instruments are electrically and mechanically delicate, a common current rating is 0.1 mA (10^{-4} A).

Used by itself, such a meter could only measure currents from 0 to 0.1 mA when placed in series in the circuit. To extend the range of the meter, a lower resistance, called a *shunt*, is placed in parallel with the meter. Figure 15-16 shows these situations in which the meter could measure full scale for different currents in the line. Meters used for the measurement of current though a circuit are known by the general name of *ammeters* (meters to measure amps of current). The resistance of the coil R_c in the galvanometer is also specified by the manufacturer, in addition to the current rating. A typical value might be 1000 Ω, or one kilohm, kΩ. From Ohm's law the voltage drop

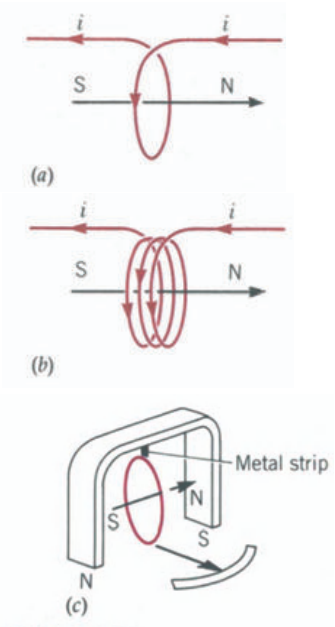


FIGURE 15-15
The basis of a galvanometer.

across the galvanometer in all cases of Fig. 15-16 must be

$$V = iR_c = 10^{-4}\ \text{A} \times 10^3\ \Omega = 0.1\ \text{V}$$

and this must also be the voltage drop across the shunt because it is in parallel with the meter. With the use of Ohm's law we may calculate the value of the shunt resistance. In Fig. 15-16*b*

$$R_s = \frac{V}{i_s} = \frac{0.1\ \text{V}}{9.9 \times 10^{-3}\ \text{A}} = 10.1\ \Omega$$

In Fig. 15-16*c*

$$R_s = \frac{V}{i_s} = \frac{0.1\ \text{V}}{99.9 \times 10^{-3}\ \text{A}} = 1.001\ \Omega$$

Many test meters have an external switch that changes the scale of the ammeter. This switch disconnects one shunt and introduces one of a different resistance into the circuit so that the same meter is used for different ranges of current.

An instrument to measure the voltage difference between two points in a circuit, say, two sides of a resistor, is called a *voltmeter* and can be made from a similar galvanometer. However, we do not want such a voltmeter to disturb the current flow through the resistor because such a change would alter the *iR* voltage drop. The ideal instrument would be one that had infinite resistance. However, the galvanometer requires that current pass through it to obtain a measurement. We will continue to use the galvanometer previously discussed, which requires 100 μA for full-scale deflection and has an internal resistance of 1 kΩ (1000 Ω). As shown before, this meter will read full scale when the voltage difference across it is $V = 10^{-4}\ \text{A} \times 10^3\ \Omega = 0.1\ \text{V}$. Suppose we wish to extend the range of the galvanometer to 10 V full scale. We would connect a resistor R_2 in series with the galvanometer so that the potential drop across the galvanometer is still 0.1 V, as in Fig. 15-17. To find the value of R_2, we note that the voltage drop across the galvanometer $V_m = 0.1$ V, plus the voltage drop across R_2, iR_2, must be 10 V, that is,

$$0.1\ \text{V} + 10^{-4}\ \text{A} \times R_2 = 10\ \text{V}$$

$$R_2 = \frac{9.9\ \text{V}}{10^{-4}\ \text{A}} = 9.9 \times 10^4\ \Omega$$

Similarly, if the meter is to read full scale across a voltage drop of 100 V, the drop across R_2 must be $100 - 0.1 = 99.9$ V and the value of R_2 must be

$$R_2 = \frac{99.9\ \text{V}}{10^{-4}\ \text{A}} = 9.99 \times 10^5\ \Omega$$

Again, an external switch on a meter connects different values of series resistances so that full-scale deflection of the meter may indicate different maximum voltages.

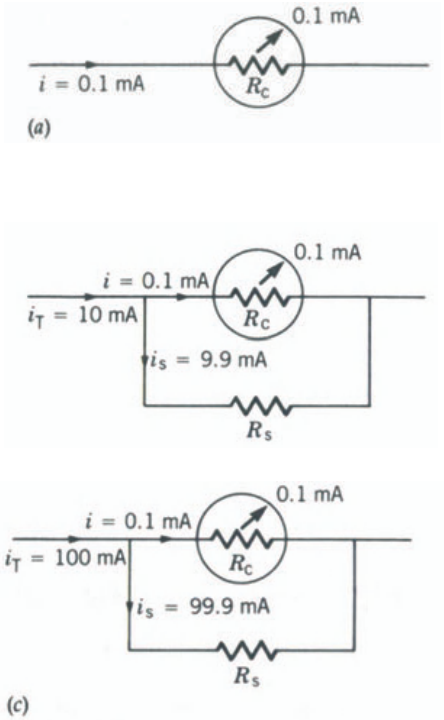


FIGURE 15-16
The construction of different ammeters from a galvanometer.

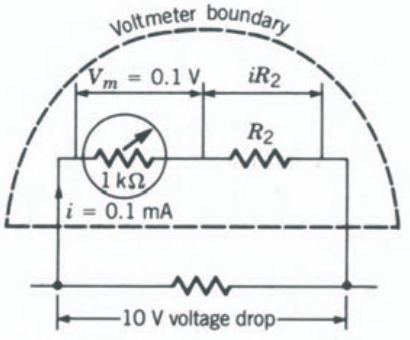


FIGURE 15-17
The construction of a voltmeter (enclosed by the dashed semicircle) from a galvanometer.

15.8 POWER DISSIPATION BY RESISTORS

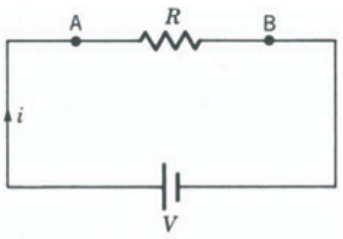


FIGURE 15-18

We saw at the end of Chapter 6 that in an elastic collision between an electron and an atom, very little energy is transferred to the atom—most of the kinetic energy is retained by the electron in its recoil (bouncing off the atom). However small, some energy is lost by the electron to the atom. This is the situation in a metal when electrons, accelerated by the electric field, collide with the atoms. Because many collisions are taking place, each small energy loss adds to a considerable amount. The kinetic energy transferred to the atoms per unit time represents an energy loss per unit time by the electrons, which is a power loss. We have seen in Chapter 9 that temperature is a measure of the average kinetic energy of the atoms (or molecules) of a system. Therefore, we expect any conductor to heat up when an electric current is passed through it. We see this phenomenon daily in electric heaters, ovens, and light bulbs.

The calculation of power dissipation P in an electrical resistor R as a result of the passage of a current i can immediately be found by considering Fig. 15-18. Let V_A and V_B represent the potentials of points A and B, respectively, and V_{AB} the potential difference. The change in potential energy of a charge Δq entering at A and leaving at B is

$$\Delta E_p = \Delta q\,(V_B - V_A)$$

This represents an energy loss because V_A is greater than V_B. In a given time Δt the amount of charge involved is $\Delta q = i\Delta t$, so

$$\Delta E_p = V_{AB}\, i\, \Delta t$$

The power dissipated in the resistor is $P_{AB} = \Delta E_p/\Delta t$, or

$$P_{AB} = V_{AB}\, i$$

We have written the voltage with the subscript AB to denote that if there is more than one resistor in a circuit, the power lost in each is the product of the voltage across each and current flowing through each. With this point in mind, we may write in more general form

$$P = Vi \tag{15.19}$$

$P = Vi$

Two other forms may be obtained by substitution of Ohm's law, $V = iR$. These are

$$P = i^2R \tag{15.20}$$

$P = i^2R$

and

$$P = \frac{V^2}{R} \tag{15.21}$$

$P = \frac{V^2}{R}$

We have shown in Chapter 5 that the unit of power is the watt or J/sec. The consistency in electrical definitions is readily seen from Eq. 15.19. The

unit of voltage is one joule per coulomb (one volt) and the unit of current is one coulomb per second (one ampere). Therefore the unit of power is

$$P = Vi = (1\ \text{JC}^{-1})\ (1\ \text{C sec}^{-1}) = 1\ \text{J/sec} = 1\ \text{W (watt)}$$

15.9 CHARGING A CAPACITOR—RC CIRCUITS

Thus far we have limited our discussion to cases where the current is constant with time, that is, direct current. In this section we consider a circuit where the current varies with time. This circuit plays an important role in the operation of computer clocks, which will be presented in Chapter 27.

In Chapter 14 (Section 14.7) we saw that when a capacitor is connected to the terminals of a battery, the plate of the capacitor connected to the positive side of the battery acquires a positive charge $q = CV$ (Eq. 14.21) and the other plate an equal but negative charge $-q$. One question that we may ask is: How long does it take for the charges to appear on the plates of the capacitor? Obviously, because resistors determine the current (that is, the rate of charge flow) in the circuit, this will depend on the resistance that is in the circuit.

Let us consider a circuit where a resistor R (the resistance of the connecting wires is assumed to be negligible) and a capacitor C are connected in series by means of switch S to a battery of emf V as in Fig. 15-19*a*. The initial condition is that when the switch is open there is no charge on the capacitor. When the switch is closed, a current is set up in the circuit and the capacitor will begin to charge (see Fig. 15-19*b*).

Let q be the charge on the capacitor at some time t after the switch is closed and let i be the current through the resistor at the same instant. V_{AD} is the voltage across the terminals of the battery, that is, $V_{AD} = V$, V_{AB} is the voltage drop iR across the resistor, and V_{BD} is the potential difference $\frac{q}{C}$ across the plates of the capacitor. We therefore write

$$V_{AD} = V_{AB} + V_{BD}$$

or

$$V = iR + \frac{q}{C} \tag{15.22}$$

By definition i is the rate at which charges flow through the resistor and, because these charges cannot cross the gap between the capacitor plates, this rate represents the rate at which the charge on the capacitor is increasing, that is, $i = \frac{dq}{dt}$. Equation 15.22 becomes

$$V = R\frac{dq}{dt} + \frac{q}{C} \tag{15.23}$$

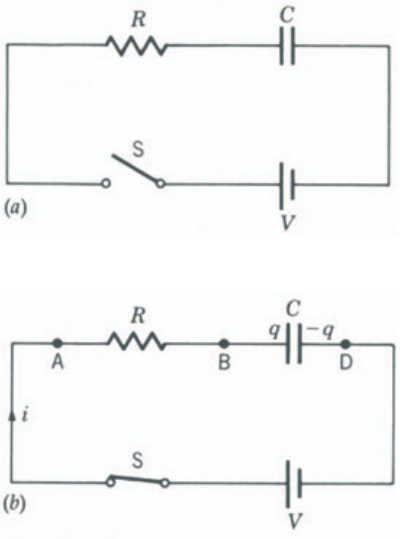


FIGURE 15-19

Dividing both sides of Eq. 15.23 by R and rearranging terms we have

$$\frac{V}{R} - \frac{q}{RC} = \frac{dq}{dt}$$

or

$$dt = \frac{dq}{\frac{V}{R} - \frac{q}{RC}} \tag{15.24}$$

We multiply both sides of Eq. 15.24 by $-1/RC$ and integrate with the limits $q = 0$ when $t = 0$ and charge q is on the plates at time t.

$$-\frac{1}{RC}\int_0^t dt = \int_0^q \frac{-\frac{1}{RC}\,dq}{\frac{V}{R} - \frac{q}{RC}}$$

$$-\frac{t}{RC} = \ln\left(\frac{V}{R} - \frac{q}{RC}\right)\Bigg|_0^q$$

Substituting the limits of integration we get

$$-\frac{t}{RC} = \ln\left(\frac{V}{R} - \frac{q}{RC}\right) - \ln\left(\frac{V}{R}\right) \tag{15.25}$$

Because $\ln x - \ln y = \ln \frac{x}{y}$, Eq. 15.25 becomes

$$-\frac{t}{RC} = \ln\left(1 - \frac{q}{CV}\right)$$

Taking the antilog of both sides we obtain

$$e^{-t/RC} = 1 - \frac{q}{CV}$$

Solving for q, we get

$$q = CV\,(1 - e^{-t/RC}) \tag{15.26}$$

Let us analyze Eq. 15.26. At $t = 0$, $q = CV\,(1 - e^{-0}) = CV\,(1 - 1) = 0$. This agrees with the fact that at $t = 0$ (when the switch was closed) the capacitor was uncharged. As t increases, the exponential term in the parenthesis decreases and consequently q increases (the capacitor is being charged). As $t \to \infty$, $e^{-t/RC} \to 0$ and $q \to CV$, the ultimate charge on the capacitor. Although it takes an infinite amount of time to fully charge the capacitor, it takes a finite amount of time to get very close to the final value $q = CV$. Moreover, this time is determined by the product RC, which is called the time constant of the circuit. For example, when $t = RC$, $q = CV\,(1 - e^{-1}) =$

0.63 CV; when $t = 4\ RC$, $q = CV\ (1 - e^{-4}) = 0.98\ CV$. We see that after a few time constants, q is very close to its ultimate value. A plot of q versus t is shown in Fig. 15-20.

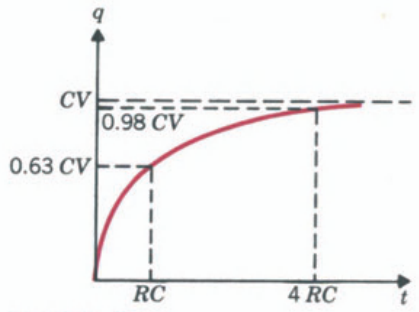


FIGURE 15-20
Charge accumulation on a capacitor in an RC circuit as a function of time.

The consistency of the electrical definitions of R and C can be readily verified in Eq. 15.26. From Ohm's law,

$$R = \frac{V}{i} = \frac{\text{volts}}{\text{amps}} = \frac{\text{volts}}{\text{C/sec}}$$

similarly from Eq. 14.21

$$C = \frac{q}{V} = \frac{\text{C}}{\text{volts}}$$

therefore

$$RC = \frac{\text{volts}}{\text{C/sec}}\frac{\text{C}}{\text{volts}} = \text{seconds}\ .$$

The mathematical solution of the circuit equation shows that the larger the value of the resistor R and of the capacitor C, the longer it will take to charge it. It is not difficult to understand the physical reason for this result. The larger R, the smaller the current through the circuit at any one time, hence the smaller the rate at which the capacitor is being charged. Similarly, the larger C, the more charge it can store for a given voltage and, obviously, the longer it will take to charge it.

PROBLEMS

15.1 A wire carries a current $i = 1$ A. How many electrons pass a fixed cross section of the wire in 1 sec?

15.2 Copper has one conduction electron per atom, that is, each atom contributes one electron that is free to move through the solid. The density of copper is 9 g/cm^3 and its molecular weight is 64 g/mole. A wire carries a current of 10 A. The cross-sectional area of the wire is 3 mm^2. (a) What is the current density? (b) What is the number of conduction electrons per m^3? (c) What is the drift velocity?
(*Answer*: (a) 3.33×10^6 A/m^2, (b) 8.47×10^{28} m^{-3}, (c) 2.46×10^{-4} m/sec.)

15.3 A copper wire 15 m long has 8×10^{26} mobile electrons. What is the drift velocity of the electrons if the current in the wire is 5 A?

15.4 The resistivity of copper at ambient temperature is $\rho = 1.7 \times 10^{-8}\ \Omega$-m. What is the resistance of a copper wire 5 m long and 2×10^{-3} m in diameter?

15.5 A copper wire ($\rho = 1.7 \times 10^{-8}\ \Omega$-m) 10 m long and 1×10^{-3} m in diameter carries a current of 2 A. What is the potential difference across the ends of the wire?

15.6 In the earth's atmosphere positive charges move toward the earth and negative charges move away from it. The total current is approximately 1800 A. The average value of the electric field responsible for this current near the surface of the earth is 100 N/C. What is the resistivity of the air at the surface of the earth? The radius of the earth is 6.37×10^6 m.
(*Answer*: $2.83 \times 10^{13}\ \Omega$-m.)

15.7 In the circuit of Example 15-4, let $R_1 = 5\ \Omega$, $R_2 = 10\ \Omega$, $R_3 = 4\ \Omega$, and $V = 2$ V. Find the currents i_1, i_2, and i_3.

15.8 In the circuit of Fig. 15-21 (a) Find the currents through each resistor. $R_1 = 3\ \Omega$, $R_2 = 6\ \Omega$, $R_3 = 6\ \Omega$, $R_4 = 12\ \Omega$, $V = 18$ V. (b) What is the total current i?

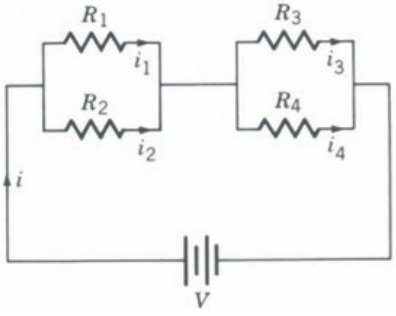


FIGURE 15-21
Problem 15.8.

15.9 The current through R_3 in the circuit of Fig. 15-22 is 0.2 A. (a) What is the current in R_1, R_2, and R_4? (b) What is the voltage of the battery?

(*Answer*: (a) 1.1 A, 0.5 A, 0.4 A, (b) 4.2 V.)

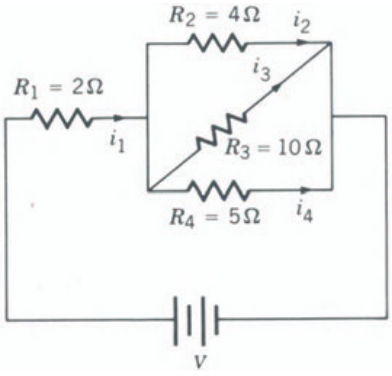


FIGURE 15-22
Problem 15.9.

15.10 How many possible resistance values can be obtained with three resistors $R_1 = 50\ \Omega$, $R_2 = 100\ \Omega$, and $R_3 = 150\ \Omega$?

15.11 In the circuit of Fig. 15-23, find i_B, the voltage drops across R_1 and R_2, and the voltage differences $V_C - V_D$ and $V_D - V_E$.

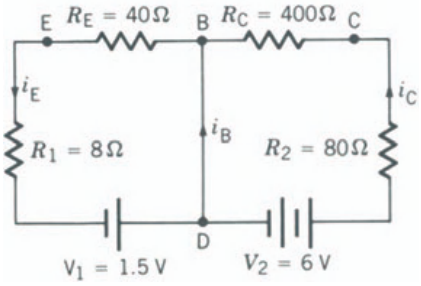


FIGURE 15-23
Problem 15.11.

15.12 The current i through R_1 in the circuit diagram of Fig. 15-24 is 40 mA. (a) What is the current through R_2, R_3, and R_4? (b) What is the potential difference between A and B?

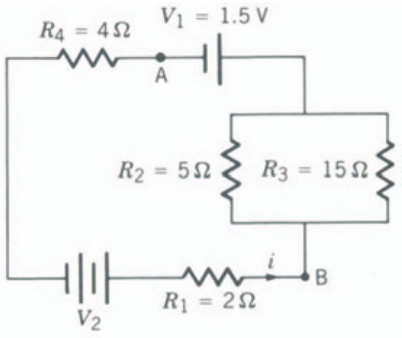


FIGURE 15-24
Problem 15.12.

15.13 The voltage drop across R_3 in the circuit diagram of Fig. 15-25 is 4 V. (a) Find the currents through the resistor R_1, R_2, and R_3. (b) What is the resistance of R_2?

(*Answer*: (a) 0.5 A, 0.3 A, 0.8 A, (b) 1.67 Ω.)

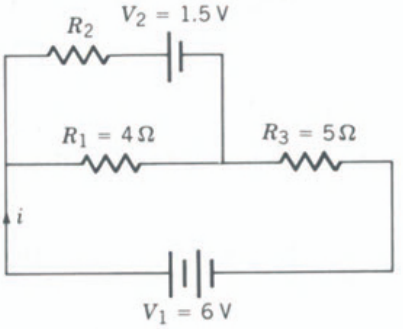


FIGURE 15-25
Problem 15.13.

15.14 Find the currents through the resistors R_1, R_2, and R_3 of the circuit of Fig. 15-26.

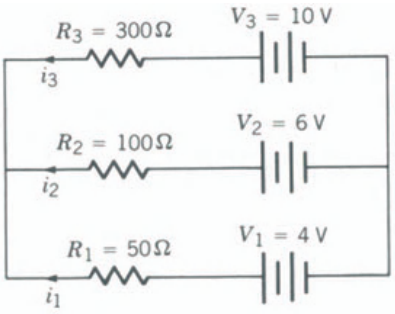


FIGURE 15-26
Problem 15.14.

15.15 The circuit of Fig. 15-27 is known as the Wheatstone Bridge. It is used to find the resistance of an unknown resistor R_x in terms of three known resistors R_1, R_2, and R_s. The value of R_s is adjusted until no current flows through the galvanometer G. (The arrow over the resistor symbol of R_s indicates

that R_s is a variable resistor.) Let $R_1 = 10\ \Omega$ and $R_2 = 100\ \Omega$. If no current flows through G when $R_s = 470\ \Omega$, what is the value of R_x?

(*Answer*: 4700 Ω.)

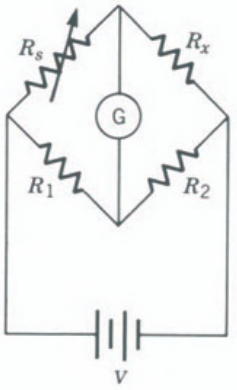


FIGURE 15-27
Problem 15.15.

15.16 A galvanometer has an internal resistance of 2000 Ω and a current of 50 μA will cause full-scale deflection. What shunt resistance is required to use it as an ammeter whose full scale reads 0.1 A?

(*Answer*: 1 Ω.)

15.17 The galvanometer of problem 15.16 is to be used as a voltmeter with a maximum scale reading of 10 V. What series resistance is required?

(*Answer*: $1.98 \times 10^5\ \Omega$.)

15.18 An electric light bulb marked 100 W is used in a home in which the wall outlet is at 120 V. What is the resistance of the filament in the bulb?

15.19 An immersion heater draws 3 A when it is plugged in a 120-V wall outlet. What is the power consumption of the heater?

15.20 If the immersion heater of problem 15.19 is used to boil a cup of water ($m = 150$ g) initially at 27°C and 80% of the power is absorbed by the water, how long will it take for the water to boil?

15.21 An electric heater of resistance 5 Ω is plugged in a 120-V outlet by means of an extension line. Compare the power loss in an extension line 5 m long when the line is made of No. 12-gauge copper wire (2.5 mm in diameter) and when it is made of No. 14-gauge wire (1.6 mm in diameter). The resistivity of copper is $1.7 \times 10^{-8}\ \Omega$-m.

(*Answer*: 9.90 W, 23.96 W.)

15.22 A megawatt (10^6 watts) of electrical power is needed to run a factory. Compare the energy losses in the transmission lines when the voltage is 120 V with when it is 6000 V.

15.23 A resistor $R = 1000\ \Omega$ and a capacitor $C = 100\ \mu$F are connected in series with a 10-V battery and a switch. (a) How long after closing the switch will the voltage across the capacitor be 1.0 V. (b) When will the capacitor be charged to 99% of its final charge?

(*Answer*: (a) 1.05×10^{-2} sec, (b) 0.46 sec.)

15.24 A capacitor C in series with a resistor R is charged by turning the switch to position a in Fig. 15-28. After the capacitor has been charged, the switch is returned to position b. Find an expression for the charge on the capacitor as a function of time. Take $t = 0$ at the moment the switch is changed from a to b.

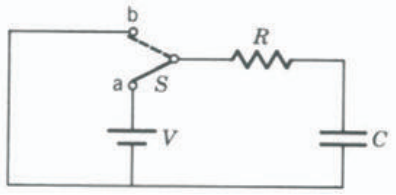
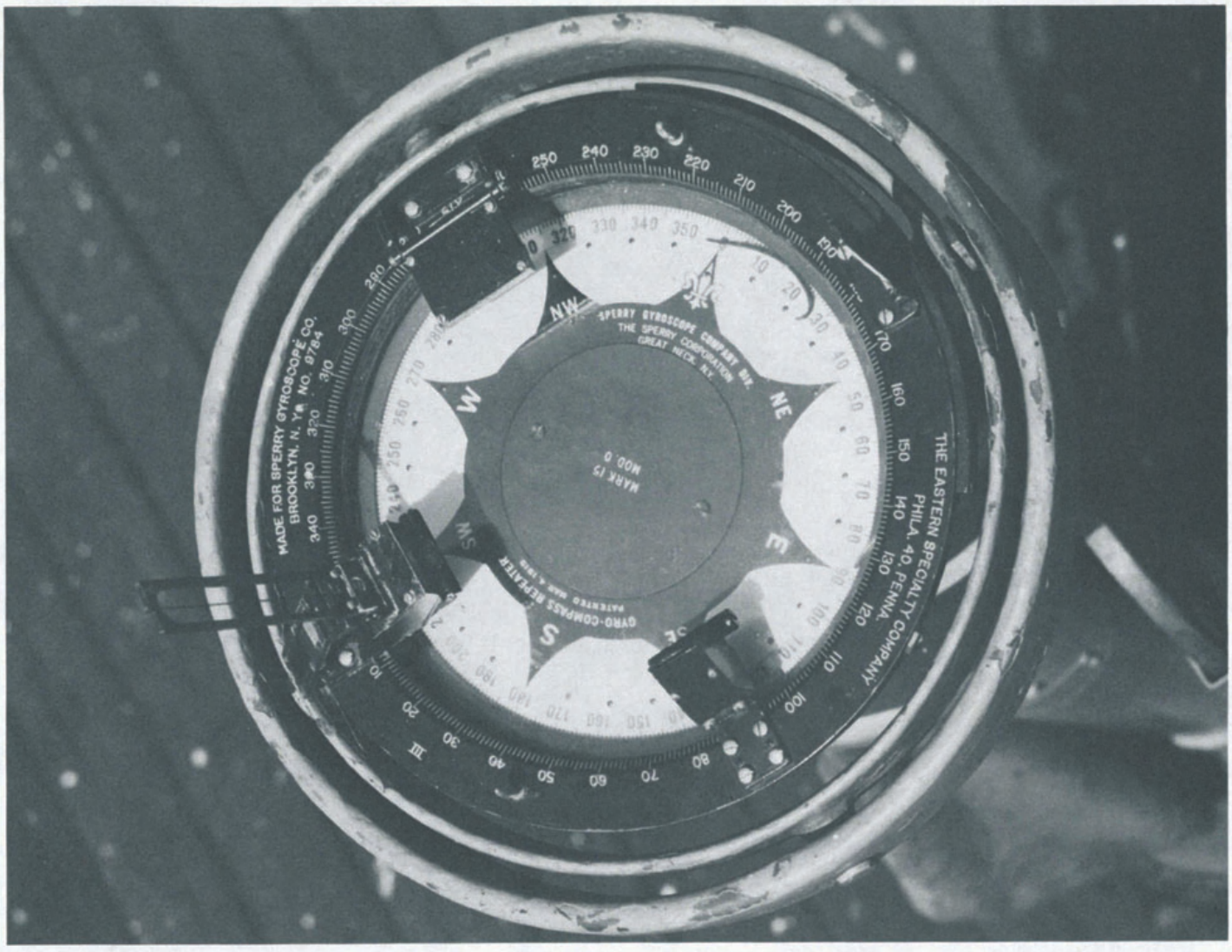


FIGURE 15-28
Problem 15.24.



CHAPTER 16

Magnetic Fields and Electromagnetic Waves

16.1 INTRODUCTION

Almost everyone has performed elementary experiments with bar magnets. If the bar magnet is suspended by a thread or supported by a pivot, one of the ends will point in a northerly direction. This end of the magnet is called the *north pole* of the magnet, with symbol N. The opposite end of the magnet is called the *south pole,* with symbol S. Elementary experiments also show that like poles repel and unlike poles attract. This suggests that there is something that we call a *magnetic field* by which poles can exert forces on each other. This field is similar to the two other fields we have already considered, the gravitational field and the electric field. There is one important difference, however: If we break a bar magnet in half, we cannot make single poles, but instead we will have two bar magnets. The broken end becomes the south pole of the half that has the north pole, and the other broken end becomes the north pole of the half that has the south pole.

There is an intimate relation between the motion of electric charges and magnetic fields, and our technological society is largely based on this relationship, from the generation of electric power to many types of electronic devices. We will not deal with all these; instead we will consider only those effects which we need for the understanding of the concepts of modern physics presented in later chapters, namely, the magnetic field of a wire coil, the magnetic moment of a current loop, the force of a magnetic field on a moving charge, and the nature of electromagnetic waves.

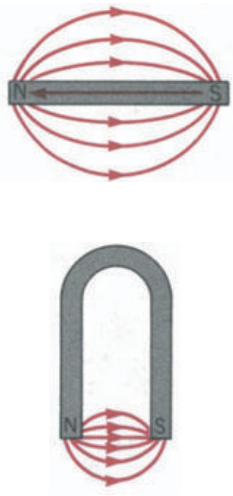


FIGURE 16-1
Magnetic field lines in between the poles of a bar magnet and a horseshoe magnet.

16.2 MAGNETIC FIELDS

We may map magnetic fields by using a small compass that we will represent by a small arrow with its head as the north pole of the compass magnet. We arbitrarily define the direction of the magnetic field at a given point to be the direction in which the compass points. Figure 16-1 shows the fields of a bar magnet and a horseshoe magnet. The fields are indicated by continuous lines from the north to the south pole, and the number of lines is arbitrary, although in the comparison of two magnets the stronger one is customarily represented by a greater number of field lines. The direction of the magnetic field at a given point is the tangent to the field line at that point. In future drawings we may represent the field direction by a single arrow on a line.

In 1820 the Danish physicist Hans Oersted (1777–1851) found that there is a magnetic field associated with current flowing in a wire. The direction of the magnetic field for a long, straight wire is schematically represented in Fig. 16-2. The field lines are circular about the wire. There are many concentric field lines, but the field becomes weaker as we move away from the wire. The direction of the magnetic field is determined by a right-hand rule. If the thumb of the right hand is pointed in the direction of the current and the fingers are curled, the circular direction of the fingers is the direction of the magnetic field.

Hans Christian Oersted (1777-1851).

Suppose we have a circular loop of wire that carries current, as in Fig. 16-3. If we apply the right-hand rule, we see that one side of the loop has a north pole perpendicular to the plane of the loop and its opposite side will become a south pole. The arrows represent the magnetic field direction *inside* the loop. These field lines return *outside* the loop so that the loop itself becomes a magnet. This situation will remain regardless of the shape of the current loop; that is, a rectangular loop will give the same result.

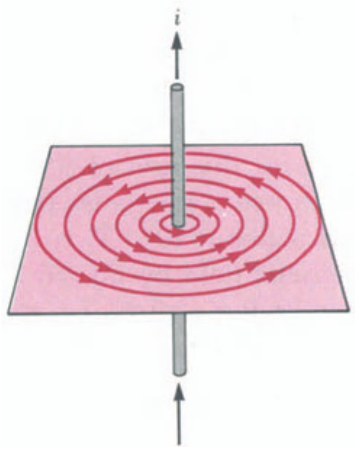


FIGURE 16-2
Magnetic field lines around a long, straight wire carrying a current i.

16.3 FORCE ON CURRENT-CARRYING WIRES

In the preceding sections, we have presented some qualitative facts about magnetic fields. We have seen that magnets exert forces on other magnets. We have also seen that a wire carrying a current produces a magnetic field, that is, becomes a magnet.

Experiment shows that when a wire carrying a current is placed in a magnetic field, it will experience a force. We can use this to define the magnitude of a magnetic field, **B**. (Remember that the direction of **B** has been defined as the direction taken by the north pole of a compass).

Figure 16-4 is a schematic drawing of an experiment which shows that when a wire carrying a current is placed in a magnetic field **B** the force **F** is in a direction that is *perpendicular* to the plane defined by the field and the direction of the current. We also find experimentally that there is no force on a wire if the wire is in the direction of the magnetic field and that the force **F** is proportional to the sine of the angle θ between the field and the wire. The experiment also shows that the force on the wire is proportional to the current in the wire i and to the length of wire Δl in the field. From these experimental results we can define the magnitude of the magnetic field **B** as follows:

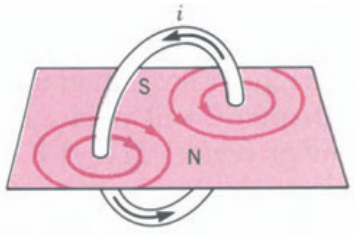


FIGURE 16-3
Magnetic field lines on a plane through a current-carrying circular wire loop.

$$B = \frac{F}{i\,\Delta l \sin\theta} \tag{16.1}$$

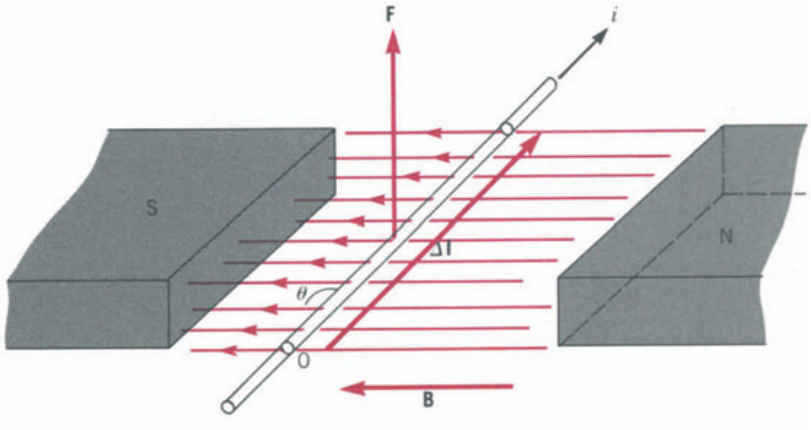


FIGURE 16-4
Force on a current-carrying wire of length Δl in a magnetic field.

It should be clear that Eq. 16.1 defines B unambiguously because, if we double or cut in half the current, the force on the wire will change accordingly. Similar arguments apply to Δl and $\sin\theta$.

From Eq. 16.1, the SI unit for B is newtons/ampere-meter (N/A-m). The name for this unit of B is the tesla (T). An older unit for the magnetic field still in use is the gauss (G), where 1 tesla $= 10^4$ gauss. The earth's magnetic field is about 10^{-4} T, so that 1 tesla is a large quantity. Having defined the magnetic field by Eq. 16.1, we can now state that when a segment of wire Δl, carrying a current i, is placed in a magnetic field of magnitude B, it will experience a force

$$F = i\,\Delta l\,B\sin\theta \tag{16.2}$$

The direction of this force is the perpendicular both to the magnetic field and to the direction of the current. We recognize that this relationship can be represented by a vector cross product. That is, Eq. 16.2 may be written as

$$\mathbf{F} = i\,\Delta\mathbf{l} \times \mathbf{B} \tag{16.3}$$

$\mathbf{F} = i\,\Delta\mathbf{l} \times \mathbf{B}$

It is conventional to let the current i be a scalar quantity and to let $\Delta\mathbf{l}$ be a vector pointing in the direction of the current. If we let the element of wire in Fig. 16-4 measure 0 at the reader's end and let $\Delta\mathbf{l}$ be the vector length in the magnetic field, then the right-hand rule for vector cross product discussed in Chapter 2 will yield a force in the upward direction perpendicular to the plane of vectors $\Delta\mathbf{l}$ and $\mathbf{B}$.

16.4 TORQUE ON A CURRENT LOOP

We are now able to understand the operation of the galvanometer that was used in Chapter 15 to construct ammeters and voltmeters.

Consider a single rectangular loop of wire connected to a pivot rod, as shown in Fig. 16-5*a*. Let the length of sides 1 and 3 be a and that of sides 2 and 4, b. We can use Eq. 16.3 to find the force on each side of the loop. For sides 1 and 3 the magnitudes of the forces are the same because the angle between $\Delta\mathbf{l}$ and $\mathbf{B}$ is 90° and both wires have the same length a, that is,

$$F_1 = F_3 = i\,\Delta l\,B\sin 90° = iaB \tag{16.4}$$

From the definition of the cross product, we see that in Fig. 16-5*a* $\mathbf{F}_1$ is out of the page toward the reader whereas $\mathbf{F}_3$ is into the page. These two forces are drawn in Fig. 16-5*b*. Similarly the magnitudes of the forces on sides 2 and 4 are equal.

$$F_2 = F_4 = i\,\Delta l\,B\sin(90° - \theta) = ibB\sin(90° - \theta)$$

The direction of $\mathbf{F}_2$ in Fig. 16-5*a* is upward, while that of $\mathbf{F}_4$ is downward.

We conclude that there is no net force in any direction. However, if we look at the top view (Fig. 16-5*b*) along the pivot rod we see that there exists

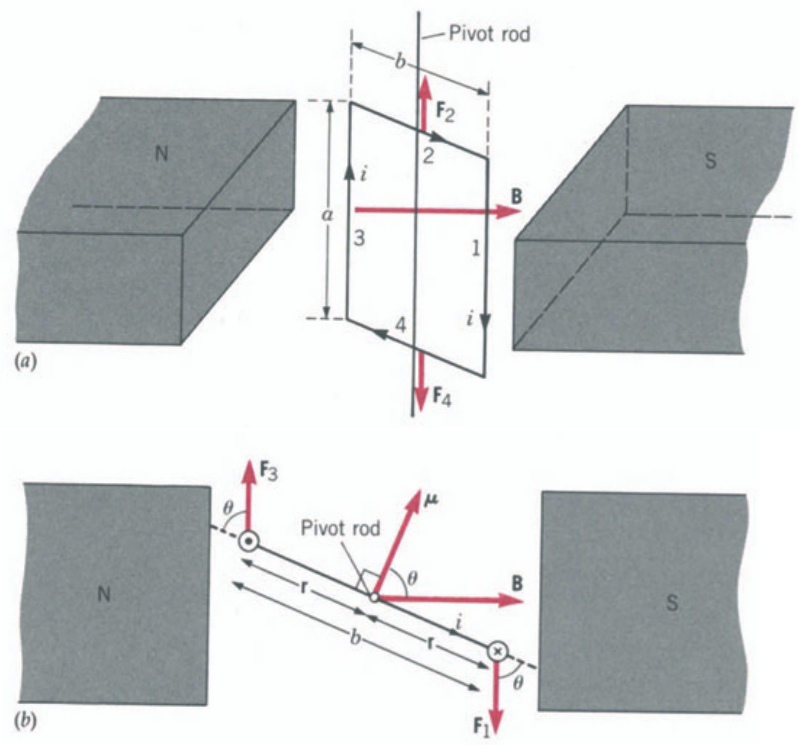


FIGURE 16-5
Torque on a current-carrying rectangular loop of wire on a pivot rod when placed in a magnetic field. (*a*) Side view. (*b*) Top view.

a torque that tends to rotate the loop about the pivot rod. The torque is given by Eq. 8.5.

$$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F} \tag{8.5}$$

Applied to the present situation, we see that $\mathbf{F}_1$ and $\mathbf{F}_3$ exert a torque on the loop. The net torque is the sum of the individual torques caused by $\mathbf{F}_1$ and $\mathbf{F}_3$, but because they are equal, we simply multiply the torque exerted by one of these forces by two, or

$$\boldsymbol{\tau} = 2\mathbf{r} \times \mathbf{F} \tag{16.5}$$

$$\tau = 2rF \sin\theta \tag{16.6}$$

where $\mathbf{F}$ stands for either $\mathbf{F}_1$ or $\mathbf{F}_3$. Substituting Eq. 16.4 for F in Eq. 16.6, we obtain

$$\tau = 2riaB \sin\theta \tag{16.7}$$

From Figs. 16-5*a* and 16-5*b* we see that $r = b/2$; therefore

$$\tau = iabB \sin\theta \tag{16.8}$$

We recognize the product ab as the area of the loop. Calling this area A, Eq. 16.8 may be written as

$$\tau = iAB \sin\theta \tag{16.9}$$

Although Eq. 16.9 has been derived for a rectangular loop of wire, it can be shown that the result is the same for any other geometric configuration.

Equation 16.9 tells us that the torque is a maximum when $\theta = 90°$, that is, when the plane of the loop lies in the direction of B, and it is zero when $\theta = 0°$, that is, when B is perpendicular to the plane of the loop. The magnitude of the torque can be increased by increasing the current in the loop. Because in the galvanometer this torque on the coil is opposed by the twisting (torsion) torque of the metal strip used to suspend the coil, the magnitude of the angle of rotation is a function of the current passing through the coil. An alternative way of increasing the torque on the coil is by using a coil made of several loops. To increase the sensitivity of the galvanometer at low currents, a coil of many loops of wire is used.

16.5 MAGNETIC DIPOLE MOMENT

In the preceding section, we saw that the important element in determining the torque on a wire loop, for a given field **B** in which it is placed, is the product of the area of the loop and the current through the loop (see Eq. 16.9). This quantity is called the *magnetic dipole moment* or simply the *magnetic moment* of the coil with symbol μ, where

$$\mu = iA \tag{16.10}$$

$\mu = iA$

The expression for the torque can now be written as

$$\tau = \mu B \sin \theta \tag{16.11}$$

Equations 16.11 and 16.9 give the magnitude of the torque, but they do not specify the direction of τ. The direction of the torque can be obtained by using Eq. 16.5

$$\boldsymbol{\tau} = 2\mathbf{r} \times \mathbf{F} \tag{16.5}$$

From the definition of the cross product, the direction of $\boldsymbol{\tau}$ in Fig. 16-5*b* is the perpendicular to the paper directed inward. We can specify both the magnitude and the direction of $\boldsymbol{\tau}$ in terms of the magnetic moment by assigning a vector direction to μ. We define $\boldsymbol{\mu}$ as a vector whose magnitude is given by Eq. 16.10 and whose direction is the perpendicular to the plane of the loop according to the right-hand rule. That is, we curl the fingers of the right hand in the direction of the current and the extended thumb indicates the direction of $\boldsymbol{\mu}$ (see Fig. 16-5*b*). We can now express the torque on the loop as

$$\boldsymbol{\tau} = \boldsymbol{\mu} \times \mathbf{B} \tag{16.12}$$

$\tau = \mu \times B$

From the definition of the cross product, it is clear that Eq. 16.12 yields the correct value for the magnitude (Eq. 16.11) and for the direction (Eq. 16.5) of $\boldsymbol{\tau}$.

Because a magnetic dipole experiences a torque when placed in an external magnetic field, work must be done by an external agent to change its orientation. As in the cases considered earlier (gravitational and electrical),

this work, by definition, becomes the potential energy E_p of the dipole. Recalling that only *changes* in potential energy are experimentally observed, we must define a zero or reference orientation. It is customary to set $E_p = 0$ when $\theta = 90°$, that is, when the dipole vector is perpendicular to the magnetic field. To calculate E_p for any other orientation of $\boldsymbol{\mu}$, we calculate the work using Eq. 8.13

$$W_\theta = \int_{\theta_0}^{\theta_f} \tau \, d\theta \tag{8.13}$$

Setting this work equal to E_p, and substituting Eq. 16.11 for τ in Eq. 8.13,

$$E_p = \int_{90°}^{\theta} \mu B \sin\theta \, d\theta$$

$$E_p = -\mu B \cos\theta \Big|_{90°}^{\theta}$$

$$E_p = -\mu B \cos\theta \tag{16.13}$$

This expression for E_p can be written as a dot product (see Eq. 2.1), that is,

$$E_p = -\boldsymbol{\mu} \cdot \mathbf{B} \tag{16.14}$$

$E_p = -\mu \cdot B$

We should notice that, because the cos θ varies between 1 and -1, the maximum energy is μB. This occurs when $\cos\theta = -1$ or $\theta = 180°$, that is, when $\boldsymbol{\mu}$ and $\mathbf{B}$ are antialigned. When $\boldsymbol{\mu}$ and $\mathbf{B}$ are aligned, $\theta = 0°$, $\cos\theta = 1$, and the potential energy is at its minimum value of $E_p = -\mu B$.

Example 16-1

Assume that the electron in a hydrogen atom is essentially in a circular orbit of radius 0.5×10^{-10} m, and rotates about the nucleus at the rate of 10^{14} times per second. What is the magnetic moment of the hydrogen atom due to the orbital motion of the electron?

Solution

$$\mu = \text{area} \times \text{current}$$
$$= \pi r^2 i$$

where i is the current due to a single electron. Because current is defined as the amount of charge passing per unit time, we may view the electron's orbit as a racetrack and ask how many times the electron passes a given point per second. The current is simply

$$i = e\nu$$

where ν is the frequency of rotation—that is, the number of times the electron

passes a given point in its orbit per second—and e in the magnitude of the charge of the electron.

$$\begin{aligned}\mu &= \pi r^2 e\nu \\ &= \pi\,(0.5 \times 10^{-10}\ \text{m})^2\,(1.6 \times 10^{-19}\ \text{C})(10^{14}\ \text{Hz}) \\ &= 1.26 \times 10^{-25}\ \text{A-m}^2\end{aligned}$$

Therefore, the hydrogen atom is essentially a small bar magnet and will behave as such in a magnetic field.

16.6 FORCE ON A MOVING CHARGE

We may use the development of Section 16.4 concerning the force on a current-carrying wire in a magnetic field to find the force experienced by a single charge. We saw in Eqs. 15.4 and 15.5 that we may write the current of either a *positive* or *negative* charge carrier as

$$i = q\,N\,A\,v$$

where q was the magnitude of a charge, N the number of charge carriers per unit volume, A the cross-sectional area, and v the average drift velocity of the charge carriers. If we substitute this into Eq. 16.3 we obtain

$$\begin{aligned}\mathbf{F} &= i\,\Delta\mathbf{l} \times \mathbf{B} \\ &= q\,N\,A\,v\,\Delta\mathbf{l} \times \mathbf{B}\end{aligned} \qquad (16.3)$$

We see that $A\,\Delta l$ is the volume of the wire segment that is in the magnetic field B. The product of the charge density N (number of charges per unit volume) and the volume of the wire segment gives the total number of charges. Therefore, if $NA\,\Delta l$ is the number of charges experiencing a total force $\mathbf{F}$, the force per charge carrier $\mathbf{F}_q$ is $\mathbf{F}/NA\,\Delta l$ and from Eq. 16.3

$$\mathbf{F}_q = \frac{\mathbf{F}}{NA\,\Delta l} = q\,\mathbf{v} \times \mathbf{B} \qquad (16.15)$$

$\mathbf{F}_q = q\,\mathbf{v} \times \mathbf{B}$

Note that this has been derived for a positive charge, because we implicitly assume that $\mathbf{v}$ is in the same direction as $\Delta\mathbf{l}$ and therefore the current. If the charge carrier is the electron, the direction of the force is reversed. The force is perpendicular to the plane of $\mathbf{v}$ and $\mathbf{B}$ by definition of the vector cross product. It is therefore perpendicular to the direction of motion given by the vector $\mathbf{v}$. The definition of work is the dot product of force and displacement $dW = \mathbf{F} \cdot d\mathbf{s} = F\,ds\cos\theta$. Because the infinitesimal displacement $d\mathbf{s}$ has the same direction as the instantaneous velocity $\mathbf{v}$, and because $\mathbf{F}$ is perpendicular to $\mathbf{v}$, $\theta = 90°$, $\cos\theta = 0$, and the magnetic field does no work on the charge. The magnetic field therefore does not change the magnitude of the velocity of the charged particle.

16.7 THE HALL EFFECT

Suppose we have a conducting metal strip of width d and thickness t connected in a circuit and placed in a uniform magnetic field **B** as in Fig. 16-6. Let the direction of the magnetic field be into the paper, indicated by the symbol ⊗, which suggests the tail of an arrow. The electric field $\mathscr{E}_x$ responsible for the current i will be directed to the right. If we assume for the moment that the current is caused by positive charges, their drift velocity v will be in the direction of $\mathscr{E}_x$ as shown in Fig. 16-6.

Let us consider two points D and C on the metal strip such that the line joining the two points is perpendicular to $\mathscr{E}_x$, (see Fig. 16-6). Without the magnetic field, the potential difference between these two points is zero because no work is done in moving a charge from one point to the other. When the magnetic field is turned on, the drifting charges will experience a force given by Eq. 16.15. We label this force $\mathbf{F}_B$ to indicate that this is the force caused by the magnetic field **B.**

$$\mathbf{F}_B = q\,\mathbf{v} \times \mathbf{B} = q\,v\,B \tag{16.15}$$

This force, illustrated in Fig. 16-7, causes the positive charges to move to the upper part of the conducting strip while they are moving to the right. Because the sample as a whole must remain neutral, the lower part of the strip will become negatively charged. This situation is also shown in Fig. 16-7. The accumulation of positive charges along the upper part and of negative charges along the lower part creates an electric field $\mathscr{E}_y$ that opposes the further upward drift of positive charges. There will be a potential difference V_H between D and C associated with this electric field. From Eq. 14.19

$$V_H = V_D - V_C = \mathscr{E}_y\,d \tag{16.16}$$

where it is assumed that in equilibrium $\mathscr{E}_y$ is constant and d is the width of the strip (the distance between D and C). This voltage difference is called the *Hall voltage* after the physicist who first measured it, and the phenomenon is called the *Hall effect*. It is clear that the equilibrium Hall voltage V_H will be established when the downward force of $\mathscr{E}_y$ equals the upward force resulting from the magnetic field. Because the force of the electric field is given by definition in Eq. 14.1 as $F_{\mathscr{E}} = q\mathscr{E}_y$, we can say that at equilibrium (which is quickly established)

$$F_{\mathscr{E}} = F_B$$

$$q\mathscr{E}_y = q\,v\,B$$

therefore

$$\mathscr{E}_y = v\,B$$

Substituting for $\mathscr{E}_y$ in Eq. 16.16, we obtain

$$V_H = v\,B\,d \tag{16.17}$$

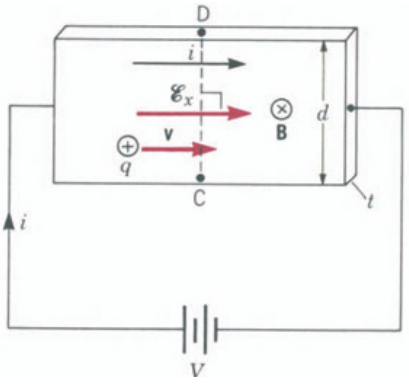


FIGURE 16-6
Experimental arrangement for the measurement of the Hall voltage.

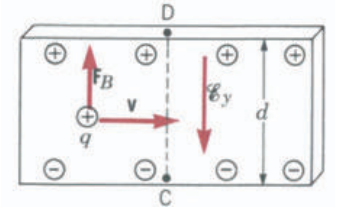


FIGURE 16-7
Behavior of mobile positive charges in the arrangement of Fig. 16-6.

Because the Hall voltage can be readily measured by connecting a voltmeter between D and C, the Hall effect permits the experimental determination of the drift velocity of the charge carriers. We can obtain additional information if we use Eqs. 15.4 or 15.5 as follows

$$i = qNAv$$

or

$$v = \frac{i}{qNA} \tag{16.18}$$

Substituting Eq. 16.18 for v in Eq. 16.17, we obtain

$$V_H = \frac{i\,B\,d}{q\,N\,A} \tag{16.19}$$

Note that A is the cross-sectional area of the foil, hence A = thickness (t) × width (d). Therefore

$$V_H = \frac{1}{qN}\frac{iB}{t} \tag{16.20}$$

$V_H = \frac{1}{qN}\frac{iB}{t}$

or

$$V_H = R_H \frac{iB}{t} \tag{16.21}$$

where $R_H = 1/qN$ is called the *Hall coefficient*. Because i, B, and t are measurable, the magnitude of the Hall voltage will yield the value of N, the density of charge carriers. In the SI system of units, this density will be the number per cubic meter.

Additional important information can be obtained from the Hall effect. In our discussion, we assumed that the charge carriers were positively charged. These charges were deflected toward the upper part of the foil, raising the potential of point D with respect to point C. Suppose, however, that the charge carriers are negatively charged particles. In this case, the drift velocity of the carriers will be opposite to that of $\mathscr{E}_x$, as in Fig. 16-8. Although the velocity vector v is reversed, the direction of the force given by Eq. 16.15 is still upward because the charge q is negative. As a result, the upper part of the strip will have an accumulation of negative charges and the lower part an accumulation of positive charges. Point D will now be at a lower potential than point C. The polarity of the Hall voltage will tell which type of carrier is responsible for conduction. We will see in Chapter 25 that the semiconductors used in logic circuits can be made to have either positive or negative charge carriers.

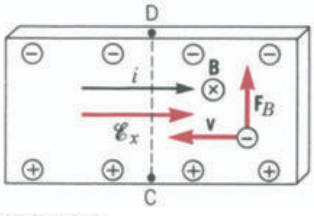


FIGURE 16-8
Behavior of mobile negative charges in the Hall effect experiment.

Example 16-2

A current of 50 A is established in a slab of copper 0.5 cm thick and 2 cm wide. The slab is placed in a magnetic field B of 1.5 T. The magnetic field is perpendicular to the plane of the slab and to the current. The free electron

concentration in copper is 8.4×10^{28} electrons/m^3. What will be the magnitude of the Hall voltage across the width of the slab?

Solution Using Equation 16.19

$$V_H = \frac{1}{Nq}\frac{i\,B\,d}{A}$$

$$= \frac{50\ \text{A} \times 1.5\ \text{T} \times 2 \times 10^{-2}\ \text{m}}{8.4 \times 10^{28}\ \text{m}^{-3} \times 1.6 \times 10^{-19}\ \text{C} \times 10^{-4}\ \text{m}^2}$$

$$= 1.12 \times 10^{-6}\ \text{V}$$

16.8 ELECTROMAGNETIC WAVES: THE NATURE OF LIGHT

In 1670, Christian Huygen had proposed that the propagation of light could be explained by assuming that light is a wave. Huygen's theory was not widely accepted until 1801, when Thomas Young performed the first successful experiment that exhibited the interference of light. Even though Young's experiment established firmly the wave nature of light, one important question remained unanswered: What is the nature of the light wave?

Starting with the fundamental laws of electromagnetism, James Maxwell (1831–1879) in 1873 showed that accelerated charges would produce electromagnetic waves whose velocity of propagation c through free space should be

$$c = 3 \times 10^8\ \text{m/sec}$$

Maxwell and other physicists of that period also showed that an electromagnetic wave consists of an electric field $\mathscr{E}$ and a magnetic field B perpendicular to each other with both $\mathscr{E}$ and B perpendicular to the direction of their propagation. The spacial and temporal behavior of the electric and magnetic fields is identical to the traverse motion of the particles in a string when a traveling wave propagates through it.

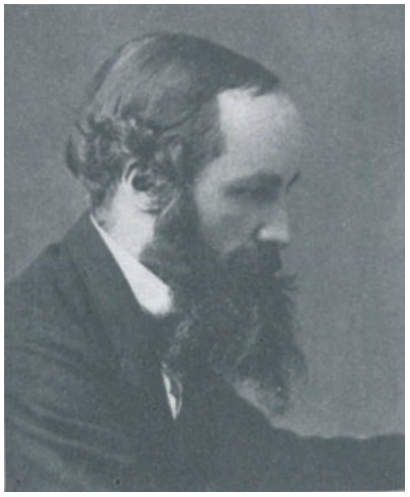
Suppose an electromagnetic wave travels along the x axis. If we measure the value of $\mathscr{E}$ and B at different points along the x axis, at some fixed time t we will observe that both $\mathscr{E}$ and B vary sinusoidally with x. This behavior is illustrated in Fig. 16-9. Similarly, if we sit at a fixed point in space and measure $\mathscr{E}$ and B at that point as a function of time, we observe that both vary sinusoidally with time.

This behavior of the electric and magnetic fields can be represented mathematically as

$$\mathscr{E} = \mathscr{E}_0 \sin(kx - \omega t) \tag{16.22}$$

and

$$B = B_0 \sin(kx - \omega t)$$



James Clerk Maxwell (1831-1879).

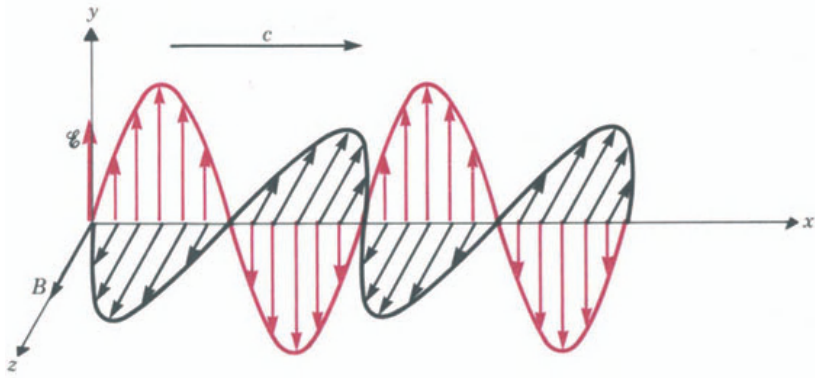


FIGURE 16-9
In an electromagnetic wave, the electric and the magnetic fields are at right angles to each other and to the direction of propagation of the wave.

Equations 16.22 have the same mathematical form as the sinusoidal traveling wave that was introduced in Chapter 11 (see Eq. 11.4 and following). James Maxwell showed that any charge distribution that oscillates sinusoidally with time should produce electric and magnetic fields that behave as described by Eqs. 16.22. Moreover, the frequency ω of the electromagnetic wave should be the same as the frequency of oscillation of the charges producing it. We should indicate at this point that no motion of material particles is involved in the electromagnetic wave, hence, there is no need for a medium of propagation.

One of the key characteristics of a particular type of wave is its velocity of propagation. Maxwell's theoretical prediction was that all electromagnetic waves should travel with velocity $c = 3 \times 10^8$ m/sec. Within the experimental uncertainty this was the value that was measured for the speed of light in 1849. This fact led Maxwell to postulate that light is an electromagnetic wave. Fifteen years after Maxwell's calculations, Heinrich Hertz (1857–1894) was able to produce waves of electromagnetic origin using a circuit with an oscillating current flowing through it. Hertz found that the speed of his electromagnetic waves agreed, within the experimental uncertainties, with the value predicted by Maxwell. The common electromagnetic spectrum is listed in the following table, although there are no upper or lower limits.

Electromagnetic Spectrum

Name	Frequency (Hz)	Wavelength (m)
Gamma rays	10^{23}–10^{19}	10^{-14}–10^{-10}
X rays	10^{20}–10^{16}	10^{-12}–10^{-8}
Ultraviolet rays	10^{17}–10^{15}	10^{-9}–10^{-6}
Visible light	$(4\text{-}7.5) \times 10^{14}$	$(7.5\text{-}4) \times 10^{-7}$
Infrared rays	10^{14}–10^{11}	10^{-5}–10^{-4}
Microwaves	10^{12}–10^{9}	10^{-4}–10^{-1}
Short radio waves	10^{9}–10^{6}	10^{-3}–10^{-2}
FM, TV	10^{8}	1
AM radio	10^{7}–10^{6}	10^{2}–10
Long radio waves	10^{6}–10^{-1}	10^{3}–10^{7}

Heinrich Hertz (1857-1894).

All the laws of electromagnetic waves apply to waves of the entire electromagnetic spectrum. Some wavelengths, such as visible light, are more accessible for experiment than are other wavelengths, but all have been at least partially checked for consistency with this model.

PROBLEMS

16.1 What force is experienced by a wire of length $l = 0.08$ m at an angle of 20° to the magnetic field direction carrying a current of 2 A in a magnetic field of 1.4 T?

16.2 The earth's magnetic field at the equator is 4×10^{-5} T and is parallel to the surface of the earth in the south-north direction. (Note that the earth's geographic north pole is the magnetic south pole.) A wire 2 m long of mass $m = 9$ g is suspended by a string. The wire is also parallel to the earth's surface and carries a current of 150 A in the east-west direction. (a) What is the tension on the string? (b) What would be the tension if the current was in the west-east direction?

(*Answer*: (a) 10.02×10^{-2} N, (b) 7.62×10^{-2} N.)

16.3 A rectangular wire loop carrying a current $i = 5$ A is hung by a string from the end of a rod pivoted about its midpoint, as in Fig. 16-10. The lower part of the loop is in a region where there is a uniform magnetic field $B = 2$ T perpendicular to the plane of the loop as shown. What weight must be placed on the other end of the pivoted rod to balance it?

(*Answer*: 8 N.)

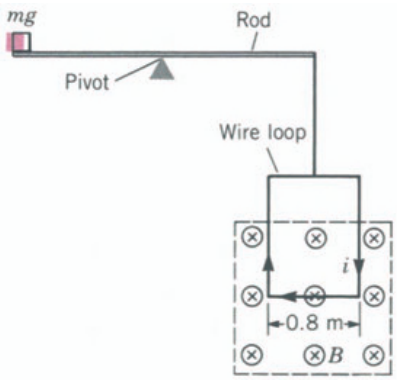


FIGURE 16-10
Problem 16.3.

16.4 The wire of Fig. 16-11 carries a current $i = 2$ A. Find the force on each segment of the wire when it is placed in a region where there is a magnetic field $B = 1.5$ T directed along the positive y axis.

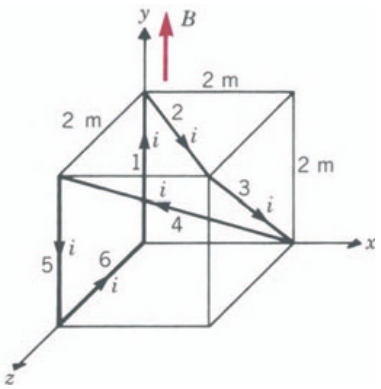


FIGURE 16-11
Problem 16.4.

16.5 The wire loop of Fig. 16-12 carries a current of 2 A. It is placed in a region where there is a magnetic field $B = 0.5$ T parallel to the plane of the loop. (a) Calculate the force on each side of the wire loop. (b) What is the torque on the wire loop?

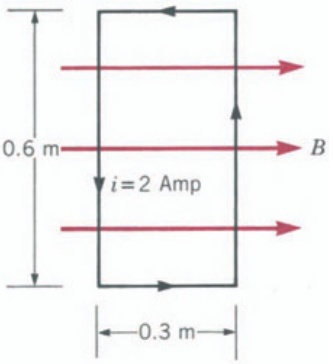


FIGURE 16-12
Problem 16.5

16.6 The wire loop of Fig. 16-13 carries a current $i = 10$ A. It is placed in a uniform magnetic field $B = 1.2$ T. Let $r = 2$ m. (a) Find the net force on

the circular part of the loop. (b) What is the net force on the loop?

(*Answer*: (a) 48 N, in the upward direction, (b) 0.)

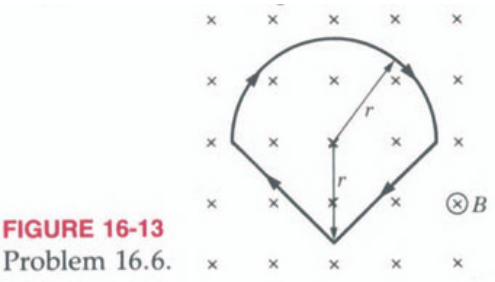


FIGURE 16-13
Problem 16.6.

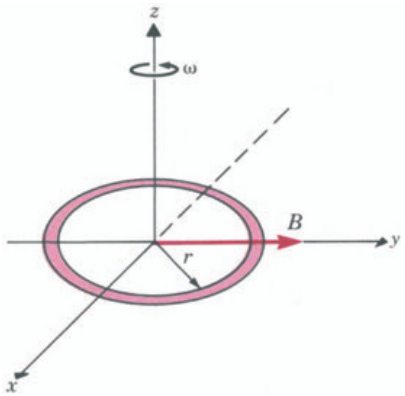


FIGURE 16-14

16.7 What is the maximum torque that acts on a coil of wire that consists of 10 loops of diameter 0.04 m and carries a current of 2×10^{-3} A in a magnetic field of 3×10^{-2} T?

16.8 A wire of length l is to be used to make a coil of one or several circular loops through which a current i will be passed. (a) Show that the maximum magnetic dipole is obtained by making a single loop. (b) What is the dipole moment?

16.9 How much work must be done to rotate the loop of Problem 16.6 from the position shown in Fig. 16-13 to a position where the magnetic field is parallel to the plane of the loop?

16.10 A magnetic dipole μ with a moment of inertia I is placed in a uniform magnetic field B. Initially μ is in the equilibrium position, that is, parallel to B. The dipole is then rotated by a *small* angle θ and then released. (a) Show that the subsequent motion of the dipole is approximately simple harmonic. (b) What is the period of the motion? (*Hint*: For small angles θ, $\sin\theta \approx \theta$.)

16.11 A ring made of an insulator with radius $r = 0.2$ m has a uniformly distributed charge $q = 4 \times 10^{-4}$ C. The ring is placed in the x-y plane of a region where there is a uniform magnetic field $B = 1.2$ T directed along the positive y axis (see Fig. 16-14). The ring is rotated about the z axis with constant angular velocity $\omega = 20$ rev/sec. (a) What is the magnetic moment of the rotating ring? (b) What is the torque exerted by the magnetic field on the ring?

(*Answer*: (a) 1.01×10^{-3} A-m^2, (b) 1.21×10^{-3} N-m.)

16.12 A proton is moving with a velocity $\mathbf{v} = (3 \times 10^5\,\mathbf{i} + 7 \times 10^5\,\mathbf{k})$ m/sec in a region where there is a magnetic field $\mathbf{B} = 0.4\,\mathbf{j}$ T. What is the force experienced by the proton?

(*Answer*: $(1.92\,\mathbf{k} - 4.48\,\mathbf{j}) \times 10^{-14}$ N.)

16.13 A proton is accelerated through a potential difference of 200 V. It then enters a region where there is a magnetic field $B = 0.5$ T. The magnetic field is perpendicular to the direction of motion of the proton. What is the force experienced by the proton?

16.14 A charged particle q is projected in the region between two parallel plates. In the region of the plates there is an electric field $\mathscr{E} = 50{,}000$ N/C and a magnetic field $B = 0.1$ T. The electric field is perpendicular to the magnetic field, and both are perpendicular to the direction of motion, as shown in Fig. 16-15. If the particle goes through the plates undeflected, what is the velocity of the particle?

(*Answer*: 5×10^5 m/sec.)

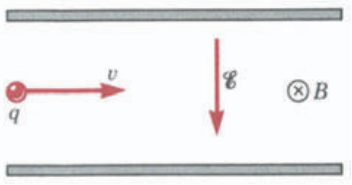


FIGURE 16-15
Problem 16.14.

16.15 A proton is accelerated through a potential difference of 300 V. It then enters a region where there is a magnetic field $B = 0.8$ T and an electric field $\mathscr{E}$. The electric field is perpendicular to the magnetic field, and both are perpendicular to the direction of motion of the proton (see Fig. 16-16). The

proton moves through undeflected. What is the value of $\mathscr{E}$?

(*Answer*: 1.92×10^5 N/C.)

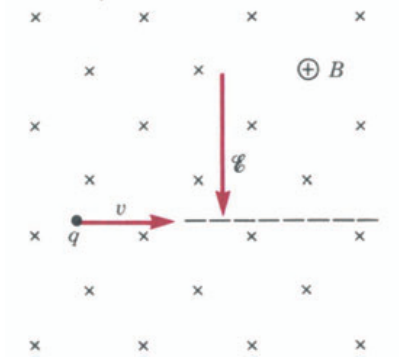


FIGURE 16-16
Problem 16.15.

16.16 A particle of mass $m = 15$ g and charge $q = 3 \times 10^{-3}$ C is moving horizontally near the surface of the earth with a velocity $v = 50$ m/sec. What is the magnitude and direction of the smallest magnetic field B that will keep the particle moving in a straight line? Ignore the electric field mentioned in Problem 15.6.

16.17 An electron is moving with a velocity $\mathbf{v_1} = (2 \times 10^6\,\mathbf{i} + 4 \times 10^6\,\mathbf{j})$ m/sec in a region where there is a uniform magnetic field, it experiences a force $\mathbf{F_1}$ along the z axis. A second electron with velocity $\mathbf{v_2} = 3 \times 10^6\,\mathbf{k}$ m/sec experiences a force $\mathbf{F_2} = 7 \times 10^{-13}\,\mathbf{i}$ N. (a) What is the direction and the magnitude of the magnetic field? (b) What is the magnitude and direction of $\mathbf{F_1}$?

(*Answer*: (a) $1.46\,\mathbf{j}$ T, (b) $4.67 \times 10^{-13}\,\mathbf{k}$ N.)

16.18 A strip of copper 1 cm wide and 1 mm thick has 50 A of current passing through it. The strip is in a magnetic field of 0.5 T directed into the paper (see Fig. 16-17). The voltage difference $V_H = V_C - V_D = 2 \times 10^{-6}$ V and the observation that V_C is larger that V_D indicates that the conduction is by electrons. What is the density of the electrons responsible for the current? ($e = 1.6 \times 10^{-19}$ C.)

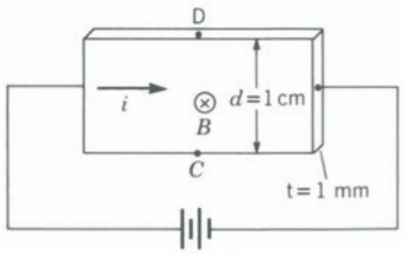


FIGURE 16-17
Problem 16.18.

16.19 A rectangular slab of silicon of thickness 1 mm is used to measure an unknown magnetic field B. The free electron concentration of that particular type of silicon is 6×10^{24} electrons per m^3. When the slab is placed in the region of the magnetic field, perpendicular to the field, and the current in the slab is 20 mA, the Hall voltage is 150 μV. What is the strength of the magnetic field?

(*Answer*: 7.2 T.)

16.20 A long metal plate of width $d = 1$ cm is moved with constant velocity v in a region where there is a magnetic field $B = 0.9$ T (see Fig. 16-18). A potential difference $V = 4.5 \times 10^{-3}$ V appears across two points D and C. The line joining the two points is perpendicular to the direction of motion of the plate. What is the velocity v?

(*Answer*: 0.5 m/s.)

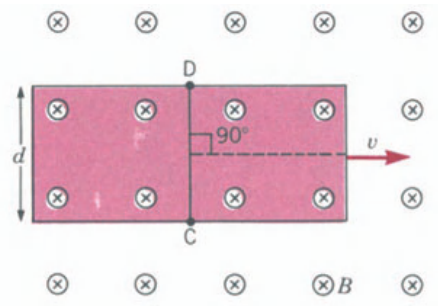
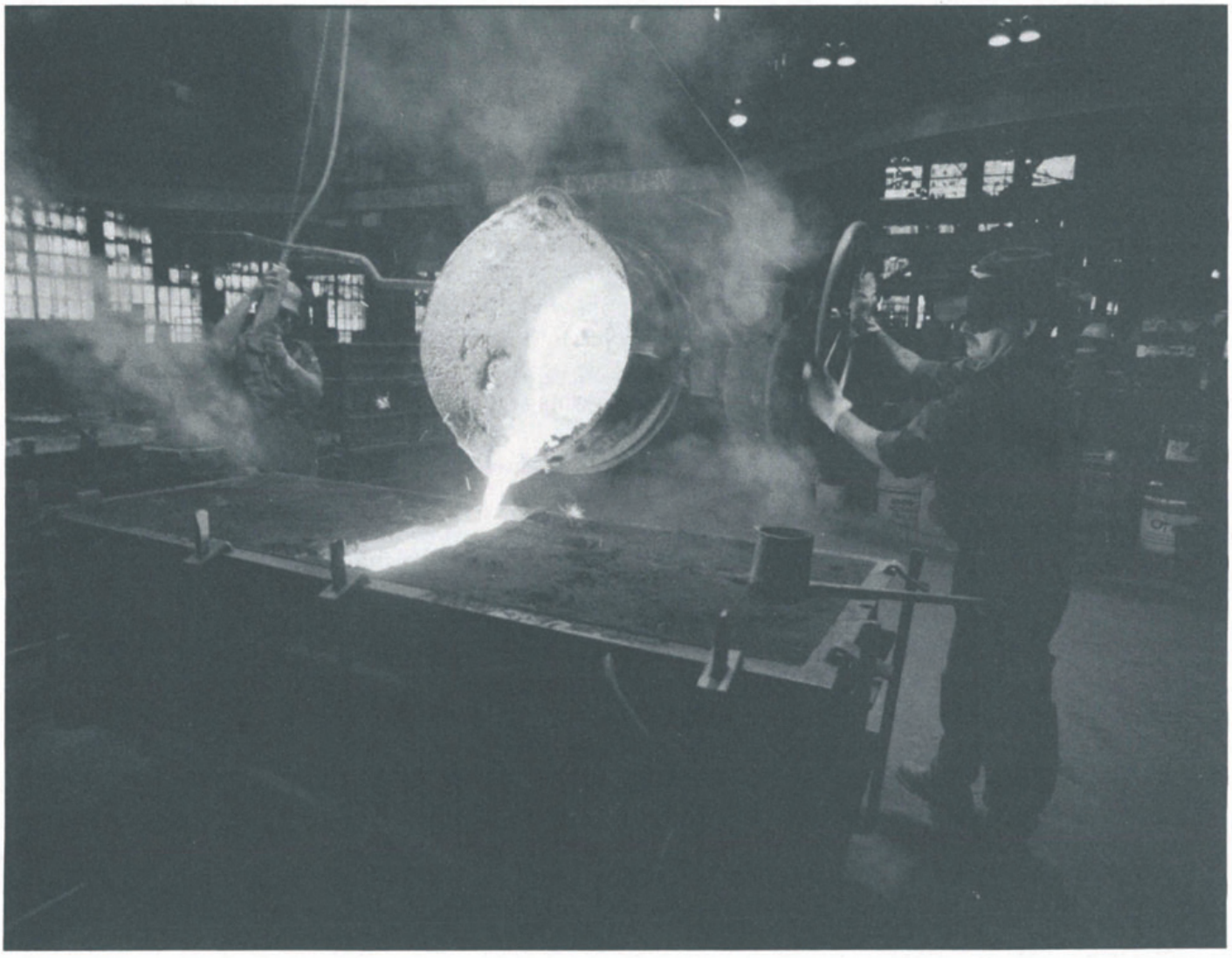


FIGURE 16-18
Problem 16.20.



CHAPTER 17

The Beginning of the Quantum Story

17.1 INTRODUCTION

By the end of the nineteenth century, most physicists felt rather good about the state of their art. In fact, some felt that their successors would spend their time simply taking measurements to the next decimal place. There were reasons for this complacent attitude. Most of the astronomical data about the motion of the planets, as well as the behavior of ordinary mechanical systems, could be explained using Newton's laws of motion and his law of universal gravitation. The empirical laws concerning electric and magnetic fields had been discovered and fused together by Maxwell, and his prediction of the existence of electromagnetic waves had been experimentally verified by Heinrich Hertz: The nature of light was no longer a mystery. More important, the same laws used to explain the behavior of macroscopic systems were also able to explain the behavior of submicroscopic objects (atoms and molecules). This came about with the development of the techniques of statistical mechanics. By applying Newton's laws statistically the ideal gas law, $PV = nRT$ could be derived. Similarly, the specific heat of gases could be predicted in agreement with the available experimental data.

There were a few minor problems. We will mention two of them that were instrumental in the advent of the scientific revolution that today we call modern physics. The principle of relativity seemed to fail when applied to electromagnetism. The principle states that the laws of physics should be the same in all inertial frames of reference. Someone performing experiments in a spaceship moving with constant velocity with respect to the earth obtains the same results from the experiments as does an experimenter on earth. Because physical laws reflect the results of the experiments, it follows that these laws have to be the same (must have the same mathematical form) in all inertial frames of reference. This mathematical invariance was shown to be preserved with the laws of mechanics, but it broke down with the laws of electricity and magnetism. This "minor" problem eventually led to development of Einstein's *special theory of relativity*.

Another problem that baffled physicists at the beginning of the twentieth century was the nature of the spectrum emitted by a class of objects called *blackbodies*. The predictions of classical ideas did not fit the experimental results. This problem led to the development of what we now call *quantum mechanics*. Relativistic effects do not generally affect computer operation, so we will address only the quantum mechanical part of modern physics in the remainder of this book.

17.2 BLACKBODY RADIATION

All substances at finite temperatures radiate electromagnetic waves. Isolated atoms (in a gas) emit discrete frequencies, molecules emit bands of frequencies, and solids radiate a continuous spectrum of frequencies.

The details of the spectrum emitted by a solid depend on its temperature and to some extent on its composition. At room temperature the spectrum is centered around the infrared; that is, most of the radiation emitted lies in the infrared part of the electromagnetic spectrum. As the temperature of the solid increases, more and more of the emitted radiation is in the visible part of the spectrum; we see it first glow red and then approach white as the temperature is increased.

Objects that emit a spectrum of *universal* character, one that does not depend on the composition of the object, are called blackbodies. The reason for the name is that these objects absorb all the radiation incident on them. They do not reflect light, and hence they appear black. We see them by contrast with other objects or their background. Any object painted with a dull black pigment is a good approximation to an ideal blackbody. Another type of blackbody is a metallic cavity with a small hole (see Fig. 17-1). Any radiation entering the hole has a very small probability of being reflected out, hence the object (hole) is "black." After multiple partial reflections by the inner walls of the cavity, the radiation is eventually absorbed by the atoms in the walls of the cavity. These atoms, in turn, will reradiate electromagnetic waves into the cavity and some of it will leak out through the hole. Theoretically, the character of this radiation that leaks out is the same as that of the other type of blackbody.

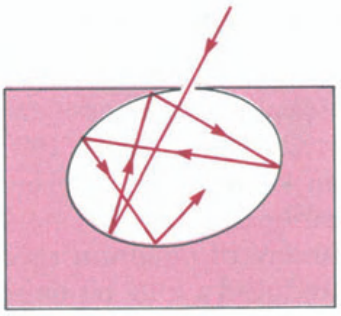


FIGURE 17-1
A metallic cavity with a small hole is an example of a blackbody. Radiation entering the hole is eventually absorbed after successive reflections at the inner walls of the cavity. Some of the radiation reemitted by the atoms in the walls of the cavity leaks out through the hole. This radiation has the same character as that of any other blackbody.

17.2a Character of the Spectrum of a Blackbody

The main features of the spectrum emitted by a blackbody are:

1. The spectrum is continuous with a broad maximum. This fact is shown in Fig. 17-2, which is a plot of $I(\nu)$, the spectral radiancy at each frequency, versus the frequency of the radiation. The spectral radiancy is the energy per frequency emitted per unit time per unit area of the blackbody. The two curves correspond to two different temperatures of the object.
2. The integral of $I(\nu)$ over all ν, which we call I_T, represents the energy emitted per unit time per unit area, regardless of the frequency, and it is found to increase with the fourth power of the temperature. This empirical fact, known as the Stefan-Boltzmann law, states that,

$$I_T = \int_0^\infty I(\nu)\, d\nu = \sigma T^4$$

where the constant $\sigma = 5.67 \times 10^{-8}$ W/m^2-K^4. This integral is clearly the area under the curve for each temperature.

3. Figure 17-2 also shows that the spectrum shifts toward higher frequencies as the temperature increases. In fact, one finds experimentally that the frequency ν_{max}, at which $I(\nu)$ is a maximum, increases linearly with the temperature of the cavity (blackbody), that is,

$$\nu_{max} \propto T$$

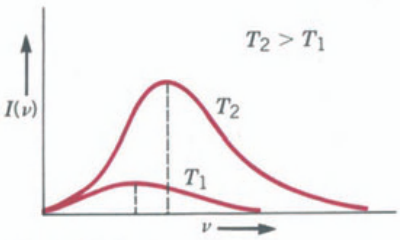


FIGURE 17-2
The intensity of the radiation emitted by a blackbody as a function of the frequency of the radiation for two different temperatures of the blackbody. Note that the total energy (area under the curve) and the frequency at which the intensity is a maximum increase with increasing temperature.

17.2b Planck's Theory

Attempts by physicists to explain the blackbody spectrum using the laws of classical electromagnetism and thermodynamics proved unsuccessful.

On December 14, 1900, at a meeting of the German Physical Society, Max Planck (1858–1947) presented a paper entitled, "On the Theory of the Energy Distribution Law of the Normal Spectrum." This event is considered the birthday of quantum mechanics. As we will see, these ideas were at first introduced a little bit haphazardly, with no justification other than that they accounted for the experimental facts. Eventually, these ideas were fused together into a set of fundamental principles by Erwin Schrödinger and Werner Heisenberg.

Planck's approach to the problem was to find an empirical mathematical expression for $I(\nu)$ that would fit the experimental data. He then observed that he could derive the expression by making a revolutionary physical hypothesis, namely; *a system undergoing simple harmonic motion with frequency ν can only have and therefore can only emit energies given by $E = nh\nu$, where $n = 1, 2, 3, \ldots$* and h is a constant now known as Planck's constant. The value of h, which resulted in a good fit between the data and the expression found by Planck for $I(\nu)$, is 6.63×10^{-34} Joule second (J-sec).

We know that the energy of a harmonic oscillator is proportional to the square of the amplitude of the motion (Section 10.6). In a classical treatment, such as an oscillating spring, this amplitude may vary continuously from zero to infinity. In contrast, Planck postulated that atomic oscillators can have only *discrete* energy values. The classical and Planck's energy spectra for an oscillator are contrasted in Fig. 17-3. By Planck's hypothesis, because an oscillator (such as the atoms in the walls of the cavity) can take only certain values for the energy, when they lose that energy (by emitting electromagnetic waves), they lose it in multiples of $h\nu$. These small quantities of energy are called *quanta* (singular, *quantum*).

Using the energy spectrum just described for the atomic oscillators, and consequently for the electromagnetic waves emitted by them, together with simple thermodynamic arguments, Planck derived a rather complicated expression for $I(\nu)$ that matched the experimental data:

$$I(\nu) = \frac{2\pi h\nu^3}{c^2} \frac{1}{\exp(h\nu/k_B T) - 1} \tag{17.1}$$

where c is the velocity of light, k_B is the Boltzmann constant, ν is the frequency of the electromagnetic wave, and T is the absolute temperature of the blackbody.

Every day experience shows that an oscillator, for example, a pendulum or a mass connected to a spring, stops oscillating progressively and smoothly, not in jumps. Is Planck's hypothesis in conflict with this macroscopic experimental observation? Not really. Let us consider a mass $m = 10$ kg, attached to a spring of force constant $k = 10^3$ N/m. Let the initial amplitude of the

$$E = nh\nu$$

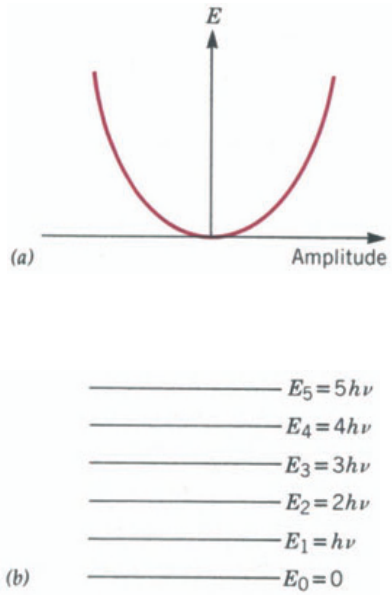


FIGURE 17-3
(*a*) Dependence of the energy of a classical oscillator on the amplitude of the motion. Because the amplitude of the motion can be varied continuously from zero to infinity, the energy of an oscillating body can have any value between zero and infinity. (*b*) Discrete (quantized) energy spectrum of atomic oscillators as proposed by Planck to explain the spectrum of frequencies emitted by a blackbody.

motion be $A = 0.1$ m. From elementary mechanics, we know that

$$E = \frac{1}{2}kA^2 = \frac{1}{2} \times 10^3 \text{ N/m} \times (0.1 \text{ m})^2 = 5 \text{ J}$$

and

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} = 1.59 \text{ Hz}$$

The separation between adjacent energy levels is, by Planck's hypothesis

$$\Delta E = h\nu = 6.63 \times 10^{-34} \text{ J-sec} \times 1.59 \text{ sec}^{-1} \approx 10^{-33} \text{ J}$$

This example shows that, for a macroscopic oscillator, the separation between the allowed energy states (as postulated by Planck) is extremely small compared with the energy of the oscillator. The oscillator may be losing energy in jumps, but the effect is not noticeable.

Max Planck (1858–1947).

17.3 THE PHOTOELECTRIC EFFECT

The quantum idea, introduced by Planck to explain the spectrum of a blackbody, was further expanded by Albert Einstein (1879–1955) in connection with the photoelectric effect. Under certain conditions, which we will discuss shortly, light incident on a metal will cause electrons to be ejected from the surface of the metal. This is known as the *photoelectric effect*.

We will summarize an experiment that can be used to study the properties of the electrons ejected from the metal. We will then see the failure of classical ideas to explain the results and, finally, we will introduce Einstein's hypothesis about the nature of electromagnetic radiation and how this hypothesis accounts for the experimental facts.

17.3a Experimental Facts

An experimental arrangement that can be used to study some of the properties of the photoelectric effect is shown in Fig. 17-4. The apparatus consists of an evacuated tube with two metal plates, C, the cathode, and A, the anode. Monochromatic light (single wavelength) is sent through a quartz window onto the cathode C. Because the anode is at a negative potential V with respect to the cathode, the electrons, on being emitted by the incident light striking the cathode, face a *retarding* voltage V. To reach the anode the photoelectrons must be ejected with a kinetic energy E_k that is greater than the difference in potential energy, $|e|V$, between the anode and the cathode. When $E_k \geq |e|V$, the electrons, on reaching the anode, will be able to contribute to the current through the circuit, which is measured by a galvanometer G. The tube is evacuated to minimize the collisions between the photoelectrons and the gas molecules in the tube. By varying the retarding voltage V, the spectrum of

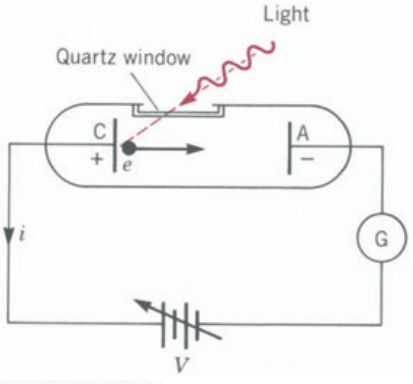


FIGURE 17-4
Experimental arrangement for the study of the photoelectric effect.

energies with which the electrons are emitted can be determined. Other parameters can be varied to see how they affect the energy and the number of photoelectrons emitted; these include the intensity I and the frequency ν of the incident light and the nature of the cathode. A summary of the experimental results is shown in Fig. 17-5.

If the value of the retarding voltage V and of the frequency of the light are kept constant (Fig. 17-5*a*), it is seen that the photocurrent i through the circuit increases linearly with increasing intensity I of the incident radiation. This in turn means that the number of electrons emitted with energies $E_k \geq |e|V$ increases with I, because i is proportional to the number of electrons that are collected by the anode.

Figure 17-5*b* shows the dependence of i, and hence of the number of emitted electrons capable of reaching the anode, on the value of the retarding voltage. The experiment is performed while keeping both the intensity and the frequency of the radiation constant. The two curves correspond to two different values of I. The result can be easily understood. For small V's, only the electrons emitted with small energies are turned back by the retarding voltage, and therefore do not contribute to the current i. As V is increased, electrons with higher energies will be turned back, and the current will decrease. When $V = V_0$ (V_0 is called the *stopping potential*), all the electrons, including the most energetic ones, are turned back and the current drops to zero. V_0 is therefore a measure of the maximum energy with which the electrons are ejected from the cathode,

$$|e|V_0 = E_{k\ \max} \tag{17.2}$$

Figure 17-5*b* shows that V_0, and therefore *the maximum energy of the photoelectrons, is independent of the intensity of the light.*

The dependence of V_0 on the frequency ν of the light can be examined by repeating the previous experiment with different ν's. Figure 17-5*c* shows that V_0 (hence $E_{k\ \max}$) increases linearly with ν. That is,

$$V_0 = a\,\nu \tag{17.3}$$

The value of the slope a is found to be 4.1×10^{-15} J-sec/C. The linear dependence of V_0 on ν and the value of the slope remain unchanged if the experiment is repeated with a cathode made of a different metal. Figure 17-5*c* also shows that for frequencies $\nu \leq \nu_c$ V_0 is zero. This means that no voltage is necessary to stop the most energetic electrons; no voltage is needed because no electrons are emitted when $\nu \leq \nu_c$. As the graph indicates, the value of ν_c depends on the material used for the cathode.

There is one final experimental fact that is crucial to the discussion that will follow. When the conditions for photoemission are favorable (high enough ν, low enough V), the emission is almost instantaneous. The photocurrent has been observed to occur within 10^{-9} sec from the onset of illumination, which is the limit of experimental accuracy. This essentially instantaneous emission has been observed to take place with extremely low intensities of light, as low as 10^{-10} W/m^2.

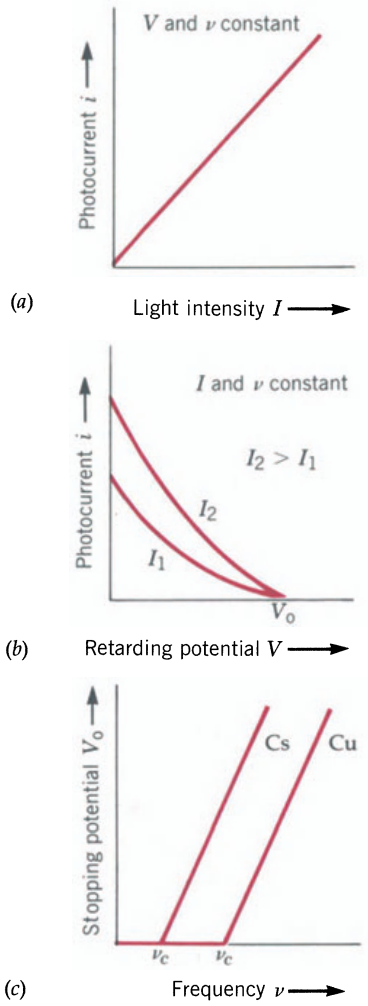


FIGURE 17-5

(*a*) The photocurrent (the number of emitted electrons) increases with increasing light intensity. (*b*) The number of emitted electrons able to reach the anode A decreases as the retarding voltage increases. The stopping potential V_o is independent of the light intensity. (*c*) The stopping potential increases linearly with increasing frequency of the light. For frequencies below ν_c, V_o is zero because no electrons are emitted. The value of ν_c depends on the material being illuminated (the cathode). The two curves are for cesium (Cs) and copper (Cu) cathodes respectively.

17.3b Failure of Classical Physics to Explain the Results

According to classical physics, light is an electromagnetic wave (see Section 16.8). To understand the failure of classical physics to explain the experimental results just presented, we need to remind ourselves of two facts about waves:

1. The energy of a wave is continuously distributed over the entire space traversed by the wave. For example, when the ripples in the pond move outward from their source, all the water in their path is displaced.
2. The intensity of a wave, which represents the energy carried by the wave per unit area perpendicular to the direction of propagation of the wave, per unit time is proportional to the square of the amplitude of the wave (Eq. 11.20). In the case of electromagnetic waves it can be shown that

$$I = \frac{1}{2}\epsilon_0 c \mathscr{E}_0^2$$

where ϵ_0 is the permittivity of free space, c is the velocity of light, and $\mathscr{E}_0$ is the amplitude of the electric field of the wave.

With these two facts in mind, let us consider an electron that is bound with some energy E_b to the metal. An electric field $\mathscr{E} = \mathscr{E}_0 \sin(kx - \omega t)$ impinges on the bound electron. The electric field will exert a force $F = |e|\mathscr{E} = |e|\mathscr{E}_0 \sin(kx - \omega t)$ on the electrons. This force will do work on the electrons, the amount of work depending on the magnitude of the force. As a result, the electric field will increase the energy of the electrons and, if the energy that the electrons pick up from the electromagnetic field is greater than the binding energy that keeps them in the metal, the electrons will come out of the metal with a kinetic energy E_k, which is the difference between the energy absorbed from the wave and the binding energy E_b. As the amplitude of the wave increases, the magnitude of the force increases, and so does the work done by the electromagnetic field on a given electron. We therefore expect, from classical physics, that the energy given to the electron will increase as the intensity of the wave increases; detailed calculations show that the energy absorbed is, indeed, proportional to the intensity.

Let us now reexamine the experimental results. The results of Fig. 17-5*a* can be explained in terms of classical concepts. The electrons in the metal are bound differently: some more tightly than others. Given a certain intensity of the wave and therefore a certain amount of energy available to them, the electrons will use it to liberate themselves from the metal; any remaining energy will be in the form of kinetic energy of the electrons. For small intensities only those electrons that are weakly bound will come out with sufficient kinetic energy to overcome the retarding potential and to contribute to the current. As the intensity is increased the energy available will increase and more electrons will come out with sufficient energy to reach the anode. The current should increase with increasing intensity, and it does.

The fact that $E_{k\ \max}$ is independent of the intensity is difficult to explain by classical theory. If you increase I, you increase the energy available to all

the electrons, including those that are the least tightly bound and that therefore come out with the maximum kinetic energy. Thus the fact that V_0 is independent of I (see Fig. 17-5*b*) cannot be explained by classical ideas.

The fact that $E_{k\ \max}$ increases with ν (see Fig. 17-5*c*) cannot be accounted for by classical physics. As we have seen, the energy of the electromagnetic wave depends on its intensity (amplitude squared), not on its frequency. Why should V_0 depend on ν? Why is there a ν_c below which no electrons are emitted, no matter how intense the wave is? Classical physics provides no answer.

Finally, the fact that the emission is almost instantaneous plays a key role in the rejection of the classical ideas about the nature of electromagnetic radiation. If we consider radiation with intensity $I = 10^{-10}$ W/m^2, there is no way that the electrons can be emitted in 10^{-9} sec. It should take considerably longer. Let us consider a sheet of some metal with an area of 1 m^2, as shown in Fig. 17-6, and let us assume that light of intensity $I = 10^{-10}$ W/m^2 shines on it. As we mentioned, the energy of the beam is spread continuously over the entire wave front. Let us be optimistic and assume that all the energy falling on a certain atomic site of the metal sheet is absorbed by only one of the electrons of the atom, the most loosely bound. It is well known, from X-ray studies, that the interatomic separation d in a metal is about 2 Å (1 Å = 10^{-10} m). We can use this fact to find out how many atomic sites there are in the first layer of the metal surface.

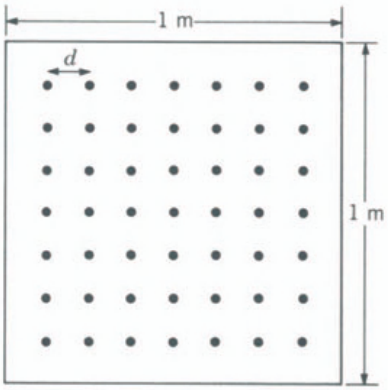


FIGURE 17-6

$$\text{The number of atoms in a 1-m-long row} = \frac{1\text{ m}}{d} = \frac{1\text{ m}}{2 \times 10^{-10}\text{ m}} = 5 \times 10^9 \text{ atoms/row}$$

For simplicity, let the metal have cubic structure. This means that there are as many rows as we have atoms in one row. And, consequently, the number of atoms in the first layer of the metal sheet will be $(5 \times 10^9)^2 = 2.5 \times 10^{19}$ atoms/layer. According to our generous assumption 10^{-10} J/sec-m^2 are shared by 2.5×10^{19} electrons. That is,

$$\text{Energy/second-electron} = \frac{10^{-10}\text{ J/sec}}{2.5 \times 10^{19}\ e} = 4 \times 10^{-30}\text{ J/sec-}e$$

It is known from other types of experiments that the minimum binding energy of an electron in a metal is typically a few eV's. Let us take 1 eV. The time required for the electron to collect 1 eV from the electromagnetic wave will be,

$$t = \frac{1\text{ eV} \times 1.6 \times 10^{-19}\text{ J/eV}}{4 \times 10^{-30}\text{ J/sec}} = 4 \times 10^{10}\text{ sec}$$

$$\text{or} \sim 10^5 \text{ days}$$

Thus, we see that classical physics also fails to explain the short time release of electrons in the photoelectric effect.

17.3c Einstein's Theory

Albert Einstein (1879-1955).

In 1905, five years after Planck's historic paper, Einstein was able to explain the photoelectric effect by proposing a theory about the nature of electromagnetic radiation that was dramatically different from that of classical electromagnetism.

According to Einstein, *the energy of an electromagnetic wave of frequency ν is not continuously distributed over the entire wave front, but instead it is localized in small bundles (particle-like entities) called photons*. The energy of each photon is $E_{\text{photon}} = h\nu$, where h is Planck's constant (see Section 17.2b).

Basically, according to Einstein, a beam of electromagnetic radiation carries energy like a beam of particles, not like a wave. Within the Einstein hypothesis, the intensity of the beam is a measure of the density of photons in the beam. Increasing the intensity without changing the frequency does not change the energy of the individual photons, but rather the number of photons per unit volume of the beam, and thus the energy density of the beam.

Einstein visualized the photoelectric effect as a particle-particle collision in which a photon of energy $h\nu$ collides with an electron in the metal and imparts all its energy to the electron. From conservation of energy principles,

$$h\nu = E_k + E_b \tag{17.4}$$

$E_{\text{photon}} = h\nu$

where E_b is the energy with which the particular electron is bound to the metal and E_k is the kinetic energy with which that electron is ejected. From Eq. 17.4 it is clear that the value of E_k will depend on how tightly bound a given electron is. The smaller E_b, the larger E_k will be. We can rewrite Eq. 17.4 as

$$E_{k\ \max} = h\nu - \phi \tag{17.5}$$

$E_{k\ \max} = h\nu - \phi$

where ϕ is the minimum binding energy and is called the *work function* of the metal. (Note that ϕ is not an angle but a symbol for energy in this case.) Experimentally, $E_{k\ \max}$ is measured by determining the stopping potential V_0 and $E_{k\ \max} = eV_0$, thus Eq. 17.5 can be rewritten as

$$V_0 = \frac{h}{e}\nu - \frac{\phi}{e} \tag{17.6}$$

We can now explain all the experimental data presented at the beginning of this section. The greater the intensity I, the larger the number of photons that will strike the metal cathode every second. This will result in a greater number of photon-electron collisions and a subsequent increase in the number of electrons emitted. Thus the results of Fig. 17-5*a* are explained.

Increasing the intensity increases the number of photons, not the energy of the individual photons. The maximum energy (and therefore V_0) should not depend on the number of photons (on intensity I) but rather on the energy of each photon, that is, on the frequency of the wave. In fact, Eq. 17.6 shows that V_0 should increase linearly with ν. This is in agreement with the result

presented in Fig. 17-5*c*. The slope of the curve according to Eq. 17.6 should be equal to $h/e = 4.1 \times 10^{-15}$ J-sec/C, which is the observed value. It should be clear that if the energy of the photons is less than the work function of the metal, no electrons can be ejected. That is, if

$$h\nu < \phi \quad \text{or} \quad \nu < \frac{\phi}{h} = \nu_c \tag{17.7}$$

$\nu_c = \frac{\phi}{h}$

no photoemission will take place. The cut-off frequency ν_c is accounted for.

Finally, the emission is instantaneous because the process is not one in which the electrons progressively gather energy until they have enough to come out. It is a particle-particle collision. If just one photon with energy $h\nu \geq \phi$ collides with an electron, the latter will be immediately ejected.

Example 17-1

The eye is capable of detecting 10 eV of light energy. If we take as the average wavelength of light 6000 Å, how many photons is the eye capable of detecting?

Solution According to Einstein's theory, the energy of a photon is given by $h\nu$. Using this, together with the fact (Eq. 11.2) that the product of the wavelength and the frequency equals the velocity of propagation of the wave, which in the case of light is c, that is,

$$\lambda\nu = c \tag{17.8}$$

we obtain

$$\text{Energy/photon} = h\nu = \frac{hc}{\lambda}$$

$$= \frac{6.63 \times 10^{-34}\ \text{J-sec} \times 3 \times 10^{8}\ \text{m/sec}}{6000 \times 10^{-10}\ \text{m}}$$

$$= 3.32 \times 10^{-19}\ \text{J} = 2.07\ \text{eV}$$

$$\text{Number of photons} = \frac{10\ \text{eV}}{2.07\ \text{eV/photon}} = 5\ \text{photons}$$

Example 17-2

The cut-off frequency for photoemission in copper is 1.0×10^{15} Hz. What is the maximum kinetic energy of the photoelectrons emitted when light of wavelength 1000 Å is shone on a copper surface?

Solution The work function is given by Eq. 17.7

$$\phi = h\nu_c = 6.63 \times 10^{-34}\ \text{J-sec} \times 1.0 \times 10^{15}\ \text{Hz} = 6.63 \times 10^{-19}\ \text{J}$$

From Eq. 17.5

$$E_{k\,\max} = h\nu - \phi = \frac{hc}{\lambda} - \phi$$

$$= \frac{6.63 \times 10^{-34}\ \text{J-sec} \times 3 \times 10^{8}\ \text{m/sec}}{1000 \times 10^{-10}\ \text{m}} - 6.63 \times 10^{-19}\ \text{J}$$

$$= 13.26 \times 10^{-19}\ \text{J} = 8.29\ \text{eV}$$

17.4 FURTHER EVIDENCE FOR THE PHOTON THEORY

There exists today a large number of experimental results that confirm the particle nature of electromagnetic radiation. In this section, we will discuss qualitatively two effects that contribute further evidence to the theory.

17.4a X-ray Production

In 1895, Wilhelm K. Roentgen (1845–1923) discovered that when highly energetic electrons struck a solid target, a strange (hence the name X rays) kind of radiation was produced. The radiation was highly penetrating: It passed through objects that were opaque to light. Moreover, the radiation was not deflected by either electric or magnetic fields, indicating that it did not consist of charged particles. The mysterious nature of X rays vanished a few years later when in 1912 Max von Laue (1879–1960) found that they could be diffracted. The diffraction of X rays by crystalline solids was discussed in Chapter 12. These experiments proved that X rays are a form of electromagnetic radiation. Their wavelength, however, is much smaller than that of light waves: typically $\lambda \sim 1$ Å.

Figure 17-7 shows a schematic of the experimental arrangement used to produce X rays. Electrons from a heated filament F are accelerated by a large potential difference V (several thousand volts). As a result, they enter the target T with a kinetic energy $E_k = eV$. On striking the target, X rays are emitted.

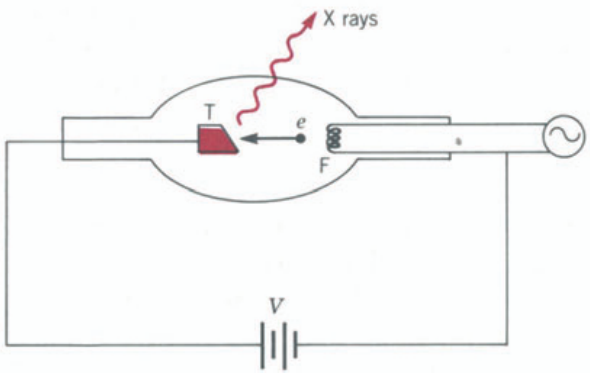


FIGURE 17-7
Apparatus for the production of X rays.

An analysis of the spectrum of the emitted X rays reveals several interesting features. We will concern ourselves primarily with one of them that is pertinent to the photon hypothesis. The emission spectrum is continuous, with one very important feature: a sharp, well-defined cut-off on the small wavelength side. These facts are shown in Fig. 17-8, which is a schematic plot of the intensity I of the emitted X rays versus wavelength λ. The value of the cut-off wavelength λ_{min} is independent of the target material but depends on the accelerating voltage V by Eq. 17.9

$$\lambda_{min} \propto \frac{1}{V} \tag{17.9}$$

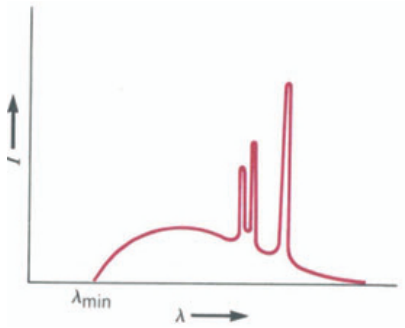


FIGURE 17-8
Intensity of X rays emitted versus the wavelength of the X rays. Note that no X rays of wavelength less than a critical value λ_{min} are emitted.

Using the fact that $\lambda\nu = c$ (Eq. 17.8), we can rewrite this result as,

$$\nu_{max} \propto V \tag{17.10}$$

Although it is not relevant to our present discussion, we should point out, for the sake of completeness, that in addition to the continuous spectrum there are several intensity peaks, as shown in Fig. 17-8, called the *characteristic X rays*. The wavelengths of the characteristic X rays are independent of the voltage V but depend on the material of the target.

Let us try to understand the origin of the spectrum presented in Fig. 17-8. When the incoming electron enters the target, it will interact with the atomic electrons and with the nuclei present there. The interaction with the atomic electrons is the process that is primarily responsible for the slowing down of the incident electrons. Through multiple collisions, an incident electron is progressively slowed down and loses its energy to the target: The kinetic energy of the bombarding electron becomes heat. Occasionally, an electron-electron collision may occur, which results in a large transfer of energy from the incident electron to an atomic electron. As a result of this collision, the atomic electron will be knocked out of an atom. In subsequent chapters we will see that the electrons in the atom occupy discrete energy levels. When one of these energy levels is vacated as a result of a collision, one of the outer electrons in the atom falls down into the vacated energy state and in the process gives off a photon. The energy of this photon $h\nu$, will be equal to the energy difference between the two atomic levels involved in the transition. These photons account for the characteristic X-ray peaks. Because, as will be shown in later chapters, the atomic energy levels are determined by the structure of the particular atom, we can understand why the frequency of the characteristic X rays depends on the target material and not on the energy of the incident electrons, that is, on the accelerating voltage V.

The incident electrons can also interact with the nuclei in the target. This interaction is responsible for the continuous spectrum or, its customary name, *bremsstrahlung* (braking radiation). Let us consider an electron with energy E_k approaching a positively charged nucleus. The electron, as a consequence of the coulombic attraction of the nucleus, will be deflected from its straight-line path; that is, it will be radially accelerated (see Fig. 17-9). Classical electromagnetic theory predicts that an accelerated charge will radiate electro-

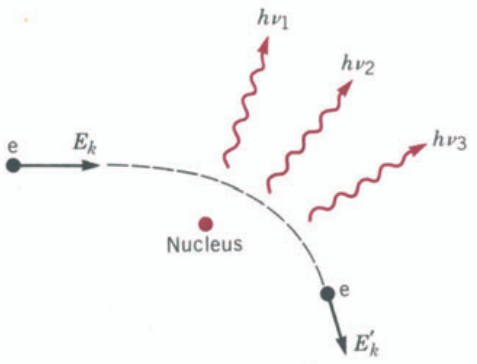


FIGURE 17-9
Schematic representation of an electron interacting with a nucleus in the target and emitting X rays in the process.

magnetic waves *continuously and of all frequencies*. Thus, from the classical view of radiation, it is impossible to understand why there is a wavelength cut-off in the emission spectrum. The cut-off can be explained rather simply by the photon model of electromagnetic radiation. The accelerated electron will radiate energy not continuously but in quanta of energy $h\nu$. If now we consider the electron of Fig. 17-9 approaching the nucleus with energy E_k, emitting several photons of energy $h\nu_1$, $h\nu_2$, and so on, and finally leaving the nucleus with energy E'_k, we can, from energy conservation considerations, write

$$E_k - E'_k = h\nu_1 + h\nu_2 + \cdots$$

We have assumed that the nucleus does not acquire any energy during the collision. This is a good approximation because the nucleus is much heavier than the electron. It is now easy to understand the existence of the cut-off frequency. The most energetic photon that can be produced by the interaction of the electron with the nucleus is the one that is produced when the electron loses all its energy in the emission of a single photon. In such a case E'_k is zero and therefore

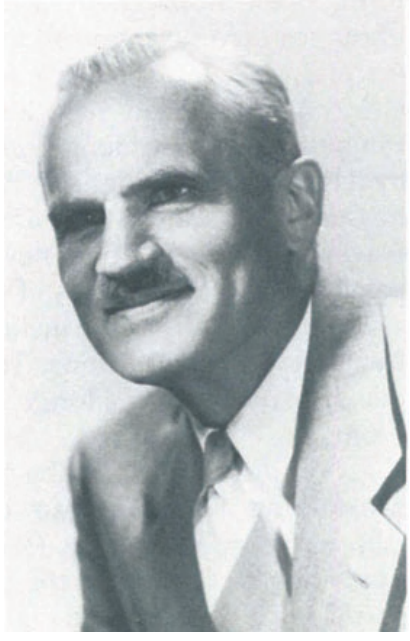
$$E_k = h\nu_{\max}$$

But because $E_k = eV$, we conclude that

$$\nu_{\max} = \frac{e}{h} V \tag{17.11}$$

$$\nu_{\max} = \frac{e}{h} V$$

Equation 17.11 predicts the experimental observation given by Eq. 17.10. Not only can the photon model predict the existence of the frequency cut-off and its proper dependence on the accelerating voltage, but the experimental results can be used to determine the value of Planck's constant. The value of h obtained in this case agrees with the values obtained by Planck in connection with blackbody radiation and by Einstein in the case of the photoelectric effect.



Arthur Holly Compton (1892-1962).

17.4b Compton Effect

In 1923, Arthur H. Compton (1892–1962) performed a series of experiments that provided a dramatic confirmation of the photon nature of electromagnetic radiation. A schematic of the experimental arrangement is shown in Fig.

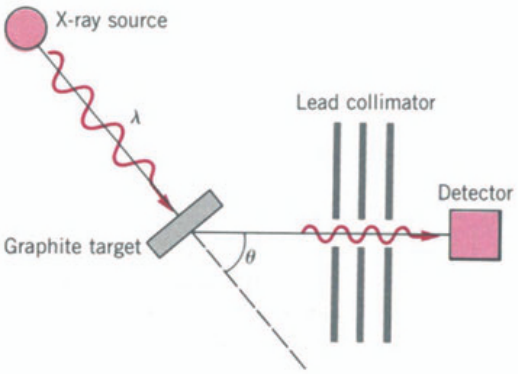


FIGURE 17-10
Apparatus used for the Compton effect experiment. Monochromatic X rays of wavelength λ are scattered by a graphite target. The wavelength of the scattered X rays is determined for different angles θ.

17-10. A beam of X rays of sharply defined wavelength was sent onto a graphite target. Compton then studied the scattered radiation to see what wavelengths were present in it. This was done for different angles θ, between the incident and the scattered beams. Whereas the incident beam consisted of X rays of wavelength $\lambda = 0.709$ Å, Compton observed that the scattered beam contained two intensity maxima: one at $\lambda = 0.709$ Å and the other at a λ' greater than that of the incident radiation. The value of λ' increased as the angle of scattering increased and reached a maximum value of 0.758 Å for $\theta = 180°$. Figure 17-11 shows the intensity of the scattered radiation as a function of wavelength for four particular values of θ.

To understand how these results support the photon hypothesis, let us first consider what classical electromagnetic theory predicts about the scattering of electromagnetic waves. When an electric field $\mathscr{E} = \mathscr{E}_0 \sin(kx - \omega t)$ impinges on an electron in the target material, the electron will experience a force $\mathbf{F} = e\mathscr{E} = e\mathscr{E}_0 \sin(\omega t - kx)$, see Eq. 14.1. As a result of this force, the electron will oscillate with a frequency equal to that of the force, that is, the frequency of the incident electromagnetic wave. On the other hand, according to classical electromagnetic radiation theory, a charged particle (the electron in this case) undergoing simple harmonic motion radiates electromagnetic waves of the same frequency as the frequency of the motion of the charged particle (see Section 16.8). The electron plays the role of a transfer agent: It absorbs energy from the incident beam and reradiates this energy at the same frequency in all directions. Thus, classical electromagnetic theory cannot explain the presence of a longer wavelength (smaller frequency) in the scattered beam.

Compton explained the shift in the wavelength of the scattered beam by considering it to be a beam of photons, each with energy $E = h\nu$ and momentum $p = h\nu/c = h/\lambda$ (see the Supplement at the end of this chapter). According to Compton, the photons collide with the electrons in a particle-particle-like collision (see Fig. 17-12). In the collision, the electron will acquire some momentum and energy at the expense of the photon. From consider-

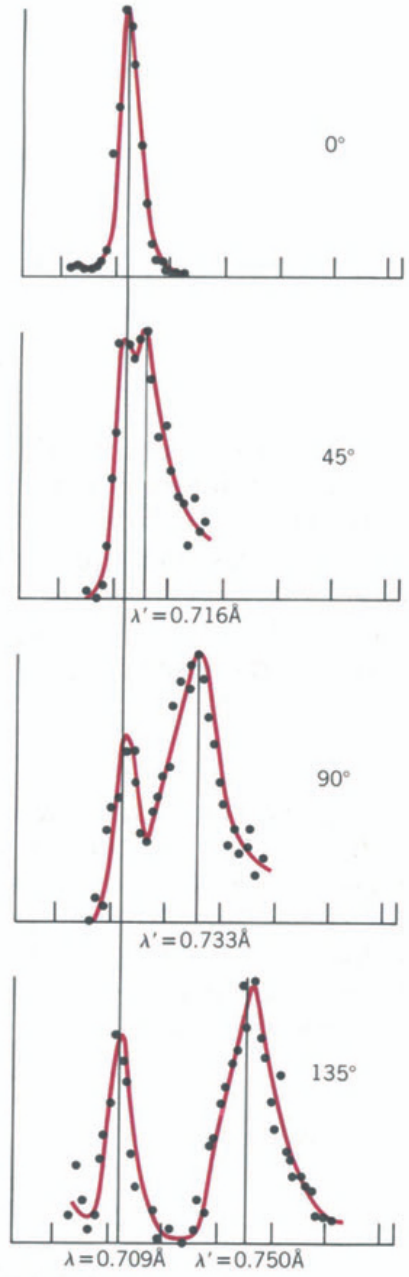


FIGURE 17-11
Intensity of the scattered X rays in the Compton experiment of Fig. 17-10 as a function of the wavelength for four different angles θ. (*Source:* Kenneth Krane, *Modern Physics*. Copyright © 1983 by John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.)

ation of the conservation of energy we conclude that the energy of the photon and, hence, its frequency will decrease: the wavelengths of the scattered photons will be longer than that of the incident photons. Clearly, the stronger the collision, the larger the angle of scattering of the photon and the greater the energy lost to the electron; the shift in frequency should increase with increasing θ. In fact, we expect that the maximum energy transfer will occur in the case of head-on collision, which will result in backward scattering ($\theta = 180°$) of the photon. These arguments can be made quantitative by writing explicitly the conservation of energy and momentum equations. The solution of these equations, which can be found in most modern physics textbooks, yields the frequency and wavelength of the scattered photons as a function of θ. The result, originally derived by Compton, is

$$\lambda' - \lambda = \frac{h}{mc}(1 - \cos\theta) \tag{17.12}$$

$$\lambda' - \lambda = \frac{h}{mc}(1 - \cos\theta)$$

where λ' and λ are, respectively, the wavelengths of the scattered and the incident photons; m is the mass of the scattering particle, the electron, and c is the velocity of light. An inspection of Eq. 17.12 corroborates the qualitative arguments presented earlier. In particular, $\lambda' - \lambda$ is a maximum when $\cos\theta = -1$, that is, when $\theta = 180°$. This maximum shift in wavelength is therefore

$$(\lambda' - \lambda)_{\max} = \frac{2h}{mc} = \frac{2 \times 6.63 \times 10^{-34}\ \text{J-sec}}{9.1 \times 10^{-31}\ \text{kg} \times 3 \times 10^{8}\ \text{m/sec}}$$

$$= 0.049 \times 10^{-10}\ \text{m} = 0.049\ \text{Å}$$

For $\theta = 90°$

$$\lambda' = \lambda + \frac{h}{mc}(1 - \cos 90°)$$

$$= 0.709 \times 10^{-10}\ \text{m}$$

$$+ \frac{6.63 \times 10^{-34}\ \text{J-sec}}{9.1 \times 10^{-31}\ \text{kg} \times 3 \times 10^{8}\ \text{m/sec}}(1 - 0)$$

$$= 0.733 \times 10^{-10}\ \text{m} = 0.733\ \text{Å}$$

which agrees with the results shown in Fig. 17-11*c*.

The unshifted wavelength present in the scattered beam can also be explained in terms of the photon model. Up to now, we have assumed that the photons collide with the electrons in the target material. The photons can also collide with the atoms in the graphite. In this case, the only change to be made in the calculations consists in replacing in Eq. 17.12 the mass of the electron by that of the carbon atom. For graphite, $m_{\text{atom}} \approx 24{,}000\ m_{\text{electron}}$. From Eq. 17.12, we can see that the shift in wavelength will be 24,000 times

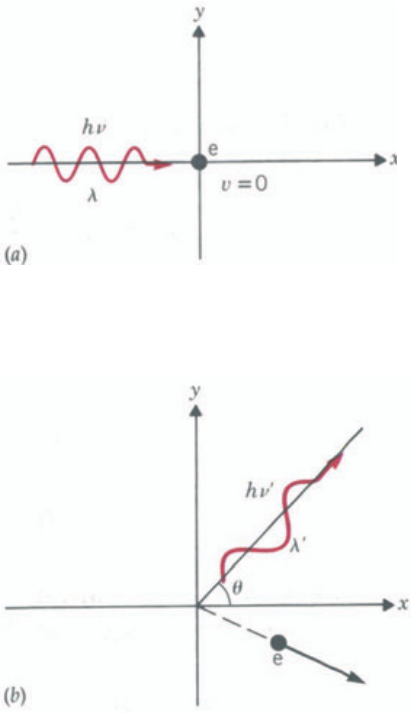


FIGURE 17-12
Scattering process of the X rays in the Compton experiment of Fig. 17-10. A photon collides with an electron in the graphite target in a particle-particle-like collision. The photon imparts some energy to the electron, resulting in a decrease in its own energy (and therefore in its frequency) and a concommitant increase in its wavelength. (*a*) Before the collision. (*b*) After the collision.

smaller, that is,

$$(\lambda' - \lambda)_{max} \approx \frac{0.049 \text{ Å}}{24{,}000} = 2 \times 10^{-6} \text{ Å}$$

This is an insignificant and unobservable amount when we compare it with $\lambda = 0.709$ Å.

Example 17-3

X rays of wavelength $\lambda = 0.700$ Å are Compton-scattered by the electrons in a graphite target. (a) What is the wavelength of the X rays scattered at an angle $\theta = 120°$? (b) What is the kinetic energy of the scattering electrons if they were originally at rest? (c) What is the scattering angle of the electrons?

Solution

(a) From Eq. 17.12

$$\lambda' = \lambda + \frac{h}{mc}(1 - \cos 120°)$$

$$= 0.700 \times 10^{-10} \text{ m} + \frac{6.63 \times 10^{-34} \text{ J-sec}}{9.1 \times 10^{-31} \text{ kg} \times 3 \times 10^{8} \text{ m/sec}}(1 + 0.5)$$

$$= 0.736 \times 10^{-10} \text{ m} = 0.736 \text{ Å}$$

(b) From conservation of energy principles, the kinetic energy of the electron is equal to the energy lost by the photon, that is,

$$E_k = h\nu - h\nu'$$

which from Eq. 17.8, $\lambda\nu = c$, can be written as

$$E_k = \frac{hc}{\lambda} - \frac{hc}{\lambda'} = hc\left(\frac{1}{\lambda} - \frac{1}{\lambda'}\right)$$

$$= 6.63 \times 10^{-34} \text{ J-sec} \times 3 \times 10^{8} \text{ m/sec}\left[\frac{1}{0.700 \times 10^{-10} \text{ m}} - \frac{1}{0.736 \times 10^{-10} \text{ m}}\right]$$

$$= 1.39 \times 10^{-16} \text{ J} = 869 \text{ eV}$$

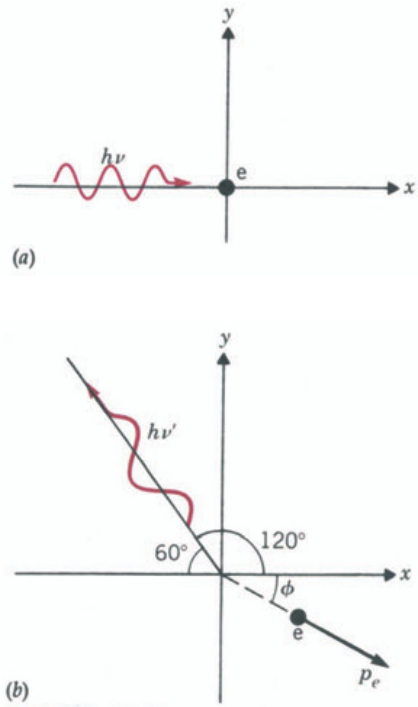


FIGURE 17-13
Example 17-3. (*a*) Before the collision. (*b*) After the collision.

(c) From conservation of linear momentum principles

$$p_x \text{ (before the collision)} = p_x \text{ (after the collision)}$$

Writing this explicitly (see Fig. 17-13 and Eq. 17.18 in Supplement 17-1).

$$\frac{h\nu}{c} = p_e \cos \phi - \frac{h\nu'}{c} \cos 60°$$

therefore

$$p_e \cos \phi = \frac{h\nu}{c} + \frac{h\nu'}{c} \cos 60°$$

$$= \frac{h}{\lambda} + \frac{h}{\lambda'} \cos 60°$$

$$p_e \cos \phi = 6.63 \times 10^{-34} \text{ J-sec} \left[\frac{1}{0.700 \times 10^{-10} \text{ m}} + \frac{0.5}{0.736 \times 10^{-10} \text{ m}} \right]$$

$$p_e \cos \phi = 13.98 \times 10^{-24} \text{ kg m/sec} \qquad (17.13)$$

Again from conservation of momentum,

$$p_y \text{ (before the collision)} = p_y \text{ (after the collision)}$$

$$0 = -p_e \sin \phi + \frac{h\nu'}{c} \sin 60°$$

Therefore

$$p_e \sin \phi = \frac{h\nu'}{c} \sin 60° = \frac{h}{\lambda'} \sin 60°$$

$$p_e \sin \phi = \frac{6.63 \times 10^{-34} \text{ J-sec}}{0.736 \times 10^{-10} \text{ m}} \times (0.866)$$

$$p_e \sin \phi = 7.80 \times 10^{-24} \text{ kg m/sec} \qquad (17.14)$$

Combining Eqs. 17.13 and 17.14, we obtain

$$\tan \phi = \frac{7.80 \times 10^{-24} \text{ kg m/sec}}{13.98 \times 10^{-24} \text{ kg m/sec}} = 0.56$$

$$\phi = 29.2°$$

SUPPLEMENT 17–1

Momentum of the Photon

From Einstein's theory of special relativity, the total relativistic energy of a particle is given by

$$E = E_k + E_0$$

$$E = mc^2 \tag{17.15}$$

where E_k is the kinetic energy of the particle, $E_0 = m_0c^2$ is the rest energy (m_0 is the rest mass of the particle, that is, the mass of the particle when its velocity is zero), m is the relativistic mass (the mass when the particle is moving), and c is the velocity of light. For the photon, the rest energy and therefore the rest mass is zero; a photon at rest does not exist.

We can now use Einstein's postulate, namely, $E_{\text{photon}} = h\nu$ and, combining this with the expression for the total relativistic energy, Eq. 17.15, we can get an expression for the momentum of the photon.

$$E_{\text{photon}} = h\nu = mc^2$$

Dividing by c, we obtain

$$mc = \frac{h\nu}{c} \tag{17.16}$$

Using Eq. 17.8, we obtain

$$mc = \frac{h}{\lambda} \tag{17.17}$$

We recognize the left side of Eqs. 17.16 and 17.17 as the momentum p of the photon, that is, the product of the mass and the velocity of the photon,

$$p_{\text{photon}} = \frac{h}{\lambda} = \frac{h\nu}{c} \tag{17.18}$$

$p_{\text{photon}} = \frac{h}{\lambda} = \frac{h\nu}{c}$

PROBLEMS

17.1 The wavelength λ_{max} for which the spectral radiancy, I, of a blackbody is a maximum, is given by Wien's displacement law, which states:

$$\lambda_{\text{max}}T = 2.898 \times 10^{-3}\ \text{m-K}$$

where T is the absolute temperature of the blackbody. The temperature of the surface of the sun is roughly 5800 K. (a) What is the wavelength of the most intense radiation emitted by the sun? (b) In what part of the electromagnetic spectrum is this radiation?

17.2 The rate at which the sun's energy strikes the earth is known as the solar constant and has a value of 2 cal/cm^2-sec. If we assume that there is no reflection by the clouds, ice caps, or such, the earth would absorb all this energy. The earth would be a blackbody radiator and would radiate all the energy

that it receives back into space. What would be the equilibrium temperature of the earth if our assumption were true?

(*Answer*: 1100 K.)

17.3 The radius of a hydrogen atom is approximately 10^{-10} m (1 Å). Light of intensity 1.0 W/m^2 is shone on such an atom. What is the time lag for the photoelectric effect on the basis of the wave theory of light? The binding energy of the electron in the hydrogen atom is 13.6 eV.

(*Answer*: 69.3 sec.)

17.4 A 5000-W radio transmitter emits radiation of frequency $\nu = 1100$ kHz. How many photons per second does it emit?

17.5 Consider a 100-W sodium vapor lamp radiating energy uniformly in all directions. Assume that 80% of the energy radiated is in the form of photons of wavelength 5890 Å. (a) What is the rate of photon emission by the lamp? (b) How far from the lamp will the average density of photons be 2 photons/cm^2-sec? (c) What is the photon flux (that is, the number of photons per unit time per unit area) 2.0 m from the lamp?

(*Answer:* (a) 2.37×10^{20} sec^{-1}, (b) 3.07×10^7 m, (c) 4.71×10^{14} sec^{-1} cm^{-2}.)

17.6 For a quick estimate of the energy of a photon in eV, physicists use the relation $E(\text{eV}) = 12{,}345/\lambda(\text{Å})$. By what percentage is this relation inaccurate?

17.7 The basis for the creation of the latent image on a photographic negative is the dissociation of molecules of silver bromide, AgBr. The heat of dissociation of AgBr is 24 kcal/mole. Find the longest wavelength of light that is just able to expose the negative, that is, dissociate AgBr.

(*Answer*: 11,900 Å.)

17.8 A metal has a work function $\phi = 1.5$ eV. (a) What is the stopping potential for light of wavelength 3000 Å? (b) What is the maximum velocity of the emitted photoelectrons?

17.9 Light of wavelength 1500 Å falls on an aluminum surface having a work function of 4.2 eV. (a) What is the kinetic energy of the fastest emitted photoelectrons? (b) What is the stopping potential? (c) What is the cut-off frequency ν_c for aluminum?

(*Answer*: (a) 4.09 eV, (b) 4.09 V, (c) 1.01×10^{15} Hz.)

17.10 When light of wavelength 2000 Å is incident on the surface of a metal, the electrons are emitted with a maximum kinetic energy of 2.0 eV. (a) Calculate the energy of the incident photons. (b) What is the work function of the metal? (c) If the incident light had a wavelength of 6000 Å, what would be the stopping potential?

17.11 The cut-off frequency for photoemission for a given metal is ν_0. What is the maximum energy of the emitted electrons when the metal is illuminated with light of frequency $3\nu_0$?

17.12 When light of frequency ν_0 is incident on a certain metal surface electrons are emitted with a maximum kinetic energy of 15 eV. When the frequency is reduced to $\nu_0/2$, the maximum kinetic energy is 3 eV. What is the cut-off frequency for photoemission for this metal?

(*Answer*: 2.17×10^{15} Hz.)

17.13 In a photoelectric effect experiment, it is found that when the surface of sodium metal is illuminated with light of wavelength $\lambda = 4200$ Å, the stopping potential $V_0 = 0.65$ V. When the metal is illuminated with light of wavelength $\lambda = 3100$ Å, the stopping potential is $V_0 = 1.69$ V. Calculate Planck's constant from these data.

17.14 Not every photon striking the surface of a metal undergoes a collision with the electrons in the metal. An important quantity in the extension of the photoelectric effect theory is the quantum efficiency, namely, how many photons are required on the average to yield one photoelectron. In a typical experiment, light of wavelength $\lambda = 4366$ Å is shone on a potassium surface. The observed yield is 8×10^{-3} C / J. How many photons are required to yield one photoelectron?

(*Answer*: 44 photons.)

17.15 Electrons in an X-ray tube are accelerated through a potential difference of 5000 V. What is the

maximum frequency and the minimum wavelength of the X rays produced?

17.16 Alpha particles (charge $+2e$) are accelerated by an electric potential difference of 20,000 V. The α particles strike a metal target and in the process produce X rays. Find the smallest wavelength of the X rays emitted by the target.

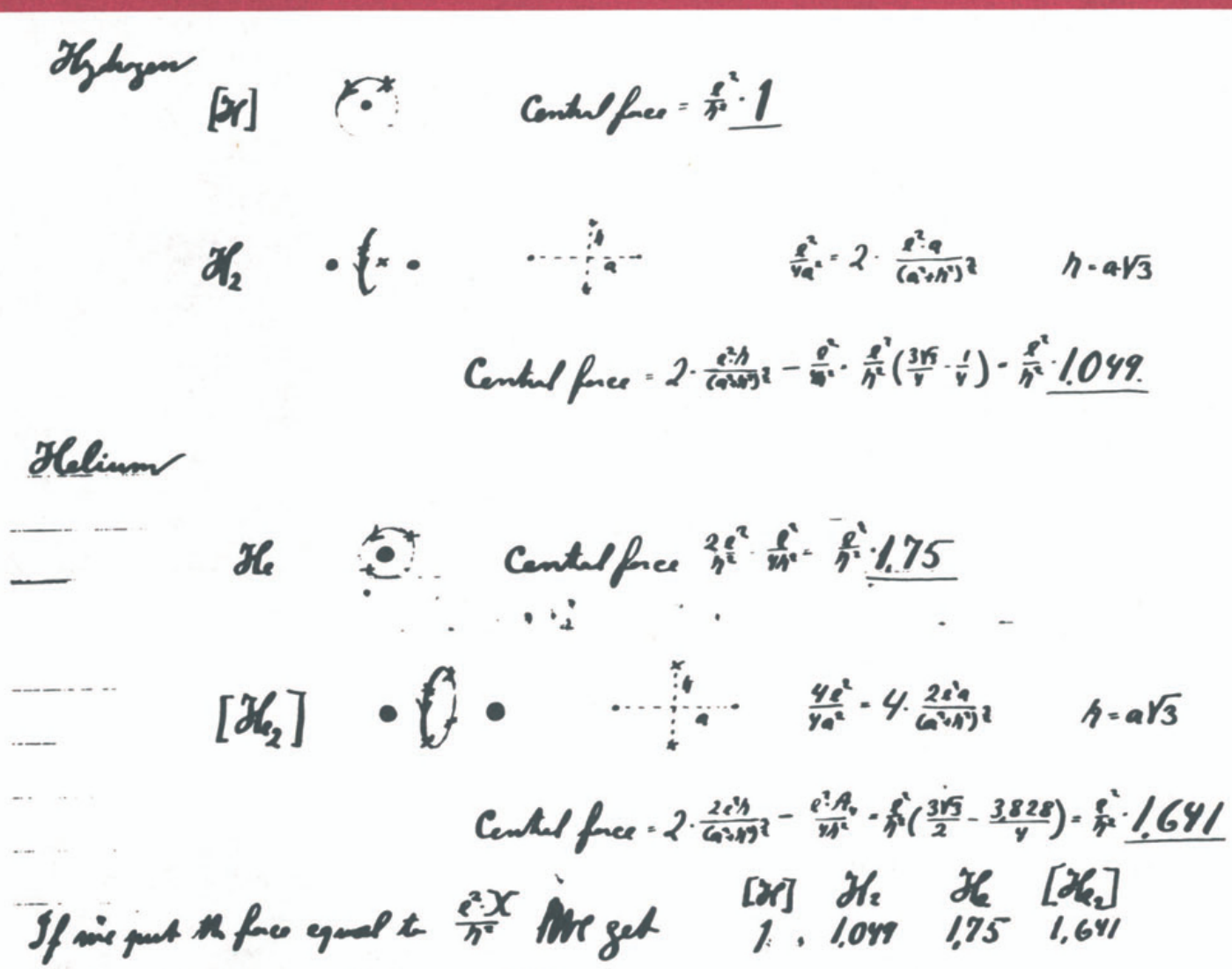
(*Answer*: 0.311 Å.)

17.17 A photon of frequency $\nu = 3 \times 10^{18}$ Hz is Compton-scattered by an electron initially at rest. After the collision, the electron moves in the direction of the incident photon. (a) Find the wavelength of the scattered photon. (b) What is the energy of the scattered electron?

17.18 X rays of wavelength 1 Å are Compton-scattered by the electrons in a carbon target. (a) Calculate the wavelength of the X rays scattered at 90° with respect to the incident X rays. (b) What is the energy of the electrons causing the scattering?

17.19 In a Compton experiment the wavelength of the incident photon is 1.3249 Å, whereas that of the scattered photon is 1.3461 Å. (a) At what angle is the photon scattered? (b) At what angle is the electron scattered? (c) What is the kinetic energy of the scattered electron?

(*Answer*: (a) 82.7°, (b) 48.1°, (c) 148 eV.)



Neils Bohr's original sketches of Hydrogen and Helium.

CHAPTER 18

Atomic Models

18.1 INTRODUCTION

Joseph John Thompson (1856-1940).

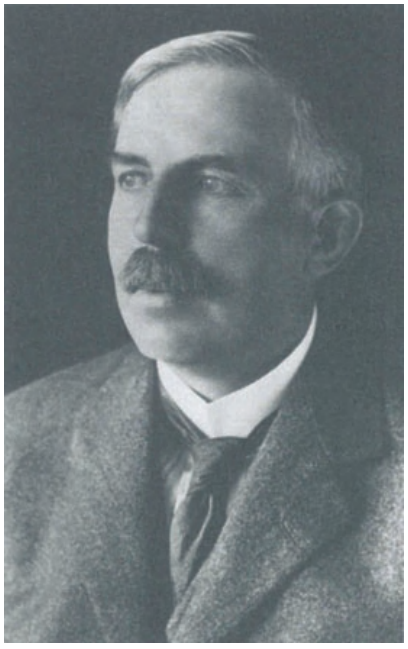
By the beginning of the twentieth century, several types of experimental data existed, including the photoelectric effect, that indicated that the atom contained electrons. On the other hand, the atoms were known to be electrically neutral. This implies that the atom carries an amount of positive charge equal in magnitude to the charge of its electrons.

Even for the lightest of elements, the mass of the atom is thousands of times greater than that of the electron. The conclusion is obvious. Most of the mass of the atom resides either with the positively charged matter or with neutral matter. However, at that time neutral matter was neither known nor conceived of.

Out of these facts, J. J. Thompson (1856–1940) proposed an atomic model. The positively charged matter was assumed to have a *continuous*, *uniform*, *spherical* distribution. The radius of this charge and mass distribution was assumed to be the known radius of the atom, ($\sim 10^{-10}$ m). Within this positively charged mass, the electrons were uniformly "sprinkled" like the "plums in the pudding." Although this model is in agreement with the facts just mentioned, it was not consistent with the known facts of electromagnetic theory. However, the real blow to the Thompson model came as a result of the work of one of Thompson's former graduate students, E. Rutherford.

18.2 THE RUTHERFORD MODEL

18.2a The Scattering of α Particles



Ernest Rutherford (1871-1937).

In 1911, E. Rutherford (1871–1937) and two co-workers, H. Geiger and E. Marsden, did a series of experiments in which positively charged α (alpha) particles, which are actually the nuclei of the helium atoms, emitted by certain radioactive materials were sent through a collimator and then shot through thin metallic foils. (A *collimator* is a set of closely spaced plates with small aligned holes. The plates are opaque to the radiation falling on them. As a result of this geometric arrangement of the plates, only the radiation incident perpendicularly on the collimator is transmitted through the holes in the plates.) They studied the angular dependence of the scattered α particles, that is, what fraction of the α particles were scattered at a given angle θ (see Fig. 18-1).

What they found was that most of the α particles were scattered in the forward direction at very small angles. A few, however, were deflected through large angles. For example, in one experiment α particles with kinetic energy $E_k = 7.7 \times 10^6$ eV were shot through a gold foil 10^{-6} m thick. It was then observed that approximately 1 in 10^4 α particles were scattered through an angle $\theta \geq 90°$. When a detailed calculation using the Thompson model was

made, it was found that this fraction should not be 1 in 10^4 but rather 1 in 10^{3500}. This was the fatal blow to the Thompson model.

Rutherford suggested that the experimentally observed large-angle scattering of α particles could be explained on the basis of a "nuclear planetary model". This model proposes the following arrangement. All the positive charge and consequently almost all of the mass of the atom is concentrated in a very small volume at the center of the atom ($r \sim 10^{-14}$ m). The electrons orbit this nucleus, like planets around the sun, with the electrostatic attraction of the positive nucleus supplying the centripetal force. The radius of the orbits is of the same order of magnitude as the size of the atom, that is, $r \sim 10^{-10}$ m. Using this model and newtonian mechanics, Rutherford was able to predict the experimental angular dependence for the scattered α particles.

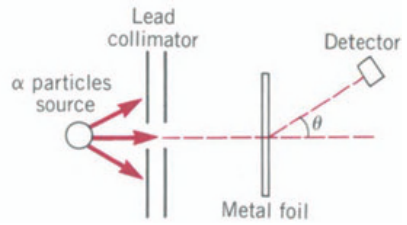


FIGURE 18-1
Schematic of Rutherford α particles scattering experiment. The number of α particles scattered by the atoms in the metal foil is measured for different values of the scattering angle θ.

18.2b Difficulties of the Nuclear Planetary Model

Although the planetary model proposed by Rutherford was able to explain properly the scattering of α particles, it presented new problems. Solution of these problems required a drastic modification of physical concepts and led to the first quantum mechanical model, the Bohr model, which we will examine shortly.

Let us consider the simplest of atoms, the hydrogen atom. An electron of charge $q = -e$ rotates in a circular orbit of radius r under the electrostatic attraction of the nucleus ($q = +e$) (see Fig. 18-2). For simplicity we will assume that the nucleus remains stationary. To understand the difficulties of the model, we will derive expressions for the total energy of the atom and for the frequency of rotation.

$$E_{\text{TOTAL}} = \text{kinetic energy} + \text{potential energy}$$

$$E_{\text{TOTAL}} = \frac{1}{2}mv^2 - \frac{1}{4\pi\epsilon_0}\frac{e^2}{r} \tag{18.1}$$

We can express E_{TOTAL} in terms of the radius of the orbit alone, by making use of the fact that it is the electrostatic attraction of the nucleus that supplies the centripetal force that causes the circular motion of the electron.

$$F_{\text{radial}} = ma_{\text{radial}}$$

$$\frac{1}{4\pi\epsilon_0}\frac{e^2}{r^2} = m\frac{v^2}{r} \tag{18.2}$$

Multiplying both sides by (1/2) r

$$\frac{1}{2}mv^2 = \frac{1}{2}\frac{1}{4\pi\epsilon_0}\frac{e^2}{r} \tag{18.3}$$

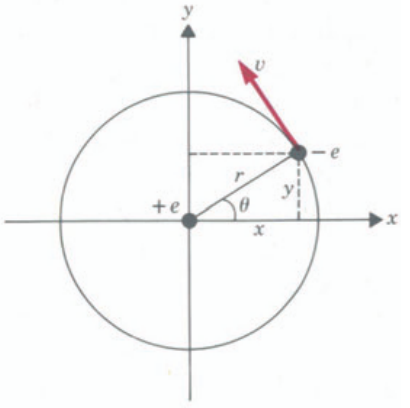


FIGURE 18-2
Electron in the hydrogen atom moving around the positive nucleus in a circular orbit.

Substituting this result for $1/2\ mv^2$ in Eq. 18.1, we get

$$E_{\text{TOTAL}} = -\frac{1}{2}\frac{1}{4\pi\epsilon_0}\frac{e^2}{r} \tag{18.4}$$

The frequency of rotation ν can be expressed as

$$\nu = \frac{\text{velocity}}{\text{distance}} = \frac{v}{2\pi r} \tag{18.5}$$

We can now express ν as a function of the radius of the orbit alone. From Eq. 18.2,

$$v = \left\{\frac{1}{4\pi\epsilon_0}\frac{e^2}{mr}\right\}^{1/2} \tag{18.6}$$

We substitute for v in the expression for the frequency (Eq. 18.5) and obtain

$$\nu = \left\{\frac{e^2}{16\pi^3\,\epsilon_0 m}\right\}^{1/2}\frac{1}{r^{3/2}} \tag{18.7}$$

The difficulties presented by the Rutherford model with respect to classical physics will now be explained. A particle moving in a circle with constant frequency $\omega = 2\pi\nu$ is equivalent to a particle simultaneously undergoing simple harmonic motion in two mutually perpendicular directions. This can be readily seen. From Fig. 18-2, the x and y coordinates of the electron are given by

$$x = r\cos\theta = r\cos\omega t = r\cos(2\pi\nu t)$$

$$y = r\sin\theta = r\sin\omega t = r\sin(2\pi\nu t)$$

which are expressions for simple harmonic motion in the x and y direction, respectively, see Eq. 10.8.

Classical electromagnetism predicts that a charged particle undergoing simple harmonic motion should radiate electromagnetic waves of the same frequency as the frequency of vibration of the particle. The electron moving in an orbit of radius r should radiate electromagnetic radiation. The frequency of this radiation will be given by Eq. 18.7. Because this electromagnetic radiation carries away energy, the energy of the atom must decrease as energy is radiated from it. From Eq. 18.4 this implies that the radius of the orbit of the electron will decrease continuously as the atom loses energy, until the electron eventually collapses into the nucleus. Thus, according to principles of classical physics, the Rutherford model is unstable. Moreover, because the frequency of the motion depends on r (see Eq. 18.7), as the radius of the orbit decreases continuously, the frequency of the motion, and correspondingly the frequency of the emitted radiation, will increase continuously. By the classical model the spectrum emitted by such an atom should be a continuous one. The experimental facts are: The hydrogen atom is not unstable and, as will be discussed in Section 18.3, the spectrum of hydrogen is not continuous. Instead, it consists of certain discrete frequencies.

18.3 THE SPECTRUM OF HYDROGEN

A typical experimental arrangement used to study the electromagnetic spectrum emitted by an element is shown in Fig. 18-3. A small amount of the element in the gaseous state is introduced into a glass tube containing two metallic electrodes. An electronic discharge is produced by applying a high voltage between the two electrodes. This discharge causes the gas in the tube to emit electromagnetic radiation. A sample of the radiation emitted is sent through a collimator and through a diffraction grating. We may consider the resulting interference pattern as it is projected onto a screen. The interference pattern of a grating is similar to that of a double slit (see section 12.4). It consists of a central maximum ($\theta = 0$) that contains all the wavelengths present in the incident radiation. In addition, however, secondary maxima will be observed on the screen at locations where the condition is the same as that of a double-slit pattern (see Eq. 12.6)

$$d \sin \theta = n\lambda \qquad n = 1, 2, 3, \cdots \tag{18.8}$$

Recall that d is the separation between slits in the double-slit condition, and this is also the separation between adjacent slits in the diffraction grating. If the incident radiation contains several wavelengths (and frequencies), the angles at which these interference maxima occur will be different for different wavelengths because from Eq. 18.8, $\sin \theta = n\dfrac{\lambda}{d}$. We can therefore separate and measure the wavelengths present in the radiation.

Using methods similar to the one described, scientists in the nineteenth century had investigated the spectra of a large number of substances. For example, in the case of hydrogen, they had found that the spectrum consists of families of lines (wavelengths); the lines of a given family could be fitted to a simple empirical relation known as the Rydberg-Ritz formula, which yields a series of lines

$$\frac{1}{\lambda} = R\left\{\frac{1}{n_k^2} - \frac{1}{n_j^2}\right\} \tag{18.9}$$

$$\frac{1}{\lambda} = R\left\{\frac{1}{n_k^2} - \frac{1}{n_j^2}\right\}$$

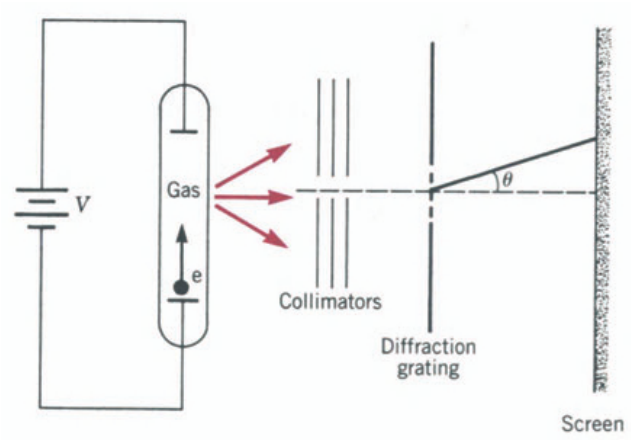


FIGURE 18-3
Apparatus used to measure the wavelengths present in the radiation emitted by the atoms of a gas. An electronic discharge, produced by a high voltage between two electrodes in the tube containing the gas, causes the emission of electromagnetic radiation by the atoms of the gas. A sample of this radiation is incident on a diffraction grating. The angle θ at which the interference maxima for the different wavelengths occur is measured.

where $R = 1.0967757 \times 10^7\ \text{m}^{-1}$ (R is known as the Rydberg constant) and n_k and n_j are integers. For a given n_k, $n_j = n_k + 1, n_k + 2, n_k + 3, \ldots$. For example,

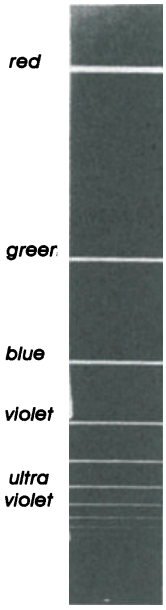
If $n_k = 1$, then $n_j = 2, 3, 4, \cdots$ This family is known as the Lyman series

If $n_k = 2$, then $n_j = 3, 4, 5, \cdots$ This family is known as the Balmer series

If $n_k = 3$, then $n_j = 4, 5, 6, \cdots$ This family is known as the Paschen series

If $n_k = 4$, then $n_j = 5, 6, 7, \cdots$ This family is known as the Brackett series

The Lyman series corresponds to frequencies in the ultraviolet part of the spectrum, the Balmer series to frequencies in the visible, and the other series to those in the infrared, each named after its discoverer. It should be emphasized that Eq. 18.9 was a purely empirical relation; no physical explanation existed.



Photograph of the visible part of the spectrum emitted by hydrogen (the Balmer series).

Example 18-1

What are the shortest and longest wavelengths of the Lyman series?

Solution From Eq. 18.9 the longest wavelength corresponds to the smallest value of n_j, which for the Lyman series is 2. Therefore

$$\frac{1}{\lambda_{\text{longest}}} = R\left(\frac{1}{1^2} - \frac{1}{2^2}\right)$$

$$= 1.0968 \times 10^7\ \text{m}^{-1}\,[1 - 0.25]$$

$$= 0.8226 \times 10^7\ \text{m}^{-1}$$

$$\lambda_{\text{longest}} = \frac{1}{0.8226 \times 10^7\ \text{m}^{-1}}$$

$$= 1.215 \times 10^{-7}\ \text{m} = 1215\ \text{Å}$$

The shortest wavelength corresponds to the largest value of n_j that is ∞.

$$\frac{1}{\lambda_{\text{shortest}}} = 1.0968 \times 10^7\ \text{m}^{-1}\left(\frac{1}{1^2} - \frac{1}{\infty}\right)$$

$$= 1.0968 \times 10^7\ \text{m}^{-1}$$

$$\lambda_{\text{shortest}} = \frac{1}{1.0968 \times 10^7\ \text{m}^{-1}}$$

$$= 0.9117 \times 10^{-7}\ \text{m} \approx 912\ \text{Å}$$

18.4 THE BOHR ATOM

18.4a Bohr's Postulates

To avoid the two problems encountered by the Rutherford model (instability and continuous spectrum) and to explain the spectral data that we have discussed in the previous section, Neils Bohr (1885–1962) proposed a model of the hydrogen atom that can be summarized in three postulates.

Postulate 1. Instead of the infinite number of orbits, with different radii, which are possible in classical mechanics, the electron can take only certain orbits: those for which the angular momentum L takes values given by

$$L = mvr = n\hbar \qquad n = 1,2,3,\cdots \tag{18.10}$$

$L = mvr = n\hbar$

where $\hbar$ is Planck's constant divided by 2π.

Postulate 2. Contrary to the predictions of classical electromagnetic theory, an electron in one of the allowed orbits does not radiate electromagnetic radiation.

Postulate 3. If an electron is initially in an allowed orbit of energy E_i and goes into another orbit of lower energy E_f, electromagnetic radiation will be emitted with a precise frequency given by

$$\nu = \frac{E_i - E_f}{h} \tag{18.11}$$

$\nu = \dfrac{E_i - E_f}{h}$

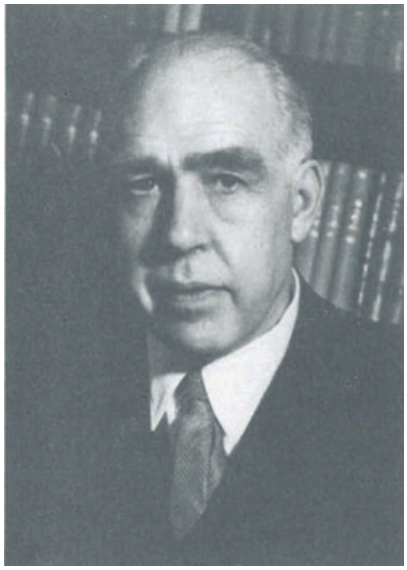
The first postulate uses the quantization idea, that is, the fact proposed by Planck that the energy of a harmonic oscillator takes discrete values. Unlike Planck, Bohr quantizes the orbital angular momentum, not the energy. However, we will see that by quantizing the angular momentum, the energy will also be quantized. The second postulate is needed to prevent the instability predicted by electromagnetic theory. The third postulate is basically a reaffirmation of the photon concept, introduced by Einstein, coupled with the requirement of energy conservation. If light is made of photons of energy $h\nu$, then, when a photon is emitted by the atom, the atom must lose an amount of energy equal to the energy of the photon, that is, $E_i - E_f = h\nu$.

18.4b Energy Spectrum of the Bohr Atom

Using the first postulate, Bohr showed that the energy E of an electron in the hydrogen atom can have only certain values; that is, the energy spectrum is quantized.

From the first postulate, Eq. 18.10

$$v = n\frac{\hbar}{mr}$$



Neils Bohr (1885-1962).

If we substitute this for the velocity v into Eq. 18.2, we have

$$\frac{1}{4\pi\epsilon_0}\frac{e^2}{r^2} = \frac{m}{r}n^2\frac{\hbar^2}{m^2r^2} \tag{18.12}$$

we can solve for the possible radii r_n of the orbits. We obtain

$$r_n = n^2 r_0 \tag{18.13}$$

where

$$r_0 = \frac{4\pi\hbar^2\epsilon_0}{e^2 m} \tag{18.14}$$

We see that only orbits with radius r_0, $4r_0$, $9r_0$, . . ., are allowed.

We can use this result to find the predicted size of the atom. Any system, if left alone, will tend to go to the lowest energy state available to it. In this case, the lowest state corresponds to the orbit of smallest radius, that is, r_0 (see Eq. 18.4). Thus, according to the Bohr model, the normal state of the hydrogen atom is a circular orbit of radius r_0, or (recall that $1/4\pi\epsilon_0 = 9 \times 10^9$ N-m^2/C^2)

$$r_0 = \frac{4\pi\epsilon_0\hbar^2}{e^2 m} = \frac{(1.11 \times 10^{-10}\ \text{C}^2/\text{N-m}^2)(1.05 \times 10^{-34}\ \text{J-sec})^2}{(1.6 \times 10^{-19}\ \text{C})^2\ (9.1 \times 10^{-31}\ \text{kg})}$$

$$= 0.53 \times 10^{-10}\ \text{m} = 0.53\ \text{Å}$$

This result is in agreement with the experimentally known size of the atom.

The restriction on the allowed radius for the orbit leads immediately to the quantization of the energy spectrum. From Eq. 18.4

$$E = -\frac{1}{2}\frac{1}{4\pi\epsilon_0}\frac{e^2}{r} \tag{18.4}$$

Substituting Eqs. 18.13 for r into Eq. 18.4 we get

$$E = -\frac{e^2}{8\pi\epsilon_0 n^2 r_0}$$

Using the value for r_0 from Eq. 18.14 we obtain

$$E_n = -\frac{e^4 m}{8\epsilon_0^2 h^2}\frac{1}{n^2}$$

or

$$E_n = -\frac{E_0}{n^2} \qquad n = 1, 2, 3, \cdots \tag{18.15}$$

where

$$E_0 = \frac{e^4 m}{8\epsilon_0^2 h^2} \tag{18.16}$$

We may evaluate the energy for the smallest orbit, that is, $n = 1$. Note that for this computation, because $\frac{1}{4\pi\epsilon_0} = 9 \times 10^9$ N-m^2/C^2, we may evaluate ϵ_0

$E_n = -\frac{E_0}{n^2}$

$E_0 = \frac{e^4 m}{8\epsilon_0^2 h^2}$

as 8.84×10^{-12} C^2/N-m^2.

$$E_0 = \frac{(1.6 \times 10^{-19}\ \text{C})^4(9.1 \times 10^{-31}\ \text{kg})}{(8)(8.84 \times 10^{-12}\ \text{C}^2/\text{N-m}^2)^2(6.63 \times 10^{-34}\ \text{J-sec})^2}$$

$$= 2.17 \times 10^{-18}\ \text{J} = 13.56\ \text{eV}$$

and from Eq. 18.15

$$E_{n=1} = -13.56\ \text{eV}$$

which is the lowest energy state of the hydrogen atom.

Unlike the classical planetary model where all orbits and corresponding energies are allowed, the Bohr model restricts (quantizes) the spectrum of energies that the atom can have. A few of the energy levels of the hydrogen spectrum are shown in Fig. 18-4. The lowest state available to the atom corresponds to the quantum number $n = 1$; the energy of this level, called the *ground state*, is -13.56 eV, which is the value we have just calculated using Eq. 18.15. The other levels, corresponding to higher values of n, are called the *excited states* of the atom. Normally, the electron is in the ground state. This is due to the tendency of all physical systems to seek the lowest energy state available to them.

To ionize the atom (to separate the electron from the nucleus, that is, to make $r = \infty$), 13.56 eV of energy must be given to it according to the model. This agrees with the experimentally measured ionization energy of hydrogen. Once the atom has been ionized, a free electron near the nucleus can be captured by it; that is, an electron will fall into one of the allowed orbits. This capture can take place in one step—namely, the electron in one jump falls down to the ground state—or in a series of steps in which the electron falls into one or more excited states before it winds up in the ground state. In each

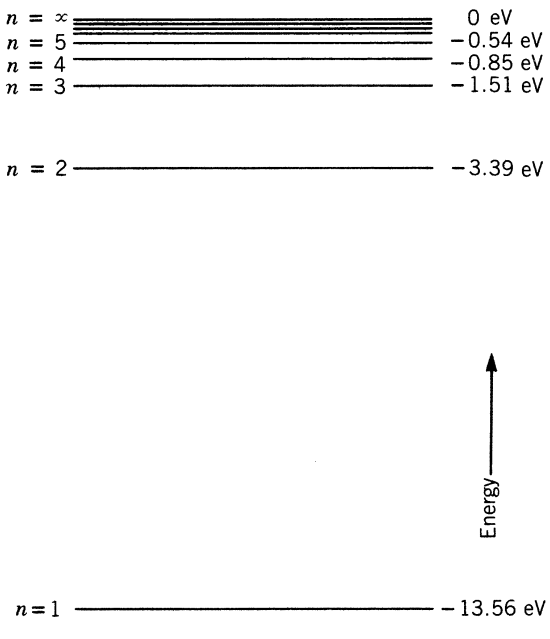


FIGURE 18-4
Some of the discrete energy levels for the electron in the hydrogen atom according to the Bohr model. An infinite number of closely spaced energy levels exists between $n = 5$ and $n = \infty$.

transition toward a lower energy state the electron loses a precise quantity of energy, called a *quantum* (that is, a photon is emitted). The energy, and consequently the frequency of this photon, will depend on the energy difference of the levels between which the transition takes place. It should be clear at this point that, because the energy difference between levels is not continuous, the frequency spectrum of the emitted photons will not be continuous either, but rather it will be discrete. In a typical spectroscopic experiment, such as the one described in Section 18.3, we deal with a very large number of atoms. These atoms are ionized by the electronic discharge and subsequently decay to the ground state in one or several steps. Because the number of atoms involved is large, all possible transitions and corresponding frequencies will be observed. We will now show that the wavelengths (and frequencies) predicted by the Bohr model for the hydrogen atom are the same as those observed experimentally, that is, those given by the Rydberg-Ritz formula of Eq. 18.9.

18.4c Spectral Lines Predicted by the Bohr Model

We have seen that the energy levels of the Bohr atom are given by

$$E_n = -\frac{E_0}{n^2} \tag{18.15}$$

where

$$E_0 = \frac{e^4 m}{8\epsilon_0^2 h^2} \tag{18.16}$$

Suppose that an electron is initially in a state of energy $E_i = -\frac{E_0}{n_i^2}$ and then makes a transition to a state of energy $E_f = -\frac{E_0}{n_f^2}$. According to the third postulate, Eq. 18.11, a photon of frequency

$$\nu = \frac{E_i - E_f}{h}$$

will be emitted (see Fig. 18-5). Substituting for E_i and E_f, we get

$$\nu = \frac{\left\{-\frac{E_0}{n_i^2}\right\} - \left\{-\frac{E_0}{n_f^2}\right\}}{h} = \frac{E_0}{h}\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right) \tag{18.17}$$

If we use the fact that $\lambda\nu = c$ (Eq. 17.8), we can obtain an expression for the wavelengths of the emitted photons that is quite similar to the empirical relation of Eq. 18.9.

$$\frac{1}{\lambda} = \frac{E_0}{hc}\left(\frac{1}{n_f^2} - \frac{1}{n_i^2}\right) \tag{18.18}$$

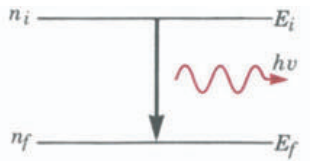


FIGURE 18-5
An electron in the atom decaying from an energy level E_i to a lower energy level E_f emits a photon of energy $h\nu$.

$$\frac{1}{\lambda} = R_{\text{Bohr}} \left(\frac{1}{n_f^2} - \frac{1}{n_i^2} \right) \tag{18.19}$$

where R_{Bohr} is a constant equal to $\dfrac{E_0}{hc}$ and is numerically equal to

$$R_{\text{Bohr}} = \frac{E_0}{hc} = \frac{e^4 m}{8\epsilon_0^2 h^3 c}$$

$$= \frac{(1.6 \times 10^{-19}\ \text{C})^4 \times (9.1 \times 10^{-31}\ \text{kg})}{8 \times (8.84 \times 10^{-12}\ \text{C}^2/\text{N-m}^2)^2 \times (6.63 \times 10^{-34}\ \text{J-sec})^3 \times (3 \times 10^8\ \text{m/sec})}$$

$$= 1.09740 \times 10^7\ \text{m}^{-1}$$

We see that Eq. 18.19 is similar to the Rydberg-Ritz formula (Eq. 18.9) and the constant R_{Bohr} is almost identical to the Rydberg constant R. (The difference between R_{Bohr} and the Rydberg constant is further reduced if the motion of the proton is considered.) All that remains to be shown is that the integers in both equations coincide. If we consider transitions where the final state is in the ground state, then n_f in Eq. 18.19 is equal to 1. Transitions to the ground state can take place when the atom is initially in the first excited state (n_i = 2) or the second excited state (n_i = 3) or the third excited state (n_i = 4), and so on. Thus when n_f = 1, n_i can be 2, 3, 4, ... These are the values taken by the integers n_k and n_j of the Rydberg-Ritz formula for the Lyman series. Similarly, if we consider transitions to the first excited state (n_f = 2), then the initial state can be the second excited state (n_i = 3), the third excited state (n_i = 4), and so on. The Balmer series corresponds to transitions to the first excited state from higher energy states. The other families of wavelengths can be explained with similar arguments. Some of the transitions that account for each family of lines are shown in Fig. 18-6.

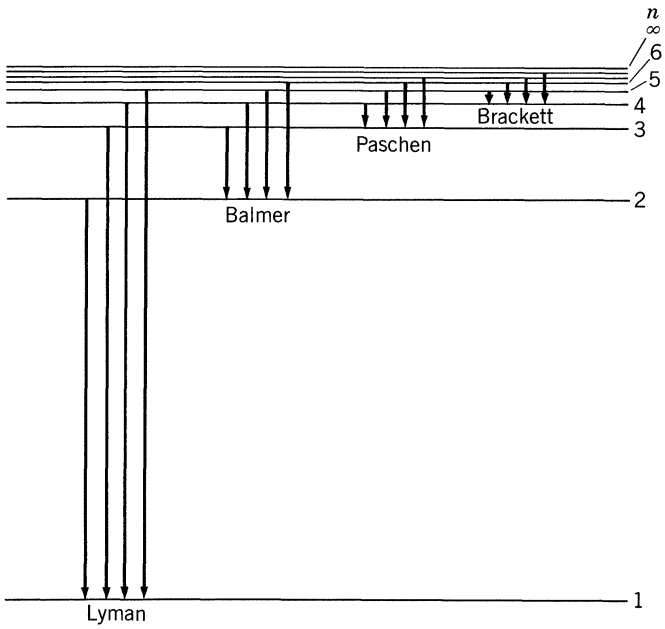


FIGURE 18-6
Transitions responsible for some of the wavelengths in each of the spectral series of the hydrogen spectrum. Additional transitions have been omitted for the sake of clarity in the diagram.

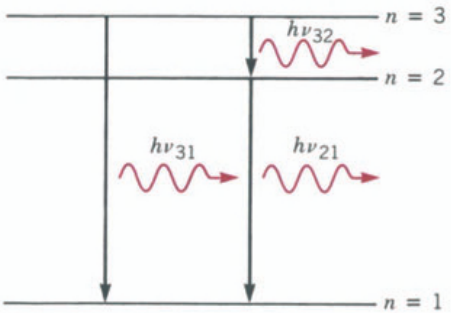


FIGURE 18-7
FIGURE 18-7

Example 18-2

After being excited, the electron of a hydrogen atom eventually falls back to the ground state. This can take place in one jump or in a series of jumps, the electron falling into lower excited states before it ends up in the ground state. Consider a hydrogen atom that has been raised to the second excited state, that is, $n = 3$. Calculate the different photon energies that may be emitted as the atom returns to the ground state.

Solution The possible transitions are shown in Fig. 18-7. Using the results of Eqs. 18.11, 18.15, and 18.16, we can write

$$h\nu_{31} = E_3 - E_1 = E_0\left(\frac{1}{1^2} - \frac{1}{3^2}\right)$$

$$= 13.56 \text{ eV}\left(1 - \frac{1}{9}\right) = 12.05 \text{ eV}$$

$$h\nu_{32} = E_3 - E_2 = E_0\left(\frac{1}{2^2} - \frac{1}{3^2}\right)$$

$$= 13.56 \text{ eV}\left(\frac{1}{4} - \frac{1}{9}\right) = 1.88 \text{ eV}$$

$$h\nu_{21} = E_2 - E_1 = E_0\left(\frac{1}{1^2} - \frac{1}{2^2}\right)$$

$$= 13.56 \text{ eV}\left(\frac{1}{1} - \frac{1}{4}\right) = 10.17 \text{ eV}$$

18.5 THE FRANCK-HERTZ EXPERIMENT

By postulating the existence of a discrete spectrum of energy levels through the angular momentum condition, Eq. 18.10, Bohr was able to predict the correct electromagnetic spectrum for hydrogen. The existence of discrete energy levels in atoms was demonstrated directly by James Franck and Gustav Hertz in 1914.

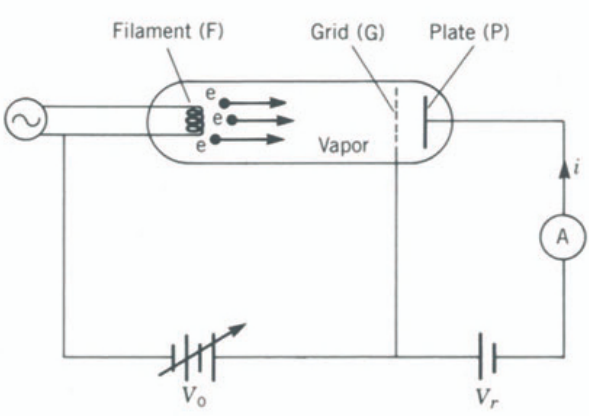


FIGURE 18-8
Schematic of the apparatus used in the Franck-Hertz experiment to show the quantization of the internal energy of atoms.

The experiment that they performed is not very complicated. In fact, today it is a standard experiment in an undergraduate modern physics laboratory. But in spite of its simplicity, the importance can be appreciated by the fact that Franck and Hertz were awarded the Nobel Prize in Physics in 1925.

A schematic of the experimental set-up is shown in Fig. 18-8. The essential part of the apparatus consists of a tube containing vapor of the element under study. The tube contains three electrodes: a filament (F) that provides electrons when heated, a plate (P), and a grid (G). A grid is a charged screen that can attract or repel electrons but, because most of it is open space, the majority of the electrons pass through it. A variable accelerating voltage V_0 is applied between the filament and the grid. As a consequence of this potential difference, the electrons will reach the grid (in the absence of collisions) with a kinetic energy $E_k = eV_0$. After reaching the grid the majority of these electrons will go through the holes in the grid, be collected by the plate P, and contribute to the plate current i, which can be measured by the ammeter A. A small, constant retarding voltage V_r ($\sim$ 1 V) is applied between the plate and the grid. If $V_r > V_0$, the electrons will be turned back before they can reach the plate and they will not contribute to the current measured by A. But even if $V_r < V_0$, the electrons will not be able to reach the plate if they lose enough kinetic energy through collisions with the atoms in the tube as they travel between the filament and the grid.

In the absence of any vapor, that is, a vacuum, the i-V_0 characteristics are those of a typical vacuum tube. This dependence is shown by the dashed line of Fig. 18-9. We should note that a vacuum tube is an example of a nonohmic device, that is, as indicated in Section 15.4, a device in which the current does not vary linearly with voltage; therefore, a straight line behavior between i and V_0 is not to be expected. If vapor of some element is present in the tube, one observes a series of fairly sudden dips superimposed on the monotonic vacuum curve. The solid curve in Fig. 18-9 shows this effect for the case where mercury vapor is present in the tube.

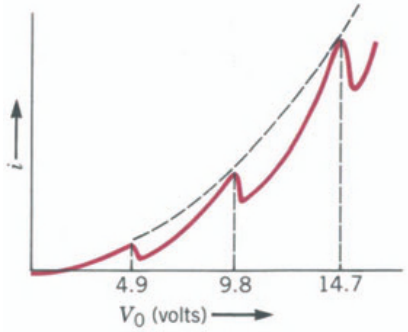


FIGURE 18-9
Dependence of the plate current i (measured by the ammeter A in the apparatus of Fig. 18-8) on the accelerating voltage V_0.

To interpret the results of Fig. 18-9, let us keep in mind the fact that electrons will not contribute to the plate current, even though $V_r < V_0$, if they lose enough kinetic energy through collisions with the atoms in the tube. The

fact that there is no drop in the current until $V_0 = 4.9$ V indicates that the electrons do not lose energy through collisions until they have 4.9 eV of kinetic energy.

Let us digress for a moment and consider what kind of collisions may occur. When the electron collides with a heavy atom such as mercury (Hg), there are two possible kinds of collisions: *elastic* and *inelastic*. In the case of elastic collisions, the total energy of both particles before and after the collision is the same. But when one particle (the Hg atom) is much heavier than the other (the electron), the requirement that the total energy and momentum be conserved leads to the fact that the kinetic energy of the light particle is hardly changed, its velocity is simply reversed. By writing down the equations of conservation of momentum and energy, it was shown (see Eq. 6.16) that if a small particle of mass m and velocity v strikes head-on a large stationary particle of mass M, then the velocity v' of the small particle after the collision is

$$v' = \frac{m - M}{m + M} v \tag{6.16}$$

Thus if $m << M$, then $v' \approx -v$. The kinetic energy of m does not change appreciably. In the case of an elastic collision, the electrons may momentarily be deflected, but because they keep their energy, they will eventually be able to overcome the small retarding voltage V_r and contribute to the plate current i. No drop in the current will be caused by this kind of collision.

In the case of inelastic collisions, the external kinetic energy of the colliding particles becomes internal energy (that is, the particles, in this case the Hg atoms, absorb energy and are excited). When this type of collision occurs, some of the electrons may lose enough energy to be prevented from reaching the plate by the retarding potential V_r. If the Hg atom can have a *continuous distribution of internal energy states*, the transfer of kinetic energy from the bombarding electrons to the Hg atom could and should occur regardless of the energy of the electrons, that is, the drop in the current should occur for any value of V_0. The fact that the drop occurs only when $V_0 = 4.9$ V (and therefore the E_k of the electrons is 4.9 eV) indicates that the first excited state of Hg (the smallest amount of energy that the Hg can absorb) is 4.9 eV above the ground state. As V_0 increases beyond the 4.9 V, the current begins to increase again because, although the electrons can and do collide inelastically and lose 4.9 eV of energy, they still have enough energy remaining to overcome the small retarding voltage V_r. When $V_0 = 2 \times 4.9$ V or 3×4.9 V, or so on, dips in the current occur again because now the electrons can undergo two, three, or more inelastic collisions with the Hg atom; in each collision they lose 4.9 eV.

This interpretation is corroborated by the electromagnetic radiation emitted by Hg atoms. There should be a spectral line whose frequency is given by $h\nu = 4.9$ eV or $\lambda = 2530$ Å. Such a wavelength is found in the spectrum of Hg. An energy diagram for Hg reconstructed from spectral data is shown in Fig. 18-10. The energy difference ΔE between the first excited state and

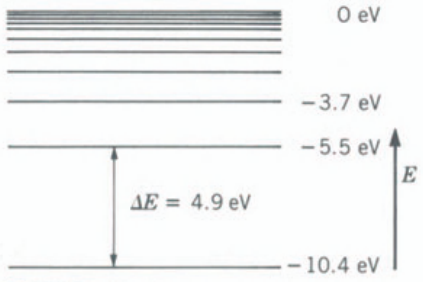


FIGURE 18-10

Some of the atomic energy levels of a mercury atom. The energy spectrum has been reconstructed from the observed wavelengths emitted by mercury atoms. Note that the energy difference between the ground state and the first excited state is 10.4 eV − 5.5 eV = 4.9 eV. Not coincidentally, the dips in the current i in the Franck-Hertz experiment occur for values of V_0 that are integral multiples of 4.9 V (see Fig. 18-9).

the ground state is $\Delta E = 10.4$ eV $- 5.5$ eV $= 4.9$ eV. We may ask why do we not observe a dip when $V_0 = 6.7$ V? This would correspond to a collision where the Hg atom is excited from the ground state to the second excited state. The amount of energy needed to produce this transition is $\Delta E = 10.4$ eV $- 3.7$ eV $= 6.7$ eV. The answer to the question is that such a dip can be observed. However, to observe it, the sensitivity of the experiment must be increased. The reason is that the electrons can and do lose energy as soon as they have 4.9 eV. As a result, the number of electrons that are able to avoid the 4.9 eV collision and gather 6.7 eV is a small fraction of the total number of incident electrons. This reduces the magnitude of the 6.7 eV dip and makes it more difficult to observe.

PROBLEMS

18.1 Calculate the shortest and the longest wavelength of the Balmer series of hydrogen.

18.2 What are (a) the energy, (b) the momentum, and (c) the wavelength of the photon that is emitted when a hydrogen atom undergoes a transition from the state $n = 3$ to $n = 1$? (The momentum of the photon is given by $h\nu/c$).

(*Answer*: (a) 12.07 eV, (b) 6.44×10^{-27} kg-m/sec, (c) 1030 Å.)

18.3 The shortest wavelength of the Paschen series from hydrogen is 8204 Å. From this fact, calculate the Rydberg constant.

18.4 When an atom emits a photon it must recoil with a momentum equal but in opposite direction to that of the photon. (a) Show that the recoil energy of an atom when a transition from one energy state to another takes place is given by

$$E_r = \frac{(h\nu)^2}{2\,mc^2}$$

where m is the mass of the atom and $h\nu$ is the energy of the photon. (b) How does E_r compare with $h\nu$ in the case of a hydrogen atom?

(*Answer*: (b) $E_r \sim 10^{-8}\, h\nu$.)

18.5 A hydrogen atom is excited from a state $n = 1$ to a state $n = 4$. (a) Calculate and display on an energy-level diagram the different photon energies that may be emitted as the atom returns to the $n = 1$ state. (b) Calculate the recoil speed of the hydrogen atom, assumed initially at rest, if it makes the transition from $n = 4$ to $n = 1$ in a single quantum jump.

18.6 Calculate the ionization energy of the hydrogen atom from the following data. The wavelength of the Balmer series limit ($n = \infty$ to $n = 2$) is 3645 Å and the wavelength of the first line of the Lyman series ($n = 2$ to $n = 1$) is 1215 Å.

18.7 A hydrogen atom initially at rest in its ground state absorbs a 100 eV photon and the ejected photoelectron moves in the same direction as the initial direction of the photon. What is the recoil velocity of the proton?

18.8 Monochromatic light is shone on a hydrogen gas with all its atoms in the ground state. As a consequence, the gas emits radiation containing six different frequencies. What is the minimum frequency of the incident light?

(*Answer*: 3.07×10^{15} Hz.)

18.9 The average lifetime of an electron in an excited state of an atom is 10^{-8} sec. How many revolutions does an electron make in the $n = 2$ state of hydrogen before dropping down to the $n = 1$ state?

(*Answer*: 8.3×10^6 rev.)

18.10 Compare the frequency of revolution in the hydrogen atom with the frequency of the emitted photon for a transition in which $\Delta n = 1$ when the initial state is (a) $n = 10$, (b) $n = 100$, (c) $n = 1000$.

18.11 A particle called the mu meson is emitted when a nucleus is struck by a high-energy particle from an accelerator. The mu meson has the same charge as an electron, but its mass is 207 times larger. Sometimes this mu meson is captured by a proton to form muonic hydrogen. What will be (a) the radius of the first Bohr orbit of the mu meson? (b) the ground state energy, and (c) the wavelength emitted in the transition $n = 2$ to $n = 1$? Neglect the motion of the proton.

(*Answer*: (a) 2.6×10^{-3} Å, (b) -2.81×10^3 eV, (c) 5.87 Å.)

18.12 Hydrogen gas is diatomic, that is, H_2. But suppose that we have monoatomic hydrogen. (a) What energy is needed to raise the electron from its ground state $n = 1$ to the first excited state $n = 2$? (b) At what temperature will the average kinetic energy of a monoatomic hydrogen gas be equal to that energy?

(*Answer*: (a) 10.19 eV, (b) 7.87×10^4 K.)

18.13 Consider transitions in hydrogen atoms in which $\Delta n = 1$. (a) Show that for very large n the energy change, $\Delta E \approx \alpha^2(mc^2/n^3)$, where m is the mass of the electron, c is the velocity of light, and $\alpha = 2\pi\left(\frac{1}{4\pi\epsilon_0}\frac{e^2}{hc}\right)$. (b) α is named the *fine structure* constant; show that it has a numerical value of 1/137 and that it is dimensionless.

18.14 Suppose hydrogen atoms are placed in a medium in which the dielectric constant $\kappa = 6$. (a) Calculate the radius of the first orbit and the ionization energy. (b) Suppose $\kappa = 12$; calculate the radius and ionization energy. (*Hint:* When two charges are placed in a medium of dielectric constant κ, the constant $1/4\pi\epsilon_0$ in Coulomb's force law is replaced by $1/4\pi\kappa\epsilon_0$).

(*Answer*: (a) 3.18 Å, 0.38 eV, (b) 6.36 Å, 0.094 eV.)

18.15 The emission spectrum of a given system consists of six different frequencies: 1.5×10^{15} Hz, 2.7×10^{15} Hz, 3.5×10^{15} Hz, 4.2×10^{15} Hz, 6.2×10^{15} Hz, and 7.7×10^{15} Hz. Choose the ground-state energy to be $E = 0$. (a) What are the energies of the excited states of the system? (b) Show which transitions give rise to each of the six frequencies. (Assume that transitions between all available energy levels are possible.)

18.16 Two hydrogen atoms having the same energy undergo an inelastic head-on collision. After the collision, the atoms come essentially to rest and two photons are emitted. The wavelengths of the emitted photons are 3667 Å and 1220 Å. What was the velocity of the hydrogen atoms? Note that momentum will be conserved if the hydrogen atoms keep a very small fraction of the kinetic energy (see problems 18.4 and 18.5).

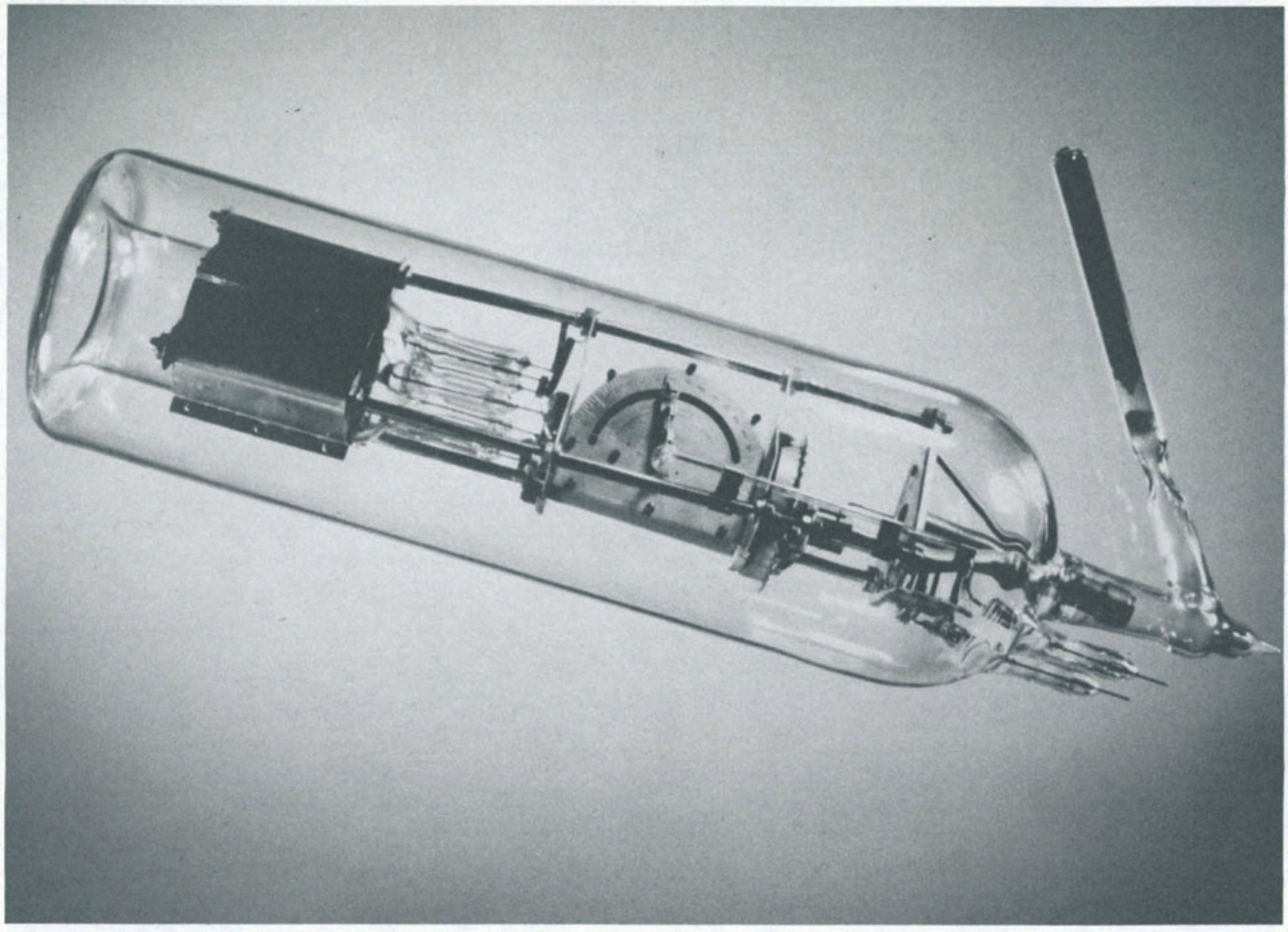
(*Answer*: 3.6×10^4 m/sec.)

18.17 Assume that the orbit of the moon around the earth is a circular orbit of radius 3.8×10^8 m. The mass of the moon is 6×10^{22} kg, and its orbital velocity is 10^3 m/sec. (a) If the first Bohr postulate applies to the motion of the moon around the earth, find the quantum number n of the orbit. (b) Use Newton's law of gravitation and Bohr's postulates to show that the allowed orbital radius of the moon is $r_n = n^2 r_0$ where r_0 is a constant. (c) What is a fractional difference between the assumed radius of 3.8×10^8 m and the next allowed orbital radius?

(*Answer*: (a) 2.2×10^{68}, (c) $\sim 10^{-68}$.)

18.18 Find the smallest electric potential difference through which an electron must be accelerated to enable it to excite a hydrogen atom out of its ground state.

18.19 The ground-state and the first excited-state energies of potassium atoms are -4.3 eV and -2.7 eV, respectively. If we use potassium vapor in the Franck-Hertz experiment, at what voltages would we see drops in the plot of current versus voltage?



Photograph of one of the original tubes used by Davisson and Germer in their electron diffraction experiments.

CHAPTER 19

Fundamental Principles of Quantum Mechanics

19.1 INTRODUCTION

In this chapter we will present two principles that form the cornerstones of quantum mechanics: *de Broglie's hypothesis* and the *uncertainty principle*.

In Chapter 17 we saw that experiments such as the photoelectric effect and the Compton effect could be explained by looking at electromagnetic radiation as being made up of particle-like entities called photons. On the other hand, the experiments of Young and others, demonstrating the interference effects of light, are as valid today as when they were originally performed; these interference and diffraction experiments require a wave model. We must therefore consider electromagnetic radiation as having dual properties: It is both a wave and a particle. Although the wave and particle properties seem to be irreconcilable, they are not. These two models are, in the words of Neils Bohr, complementary. The *principle of complementarity* is discussed at the end of this chapter.

Louis de Broglie (1892-).

19.2 DE BROGLIE'S HYPOTHESIS AND ITS EXPERIMENTAL VERIFICATION

In 1925, Louis de Broglie (1892–) assumed the existence of a natural symmetry in nature and proposed that the dual character exhibited by photons should equally apply to all material particles. Accordingly, he hypothesized: *The motion of a particle is governed by the wave propagation properties of a "pilot" wave (also called matter wave). The wavelength λ and the frequency ν of the pilot wave associated with a particle of momentum p and energy E are*

$$\lambda = \frac{h}{p} \quad \text{and} \quad \nu = \frac{E}{h} \tag{19.1}$$

To prove de Broglie's hypothesis we have to show experimentally that a beam of particles exhibits wave-like properties, such as interference and diffraction.

Let us look at the feasibility of such experiments. Consider a bullet of mass $m = 0.1$ kg moving with a velocity $v = 10^3$ m/sec. According to de Broglie, the wavelength of the pilot wave will be,

$$\lambda = \frac{h}{p} = \frac{6.63 \times 10^{-34} \text{ J-sec}}{0.1 \text{ kg} \times 10^3 \text{ m/sec}} = 6.63 \times 10^{-36} \text{ m}$$

which is 26 orders of magnitude smaller than the diameter of an atom. As we saw in Chapter 12, in order to observe interference effects, the size of the diffracting slit d must be comparable to the wavelength of the wave. It is clear, therefore, that we cannot hope to see a diffraction pattern when a beam of macroscopic objects goes through a hole. The smallest slit or diffraction grating available in nature is a crystalline solid where d is a few angstroms. If we are going to have any success in our attempts to observe the wave

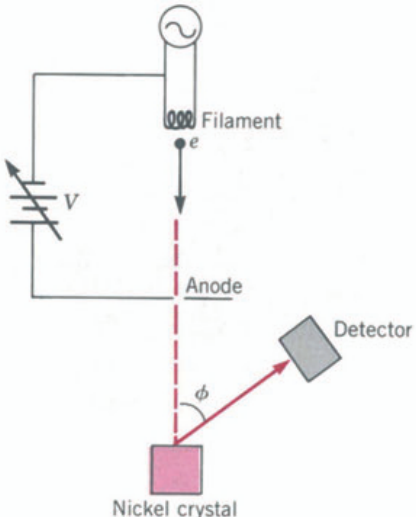


FIGURE 19-1
Schematic of the apparatus used in the Davisson-Germer experiment. Electrons from a heated filament are accelerated through a variable potential difference V. They strike a nickel crystal, and the number of electrons scattered at a given angle ϕ is measured as a function of the accelerating voltage. The experiment can be repeated for different values of the angle ϕ.

nature of material particles, we must have a beam with a wavelength not much smaller than a few angstroms. From Eq. 19.1 this means that the momentum of the particles has to be relatively small, and the most practical way of achieving this condition is to choose a particle of small mass such as the electron.

$$\lambda = \frac{h}{p}$$

$$\nu = \frac{E}{h}$$

In 1927, C. J. Davisson and L. H. Germer performed such an experiment. The experimental setup used is shown in Fig. 19-1. Although not shown, the electron path must be through a vacuum. Electrons from a heated filament are accelerated by a variable voltage V. These electrons emerge through a small hole in the anode with a kinetic energy $E_k = eV$. The beam then strikes a single crystal of nickel, and the intensity of the scattered beam can be measured for different angles ϕ and various voltages V. Although these investigators used a nickel crystal, almost any crystalline solid will produce the same results.

If the propagation of the beam is particle-like, we may expect that the intensity I of the scattered beam will have a smooth monotonic dependence of both ϕ and V because only elastic collisions with the atoms of the crystal are involved. Thus, if we consider the plane of atoms in the crystal to act like a wall, the incident particles will carom off so that the incoming angle (angle of incidence) is equal to the outgoing angle (angle of reflection) (see Fig. 19-2*a*). Then for any incoming angle we will observe the same number of particles caroming off at a similar angle regardless of their velocity and thus of the potential V through which they have been accelerated. On the other hand, if the incoming electrons are not particles but are actually waves, we would expect a diffraction effect like the one observed with X rays when the Bragg scattering condition discussed in Chapter 12 is satisfied, that is, when

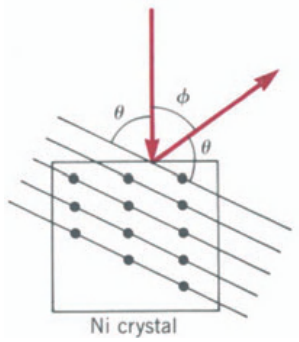
$$n\lambda = 2d \sin \theta \tag{12.11}$$

From X-ray studies the physicists knew that the spacing between one set of atomic planes of nickel is 0.91 Å. For X rays of wavelength 1.65 Å the angle of incidence to the plane of atoms that satisfies Eq. 12.11 for first-order diffraction ($n = 1$) is

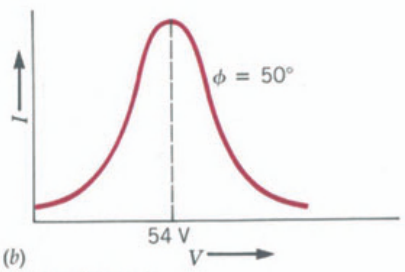
$$\sin \theta = \frac{1.65 \text{ Å}}{2 \times 0.91 \text{ Å}} = 0.907$$

$$\theta = 65°$$

It is seen in Fig. 19-2*a* that the angle ϕ between the two beams is defined as $2\theta + \phi = 180°$. The Bragg angle θ is 65° and therefore $\phi = 50°$. This was the angle selected for the experimental arrangement in Fig. 19-1. Now de Broglie's hypothesis, Eq. 19.1, states that the wavelength of a particle is a function of its momentum p. The momentum is related to the kinetic energy of the particle and the kinetic energy, as shown in section 14.4, depends on its accelerating voltage through the relation $1/2\ mv^2 = eV$. The experimental results of Davisson and Germer are shown in Fig. 19-2*b*. The angle ϕ between the incoming and the scattered beams of electrons was set at 50°. The accel-



(*a*)



(*b*)

FIGURE 19-2
(*a*) Electrons being scattered by the atomic planes of the nickel crystal. (*b*) Intensity of the scattered beam, that is, number of scattered electrons, as a function of the accelerating voltage V for a fixed value of the angle $\phi = 50°$.

erating voltage was increased, and the intensity I of the reflected beam was measured for different values of this voltage. A maximum was observed when the voltage was $V = 54$ V. Davisson and Germer then calculated the wavelength that the electrons would have from the de Broglie hypothesis, Eq. 19.1. The momentum p can be obtained from the kinetic energy E_k. By definition

$$E_k = \frac{1}{2} mv^2$$

$$= \frac{1}{2} \frac{m^2v^2}{m}$$

$$= \frac{p^2}{2\,m}$$

Therefore

$$p = \sqrt{2m\,E_k}$$

but $E_k = eV$, and it follows that

$$p = \sqrt{2\,meV}$$

Substituting this relation for the momentum into the de Broglie equation, Eq. 19.1,

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2\,meV}}$$

they obtained a wavelength for the electrons of

$$\lambda = \frac{6.63 \times 10^{-34}\ \text{J-sec}}{(2 \times 9.1 \times 10^{-31}\ \text{kg} \times 1.6 \times 10^{-19}\ \text{C} \times 54\ \text{V})^{1/2}}$$

$$= 1.67 \times 10^{-10}\ \text{m} = 1.67\ \text{Å}$$

The agreement between the two wavelengths—that is, the predicted wavelength for electrons of this energy and the wavelength of 1.65 Å for X rays to be diffracted from nickel—is excellent. Thus the experiment confirms de Broglie's hypothesis that particles have wave properties and that the wavelength of the wave is given by Eq. 19.1.

Since the original experiment by Davisson and Germer, other people have found convincing evidence for the wave properties of material particles. Estermann, Stern, and Frisch have shown inteference effects with beams of hydrogen and helium atoms. Fermi, Marshall, and Zinn have done similar experiments with neutrons. Today, the existence of matter waves rests on solid experimental foundations, and electron and neutron diffraction measurements are standard techniques for the study of crystalline materials. The design and construction of the electron microscope are based on the principle that electrons propagate as waves.

Example 19-1

A beam of monochromatic neutrons is incident on a KCl crystal with lattice spacing of 3.14 Å. The first-order diffraction maximum is observed when the angle θ between the incident beam and the atomic planes is 37°. What is the kinetic energy of the neutrons?

Solution Using the Bragg condition (Eq. 12.11), with $n = 1$, we can find the wavelength of the neutron beam.

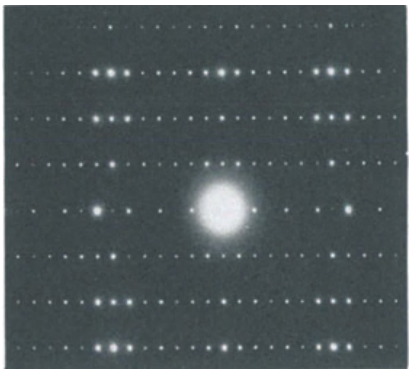
$$\lambda = 2d \sin \theta$$
$$= 2 \times 3.14 \text{ Å} \sin 37° = 3.78 \text{ Å}$$

From de Broglie's hypothesis (Eq. 19.1), the momentum of the neutrons is

$$p = \frac{h}{\lambda} = \frac{6.63 \times 10^{-34} \text{ J-sec}}{3.78 \times 10^{-10} \text{ m}} = 1.75 \times 10^{-24} \text{ kg-m/sec}$$

The kinetic energy of the neutrons will be

$$E_k = \frac{1}{2} mv^2 = \frac{p^2}{2m}$$
$$= \frac{(1.75 \times 10^{-24} \text{ kg-m/sec})^2}{2 \times 1.67 \times 10^{-27} \text{ kg}} = 9.21 \times 10^{-22} \text{ J}$$
$$= 5.75 \times 10^{-3} \text{ eV}$$



Electron diffraction pattern of a $Ti_2Nd_{10}O_{29}$ crystal. As in the X-ray diffraction on page 172, each bright dot is produced by a set of atomic planes that satisfies the Bragg condition for the de Broglie wavelength of the electrons.

19.3 NATURE OF THE WAVE

If we accept the hypothesis that a particle is guided by a wave as it moves through space, what is the nature of this wave? What is waving? To understand the accepted interpretation of the nature of the wave, let us first consider the case of electromagnetic radiation.

Classically, electromagnetic (EM) radiation is an electric and a magnetic field whose behavior in time and space is comparable to that of a wave in a string (see Section 16.8). We know, however, that this model of EM radiation cannot explain the photoelectric effect or the Compton effect. To explain these effects, we need a particle model. The fact remains that when a beam of photons passes through a double slit, an interference pattern is observed; this is a unique wave property. The question is therefore: *Within the photon picture, what is the nature of the wave that governs the propagation of photons?*

The link between the classical (wave) picture and the photon (particle) picture is the intensity I. The intensity has meaning in both models; it is simply the energy per unit time, per unit area of the beam (see Section 11.4). The difference is the way the energy is carried by the beam: localized (in the particle model) versus continuous (in the wave model). We saw in Chapter 11, Eq. 11.20, that the intensity of a wave is proportional to the square of the

amplitude of the wave, that is,

$$I \propto (\text{amplitude of the wave})^2 \qquad \text{(Wave model)}$$

On the other hand, it is relatively easy to see that, in the photon model, the intensity is given by

$$I = (c)(N)(h\nu) \qquad \text{(Particle model)}$$

where N is the number of photons per unit volume of the beam, $h\nu$ the energy of each photon, and c is the velocity of the beam, which is the velocity of light.

Classically, one can increase the intensity of the EM radiation by increasing the amplitude of the wave. In the photon scheme, the intensity of a monochromatic beam is increased by increasing the number of photons per unit volume, N. This comparison tells us that the number of photons in the beam must be proportional to the square of the amplitude of the wave.

$$N \propto (\text{amplitude of the wave})^2 \qquad (19.2)$$

When we talk about the density of photons, we should not think of N as an exact number in the sense that one finds exactly the same number of photons in a given volume of the beam. There are fluctuations. We do not notice these fluctuations with ordinary intensities, such as the intensity from a nearby lamp, because the number of photons is very large. But if one were to shine light on a screen with an intensity corresponding to 100 photons per second per meter^2, one would make the following observations:

1. The 100 photons per second is an average value: one second, we may have 105 photons landing on the screen, another 92, yet another 97, and so forth.
2. One cannot predict where the next photon striking the screen will land. We will find that if we wait long enough, the same number of photons will land on one particular section of the screen as on any other.

The reason for this randomness in the behavior of photons is the process of photon emission itself. It is a statistical process in which a large number of atoms, after being excited or ionized, eventually decay (with a half-life that is typically 10^{-8} sec) to the ground state while emitting photons. When we say that the half-life of the excited state is 10^{-8} sec, we mean that half of the excited atoms will decay in 10^{-8} sec. There is no way of predicting whether a particular atom will decay before or after 10^{-8} sec.

As a result of this randomness we cannot state with certainty whether or not we will find a photon within a small volume of the beam: We can nevertheless talk about the probability of finding a photon within such a volume. Suppose that $N = 1$ photon/m^3. Because this photon can be anywhere in the m^3, the probability of finding the photon within a $\Delta V = 1$ cm^3 will be 1 in 10^6 ($1\ \text{m}^3 = 10^6\ \text{cm}^3$). If N is 10 photons/m^3, the probability of finding a photon within the same ΔV will be 10 times greater. We see that the probability of finding a photon, the probability density, within a given volume of the

Max Born (1882-1970).

beam is proportional to the density of the photons N. But from Eq. 19.2, $N \propto$ (amplitude of the wave)^2; we therefore conclude, as did Einstein, that the probability of finding a photon within a given volume of the beam is proportional to the square of the amplitude of the wave associated with the beam. *Within the photon model the wave is not an electromagnetic field; it is represented by a mathematical function that measures the photon probability density.*

In 1926, Max Born (1882–1970) extended this probabilistic interpretation to the matter waves proposed by de Broglie for material particles. According to this interpretation, accepted by most scientists today, "what is waving" is nothing physical like the particles in a string or the water molecules in a pond when the ripples go by: The guiding wave is represented by a mathematical function, $\psi(r,t)$, called a *wavefunction*. The physical significance of the wavefunction is the following: *At some instant t, a measurement is made to locate the particle associated with the wavefunction ψ. The probability $P(r,t)\,dV$ that the particle will be found within a small volume dV centered around a point with position vector r (with respect to a prechosen set of coordinates) is equal to $|\psi|^2\,dV$,* that is,

$$P(r,t)\,dV = |\psi|^2\,dV \tag{19.3}$$

$P(r,t)\,dV = |\psi|^2\,dV$

We should note that the probability of finding the particle somewhere in space must be unity. Therefore, the wavefunction ψ must be *normalized;* that is, it must satisfy the condition

$$\int_{-\infty}^{\infty} |\psi|^2\,dV = 1 \tag{19.4}$$

$\int_{-\infty}^{\infty} |\psi|^2\,dV = 1$

where the integration is carried over all space.

The methods of quantum mechanics consist in first finding the wavefunction associated with a particle or a system of particles. As we will see in Chapter 20, all the information about the physical properties of the system can be obtained from it.

It is interesting to note that among the critics of this probabilistic interpretation of matter waves we find Einstein ("God does not play dice with the Universe") and de Broglie. The reason for the opposition is that the interpretation introduced by Born goes against one of the fundamental philosophical tenets of classical physics: the concept of "determinism." This brings us to one of the fundamental principles of quantum mechanics: the *uncertainty principle,* first enunciated in 1927 by Werner Heisenberg (1901–1976).

19.4 THE UNCERTAINTY PRINCIPLE

The use of probability functions to describe the behavior of a system is not foreign to classical physics. In statistical mechanics, where one often deals with a very large number of particles (for example, the study of a gas), we do not deal with the individual velocities, energies, or such, of the particles. We talk about the average value of these quantities and we deal with velocity and energy distribution functions, that is, the number or fraction of a group

at each velocity or energy (Chapter 9). In classical physics this recourse to probability functions is the result of practical necessity. If we want to study the behavior of 10^{23} molecules in a gas we could, in principle, write down and solve 10^{23} coupled differential equations for the coordinates of each molecule. But this would be beyond the most advanced computer and would not yield any useful information. In quantum mechanics, the use of probability functions is one of fundamental necessity: We have no choice. Even for the simplest of systems, the one particle system, we cannot determine its position as a function of time, $x(t)$.

Werner Heisenberg (1901-1976).

What is the root of this basic disagreement? According to classical physics, if you know the forces acting on a particle and you know the initial conditions (that is, the position x_0 and the velocity v_0 at some initial time $t = 0$), then by solving Newton's laws of motion, the position $x(t)$ of the particle at any other time can be predicted or determined (hence "determinism"). The assumption here is that the initial position and velocity can be determined exactly. Classical physics assumes that given the proper measuring instrument, this can be done. The answer of quantum mechanics is different and was presented in the famous uncertainty principle by Werner Heisenberg: *An experiment cannot simultaneously determine a component of the momentum of a particle, for example p_x and the exact value of the corresponding coordinate x.* The best one can do is

$$\Delta p_x \, \Delta x \geq \hbar \tag{19.5}$$

$\Delta p_x \, \Delta x \geq \hbar$

where Δ represents the uncertainty of measurement.

There are three points that we should note about the uncertainty principle.

1. The limitations imposed by the uncertainty principle have nothing to do with the quality of the experimental instrument. With the ideal instrument, the best one can do is given by Eq. 19.5.
2. The uncertainty principle does not say that one cannot determine the position or the momentum exactly. However, if $\Delta x = 0$, then the uncertainty in the momentum will be infinite, and vice versa.
3. Later in this chapter we will show that the uncertainty principle is a direct consequence of de Broglie's hypothesis, which, as we have seen, is confirmed by experiment. Thus the uncertainty principle is based on experiment.

In classical physics we assume that the position and the velocity of a particle can be determined with any desired degree of accuracy. Because of the smallness of Planck's constant, this assumption is a very good approximation when dealing with macroscopic objects; the approximation breaks down in what Born called the "restless universe" of the atom. Let us see what the implications of the uncertainty principle are when we try to predict the trajectory (the future) of the moon around the earth. The mass of the moon is 6×10^{22} kg, and its average orbital velocity is 10^3 m/sec. Suppose that we are able to determine the position of the moon with an uncertainty $\Delta x =$

10^{-6} m, a very small uncertainty in this case. What is the limit of our ability to determine its velocity simultaneously? From Eq. 19.5

$$\Delta p_x \geq \frac{\hbar}{\Delta x} = \frac{10^{-34}\ \text{J-sec}}{10^{-6}\ \text{m}} = 10^{-28}\ \text{kg m/sec}$$

Because by definition $p_x = mv_x$, it follows that

$$\Delta v_x = \frac{\Delta p_x}{m} \geq \frac{10^{-28}\ \text{kg m/sec}}{6 \times 10^{22}\ \text{kg}} \approx 10^{-50}\ \text{m/sec}$$

This is an insignificant error when we compare it with the measured value of $v = 10^3$ m/sec.

Now, let us consider an electron in the hydrogen atom. We saw in Chapter 18 that the smallest radius of a Bohr orbit is approximately 0.5×10^{-10} m. Suppose that, in the hope of being able to determine p_x as accurately as possible, we decide to give up all knowledge about the coordinates of the electron, that is, let $\Delta x \approx 10^{-10}$ m. (The electron can be anywhere in the orbit, and therefore x can take any value between 0.5×10^{-10} m and -0.5×10^{-10} m.) With what accuracy can the x component of p be determined? From Eq. 19.5

$$\Delta p_x \geq \frac{\hbar}{\Delta x} = \frac{10^{-34}\ \text{J-sec}}{10^{-10}\ \text{m}} = 10^{-24}\ \text{kg m/sec}$$

To understand the significance of this uncertainty, let us compare it with the magnitude of the momentum of the electron. In Chapter 18 we showed that the kinetic energy of the electron was equal to the magnitude of the total energy (see Eqs. 18.3 and 18.4). The latter is of course negative and equal to -13.6 eV; hence $E_k = 13.6$ eV $= 13.6\ \text{eV} \times 1.6 \times 10^{19}\ \text{J/eV} = 2.18 \times 10^{-18}$ J. The corresponding momentum is

$$p = (2m\,E_k)^{1/2} = (2 \times 9.1 \times 10^{-31}\ \text{kg} \times 2.18 \times 10^{-18}\ \text{J})^{1/2}$$
$$= 2 \times 10^{-24}\ \text{kg m/sec}$$

Then the ratio of the uncertainty of the momentum Δp_x to the momentum itself is

$$\frac{\Delta p_x}{p} = \frac{10^{-24}\ \text{kg m/sec}}{2 \times 10^{-24}\ \text{kg m/sec}} = 0.5 \quad \text{or 50\% uncertainty}$$

This simple example shows that the classical concept of determinism cannot be carried over to atomic phenomena. Note that if we try to specify the location of the electron more precisely than the total orbit, that is, make Δx smaller, the uncertainty in the momentum becomes even greater than 50%. With such a limitation on simultaneous knowledge of position and momentum, we must develop and use the concept of the probability of location and momentum.

19.5 PHYSICAL ORIGIN OF THE UNCERTAINTY PRINCIPLE

In the next section we will show that the uncertainty principle is a direct consequence of de Broglie's hypothesis. Here we present a "gedanken" (thought) experiment proposed by Bohr, which illustrates the physical origin of the principle, namely, the measuring process itself introduces the uncertainty. We can determine very accurately the position of a planet without significantly altering its momentum. The same is not true in the submicroscopic world of atoms and molecules.

Suppose that we want to determine the position of an electron and that we can look at the electron with a hypothetical microscope. To see it, we must shine light on the electron; it is the scattered photons that the eye really sees. Any photon scattered by the electron within an angle equal to 2ϕ (see Fig. 19-3) will be focused by the microscope lenses and will be detected by the eye. It is clear that the collision of the photon with the electron will change the momentum of the electron. We can minimize this disturbance and the resulting uncertainty in the momentum of the electron by reducing the intensity of the light. The smallest intensity we can use, according to the quantum mechanical model of light, is one photon. This photon may enter the objective lens anywhere within the angular range $+\phi$ to $-\phi$. This implies that the x-component of the momentum of this photon could have any value between $-p \sin \phi$ and $p \sin \phi$ (where p is the magnitude of the momentum of the photon). We conclude, therefore, that the uncertainty of the x-component of the momentum of the photon is

$$\Delta p_x \text{ (photon)} = 2p_{\text{photon}} \sin \phi$$

Conservation of linear momentum necessitates that if the photon acquires a certain momentum in the x direction, the electron must acquire the same amount in the opposite direction. This means that the uncertainty in the x-momentum of the electron also has the same magnitude, that is,

$$\Delta p_x \text{ (electron)} = 2p_{\text{photon}} \sin \phi$$

Because the momentum of the photon is given by $p = h/\lambda$ (see Eq. 17.18) we get

$$\Delta p_x \text{ (electron)} = 2\frac{h}{\lambda} \sin \phi \qquad (19.6)$$

Thus, we see that in the process of locating the electron, we have introduced an uncertainty in its momentum.

We could reduce the uncertainty in p_x of the electron in two ways. We can use photons of longer wavelength: Instead of using visible light, we could use microwaves or radio waves and a detector that is sensitive to them. Simultaneously, we could reduce ϕ, the angle subtended by the lens: We can

Standing (from left): William Osler, Neils Bohr, James Frank, Oscar Klein. Seated: Max Born.

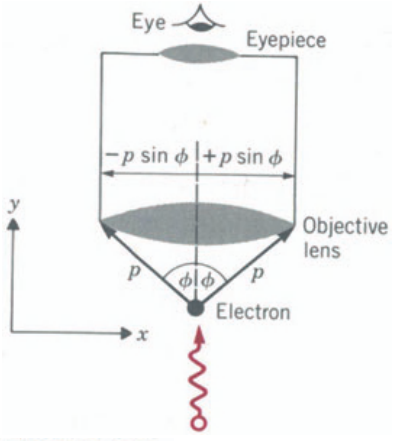


FIGURE 19-3
"Looking" at an electron with a hypothetical microscope in Bohr's gedanken experiment. The electron is illuminated with light (photons) The photons scattered by the electron that enter the objective lens of the microscope are detected by the eye of the observer.

make the aperture of the lens small. Unfortunately, these two factors that would reduce the uncertainty in p_x lead to an increase in the uncertainty of the position of the electron that we are trying to locate. In Chapter 12, Section 12.6, we showed that owing to diffraction effects, our ability to resolve two separate sources of light decreases as the wavelength of the light emitted by the sources increases and as the size of the opening used to "look" at the sources decreases, Eq. 12.10. Thus, in the experiment under consideration, longer wavelengths and smaller lens aperture will increase the uncertainty about the origin of the photon entering the microscope, that is, about the position of the scattering electron.

19.6 MATTER WAVES AND THE UNCERTAINTY PRINCIPLE

We have seen that the spacial propagation of a particle is governed by a wave $\psi(x,t)$. Let us try to find the specific mathematical form of the wavefunction $\psi(x,t)$. We can start by looking at a very common type of wave: the sinusoidal traveling wave that we discussed in Chapter 11, Eq. 11.5′

$$\psi(x,t) = A \sin (kx - \omega t) \tag{11.5′}$$

We recall that this particular wave has the following properties:

1. The amplitude A is the same at all points in space.
2. It has a well-defined wavelength, $\lambda = 2\pi/k$ (Eq. 11.12).
3. It has a well-defined frequency, $\nu = \omega/2\pi$ (Eq. 11.14).
4. It travels toward increasing values of x with a velocity

$$v = \lambda\nu = \frac{\omega}{k} \tag{11.2}$$

Can we associate this wave with a free particle? If the particle has a well-defined momentum and energy, the wave of Eq. 11.5′ satisfies the requirements set forth by de Broglie, namely, a well-defined λ and ν. We should note, however, that because the amplitude of the wave is the same for all values of x, the particle can be found with equal probability at any point in space, that is, the particle is completely unlocalized, $\Delta x = \infty$. Actually, we should not be surprised by this result. A well-defined λ implies a well-defined momentum for the particle, that is, $\Delta p_x = 0$. This in turn, according to the uncertainty principle, should lead to an infinite uncertainty in the coordinate of the particle. Thus, although the wavefunction of Eq. 11.5′ can be used to describe a particle of well-defined momentum and energy, the price paid is a complete lack of localization.

Before we consider the type of wave to be used to describe a partially localized particle, we will show that the velocity of the wave given by Eq.

11.5′ is not the same as the velocity of the particle that it guides. The velocity of this wave is given by Eq. 11.2

$$v = \lambda\nu$$

De Broglie's relations are $\lambda = h/p$ and $\nu = E/h$, where p is the momentum of the particle and is equal to mv_{particle}, and E is the energy of the particle $\frac{1}{2}\, mv^2_{\text{particle}}$. Substituting for ν and λ into the expression for v, we obtain

$$v = \frac{h}{p} \times \frac{E}{h} = \frac{E}{p} = \frac{\frac{1}{2}\, mv^2_{\text{particle}}}{mv_{\text{particle}}} = \frac{1}{2}\, v_{\text{particle}}$$

The velocity of the wave is half the velocity of the particle. In this case, because the amplitude of the wave is the same everywhere in space, it is not a serious problem. However, this condition would not be acceptable in the case of a partially localized particle, which we discuss next.

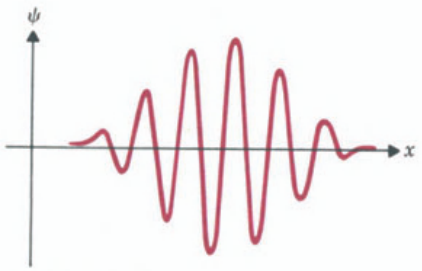


FIGURE 19-4
A wave packet may be used to describe a partially localized particle. The particle is located in the region of space where the amplitude of the wave packet is not zero.

To describe a particle that is partially localized, we need a wave with an amplitude that is different from zero only over a small region of space where there is a chance of finding the particle. We need a wave that looks like the *wave packet* shown in Fig. 19-4. The more localized the particle is, the narrower the wave packet must be. Mathematically, such a wave packet is obtained by mixing (adding) together an infinitely large number of sinusoidal traveling waves of the type described by Eq. 11.5′. Each of these waves differs in wavelength and frequency by an infinitesimally small amount from the previous one. The amount of a given wave included in the mixing is determined by the amplitude A of that particular wave, that is, the amplitude becomes a function of the wavelength and of the frequency, $A(\lambda,\nu)$. Because the sum is carried over waves of infinitesimally different λ's and ν's, it can be written as an integral with respect to λ and ν. Alternatively, we can express the wave packet as an integral with respect to k and ω because these are related to λ and ν, respectively, by Eqs. 11.12 and 11.14. The wave packet can be written as

$$\psi(x,t) = \int_0^\infty \int_0^\infty A(k,\omega) \sin (kx - \omega t)\, dk\, d\omega \qquad (19.7)$$

For a given $k(\lambda)$ and $\omega(\nu)$, the amplitude A has a definite value. However, A can be different for different k's and ω's. The coefficient A basically determines how much of a particular wavelength and frequency we mix with the others. These coefficients determine the shape of the wave packet. Because of mathematical complexity, we will not go over the details involved in the calculation of the integral, Eq. 19.7. However, as an illustration, Fig. 19-5 shows the spacial variation of ψ at $t = 0$ when we mix various amounts of sinusoidal waves of different wavelength. The graphs to the left show the form of A (that is, the amount of mixing), and the graphs to the right, the resulting wave packet ψ.

The salient feature of these graphs is fairly obvious: The greater the range of k's (and therefore λ's) that we mix, the narrower is the width of the resulting

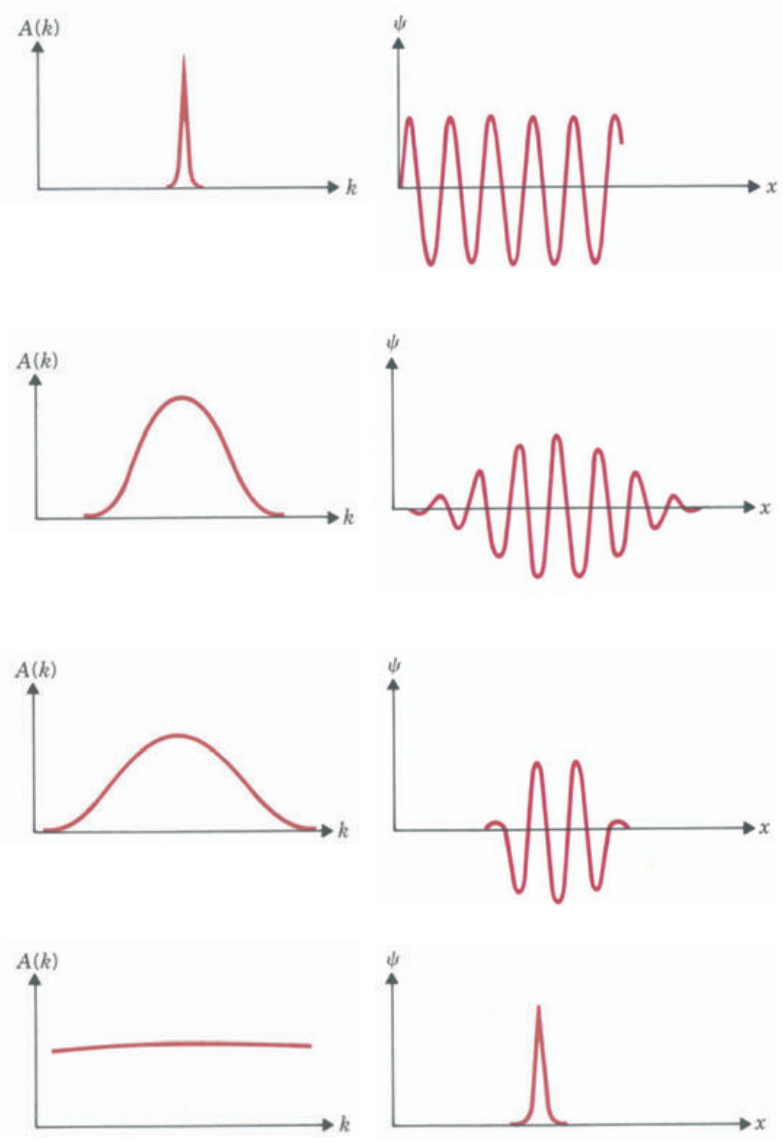


FIGURE 19-5
Wave packets obtained by mixing (adding) sinusoidal traveling waves of different wave vector k (and therefore, different wavelength). The graphs to the left show the amplitude (the amount of mixing) of the sinusoidal waves being mixed as a function of k, the graphs to the right show the resulting wave packet. The salient feature of these graphs is obvious: The greater the range of k's being mixed, the narrower the resulting wave packet.

wave packet. The physical significance of this mathematical result is clear, and we can see the uncertainty principle coming through. When ψ contains only one k, the wave has a well-defined wavelength and consequently the particle with which this wave is associated has a well-defined momentum. The amplitude of the wave is the same for all points in space: The particle is completely unlocalized. To localize the particle, we must mix waves with a range of k's (and correspondingly a range of wavelengths). Because the wave in this case does not have a well-defined wavelength, the particle with which this wave is associated will not have a well-defined momentum. The wider the range of wavelengths that we mix, the wider the range of values that the momentum of the particle can take. But, at the same time, the narrower the wave packet will be and, consequently, the better localized the particle will be.

19.7 VELOCITY OF THE WAVE PACKET: GROUP VELOCITY

We have seen that to describe a localized particle we must use a wave packet. Because the wave packet accompanies the particle and tells us approximately where the particle may be found, it must travel with the same velocity as the particle.

Rather than looking at the wave packet of Eq. 19.7, let us consider a simpler case that is mathematically easier to handle. Let us mix two traveling waves that differ slightly in wavelength and frequency,

$$\psi_1 = A \sin [kx - \omega t]$$

$$\psi_2 = A \sin [(k + \Delta k)x - (\omega + \Delta\omega)t]$$

where

$$\Delta k << k \quad \text{and} \quad \Delta\omega << \omega$$

The resulting $\psi(x,t)$ will be

$$\psi(x,t) = \psi_1 + \psi_2$$

$$\psi(x,t) = A \sin [kx - \omega t] + A \sin [(k + \Delta k)x - (\omega + \Delta\omega)t]$$

If we use the trigonometric relation

$$\sin A + \sin B = 2 \cos \frac{A - B}{2} \sin \frac{A + B}{2}$$

we get,

$$\psi(x,t) = 2A \cos \left(\frac{\Delta k x - \Delta\omega t}{2} \right) \sin (kx - \omega t) \tag{19.8}$$

where we have used the approximations $2k + \Delta k \approx 2k$, $2\omega + \Delta\omega \approx 2\omega$. The resulting wave is thus the product of two traveling waves. The second term of Eq. 19.8 represents a wave having roughly the same frequency and wavelength as the original waves. The first term of Eq. 19.8 represents a wave having a much larger wavelength and much smaller frequency. (Because $\lambda = 2\pi/k$, if k is very small the corresponding λ will be very large. Thus, in Eq. 19.8 the λ represented by Δk must be much larger than the λ for k. The opposite is true for the frequency. Because $\omega = 2\pi\nu$, a small $\Delta\omega$ has a smaller frequency ν associated with it than does a large ω). We can consider ψ as a wave similar to the original ones except that its amplitude is modulated by the first term, giving rise to a periodically varying amplitude. A "snapshot" of ψ is shown in Fig. 19-6. Here we do not have a single wave packet as in the cases shown in Fig. 19-5, but rather a series of packets. Such a ψ could be used to describe a beam of particles, with one particle in each wave packet. The velocity of the wave inside the envelopes is the same as the velocity of the individual waves, $v = \omega/k$, Eq. 11.2. By analogy to this, and on inspection

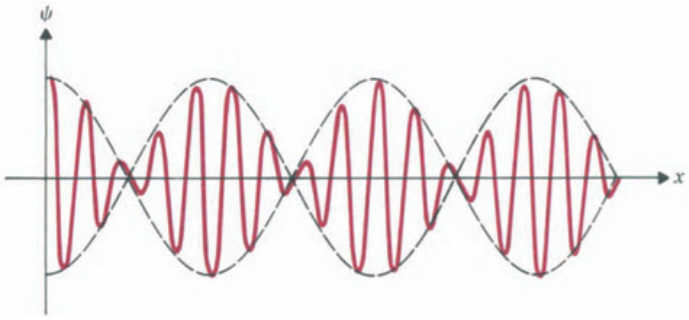


FIGURE 19-6
A "snapshot" of a wave with a periodically varying amplitude. Such a wave is obtained by adding two sinusoidal traveling waves of slightly different frequency and wavelength.

of Eq. 19.8, the envelopes (that is, the wave packets) travel with a velocity, called the group velocity, v_{group}, given by

$$v_{\text{group}} = \frac{\frac{\Delta\omega}{2}}{\frac{\Delta k}{2}} = \frac{\Delta\omega}{\Delta k} \approx \frac{d\omega}{dk} \tag{19.9}$$

We can show that v_{group} is the same as the velocity of the particle, by using de Broglie's relations,

$$\lambda = \frac{h}{p} \quad \text{or because } \lambda = \frac{2\pi}{k}, \quad k = \frac{p}{\hbar} \quad \text{hence } dk = \frac{dp}{\hbar}$$

and

$$\nu = \frac{E}{h} \quad \text{or because } \nu = \frac{\omega}{2\pi}, \quad \omega = \frac{E}{\hbar} \quad \text{hence } d\omega = \frac{dE}{\hbar}$$

Substituting for dk and $d\omega$ into the expression for v_{group}, Eq. 19.9, we get

$$v_{\text{group}} = \frac{dE}{dp} \tag{19.10}$$

But we know that $E = \frac{1}{2} m v_{\text{particle}}^2 = p^2/2m$; it follows that

$$v_{\text{group}} = \frac{dE}{dp} = \frac{p}{m} = \frac{m v_{\text{particle}}}{m} = v_{\text{particle}}$$

$$v_{\text{group}} = \frac{dE}{dp}$$

The wave packet moves with the particle. Although we have derived this result for a particularly simple case, it holds for the more general case described by Eq. 19.7.

19.8 THE PRINCIPLE OF COMPLEMENTARITY

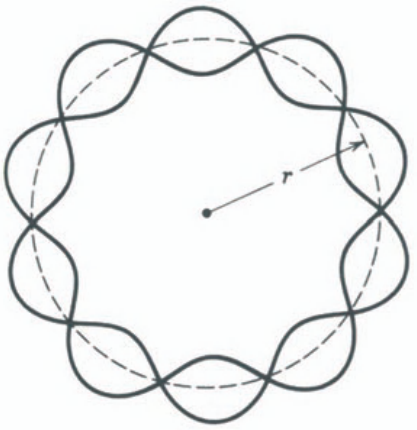
When we summarize our knowledge of nature, we see that energy can be carried two ways: *wave-like*, such as in the case of the ripples in a pond, or *particle-like*, such as in the transfer of energy from a gun to the target by the bullets. To explain these diverse phenomena, we construct two distinct math-

ematical models: a wave model and a particle model, each satisfying certain laws—Newton's laws of motion for particles and the superposition principle for waves.

As scientists, we try to extend the models into regions that are visually less accessible. We explain the behavior of sound using the wave model, and we do so successfully. Similarly, we use the particle model to explain the interaction of the molecules of a gas with the walls of the container, and again, we are successful.

These successes condition us to expect all entities to behave either as the particle model predicts or as the wave model does. Unfortunately, when we deal with electromagnetic radiation or with elementary particles such as electrons, protons, or neutrons, neither model is able to account for their complete behavior. We need a more complex model, one that encompasses aspects of both the particle and the wave models. This dual character of matter is summarized in Bohr's principle of complementarity: *The particle and the wave models are complementary.* Certain measurements reveal the wave aspects of electromagnetic radiation and of material particles; measurements that involve their spacial distribution. Other measurements reveal their particle aspects; measurements dealing with the interaction with one another or with other entities.

Because the predictions of the wave model concerning either the propagation through space or interactions are incompatible with those of the particle model, no measurement can simultaneously reveal the particle and the wave properties of matter. In a given situation, electrons are either particles or waves, not both. However, to understand their overall behavior, both models are needed and, as we have shown, both models can be made mathematically equivalent through the connecting link of de Broglie's hypothesis.



The allowed orbits for the electron in the hydrogen atom (as proposed by Bohr) are those in which the de Broglie wavelength of the electron for that orbit fits an integral number of times (see problem 19.15).

PROBLEMS

19.1 The wavelength of one of the yellow emission lines of sodium is 5890 Å. What is the kinetic energy of an electron having the same de Broglie wavelength?

19.2 The de Broglie wavelength of a proton is 10^{-13} m. (a) What is the speed of the proton? (b) Through what potential difference must the proton be accelerated to acquire such a speed?

19.3 Through what potential difference must an electron be accelerated for its de Broglie wavelength to be equal to the smallest wavelength of the X rays produced by an X-ray tube operating at 50,000 V?

(*Answer*: 2440 V.)

19.4 For quick calculations of the wavelength of electrons accelerated through a potential difference V, physicists use the formula $\lambda(\text{Å}) = \sqrt{150/V}$. Derive this formula, and find the percentage inaccuracy of the equation.

(*Answer*: 0.32%.)

19.5 The limit of resolution of an object is of the same order of magnitude as the wavelength used to "see" the object. An electron microscope operates at 60,000 V. What is the approximate size of the smallest object that can be seen with such a microscope?

19.6 An electron, a neutron, and a photon have

the same wavelength, $\lambda = 1$ Å. (a) Calculate and compare the frequency of each. (b) Calculate the energy of each.

(*Answer*: (a) 3.64×10^{16} Hz (electron), 1.98×10^{13} Hz (neutron), 3.00×10^{18} Hz (photon), (b) 151 eV (electron), 8.2×10^{-2} eV (neutron), 1.24×10^{4} eV (photon).)

19.7 An α particle is emitted from a nucleus with an energy of 5 MeV (5×10^{6} eV). Calculate the wavelength of an α particle with such energy and compare it with the size of the emitting nucleus that has a radius of 8×10^{-15} m.

(*Answer*: 6.4×10^{-15} m.)

19.8 What is the energy of an electron with a wavelength equal to the nuclear diameter, about 10^{-14} m?

19.9 A beam of electrons having a range of velocities enters a region of space with an electric field perpendicular to a magnetic field and both fields perpendicular to the direction of the beam (see Fig. 19-7). What is the de Broglie wavelength of the emerging electrons? The symbol $\cdot$ next to B in Fig. 19-7 indicates that the direction of the magnetic field is out of the page.

(*Answer*: $\frac{hB}{m\mathscr{E}}$.)

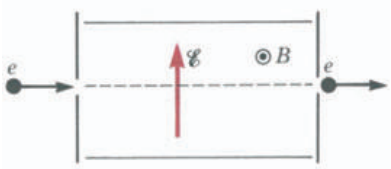


FIGURE 19-7
Problem 19.9.

19.10 A neutron of wavelength 3 Å undergoes a first-order Bragg reflection from a calcite crystal with lattice spacing 3.036 Å. Calculate the reflection angle.

19.11 In neutron spectroscopy a beam of monoenergetic neutrons is obtained by reflecting reactor neutrons from a beryllium crystal. If the separation between the atomic planes of the beryllium crystal is 0.732 Å, what is the angle between the incident neutron beam and the atomic planes that will yield a monochromatic beam of neutrons of wavelength 0.1 Å?

(*Answer*: 3.9°.)

19.12 In the Davisson and Germer experiment described in Section 19.2, at what angles would the second- and third-order diffraction maxima occur?

(*Answer*: They cannot be observed.)

19.13 A beam of α particles ($q = +2e$) is accelerated through a potential difference of 10 V. It then strikes a NaCl crystal ($d = 2.82$ Å). What is the highest order Bragg reflection that can be observed?

19.14 Show that radius of the nth electron orbit in the Bohr atom is equal to $n/2\pi$ times the de Broglie wavelength of the electron in that orbit.

19.15 Suppose we wish to develop the Bohr model from the de Broglie hypothesis by assuming that the allowed orbits are those that contain an integral number of de Broglie wavelengths. Show that this condition leads to Bohr's first postulate, Eq. 18.10, $mvr = n\hbar$.

19.16 A small particle of mass 10^{-6} g moves along the x axis; its speed is uncertain by 10^{-6} m/sec. (a) What is the uncertainty in the x coordinate of the particle? (b) Repeat the calculation for an electron assuming that the uncertainty in its velocity is also 10^{-6} m/sec.

(*Answer*: (a) 1.05×10^{-19} m, (b) 115 m.)

19.17 What is the uncertainty in the position of an electron that has been accelerated through a potential difference $V = 1000 \pm 0.1$ V. (*Hint:* Express the momentum p of the electron in terms of V; differentiate p (V) with respect to V to find Δp in terms of ΔV).

19.18 In an experiment the position of an electron is determined with an accuracy of $\pm 10^{-7}$ m. What is the uncertainty in the position of the electron 10^{-6} sec later?

(*Answer*: 0.6×10^{-3} m.)

19.19 The uncertainty in the position of a particle is equal to the de Broglie wavelength of the particle. Calculate the uncertainty in the velocity of the par-

ticle in terms of the velocity of the de Broglie wave associated with the particle.

(*Answer*: v_{wave}/π.)

19.20 A beam of monochromatic electrons is sent through a double slit. The width of the slits is smaller than the de Broglie wavelength of the electrons. The separation between the slits d is greater than but of the same order of magnitude as λ. A detector is placed to the right of the slits at an angle θ with respect to the incident direction of the electrons (see Fig. 19-8). When slit 1 is closed, N_2 electrons reach the detector every second. When slit 2 is closed, N_1 electrons reach the detector every second. Find the number of electrons reaching the detector per second when both slits are open if θ is such that $d \sin \theta = \lambda$, where λ is the de Broglie wavelength of the electrons.

(*Answer*: $N_1 + N_2 + 2\,(N_1N_2)^{1/2}$)

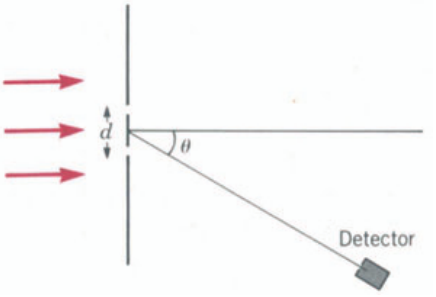


FIGURE 19-8
Problem 19.20.

19.21 In elementary statistics it is shown that the standard deviation σ_x associated with a quantity x is equal to

$$\sigma_x^2 = \overline{x^2} - \bar{x}^2$$

where $\overline{x^2}$ is the average of the squares of x, that is,

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}$$

and $\bar{x}^2$ is the square of the average of x, that is,

$$\bar{x}^2 = \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

The standard deviation is often taken as a measure of the uncertainty in the quantity involved. (a) Use this to estimate the minimum energy of an electron confined in an atom of radius 10^{-10} m. (b) How does the answer to part (a) compare with the kinetic energy of the electron in the hydrogen atom? (c) Repeat the calculation of part (a) for an electron confined to a nucleus of radius 10^{-14} m. (d) The binding energy of the nuclear constituents is a few MeV (10^6 eV). What conclusion can we draw about the presence of electrons in the nucleus?

(*Answer*: (a) 11 eV, (c) 1.1×10^9 eV.)

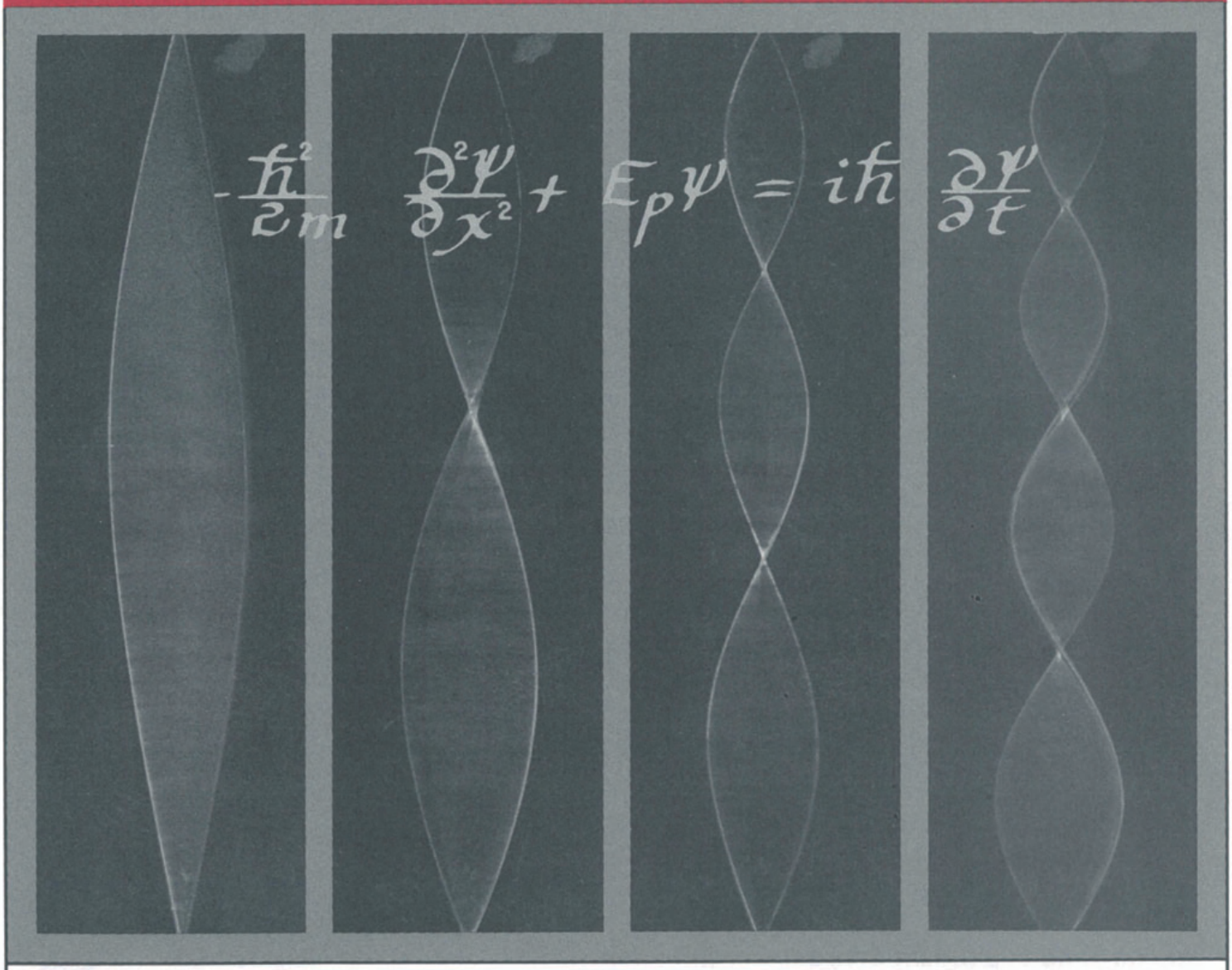
19.22 Use the uncertainty principle to show that the minimum energy of a particle undergoing simple harmonic motion in one dimension is of the order of $h\nu$, where h is Planck's constant and ν is the frequency of the motion. Take the standard deviation presented in Problem 19.21 as the uncertainty in the position and in the momentum of the particle, that is,

$$(\Delta x)^2 = \sigma_x^2 = \overline{x^2} - \bar{x}^2$$

and

$$(\Delta p_x)^2 = \sigma_{px}^2 = \overline{p_x^2} - \bar{p}_x^2$$

Note that in this case $\bar{x}^2 = 0$ and $\bar{p}_x^2 = 0$.



CHAPTER 20

An Introduction to the Methods of Quantum Mechanics

20.1 INTRODUCTION

So far, we have discussed some of the experimental evidence that led to the breakdown of classical physics and to the beginning of quantum mechanics. We have seen how the introduction of the quantization postulates explained the experimental facts concerning blackbody radiation, the photoelectric effect, and the hydrogen spectrum. These theories constitute what we call today the old quantum theory (OQT).

Despite its successes, the OQT has some serious deficiencies:

1. The theory can be applied only to periodic systems (harmonic oscillators, circular motion, and such), although there are many important physical systems that are not periodic.
2. Although the Bohr theory predicts the observed wavelengths of the spectrum of hydrogen, it does not explain why certain wavelengths are more intense than others; that is, it does not account for the rate of transition between different energy levels.
3. The Bohr theory explains well the spectrum of monatomic hydrogen (H), singly ionized helium (He^+), and reasonably well those of the alkali elements (lithium Li, sodium Na, potassium K, . . .), but only because these are H-like atoms (as will be shown in Chapter 21). It fails to explain the spectrum of even the simplest of the multielectron atoms, He.
4. But perhaps the most serious criticism of the OQT is that it is intellectually unsatisfying; it is not a unified or a general theory. It assigns to microscopic particles a well-defined path that, by the uncertainty principle, is not possible.

20.2 THE SCHRÖDINGER THEORY OF QUANTUM MECHANICS

20.2a The Time-dependent Schrödinger Equation

The basis of the modern theory of quantum mechanics was developed in 1925 by Erwin Schrödinger (1887–1961); an equivalent, but mathematically different, theory was presented just about the same time by Werner Heisenberg. In this book we will deal only with the Schrödinger theory.

The most important fact that we have presented so far is that the behavior of a microscopic particle is governed by the wave associated with it. In trying to find the wave associated with the particle, de Broglie's postulates give us the first guideline. We have seen that if a particle has a well-defined momentum and energy, we can use a sinusoidal traveling wave, that is, either

$$\psi = A \sin(kx - \omega t) \quad \text{or} \quad \psi = A \cos(kx - \omega t)$$

or a linear combination of both. As we have seen, if we want to describe a free particle, which is partially localized, we could use a wave packet.

Erwin Schrödinger (1887-1961).

De Broglie's hypothesis does not tell us what type of wave one can associate with a particle that is not free and that is acted on by a force. If a particle is acted on by a force, its momentum and its energy will not be constant. Therefore, it is meaningless to talk about a λ and a ν, because these quantities are changing. The Schrödinger theory tells us how to obtain the wavefunction $\psi(x,t)$ associated with a particle, when we specify the forces acting on the particle, by giving the potential energy associated with the forces. (In quantum mechanics the potential energy is often referred to simply as the potential). The Schrödinger theory also tells us how to extract information about the particle from the associated wavefunction.

Schrödinger developed a differential equation whose solutions yield the possible wavefunctions that can be associated with a particle in a given physical situation. This equation, known as the *Schrödinger equation*, tells us how the wavefunction changes as a result of the forces acting on the particle. Because the wavefunction ψ is a function of space and time, the equation contains derivatives (remember that a derivative represents the rate of change) with respect to x, y, and z and with respect to t. In this chapter we will primarily consider motion in only the x direction. In mathematics, when a function depends on more than one independent variable, the derivative of the function with respect to one of them, while treating the other variables as constants, is called the *partial derivative* with respect to that variable. The partial derivative of ψ with respect to x is written as $\partial\psi/\partial x$, with similar notation for the partial derivative with respect to t. Several plausibility arguments can be used to arrive at the Schrödinger equation. We present here one that is based on the idea, to be demonstrated later, that there is a relation between mathematical *operators* and physical quantities. (In mathematics an operator is represented by a symbol or group of symbols and indicates an operation to be performed. Thus, for example, when the operator $-i\hbar\,\partial/\partial x$ is placed in front of a function, it indicates that the function is to be differentiated with respect to x and then multiplied by $-i\hbar$, where i is the imaginary number $\sqrt{-1}$.)

The total energy of a particle is equal to the kinetic energy plus the potential energy,

$$E = \frac{1}{2}mv^2 + E_p$$

$$= \frac{p^2}{2m} + E_p$$

Multiplying both sides of this equation by the wavefunction ψ we obtain

$$E\psi = \frac{p^2}{2m}\psi + E_p\psi \tag{20.1}$$

Later in this chapter (see Section 20.2c) we will see that there is a relation between the energy E and the operator $i\hbar\,\partial/\partial t$ and between the momentum p and the operator $-i\hbar\,\partial/\partial x$. By this we mean that, operating on the function

with $i\hbar\,\partial/\partial t$ and with $-i\hbar\,\partial/\partial x$ is the same as multiplying the function by E and p, respectively. Substituting these operators for E and p in Eq. 20.1, we have

$$i\hbar\frac{\partial\psi}{\partial t} = \frac{1}{2m}\left(-i\hbar\frac{\partial}{\partial x}\right)\left(-i\hbar\frac{\partial}{\partial x}\right)\psi + E_p\psi$$

or, because i and $\hbar$ are constants,

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + E_p\psi$$

which is conventionally written as

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + E_p\psi = i\hbar\frac{\partial\psi}{\partial t} \tag{20.2}$$

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + E_p\psi = i\hbar\frac{\partial\psi}{\partial t}$$

Equation 20.2 is known as the *one-dimensional time-dependent Schrödinger equation*. If the potential energy E_p is known, this equation can be solved in principle, and the solution will yield the possible wavefunctions that we can associate with the particle. The Schrödinger equation is to quantum mechanics what Newton's second law is to classical physics.

One sometimes finds "proofs" of the Schrödinger equation. These so-called proofs are simply plausibility arguments. Somewhere along the argument, at a critical point, the argument becomes a postulate. For example, in the argument we just presented to arrive at Eq. 20.2, we used the fact that there is a relation between E and $i\hbar\,\partial/\partial t$ and between p and $-i\hbar\,\partial/\partial x$. We will prove this for the free particle case. When the particle is not free, the relationship cannot be proved, but we postulate that it still holds and experiment bears this out. The Schrödinger equation cannot be derived from first principles; it is a first principle, which cannot be mathematically derived, just as Newton's laws of motion are not derivable. *The justification lies in the fact that its predictions agree with the experiment.*

20.2b The Schrödinger Equation for a Free Particle

Let us consider a free particle moving along the x axis with definite momentum $p = mv$ and definite energy $E = 1/2\ mv^2$. If no force acts on the particle, that is, $F = 0$, the potential energy is $E_p =$ constant, which we can choose to be 0. Thus, the condition $F = 0$ requires that $E =$ constant. The Schrödinger equation (Eq. 20.2) in this case may be written for $E_p = 0$ as

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} = i\hbar\frac{\partial\psi}{\partial t} \tag{20.3}$$

As mentioned at the beginning of Section 20.2a, because we are dealing with a particle of well-defined momentum and energy, we might expect that the solution would be in the form of a traveling wave, that is, either

$$\psi = A\sin(kx - \omega t) \quad \text{or} \quad \psi = A\cos(kx - \omega t)$$

or some linear combination of these two functions. If one tries either of these by substitution into Eq. 20.3, one finds that neither satisfies the Schrödinger equation. The reason is that when you differentiate a sine function twice with respect to x, you get the sine function back, but when you differentiate it with respect to time once, you get a cosine function. For example, let us consider $\psi = A \sin(kx - \omega t)$. Differentiating with respect to x, we first get

$$\frac{\partial \psi}{\partial x} = \frac{\partial}{\partial x}[A \sin(kx - \omega t)] = kA \cos(kx - \omega t)$$

and the second differentiation yields

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{\partial}{\partial x}[kA \cos(kx - \omega t)] = -k^2 A \sin(kx - \omega t)$$

The derivative of ψ with respect to t yields

$$\frac{\partial \psi}{\partial t} = \frac{\partial}{\partial t}[A \sin(kx - \omega t)] = -\omega A \cos(kx - \omega t)$$

Substituting these results for $\partial^2\psi/\partial x^2$ and $\partial\psi/\partial t$ in Eq. 20.3, we obtain

$$\frac{k^2\hbar^2}{2m} A \sin(kx - \omega t) = -i\hbar\omega A \cos(kx - \omega t)$$

Because the sine and cosine functions are equal only for certain angles (for example, 45°), this cannot be satisfied for all x's and t's. Similar results are obtained with the cosine function. There is, however, a particular combination of these two functions that does satisfy the Schrödinger equation. This combination is

$$\psi = A[\cos(kx - \omega t) + i \sin(kx - \omega t)]$$

which, by mathematical definition, may be written as

$$\psi = A e^{i(kx - \omega t)} \tag{20.4}$$

We may show that Eq. 20.4 satisfies the Schrödinger equation by substitution into Eq. 20.3. Differentiate Eq. 20.4 twice with respect to x

$$\frac{\partial \psi}{\partial x} = \frac{\partial}{\partial x}[A e^{i(kx - \omega t)}] = ik A e^{i(kx - \omega t)}$$

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{\partial}{\partial x}[ik A e^{i(kx - \omega t)}] = (ik)^2 A e^{i(kx - \omega t)} = -k^2 A e^{i(kx - \omega t)}$$

The derivative with respect to t yields

$$\frac{\partial \psi}{\partial t} = \frac{\partial}{\partial t}[A e^{i(kx - \omega t)}] = -i\omega A e^{i(kx - \omega t)}$$

and, substituting into Eq. 20.3, we get

$$\frac{\hbar^2 k^2}{2m} A\, e^{i(kx-\omega t)} = \hbar\omega\, A\, e^{i(kx-\omega t)}$$

or

$$\frac{\hbar^2 k^2}{2m} = \hbar\omega \tag{20.5}$$

which means that Eq. 20.4 is a solution for this relation of the constants. We must now show that Eq. 20.5 is consistent with de Broglie's hypothesis. The kinetic energy of a particle is given by $E = 1/2\, mv^2 = p^2/2m$. From de Broglie's hypothesis,

$$E = h\nu = h\frac{\omega}{2\pi} = \hbar\omega$$

$$p = \frac{h}{\lambda} = \frac{h}{\frac{2\pi}{k}} = \hbar k$$

Substituting these relations into the expression for the energy, $E = \frac{p^2}{2m}$, we get

$$\hbar\omega = \frac{\hbar^2 k^2}{2m}$$

which is the same as Eq. 20.5, and the consistency of the formulation has been demonstrated. That is, for a free particle of definite momentum and energy, the wavefunction ψ that satisfies the Schrödinger equation and de Broglie's postulate is given by Eq. 20.4. The fact that Eq. 20.4 is not a real function (that is, it has an imaginary component) is not a problem. The wavefunction itself has no physical meaning. We have seen (see Eq. 19.3) that it is $|\psi|^2$ that has physical significance. When ψ is a complex function we write by mathematical definition that

$$|\psi|^2 = \psi^*\psi$$

$|\psi|^2 = \psi^*\psi$

where ψ^* is the complex conjugate of ψ. (The complex conjugate of a function ψ is a function whose real part is the same as the real part of ψ and whose imaginary part is the negative of the imaginary part of ψ.) In the preceding case,

$$|\psi|^2 = \psi^*\psi = A^* e^{-i(kx-\omega t)}\, A\, e^{i(kx-\omega t)}$$
$$= A^* A$$

This result is a real quantity, which it must be by Eq. 19.3, because an imaginary probability is not defined mathematically. In this equation the probability of finding the particle at any point in space is given as $|\psi|^2 dV$, and

therefore $|\psi|^2$, when properly normalized, (Eq. 19.4), may be considered a probability density that must be both real and positive.

20.2c Expectation Values

The preceding solution of Eq. 20.2 was for the case of $E_p = 0$. The method, however, is valid for all functional forms of E_p, although the solution of $\psi(x, t)$ is more complicated. The following discussion applies to all solutions of Eq. 20.2, regardless of the form of E_p.

One of the consequences of the uncertainty principle is that the position, momentum, energy, and so forth of a particle cannot, in general, be precisely determined. We can, however, talk about the average value of these quantities. Average values in quantum mechanics are called *expectation values*. The physical significance of the expectation value is the following. Suppose we take a large number of particles, all of them described by the same $\psi(x, t)$. If we measure the value of a particular dynamical quantity, for example, the momentum p, we will find that, in general, p is different for different particles even though all are described by the same $\psi(x, t)$. We can, however, calculate the average value of p. If we repeat our experiment with another system of a large number of particles each described by the same $\psi(x, t)$, we will find that the average value of p will be the same as in the first experiment. Expectation values do have physical significance as a statistical average even in cases where the exact value of the individual dynamical quantity is not well defined.

The question that we may ask is: Can we determine these expectation values without doing the experiment if we know the wavefunction $\psi(x, t)$ associated with the system? The answer is yes.

The average position of a system of particles is usually obtained by the standard statistical method of adding the positions of all the particles and dividing this sum by the total number of particles. If we have N_{Total} particles distributed along the x axis in such a way that there are N_1 particles at x_1, N_2 at x_2, N_3 at x_3, . . ., then the average particle position $\bar{x}$ will be given by

$$\bar{x} = \frac{x_1N_1 + x_2N_2 + x_3N_3 + \cdots}{N_{\text{Total}}}$$

$$= \frac{\sum x_iN_i}{\sum N_i}$$

This expression is analogous to the one used in Chapter 6 (Eq. 6.2) for the center of mass of a system of particles. The only difference is that the weighting factor here is the number of particles at each location rather than the mass.

Suppose that instead of a number of particles we have a single particle and that it has different probabilities of being found at different locations, for example, probability P_1 of being found at position x_1, P_2 at x_2, ... Also,

because the particle must be somewhere, we know that the sum of all the probabilities must be unity, that is,

$$P_1 + P_2 + P_3 + \cdots = 1$$

This statement about the sum of probabilities bears the term that they are *normalized*. The expression for the average or mean particle position can now be written as

$$\bar{x} = x_1P_1 + x_2P_2 + x_3P_3 + \cdots$$

or

$$\bar{x} = \sum x_iP_i$$

When the intervals $\Delta x = x_i - x_{i-1}$ approach zero—that is, the position that the particle may be at varies continuously—the preceding summation becomes an integral and

$$\bar{x} = \int_{-\infty}^{\infty} xP(x)\,dx$$

As in the case of the discrete probabilities, P_i's, the continuous sum of probabilities must be normalized, that is,

$$\int_{-\infty}^{\infty} P(x)\,dx = 1$$

If the particle's position varies with time, then the probability of finding the particle between x and $x+dx$ is a function of time and is written as $P(x,t)\,dx$. The average value of x will be

$$\bar{x} = \frac{\int_{-\infty}^{\infty} xP(x,t)\,dx}{\int_{-\infty}^{\infty} P(x,t)\,dx}$$

where the denominator has been included to ensure that the P's are properly normalized.

In Chapter 19, Eq. 19.3, we saw that the probability $P(r,t)\,dV$ that a particle be found within a small volume dV centered around a point with position vector r was $|\psi|^2dV$. In the special case under consideration, where the particle is restricted to the x axis, the probability $P(x,t)\,dx$ of finding the particle at time t between x and $x + dx$ will be $|\psi|^2dx$. We mentioned earlier that when ψ is a complex function, $|\psi|^2 = \psi^*\psi$, where ψ^* is the complex conjugate of ψ. Thus,

$$\bar{x} = \frac{\int_{-\infty}^{\infty} x\psi^*\psi\,dx}{\int_{-\infty}^{\infty} \psi^*\psi\,dx}$$

where the denominator ensures that the probability is normalized. If we have wavefunctions ψ that we know are normalized we do not bother to write the denominator because we know that it equals unity. It is customary to write the numerator with the x between the two wavefunctions as

$$\bar{x} = \int_{-\infty}^{\infty} \psi^* x \psi \, dx \tag{20.6}$$

$\bar{x} = \int_{-\infty}^{\infty} \psi^* x \psi \, dx$

Equation 20.6 tells us that the average value of x is the sum of all the possible values of x weighted by the probability that the system may be found at each possible value. For this relation to be valid, ψ must be normalized so that the sum of their probabilities for all x is unity, that is,

$$\int_{-\infty}^{\infty} \psi^* \psi \, dx = 1$$

$\int_{-\infty}^{\infty} \psi^* \psi \, dx = 1$

Not only can we get the average value of x from ψ by this method, but we can similarly find the average value of any dynamical quantity that can be written as a function of x, for example, the average value of the potential energy $E_p(x)$ is

$$\overline{E_p} = \int_{-\infty}^{\infty} \psi^* E_p(x) \psi \, dx$$

$\overline{E_p} = \int_{-\infty}^{\infty} \psi^* E_p(x) \psi \, dx$

It should be clear why this expression is correct. If the particle is at a given x, it will have the potential energy $E_p(x)$ associated with that x. Thus $\psi^*\psi dx$ is not only the probability that the particle be found between x and $x + dx$ but also the probability that it will have a potential energy between $E_p(x)$ and $E_p(x + dx)$. The same applies to any other dynamical quantity that can be expressed in terms of x.

There are quantities, such as the momentum and the total energy, that in classical physics can be expressed in terms of the coordinates of the particles. In quantum mechanics, however, this is not permitted because it would violate the uncertainty principle. In Chapter 19, when we introduced the uncertainty principle, we indicated that it is not possible to determine exactly both the position and the momentum of a particle. Therefore, we cannot express the momentum as a function of x, that is, $p(x)$. How can we get information about the momentum of the particle? Let us consider the free particle wavefunction that we found previously (see Eq. 20.4). This wavefunction represents a particle with definite momentum $p = \hbar k$ and definite energy $E = \hbar\omega$. Let us differentiate the wavefunction given in Eq. 20.4 with respect to x and then multiply both sides by $-i\hbar$.

$$\frac{\partial \psi}{\partial x} = ik \, A \, e^{i(kx - \omega t)}$$

$$-i\hbar \frac{\partial \psi}{\partial x} = (-i\hbar) ik \, A \, e^{i(kx - \omega t)} = \hbar k \, A \, e^{i(kx - \omega t)}$$

$$-i\hbar \frac{\partial \psi}{\partial x} = p\psi \tag{20.7}$$

This result tells us that there is an association between the dynamical quantity p and the operator $-i\hbar \dfrac{\partial}{\partial x}$. The effect of multiplying the wavefunction ψ by the momentum p is the same as operating on ψ with the operator $-i\hbar \dfrac{\partial}{\partial x}$, and the equivalence may be expressed as

$$p \longleftrightarrow -i\hbar \frac{\partial}{\partial x}$$

$$p \longleftrightarrow -i\hbar \frac{\partial}{\partial x}$$

Let us now return to our method of calculating expectation values, Eq. 20.6. The method consists in first multiplying the wavefunction ψ by the quantity in question (x, $E_p(x)$, p, . . .), we then multiply the result by the complex conjugate of the wavefunction ψ^* and integrate over all space. When we try this method for the momentum p, we run into trouble because we cannot express p in terms of x. However, as indicated by Eq. 20.7, instead of multiplying ψ by p, we could operate on ψ with $-i\hbar \dfrac{\partial}{\partial x}$, because the effect is the same. We obtain the expectation value of p for a particle moving along the x axis as follows

$$\bar{p} = \int_{-\infty}^{\infty} \psi^* \left(-i\hbar \frac{\partial}{\partial x}\right) \psi \, dx \tag{20.8}$$

$$\bar{p} = \int_{-\infty}^{\infty} \psi^* \left(-i\hbar \frac{\partial}{\partial x}\right) \psi \, dx$$

Let us apply this to the free particle wavefunction, Eq. 20.4,

$$\bar{p} = \int_{-\infty}^{\infty} \psi^* \left(-i\hbar \frac{\partial}{\partial x}\right) A\, e^{i(kx-\omega t)}\, dx$$

$$\bar{p} = \int_{-\infty}^{\infty} \psi^* \left(-i\hbar\, ik\, A\, e^{i(kx-\omega t)}\right) dx$$

$$\bar{p} = \hbar k \int_{-\infty}^{\infty} \psi^* \psi \, dx$$

$$\bar{p} = \hbar k$$

which is a form of de Broglie's relation. In the last step, we made use of the normalization condition mentioned earlier; that is, the probability of finding the particle anywhere in space must be 1.[1]

$$\int_{-\infty}^{\infty} \psi^* \psi \, dx = 1$$

This result is mathematically consistent; that is, the average or expectation value of something that has a definite value is that value itself.

A similar method can be used to find the average value of the energy. Let us again consider the free particle wavefunction, Eq. 20.4, associated with

[1]Although the wavefunction $\psi = A\, e^{i(kx-\omega t)}$ is not normalizable, it can be normalized by making the limits of integration very large but not infinite.

a particle of definite energy $E = \hbar\omega$.

$$\psi = A\, e^{i(kx - \omega t)}$$

differentiate both sides with respect to t

$$\frac{\partial \psi}{\partial t} = -i\omega\, A\, e^{i(kx - \omega t)}$$

multiply both sides by $i\hbar$

$$i\hbar \frac{\partial \psi}{\partial t} = (i\hbar)(-i\omega)\, A\, e^{i(kx - \omega t)} = \hbar\omega\, A\, e^{i(kx - \omega t)}$$

$$i\hbar \frac{\partial}{\partial t} \psi = E\, \psi \tag{20.9}$$

We see that just as in the case of the momentum, there is an association between the energy of the particle and an operator, in this case the operator is $i\hbar \dfrac{\partial}{\partial t}$. The effect of multiplying the wavefunction ψ by the energy E is the same as operating on ψ with the operator $i\hbar \dfrac{\partial}{\partial t}$, and the equivalence may be expressed as

$$E \longleftrightarrow i\hbar \frac{\partial}{\partial t}$$

$E \longleftrightarrow i\hbar \dfrac{\partial}{\partial t}$

Therefore, the average or expectation value of E is obtained by the same method of finding averages, namely

$$\overline{E} = \int_{-\infty}^{\infty} \psi^*\, E\, \psi\, dx$$

$$\overline{E} = \int_{-\infty}^{\infty} \psi^* \left(i\hbar \frac{\partial}{\partial t} \right) \psi\, dx \tag{20.10}$$

Let us apply this to our free particle wavefunction, Eq. 20.4.

$$\overline{E} = \int_{-\infty}^{\infty} \psi^* \left(i\hbar \frac{\partial}{\partial t} \right) A\, e^{i(kx - \omega t)}\, dx$$

$$\overline{E} = \int_{-\infty}^{\infty} \psi^*\, (i\hbar)(-i\omega)\, A\, e^{i(kx - \omega t)}\, dx$$

$$\overline{E} = \hbar\omega \int_{-\infty}^{\infty} \psi^* \psi\, dx$$

$$\overline{E} = \hbar\omega$$

This result is mathematically consistent because the wavefunction represents a particle of definite energy $E = \hbar\omega$ and the average value of something that has a definite value is the value itself.

20.2d Time-Independent Schrödinger Equation

The Schrödinger equation in one dimension is a partial differential equation that involves t and x as the independent variables. One standard technique for solving partial differential equations is the method of *separation of variables*. The method consists in trying a solution that is a product of two functions: one a function of x alone and the other a function of t alone; that is, let

$$\psi(x,t) = \chi(x)\,\Gamma(t) \tag{20.11}$$

$\psi(x,t) = \chi(x)\,\Gamma(t)$

In the cases in which this method works, the partial differential equation is reduced to two ordinary differential equations, each one involving only one of the independent variables. In the case of the Schrödinger equation, Eq. 20.2, the method works if the potential energy E_p is a function of x alone, that is, $E_p(x)$.

In the remainder of this section we are going to show that Eq. 20.11 satisfies the Schrödinger equation. Let us substitute it into Eq. 20.2,

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + E_p(x)\psi = i\hbar\frac{\partial\psi}{\partial t}$$

$$-\frac{\hbar^2}{2m}\Gamma(t)\frac{d^2\chi(x)}{dx^2} + E_p(x)\chi(x)\Gamma(t) = i\hbar\chi(x)\frac{d\Gamma}{dt}$$

If we divide through by $\chi(x)\Gamma(t)$, we get

$$-\frac{\hbar^2}{2m}\frac{1}{\chi(x)}\frac{d^2\chi(x)}{dx^2} + E_p(x) = i\hbar\frac{1}{\Gamma(t)}\frac{d\Gamma(t)}{dt}$$

The right side of the equation is a function of t alone, and the left side is a function of x alone. Because x and t are independent variables, the equation will be correct if both sides are equal to the same constant G which is called the *separation constant*, that is,

$$-\frac{\hbar^2}{2m}\frac{1}{\chi(x)}\frac{d^2\chi}{dx^2} + E_p(x) = G \tag{20.12}$$

and

$$i\hbar\frac{1}{\Gamma(t)}\frac{d\Gamma}{dt} = G \tag{20.13}$$

We now have two ordinary differential equations, one involving x alone and the other involving t alone. We should point out that the equation for $\Gamma(t)$, Eq. 20.13, does not involve $E_p(x)$. This means that the solution for $\Gamma(t)$ is the same in all physical situations where E_p does not depend on time. The differential equation for $\Gamma(t)$ can be solved easily by simple integration,

$$\frac{d\Gamma}{\Gamma} = \frac{G}{i\hbar}\,dt = -\frac{iG}{\hbar}\,dt$$

Integrating, we get

$$\ln \Gamma = -i\frac{G}{\hbar}t + \text{constant}$$

$$\Gamma(t) = Ke^{-iGt/\hbar} \tag{20.14}$$

where K is the natural antilog of the integration constant. Because it is an arbitrary constant, we will set it equal to unity (1) and drop it from the equation. We may evaluate the constant G by the following method. Let us operate on the wavefunction $\psi(x,t)$ with the energy operator $i\hbar\,\partial/\partial t$

$$\begin{aligned} i\hbar\frac{\partial}{\partial t}\psi(x,t) &= i\hbar\frac{\partial}{\partial t}\chi(x)\Gamma(t) \\ &= i\hbar\frac{\partial}{\partial t}\chi(x)\,e^{-iGt/\hbar} \\ &= i\hbar\,\chi(x)\left(-i\frac{G}{\hbar}\right)e^{-iGt/\hbar} \\ &= G\,\chi(x)\Gamma(t) \end{aligned}$$

$$i\hbar\frac{\partial}{\partial t}\psi(x,t) = G\,\psi(x,t) \tag{20.15}$$

Comparing Eq. 20.15 with Eq. 20.9, we see that the separation constant G is the total energy E of the system. Thus the time-dependent part of the wavefunction $\psi(x,t)$ is,

$$\Gamma(t) = e^{-iEt/\hbar} \tag{20.16}$$

$\Gamma(t) = e^{-iEt/\hbar}$

If we now wish to find what the space-dependent part of the wavefunction is, we have to solve the differential equation (20.12), where G is now E.

$$-\frac{\hbar^2}{2m}\frac{d^2\chi}{dx^2} + E_p(x)\,\chi = E\,\chi \tag{20.17}$$

$-\frac{\hbar^2}{2m}\frac{d^2\chi}{dx^2} + E_p(x)\,\chi = E\,\chi$

Equation 20.17 is called the *time-independent* Schrödinger equation. For a given $E_p(x)$, the equation has to to be solved to find the possible $\chi(x)$'s that can be associated with the system. So far, there are no restrictions on the values of the total energy E of the system. Thus, for different E's we expect to get different functions $\chi(x)$. We will see, however, that $\chi(x)$ must satisfy certain requirements that we discuss next. These requirements will, in some cases, restrict the number of physically acceptable $\chi(x)$'s and consequently the values of E that the system may have. In the Schrödinger theory, energy quantization is not introduced as an arbitrary postulate, as in the Bohr theory; it is a consequence of the fact that the wavefunction associated with the particle must be *well-behaved*, a term defined in the next section. The solutions χ of Eq. 20.17 are called the *eigenfunctions* or *eigenstates;* the corresponding values of E are called the *eigenvalues* (in German, "eigen" means "proper").

20.2e Required Properties of the Eigenfunction $\chi(x)$ and its Derivative

We just mentioned that $\chi(x)$ must be *well-behaved*. This means that the eigenfunction $\chi(x)$ and its derivative $d\chi/dx$ must have the following properties:

1. They must be *finite* everywhere. Figure 20-1*a* shows a function f that becomes infinite at $x = x_0$. Such a behavior for χ or $d\chi/dx$ is not acceptable.
2. They must be *single-valued* everywhere. In Figure 20-1*b* we show a function f that is not single-valued everywhere. A χ or $d\chi/dx$ exhibiting such behavior would not be acceptable.
3. They must be *continuous* everywhere. Figure 20-1*c* illustrates a function f with a discontinuity at $x = x_0$. Again, if χ or $d\chi/dx$ behaved in such a manner, that eigenfunction would not be acceptable. (In some special cases $d\chi/dx$ is not continuous.)

When these conditions are satisfied, the eigenfunction and the associated wavefunction (remember that $\psi(x, t) = \chi(x)\Gamma(t)$) are said to be well-behaved. Why must these conditions be satisfied?

Recall that (see Eqs. 20.6 and 20.8)

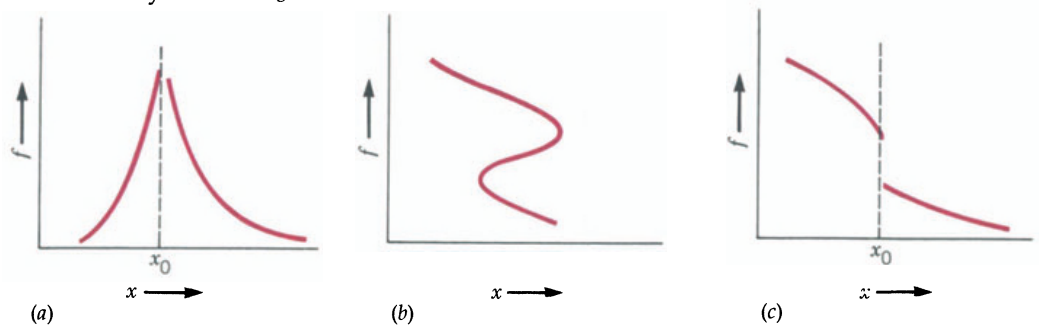
$$\bar{x} = \int_{-\infty}^{\infty} \psi^* \, x \, \psi \, dx$$

and

$$\bar{p} = \int_{-\infty}^{\infty} \psi^* \left(-i\hbar \frac{\partial}{\partial x} \right) \psi \, dx$$

If χ or $d\chi/dx$ are not finite, the average value of the momentum will be infinite; infinite momentum means infinite energy; this is not physically possible. Similarly, if χ or $d\chi/dx$ were not single-valued, the average position and the average momentum would not be defined. The *exact* value of x and p may

FIGURE 20-1
Examples of functions that are not well-behaved. (*a*) A function that becomes infinite at $x = x_0$. (*b*) A function that is not single-valued everywhere. (*c*) A function with a discontinuity at $x = x_0$.



be uncertain, but the averages have to be defined. The reason is that if we make a set of measurements on a very large number of identical systems, that is, systems described by the same $\psi(x, t)$, we will get a consistent, well-defined value for the average x and the average p. The calculation of these averages must reflect the experimental fact, that is, $\chi(x)$ and $d\chi/dx$ must be single-valued.

The requirement of continuity is closely tied to the requirement of finiteness. If χ is not continuous at a given point in space, then $d\chi/dx$ would be infinite at that point. This, as we have indicated, is not allowed. If $d\chi/dx$ is discontinuous at some point, then $d^2\chi/dx^2$ would be infinite at that point. From the time-independent Schrödinger equation, Eq. 20.17, this would imply that either E or E_p is infinite, which is not physically possible.

In the case of *bound* systems (one in which the particle is restricted to move over a finite region of space), these requirements on χ lead to the quantization of the energy and of other physical quantities of the system. This fact will become clear when we consider some simple cases.

20.3 APPLICATION OF THE SCHRÖDINGER THEORY

20.3a Particle Inside an Infinite Potential Well

Let us consider a particle such as an electron confined inside a one-dimensional well of length a with infinitely high walls (see Fig. 20-2).

$$E_p(x) = \begin{cases} \infty \text{ for } 0 > x > a \\ 0 \text{ for } 0 \leq x \leq a \end{cases}$$

Let us write E_p (outside) as E_{po}. Since E_{po} is ∞, one must provide an infinite amount of energy to pull the particle out of the well. This means that the particle cannot get out. This, obviously, is a highly idealized model. The reason for first considering the idealized model is that it is mathematically much easier to solve than are physically realistic ones. However, the main features of both idealized and realistic models are not really very different because E_{po} can be very large while remaining finite. We will see when we apply the model to the electrons in a solid that the predictions based on the present approximation of $E_{po} = \infty$ agree quite well with the experimental results.

Our task is to find all the possible wavefunctions that can be associated with the particle inside such a well. Before we do this, let us answer the question: What is the eigenfunction χ outside? The answer is $\chi = 0$. The reason is that for the particle to get out, an infinite amount of energy must be provided. This is physically impossible, and therefore $|\chi|^2 = 0$ outside the well, and we may state that

$$\chi = 0 \text{ for } 0 \geq x \geq a$$

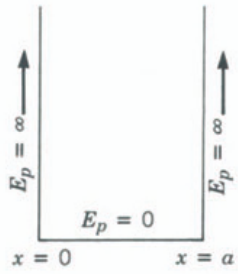


FIGURE 20-2
An infinite potential well.

To evaluate χ inside, we must solve Eq. 20.17 for the case when $E_p = 0$,

$$-\frac{\hbar^2}{2m}\frac{d^2\chi}{dx^2} = E\chi \tag{20.18}$$

Divide both sides by $-\hbar^2/2m$ and put both terms on the left side to get

$$\frac{d^2\chi}{dx^2} + \frac{2mE}{\hbar^2}\chi = 0$$

or

$$\frac{d^2\chi}{dx^2} + k^2\chi = 0 \tag{20.19}$$

where

$$k^2 = \frac{2mE}{\hbar^2} \tag{20.20}$$

Note that k in Eq. 20.20 is the wave vector $k = 2\pi/\lambda$ of Eq. 11.12. This can be readily shown. Because in this case the total energy E is simply the kinetic energy,

$$E = \frac{1}{2}mv^2 = \frac{p^2}{2m}$$

From de Broglie's hypothesis, Eq. 19.1,

$$p = \frac{h}{\lambda} = \frac{h}{\frac{2\pi}{k}} = \hbar k$$

Therefore

$$E = \frac{p^2}{2m} = \frac{\hbar^2 k^2}{2m}$$

and

$$k^2 = \frac{2mE}{\hbar^2}$$

which is the same relation as Eq. 20.20.

Equation 20.19 is a second-order differential equation with constant coefficients. Although mathematical methods are available for the solution, we may solve it simply by trial substitution of an assumed solution. Try a solution of the form,

$$\chi = e^{\alpha x} \tag{20.21}$$

where α is a constant as yet undetermined. Differentiating Eq. 20.21 with

respect to x, we obtain

$$\frac{d\chi}{dx} = \frac{d}{dx} e^{\alpha x} = \alpha e^{\alpha x}$$

$$\frac{d^2\chi}{dx^2} = \frac{d}{dx} \alpha e^{\alpha x} = \alpha^2 e^{\alpha x} = \alpha^2 \chi$$

Substituting for $d^2\chi/dx^2$ into Eq. 20.19 we have

$$\alpha^2 \chi + k^2 \chi = 0$$

$$\alpha^2 + k^2 = 0$$

$$\alpha = \pm ik$$

The two solutions are therefore e^{ikx} and e^{-ikx}. The general solution will be the sum of these two or

$$\chi(x) = a\, e^{ikx} + b\, e^{-ikx}$$

where a and b are arbitrary constants. By mathematical definition

$$a\, e^{ikx} = a \cos kx + ia \sin kx$$

$$b\, e^{-ikx} = b \cos kx - ib \sin kx$$

Therefore $\chi(x)$ can be written as

$$\chi(x) = (a + b) \cos kx + i(a - b) \sin kx$$

and, because a and b are arbitrary constants, we simplify the notation by introducing new constants A and B,

$$\chi(x) = A \cos kx + B \sin kx \qquad (20.22)$$

A and B are now the arbitrary constants to be determined from physical considerations.

Thus far, no restrictions have been found as to the values that k and therefore E can take. However, as soon as we require that χ be well-behaved, the restrictions on k and E will appear.

The conditions of finiteness and single-valuedness on χ and $d\chi/dx$ are satisfied by the χ of Eq. 20.22. The function is finite and single-valued for all values of x.

The condition that χ be continuous requires that Eq. 20.22 be 0 at $x = 0$ and at $x = a$, because χ is zero outside the well and from this we evaluate the constant A and the values of k. Because

$$\chi(x) = 0 \quad \text{for } x = 0$$

then Eq. 20.22 becomes

$$0 = A(1) + B(0)$$

It follows that

$$A = 0$$

and therefore

$$\chi = B \sin kx \tag{20.23}$$

Furthermore, because

$$\chi = 0 \quad \text{for } x = a$$

then Eq. 20.23 becomes

$$0 = B \sin ka$$

There are two possible ways to satisfy this latter condition. Either $B = 0$ or $\sin ka = 0$. The first choice is not an attractive one; if $B = 0$, then $\chi = 0$ everywhere. This means that the particle is not in the well. The only meaningful way to satisfy the condition of continuity is when

$$\sin ka = 0, \quad \text{that is,} \quad ka = 0, \pi, 2\pi, 3\pi, \cdots$$

or

$$k = n\frac{\pi}{a} \qquad n = 1, 2, 3, \cdots \tag{20.24}$$

Notice that $n = 0$ is not an acceptable choice. If $n = 0$, then $k = 0$ and from Eq. 20.23 χ would be zero everywhere; that is, the particle could not be in the well.

If we substitute the value for k from Eq. 20.24 into Eqs. 20.23 and 20.20, we will get a set of eigenfunctions and the corresponding set of eigenvalues that we can associate with the particle in the well.

$$\chi_n = B \sin n\frac{\pi}{a}x \tag{20.25}$$

$\chi_n = B \sin n\frac{\pi}{a}x$

and from Eq. 20.20,

$$E = \frac{k^2\hbar^2}{2m} = \frac{n^2\pi^2}{a^2}\frac{\hbar^2}{2m}$$

and the possible values of E are

$$E_n = n^2 E_0 \quad \text{where} \quad E_0 = \frac{\pi^2\hbar^2}{2ma^2} \tag{20.26}$$

$E_n = n^2\frac{\pi^2\hbar^2}{2ma^2}$

We should note that the first derivative of the eigenfunctions in this case is not continuous at $x = 0$ and at $x = a$. In this idealized example $E_p = \infty$ at $x = 0$ and $x = a$, and the requirement that the first derivative be continuous does not hold.

Let us analyze the results that we have obtained. The obvious result is that unlike the classical case of a particle bouncing between two walls, where the velocity and therefore the energy can have any value whatsoever, the

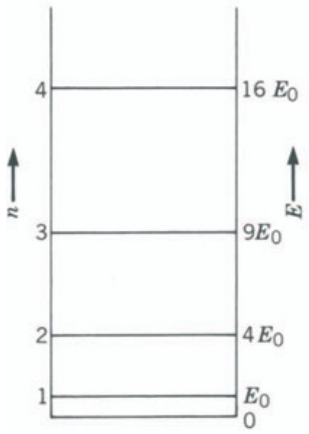


FIGURE 20-3
The first four allowed energy levels for a particle in an infinite potential well.

quantum mechanical result restricts (quantizes) the values of the energy. The first few allowed energy levels are represented schematically in Fig. 20-3. Thus, we see that in the Schrödinger theory energy quantization does not come in as a postulate; rather, it is a direct consequence of the fact that a particle is described by a wave and that the wave must satisfy certain conditions to be a good representation of the particle.

The model that we have been discussing could be used to describe a particle or system of particles confined inside a container with rigid walls. Classically, the internal energy of such a system of particles can be controlled by changing the temperature. Recall that the average energy of an atom in a gas is $E = 3/2\ k_B T$ (Eq. 9.21), where k_B is the Boltzmann constant and T the absolute temperature. At absolute zero temperature the energy would be zero. In quantum mechanics, as we can see, this is not possible. Even at absolute zero the energy cannot be zero, because $E = 0$ is not an allowed energy state if the particle is in the well. The smallest energy that the particle can have is E_0. This result comes out of the solution of the Schrödinger equation, but we could have guessed it from the uncertainty principle. If $E = 0$, then $p = 0$, and therefore $\Delta p = 0$. From the uncertainty principle it follows that $\Delta x = \infty$. But Δx cannot be ∞ because the particle is in the well; the largest Δx can be is a, the size of the well. It follows that $\Delta p \neq 0$, hence $p \neq 0$ and $E \neq 0$.

The solution of the Schrödinger equation has given us a set of eigenfunctions χ that can be used to describe a particle in the potential well. It does not tell us what particular χ is associated with the particle. Which particular χ one assigns to it depends on how the particle was placed in the well and what is done to the particle afterward. If we leave the particle alone, it will, following the tendency of all physical systems, tend to go to the lowest energy state available, which is called the *ground state* or $E = E_0$ (provided, as we will see later in Chapter 23, that that state is not already occupied by another particle). In the present case of lowest energy, the eigenfunction representing the x-position of the particle, will be, from Eq. 20.25

$$\chi(x) = B \sin \frac{\pi}{a} x \tag{20.27}$$

We note that when the $\chi(x)$ of Eq. 20.27 is multiplied by the time part of the wavefunction, $\Gamma(t)$ of Eq. 20.16, the resulting wavefunction $\psi(x, t)$ represents a standing wave of the type discussed in Section 12.8. In fact, a plot (see Fig. 20-4) of the first four eigenfunctions ($n = 1, 2, 3, 4$), Eq. 20.25, reveals that they are identical to the allowed standing waves in a string (Fig. 12-19).

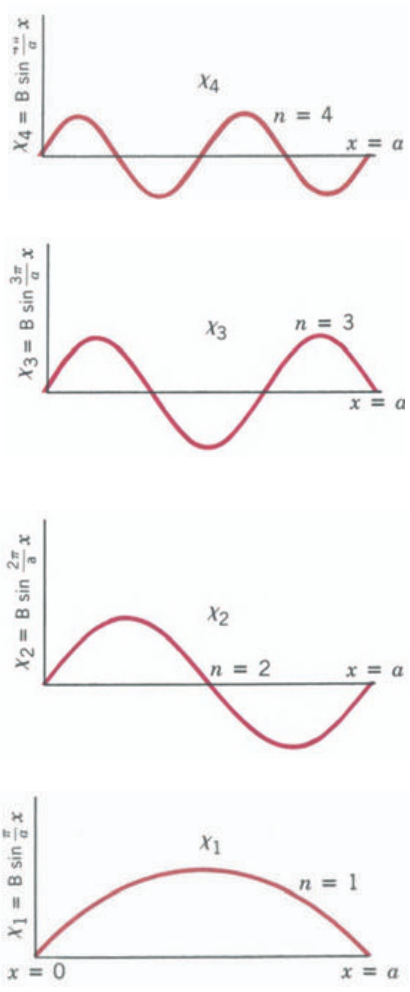


FIGURE 20-4
Plot of the first four eigenfunctions for the particle in the infinite potential well.

Example 20-1 Normalization of the Wavefunction

The arbitrary constant B in Eq. 20.25 is determined by the normalization condition; that is, the probability of finding the particle *somewhere* in space must be 1. In mathematical terms this fact is stated as follows:

$$\int_{-\infty}^{\infty} \psi^* \psi \, dx = 1$$

Evaluate the constant B for the ground state wavefunction of a particle in a one-dimensional well, that is, for the wavefunction

$$\psi_1(x, t) = \chi_1(x)\Gamma(t) = B \sin\left(\frac{\pi x}{a}\right) \exp\left(-\frac{iE_0}{\hbar} t\right)$$

Solution For the particle in the well the limits of integration must be from 0 to a because the probability of finding the particle in the well must be unity. Therefore,

$$\int_0^a B^2 \sin^2\left(\frac{\pi x}{a}\right) \exp\left(\frac{iE_0}{\hbar} t\right) \exp\left(-\frac{iE_0}{\hbar} t\right) dx = 1$$

$$\int_0^a B^2 \sin^2\left(\frac{\pi x}{a}\right) dx = 1$$

$$\frac{a}{\pi} \int_0^a B^2 \sin^2\left(\frac{\pi x}{a}\right) \left(\frac{\pi}{a} dx\right) = 1$$

The integral can be found in standard integration tables. From them, we get

$$\frac{a}{\pi} B^2 \left\{\frac{\pi x}{2a} - \frac{1}{4} \sin\left(\frac{2\pi x}{a}\right)\right\}_0^a = 1$$

Substituting the limits, we obtain

$$\frac{a}{\pi} B^2 \left(\frac{\pi}{2}\right) = 1$$

or

$$B = \left(\frac{2}{a}\right)^{1/2}$$

Example 20-2 Probability Calculations

We indicated previously that all the information about the particle can be found from the wavefunction. Consider a particle in the ground state, that is, one represented by the wavefunction of Example 20-1. What is (a) the average position, (b) the average momentum, and (c) the average energy of such a particle?

Solution

(a) From Eq. 20.6, the average position is given by

$$\bar{x} = \int_0^a \psi_1^* \, x \, \psi_1 \, dx$$

which in the present case becomes

$$\bar{x} = \frac{2}{a}\int_0^a \sin^2\left(\frac{\pi}{a}x\right) x\,dx$$

$$= \frac{2}{a}\frac{a^2}{\pi^2}\int_0^a \left(\sin^2\left(\frac{\pi x}{a}\right)\right)\left(\frac{\pi x}{a}\right)\left(\frac{\pi}{a}\,dx\right)$$

From the integration tables, we get

$$\bar{x} = \frac{2a}{\pi^2}\left[\frac{1}{4}\left(\frac{\pi x}{a}\right)^2 - \frac{\pi x \sin\left(\frac{2\pi x}{a}\right)}{4a} - \frac{\cos\left(\frac{2\pi x}{a}\right)}{8}\right]_0^a$$

$$= \frac{2a}{\pi^2}\left\{\left(\frac{\pi^2}{4} - 0 - \frac{1}{8}\right) - \left(0 - 0 - \frac{1}{8}\right)\right\}$$

$$\bar{x} = \frac{a}{2}$$

This result makes physical sense and would have been predicted by looking at a graph of the probability density $\psi_1^*\psi_1$, which is plotted in Fig. 20-5. It is clear that the probability of finding the particle on the left side of the well is the same as the probability of finding it on the right side. The average position is therefore the midpoint $x = a/2$, which is the result found in the preceding calculation. This result is true not only for $n = 1$ but for all n's. For example, the square of the eigenfunctions shown in Fig. 20-4 will be symmetric about the midpoint of the well.

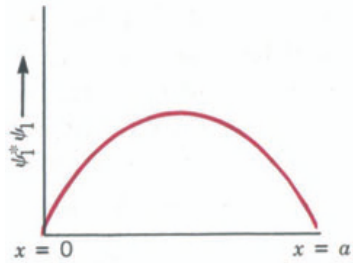


FIGURE 20-5
Example 20-2.

(b) The average value of the momentum can be found using the method outlined in Section 20.2c of this chapter, Eq. 20.8.

$$\bar{p} = \int_0^a \psi_1^*\left(-i\hbar\frac{\partial}{\partial x}\right)\psi_1\,dx$$

$$= \frac{2}{a}\int_0^a \left(\sin\frac{\pi x}{a}\right)\left(-i\hbar\frac{\partial}{\partial x}\right)\left(\sin\frac{\pi x}{a}\right)dx$$

$$= \frac{2}{a}\int_0^a \left(\sin\frac{\pi x}{a}\right)\left(-i\hbar\frac{\pi}{a}\cos\frac{\pi x}{a}\right)dx$$

$$= \frac{2}{a}(-i\hbar)\int_0^a \left(\sin\frac{\pi x}{a}\right)\left(\cos\frac{\pi x}{a}\right)\left(\frac{\pi}{a}\,dx\right)$$

$$= \frac{2}{a}(-i\hbar)\frac{\sin^2\frac{\pi x}{a}}{2}\Bigg|_0^a$$

$$\bar{p} = 0$$

This result also makes physical sense. The particle is moving back and forth between the walls of the well. The probability of finding the particle moving toward the right is the same as the probability of finding it moving toward the left. Thus the average value of the momentum has to be zero.

(c) The average value of the energy E can be calculated using Eq. 20.10

$$\begin{aligned}
\bar{E} &= \int_0^a \psi_1^* \left(i\hbar \frac{\partial}{\partial t} \right) \psi_1 \, dx \\
&= \int_0^a \psi_1^* \left(i\hbar \frac{\partial}{\partial t} \right) \chi_1(x) \exp\left(-i \frac{E_0}{\hbar} t \right) dx \\
&= \int_0^a \psi_1^* \, (i\hbar) \left(-i \frac{E_0}{\hbar} \right) \chi_1(x) \exp\left(-i \frac{E_0}{\hbar} t \right) dx \\
&= \int_0^a E_0 \, \psi_1^* \, \psi_1 \, dx \\
\bar{E} &= E_0
\end{aligned}$$

In the last step, we made use of the fact that $\int_0^a \psi_1^* \, \psi_1 \, dx = 1$. The result should come as no surprise. We indicated that when the particle is in the ground state, the eigenfunction associated with the particle was given by Eq. 20.27. In this case the particle has a well-defined energy $E = E_0$. We expect therefore that the average value will be the actual value.

PROBLEMS

20.1 Show by direct substitution into the time-dependent Schrödinger equation for the free particle, Eq. 20.3, that $\psi(x, t) = A \cos(kx - \omega t)$ is not a solution.

20.2 Show by direct substitution that the wavefunction $\psi(x, t) = A \cos kx \, e^{-i\omega t}$ satisfies the time-dependent Schrödinger equation for the free particle, Eq. 20.3.

20.3 Explain why the following eigenfunctions are not acceptable solutions of the Schrödinger equation

(a) $\chi(x) = 0 \quad$ for $x \leq 0$

$\chi(x) = A \cos kx \quad$ for $x \geq 0$

(b) $\chi(x) = A \dfrac{e^{ikx}}{x}$

(c) $\chi(x) = A \ln kx$

20.4 We saw that there are two solutions for the eigenfunction of a particle in an infinite potential well, e^{ikx} and e^{-ikx}. Show that the sum $\chi = a \, e^{ikx} + b \, e^{-ikx}$ is also a solution, where a and b are arbitrary constants.

20.5 (a) What is the minimum energy of an electron confined in a one-dimensional well of width 1 Å, that is, the approximate size of the atom? (b) How does this energy compare with the binding energy

of an electron in an atom? (c) What is the minimum energy of an electron confined in a one-dimensional well of width comparable to the nuclear diameter, that is, 10^{-14} m? (d) The binding energy of the nuclear constituents is a few MeV (10^6 eV), what can you conclude about the existence of electrons in the nucleus?

(*Answer*: (a) 37 eV, (c) 3.7×10^9 eV.)

20.6 (a) What is the minimum energy of a neutron confined in a one-dimensional well of width comparable to the nuclear diameter, that is, 10^{-14} m? (b) Is it reasonable to assume that there are neutrons in the nucleus? The binding energy of the nuclear constituents is a few MeV (10^6 eV).

20.7 (a) Under quantum conditions, what is the minimum energy that a 100-kg person can have when confined to a closet 1 m long? (b) If the person moves with a speed of 10^{-4} m/sec, what would be his quantum number n?

(*Answer*: (a) 5.4×10^{-70} J, (b) 3×10^{31}.)

20.8 Evaluate the energy of an electron inside a one-dimensional well of width 2 Å for the states $n = 1, 2, 10, 11, 100, 101, 1000, 1001$. What conclusion can you draw about the fractional difference in the energy

$$\frac{E_{n+1} - E_n}{E_n}$$

between adjacent states?

(*Answer*: $E_1 = 9.34$ eV, $E_2 = 37.47$ eV, $E_{10} = 9.34 \times 10^2$ eV, $E_{11} = 11.30 \times 10^2$ eV, $E_{100} = 9.34 \times 10^4$ eV, $E_{101} = 9.53 \times 10^4$ eV, $E_{1000} = 9.34 \times 10^6$ eV, $E_{1001} = 9.36 \times 10^6$ eV.)

20.9 For the one-dimensional well, show that

$$\lim_{n\to\infty} \frac{E_{n+1} - E_n}{E_n} = 0$$

20.10 An electron in one of the excited states of the infinite potential well can emit photons when it decays to lower energy states. Assume that only transitions in which $\Delta n = \pm 1$ are allowed. Show that the frequencies of the photons emitted in these transitions are

$$\nu = \frac{3E_0}{h}, \frac{5E_0}{h}, \frac{7E_0}{h}, \cdots \frac{(2n+1)E_0}{h},$$

$$\text{where } n = 1, 2, 3, \cdots \text{ and } E_0 = \frac{\hbar^2\pi^2}{2ma^2}$$

20.11 An electron is confined in a one-dimensional well. Assume as in Problem 20.10 that only transitions in which $\Delta n = \pm 1$ are allowed. The electron initially in the ground state absorbs a photon of frequency 2.4×10^{14} Hz. (a) What is the energy spectrum of the electron? (b) What is the width of the well?

(*Answer*: (a) $E_n = 0.33\, n^2$ eV, where $n = 1, 2, 3, \ldots$, (b) 10.6 Å.)

20.12 What is the probability of finding a particle in a well of width a at a position $a/4$ from the wall if $n = 1$, if $n = 2$, if $n = 3$. Use the normalized wavefunctions $\psi(x, t) = (2/a)^{1/2} \sin n\pi x/a\, e^{-iEt/\hbar}$.

(*Answer*: $1/a$, $2/a$, $1/a$.)

20.13 Calculate the expectation value for the coordinate x, the momentum p, and the energy E of a particle in an infinite potential well represented by the wavefunction of the first excited state, that is,

$$\psi(x, t) = (2/a)^{1/2} \sin 2\pi x/a\, e^{-iE_2t/\hbar}$$

(*Answer*: $a/2$, 0, E_2.)

20.14 An electron in an infinite potential well is in the third excited state, that is, $\chi_4 = (2/a)^{1/2} \sin 4\pi x/a$. What is the probability of finding the electron in the region between $x = 0$ and $x = a/4$? Answer the question directly from the graph of Fig. 20-4 without performing any calculations. Verify your answer by evaluating the integral

$$P\left(0 \le x \le \frac{a}{4}\right) = \int_0^{a/4} \chi_4^2\, dx$$

(*Answer*: $\frac{1}{4}$.)

20.15 As we saw in Section 20-2c, we can associate the operator $-i\hbar\, \partial/\partial x$ with the momentum p of a particle. Similarly, we can associate the operator $(-i\hbar\, \partial/\partial x)\,(-i\hbar\, \partial/\partial x) = -\hbar^2\, \partial^2/\partial x^2$ with the square of the momentum p^2. Use this to find the expectation value of p^2, that is, $\overline{p^2}$, for the lowest energy state of

the particle in the one-dimensional well. Use the normalized wavefunction of Example 20-1. (*Hint:* If you are thoughtful, there is no need to evaluate any integral.)

(*Answer*: $\left(\frac{\hbar\pi}{a}\right)^2$.)

20.16 Use the result of Problem 20.15 together with the fact that the kinetic energy of a particle is $E_k = 1/2\ mv^2 = p^2/2m$ to evaluate the expectation value of the kinetic energy of a particle in the lowest energy state of the one-dimensional infinite potential well. How does your result compare with the energy E_0 of that state?

20.17 In Problem 19.21 we saw that the standard deviation σ of a quantity is often used as a measure of the uncertainty in that quantity. The standard deviation of a quantity, for example, the momentum p is defined as $\sigma_p^2 = \overline{p^2} - \bar{p}^2$. (a) Use the result of Problem 20.15 for $\overline{p^2}$ and that of Example 20-2 for $\bar{p}$ to find the uncertainty in the momentum of the particle in the lowest energy state of the well. (b) The particle can be anywhere in the well, therefore $\Delta x = a$. Calculate $\Delta p\ \Delta x$, and compare the result with the uncertainty principle, Eq. 19.5.

(*Answer*: (a) $\hbar\pi/a$, (b) $h/2$.)

20.18 Alpha particles (He nuclei) are emitted from certain nuclei. Consider an α particle with energy 6 MeV (6×10^6 eV) in a nuclear well of width 10^{-14} m (see Fig. 20-6). How many collisions per second does it make with the walls? The mass of an α particle is $4 \times 1.67 \times 10^{-27}$ kg.

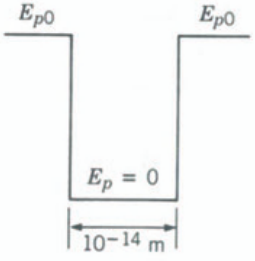


FIGURE 20-6
Problem 20.18.

20.19 The two-dimensional Schrödinger equation is

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\chi}{\partial x^2} + \frac{\partial^2\chi}{\partial y^2}\right) + E_p\chi = E\chi$$

Solve the Schrödinger equation for a particle confined in a two-dimensional infinite potential well of width and length a. In particular, show that the eigenfunctions χ and the corresponding eigenvalues E are given by

$$\chi_{n_1n_2} = A \sin\left(\frac{n_1\pi x}{a}\right) \sin\left(\frac{n_2\pi y}{a}\right)$$

$$E_{n_1n_2} = \frac{\pi^2\hbar^2}{2ma^2}(n_1^2 + n_2^2)$$

where $n_1 = 1, 2, 3, \ldots$ and $n_2 = 1, 2, 3, \ldots$

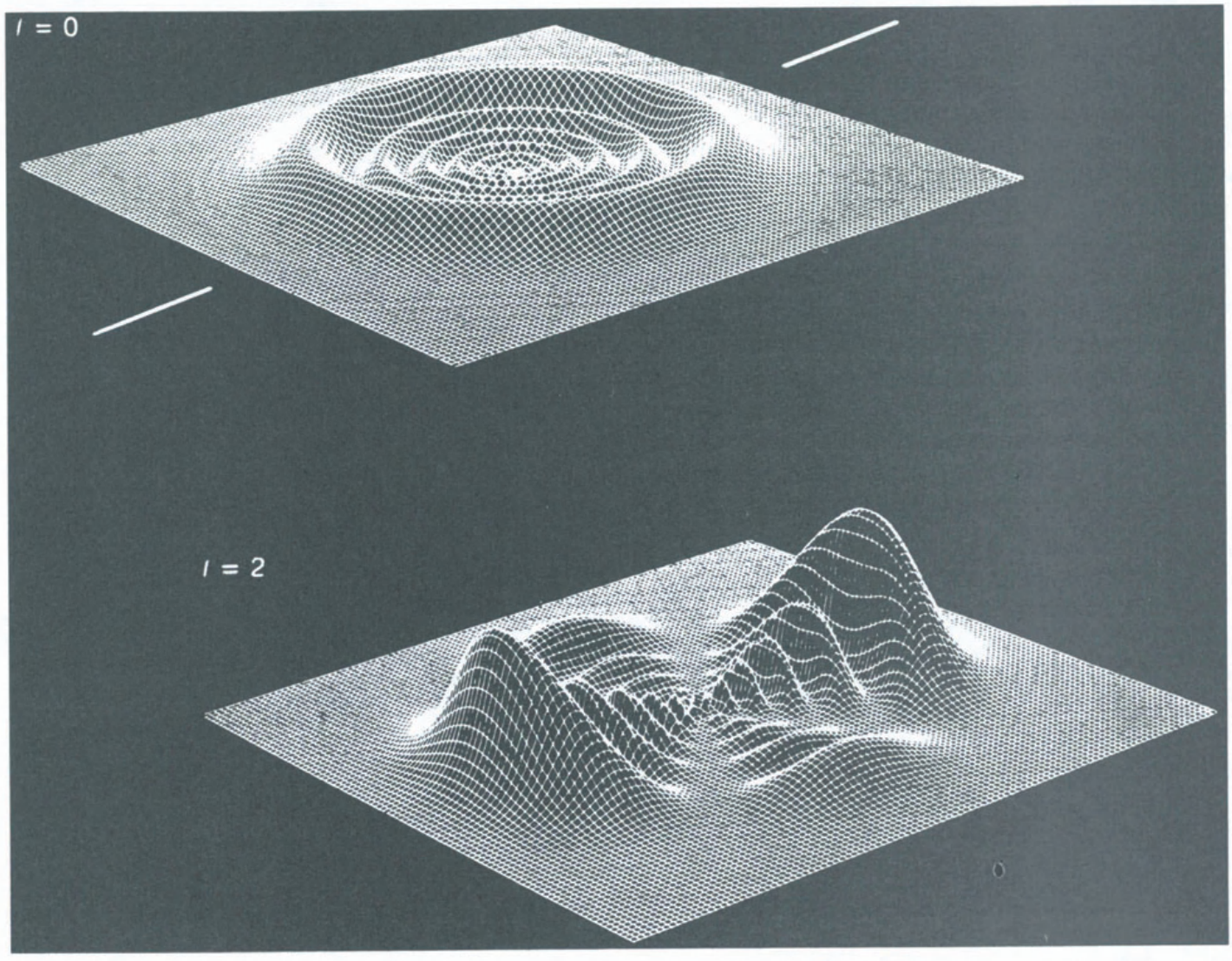
(*Hint:* Assume $\chi(x, y) = \chi_1(x)\ \chi_2(y)$, and then use the separation of variables method discussed in Section 20-2d.)

20.20 As we have seen in Chapter 19, one of the consequences of the uncertainty principle is that we cannot ascribe to a subatomic particle, such as an electron, a well-defined path. We can only talk about the spacial probability density $|\psi|^2 = \psi^*\psi$. If the particle is a charged particle, then we can associate with its probability density a spacial charge density $q|\psi|^2$, where q is the charge of the particle. Consider an electron in an infinite potential well. Let the electron be in an eigenstate χ_n. (a) Calculate the spacial charge density associated with that particle. (b) What conclusion do you draw about the time dependence of that charge density? (c) What is the significance of this result in connection with Bohr's second postulate about the hydrogen atom?

(*Answer*: (a) $e\chi_n^2(x)$.)

20.21 When an electron is making a transition from some initial state $\psi_i(x, t)$ to some final state $\psi_f(x, t)$, one cannot assert with certainty whether or not the transition has already taken place. As a consequence, the wavefunction describing the electron must be a mixture of the initial and final state wavefunctions, that is, $\psi(x, t) = A\psi_i(x, t) + B\psi_f(x, t)$, where A and B are constants. Calculate the spatial charge density associated with such an electron. What is the physical significance of the result in connection with Bohr's third postulate about the hydrogen atom?

(*Answer*: $e[A^2\chi_i^2(x) + B^2\chi_f^2(x)] + e\left[2\ AB\ \chi_i(x)\chi_f(x) \cos\left(\frac{E_i - E_f}{\hbar}\right) t\right]$.)



CHAPTER 21

Quantum Mechanics of Atoms

Probability density (vertical axis) of the electron in the hydrogen atom for two angular momentum states. Both graphs are for $n = 8$, $m_l = 0$. The nucleus is at the center of the graphs and the white lines correspond to the z-axis.

21.1 INTRODUCTION

Thus far, we have applied the Schrödinger theory to an ideal, one-dimensional case. One of the main features of the theory is that the quantization of the energy does not come in as a postulate but instead it comes naturally from the solution of the Schrödinger equation and the physically justified requirements that the wavefunction be well-behaved. The great success of the Schrödinger theory, which is a postulate, but a fundamental one, is that one is able to derive from mathematical principles the postulates that had originally been presented to explain the classically unexplainable experimental results. Thus, if one solves the Schrödinger equation for a particle acted on by a linear restoring force $F = -kx$, that is, for a classical harmonic oscillator with potential energy $E_p = \frac{1}{2} kx^2$, such as a mass attached to a spring, the only physically acceptable solutions are those for which the energy E has values given by

$$E_n = (n + \tfrac{1}{2})\, h\nu \quad \text{where } n = 0, 1, 2, 3, \ldots$$

This result leads to the same conclusion as that of Planck concerning the energy of the electromagnetic spectrum produced by such oscillators, namely, $E_{\text{waves}} = nh\nu$, and therefore the Schrödinger theory explains the spectrum of a blackbody without the need for a postulate. Note that the solution of the spectrum given here has a factor of $\frac{1}{2}$ in it. This means that in the lowest energy state $n = 0$ a system still has vibrational energy. Therefore, even at a temperature of absolute zero, electrons, atoms, and such, will still vibrate.

Similarly, when one solves the Schrödinger equation for an electron in the potential of a proton (a hydrogen atom), one again finds that the only physically acceptable solutions for the wavefunctions are those where the energy E takes certain discrete values. The values are those postulated by Bohr, which in turn led to the explanation of the hydrogen spectrum. The Schrödinger theory provides much more information than only the quantization of the energy spectrum. It explains why certain spectral lines are brighter than others; that is, it can be used to calculate transition rates. It explains some fine details about the spectrum; that is, some of the spectral lines under a high resolution spectrometer turn out to be two or more very closely spaced lines (two or more different wavelengths differing by a few angstroms).

The mathematical complexities do not permit us to solve in detail the Schrödinger equation for the coulomb potential of the hydrogen atom. We will, however, outline the solution of the problem and present the results that are obtained by solving the Schrödinger equation. We will then discuss the physical significance of these results and use these results together with a very important principle of quantum mechanics, the *exclusion principle*, to gain some understanding of the ground state of multielectron atoms. Then we will be ready to apply the quantum theory to solids.

21.2 OUTLINE OF THE SOLUTION OF THE SCHRÖDINGER EQUATION FOR THE H ATOM

We saw in connection with the Bohr model of the hydrogen atom, Eq. 18.1, that the potential energy of the electron in the electric field of a nucleus is

$$E_p = -\frac{1}{4\pi\epsilon_0}\frac{e^2}{r}$$

where $r = (x^2 + y^2 + z^2)^{1/2}$ is the distance between the electron and the proton. If the nucleus has more than a single positive charge we must include the additional charges in the Coulomb energy. We do this by a multiplicative term Z, called the *atomic number*, that represents the number of protons in the nucleus and also the number of electrons in neutral atoms. Thus the charge part of the Coulomb equation for a single electron becomes eZe or Ze^2, and the above energy equation, which is for $Z = 1$, is written as

$$E_p = -\frac{1}{4\pi\epsilon_0}\frac{Ze^2}{r} \tag{21.1}$$

We extend the time-independent Schrödinger equation (Eq. 20.17) to a situation where a particle is free to move in three dimensions. To do this we simply include derivatives of the space part of the wavefunction, χ with respect to the other two coordinates, y and z, namely,

$$-\frac{\hbar^2}{2m}\left[\frac{\partial^2\chi}{\partial x^2} + \frac{\partial^2\chi}{\partial y^2} + \frac{\partial^2\chi}{\partial z^2}\right] + E_p\,(x, y, z)\,\chi = E\,\chi$$

This is a partial differential equation and, because E_p is a function of x, y and z, it is difficult to solve in the form shown here. However, because E_p is a function of the separation between the protons and the electron only, it is much simpler to solve the equation if we write it in spherical coordinates (see Fig. 21-1). In spherical coordinates the Schrödinger equation becomes

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\chi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\chi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\chi}{\partial\phi^2} + \frac{2m}{\hbar^2}[E - E_p(r)]\,\chi = 0 \tag{21.2}$$

This is an awesome looking equation, but one that can be solved by the standard technique of separation of variables; we try a solution of the form

$$\chi(r,\theta,\phi) = R(r)\,\Theta(\theta)\,\Phi(\phi)$$

where R, Θ, and Φ are functions of only one coordinate. On substitution into Eq. 21.2, rearrangement of terms, and division of both sides of the resulting

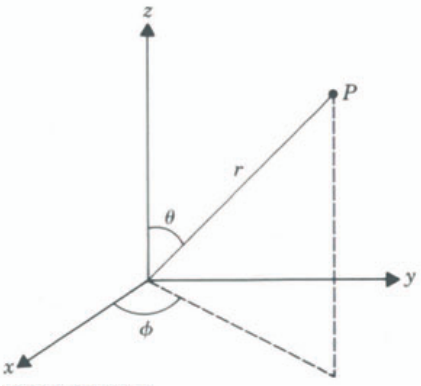


FIGURE 21-1
Spherical coordinates r, θ, and ϕ of a point P.

equation by $R(r)$, $\Theta(\theta)$, and $\Phi(\phi)$, we obtain

$$-\frac{\sin^2\theta}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)-\frac{2m}{\hbar^2}r^2\sin^2\theta\,[E-E_p(r)]$$

$$-\frac{\sin\theta}{\Theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)=\frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2} \tag{21.3}$$

The left side of Eq. 21.3 is a function of r and θ alone, whereas the right side is a function of ϕ alone. The only way the equation can be valid is when both sides of the equation are equal to the same constant. For convenience, we let this constant be $-m_l^2$, where this m_l is not to be confused with the m used for mass. The right side of the equation gives us an ordinary differential equation for Φ

$$\frac{1}{\Phi}\frac{d^2\Phi}{d\phi^2}=-m_l^2 \tag{21.4}$$

The constant is chosen as $-m_l^2$ in order that m_l will be an integer, as will be indicated later. The left side of Eq. 21.3, after being set equal to the constant $-m_l^2$ and rearranged, may be written as

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)+\frac{2mr^2}{\hbar^2}[E-E_p(r)]$$

$$=\frac{m_l^2}{\sin^2\theta}-\frac{1}{\Theta}\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)$$

Again, one side of this equation is a function of the variable r only, whereas the other side is a function of the variable θ only. The only way the equality can be valid is when both sides are equal to the same constant. If we call this constant $l\,(l+1)$, l will be an integer. We now get two differential equations, one for $R(r)$ and another for $\Theta(\theta)$.

$$\frac{m_l^2}{\sin^2\theta}-\frac{1}{\Theta\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)=l\,(l+1) \tag{21.5}$$

and

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)+\frac{2mr^2}{\hbar^2}[E-E_p(r)]=l\,(l+1) \tag{21.6}$$

The next step is to solve the three differential equations (21.4, 21.5, and 21.6). These three equations and their solutions are well known in mathematics. We will simply mention what happens when one solves them subject to the constraints on the wavefunctions discussed in Chapter 20.

1. When we solve Eq. 21.4, we find that the only solutions for Φ that are *single-valued* are those for which (see Problems 21.1 and 21.2)

$$m_l=0,\ \pm1,\ \pm2,\ \ldots \tag{21.7}$$

2. When we solve Eq. 21.5 for Θ, we find that the only solutions that are finite everywhere are those for which

$$l = 0, 1, 2, \ldots \tag{21.8}$$

and

$$l \geq |m_l|$$

3. Finally, when we solve Eq. 21.6, we find that the only solutions for R that remain finite everywhere are those for which

$$E_n = -\frac{Z^2 e^4 m}{8\epsilon_0^2 h^2}\frac{1}{n^2} \quad n = 1, 2, 3, \ldots \tag{21.9}$$

and

$$l < n$$

The restrictions on the values that m_l and l can take can now be restated as follows:

$$m_l = 0, \pm 1, \pm 2, \ldots, \pm l \tag{21.7a}$$

and

$$l = 0, 1, 2, \ldots, n-1 \tag{21.8a}$$

21.3 PHYSICAL SIGNIFICANCE OF THE RESULTS

The most important result of the solution of the Schrödinger equation for the H atom is the fact that the energy of the atom is quantized (Eq. 21.9). The spectrum of energy levels is the same as the one postulated by Bohr. Because the energy depends only on the quantum number n, it is called the *principal quantum number.*

The spacial variation of χ depends, however, on the three quantum numbers n, l, m_l, which arise from the integers of Eqs. 21.7*a*, 21.8*a*, and 21.9, and the wavefunction is written with the quantum numbers as subscripts, χ_{nlm_l}. Because for a given n the other two numbers can take several values, this means that it is possible for the electron to have quite different characteristics while maintaining the same energy. States χ having the same energy but different values for the quantum numbers l and m_l are called *degenerate states.* The degree of degeneracy depends on the value of n. For example, from Eqs. 21.7a and 21.8a, if $n = 1$, $l = 0$, and $m_l = 0$; we have only one state χ_{100}. If $n = 2$, $l = 0$ or 1; for $l = 0$, $m_l = 0$; for $l = 1$, $m_l = 0$, 1, or -1. Thus for $n = 2$, we have a fourfold degeneracy with wavefunctions χ_{200}, χ_{210}, χ_{211}, and χ_{21-1}, all having the same energy. This degeneracy is often represented (partially) in a two-dimensional diagram (see Fig. 21-2). The vertical axis shows the quantum number n. Equation 21.9 shows that as n increases E becomes less negative and therefore larger. The horizontal axis is the quantum number l.

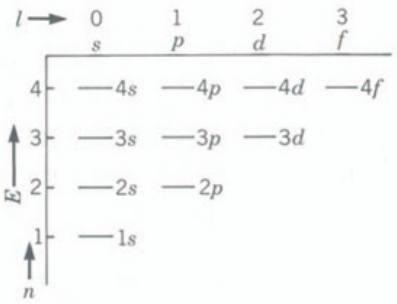


FIGURE 21-2
Partial representation of the degeneracy of the eigenfunctions in the hydrogen atom. The energy E is determined by the quantum number n alone. The spacial variation of the eigenfunctions depends on the three quantum numbers n, l, m_l. States with the same n but different l or m_l have the same energy and are said to be degenerate states. The degeneracy with respect to m_l is not shown in the figure.

States for which $l = 0$ are called s states, those with $l = 1$ are called p states, those with $l = 2$ are called d states, those with $l = 3$ are called f states, following the initials of the original names given by spectroscopists to groups of spectral lines; *sharp, principal, diffuse, fundamental.*

Although further details of calculations will not be shown, certain results of these calculations are important and will be listed and discussed.

1. The quantum number l is called the *orbital quantum number,* because l determines the magnitude of the angular momentum **L** of the atom.
2. The quantum number m_l is called the *magnetic quantum number,* because m_l determines the orientation of the angular momentum **L** in a magnetic field. It can be shown that if an atom is placed in a magnetic field directed along the z direction, the z component of the angular momentum **L** of the atom is given by

$$L_z = m_l \hbar \tag{21.10}$$

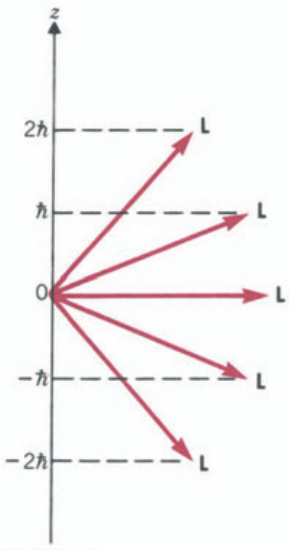


FIGURE 21-3
Possible orientations of the angular momentum L of the electron in the hydrogen atom for the case where the orbital quantum number $l = 2$.

To have a feeling of what we mean by the orientation of **L**, we must remember that in order to specify the state of rotation of an object, it is not enough to say how fast the object is rotating, we must also specify the plane of rotation, that is, we must give a direction to **L**. The convention is that the vector **L** is perpendicular to the plane of rotation, and the sense is given by the right-hand rule, Eq. 8.25. The result of Eq. 21.10 tells us that in an atom, **L** cannot have any arbitrary orientation with respect to the z axis (as would be the case classically), but rather it can have only certain discrete orientations. This is known as *space quantization.* Suppose $l = 2$, then m_l can be 2, 1, 0, -1, -2. Thus $L_z = 2\hbar, \hbar, 0, -\hbar, -2\hbar$. See Fig. 21-3.

So long as there is no preferred direction, space quantization is meaningless. If, however, you perform an experiment in which a particular direction becomes preferred, such as by the application of a magnetic field, then direction becomes meaningful. If the field tends to align the angular momentum of the atom along a given direction, then that direction becomes the z axis.

Is this prediction of the Schrödinger theory concerning space quantization borne out by the experiment? We will now examine the behavior of an atom in a magnetic field.

$L_z = m_l \hbar$

21.4 SPACE QUANTIZATION: THE EXPERIMENTS

21.4a The Zeeman Effect

We know that if a current i flows through a loop of area A, a magnetic dipole moment $\boldsymbol{\mu} = iA$ is associated with it (see Eq. 16.10 et. seq.) If we place this dipole in an external uniform magnetic field **B** (see Fig. 21-4), it will experience a torque $\boldsymbol{\tau} = \boldsymbol{\mu} \times \mathbf{B} = \mu B \sin\theta$, where θ is the angle between the two vectors, Eq. 16.12. The torque is a maximum when the dipole is perpendicular

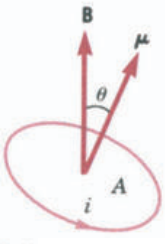


FIGURE 21-4
A magnetic dipole moment $\boldsymbol{\mu} = iA$ in a magnetic field **B**.

to the field and zero when it is parallel or antiparallel to it. As a consequence, we can associate with such a dipole $\boldsymbol{\mu}$ in a magnetic field $\mathbf{B}$ a potential energy that will depend on the orientation of $\boldsymbol{\mu}$ with respect to $\mathbf{B}$. The potential energy E_p is,

$$E_p = -\boldsymbol{\mu} \cdot \mathbf{B} = \mu B \cos\theta \tag{16.14}$$

Let us consider an electron in one of the Bohr orbits (see Fig. 21-5); the circulating electron represents a current i given by

$$i = \frac{\text{charge}}{\text{time}} = \frac{|e|}{T} = \frac{|e|}{\frac{2\pi r}{v}} = \frac{|e|v}{2\pi r}$$

where T is the period of rotation, v is the tangential speed of the electron, and r is the radius of the orbit. Then

$$\mu = Ai = \pi r^2 i = \pi r^2 \frac{|e|v}{2\pi r} = \frac{|e|rv}{2}$$

In quantum mechanics we cannot talk about a particle moving in a circular loop of radius r with a velocity v. We can rewrite this expression in a form that is meaningful quantum mechanically by multiplying the numerator and the denominator by the mass of the electron and recalling that mvr is the angular momentum L of a particle of mass m rotating with a tangential velocity v about a point a distance r away.

$$\mu = \frac{|e|rv}{2} = \frac{|e|mvr}{2m} = \frac{|e|}{2m} L$$

$$\boldsymbol{\mu}_l = -\frac{|e|}{2m} \mathbf{L} \tag{21.11}$$

We have included a minus sign because the direction of i is opposite to the direction of motion of the electron, hence of $\mathbf{L}$ (see Fig. 21-6); and we have written $\boldsymbol{\mu}$ with the subscript l to indicate that the dipole is associated with the orbital motion of the electron. Thus, we have established that an atom with angular momentum $\mathbf{L}$ has associated with it a magnetic dipole moment. If we now place the atom in a magnetic field $\mathbf{B}$, there will be an extra contribution to the energy given by Eq. 16.14

$$E_p = -\boldsymbol{\mu}_l \cdot \mathbf{B} = \frac{|e|}{2m} \mathbf{L} \cdot \mathbf{B} = \frac{|e|}{2m} BL\cos\theta$$

(Note that this is not the Coulomb potential energy term of Eq. 21.1.) From Fig. 21-6, $L\cos\theta = L_z$, therefore

$$E_p = \frac{|e|}{2m} BL_z \tag{21.12}$$

The total energy of the atom, E_{Total}, will be the sum of the energy resulting from the interaction of the electron with the nucleus, that is, E_n of Eq. 21.9,

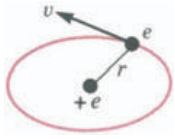


FIGURE 21-5
An electron orbitting the nucleus in one of the Bohr orbits.

$$\boldsymbol{\mu}_l = -\frac{|e|}{2m} \mathbf{L}$$

$$E_p = \frac{|e|}{2m} BL_z$$

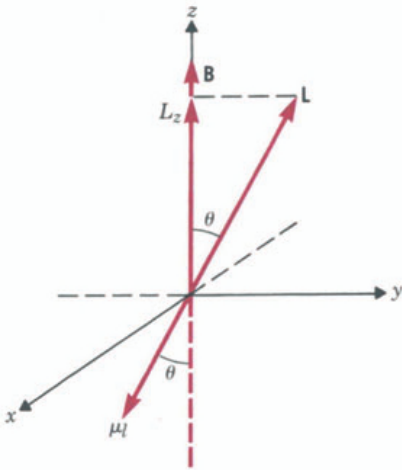


FIGURE 21-6
Relative orientation between the angular momentum of the electron $\mathbf{L}$ and the associated magnetic dipole moment $\boldsymbol{\mu}_l$.

and the energy resulting from the interaction of the dipole moment with the magnetic field B, that is, E_p of Eq. 21.12.

$$E_{\text{Total}} = E_n + E_p \tag{21.13}$$

Classically, L_z can take any value between L and $-L$. This means that an energy level E_n would become a band of width 2 ($|e|/2m$) (BL) (see Fig. 21-7). Quantum mechanically, however, L_z can take only certain values, that is, $L_z = m_l\hbar$. Thus the total energy will be given by

$$E_{\text{Total}} = E_n + \frac{|e|}{2m} B\hbar\, m_l \tag{21.14}$$

Because m_l takes values 0, ± 1, ± 2, . . . $\pm l$, the energy levels of quantum states with orbital quantum number l larger than 0 and thus, with several values for m_l, split into several discrete sublevels (see Fig. 21-8). If we now reexamine the spectrum of hydrogen (Eq. 21.9, Fig. 21-2), the $n = 2$ to $n = 1$ transition should give rise to three different lines (three different λ's). In the transitions from higher n states, we may expect even a greater number of lines. However, there is a selection rule that initially was found experimentally and was later derived from the Schrödinger theory which states that the only allowed transitions are those for which $\Delta m_l = 0$ or ± 1. Thus, regardless of the number of sublevels, transitions from a given n to another should give rise to only three distinct lines. This effect is known as the *normal Zeeman effect*. It had been observed experimentally in the spectrum of such elements as helium, calcium, zinc, and a few others long before quantum mechanics was formulated. Often more than three lines are observed. For example, in hydrogen the $n = 2$ to $n = 1$ transition shows seven closely spaced lines. This effect is known as the *anomalous Zeeman effect*. Thus, the experiment shows that the idea of space quantization is correct. However, it also shows that the Schrödinger theory as originally formulated was incomplete. In Section 21.5 we will show what was necessary to complete the theory.

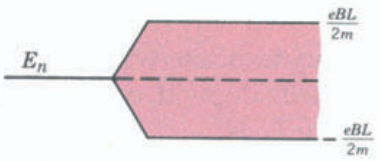


FIGURE 21-7
Classically, the energy levels of the electron in the atom should become energy bands of width $2\ eBL/2m$ when placed in an external magnetic field B.

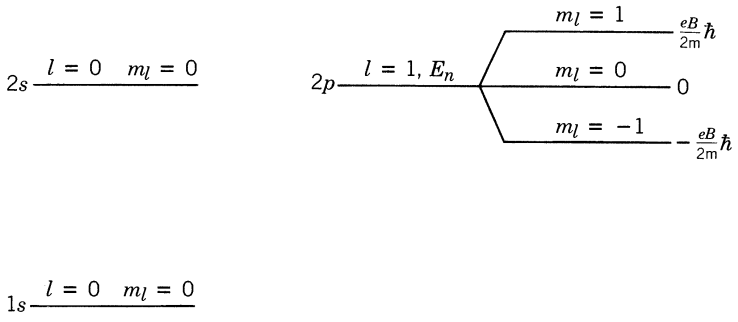


FIGURE 21-8
Schematic representation of the splitting of the $2p$ state of the hydrogen atom into three Zeeman energy levels corresponding to the three possible orientations (three possible values of m_l, 0, 1, -1) of the angular momentum **L** with respect to the external magnetic field **B**. Note that the two s states are not affected by the magnetic field because for those states $m_l = 0$.

Example 21-1

Calculate the normal Zeeman splitting of the calcium 4226 Å line when the atoms are placed in a magnetic field of 1.2 T (tesla).

Solution From Eq. 21.14, the energy between adjacent Zeeman levels is given by

$$\Delta E = \frac{|e|}{2m} B\hbar$$

$$= \frac{1.6 \times 10^{-19}\ \text{C} \times 1.2\ \text{T} \times 1.05 \times 10^{-34}\ \text{J-sec}}{2 \times 9.1 \times 10^{-31}\ \text{kg}}$$

$$= 1.11 \times 10^{-23}\ \text{J} = 6.92 \times 10^{-5}\ \text{eV}$$

Pieter Zeeman (1865–1943).

The relation between the energy difference of two levels and the associated difference in the wavelength of the photons emitted as a result of transitions from those levels to the same final state can be found by using Einstein's relation: $E_{\text{photon}} = h\nu = hc/\lambda$. Differentiating this expression with respect to the wavelength, we obtain

$$dE = -\frac{hc}{\lambda^2} d\lambda$$

or

$$|dE| = \frac{hc}{\lambda^2} |d\lambda|$$

Because the shift in energy, ΔE, is small compared with the energy of the level, we can make the approximation $dE \approx \Delta E$. Thus,

$$|\Delta\lambda| = \frac{\lambda^2}{hc} |\Delta E|$$

$$= \frac{(4.226 \times 10^{-7}\ \text{m})^2 \times 1.11 \times 10^{-23}\ \text{J}}{6.63 \times 10^{-34}\ \text{J-sec} \times 3 \times 10^{8}\ \text{m/sec}} = 9.96 \times 10^{-12}\ \text{m} = 0.0996\ \text{Å}$$

21.4b Stern-Gerlach Experiment

Another experiment that shows the idea of space quantization of the Schrödinger theory as being correct although incomplete is the Stern-Gerlach experiment.

We know that when a magnetic dipole is placed in a *uniform* magnetic field **B** it experiences a torque. However, if the field **B** is not uniform, the dipole will also experience a net force. Figure 21-9 is a schematic of a nonuniform magnetic field where the B arrows represent separate parts of the magnetic field where the vector sum at any point is the total magnetic field at that point. Because the density of B lines represents the strength of **B**, it is clear from the figure that **B** increases in the z direction; thus there is gradient $d\mathbf{B}/dz$ along the z axis. It is also clear from the picture that a net force directed upward will act on a magnetic dipole placed in the field. In general, it can be shown that if there is a magnetic gradient along the z axis the force will be directed along the z axis and the magnitude of the force will depend on

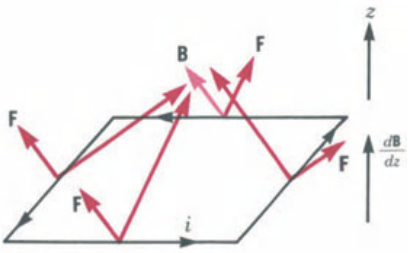


FIGURE 21-9

Forces on a current-carrying loop (a magnetic dipole) placed in an inhomogeneous magnetic field with a gradient dB/dz in the z direction.

the magnitude of the gradient and on the orientation of $\boldsymbol{\mu}_l$ with respect to the gradient. The actual value of the force is

$$F_z = \boldsymbol{\mu}_l \cdot \frac{d\mathbf{B}}{dz}$$

$$F_z = \mu_l \frac{dB}{dz} \cos\theta$$

where θ is the angle between $\boldsymbol{\mu}_l$ and the gradient $d\mathbf{B}/dz$. From Fig. 21-10, $\mu_l \cos\theta = \mu_{lz}$

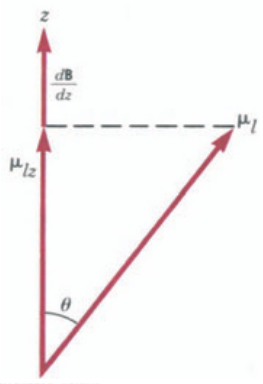


FIGURE 21-10

$$F_z = \mu_{lz} \frac{dB}{dz}$$

From Eq. 21.11 we obtain

$$F_z = -\frac{|e|}{2m} L_z \frac{dB}{dz} \tag{21.15}$$

Classically, L_z can have any value between L and $-L$; that is, if you send a beam of magnetic dipoles μ_l through such an inhomogeneous magnetic field $B(z)$, they will experience a force that can take any value between

$$\frac{|e|}{2m} L \frac{dB}{dz} \quad \text{and} \quad -\frac{|e|}{2m} L \frac{dB}{dz}$$

In the experiment, Stern and Gerlach sent a beam of neutral silver atoms from a heated oven through a collimator and then through a region with a nonuniform magnetic field $B(z)$ with a gradient along the z axis (see Fig. 21-11). The silver atoms were collected on a screen. Although the atoms were uncharged, they did possess a magnetic dipole. Therefore, the only force acting on them was the one resulting from the inhomogeneity of the magnetic field. Depending on the magnitude of this force (that is, on the orientation between $\boldsymbol{\mu}_l$ and $d\mathbf{B}/dz$), the atomic magnetic dipoles will be deflected as they travel through the magnet. The stronger the force, the greater the deflection, and therefore the farther away from the center they would land on the screen.

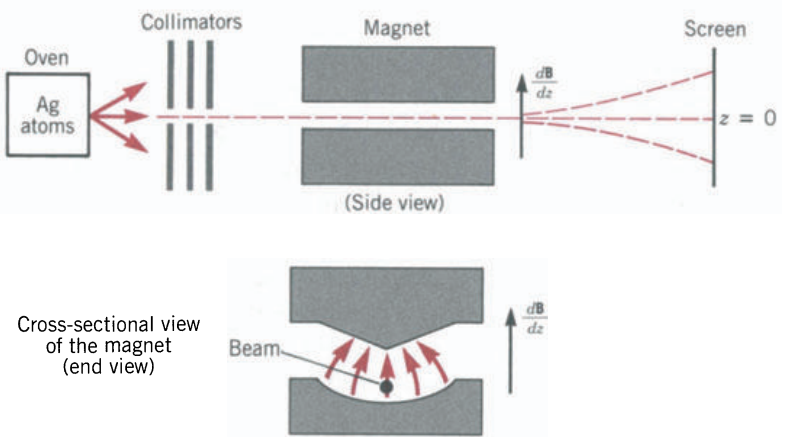


FIGURE 21-11
Schematic of the experimental setup used in the Stern-Gerlach experiment.

Because from classical physics the force F varies continuously between the two values mentioned earlier, one would expect that the distribution of atoms on the screen would be more or less uniform between some z_{max} and $-z_{max}$; that is, we would expect the distribution to be of the form shown in Fig. 21-12. On the other hand, if the Schrödinger theory is correct, and L_z is quantized, then the force in the z direction, F_z, can have only a discrete number of values, one for each m_l. We would therefore expect that the atoms would land at some discrete distances z from the center of the screen. The experimental results for silver (Ag) atoms were as shown in Fig. 21-13. Because all atoms landed at a given distance z above zero or at the equivalent distance $-z$ below zero, this result shows that the force F_z was either plus or minus a given value and that L_z had only two possible values. The experiment confirmed the concept of space quantization. This experiment was repeated by Phipps and Taylor with hydrogen atoms, and the results were identical in form to those shown in Fig. 21-13.

FIGURE 21-12
Expected distribution (according to classical principles) of the silver atoms as they land on the screen of the experiment in Fig. 21-11. The $z = 0$ position corresponds to atoms that go through the magnet undeflected.

Although the experimental results showed splitting, which demonstrates space quantization, they were in quantitative disagreement with the Schrödinger theory. Because the magnetic quantum number values are $m_l = 0, \pm 1, \pm 2, \ldots \pm l$, there will be $(2l + 1)$ possible values, and because l is an integer, $2l + 1$ is an odd integer. Therefore, *m_l should take an odd number of values, and one of them should be 0*. Because m_l determines L_z and hence F_z, the same statement applies to F_z. In fact, in the case of hydrogen the expected experimental result is obvious: one line at $z = 0$. The reason is that at the temperature of a few hundred degrees, all the atoms should be in the ground state, which is called the 1*s* state (characterized by the quantum numbers $n = 1$ and $l = 0$). In this state $l = 0$, and therefore $m_l = 0$ and $\mu_{lz} = 0$. However, because the Phipps-Taylor experiment on hydrogen showed two well-defined z spacings and no value at $z = 0$, it is evident that *there is a magnetic dipole other than the orbital dipole* that has been overlooked. We will next consider this new modification.

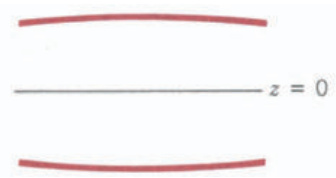


FIGURE 21-13
Experimentally observed positions of the landings of silver atoms on the screen in the experiment of Fig. 21-11.

One final piece of experimental evidence that illustrates this oversight is the following: In the absence of an external magnetic field, the transition in hydrogen from $n = 2$ to $n = 1$ should give rise to just one line, one λ, Fig. 21-2. This transition in the absence of a magnetic field is shown in Fig. 21-14. Actually, the experimental observation of this spectral line shows two very closely spaced wavelengths that correspond to two very close frequencies, that is, energies. This phenomenon is known as the *fine structure* of a spectral line.

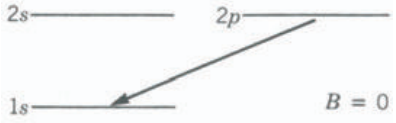


FIGURE 21-14
Fine structure. In the absence of an external magnetic field, transitions in hydrogen atoms from the state $n = 2$ to the state $n = 1$ should yield photons of a single frequency because the 2*s* and the 2*p* states are degenerate states. Two slightly different wavelengths are observed.

21.5 THE SPIN

What has the original Schrödinger theory overlooked? It was while studying the fine structure of the spectrum that G. Uhlenbeck and S. Goudsmit proposed the idea of the electron spin: *The electron has an intrinsic angular momentum called the spin* **S**. *Just as the orbital angular momentum* **L** *has a magnetic*

dipole associated with it, so does the spin; that is, the electron has a spin dipole moment $\boldsymbol{\mu}_s = -e/m\,\mathbf{S}$. By analogy with the behavior of L and in order to explain the experimental results, they postulated that the magnitude of S and its z component were quantized as follows

$$S = [s(s+1)]^{1/2}\hbar \quad \text{where } s = \tfrac{1}{2}$$

$$S_z = m_s\hbar \quad \text{where } m_s = \pm\tfrac{1}{2}$$

$S_z = m_s\hbar$
where $m_s = \pm\frac{1}{2}$

The existence of the spin and its behavior enter as a postulate in the Schrödinger theory. The reason is that this theory is nonrelativistic. In the relativistic quantum mechanical theory developed later by P. Dirac, the existence of the spin and the rules governing its behavior are a natural consequence of its formulation. The development of relativistic quantum mechanics is beyond the scope of this book.

Using the spin postulate, the experimental results concerning the anomalous Zeeman effect, the Stern-Gerlach experiment and the Phipps-Taylor experiment, as well as the fine structure, can be explained. For us the main conclusion is that *the state of an electron is now specified by four quantum numbers:* n, l, m_l, m_s. Note that the quantum number s is $\frac{1}{2}$ for all individual electrons and thus we do not need to specify it further.

Example 21-2

A beam of hydrogen atoms is used in a Stern-Gerlach type experiment. The atoms emerge from the oven with a velocity $v = 10^4$ m/sec. They enter a region 20 cm long where there is a magnetic field gradient $dB/dz = 3 \times 10^4$ T/m. The field gradient is perpendicular to the incident velocity of the atoms. The mass of the hydrogen atom is 1.67×10^{-27} kg. What is the separation of the two components of the beam as they emerge from the magnet?

Solution In the ground state, hydrogen atoms have no net *orbital* magnetic dipole moment. The only dipole moment is the one associated with the *spin* of the electron in the 1s state, that is, $\boldsymbol{\mu}_s = -|e|/m\,\mathbf{S}$, where m is the mass of the electron. From our previous discussion of the Stern-Gerlach experiment,

$$F_z = \boldsymbol{\mu}_s \cdot \frac{d\mathbf{B}}{dz} = -\frac{|e|}{m} S_z \frac{dB}{dz} = \pm\frac{1}{2}\frac{|e|}{m}\hbar\frac{dB}{dz}$$

We can use Newton's second law to find the acceleration a_z of the hydrogen atoms as they traverse the magnet.

$$a_z = \frac{F_z}{m_{\text{atom}}} = \frac{|e|\,\hbar\,\dfrac{dB}{dz}}{2\,m\,m_{\text{atom}}}$$

$$= \frac{1.60 \times 10^{-19}\ \text{C} \times 1.05 \times 10^{-34}\ \text{J-sec} \times 3 \times 10^4\ \text{T/m}}{2 \times 9.1 \times 10^{-31}\ \text{kg} \times 1.67 \times 10^{-27}\ \text{kg}}$$

$$= 1.65 \times 10^8\ \text{m/sec}^2$$

The deflection of each component in the direction of the force (z axis) will be

$$\Delta z = \frac{1}{2} a_z t^2$$

where t is the time that the atoms spend in the magnet. This time can be found by dividing the length of the magnet by the incident velocity of the atoms

$$t = \frac{0.20 \text{ m}}{10^4 \text{ m/sec}} = 2 \times 10^{-5} \text{ sec}$$

Therefore

$$\Delta z = \frac{1}{2} \times 1.65 \times 10^8 \text{ m/sec}^2 \times 4 \times 10^{-10} \text{ sec}^2$$

$$= 3.3 \times 10^{-2} \text{ m} = 3.3 \text{ cm}$$

Because of the two possible values for m_s, some atoms will be deflected upward and some downward. Therefore, the separation between the two components of the beam will be $2\ \Delta z$ or 6.6 cm.

21.6 SOME FEATURES OF THE ATOMIC WAVEFUNCTIONS

We now have almost all the information necessary to understand the important features of multielectron atoms and to proceed to the study of solids. A couple of additional facts that we need are found by looking at the wavefunctions obtained in the solution of the Schrödinger equation for the hydrogen atom.

Figure 21-15 shows simulated densities representing $|\chi|^2$ as a function of x, y, and z for the lowest energy states. One important feature is that, except for s states (that is, $l = 0$), $|\chi|^2$ does not have spherical symmetry; it has axial symmetry (a consequence of L_z having a definite value). However, if we take a subset of states made up of states having the same n and l but different m_l and we add up $|\chi|^2$ for all those states, we will find that the sum has spherical symmetry. This can be shown directly by taking the actual wavefunctions and then evaluating the sum

$$\sum_{m_l} \chi^*_{nlm_l} \chi_{nlm_l}$$

This sum is independent of the orientation angles θ and ϕ. But the result can be seen to be true by simply looking at Fig. 21-15. It turns out that the sum of all the wavefunctions of a subset of states having the same n and l is spherically symmetric not only for the hydrogen atom wavefunctions but for those of multielectron atoms as well. What this means is that if you have one electron in each of the m_l states associated with a given l, the resulting charge distribution will have spherical symmetry. This is important because it

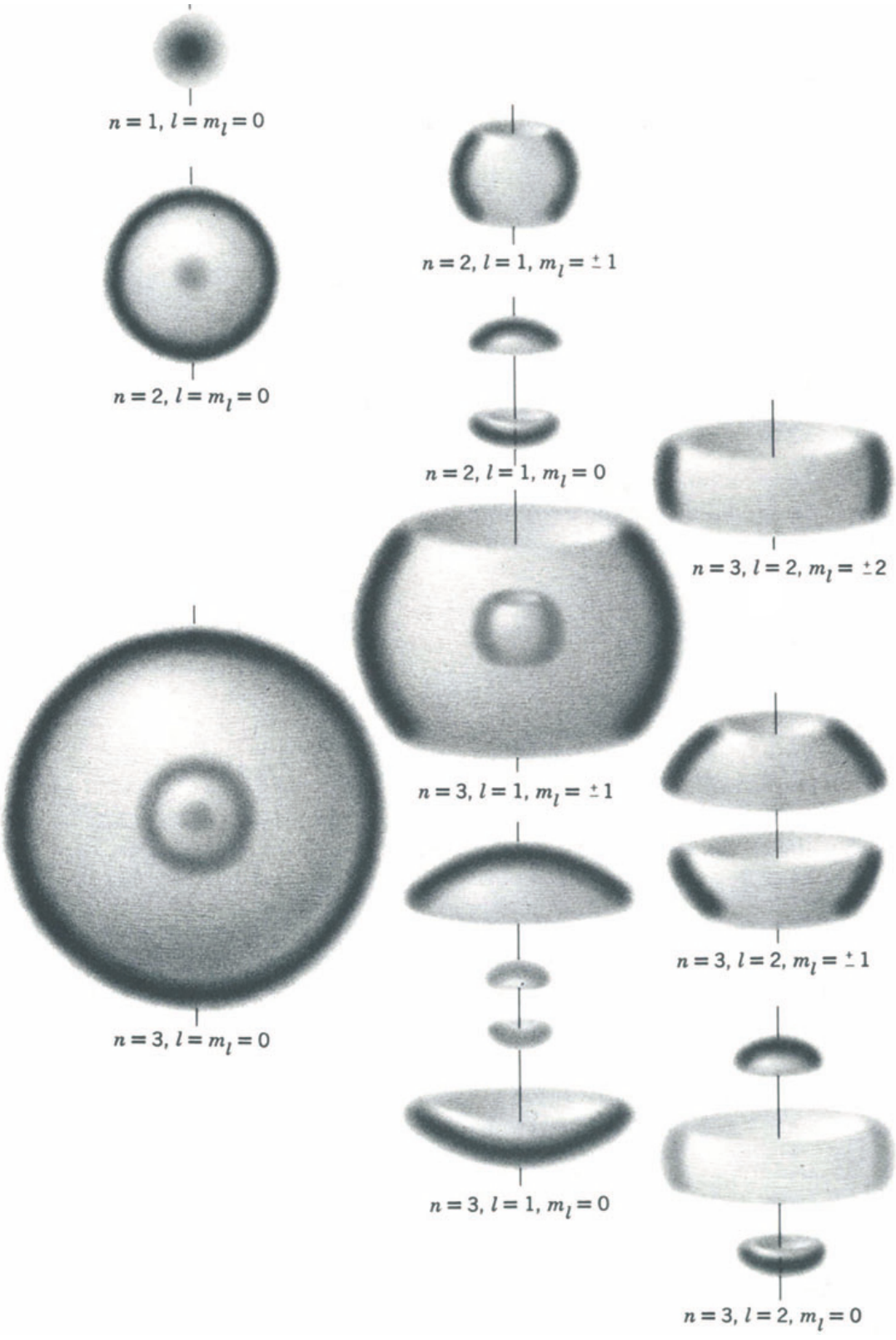


FIGURE 21-15
An artist simulation of the probability densities $|\chi|^2$ as a function of x, y, and z for the lowest energy states of the hydrogen atom. The vertical lines correspond to the z axis. (*Source*: Robert Eisberg and Robert Resnick, *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles*, 2nd ed. Copyright © 1985 by John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc).

explains partially why the rare gases helium (He), neon (Ne), argon (A), . . . are inert. It will also allow us to understand why atoms such as lithium (Li), sodium (Na), potassium (K), . . . are hydrogen-like. We will use these facts later.

Another important feature of the wavefunctions can be found by looking at Fig. 21-16. It shows the radial probability density, $P(r)$, where $P(r)\,dr$ is the probability of finding the electron between r and $r + dr$ (that is, summed over all angles). These curves are obtained by integrating the probability density per unit volume $\psi^*\psi$ over the volume enclosed by the mathematical spherical shells with radii r and $r + dr$, respectively. As can be seen, $P(r)$ varies with the distance from the nucleus r, with a maximum at certain r values that depends on the particular wavefunction. The horizontal axis is in units of r_0, where r_0 is the radius of the smallest orbit in the Bohr model. Notice that an electron in quantum state $n = 1$ has an average r value of about r_0; an electron in state $n = 2$ has an average $r \approx 4\,r_0$, and so on. Looking at it differently, an electron in state $n = 2$ is more likely to be found outside the region that would be occupied by an electron in state $n = 1$. An electron in state $n = 3$ is more likely to be found outside the region occupied by an

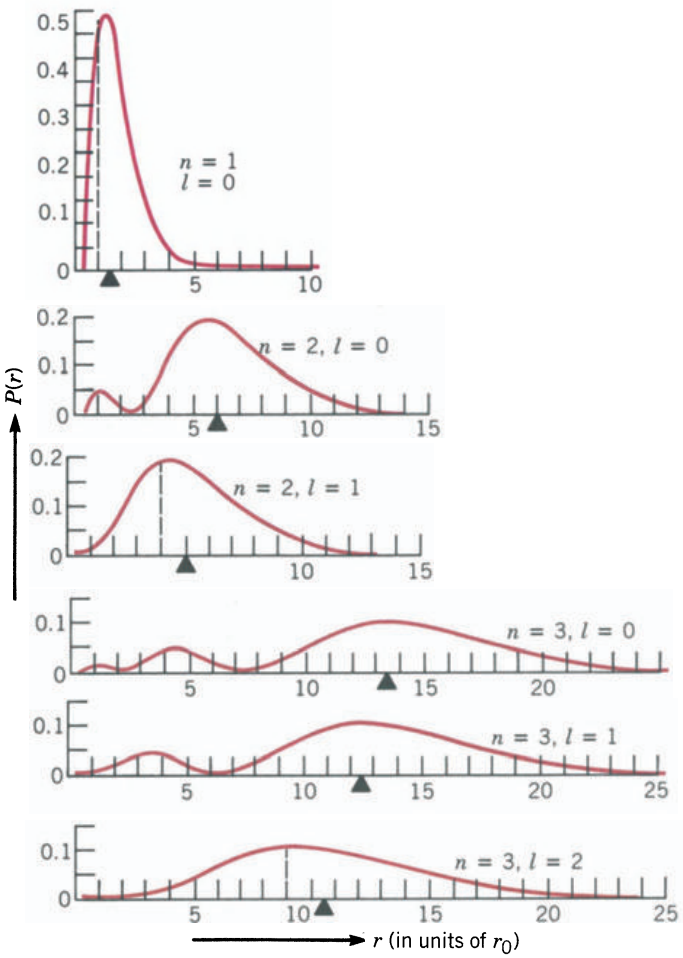


FIGURE 21-16
Radial dependence of the probability density (probability of finding the electron at a certain distance from the nucleus regardless of the angular position) versus distance from the nucleus for the lowest energy states of the hydrogen atom. The triangles on the horizontal axes indicate the average value of r for that state. (*Source*: Robert Eisberg and Robert Resnick, *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles*, 2nd ed. Copyright © 1985 by John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc).

electron in states $n = 1$ and $n = 2$, and so on. But an electron in state $n = 2$, $l = 0$ has a small but finite probability of being inside the space occupied by the electron in state $n = 1$, that is, being very close to the nucleus. This tendency is less pronounced for the electron in state $n = 2$, $l = 1$. Similar remarks apply for the three states with $n = 3$. An electron with $n = 3$, $l = 0$ has a small chance of being inside the region occupied by electrons in state $n = 2$ and also a chance of being very close to the nucleus inside the region occupied by electrons in state $n = 1$. The electron with $n = 3$, $l = 1$ shows the same tendency for the region occupied by electrons with $n = 2$ but not for the electrons with $n = 1$. Finally, electrons with $n = 3$, $l = 2$ do not have either of the two tendencies. We can summarize this effect as follows: *For a given state n, the electrons in the lower angular momentum states are more likely to be found near the nucleus than those in states with higher angular momentum.* In the case of hydrogen, this effect is of no consequence; states with the same n and different l are degenerate and therefore have the same energy. In the case of a multielectron atom, this effect is important because it will explain why states with the same n but different l have different energies. This fact is important in understanding some of the general features of the periodic table.

21.7 THE PERIODIC TABLE

21.7a Pauli's Exclusion Principle

The lowest energy state of an electron is called the *ground state*. In the hydrogen atom this is specified by the following quantum numbers.

$$n = 1,\ l = 0,\ m_l = 0, \text{ and } m_s = \tfrac{1}{2} \text{ or } -\tfrac{1}{2}$$

What happens in an atom with 70 electrons? Are all 70 electrons in a state specified by these four quantum numbers? The experimental evidence says not. This conclusion is reflected in *Pauli's exclusion principle*, which states: *No two electrons in a system (be it an atom or a solid) can be in the same quantum state. That is, no two electrons can have identical values for the set of quantum numbers specifying the state, which in the case of an atom are n, l, m_l, m_s. At least one of the quantum numbers must be different.*

Professor Banesh Hoffman of the Queens College Mathematics Department has put it in a very colorful way: "It is as if the atom were a large city where electrons live in separate apartments. Each apartment has a different address, one quantum number indicating the street, another the house, a third the floor, and the fourth the apartment. These four quantum numbers are then the complete address of each apartment, and Pauli's principle is a regulation against overcrowding."[1]

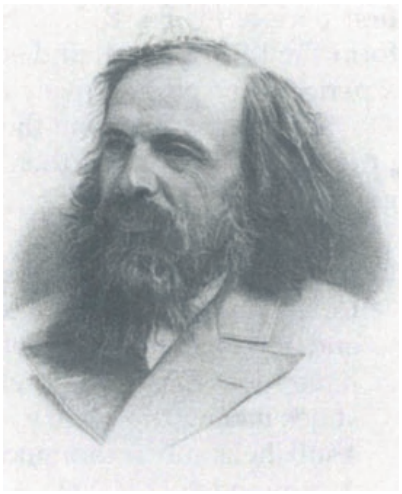
Wolfgang Pauli (1900–1958).

[1] *The Strange Story of the Quantum*. Banesh Hoffmann. Dover Publications Inc. N.Y. (1959).

21.7b Understanding the Periodic Table

Many physical and chemical properties of the elements are periodic functions of the atomic number Z, which is the number of positive charges (protons) in the nucleus and is also equal to the number of electrons for neutral atoms. This periodicity has to be taken loosely, in the sense that the period is not constant. But the fact remains that elements such as H, Li, Na, and K have very similar properties and so do elements such as He, Ne, and Ar. A standard form of the periodic table is presented in Fig. 21-17. If we leave aside for the moment the central block (see the scandium to zinc part of this section), the table is divided into eight groups (indicated by roman numerals), all elements in a group are in the same column. H, Li, Na, . . . form the first group; beryllium (Be), magnesium (Mg), calcium (Ca) . . . belong to group II; boron (B), aluminum (Al), gallium (Ga) . . . form group III; and so on until we get to He, Ne, Ar, . . ., which form group VIII. *Elements in a group have very similar chemical and optical properties*. The reason for this is that these properties are determined by the electron configuration. Atoms in a group have a similar electron configuration, as we will soon see.

In addition to dividing the elements into groups, we can also divide them into periods (indicated by arabic numerals in Fig. 21-17). H and He form the



Dimitri Mendeleev (1834–1907) was the first to indicate in 1869 that the chemical and physical properties of the elements are periodic functions of the atomic number Z.

FIGURE 21-17

Periodic table of the elements.

Period	Group I	Group II	Group III	Group IV	Group V	Group VI	Group VII	Group VIII
1	1 H 1.008							2 He 4.003
2	3 Li 6.941	4 Be 9.012	5 B 10.81	6 C 12.01	7 N 14.01	8 O 16.00	9 F 19.00	10 Ne 20.18
3	11 Na 22.99	12 Mg 24.31	13 Al 26.98	14 Si 28.09	15 P 30.97	16 S 32.06	17 Cl 35.45	18 Ar 39.95
4	19 K 39.10	20 Ca 40.08	21 Sc 44.96	22 Ti 47.90	23 V 50.94	24 Cr 52.00	25 Mn 54.94	26 Fe 55.85
	27 Co 58.93	28 Ni 58.71	29 Cu 63.55	30 Zn 65.38	31 Ga 69.72	32 Ge 72.59	33 As 74.92	34 Se 78.96
	35 Br 79.90	36 Kr 83.80						
5	37 Rb 85.47	38 Sr 87.62	39 Y 88.91	40 Zr 91.22	41 Nb 92.91	42 Mo 95.94	43 Tc (99)	44 Ru 101.1
	45 Rh 102.9	46 Pd 106.4	47 Ag 107.9	48 Cd 112.4	49 In 114.8	50 Sn 118.7	51 Sb 121.8	52 Te 127.6
	53 I 126.9	54 Xe 131.3						
6	55 Cs 132.9	56 Ba 137.3	72 Hf 178.5	73 Ta 180.9	74 W 183.9	75 Re 186.2	76 Os 190.2	77 Ir 192.2
	78 Pt 195.1	79 Au 197.0	80 Hg 200.6	81 Tl 204.4	82 Pb 207.2	83 Bi 209.0	84 Po (210)	85 At (210)
	86 Rn (222)							
7	87 Fr (223)	88 Ra 226.0						
Lanthanide series			57 La 138.9	58 Ce 140.1	59 Pr 140.9	60 Nd 144.2	61 Pm (145)	62 Sm 150.4
	63 Eu 152.0	64 Gd 157.3	65 Tb 158.9	66 Dy 162.5	67 Ho 164.9	68 Er 167.3	69 Tm 168.9	70 Yb 173.0
	71 Lu 175.0							
Actinide series			89 Ac (227)	90 Th 232.0	91 Pa 231.0	92 U 238.0	93 Np 237.0	94 Pu (244)
	95 Am (243)	96 Cm (247)	97 Bk (247)	98 Cf (251)	99 Es (254)	100 Fm (253)	101 Md (256)	102 No (254)
	103 Lw (257)							

KEY: Atomic number / Atomic Symbol / Atomic weight*

*The number in ( ) = Mass Number of the most stable isotope.

first period; Li, Be, B, . . . Ne form the second period; Na, Mg, Al, . . . Ar form the third period, and so on. The chemical properties of the elements in a period vary progressively as we go from left to right.

Let us now examine the electron configuration of some different atoms in order to understand their chemical behavior. We must keep in mind two guiding principles:

1. The electron configuration must satisfy the exclusion principle. Because for each quantum number l, the value of m_l has $(2l + 1)$ possible values and, because for each value of m_l, the value of m_s can be $+\frac{1}{2}$ or $-\frac{1}{2}$, we have room for $2(2l + 1)$ electrons in each *subshell*. (A subshell is the set of states having the same n and l quantum numbers.) Thus, for example, the s subshells can accommodate $2\,[(2)\,(0) + 1] = 2$ electrons; p subshells can accommodate $2\,[(2)\,(1) + 1] = 6$ electrons; d subshells can accommodate $2\,[(2)\,(2) + 1] = 10$ electrons, and so on.

2. The ground-state configuration is the one in which the total energy is a minimum. This means that we must know the energy of each electron when it is in a given state, that is, in a state characterized by a given n and l. In the case of H the energy of the atom is caused by the interaction between a single electron and the single positive charge in the nucleus. In multielectron atoms the energy is caused by the interaction of Z electrons with the positive nucleus of charge $+ Z|e|$ and the interaction of the Z electrons among themselves. (Note that for an atom to be electrically neutral the number of electrons must equal the number of positive charges.) The calculation of the energy associated with these interactions is a long and tedious task that cannot be solved analytically but can be solved by a numerical method. However, important results from these calculations can be understood with intuitive arguments.

   a. *The total energy of an electron in a multielectron atom becomes less negative (increases) with increasing n. The increase from a given n to the next is less pronounced as we go to higher n's.* Everything being equal, the state $n = 1$ has a lower energy (is more negative) than the state $n = 2$, the state $n = 2$ is lower than $n = 3$, and so on. This can be seen from Fig. 21-16, in which higher n is associated with larger r, and therefore the electrons in increasing higher n states have decreasingly lower binding energies. In addition, it can be seen from Fig. 18.4 that $\Delta E_{1\to2} > \Delta E_{2\to3}$, and so on. We saw this is true for hydrogen, and calculations show that it is also true for multielectron atoms.

   b. *For a given n, the electron with the lowest l has the lowest energy.* In the case of hydrogen, we saw that states with the same n but different l were degenerate (had the same energy). In multielectron atoms the situation is different. Consider an atom with Z electrons; let us say that $Z - 2$ electrons fill a certain number of subshells. We saw previously that a filled subshell has spherical symmetry. This means that outside the filled subshells, the remaining electrons will see a net charge of

$(+ Z|e|) - (Z - 2)|e| = 2|e|$. This will be true as long as they stay outside the space occupied by the inner electrons. We saw, however, that the electrons in the outer subshells with small values of l have a small chance of being inside the inner subshells, very close to the nucleus. This tendency decreases as l increases. If the electron gets inside, close to the nucleus, the nuclear charge will not be shielded by the other electrons and the electron will be more tightly bound; that is, its energy will be more negative, lower. Because this tendency to get inside is greater for the low l states, the energy will be lower for those electrons for which l is small.

With these two guiding principles in mind, let us proceed to determine the lowest energy configuration of electrons for the different elements. Consider first hydrogen, ^{1}H, where the superscript to the left of the chemical symbol represents the total number of electrons for that atom. The one electron will go to the 1s subshell. (The numeral in front of the letter s indicates the value of the quantum number n; for example, the 1s subshell is one for which $n = 1$ and $l = 0$.) The configuration will be $1s^1$, where the superscript 1 on the right side of s indicates the number of electrons in that state. For ^{2}He, the configuration is $1s^2$, that is, two electrons in the 1s subshell. Because m_l can have $(2l + 1)$ values, in the s subshells m_l has $[(2)\ (0) + 1] = 1$ value, this value is $m_l = 0$. By Pauli's principle, the two electrons in the s subshell must have opposite spins. The 1s subshell is full. Because for $n = 1$, l can only be zero, there are no more subshells with $n = 1$. For the next atoms, ^{3}Li, ^{4}Be, . . ., we must begin to put electrons in the subshells with $n = 2$. For $n = 2$, we have two possible subshells: 2s ($l = 0$) and 2p ($l = 1$). As indicated herein, the subshell with the smaller l value has lower energy and we begin filling the 2s and then the 2p. The electronic configurations for the next few atoms are as follows: For ^{3}Li the configuration is $1s^22s^1$; for ^{4}Be the configuration is $1s^22s^2$. The 2s is full. For ^{5}B, the configuration is $1s^22s^22p^1$. The 2p subshell has room for 5 more electrons. The 2p subshell will be progressively filled for the successive elements ^{6}C (carbon), ^{7}N (nitrogen), ^{8}O (oxygen), ^{9}F (fluorine) until we get to ^{10}Ne (neon) whose configuration will be $1s^22s^22p^6$. The 2p is full, and there are no more subshells in $n = 2$. We now begin to fill subshells with $n = 3$, starting with the 3s subshell and then the 3p; that is, for ^{11}Na, the electron configuration is $1s^22s^22p^63s^1$. For ^{12}Mg, the electron configuration is $1s^22s^22p^63s^2$. For ^{13}Al the electron configuration is $1s^22s^22p^63s^23p^1$. . . . . For ^{18}Ar the electron configuration is $1s^22s^22p^63s^23p^6$. The 3p is full.

We might expect that after the 3p state is full the 3d would begin to fill. However, it turns out that the 4s subshell has a lower energy than the 3d subshell. Why? There is no simple explanation; it just comes out of the calculations. We can get a feeling for the reason from the following argument. We have seen that the difference in energy in going from a given principal quantum number n to the next higher n becomes less pronounced as n increases. We have also seen that states with low l have lower energy because

they have a small probability of being found inside the inner shells, close to the nucleus. It is relatively simple to see that the dependence of the total energy E on l becomes more important as n increases, and there is an increasing number of electrons. The reason is that the difference in the charge that the electron experiences, when it is close to the nucleus or far from it, is greater when there are many electrons than when there are just a few. Consider, for example, ^{3}Li. The electron in the $2s$ state sees a net charge of only $+|e|$ when it is outside the orbit of the $1s$ electrons, and $+3|e|$ when it is inside their orbit, a threefold increase. Now consider ^{19}K, the electron in the $4s$ state sees a net charge of only $+|e|$ when it is outside the orbits of the other electrons, but experiences the full charge of the nucleus of $+19|e|$ when it is inside the other orbits, a nineteenfold increase. This argument gives a feeling of why the inversion of the energy of shells may take place particularly with the more complex wavefunction shapes of the higher energy electrons. This inversion of shell energies, that is, $4s$ states having lower energy than $3d$ states, begins to take place with ^{19}K and is predictable from detailed calculations. Because of this, ^{19}K has the electron configuration $1s^2 2s^2 2p^6 3s^2 3p^6 4s^1$. ^{20}Ca has the configuration $1s^2 \ldots 4s^2$. The $4s$ is now full. Next the $3d$ subshell begins to fill. The $3d$, with a capacity for 10 electrons is progressively filled as we go from ^{21}Sc (scandium) to ^{30}Zn (zinc). One should notice that the $4s$ and the $3d$ must be very close in energy, as evidenced by the fact that for Cr (chromium) and Cu (copper) the $4s$ loses one of the electrons to the $3d$. Inversions similar to that between the $3d$ and the $4s$ occur later. With this understanding of the configurations of electrons in the ground states of the simpler elements, let us now consider the general properties of some of the groups of elements in Fig. 21-17.

Rare Gases

Atoms in group VIII, He, Ne, Ar, . . . are inert gases: They interact very poorly with one another or with other atoms; as a result, they do not form molecules or solids easily. They are, except under very extreme circumstances (low temperature and high pressure), monoatomic gases. Why?

In these atoms the p subshell (except for He) is filled, and there are no additional electrons outside this subshell. Because the electron configuration is made up of closed subshells, their spacial distribution (and associated charge) has spherical symmetry. Spherically symmetric charge distribution, to an external observer (another atom), looks as if it is all concentrated at the center, that is, at the nucleus. The net charge of these atoms behaves as a *neutral point particle* at the center of the atom. As a result, these atoms do not have a net electric dipole. They do not have a magnetic dipole either. The reason is that for each electron with a certain positive m_l, there is one with an equal negative value of m_l, and therefore the net orbital angular momentum L is zero. If there is no orbital angular momentum, there can be no magnetic dipole moment because μ is proportional to L. Not only is there no orbital dipole moment, but there is also no net spin dipole moment. For every electron

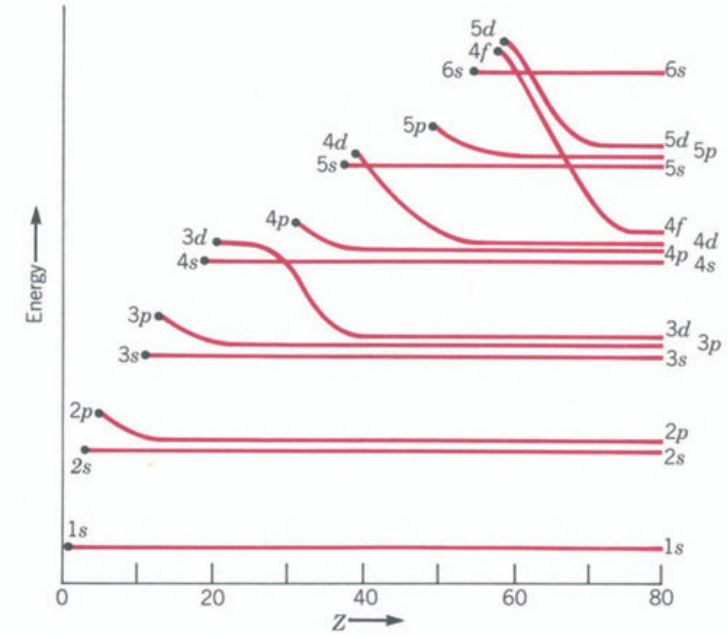


FIGURE 21-18
Relative energy of the subshells in multielectron atoms as a function of the atomic number Z. The curve showing the energy of a given subshell starts at the value of Z for which that particular subshell begins to be occupied. (*Source*: Robert Eisberg and Robert Resnick, *Quantum Physics of Atoms, Molecules, Solids, Nuclei, and Particles*, 2nd ed. Copyright © 1985 by John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc).

with $m_s = \frac{1}{2}$, there is one with $m_s = -\frac{1}{2}$. In addition, there is a large energy difference between the p subshell occupied by the outer electrons and the next available subshell, the s subshell. Figure 21-18 shows the energies of the shells as a function of Z. It is seen that it takes considerable energy to excite (raise to a higher energy level) the electrons of these atoms and thereby destroy their spherical electrical symmetry. Such excitation energy is not available at ordinary temperatures.

The standard mechanisms for molecule and crystal formation, which will be discussed in the next chapter, are electron exchange and sharing, and electric dipole and magnetic dipole interaction. These mechanisms are not present when there is spherical electrical symmetry. The electron configuration of these rare gas atoms is a very stable, low-energy configuration.

Alkali Elements

Now that we understand the rare gases, we can also understand many features of the elements in group I called the *alkali* elements. All these elements consist of a rare gas electron configuration plus one extra electron. Because the inert configuration is difficult to disturb, the chemical and some physical properties are determined by the behavior of this lone electron. For example, the optical spectrum is due to excitations of this extra electron. The atom has a configuration of electrons very similar to that of hydrogen but with some important differences.

When the extra electron is outside the inert configuration, it sees a spherical charge distribution that consists of

$$+Z\,|e| - (Z - 1)\,|e| = +|e|$$

In other words, all the positively charged protons in the nucleus are shielded

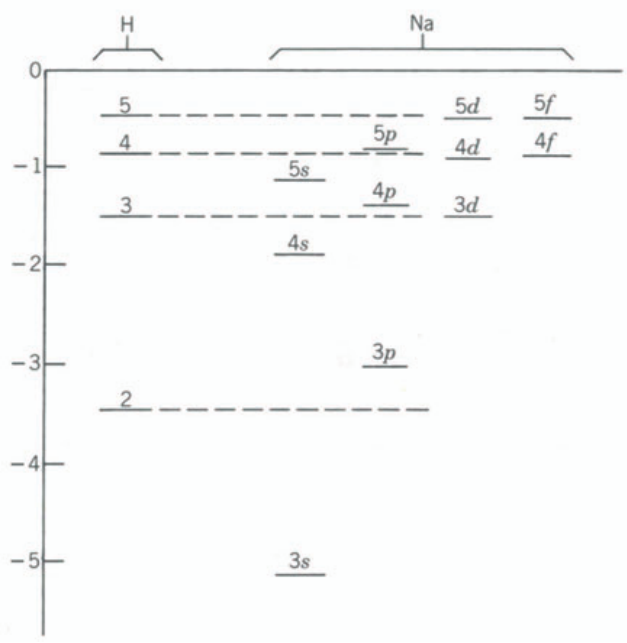


FIGURE 21-19
Schematic comparison of the energy levels of hydrogen and sodium atoms.

except one. A net charge of $+|e|$ acts as if it were at the center of the atom. This is the situation of the hydrogen atom. We may therefore expect that the energy levels for the outer electron should be very similar to those of hydrogen. They are in some cases; those cases for which l is large. However, they are different in cases where l is small. The reason for the similarities and the differences is that electrons in states with a large l tend to stay outside the inert configuration, as we have seen previously (see Fig. 21-16). Those electrons in states with small l spend some time inside the inner shells where the positive nuclear charge is not fully shielded. This means that the electron in these states will be more tightly bound and consequently will have a lower energy, that is, be more negative.

In Fig. 21-19 we show a few of the experimentally determined energy levels of Na (sodium). For purposes of comparison, some of the excited levels of H are shown on the left. As we indicated earlier, in the ground state the electronic configuration of Na ($1s^2 2s^2 2p^6 3s^1$) consists of 10 electrons having the same inert configuration as those of Ne ($Z = 10$) and one electron in the $3s$ subshell. As expected from the arguments just mentioned, in Na the other subshells with $n = 3$, that is, the $3p$ ($l = 1$) and the $3d$ ($l = 2$) are subshells of higher energy than the $3s$. It is clear from Fig. 21-19 that the $3s$ state of Na is quite different in energy from the $n = 3$ level of hydrogen, approximately 3.7 eV (remember that for hydrogen, states having the same n but different l are degenerate, Fig. 21-2). The difference between the $3p$ state of Na and the $n = 3$ state of hydrogen is less, $\approx$1.5 eV, and the $3d$ state of Na is almost identical to the $n = 3$ state of hydrogen. Similar conclusions are drawn if we compare the $n = 4$ states of Na and the $n = 4$ state of hydrogen. Experiment confirms the qualitative arguments concerning the similarities and the differences between hydrogen and the alkali elements.

Because the outer electron is either fully or partially shielded from the full charge of the nucleus by the other electrons, the energy needed to remove it from the atom, that is, to ionize the atom, is small. As a result, the alkali atoms are very active chemically: They easily lose the outer electron to other atoms and can therefore interact electrically with them. Most of these elements form ionic compound crystals, or metallic solids. We will discuss this in some detail in the next chapter.

Example 21-3

Let us assume that Na atoms are hydrogen-like atoms. The electron configuration of Na ($Z = 11$) is $1s^2 2s^2 2p^6 3s^1$. (a) What would be the expected energy needed to remove the $3s$ electron? (b) The measured ionization energy of Na is 5.14 eV. What is the effective positive charge seen by the $3s$ electron?

Solution

(a) The electron configuration of Na atoms consists of 10 electrons having a rare gas configuration plus one electron in the $3s$ subshell. Because the rare gas configuration has spherical symmetry, when the $3s$ electron is outside this configuration it sees a net positive charge of $+|e|$ at the center of the atom. As a consequence, the energy spectrum of this electron should be the same as that of hydrogen, that is,

$$E_n = -\frac{13.56 \text{ eV}}{n^2}$$

In particular, the ground-state energy, and therefore, the energy needed to remove the $3s$ electron should be

$$E_3 = -\frac{13.56 \text{ eV}}{3^2} = -1.51 \text{ eV}$$

(b) From Eq. 21.9, $E_n \propto Z^2$. We can make use of the fact that the ionization energy is 5.14 eV, rather than 1.51 eV, to find the effective charge seen by the $3s$ electron.

$$E_3 = -5.14 \text{ eV} = Z^2_{\text{effective}} \left(-\frac{13.56}{3^2} \text{ eV} \right)$$

Solving for $Z_{\text{effective}}$,

$$Z_{\text{effective}} = \left[\frac{5.14 \text{ eV} \times 9}{13.56 \text{ eV}} \right]^{1/2} = 1.85$$

The atoms that readily accept the electron that the alkali elements may easily give away are the elements in group VII. The reason is simple; these atoms

find themselves in the opposite situation. They have an imperfectly shielded nuclear charge; they lack one electron in the p subshell to have an inert low-energy configuration.

When chlorine (Cl) acquires an extra electron to become the negative ion Cl^-, this electron is bound by 3.62 eV. This means that in the process of capturing this electron, 3.62 eV of energy are released,

$$Cl + e^- \rightarrow Cl^- + 3.62 \text{ eV}$$

On the other hand, to ionize Na, 5.14 eV must be provided

$$Na + 5.14 \text{ eV} \rightarrow Na^+ + e^-$$

It would seem that energetically, this cannot occur naturally because an amount of energy equal to $5.14 \text{ eV} - 3.62 \text{ eV} = 1.52 \text{ eV}$ must be provided. However, there is an additional coulombic attraction energy to be considered. The binding between Na^+ and Cl^- involves a net release of energy, that more than compensates for the difference (see Section 22.3a). The consequence is that elements in group I combine readily with those in group VII to form very stable compounds, for example, NaCl, LiF, NaBr.

Scandium to Zinc (Sc–Zn)

A complete discussion of the properties of each group in the periodic table of elements is beyond the scope of this book. However, there is one more point of pertinent interest. As mentioned in connection with the alkali elements, the chemical and physical properties of an atom are determined by the outermost electrons (electrons outside the inert configuration). These are called the *valence* electrons. We can see, qualitatively at least, that in going from left to right in a period, the chemical properties change more or less progressively because the number of valence electrons increases. This, in physical terms, means that the net electric and magnetic dipole moments change, which in turn changes the interactions with other atoms. This progressive change in the chemical properties breaks down at the central block. In fact, the chemical properties of the atoms from Sc through Zn are very similar, and it is not difficult to see why. What happens in going from Sc to Zn is the addition of one more electron in the $3d$ subshell for each element in this progression. The $3d$ is being filled while the $4s$ is already full. The reason for the inversion, given earlier, is that the $4s$ state is lower in energy than the $3d$ state. However, the average radius of the $4s$ is greater than that of the $3d$. As a consequence, the $4s$ electrons *shield* the $3d$ electrons from external influences. Thus, adding one more electron to the $3d$ subshell inside the filled $4s$ subshell does not appreciably change the interaction between different atoms. It is not the $3d$ electrons that participate in the bonding but rather the $4s$ electrons. The $3d$ electrons remain localized, that is, attached to the original atom.

PROBLEMS

21.1 The differential equation for $\Phi(\phi)$, Eq. 21.4, can be rewritten as

$$\frac{d^2\Phi}{d\phi^2} + m_l^2 \Phi = 0$$

In this form, the equation is similar to the differential equation for the particle in the infinite potential well, Eq. 20.19. Assume a solution of the form $\Phi = e^{\alpha\phi}$ and show by direct substitution that such a function is a solution provided $\alpha = \pm i m_l$, where $i = \sqrt{-1}$.

21.2 In Fig. 21-1 it is seen that a point P with coordinates r, θ, and ϕ is the same point as a point with coordinates r, θ, and $\phi + 2\pi$. Because the eigenfunctions must be single-valued $\Phi(\phi) = \Phi(\phi + 2\pi)$. Use the solution found in Problem 21.1 to show that this requirement limits the values that m_l can take. That is, show that $m_l = 0, \pm 1, \pm 2, \pm 3, \ldots$.

21.3 In the Bohr model of the hydrogen atom, the electron is assumed to move in circular orbits around the proton, that is, the motion takes place in a plane that we can call the x-y plane. Use the uncertainty principle in the z direction, that is, $\Delta p_z \, \Delta z \geq \hbar$ and the fact that $\overline{p_z^2} \geq (\Delta p_z)^2$ (see Problems 19.21 and 19.22) to show that the motion of the electron cannot be planar motion.

21.4 Singly ionized helium behaves as a hydrogen atom but with twice the nuclear charge of the hydrogen atom. What is the ground-state energy of the remaining electron in singly ionized helium?

(*Answer*: 54.24 eV.)

21.5 (a) Use Eq. 21.9 to compare the spectrum of energies of singly ionized helium with the spectrum of hydrogen. (b) Draw an energy diagram showing the first few energy levels for hydrogen and for singly ionized helium. (c) What is the wavelength of the photon emitted when the remaining electron in singly ionized helium makes a transition from $n = 2$ to $n = 1$?

21.6 (a) How many atomic states are there in hydrogen with $n = 3$? (b) How are they distributed among the subshells? Label each state with the appropriate set of quantum numbers n, l, m_l, m_s. (c) Show that the number of states in a shell, that is, states having the same n, is given by $2n^2$. (*Hint*: $1 + 2 + 3 + \ldots + n = n(n + 1)/2$.)

21.7 A hydrogen atom is in a state with $l = 3$. What are the allowed values of L_z and μ_{lz}?

(*Answer*: $0, \pm\hbar, \pm 2\hbar, \pm 3\hbar$; $0, \pm e\hbar/2m, \pm e\hbar/m, \pm 3e\hbar/2m$.)

21.8 A ball with a mass of 100 g rotates in a circle of radius 1 m with a speed of 4 m/sec. How many possible orientations can the angular momentum of the ball take?

(*Answer*: 4×10^{33}.)

21.9 Draw an energy-level diagram and calculate the separation between adjacent normal Zeeman energy levels for the $3d$ state of an atom placed in a magnetic field of 1.5 T.

21.10 A certain transition in an atom of wavelength $\lambda = 4226$ Å, when observed in a magnetic field of 1 T, exhibits a normal Zeeman pattern whose components are separated by 1.4×10^{10} Hz. Calculate the ratio of the charge to the mass of the electron from these data.

21.11 A spectrometer is capable of resolving spectral lines separated by 0.1 Å when the wavelength is 5000 Å. What is the smallest magnetic field needed to observe the normal Zeeman effect in such a case?

(*Answer*: 0.86 T.)

21.12 A beam of hydrogen atoms emerges from an oven with a velocity $v = 10^4$ m/sec. The beam of atoms passes through a region 5 cm long with a magnetic field gradient $dB/dz = 10^5$ T/m. The direction of the field gradient is perpendicular to the incident velocity. The atoms then continue through a field-free region for another 20 cm before being deposited on the screen. What is the separation between the two lines on the screen?

(*Answer*: 12.5 cm.)

21.13 Electron spin resonance refers to the absorption of electromagnetic radiation by the electrons in atoms as they make transitions from the state where their spins are parallel to an external magnetic field to the state where the spins are antiparallel. At what frequency ν would spin resonance occur with hydrogen atoms in a magnetic field $B = 1.5$ T?

(*Answer*: 4.2×10^{10} Hz.)

21.14 The radial probability function for the ground state of the hydrogen atom is $P(r) = A\, r^2\, e^{-2r/r_0}$, where A is a constant and r_0 is the radius of the smallest orbit in the Bohr model. For what value of r is $P(r)$ a maximum?

(*Answer*: r_0.)

21.15 (a) Calculate the ground-state energy of a system of 10 electrons in a one-dimensional infinite potential well of width $a = 1$ Å? (b) What is the average energy per electron?

(*Answer*: (a) 4.1×10^3 eV, (b) 4.1×10^2 eV/electron.)

21.16 Repeat Problem 21.15 for a system of 10 electrons in a two-dimensional well of width 1 Å and length 1 Å. See Problem 20.19.

(*Answer*: (a) 2.2×10^3 eV, (b) 2.2×10^2 eV/electron.)

21.17 Suppose there were atoms having electrons with the principal quantum number up to and including $n = 7$. What would be the maximum number of elements?

(*Answer*: 280.)

21.18 What is the electronic configuration of Al ($Z = 13$), Si ($Z = 14$), P ($Z = 15$), Ca ($Z = 20$), and Br ($Z = 35$)?

21.19 The electron configuration of sulfur (16 electrons) is $1s^2 2s^2 2p^6 3s^2 3p^4$. Write a complete set of quantum numbers for the four electrons in the $3p$ subshell.

21.20 No two electrons in a subshell pair up with opposite spins until there is at least one electron in each state (each value of m_l) of the subshell. This is known in spectroscopy as *Hund's rule*. In view of this, what combinations of four among the set of quantum numbers found in Problem 21.19 are acceptable for the four electrons in the $3p$ subshell of the sulfur atom?

21.21 (a) What is the electron configuration of Li ($Z = 3$)? (b) Assuming that Li is a hydrogen-like atom, calculate the ionization energy of the $2s$ valence electron. (c) The experimentally determined ionization energy of Li is 5.39 eV; what is the effective positive charge seen by the $2s$ electron?

21.22 Potassium belongs to the alkali elements group with one valence electron in the $4s$ subshell. The ionization energy of potassium is 4.34 eV. What is the effective charge seen by the $4s$ valence electron?

(*Answer*: 2.26 $|e|$.)

CHAPTER 22

Crystal Structures and Bonding in Solids

22.1 INTRODUCTION

Matter comes in three forms or phases: gases, liquids, and solids; the last two are called *condensed systems*. Semiconductors are crystalline solids. To understand their electrical properties we need a basic knowledge of their crystal structure.

Gases

In a molecular gas the average distance between the molecules is large compared with their size. As a result, the intermolecular forces (forces between molecules) are small compared with the forces that bind the molecules together. Because of this, the relative positions of the molecules are completely random.

Liquids and Solids

The atoms and molecules in condensed systems are close together; the forces between molecules are comparable to the forces within the molecules. In a liquid the molecules form temporary short-range order arrangements that are continuously broken by the high thermal energy of the molecules. When a liquid is slowly cooled, the molecules will arrange themselves in a crystalline array that produces the maximum number of bonds and thereby leads to a state of minimum energy. If the liquid is cooled too fast, the internal energy will be removed before the molecules have a chance to arrange themselves. As a result, a solid is formed that can be considered a "snapshot" of the liquid: We have an *amorphous* solid that displays short-range order but not long-range order. Short-range order usually means a few atoms and long-range order thousands of atoms. Amorphous solids are called *glasses* or glassy solids. Because of the lack of long-range order in glasses, the bonds between different atoms or molecules vary in strength. When such a solid is heated, the weaker bonds break first, leading to a gradual softening before a complete meltdown occurs. Because of this, amorphous materials lack a well-defined melting point.

We will be dealing primarily with crystalline solids and, although a detailed knowledge of the crystal structure is not necessary to obtain a semiquantitative understanding of electrical conduction processes, we will discuss briefly a few facts about crystal structures so that the structure of the semiconductor silicon can be understood.

22.2 CRYSTAL STRUCTURES

A crystalline solid is a three-dimensional, periodic array of atoms or molecules called a *crystal structure*. The most convincing evidence concerning the regular arrangement of atoms in a solid comes from X-ray diffraction (Section 12.7). The diffraction pattern not only confirms the periodic arrangement, but also has allowed crystallographers to determine the arrangement.

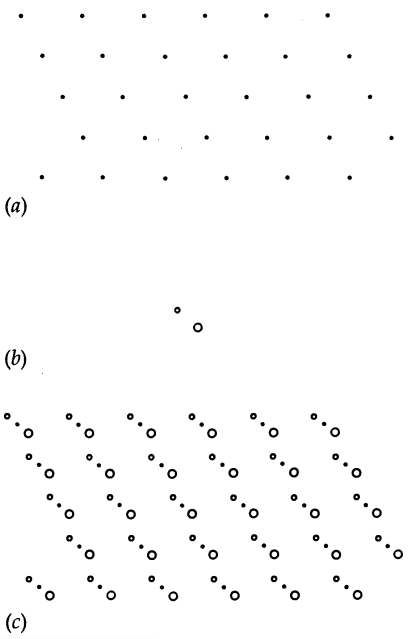


FIGURE 22-1
Specification of a crystal structure. (*a*) Regular periodic arrangement of mathematical points in space called the *space lattice*. (*b*) *Basis*: atom or group of atoms. (*c*) Basis placed at or near each lattice point yields a *crystal structure*.

One might guess from seeing snowflakes that there is an infinite number of possible lattice arrangements, but such is not the case. It turns out that all crystal structures can be included in one of fourteen lattice arrangements.

A crystal structure can be specified by a *periodic space lattice* and an atom or group of atoms placed at or around each lattice point. The atom or group of atoms constitutes the *basis* (see Fig. 22-1*b*). The *space lattice* is a regular periodic arrangement of points in space and is purely a mathematical abstraction (see Fig. 22-1*a*). To obtain a crystal structure, we must place at or around each lattice point a basis of atoms, Fig. 22-1*c*. This group of atoms must be identical in composition, arrangement, and orientation.

One important fact about a space lattice that can be readily seen from Fig. 22-1*a* is that every lattice point has identical surroundings. The grouping of lattice points about any given lattice point is the same as the grouping about any other point. It turns out that geometrically there are only 14 unique ways in which this can be done: This gives rise to the 14 *Bravais lattices* shown in Fig. 22-2, named after the discoverer. Any lattice can be constructed by putting side by side one of these 14 Bravais lattices. These lattices are all parallelepipeds that differ from one another in the relative sizes of the three basic sides and/or the angles between them.

Let us consider some crystal structures.

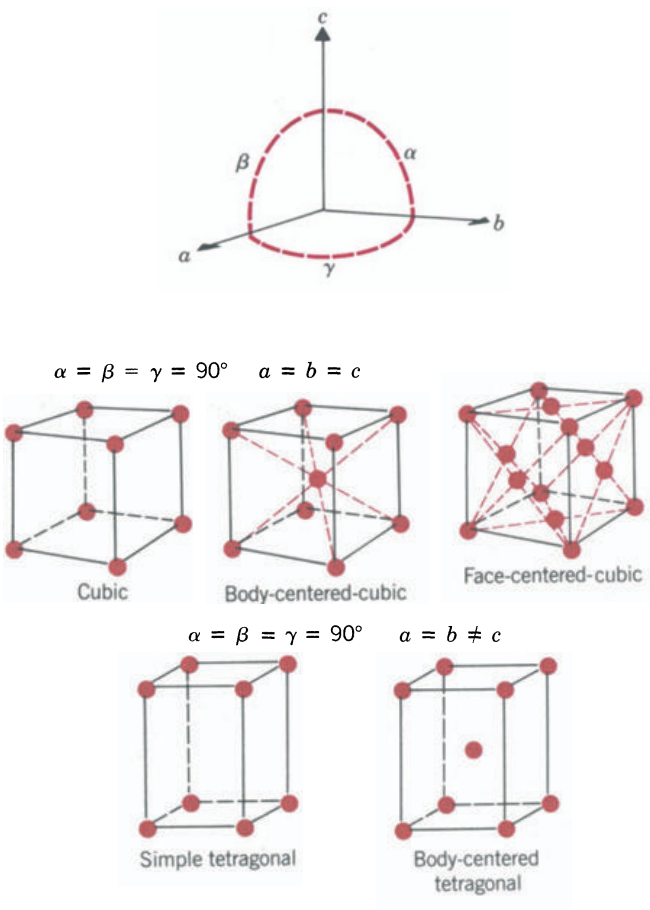


FIGURE 22-2
The 14 Bravais space lattices.
(continued on page 350)

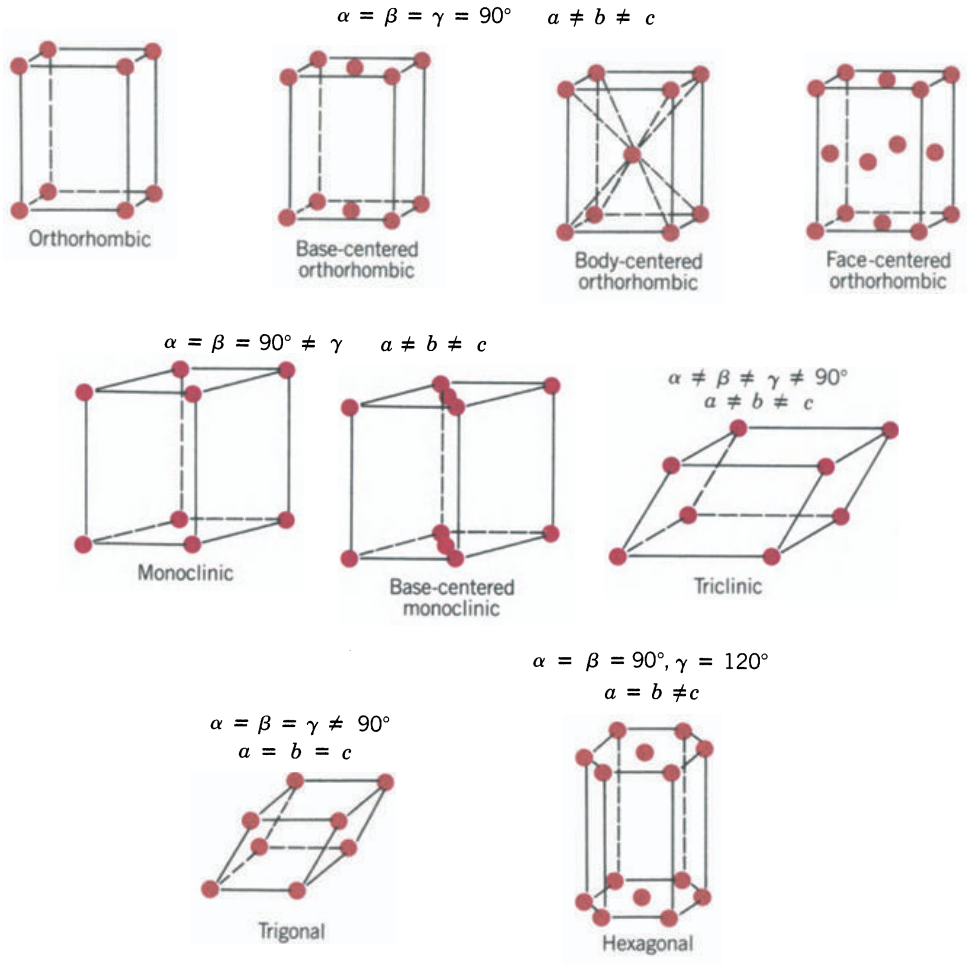


FIGURE 22-2
(Continued)

NaCl Structure

This is a face-centered cubic (fcc) Bravais lattice (Fig. 22-3*a*). The basis consists of a Na atom and a Cl atom separated by one-half the body diagonal, Fig. 22-3*b*. Other materials having the same structure include KBr, KCl, AgBr, MgO, MnO, and PbS.

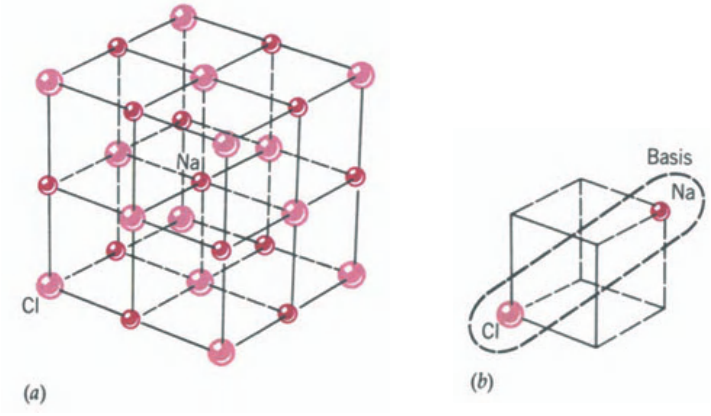


FIGURE 22-3
(*a*) Face-centered crystal structure of sodium chloride (NaCl). (*b*) Basis of the NaCl crystal structure.

CsCl Structure

This is a simple cubic Bravais lattice (Fig. 22-4*a*). The basis consists of a Cl atom at the corner and a Cs atom separated by one-half the body diagonal, Fig. 22-4*b*. Other materials having this structure include AgMg, AlNi, CuZn (brass), and BeCu.

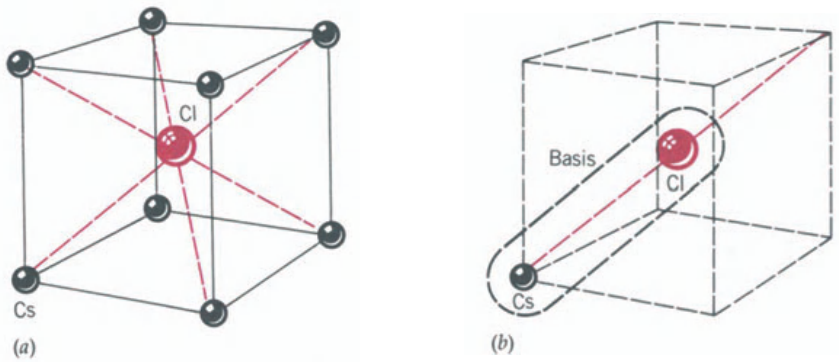


FIGURE 22-4
(*a*) Cubic crystal structure of cesium chloride (CsCl). (*b*) Basis of the CsCl crystal structure.

Diamond Structure

It is a face-centered cubic (Fig. 22-5*a*). The basis consists of two identical atoms, one is situated at the corner of the cube and the other is displaced by one-quarter the body diagonal along that diagonal (see Fig. 22-5*b*). The diamond structure can be thought of as being made of two interlocked fcc structures, each with one atom per lattice point. These two fcc structures are displaced from each other by $\frac{1}{4}d$ along the body diagonal.

The structure of diamond gives rise to a tetrahedral bond arrangement where each atom can be considered to be at the center of a tetrahedron forming one bond with each of its four nearest neighbors. These appear to be located at the four corners of the tetrahedron (see Fig. 22-6). Materials having this type of structure include C (carbon in the diamond structure), Si (silicon), Ge (germanium), and Sn (tin).

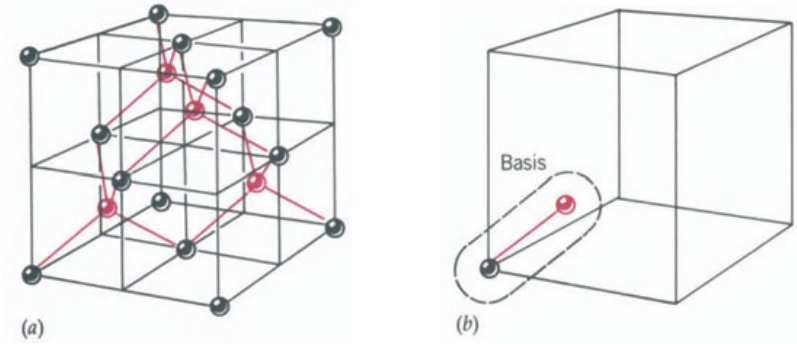


FIGURE 22-5
(*a*) Face-centered crystal structure of diamond. (*b*) Basis of the diamond crystal structure. The diamond structure can be thought of as two interlocked face-centered structures displaced from each other along the body diagonal by one quarter of the diagonal of the unit cube. For the sake of clarity in the drawing, only some of the atoms in the unit cube of the structure have been drawn in the figure.

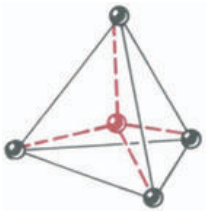


FIGURE 22-6
Tetrahedral arrangement of atoms in diamond resulting from the crystal structure of Fig. 22-5*a*. Each carbon atom in the diamond structure is at the center of a tetrahedron with its four nearest neighbors located at the corners of the tetrahedron.

22.3 CRYSTAL BONDING

As mentioned before, the exact knowledge of the crystal structure is not necessary in order to obtain a semiquantitative understanding of conduction processes in solids. However, the scientist who wants to make exact calculations of some property of a particular crystal must have a detailed knowledge of the crystal structure of that material. For our purposes it is more important to achieve an understanding of the different mechanisms that hold the structure together, that is, the different types of bonding. There are four general types of bonding that we discuss next. The actual bonding is usually a mixture of two or three.

22.3a Ionic Crystals (Ionic Bonding)

Ionic crystals consist of positive and negative ions in a periodic array, for example, Figs. 22-3 and 22-4. This is a result of an atom losing one or more electrons and another atom capturing them. For this to occur, the exchange must be energetically favorable. This means that if one starts with two neutral atoms, say Na and Cl, and goes through the process of ionizing one of them, then giving the resulting electron or electrons to the other and bringing the two ions together, the overall energy of the bound pair must be less than the energy of the initial state of two neutral atoms. As we have discussed in Chapter 21 in connection with the periodic table, this situation occurs (although not exclusively) with the alkali elements that have a weakly bound *s* electron and the halogen elements that need one electron in order to have an inert configuration. To ionize Na, an energy of 5.14 eV must be provided, which may be written as

$$\mathrm{Na} + 5.14\ \mathrm{eV} \rightarrow \mathrm{Na}^+ + \mathrm{e}^-$$

On the other hand, when Cl captures one electron, this electron will become bound, and the energy released will be 3.62 eV. We may write the reaction as

$$\mathrm{Cl} + \mathrm{e}^- \rightarrow \mathrm{Cl}^- + 3.62\ \mathrm{eV}$$

If now we bring Na^+ and Cl^- together until the separation between them is 2.51 Å (this is the equilibrium separation between the centers of the two ions in the NaCl molecule), the Coulomb attraction potential energy E_p will be (see Eq. 14.9)

$$E_p = -\frac{1}{4\pi\epsilon_0}\frac{e^2}{r} = -9 \times 10^9\ \mathrm{Nm^2/C^2}\,\frac{(1.6 \times 10^{-19}\ \mathrm{C})^2}{2.51 \times 10^{-10}\ \mathrm{m}}$$
$$= -9.18 \times 10^{-19}\ \mathrm{J} = -5.73\ \mathrm{eV}$$

This represents a decrease of energy from the situation where the two ions are infinitely far apart; that is, this represents a release of energy. The net energy released is

$$E = 3.62\ \mathrm{eV} + 5.73\ \mathrm{eV} - 5.14\ \mathrm{eV} = 4.22\ \mathrm{eV}$$

If $E = 0$ is the reference energy state when the Na and Cl atoms are far apart, when they are ions at $r = 2.51$ Å their energy is -4.22 eV. In other words, they fall into a potential well of depth -4.22 eV. This is the energy needed to break apart the NaCl molecule and restore each ion to electrical neutrality.

In an ionic crystal, the binding energy is the sum of all the attractive forces (between all ions of opposite sign) and all the repulsive forces (between ions of like sign). For an ionic crystal this energy can be written as

$$E = -\alpha \frac{1}{4\pi\epsilon_0} \frac{e^2}{r} \tag{22.1}$$

where α is known as the *Madelung's constant* and r is the distance between two nearest ions. The Madelung constant is different for different crystal structures. For the fcc NaCl structure, $\alpha = 1.7476$. Let us look at this particular structure. The Na^+ ion has six nearest neighbors (Fig. 22-3*a*) at a distance r (six Cl^- ions). Therefore the energy contribution is

$$E_6 = -6 \frac{1}{4\pi\epsilon_0} \frac{e^2}{r} \tag{22.2}$$

The next nearest neighbors are 12 Na^+ at a distance $(2)^{1/2}r$ (Fig. 22-3*a*) and contribute a repulsion energy of

$$E_{12} = +12 \frac{1}{4\pi\epsilon_0} \frac{e^2}{2^{1/2}r} \tag{22.3}$$

At the next distance there are eight Cl^- at $(3)^{1/2}r$

$$E_8 = -8 \frac{1}{4\pi\epsilon_0} \frac{e^2}{3^{1/2}r} \tag{22.4}$$

Then there are six Na^+ at $2r$

$$E_6 = +6 \frac{1}{4\pi\epsilon_0} \frac{e^2}{2r} \tag{22.5}$$

and so on. The total binding energy obtained by combining all these contributions, Eqs. 22.2, 22.3, 22.4, 22.5, . . ., is

$$E_{\text{Total}} = -\frac{1}{4\pi\epsilon_0} \frac{e^2}{r} \left[6 - \frac{12}{\sqrt{2}} + \frac{8}{\sqrt{3}} - \frac{6}{2} + \cdots \right]$$

and the summation of the series was evaluated by Madelung for fcc crystals as 1.7476. The total binding energy is then

$$E_{\text{Total}} = -1.7476 \frac{1}{4\pi\epsilon_0} \frac{e^2}{r} \tag{22.6}$$

In arriving at Eq. 22.6, we have considered a Na^+ ion and have calculated the electrostatic potential energy due to the other ions. We have assigned all this energy to the Na^+ ion. Actually, when we consider two charges q_1 and q_2

separated by a distance r and write that the electrostatic potential energy is

$$\frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r}$$

this energy is the energy of both charges. The electrostatic potential energy associated with one charge is half that amount. Thus, in the present case, the binding energy of the Na^+ ion under consideration is half the amount given by Eq. 22.6. When the values for e and r (2.8 Å for crystalline NaCl) are substituted into the Eq. 22.6 and the division by two is performed, we obtain a binding energy per ion as $\approx$ 4 eV/ion. This is a rather strong bond. To get an idea, consider that the thermal energy at room temperature k_BT is 0.025 eV. This is the reason why ionic crystals are extremely hard and have a high melting point (801° C for NaCl, 845° C for LiF as examples). We should also note that all the electrons are bound to the ions, which, with the electron exchange, now have completely filled electron subshells. Because there are no free or loosely bound electrons to transport charge when an electric field is applied across the crystal, simple ionic crystals do not conduct electricity and belong to the class of *insulators*.

22.3b Covalent Bond

Another important and very strong type of bond is the *covalent bond*. A good example of this is the H_2 molecule. The bond comes about because both electrons are in orbit (so to speak) around both nuclei, Fig. 22-7. Because electrons repel each other, the largest average distance apart can be maintained when their motion is coupled so that at a given time each is on the same side of each nucleus. Therefore, one or the other of the electrons spends part of its time between the two nuclei, attracting both of them and thus binding them. One way to understand why this takes place is by considering the tendency of the atoms to have filled subshells. The hydrogen atom needs one more electron to fill the 1*s* subshell and thus achieve the inert electronic configuration of helium. Lacking any other source of electrons, it will try to capture the extra electron from the adjacent hydrogen atom. Of course, the other atom will try to do the same, and they compromise by sharing both electrons.

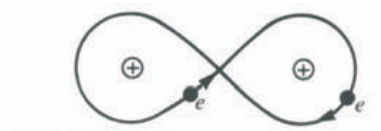


FIGURE 22-7
Schematic representation of the electrons of two bound hydrogen atoms orbiting around the nuclei and in the process creating a covalent bond between the two atoms.

This can be put into a more rigorous quantum mechanical language. Consider two identical H atoms, each with its electron in the 1*s* state. When the two atoms are far apart the wavefunctions do not overlap, and therefore the electrons remain with their parent atoms. The situation is shown in Fig. 22-8, which represents the radial dependence of the 1*s* wavefunction for

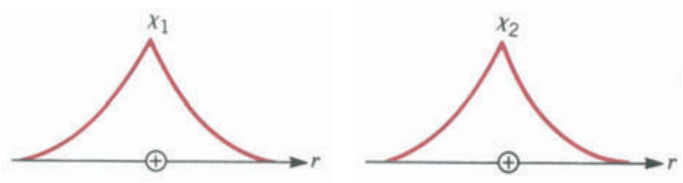


FIGURE 22-8
Radial part of the ground-state eigenfunctions of two isolated hydrogen atoms as a function of the distance of the electrons from their respective nuclei.

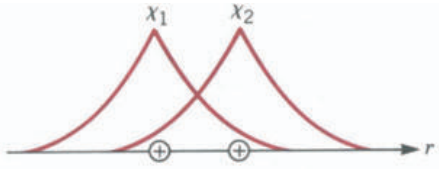


FIGURE 22-9
Overlap of the eigenfunctions of two individual atoms when the atoms are brought close together to form a molecule.

hydrogen atoms numbered 1 and 2. Note that each wavefunction is radially symmetric about a nucleus. If, however, the two atoms are brought close together until the nuclear separation is ≈ 0.7 Å (the equilibrium separation between the two hydrogen atoms in the H_2 molecule), the two wavefunctions overlap significantly, see Fig. 22-9. Because the electron from one atom can be found close to the nucleus of the other, it can be captured by the latter, and as a result, for a certain fraction of its time a given electron will be found with each proton. Because the two electrons are indistinguishable, it can be said that they spend equal time with each proton. The wavefunction associated with the electrons must reflect this fact. There are two possibilities, which will be discussed in detail in Chapter 24 (Section 24.4). The one that leads to the *covalent bond* is the sum of the two wavefunctions $\chi = \chi_1 + \chi_2$, shown in Fig. 22-10. This wavefunction satisfies the condition that each electron can be found equally with either of the two protons because it is symmetric about the midpoint between the two protons. But it also has the interesting feature that electrons represented by such a wavefunction spend a considerable amount of time between the two protons. When that happens, both protons are attracted toward the center by the electron and thus toward each other, thereby producing the bonding. In a single covalent bond each atom contributes one electron to the bond.

The best examples of covalent bonding in solids are offered by carbon in the diamond structure, silicon, and germanium. These elements belong to group IV of the periodic table (see Fig. 21-17). The outermost electron configuration consists of a filled s subshell followed by a p subshell with two electrons (C: $1s^2 2s^2 2p^2$; Si: $1s^2 2s^2 2p^6 3s^2 3p^2$; Ge: $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^2$). To have a full p subshell, each needs four more electrons. They achieve this by sharing four electrons with four other atoms. Each atom provides two electrons from the outer p subshell and two from the s subshell in order to form four covalent bonds. Even though the two electrons in the s subshell form a closed subshell, they participate in the bonding because they are very close in energy to the next p subshell (see Fig. 21-18). The most stable arrangement of the four bonds occurs when the atoms are symmetrically arranged in space. This symmetry is achieved by having a tetrahedral structure. Each atom is at the center of a tetrahedron and makes a covalent bond with its four nearest

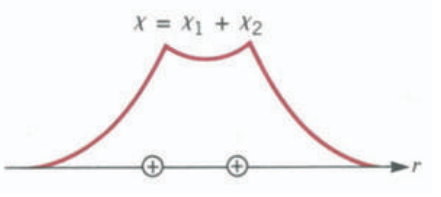


FIGURE 22-10
Quantum mechanical basis of the covalent bond. The eigenfunction $\chi = \chi_1 + \chi_2$ associated with an electron that can be found with equal probability with either of the two nuclei. Note that because the value of χ is large in the region between the two nuclei, the electron represented by such an eigenfunction spends considerable amount of time between the nuclei and in so doing binds them together.

neighbors, which are located at the corners of the tetrahedron (see Fig. 22-6). Each corner atom, in turn, finds itself at the center of another tetrahedron, so that the structure consists of a system of interlocked tetrahedra. The structure is in fact two interpenetrating face-centered cubic lattices that are displaced along the diagonal by one quarter of the diagonal of the cube (Fig. 22-5*a*).

The covalent bond is very strong, for example, 7.4 eV for diamond, 12.3 eV for SiC (carborandum). Moreover, because all available valence electrons pair up and orbit around pairs of atoms, it is difficult to dislodge them to conduct electricity and therefore covalently bonded solids belong to the class of *insulators* or *semiconductors*.

22.3c Metallic Bond

The metallic bond can be thought of as the limiting case of the covalent bond in which electrons are shared by all the ions in the crystal.

When a crystal is formed of atoms that have a few weakly bound electrons in the outer subshells, these electrons become free from the individual atoms, which thereupon acquire stable closed subshell configurations. The energy needed to liberate these loosely bound electrons is more than compensated for by the decrease in energy resulting from the binding.

We can visualize the metallic bonding as follows. The liberated electrons move through the entire crystal, visiting each positive ion at some time or other. We thus have a situation where an array of heavy positive ions is permeated by a "sea" of highly mobile negative electrons (see Fig. 22-11). It is the electrostatic attraction between the electron sea and the positive ions that prevents the breakdown of the entire structure, which would result from the repulsion of the positive ions. The electrons provide the "glue" that keeps the structure together.

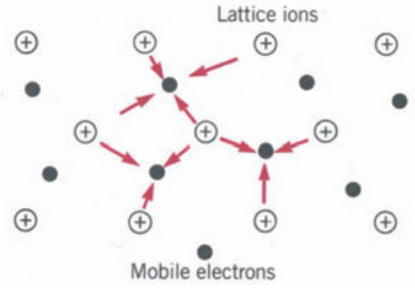


FIGURE 22-11
Nature of the metallic bond. As the free electrons in the metal move through the solid they attract the positive lattice ions toward them and thus prevent the breakdown of the structure.

The metallic bond is in many ways similar to the ionic bond in the sense that the main role is played by the electrostatic attraction between unlike charges; however, there is a big difference. Whereas in the ionic crystal the position of the positive and negative charges and therefore the directionality of the forces is fixed, in the metallic bond they are not. The electrostatic attraction comes from all directions. This is important, because it explains why a small deformation in a nearly perfect metal crystal does not cause a fracture. Whether we compress, twist, or pull a piece of metal, the cohesive forces are still there and coming from all directions. Pure metals are ductile and malleable. The presence of highly mobile charges will be used in Chapter 23 to explain the excellent thermal and electrical properties of metals.

22.3d Molecular or van der Waals Bond

We have seen that when the outer subshells of an atom are not filled, the atom tries to fill them by associating with an atom that can supply one or more electrons. The two atoms then become bonded in the process. What

happens if the atom has an outer subshell that is completely filled and difficult to excite, such as is the case with the rare gases? The answer is that it is very difficult for them to bond and therefore to form molecules and solids. However, they (except He) can form solids at very low temperatures by another mechanism. The mechanism is generally known as *dipole-dipole interaction*, and the resulting forces are called *van der Waals* or *London forces.*

Before we consider this effect in the rare gases, let us examine the case of the dipole bonding of water, H_2O. There are molecules (*polar* molecules) that possess permanent electric dipole moments. An example is H_2O, in which the molecule is formed by covalent bonding. The two 1*s* electrons from the H atoms form covalent bonds with the 2*p* electrons from the O atom. The result is that the two H atoms as well as the O atom will complete their subshells, a situation that yields a stable low-energy configuration. The resulting electron configuration, however, is not symmetric. The electron concentration around the oxygen atom makes that end of the molecule more negative than the ends where the hydrogen atoms are. The partially positive hydrogen atoms repel one another, which contributes to the angle of their bonds (see Fig. 22-12*a*). The net result is that the molecule has a permanent dipole moment, Fig. 22-12*b*. Such molecules tend to align themselves in such a way that the ends of opposite charge are adjacent and thus attract each other, forming chains and two- or three-dimensional structures, as indicated in Fig. 22-13. A polar molecule is also able to attract molecules that do not have a permanent dipole moment. The electric field of the polar molecule causes a separation of charge in the other molecule; that is, it induces a dipole moment in the same direction as its own, as shown in Fig. 22-14. The result is an attractive force. (This is analogous to what happens to an unmagnetized piece of iron in the presence of a magnet.) Two molecules or atoms with no permanent dipole moment, such as the rare gases, can still attract each other. Even though the electron distribution is *symmetric on the average*, the electrons themselves are in constant motion, and at any given instant, one part of the molecule can be more negative than another. In the polar molecule, the charge asymmetry is fixed, whereas in the nonpolar molecule there is a constantly shifting charge asymmetry. When two such nonpolar molecules are close together, their fluctuating charge distribution tends to couple and move together so that adjacent ends always have opposite signs, thus mutually inducing dipoles with a resulting attraction.

We may examine the effect of the moving charges in atoms in much the same way as we did in considering the covalent bond. When two atoms are brought close enough so that the electrons of one atom experience coulombic forces from the electrons of the other atom, and vice versa, they seek the lowest energy configuration. If the electrons circled the nuclei at random, then at times two electrons would be on the sides of the nuclei adjacent to each other, even though at other times they would be on opposite sides of their nuclei at the farthest distance from each other (see Fig. 22-15*a* and *b*). It was shown quantum mechanically that the average electron energy of the *a-b* type shown in Fig. 22-15*a* and 22-15*b* is higher than that of the coupled

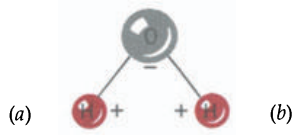


FIGURE 22-12
(*a*) Covalent bond of the oxygen and hydrogen atoms in the water molecule (H_2O). (*b*) Net electric dipole moment of the H_2O molecule.

FIGURE 22-13
Chain of molecules with permanent electric dipole moment created when the sides of opposite charge of the molecules align themselves.

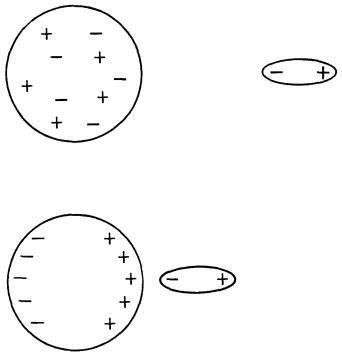


FIGURE 22-14
A molecule with permanent electric dipole moment (a polar molecule) induces a dipole moment in a nonpolar molecule when the two molecules are close together.

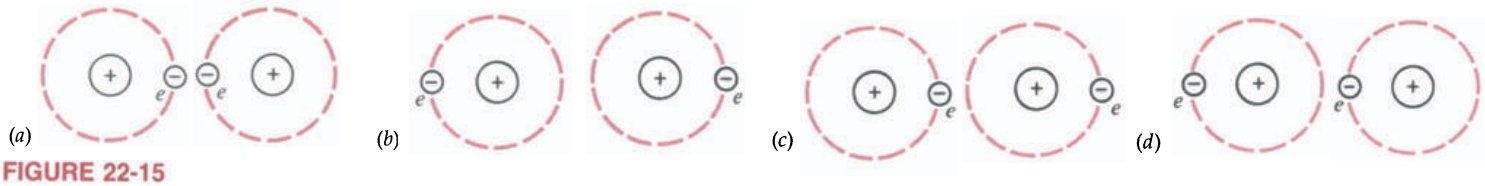


FIGURE 22-15

(*a*) and (*b*) Uncoupled motion of electrons in two individual atoms resulting in a situation where at times the electrons from the two atoms are on adjacent sides of the nuclei and at other times they are on opposite sides of the nuclei. (*c*) and (*d*) Coupled motion of electrons in two atoms as a result of which the electrons are never close together. This coupled motion gives rise to attractive forces between the two atoms, as can be seen from Fig. 22-15*c* and *d*. These forces are known as London or van der Waal forces.

types, Fig. 21-15*c* and *d*, in which the electrons "see" each other and stay as far away as possible. This reduction of energy by the coupling of electrons is called the *London force* after Fritz London, who first calculated it in 1927. This electron coupling occurs for all atoms, and this low energy attraction is also called *van der Waal's forces* after the chemist who analysed the condensation of gases. Although London forces exist between all atoms and molecules, they can be observed in their isolated state only in the rare gases because other atoms and molecules have additional forces, discussed earlier, that mask the weaker London forces. It is these forces that cause rare gases to condense, but only at very low temperatures because the resulting binding energy of the atoms is smaller than the thermal energy at higher temperatures. All the rare gases condense, and at sufficiently low temperatures they will solidify into crystalline solids, except for helium. Helium will solidify only under high pressure and low temperature.

PROBLEMS

22.1 Copper has a face-centered cubic structure with a one-atom basis. The density of copper is 8.96 g/cm^3 and its atomic weight is 63.5 g/mole. What is the length of the unit cube of the structure?

(*Answer*: 3.61 Å.)

22.2 Sodium has a body-centered cubic structure with a one-atom basis. The density and the atomic weight of sodium are 0.971 g/cm^3 and 23 g/mole, respectively. What is the length of the unit cube of the structure?

(*Answer*: 4.29 Å.)

22.3 Assuming that atoms in a crystal structure are arranged as close-packed spheres, what is the ratio of the volume of the atoms to the volume available for the simple cubic structure? Assume a one-atom basis.

(*Answer*: 0.52.)

22.4 Repeat Problem 22.3 for the body-centered cubic structure.

(*Answer*: 0.68.)

22.5 Repeat Problem 22.3 for the face-centered cubic structure.

(*Answer*: 0.74.)

22.6 A given ionic crystal has the same structure as NaCl. The separation between adjacent ions is 3.2

Å. What is the binding energy of each ion, and would you expect it to have a higher or lower melting point than NaCl?

(*Answer*: 3.93 eV.)

22.7 As we saw in Section 22-3a, the dissociation energy (energy to break up the molecule) of the NaCl molecule is 4.22 eV. Many chemical handbooks list the dissociation energy in kcal/mole. What is the dissociation energy of NaCl in kcal/mole?

(*Answer*: 97 kcal/mole.)

22.8 The equilibrium separation between the centers of the two ions in a HCl molecule is 1.27 Å. The ionization energy for the hydrogen atom is 13.56 eV. The electron affinity of Cl (that is, the energy released when the Cl atom captures an electron) is 3.62 eV. What is the dissociation energy (the energy needed to break up the molecule) for the HCl molecule?

(*Answer*: 1.40 eV.)

22.9 The dissociation energy of the KF molecule is 5.12 eV. The ionization energy for K is 4.34 eV, and the electron affinity of F is 4.07 eV. What is the equilibrium separation constant for the KF molecule?

22.10 Consider a one-dimensional lattice consisting of alternating positive and negative ions. Show that the Madelung constant is 1.386. (*Hint*: $\ln (1 + x) = x - x^2/2 + x^3/3 - x^4/4 + \ldots$)

22.11 What type of bonding mechanism can we expect for (a) the KCl molecule, (b) the CH_4 (methane) molecule, (c) the CO_2 molecule, (d) the Al atoms in a solid, and (e) the O_2 molecule?

22.12 We indicated in Section 22.3a that in the NaCl molecule the bonding is ionic. Actually, the wavefunction of the 3*s* electron that is transferred from Na to Cl, although small, is not zero in the space between the ions. As a result, the 3*s* electron spends some time between the two atoms (is shared by both atoms). Thus, the bonding, although mostly ionic, is to a small degree covalent. Consider the NaCl molecule as an electric dipole, that is, two equal charges, one positive, one negative ($q = 1.6 \times 10^{-19}$ C) separated by distance $d = 2.4$ Å (the equilibrium separation of the ions in the molecule). (a) What is the electric dipole moment, $\mu_e = qd$, of the molecule? (b) The measured electric dipole moment is 3.00×10^{-29} C-m, to what degree is the bond ionic; that is, what is the ratio of the actual dipole moment to the theoretical one?

(*Answer*: (a) 3.84×10^{-29} C-m, (b) 0.78.)

22.13 Barium has two valence electrons, whereas oxygen needs two electrons to achieve an inert electron configuration. One would expect that in the BaO molecule the binding would be ionic, the barium atom giving two electrons to the oxygen atom. The equilibrium separation between the ions in the BaO molecule is 1.94 Å, and the measured electric dipole moment of the molecule is 2.65×10^{-29} C-m. To what degree is the bond ionic? (See Problem 22.12.)

22.14 In Section 22-3b we indicated that one of the possible eigenfunctions for the electrons in the H_2 molecule is $\chi = \chi_1 + \chi_2$, where χ_1 and χ_2 are the ground-state eigenfunctions of the individual atoms (see Fig. 22-10). In the ground state of the H_2 molecule, both electrons, with opposite spins, have this particular eigenfunction associated with them. The ionization energy (energy to remove one electron) of H_2 is 15.7 eV, whereas for H it is 13.6 eV. Explain why the ionization energy is greater for the hydrogen molecule than for the hydrogen atom.

CHAPTER 23

Free Electron Theories of Solids

23.1 INTRODUCTION

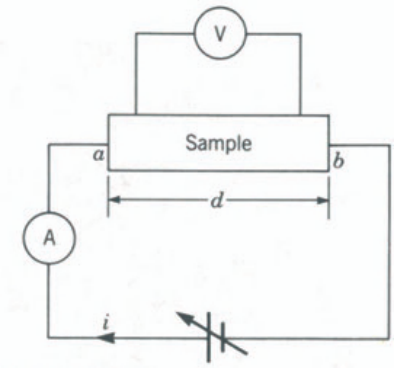


FIGURE 23-1

If a battery is connected in series with an ammeter and a piece of some material (copper, silicon, glass, or such), a deflection in the ammeter will indicate a flow of charged particles (a current i) through the circuit (see Fig. 23-1). Simultaneously there will be a voltage drop V across the sample as measured by a voltmeter. If the voltage of the battery is varied, i will vary proportionally with V. It can therefore be stated that if a voltage V is applied across a sample, a current proportional to it will flow:

$$V = Ri \tag{23.1}$$

where R is the proportionality *constant* and is called the resistance of the sample. Equation 23.1 is known as Ohm's law and was introduced in Chapter 15.

Ohm's law can be restated in a form more appropriate to the understanding of the phenomenon of conduction by concentrating attention on the sample. The fact that there is a potential difference V across the sample means that there is an electric field $\mathscr{E}$ in the sample. If the sample is uniform in geometry and quality, $\mathscr{E}$ will be constant, and it follows that

$$V_{ab} = V = \int_a^b \mathscr{E}\, dx = \mathscr{E}d$$

$$V = \mathscr{E}d \tag{23.2}$$

where d is the length of the sample.

Given a certain potential difference (and therefore a certain $\mathscr{E}$), the larger the cross-sectional area A of the sample (Fig. 23-2), the larger the current will be. (Think of the analogy of a pipe with flowing water.) We can eliminate the geometric parameter A by introducing a new quantity, the current density J, defined as the current per unit cross-sectional area (see Section 15.3).

$$J = \frac{i}{A} \quad \text{or} \quad i = JA \tag{23.3}$$

Substituting Eq. 23.2 for V and Eq. 23.3 for i into Ohm's law, Eq. 23.1, results in the relation

$$\mathscr{E}d = RAJ \quad \text{or} \quad \mathscr{E} = R\frac{A}{d}J$$

and

$$\mathscr{E} = \rho J \tag{23.4}$$

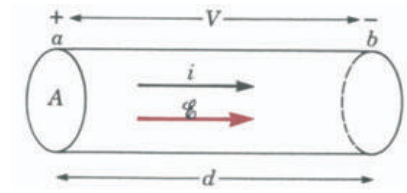


FIGURE 23-2

where the quantity $\rho = RA/d$ is called the *electrical resistivity* and has dimensions of ohm-meters (Ω-m). Ohm's law can be expressed another way by introducing the term *electrical conductivity*, $\sigma = 1/\rho$, which has dimensions of reciprocal ohm-meters $(\Omega\text{-m})^{-1}$. Equation 23.4 can then be expressed as

$$J = \sigma \mathscr{E} \tag{23.5}$$

Equations 23.4 and 23.5 are two equivalent forms of Ohm's law. Electrical resistivity ρ and electrical conductivity σ are both quantities characteristic of the material and independent of the geometry of the sample.

One of the reasons that the electrical properties of solids are so interesting and have drawn so much attention can be found by looking at Table 23-1, which lists the electrical conductivity of example materials at room temperature. The first and obvious thing one notices is the tremendous range of values, about 27 orders of magnitude (powers of 10). Any satisfactory theory of electrical conduction must be able to explain not only Ohm's law but also such large differences in conductivity. There are additional effects to be explained such as the effect of temperature, impurities, and alloying on the electrical conductivity of solids.

Paul Drude (1863-1906).

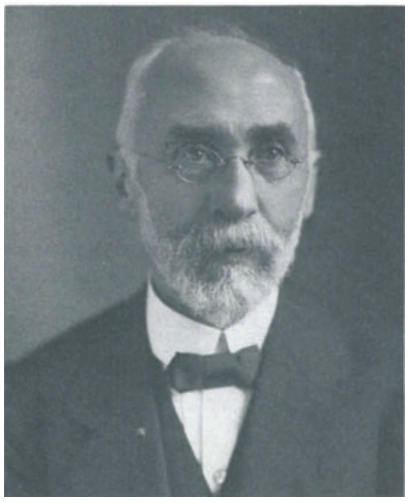
TABLE 23-1
Electrical Conductivity[a]

Material	σ $(\Omega\text{-m})^{-1}$
Copper	6×10^{7}
Aluminum	3×10^{7}
Iron	1×10^{7}
Germanium	2×10^{-2}
Silicon	4×10^{-4}
Glass	2×10^{-11}
Amber	2×10^{-15}
Polystyrene	1×10^{-20}

[a]These values of σ are at room temperature.

23.2 CLASSICAL FREE ELECTRON (CFE) MODEL

The first successful attempt to understand the electrical properties of metals was presented by P. Drude in 1900 and was extended by H. A. Lorentz in 1909. The results of these two investigators, as well as the work of others, is now called the *classical free electron model* (CFE).



H.A. Lorentz (1853–1929).

23.2a Assumptions of the Model

1. The main assumption of the CFE model is that a metal is composed of an array of ions with *valence electrons that are free* to roam through the ionic array with the only restriction that they remain confined within the boundaries of the solid. Because these valence electrons are responsible for the electrical conduction of the solid, they are called *conduction electrons.*
2. The mutual repulsion between the negatively charged electrons is neglected. In addition, the potential energy due to the ions is assumed to be everywhere constant. These free electrons are basically treated as an ideal

neutral gas that obeys classical Maxwell-Boltzmann statistics (see Supplement 9-1 at the end of Chapter 9).

3. In the absence of an electric field, these electrons are moving with their random thermal velocities given by the ideal gas result (Eq. 9.21)

$$\frac{1}{2}m\overline{v^2} = \frac{3}{2}k_BT \tag{23.6}$$

where $\overline{v^2}$ is the average of the square of the thermal speeds; k_B is Boltzmann's constant, 1.38×10^{-23} J/K; and T is the absolute temperature of the solid and therefore of the electron gas.

4. When an electric field $\mathscr{E}$ is applied to the solid, the free electrons acquire an average drift velocity in the direction opposite to the electric field, thus producing an electric current.

Despite the many simplifying assumptions, this CFE model was able to explain many properties of metals in a quantitative and satisfactory way. As we will show, it was able to predict Ohm's law, as well as the Wiedemann-Franz law, an empirical formula relating the electrical and the thermal conductivities in metals, which will be discussed later. The CFE model was not totally adequate; subsequent models started with some of the basic assumptions of the Drude-Lorentz model but modified them with new concepts. (Remember that quantum mechanics did not come onto the scene until 1926.)

23.2b Ohm's Law: Derivation from the CFE Model

Let us consider a piece of metal. According to the CFE model, there are a certain number of free electrons moving randomly with thermal velocities whose average value is obtained from Eq. 23.6. At room temperature, $T = 300$ K, this is

$$v_{\text{RMS}} = \left[\frac{3\,k_BT}{m}\right]^{1/2} = \left[\frac{3 \times 1.38 \times 10^{-23}\ \text{J/K} \times 300\ \text{K}}{9.1 \times 10^{-31}\ \text{kg}}\right]^{1/2}$$

$$= 1.2 \times 10^5\ \text{m/sec} \sim 75\ \text{miles/sec}$$

where v_{RMS} stands for root mean square (the square root of the mean of the square). Random motion means that, if you consider a hypothetical plane in the sample (Fig. 23-3), the rate at which electrons pass from the right to the left is the same as the rate at which they pass from the left to the right. Therefore, the net flow in any direction is zero, and there is accordingly no net electrical current. The way to obtain a net current is to bias the random motion of the electrons by applying an external force, that is, creating an electric field in the sample. This may be done by applying a potential difference

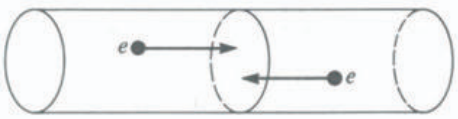


FIGURE 23-3

across the sample. This biasing will have the effect of superimposing a net constant drift velocity on the random motion in the direction of the force. This concept was first introduced in Chapter 15 but will be repeated here as the starting point of further development of the theory.

In Fig. 23-4 it is seen that, although the electrons bounce back and forth from collisions, there is a net drift velocity v_d toward the right. One can visualize what is happening by thinking of a simple mechanical analogy. Consider a smooth, frictionless table with fixed pins uniformly distributed throughout. If you drop a large number of marbles on the table when it is flat, they will hit the pins and move randomly with no net motion in any given direction; that is, on the average as many marbles will hit one end of the table as the opposite one. However, if you tilt the table, then at the same time that the marbles bounce back and forth randomly from the pins, there is a net drift of all the marbles toward the lower end. Tilting the table effectively creates a potential difference (gravitational) between the higher and lower ends of the table. Why don't the marbles acquire an acceleration in the direction of the force? Why only a constant drift velocity? The reason is the collisions with the pins. These collisions impede the downward motion of the marbles and therefore act as a retarding frictional force. Obviously, the more collisions per unit time, the greater the time rate of change of the forward momentum of the marbles—that is, the greater the retarding force. The drift velocity v_d represents an equilibrium condition in which the rate at which the energy they receive from the external gravitational force equals the rate at which they lose it through collisions with the pins, which is the energy dissipated as heat. The situation is governed by Newton's second law in which the vector sum of forces is zero, and therefore there can be no acceleration between the top and bottom of the table although there are local accelerations between the pins. This analogy is close to the free electron model of the motion of electrons through a metal. An externally applied electric field replaces the gravitational field of the tilted table, and the metal ions replace the pins. The marbles are now electrons, and their motion resulting from the force of the electric field is impeded by successive collisions with ions. We will now derive Ohm's law on the basis of this model.

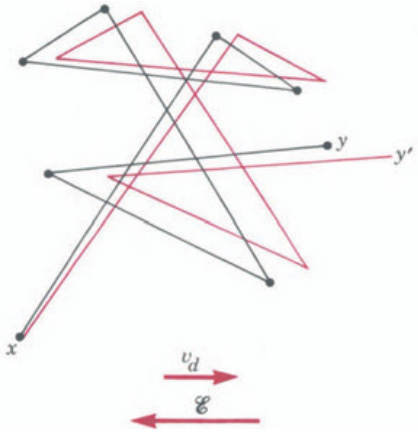


FIGURE 23-4
Random path of the electrons resulting from collisions with the ions (black lines) and the effect of an electric field on the path, which results in a net drift velocity v_d in the direction opposite to the direction of the electric field (colored lines). (*Source*: David Halliday and Robert Resnick, *Fundamentals of Physics*, 2nd ed. Copyright © 1981 by John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.)

The definition of current is the rate at which an amount of charge Δq passes a given point, or

$$i = \frac{\Delta q}{\Delta t} \tag{23.7}$$

In a time Δt the charges move a distance $x = v_d\,\Delta t$. All charges that were to the left of plane A in Fig. 23-5 up to a distance $x = v_d\,\Delta t$ will pass by A in

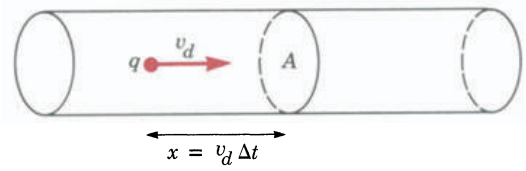


FIGURE 23-5

Δt. If the number of charges per unit volume of the metal is N, then the number of charges that will flow to the right across A in time Δt will be N times the volume of a cylinder with cross-sectional area A and length x, or

$$\text{Number of charges} = N A v_d \Delta t$$

If q is the charge of each electron, then the amount of charge Δq that crosses A in time Δt is

$$\Delta q = qNAv_d \Delta t \tag{23.8}$$

Substitution of Eq. 23.8 in Eq. 23.7 obtains

$$i = qNAv_d$$

and by the definition of current density, Eq. 23.3, we have

$$J = qNv_d \tag{23.9}$$

$J = qNv_d$

The dependence of J on the external electric field $\mathscr{E}$ is contained in v_d.

Consider the negative charge q shown in Fig. 23-6. The force acting on the charge resulting from the electric field $\mathscr{E}$ is $F = q\mathscr{E}$. From Newton's second law

$$F = ma = q\mathscr{E}$$

This may be written as

$$m \frac{dv}{dt} = q\mathscr{E} \tag{23.10}$$

and

$$dv = \frac{q\mathscr{E}}{m} dt$$

The velocity v at any time t is obtained by integration

$$\int_0^v dv = \frac{q\mathscr{E}}{m} \int_0^t dt \tag{23.11}$$

and

$$v_d = \frac{q\mathscr{E}}{m} t \tag{23.12}$$

where

$$\frac{q\mathscr{E}}{m} = a$$

by analogy to

$$v = v_0 + at$$

If t in Eq. 23.12 is taken as an average time τ between collisions and v_d from

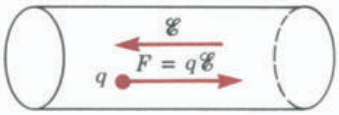


FIGURE 23-6

Eq. 23.9 is substituted into Eq. 23.12, then

$$\frac{J}{qN} = \frac{q\mathscr{E}}{m}\tau$$

or

$$J = \frac{Nq^2\tau}{m}\mathscr{E} \tag{23.13}$$

Comparing Eq. 23.13 with Eq. 23.5, we conclude that the conductivity σ is

$$\sigma = \frac{q^2N\tau}{m} \tag{23.14}$$

The Drude-Lorentz theory predicts one of the key empirical facts about the electrical conductivity, Ohm's law, and has given an explicit expression for electrical conductivity. To prove Ohm's law fully, it must be shown that the expression for σ, Eq. 23.14, is independent of $\mathscr{E}$. This is equivalent to showing τ independent of $\mathscr{E}$, because q, N, and m by definition are independent of the electric field. In Fig. 23-7, l (*mean free path*) is taken as the average separation between ions. If $\bar{v}$ is the average velocity of the electrons, which is a combination of the thermal velocity v_{RMS} (determined by the temperature alone) and the drift velocity v_d (which as we have seen depends on $\mathscr{E}$), we may write

$$\tau = \frac{l}{\bar{v}}$$

A simple calculation will show that $v_d << v_{\text{RMS}}$ in most cases; hence v_d may be neglected in the calculation of τ (see Example 23-1). Therefore the collision time τ is given as

$$\tau = \frac{l}{v_{\text{RMS}}} \tag{23.15}$$

FIGURE 23-7
In the CFE model the *mean free path* l (average distance between collisions of the electrons with the ions) is assumed to be the average interionic separation.

$$\sigma = \frac{q^2N\tau}{m}$$

Example 23-1

Consider a copper wire of cross-sectional area $A = 1$ mm^2 carrying a current $i = 1$ amp. What is the drift velocity v_d of the electrons? Cu is monovalent, that is, there is one free electron per atom. The density and the molecular weight of Cu are 9 g/cm^3 and 64 g/mole, respectively.

Solution From Eq. 23.9

$$v_d = \frac{J}{Nq}$$

The current density J is given by

$$J = \frac{i}{A} = \frac{1 \text{ amp}}{10^{-6} \text{ m}^2} = 10^6 \text{ amp/m}^2$$

Because copper is monovalent, the number of free electrons per unit volume N is equal to the number of atoms per unit volume N_{Atoms}. The latter can be found as follows:

$$N_{\text{Atoms}} = (\text{number of moles/cm}^3) \times (\text{number of atoms/mole})$$

The number of atoms per mole is given by Avogadro's number $N_A = 6.02 \times 10^{23}$ atoms/mole. Thus

$$N = N_{\text{Atoms}} = \frac{9 \text{ g/cm}^3}{64 \text{ g/mole}} \times (6.02 \times 10^{23} \text{ atoms/mole})$$

$$= 8.4 \times 10^{22} \text{ atoms/cm}^3 = 8.4 \times 10^{28} \text{ atoms/m}^3$$

Substituting for J and N into the expression for v_d, we obtain

$$v_d = \frac{10^6 \text{ amp/m}^2}{8.4 \times 10^{28} \text{ atoms/m}^3 \times 1.6 \times 10^{-19} \text{ C}} = 7 \times 10^{-5} \text{ m/sec}$$

As we saw earlier, at $T = 300$ K $v_{\text{RMS}} = 1.2 \times 10^5$ m/sec. Thus $v_d << v_{\text{RMS}}$ and can be neglected as a contribution to $\bar{v}$. Consequently, the relaxation time τ of Eq. 23.15 is determined by l and v_{RMS}, both of which are independent of $\mathscr{E}$. Therefore, σ is independent of $\mathscr{E}$.

In addition to being able to explain Ohm's law, the CFE was also able to explain the experimental fact that good electrical conductors are also good thermal conductors (see Supplement 23-1 at the end of this chapter).

23.2c Failures of the CFE Model

The CFE model is conceptually easy and succeeds in explaining certain important experimental facts. It fails, however, to explain other experimental results. We will show later how many of the failures are corrected by the quantum mechanical model. The basic form of the conductivity equation will remain the same, the main difference being in the calculation of the time between collisions. What are some of the failures of the classical theory?

1. Specific Heat

As we indicated earlier, according to the CFE model, the valence electrons in a solid behave as an ideal neutral gas. In Chapter 9, Eq. 9.32, we showed that the molar specific heat of a gas at constant volume is

$$C_v = \frac{3}{2} R \tag{9.32}$$

where R is the universal gas constant. The contribution of the conduction electrons to the specific heat of a solid should therefore be given by Eq. 9.32.

Instead, experimentally the electronic specific heat per mole is

$$C_v = 10^{-4}\, RT \tag{23.16}$$

$C_v = 10^{-4}\, RT$

Not only is the CFE model grossly incorrect in the actual value, but it predicts that the C_v is temperature independent, whereas the experiment shows that C_v is proportional to T. This represents an outstanding failure of the model. It is particularly puzzling in view of the success in explaining the electrical and thermal conductivities. In the CFE model it is impossible to comprehend how the valence electrons can participate in transport processes as if they are free and yet give such a small contribution to the specific heat, that is, be able to absorb heat.

2. Temperature Dependence of σ

Experimentally, for a metal (except at very low temperatures) the electrical conductivity σ is inversely proportional to the temperature T: $\sigma \propto T^{-1}$. According to the CFE model, Eq. 23.14,

$\sigma \propto T^{-1}$

$$\sigma = \frac{q^2 N \tau}{m}$$

The only quantity that is temperature dependent is the time between collisions $\tau = l/v_{\text{RMS}}$. The average mean free path l will become smaller with increasing temperature because the ions vibrate with larger amplitude, thus increasing the chances of collision and therefore reducing the effective l. The thermal velocity v_{RMS} is also temperature dependent and is given by Eq. 23.6.

$$\frac{1}{2} m v_{\text{RMS}}^2 = \frac{3}{2} k_B T \tag{23.6}$$

Therefore

$$v_{\text{RMS}} = \left(\frac{3k_B T}{m}\right)^{1/2}$$

which shows that v_{RMS} increases proportionally to the square root of the temperature. If we assume that the change of v_{RMS} with temperature is much larger than the change of l, we may write

$$\tau = \frac{l}{\left(\frac{3k_B}{m}\right)^{1/2}} \quad T^{-1/2}$$

Substituting this into the equation for σ, Eq. 23.14, we find that the temperature dependence of the electrical conductivity should be proportional to the reciprocal of the square root of the temperature

$$\sigma \propto T^{-1/2}$$

Thus, the CFE model predicts the incorrect temperature dependence of σ, which experimentally is $\sigma \propto T^{-1}$.

23.3 QUANTUM-MECHANICAL FREE ELECTRON MODEL (QMFE)

Arnold Sommerfeld (1868–1951).

Many of the difficulties encountered by the classical free electron model were removed with the advent of quantum mechanics. In 1928, A. Sommerfeld modified the free electron model in two important ways:

1. *The electrons must be treated quantum mechanically.* This will quantize the energy spectrum of the electron gas.
2. *The electrons must obey Pauli's exclusion principle;* that is, no two electrons can have the same set of quantum numbers.

As a result of these modifications, when we put an electron gas in a solid, we begin by putting the electrons in the lowest energy states available, while obeying the exclusion principle, until we have used all the available electrons. This is to be contrasted with the classical free electron gas in which the electrons can assume continuous energy values, with many electrons having the same energy. This has profound implications for the statistical distribution of energies (the average number of electrons having a certain energy E) that the electrons can have. Thus, whereas a classical gas will obey Maxwell-Boltzmann statistics (Supplement 9-1, Chapter 9), the quantum-mechanical gas will follow a new type of statistical distribution known as the Fermi-Dirac distribution (Supplement 23-2 at the end of this chapter). This in turn will affect the way the electron gas can absorb energy from an external source, such as a heat source, and the way it responds to an electric field.

Aside from these two key modifications, Sommerfeld kept most of the assumptions of the Drude model:

1. *The valence electrons are free* to move through the solid.
2. Aside from collisions with the ions, *the electrostatic interaction between the electrons and the lattice ions is ignored.*
3. *The interaction between the electrons is also neglected.*

Essentially, the valence electrons retain the main features of an ideal gas but a gas that must be treated quantum mechanically rather than classically.

23.3a Three-Dimensional Infinite Potential Well

If we want to treat a free electron gas quantum mechanically, the first question to ask is: What is the potential[1] in which the electrons find themselves? We can start by noting the empirical fact that there are no electrons beyond the boundaries of the metal. There is some force keeping the electrons inside. This might be a large electrical potential barrier at the boundaries. And what about inside? How will the potential energy of an electron vary in the presence

[1]As indicated in Chapter 20 (Section 20.2), the potential energy is often referred to as the potential.

of the great number of ions and other electrons? Because we have decided to neglect interactions with ions and other electrons in this first approximation, *the potential energy inside will be uniform.* These are, of course, sweeping assumptions, but the model works to a great extent. We will improve on it later. For now, let us consider the case of a uniform potential energy inside and an infinite potential energy at the boundaries of the solid so that no electrons can escape.

We have previously solved in Chapter 20 the problem of a particle that finds itself in a potential well in which the potential energy $E_p(x)$ is given by the condition (see Fig. 23-8),

$$E_p(x) = \begin{cases} \infty & \text{for} \quad 0 > x > a \\ 0 & \text{for} \quad 0 \leq x \leq a \end{cases}$$

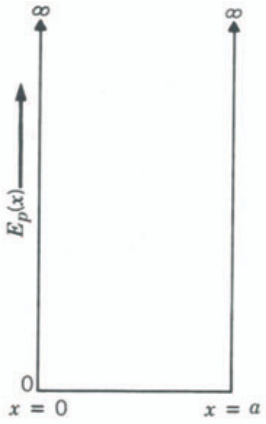


FIGURE 23-8
One-dimensional infinite potential well of width a.

The Schrödinger equation inside the well was given by Eq. 20.18.

$$-\frac{\hbar^2}{2m}\frac{d^2\chi}{dx^2} = E\chi \tag{20.18}$$

or

$$\frac{d^2\chi}{dx^2} + k^2\chi = 0 \tag{20.19}$$

where

$$k = \sqrt{\frac{2mE}{\hbar^2}}$$

The solutions for χ were $\chi = B \sin kx$, Eq. 20.23, inside and $\chi = 0$ outside. When the boundary condition $\chi = 0$ at $x = a$ was imposed, a restriction on the acceptable values of k and hence of E was found: Eqs. 20.24 and 20.26

$$k_n = n\frac{\pi}{a} \quad \text{where } n = 1, 2, 3, \ldots \tag{20.24}$$

and, correspondingly,

$$E_n = n^2\frac{\pi^2\hbar^2}{2ma^2} \tag{20.26}$$

The extension of the problem to three dimensions, that is, a particle moving in an arbitrary direction in a three-dimensional uniform potential well of infinitely high potential walls, is rather straightforward. Because each coordinate is at right angles to the others, their derivatives are independent. The Schrödinger equation in this case becomes

$$-\frac{\hbar^2}{2m}\left\{\frac{\partial^2\chi}{\partial x^2} + \frac{\partial^2\chi}{\partial y^2} + \frac{\partial^2\chi}{\partial z^2}\right\} = E\chi \tag{23.17}$$

Although this is a more complex partial differential equation, the solution may be readily found by the separation of variables method that we have

seen previously (see Section 20.2d); that is, we assume a solution of the form

$$\chi(x,y,z) = \chi_1(x)\,\chi_2(y)\,\chi_3(z) \tag{23.18}$$

When we substitute Eq. 23.18 into Eq. 23.17, we obtain three ordinary differential equations for χ_1, χ_2, and χ_3, each identical to the one-dimensional case; that is, the solutions for χ_1, χ_2, and χ_3 will be of the form $\chi = B \sin kx$ (Eq. 20.23). Therefore, the solutions for $\chi\ (x, y, z)$ will be

$$\chi(x,y,z) = A\ (\sin k_1x)(\sin k_2y)(\sin k_3z)$$

where

$$A = B^3$$

As before, when the boundary conditions $\chi(x, y, z) = 0$ for $x = a$, $y = a$, and $z = a$ are imposed, we obtain

$$k_1 = n_1 \frac{\pi}{a} \quad \text{where } n_1 = 1, 2, 3, \ldots$$

$$k_2 = n_2 \frac{\pi}{a} \quad \text{where } n_2 = 1, 2, 3, \ldots$$

$$k_3 = n_3 \frac{\pi}{a} \quad \text{where } n_3 = 1, 2, 3, \ldots$$

And just as in the one-dimensional case, the energy of the electrons becomes quantized but with three quantum numbers specifying the state:

$$E_{n_1n_2n_3} = \frac{\pi^2\hbar^2}{2ma^2}(n_1^2 + n_2^2 + n_3^2) \tag{23.19}$$

$E_{n_1n_2n_3} = \frac{\pi^2\hbar^2}{2ma^2}(n_1^2 + n_2^2 + n_3^2)$

23.3b Occupancy Conditions at T = 0 K

One important result from having three quantum numbers to specify the energy is that there may be *degenerate states* (states having the same energy but different quantum numbers). The lowest energy available will be for $n_1 = 1$, $n_2 = 1$, and $n_3 = 1$, which we represent in simplified notation as the (1, 1, 1) state. The energy of this state is $E = 3\,E_0$, where $E_0 = \pi^2\hbar^2/2ma^2$. It is nondegenerate. The next higher energy level is (2, 1, 1) or (1, 2, 1) or (1, 1, 2) for which $E = 6\,E_0$. It is a threefold degenerate level. The next one is (2, 2, 1) or (2, 1, 2) or (1, 2, 2), $E = 9\,E_0$, and so on. A summary of the first few levels is shown in Fig. 23-9.

Let us now consider, for simplicity, an electron gas made up of 34 electrons. At absolute zero ($T = 0$ K) the gas will take the lowest energy available to it.

Classically, because we know that the average energy of each electron is $3/2\ k_BT$, the average will be zero, and because the kinetic energy is always positive, this implies that each electron will have exactly zero energy.

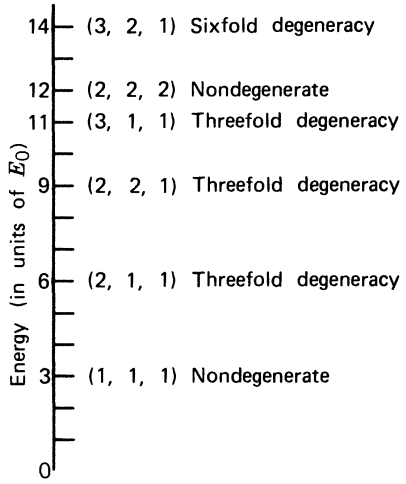


FIGURE 23-9
Schematic representation of the first few energy levels and the degree of degeneracy of each for an electron in a three-dimensional infinite potential well.

Quantum mechanically, the story is quite different. To begin with, $E = 0$ is not an allowed energy state. The solution of the three-dimensional potential well gives the lowest energy state available as $E = 3\,E_0$. We may be tempted to put all 34 electrons of our hypothetical gas in that state as the ground state. But this would violate Pauli's exclusion principle (Section 21.7a), and therefore this arrangement is not possible.

Let us see how we obtain the ground state for our gas of 34 electrons. In the state $E = 3\,E_0$ we can put two electrons: They will have the same set of quantum numbers, n_1, n_2, n_3, but their spins must be of opposite sign; one will have $m_s = \frac{1}{2}$ and the other $m_s = -\frac{1}{2}$. In the state $E = 6\,E_0$, we can put six electrons: two in the (2, 1, 1) state—these two electrons will have different spins; two in the (1, 2, 1) state with different spins; and finally two in the (1, 1, 2) state—again these two electrons will have different spins. In $E = 9\,E_0$, we can place six more electrons. In $E = 11\,E_0$, we can place six more electrons. In $E = 12\,E_0$, we can place only two because the energy state is nondegenerate. In $E = 14\,E_0$, we can place 12 electrons, two in each of the six degenerate states. All 34 electrons have been placed in the lowest energy configuration. The ground state of the gas is one in which the electrons have energies that range between $3\,E_0$ and $14\,E_0$. We can state this fact as a probability statement: At absolute zero temperature the probability that a level below $14\,E_0$ will be occupied is 1. The probability that a level above $14\,E_0$ will be occupied is 0 (all levels are empty). The highest filled energy level is called the *Fermi energy* at absolute zero temperature, E_F (0). In our example E_F (0) $= 14\,E_0$.

This probability statement is a particular case of the new type of energy distribution function that must be used to describe a quantum mechanical gas obeying the Pauli exclusion principle; it is known as *the Fermi-Dirac distribution function* $F(E)$ (Fig. 23-10). $F(E)$ is the probability that a state with energy E is occupied by an electron. At $T = 0$ K

$$F(E) = \begin{cases} 0 & \text{for } E > E_F \\ 1 & \text{for } E < E_F \end{cases} \qquad (23.20)$$

Let us attack a practical and important problem. We have considered, for pedagogical reasons, a gas made up of 34 electrons. How do you determine occupation levels when you have a gas of 10^{28} electrons/m^3, which is the problem that we face? The two important questions are: How do we determine E_F for a large number of electrons? How do we find how many electrons we can place within a certain energy range ΔE?

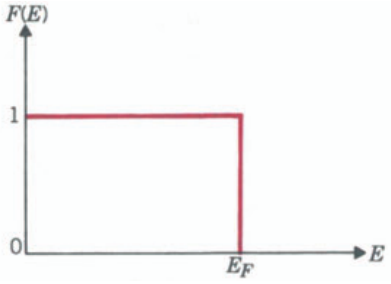


FIGURE 23-10
Fermi-Dirac distribution function at $T = 0$ K.

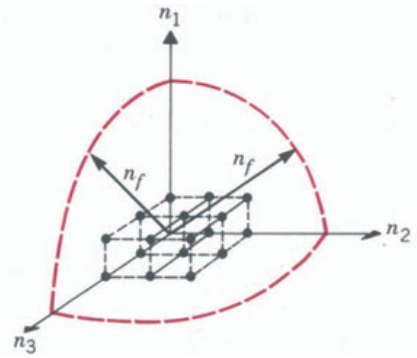


FIGURE 23-11
Scheme to represent the allowed energy states of an electron in a three-dimensional infinite potential well. The axes correspond to the three quantum numbers that specify the energy of the electron, Eq. 23-19. A point is plotted for each integral value of n_1, n_2, and n_3, that is, each point in this three-dimensional n-space corresponds to an allowed value of the energy. Here n_f is the distance of the occupied points farthest away from the origin.

23.3c Calculation of E_F at $T = 0$ K

To determine the two quantities just mentioned without counting, we can use a standard trick of physics. Let us construct a lattice in n-space; that is, let us have a set of three mutually perpendicular axes. Let one of the axes correspond to n_1, the others to n_2 and n_3 (see Fig. 23-11). Let us plot a point for each integral value of n_1, n_2, and n_3. In so doing we will get a uniform three-dimensional array of points. Each point in the array corresponds to a

particular allowed value of the energy (corresponding to a particular set of quantum numbers n_1, n_2, n_3). There are two facts that we must keep in mind.

1. In a unit of volume of this n-space there is only *one point*. That is, we can fill the whole space with $1 \times 1 \times 1$ cubes in such a way that there is one point at the center of each cube.
2. As we have seen in our early example, although each point has a different set of quantum numbers n_1, n_2, n_3, some of the points correspond to states with the same energy (they are degenerate). How do we identify them?

Suppose that we write the energy, Eq. (23.19), as follows:

$$E = E_0 (n_1^2 + n_2^2 + n_3^2) = E_0 n^2 \tag{23.21}$$

where

$$n^2 = n_1^2 + n_2^2 + n_3^2$$

We can identify the degenerate states by understanding the geometric significance of n, which is nothing else but the distance (in n-space) from the particular point to the origin; that is, the three-dimensional pythagorean formula with n_1, n_2, and n_3 as the distances instead of the usual x, y, and z. Consequently, with a little thought we can conclude that points equidistant from the origin correspond to degenerate states. In addition, we see from Eq. 23.21 that the farther away from the origin a given point is, the higher the energy of the state represented by that point.

We can now proceed to answer the question: What is the E_F for a gas of 10^{28} electrons/m^3? Because E_F (0) is the highest occupied energy level at $T =$ 0 K, we can write

$$E_F (0) = E_0 n_f^2 \tag{23.22}$$

where n_f is the distance of the occupied points farthest away from the origin. These points, being equidistant from the origin, lie on a spherical shell of radius n_f (see Fig. 23-11). Because only positive values of n are permitted, only the positive octant of the sphere is used, which is one-eighth of a whole sphere. All points inside that shell are occupied by two electrons (corresponding to the two possible values of the spin). Therefore, we may write that

Total number of electrons in the octant =

(2) $\times$ (one-eighth the volume of a sphere of radius n_f)

$\times$ (number of points per unit volume)

But we can write that the total number of electrons is the number N of free electrons per unit volume of real space times the volume of the three-dimensional well, a^3. Thus,

$$N a^3 = (2) \left(\frac{1}{8} \times \frac{4}{3} \pi n_f^3 \right) (1) \tag{23.23}$$

Solving for n_f from Eq. 23.23, we obtain the value for n_f

$$n_f = \left(\frac{3\,N\,a^3}{\pi}\right)^{1/3}$$

and, substituting this value of n_f and $E_0 = \pi^2\hbar^2/2ma^2$ into Eq. 23.22, we obtain

$$E_F\,(0) = \frac{\hbar^2\pi^2}{2ma^2}\left(\frac{3Na^3}{\pi}\right)^{2/3}$$

We may algebraically simplify this equation to

$$E_F\,(0) = \frac{\hbar^2}{2m}\,(3N\,\pi^2)^{2/3} \qquad (23.24)$$

$E_F\,(0) = \frac{\hbar^2}{2m}\,(3N\,\pi^2)^{2/3}$

This expression is the approximate value of the Fermi level in a solid at absolute zero temperature. All that is needed is N, the density of the free electrons.

Example 23-2

We showed in Example 23-1 that the number of free electrons in copper is 8.4×10^{28} electrons/m^3. (a) Calculate the Fermi energy for Cu. (b) At what temperature T_f will the average thermal energy k_BT_f of a gas be equal to that energy?

Solution

(a) Substituting for the different parameters in Eq. 23.24, we obtain

$$E_F\,(0) = \frac{(1.05 \times 10^{-34}\ \text{J-sec})^2}{2 \times 9.1 \times 10^{-31}\ \text{kg}} \times (3 \times 8.4 \times 10^{28}\ \text{m}^{-3} \times \pi^2)^{2/3}$$

$$= 11.1 \times 10^{-19}\ \text{J} = 6.95\ \text{eV}$$

(b) We can equate this energy to k_BT_f to find the temperature at which the average thermal energy would be equal to the Fermi energy of Cu.

$$T_f = \frac{E_F\,(0)}{k_B} = \frac{11.1 \times 10^{-19}\ \text{J}}{1.38 \times 10^{-23}\ \text{J/K}} = 80{,}500\ \text{K}$$

This result illustrates the profound changes introduced by the QMFE model. For a classical electron gas to have thermal energies similar to those that a quantum mechanical electron gas has at $T = 0$ K, the temperature of copper should be about 60 times higher than its melting temperature, 1356 K.

23.3d Density of States

In our previous discussion, we have seen that the energies that an electron can take in a solid form a discrete quantized spectrum, the separation between

adjacent levels being (Eq. 23.19)

$$\Delta E \sim \frac{\hbar^2 \pi^2}{2ma^2}$$

If we take a macroscopic sample, for example, $a = 1$ cm, we may calculate the approximate energy difference between levels

$$\Delta E = \frac{(1.05 \times 10^{-34}\ \text{J-sec})^2\ \pi^2}{2 \times 9.1 \times 10^{-31}\ \text{kg} \times (10^{-2}\ \text{m})^2} = 0.6 \times 10^{-33}\ \text{J}$$

$$= 4 \times 10^{-15}\ \text{eV} \qquad (23.25)$$

We have seen in Example 23-2 that the Fermi energy for copper is 6.95 eV. For most metals, E_F is of the same order of magnitude. From Eq. 23.25, we conclude that the electrons in a solid occupy very closely spaced levels starting from very close to $E = 0$ eV all the way up to several electron volts. Thus, although the energy spectrum is quantized, the energy separation between adjacent levels is so small that we often use the term *quasicontinuous* to describe the electronic energy spectrum of a solid. As a result, even in a microscopically small energy interval dE there are many discrete energy states. We can therefore introduce the concept of *density of energy states*, which will simplify calculations considerably. Let $g(E)\,dE$ be the number of energy states available with energy between E and $E + dE$. Going back to our lattice in n-space, we consider an octant of a spherical shell of radius n and one of radius $n + dn$ (see Fig. 23-12). All the states between these two octants have energies that range between $E = n^2 E_0$ and $E + dE = (n + dn)^2 E_0$, Eq. 23.21. We choose dn so that dE is small, but because of the quasicontinuous nature of the spectrum, there are a large number of points within the octant shell. The density of energy states may be calculated in the following way.

$g(E)\,dE$ = (2) (volume between the octant shells)
× (number of points/unit volume)

where 2 is for the two allowed spin orientations.

$$g(E)\,dE = (2)\left(\frac{1}{8} \times 4\pi n^2\,dn\right)(1)$$

$$= \pi n^2\,dn \qquad (23.26)$$

We can express $g(E)$ in terms of the energy E instead of the state number n by making use of Eq. 23.21, $E = n^2 E_0$, or

$$n = \left(\frac{E}{E_0}\right)^{1/2}$$

Differentiating n with respect to E gives

$$dn = \frac{E^{-1/2}}{2E_0^{1/2}}\,dE = \frac{1}{2}\left(\frac{1}{EE_0}\right)^{1/2} dE$$

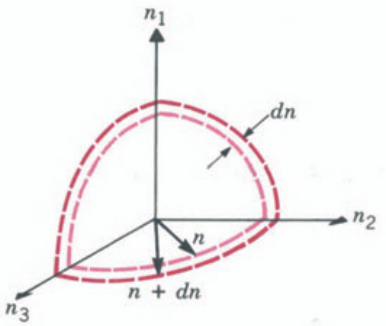


FIGURE 23-12
Two octants in the n-space lattice of Fig. 23-11 of radius n and $n + dn$, respectively. All points between these two octants have energies between $E = n^2 E_0$ and $E + dE = (n + dn)^2 E_0$.

Substituting for n and dn into Eq. 23.26, we obtain

$$g(E)\,dE = \frac{\pi}{2E_0^{3/2}} E^{1/2}\,dE$$

or

$$g(E)\,dE = C\,E^{1/2}\,dE \tag{23.27}$$

where

$$C = \frac{\pi}{2E_0^{3/2}}$$

Using the result from Eq. 23.19 that

$$E_0 = \frac{\hbar^2\pi^2}{2ma^2}$$

the constant C has the value

$$C = \frac{a^3(2m)^{3/2}}{2\hbar^3\pi^2} \tag{23.28}$$

A plot of $g(E)$ versus E is shown in Fig. 23-13. It is seen that the number of energy states $g(E)$ in an energy range dE increases with increasing energy. Although $g(E)$ gives the number of energy states available as a function of E, it does not tell us the number of electrons occupying those states. To obtain this, we must multiply $g(E)$ by the probability that a given energy state is occupied. This probability is given by the Fermi-Dirac distribution function, $F(E)$ (see Eq. 23.20). If we define $N(E)\,dE$ as the number of electrons with energy between E and $E + dE$, then

$$N(E)\,dE = g(E)\,F(E)\,dE \tag{23.29}$$

Because at $T = 0$ K, $F(E) = 1$ for $E < E_F$ and $F(E) = 0$ for $E > E_F$, Eq. 23.20, a plot of $N(E)$ versus E will look like that shown in Fig. 23-14.

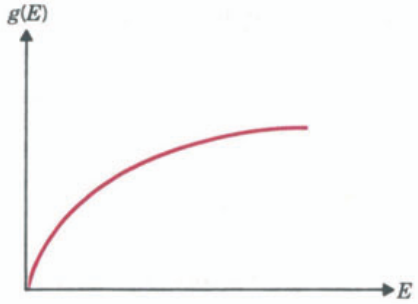


FIGURE 23-13
Density of states—that is, the number of energy states available to the electrons in a solid—as a function of the energy E.

$g(E)\,dE = C\,E^{1/2}\,dE$

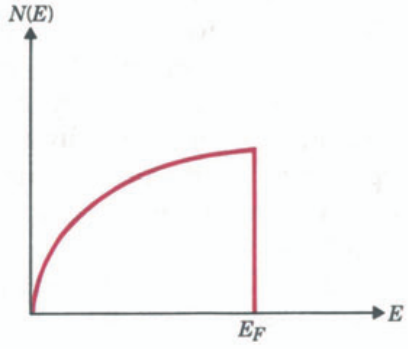


FIGURE 23-14
Energy distribution of the electrons in a solid at $T = 0$ K, that is, number of electrons having a certain energy E, as a function of the energy E.

Example 23-3

Calculate the average energy per electron in a metal at $T = 0$ K.

Solution We can use the standard method for the calculation of averages. That is, we take each value of E and multiply it by the number of electrons having that energy. We then add this product over all possible values of E and divide the result by the total number of electrons. In this case, the sum becomes an integral because E varies quasicontinuously.

$$\bar{E}\,(T = 0) = \frac{1}{N_{\text{total}}} \int_0^\infty N(E)\,E\,dE$$

Substituting for $N(E)$, Eq. 23.29,

$$\bar{E}\,(T = 0) = \frac{1}{N_{\text{total}}} \int_0^{E_F} C\,E^{1/2}\,E\,dE$$

$$= \frac{2}{5}\frac{C}{N_{\text{total}}} E_F^{5/2}$$

After substituting for C, Eq. 23.28, E_F, Eq. 23.24, and $N_{\text{total}} = Na^3$, where N is the number of free electrons per unit volume, we obtain

$$\bar{E}\,(T = 0) = \frac{3}{5} E_F \qquad (23.30)$$

$$\bar{E}\,(T = 0) = \frac{3}{5} E_F$$

23.3e Energy Distribution of Electrons for $T > 0$ K

So far we have confined our discussion concerning the energy distribution of the electrons to the case where $T = 0$ K. As we have seen, at $T = 0$ K, the electrons, seeking the lowest energy states available to them, occupy all the states below E_F, whereas the states above E_F remain empty.

What happens if the temperature is raised to some finite value of T? When an amount of thermal energy of about k_BT is available to the electrons, they will try to move to higher energy states. But this is not possible for all the electrons. Because the energy available is of the order k_BT, the electrons can jump to higher states lying up to k_BT above the states they occupied at $T = 0$ K. For ordinary temperatures the thermal energy is much smaller than the Fermi energy (for example, at $T = 300$ K, $k_BT = 0.025$ eV as compared with E_F being a few electron volts). The result is that only electrons whose energy is within a range k_BT below E_F can make transitions to higher energy levels. There are lots of *empty* levels above E_F. On the other hand, those electrons whose energy lies below E_F by an amount much greater than k_BT cannot jump to higher energy levels because these levels are occupied by other electrons and the Pauli exclusion principle prevents multiple occupancy. We can say that roughly only the fraction k_BT/E_F of the total number of electrons can be thermally excited to higher levels. Because $k_BT << E_F$, this represents a very small fraction. As a result, the probability of occupancy $F(E)$ for temperatures below the melting point of the solid does not differ greatly from that at $T = 0$ K.

The exact mathematical expression (derived in Supplement 23.2 of this chapter) for the probability of occupancy of an energy state E reflects the qualitative argument presented here and is called the *Fermi-Dirac distribution*. The expression is

$$F(E) = \frac{1}{\exp\left(\dfrac{E - E_F}{k_BT}\right) + 1} \qquad (23.31)$$

$$F(E) = \frac{1}{\exp\left(\dfrac{E - E_F}{k_BT}\right) + 1}$$

Note that if $T = 0$ K and $E < E_F$, Eq. 23.31 becomes equal to unity:

$$F(E) = \frac{1}{\exp\left(-\frac{\Delta E}{0}\right) + 1} = 1$$

In contrast, if $T = 0$ K but $E > E_F$, Eq. 23.31 is equal to zero:

$$F(E) = \frac{1}{\exp\left(\frac{\Delta E}{0}\right) + 1} = 0$$

We can also verify that if $E \leqslant E_F - k_B T$, then $F(E) \approx 1$. As an example, the probability of occupation for the state $E = E_F - 2k_B T$ is

$$F(E) = \frac{1}{\exp(-2) + 1} = 0.88$$

which is not very different from the value 1 at $T = 0$ K. Similarly, for $E = E_F + 2k_B T$, the probability of occupation is

$$F(E) = \frac{1}{\exp(2) + 1} = 0.12$$

A sketch of the effect of temperature on the Fermi distribution is shown in Fig. 23-15. At temperatures above zero there is a symmetric shift of electrons with energies slightly below E_F to states slightly above E_F, as seen in Fig. 23-15.

The Fermi-Dirac distribution gives a new definition of the Fermi level: *E_F is the energy at which the probability that the state be occupied is 1/2.* This can be seen by setting $E = E_F$ in Eq. 23.31.

$$F(E = E_F) = \frac{1}{\exp\left(\frac{E_F - E_F}{k_B T}\right) + 1} = \frac{1}{e^0 + 1} = \frac{1}{2}$$

For all T's except extremely high T's, E_F is essentially the same as $E_F(0)$. An exact calculation of $E_F(T)$ yields:

$$E_F(T) \approx E_F(0)\left\{1 - \frac{\pi^2}{12}\left(\frac{k_B T}{E_F(0)}\right)^2\right\}$$

where the second term is small compared with unity.

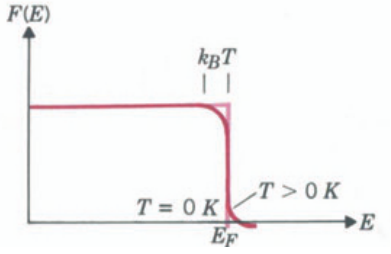


FIGURE 23-15
Comparison of the Fermi-Dirac function at $T = 0$ K and at $T > 0$ K. The probability of occupation changes only for energies such that $E - E_F \approx k_B T$. At room temperature, $k_B T << E_F$, and the function is barely changed.

23.3f Failures of the CFE Model Revisited

Let us now return to the two difficulties encountered by the CFE model that we discussed in Section 23.2c.

1. Specific Heat

The classical free electron model failed to predict correctly the molar electronic specific heat C_v. According to the CFE model (Section 23.2c)

$$C_v = \frac{3}{2} R$$

whereas experimentally (Eq. 23.16)

$$C_v = 10^{-4}\, RT \tag{23.16}$$

The heat required to increase ΔT the temperature of 1 mole of electrons is

$$\text{Heat} = C_v\, \Delta T$$

This heat input goes into increasing the internal energy of the electrons by ΔE

$$\Delta E = \text{Heat} = C_v\, \Delta T$$

or

$$C_v = \frac{\Delta E}{\Delta T} \tag{23.32}$$

This expression for C_v is correct only if C_v is constant, that is, independent of T. If it is not (as we know to be the case here), the expression holds only when all the quantities involved in the definition are infinitesimally small.

To obtain an expression for $C_v(T)$, we write the energy per mole as a function of T and differentiate with respect to T. This energy can be calculated exactly by using the density of states $g(E)$ and the probability of occupancy $F(E)$. Although this calculation can be done for T greater than 0 K, in practice it is a rather messy thing to do. We can get a good estimate of the value of E for thermally excited electrons by using a qualitative argument similar to the one we used before. At $T = 0$ K the average energy per electron is 3/5 E_F (0), Eq. 23.30. The energy of one mole of electrons, that is, N_A electrons will be 3/5 N_A E_F (0). At a temperature T, only those electrons near E_F can be excited. As we saw, for an electron with $E = E_F - 2k_BT$, the probability of occupation was 0.88 at $T = 300$ K, whereas it was 1 at $T = 0$ K. This means that the probability of being excited to higher levels is 0.12. For those electrons deeper down, the probability of excitation is essentially zero. As we indicated in Section 23.3e, we can assume as a rough estimate that only the fraction k_BT/E_F of the total number of electrons is able to gain thermal energy, and therefore, the number of electrons that can be excited is $N_A\, k_BT/E_F$. The additional internal energy ΔE gained by the electron gas when T is increased from 0 to T is therefore

$$\Delta E = (\text{number of excited electrons}) \times (\text{average energy gained})$$

If we assume that the average energy gained per electron is k_BT, we may write

$$\Delta E = \left(\frac{k_BT}{E_F} N_A\right)(k_BT)$$

Consequently,

$$E = \frac{3}{5} N_A E_F + \frac{k_B^2 N_A}{E_F} T^2$$

and from Eq. 23.32, in the case where ΔE and ΔT are infinitesimally small,

$$C_v = \frac{dE}{dT} = \frac{2k_B^2 N_A}{E_F} T$$

Or, if we use the fact that $k_BN_A = R$, we obtain

$$C_v = \frac{2k_B}{E_F} RT \tag{23.33}$$

If we take E_F at about 5 eV, a fairly typical value, then

$$\frac{2k_B}{E_F} = \frac{2 \times 1.38 \times 10^{-23} \text{ J/K}}{5 \text{ eV} \times 1.6 \times 10^{-19} \text{ J/eV}} \approx 10^{-4} \text{ K}^{-1}$$

and

$$C_v \approx 10^{-4} RT$$

which is in agreement with the experimental results, Eq. 23.16.

It is instructive to examine the details of the calculation of specific heat and try to understand why the quantum mechanical model succeeds whereas the classical one does not. In the classical model all the electrons are able to gain the thermal energy offered to them. The number is *large and constant*. In the quantum mechanical model, as a result of Pauli's exclusion principle, only the fraction k_BT/E_F can gain energy. Because $k_BT << E_F$ for the temperatures at which a solid remains solid, this is a small fraction of the total number of electrons and, moreover, the contribution to the specific heat by that fraction will depend linearly on T.

The exact result for the electronic specific heat is

$$C_{v(\text{mole})} = \frac{\pi^2 k_B}{2E_F} RT = \frac{4.93k_B}{E_F} RT$$

instead of

$$C_{v(\text{mole})} = \frac{2k_B}{E_F} RT$$

2. Electrical Conductivity

The CFE model predicted that $\sigma \propto T^{-1/2}$, whereas at most temperatures (except at very low T's) experiments show that $\sigma \propto T^{-1}$. We will now show that the quantized model yields the correct temperature dependence.

When an electric field is applied to a solid, the electrons acquire a net drift velocity $v_d = q\tau\mathscr{E}/m$ (where τ is the collision time and $\mathscr{E}$ is the electric field), Eq. 23.12. How does this affect the velocity distribution of the quantum mechanical electron gas? The number of electrons with energies between E and $E + dE$ is indicated in Fig. 23-16, which is a combination of Fig. 23-13 and Fig. 23-15. The velocities of the electrons associated with these energies are completely random; that is, there is an equal chance of finding an electron with a given v_x as with the same velocity in the opposite direction—the velocity distribution will be symmetric with respect to $v_x = 0$. This distribution can be found from the definition of kinetic energy, which in this case is the total energy E, and the density of states (Eq. 23.27). Because $E = 1/2\ mv^2$, $E^{1/2} = \pm\sqrt{m/2}\ v$ and $dE = mv\ dv$. Substituting for $E^{1/2}$ and for dE into Eq. 23.27 we obtain

$$g(v)\ dv \propto v^2\ dv$$

The distribution for the x component of v is illustrated in Fig. 23-17. If an electric field is applied in the $-x$ direction, giving rise to a force on the electrons in the $+x$ direction, the entire distribution will be shifted toward positive velocities, this is shown in Fig. 23-18. The preceding statement implies that all the electrons participate in the conduction process: They all acquire some extra energy from the electric field. This is surprising in view of what we have said concerning specific heats. In that case, only those electrons near E_F could be excited to higher energy levels. In the case of added thermal energy, the energy that the electrons acquire is randomly distributed among them and the velocity distribution is not shifted as a whole. Electrons in the $+$ side of Fig. 22-17 as well as in the $-$ side want to go to higher energy states, that is, higher positive velocities if v is positive and more negative velocities if v is negative. But they cannot because those states are already occupied. Only those electrons near the Fermi level have unoccupied states nearby in energy. With the application of an electric field, instead of thermal

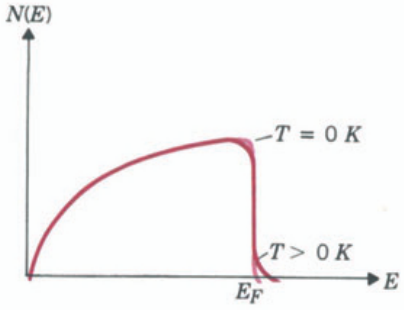


FIGURE 23-16
Comparison of the energy distribution of the electrons in a solid at $T = 0$ K and at $T > 0$ K for the case where $k_BT << E_F$.

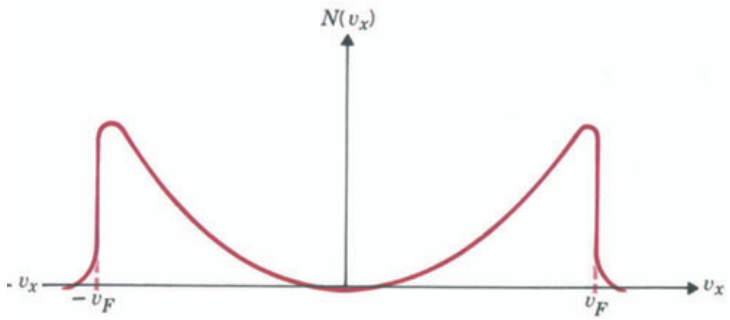
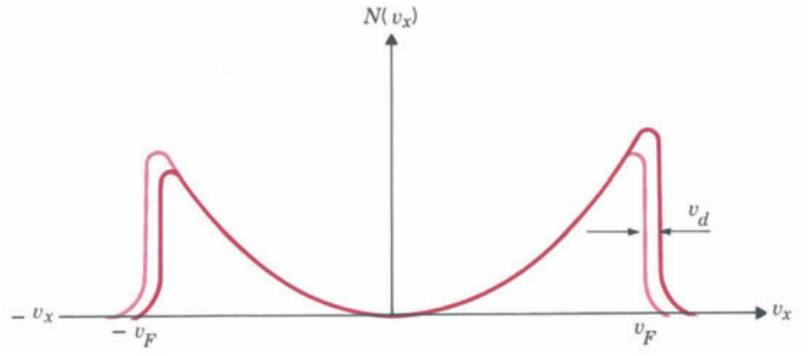


FIGURE 23-17
x-component of the velocity distribution of the electrons in a solid in the absence of an electric field.



When an external electric field is applied to the electrons in the solid, the velocity distribution of Fig. 23-17 is shifted by an amount equal to the drift velocity v_d in the direction opposite to that of the electric field.

energy, the energy from the field offered to the electrons is coherent and unidirectional: *All move toward more positive velocities together*. Thus, there is always a vacant state ready to receive an electron that is changing its state under the action of the electric field. The vacant state is created by the simultaneous change of the velocity state of another electron.

Although all the electrons participate in an electric conduction process, it is the electrons near E_F that determine the scattering time τ for all of them. The reason is the following. When the electrons are placed in the electric field, they increase their energy momentarily, then lose it on being scattered by the ions. However, this loss of energy is random: Not all the electrons lose energy at the same time, because l is a mean distance, and the deflection is not in the same direction. Because of this, not all the electrons can lose their energy on being scattered; only those that have *empty lower* states available to them, that is, only those electrons near E_F. Those deep down do not have empty lower states available to be scattered into. Do these electrons increase their energy indefinitely, as they do not seem to be able to lose it? The answer is no: The electrons ahead (higher in energy) prevent them from going to higher energy states. Therefore, they can only drift with the same velocity as the electrons near E_F. Because the drift velocity is proportional to τ (see Eq. 23.12) the average scattering time is therefore

$$\tau = \frac{l}{v_F} \tag{23.34}$$

where l is the mean free path and v_F is the velocity of the electrons at the Fermi level. If we use the expression for the electrical conductivity, Eq. 23.14,

$$\sigma = \frac{q^2 N \tau}{m} \tag{23.14}$$

and substitute for τ from Eq. 23.34, we obtain

$$\sigma = \frac{q^2 N l}{m v_F} \tag{23.35}$$

The QMFE model now has a serious problem. As we have seen, E_F and therefore v_F are essentially temperature independent. It would seem that σ is now also temperature independent unless l is a function of T.

In the CFE model l was basically determined by the position of the ions: In fact, we assumed that l was the average interionic separation. A hint that this assumption is not correct can be found immediately if we calculate l, which is done in Example 23-4.

Example 23-4

The electrical conductivity of Cu at room temperature is $5.9 \times 10^7\ \Omega^{-1}\ m^{-1}$. As we have shown earlier, the Fermi energy for copper is 6.95 eV (Example 23-2) and the carrier density $N = 8.4 \times 10^{28}$ electrons/m^3 (Example 23-1). Calculate the mean free path of the electrons.

Solution Write Eq. 23.34 as

$$l = v_F \tau \tag{23.36}$$

The Fermi velocity can be obtained from the Fermi energy because $E_F = \frac{1}{2} m v_F^2$.

$$v_F = \left(\frac{2E_F}{m}\right)^{1/2} = \left[\frac{2 \times 6.95\ \text{eV} \times 1.6 \times 10^{-19}\ \text{J/eV}}{9.1 \times 10^{-31}\ \text{kg}}\right]^{1/2}$$

$$= 1.56 \times 10^6\ \text{m/sec}$$

The scattering time τ is obtained from the electrical conductivity. From Eq. 23.14

$$\tau = \frac{\sigma m}{q^2 N}$$

$$= \frac{(5.9 \times 10^7\ \Omega^{-1}\ m^{-1})\ (9.1 \times 10^{-31}\ \text{kg})}{(1.6 \times 10^{-19}\ \text{C})^2\ (8.4 \times 10^{28}\ \text{electrons/m}^3)}$$

$$= 2.50 \times 10^{-14}\ \text{sec}$$

Substituting for τ and for v_F into Eq. 23.36 we obtain

$$l = 1.56 \times 10^6\ \text{m/sec} \times 2.50 \times 10^{-14}\ \text{sec} = 3.90 \times 10^{-8}\ \text{m}$$

$$= 390\ \text{Å}$$

This is roughly 100 times greater than the interatomic spacing.

The resolution of the problems faced by both the QMFE and CFE model, large l and the incorrect temperature dependence of the conductivity, is found by taking the wave nature of the electrons into account. The assumption that l is the average interionic separation is fine if we consider a model of small marbles (electrons) colliding with large stationary objects (ions). We know from the de Broglie wave relation that the propagation of electrons and other elementary particles cannot be properly described by a particle model but rather is governed by a wave. We saw in Chapter 12 that a wave in a crystal lattice undergoes Bragg scattering, that is, it will be reflected (constructive interference) if the condition (see Fig. 23-19)

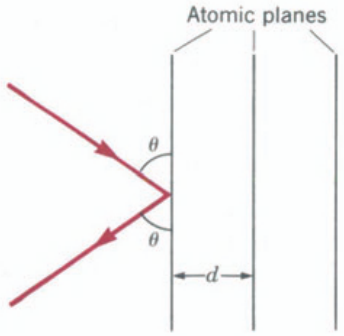


FIGURE 23-19
Bragg reflection of electron waves by the atomic planes of the solid.

$$2d \sin \theta = n\lambda \tag{12.11}$$

is satisfied. If the wavelength is long compared with the crystal spacing, Bragg scattering cannot occur. The conduction electrons must be considered as waves that may undergo Bragg scattering by the lattice. In copper $E_F \approx 7$ eV. This is the energy of the most energetic electrons. Because the de Broglie wavelength is $\lambda = h/p$, these electrons will have the smallest wavelength, which is

$$\lambda_F = \frac{h}{p_F}$$

but

$$E_F = \frac{p_F^2}{2m}$$

therefore

$$E_F = \frac{h^2}{2m\lambda_F^2}$$

and

$$\lambda_F = \frac{h}{(2mE_F)^{1/2}}$$

$$= \frac{6.63 \times 10^{-34} \text{ J-sec}}{(2 \times 9.1 \times 10^{-31} \text{ kg} \times 7 \text{ eV} \times 1.6 \times 10^{-19} \text{ J/eV})^{1/2}}$$

$$\lambda_F = 4.65 \times 10^{-10} \text{ m} = 4.65 \text{ Å}$$

The spacing between planes of copper atoms is $d = 2.09$ Å; therefore $\lambda_F > 2d$ and from Eq. 12.11, $\sin \theta > 1$. The Bragg condition for constructive reflection cannot be satisfied for any angle of incidence. This means that the electron waves in copper will not be reflected by the atomic planes but can traverse

the crystal lattice without being scattered. As a consequence, the electrons would have a mean free path equal to the dimensions of the crystalline sample.

The scattering of electron waves arises because of the deviations in the crystal from perfect periodicity: Such deviations include crystal lattice defects, impurities and, more important, the vibrations of the ions in the lattice. Except at very low temperatures, the main mechanism for the scattering in a pure crystal is the lattice vibrations. In the absence of such vibrations, the ions constitute a perfect lattice and, as we have seen, the mean free path of the electrons would be extremely large. We can represent this fact by visualizing the lattice ions as mathematical point particles with zero cross-sectional area, through which the electrons can move freely without being scattered. When the ions begin to vibrate, the lattice ceases to be ideal and now they offer an effective cross section for scattering. Clearly, the larger the cross-sectional area, the greater the chances of scattering and the shorter the mean free path. The mean path length of a moving object is inversely proportional to the cross-section of the target or

$$l \propto \frac{1}{\pi r^2}$$

where r, the radius of a circular target, is $r^2 = x_0^2 = y_0^2$ (see Fig. 23-20).

We know from classical physics (Section 10.6) that the total energy of a vibrating object is $E \propto (\text{amplitude})^2$, that is, $E \propto r^2$. But the energy of vibration comes as a result of the increase in temperature and is in fact proportional to the temperature, $E \propto k_B T$. Therefore,

$$r^2 \propto T$$

and therefore

$$l \propto T^{-1}$$

From Eq. 23.35, we see that

$$\sigma = \frac{q^2 N l}{m v_F} \propto T^{-1}$$

which is the observed temperature dependence of σ.

Although we have obtained this result by using a judicious mixture of simple wave and kinetic arguments, a more rigorous (and much more complicated) treatment based solely on the behavior of electron waves and on the theory of lattice vibrations leads to the same conclusion.

The QMFE model has removed most of the problems that we encountered in the CFE model, but neither has answered one of our initial questions: Why do the values of σ vary over such a wide range in different types of materials?

This leads us into the topic of our next chapter: The Band Theory of Conduction.

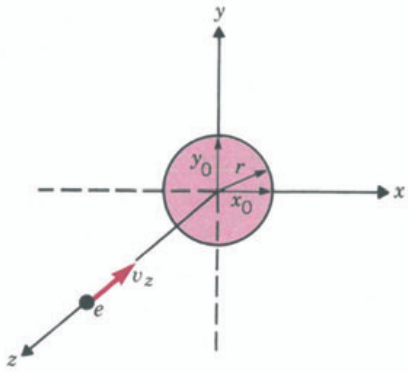


FIGURE 23-20
Scattering cross-section presented by a vibrating lattice ion to an incoming electron.

SUPPLEMENT 23-1

The Wiedemann-Franz Law

The CFE model was able to predict Ohm's law. It was also able to explain why good electrical conductors are also good thermal conductors (conductors of heat). Specifically, an empirical relation, known as the Wiedemann-Franz law, had been established in 1853 that relates the *electrical* and the *thermal conductivities*

$$\frac{\eta}{\sigma} = \mathcal{L}T \tag{23.37}$$

$\frac{\eta}{\sigma} = \mathcal{L}T$

where η is the thermal conductivity (to be defined following), σ the electrical conductivity, T the absolute temperature, and $\mathcal{L}$ is called the *Lorentz number* (a kind of universal constant, roughly the same for all materials), which has an average value of 2.3×10^{-8} W-Ω/K^2.

Let us examine the conduction of heat by electrons in a metal. Consider a thin slab, as in Fig. 23-21, of thickness Δx, cross-sectional area A, and let a temperature difference $\Delta T = T_H - T_C$ exist across Δx. Let T_H and T_C represent the temperatures of the hot and cold sides, respectively.

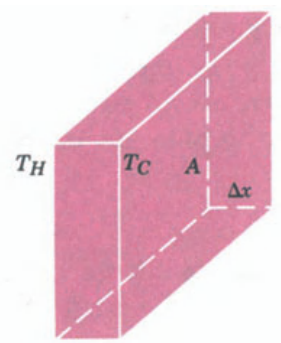


FIGURE 23-21

Experimentally one observes that heat energy is transferred from the hot to the cold side of the slab. Moreover, the heat current i_{heat} (heat per unit time) is found to be proportional to the temperature difference ΔT between the two sides and to the cross-sectional area A and inversely proportional to the thickness Δx of the slab. The proportionality can be converted to an equality by introducing a constant η, which is called the thermal conductivity of the material. That is,

$$i_{\text{heat}} = \eta A \frac{\Delta T}{\Delta x}$$

As in the case of electrical conduction, we define a heat current density

$$J_{\text{heat}} = \frac{i_{\text{heat}}}{A}$$

which is the heat transferred per unit time per unit area. The thermal conduction equation becomes

$$J_{\text{heat}} = \eta \frac{\Delta T}{\Delta x} \tag{23.38}$$

We can derive an expression for η based on the assumption that the agents that transfer the heat from one side to the other are the same that transfer the electric current when a potential difference exists across the sample, namely, the conduction electrons.

Consider, as shown in Fig. 23-22, a region filled with an electron gas with N electrons per unit volume. Let v_{RMS} be the average thermal speed and let l be the mean free path between collisions. Let the thermal energy at the central plane of this region be E and the gradient of the thermal energy be $-dE/dx$ from T_H to T_C. The thermal energy of the electrons at a plane l away

from the central plane (on the hot side) will be $E + (dE/dx)\, l$ and the energy of those at a plane l away (on the cold side) will be $E - (dE/dx)\, l$. The reason for choosing planes l away from the central plane is so that the electrons can reach and cross the central plane without colliding with the ions; therefore, as they cross the central plane they will have the same energy that they had when they were a distance l away from it.

Let us evaluate how many electrons will cross the central plane in 1 sec. For the moment, let us assume that all the electrons to the left of the plane are moving toward it with a velocity v_{RMS}. If the cross-sectional area of the plane is 1 m^2 and τ is the time to travel a distance l with an average speed of v_{RMS} in the direction of the plane, then in time τ all the electrons in the volume (1 m^2) ($v_{\text{RMS}}\, \tau$) will cross the plane. In a time of 1 sec all the electrons in the volume (1 m^2) v_{RMS} (1 sec) will cross the plane, and if the electron density is N, then in 1 sec the number of electrons crossing the plane will be, by this model, Nv_{RMS}. But this model assumes that all the electrons are moving perpendicularly toward the plane, whereas on the average only one sixth of them have that motion. This is because the number of vector directions is $\pm x$, $\pm y$, and $\pm z$ for a total of 6. With this modification we may write that the electrons from the hot side will carry a thermal current density toward the plane from the left of

$$J^+ = \frac{1}{6} N v_{\text{RMS}} \left(E + \frac{dE}{dx} l \right)$$

The electrons from the cold side will carry a thermal current density to the plane from the right of

$$J^- = \frac{1}{6} N v_{\text{RMS}} \left(E - \frac{dE}{dx} l \right)$$

The net thermal current density will be the sum of the positive and negative currents

$$J_{\text{heat}} = J^+ - J^-$$

$$J_{\text{heat}} = \frac{N v_{\text{RMS}}\, l}{3} \frac{dE}{dx} \tag{23.39}$$

The reason why there is an energy gradient dE/dx is that there is a temperature gradient dT/dx. We can relate these two by the standard mathematical chain rule,

$$\frac{dE}{dx} = \frac{dE}{dT} \frac{dT}{dx}$$

In Section 9.8 we introduced the molar specific heat at constant volume C_v, which was defined as the heat energy needed to raise 1 mole of a substance by 1 K. The heat energy needed to raise the temperature of 1 mole of electrons by dT is $C_v\, dT$, and the increase in the energy dE per electron will be

$$dE = \frac{C_v}{N_A} dT$$

where N_A is Avogadro's number.

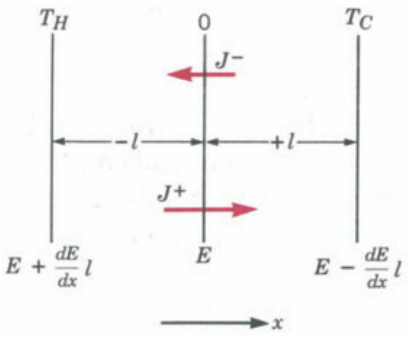


FIGURE 23-22
Three hypothetical planes having different temperatures. Since the electrons at the plane labeled T_H have a higher thermal energy than those in the plane labeled T_C, there will be a net thermal current $J^+ - J^-$ (thermal energy per unit time per unit area) across the central plane labeled 0.

Substituting C_v/N_A for dE/dT in the chain rule, and then substituting the result in Eq. 23.39, we obtain the relation

$$J_{\text{heat}} = \frac{N v_{\text{RMS}}\, l}{3} \frac{C_v}{N_A} \frac{dT}{dx}$$

Comparing this expression with the empirical one, Eq. 23.38, we see that

$$\eta = \frac{N v_{\text{RMS}}\, l\, C_v}{3N_A} = \frac{N v_{\text{RMS}}^2\, \tau\, C_v}{3N_A} \tag{23.40}$$

In the last step we made use of the CFE model result, Eq. 23.15, $l = v_{\text{RMS}}\, \tau$. We can find the ratio of the thermal and electrical conductivities by dividing the expression that we have found for η, Eq. 23.40, by the expression for σ that we found earlier, Eq. 23.14

$$\frac{\eta}{\sigma} = \frac{\dfrac{N v_{\text{RMS}}^2\, \tau\, C_v}{3N_A}}{\dfrac{N q^2\, \tau}{m}}$$

$$\frac{\eta}{\sigma} = \frac{m v_{\text{RMS}}^2\, C_v}{3q^2 N_A} \tag{23.41}$$

The molar specific heat at constant volume is given by Eq. 9.32:

$$C_v = \frac{3}{2} R \tag{9.32}$$

Substituting Eq. 9.32 for C_v into Eq. 23.41, we obtain

$$\frac{\eta}{\sigma} = \frac{m v_{RMS}^2\, R}{2q^2 N_A} = \frac{m v_{\text{RMS}}^2\, k_B}{2q^2} \tag{23.42}$$

In the last step, we made use of the fact that $R/N_A = k_B$. Because $1/2\, m v_{\text{RMS}}^2 = 3/2\, k_B T$, (Eq. 9.21), Eq. 23.42 can be written as

$$\frac{\eta}{\sigma} = \frac{3}{2}\left(\frac{k_B}{q}\right)^2 T \tag{23.43}$$

which is the Wiedemann-Franz law, Eq. 23.37. We obtain the theoretical value for the Lorentz constant by simple substitution

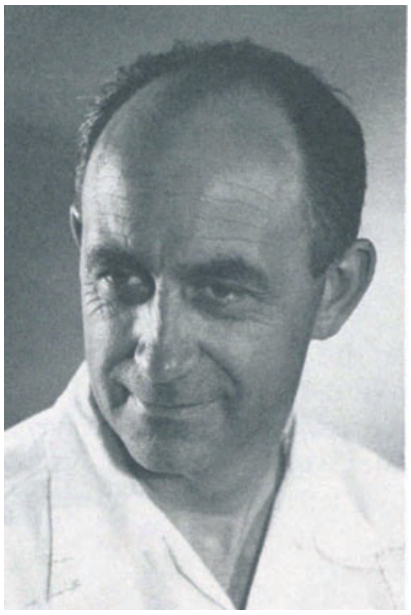
$$\mathcal{L} = \frac{3}{2}\left(\frac{k_B}{q}\right)^2 = \frac{3 \times (1.38 \times 10^{-23}\ \text{J/K})^2}{2 \times (1.6 \times 10^{-19}\ \text{C})^2}$$

$$= 1.12 \times 10^{-8}\ \text{W-}\Omega/\text{K}^2$$

whereas the experimental value is $\mathcal{L} = 2.3 \times 10^{-8}\ \text{W-}\Omega/\text{K}^2$. Although the derived value of $\mathcal{L}$ is roughly one half the experimental value, the theoretical value is nevertheless of the same order of magnitude. This is quite good considering the many simplifying assumptions and omissions of the CFE model.

SUPPLEMENT 23-2

Fermi-Dirac Statistics



Enrico Fermi (1901–1954).

In Chapter 9 we saw that the Maxwell-Boltzmann distribution gives the relative probability of occupancy of two states i and j with energies E_i and E_j by Eq. 9.40 as

$$\frac{N_i}{N_j} = \exp\left[-(E_i - E_j)/k_B T\right] \tag{9.40}$$

Eq. (9.40) contains the implicit assumption that the presence of one or several particles in an energy state does not prevent other particles from occupying that state. In a crystal, because of the exclusion principle, only two electrons may occupy each state, and we need a different statistical probability function. Although these new statistics may be derived more rigorously by methods of statistical mechanics, we will give here an elementary demonstration that the statistical function we choose does satisfy the Pauli principle. We will ignore the fact that two electrons may occupy each state and assume that only one can. It can be done for two electrons, but the factor of two thus introduced will cancel at the end anyway.

If two particles with energies E_1 and E_2 collide, they may then have energies E_3 and E_4, respectively, in a Maxwell-Boltzmann distribution. But in a distribution subject to the Pauli principle such a collision may take place only if energy states E_3 and E_4 are unoccupied. If $F(E_i)$ is the probability that energy state E_i is occupied, then $[1 - F(E_i)]$ is the probability that energy state E_i is vacant. In this collision the joint probability that both states E_1 and E_2 are occupied is the product of the two individual probabilities, $F(E_1)\,F(E_2)$, and the probability that the final states are vacant is the product of their individual probabilities $[1 - F(E_3)][1 - F(E_4)]$. The joint probability that such a collision may occur is the product of all four individual probabilities, and by introducing a constant c that contains terms such as density of particles, cross-section, and velocity, we may write an expression for f, the number of this type of collision per second as

Paul Adrien Dirac (1902–1984).

$$f = cF(E_1)F(E_2)\,[1 - F(E_3)][1 - F(E_4)] \tag{23.44}$$

The rate of the reverse type of collision f' may be written as

$$f' = cF(E_3)F(E_4)\,[1 - F(E_1)][1 - F(E_2)] \tag{23.45}$$

Under equilibrium conditions the forward and the reverse collision must be occurring at the same rate. We must therefore equate Eqs. 23.44 and 23.45. If we do so and divide by $cF(E_1)F(E_2)F(E_3)F(E_4)$ we obtain

$$\left[\frac{1 - F(E_3)}{F(E_3)}\right]\left[\frac{1 - F(E_4)}{F(E_4)}\right] = \left[\frac{1 - F(E_1)}{F(E_1)}\right]\left[\frac{1 - F(E_2)}{F(E_2)}\right] \tag{23.46}$$

Because energy is conserved in the collision process, if we write δ as the gain in energy of the particle with initial energy E_1, then $-\delta$ must be the loss in energy of the particle with initial energy E_2. Therefore, we may write $E_3 = E_1 + \delta$ and $E_4 = E_2 - \delta$. Substituting these expressions in Eq. 23.46 we obtain

$$\left[\frac{1 - F(E_1 + \delta)}{F(E_1 + \delta)}\right]\left[\frac{1 - F(E_2 - \delta)}{F(E_2 - \delta)}\right] = \left[\frac{1 - F(E_1)}{F(E_1)}\right]\left[\frac{1 - F(E_2)}{F(E_2)}\right] \tag{23.47}$$

The relation of Eq. 23.47 can be satisfied with the substitution of

$$\frac{1 - F(E_i)}{F(E_i)} = C\, e^{\beta E_i}$$

or

$$F(E_i) = \frac{1}{1 + C\, e^{\beta E_i}} \tag{23.48}$$

where C and β are arbitrary positive constants. We may show that Eq. 23.48 satisfies the relation of Eq. 23.47 by direct substitution:

$$C\, e^{\beta(E_1 + \delta)}\, C\, e^{\beta(E_2 - \delta)} = C\, e^{\beta E_1}\, C\, e^{\beta E_2} \tag{23.49}$$

Because the δ terms cancel, Eq. 23.49 is an identity, and Eq. 23.48 is a solution of Eq. 23.47.

Thus, the probability of the ith state being occupied is

$$F(E_i) = \frac{1}{C\, e^{\beta E_i} + 1} \quad \text{where } F(E_i) \leq 1 \tag{23.50}$$

For large values of E_i, $C\, e^{\beta E_i} >> 1$, so we may ignore the 1 in the denominator and write

$$F(E_i) = \frac{1}{C}\, e^{-\beta E_i} \tag{23.51}$$

Equation 23.51 is the Maxwell-Boltzmann distribution, Eq. 9.39, with $\beta = 1/k_B T$. Because C is an arbitrary mathematical constant, let us write it as $C = e^{-E_F/k_B T}$ and substitute this value and $\beta = 1/k_B T$ in Eq. 23.50. We now have

$$F(E_i) = \frac{1}{\exp\dfrac{E_i - E_F}{k_B T} + 1} \tag{23.52}$$

Equation 23.52 is usually written without the i subscript and is called the *Fermi-Dirac distribution* and the constant E_F is called the *Fermi energy*. It applies to condensed systems in which the Pauli exclusion principle is obeyed.

PROBLEMS

23.1 Consider a copper wire of cross-sectional area $A = 3$ mm^2. (a) What would the current through the wire be for the drift velocity to be comparable to the thermal velocity at room temperature, that is, $v_d \sim 10^5$ m/sec? (b) Would Ohm's law be obeyed in this case? Explain.

23.2 Gold is monovalent. The atomic weight and the density of gold are 197 g/mole and 19.3 g/cm^3, respectively. Calculate the number of free electrons per unit volume.

(*Answer*: 5.87×10^{28} m^{-3}.)

23.3 The electrical conductivity of copper at room temperature is $5.9 \times 10^7\ \Omega^{-1}\text{m}^{-1}$. (a) What is the thermal conductivity of copper at room temperature? (b) What is the thermal conductivity at 1000 K?

(*Answer*: (a) 407 W/m-K, (b) 407 W/m-K.)

23.4 The electrical conductivity of copper at 300 K is $5.9 \times 10^7\ \Omega^{-1}\text{m}^{-1}$. (a) What should be the conductivity at 77 K according to the CFE model? (b) How does it compare with the experimentally measured value of $4 \times 10^8\ \Omega^{-1}\text{m}^{-1}$?

(*Answer*: (a) $1.2 \times 10^8\ \Omega^{-1}\text{m}^{-1}$.)

23.5 The electronic energy levels in the three-dimensional infinite potential well are given by Eq. 23.19, $E(n_1, n_2, n_3) = E_0\,(n_1^2 + n_2^2 + n_3^2)$. Find the fractional difference in the energy between the pairs of states (a) $E(1,1,1)$ and $E(1,1,2)$, (b) $E(10,10,10)$ and $E(9,10,11)$ (c) E (100,100,100) and E (99,100,101). (d) What conclusion can you draw from these results?

23.6 Magnesium is a bivalent metal with an atomic weight of 24.32 g/mole and a density of 1.74 g/cm^3. (a) What is the free electron density? (b) What is the Fermi energy? (c) Calculate the Fermi velocity. (d) What is the de Broglie wavelength of the electrons at the Fermi level?

(*Answer*: (a) 8.4×10^{28} m^{-3}, (b) 7.07 eV, (c) 1.58×10^6 m/sec, (d) 4.62 Å.)

23.7 The Fermi energy of cesium is 1.55 eV. (a) Determine the number of electrons in 1 cm^3 by free electron theory. (b) How does this compare with the number of cesium atoms in 1 cm^3? (The density of cesium is 1.87 g/cm^3 and the atomic weight is 133 g/mole).

23.8 In a Hall effect experiment like the one discussed in Section 16-7, a rectangular gold slab 1 mm thick carries a current of 20 A. The magnetic field $B = 1.2$ T. It is found that the Hall voltage $V_H = 2.6\ \mu$V. (a) What is the free electron concentration? (b) How does the answer to (a) compare with that obtained by the method used in Problem 23.2? Gold is monovalent with molecular weight 197 g/mole and density 19.3 g/cm^3.

(*Answer*: (a) 5.77×10^{28} m^{-3}, (b) 5.87×10^{28} m^{-3}.)

23.9 When the sun stops producing energy by fusion it will collapse and become a white dwarf star with radius about that of the earth. The mass of the sun is 2×10^{30} kg and the radius of the earth is 6.37×10^3 km. Assume that there is one electron for every two *nucleons*. (A nucleon is the general name for the constituents of the nucleus, protons and neutrons). What will be the Fermi energy of the electrons in that white dwarf star?

(*Answer*: 2.43×10^5 eV.)

23.10 A neutron star is a giant nucleus composed of neutrons. Neutrons obey the Pauli exclusion principle and therefore have a Fermi-Dirac distribution of energies. Consider a neutron star of mass 4×10^{30} kg and radius 10 km. What is the average energy of the neutrons in that star?

23.11 (a) Estimate how much energy would be released by the conduction electrons in a 1 m^3 piece of copper if we could suddenly turn off the Pauli exclusion principle. (b) How long could a 100-W lamp be lit with this amount of energy? (The free electron density in copper is 8.4×10^{28} electrons/m^3.)

(*Answer*: (a) 3.5×10^{29} eV, (b) 5.6×10^8 sec or 17.6 years.)

23.12 Let us assume that kinetic theory applies to a quantum-mechanical electron gas. (a) What would

be the pressure of the electron gas at $T = 0$ K? (*Hints:* See Section 9.5, Eq. 9.18 and recall that the average energy of an electron obeying Fermi-Dirac statistics is 3/5 E_F.) (b) Calculate the pressure for copper ($E_F = 6.95$ eV, $N = 8.4 \times 10^{28}$ m^{-3}).

(*Answer*: (a) 2/5 NE_F, (b) 3.7×10^{10} N/m^2.)

23.13 Estimate the fraction of free electrons in copper that are in excited states at room temperature. The Fermi energy of copper is 6.95 eV.

23.14 If the Fermi energy for a given metal is 1 eV and the metal melts at 2000°C, approximately what fraction of the electrons are in excited states at the melting point?

23.15 The Fermi energy of silver is 5.1 eV. If the temperature is 300 K, what is the probability that a state be occupied for the following energies: (a) $E = 5$ eV, (b) $E = 5.2$ eV, (c) $E = 4$ eV, and (d) $E = 6$ eV? (e) At what temperature will the probability of occupation for the state of energy $E = 5.2$ eV be 0.1?

(*Answer*: (a) 0.979, (b) 0.021, (c) 1, (d) 9×10^{-16}, (e) 527 K.)

23.16 In the Fermi distribution let $\Delta E = E - E_F$. Calculate $F(E)$ for $\Delta E = 2\ k_BT$, $4\ k_BT$, $10\ k_BT$.

23.17 Show that the probability that a state with energy $E = E_F + \Delta E$ be occupied is equal to the probability that a state with energy $E = E_F - \Delta E$ be empty.

23.18 Let us define the Fermi temperature $T_f = E_F/k_B$, where k_B is the Boltzmann constant. What is T_f for silver? The Fermi energy of silver is 5.1 eV.

23.19 The electronic molar specific heat of copper is $C_v = 1.65 \times 10^{-4}\ T$ calories/mole-K. Find E_F from that information.

(*Answer*: 5.1 eV.)

23.20 The electrical conductivity of aluminum, a trivalent metal, at room temperature is $3.8 \times 10^7\ \Omega^{-1}\ m^{-1}$. The atomic weight and the density of Al are 27.0 g/mole and 2.7 g/cm^3, respectively. What is the mean free path for the electrons in Al at room temperature?

(*Answer*: 151 Å.)

23.21 A collection of N atoms have available to them two energy levels: the ground state E_1 and an excited state E_2. Assume that the atoms obey the Maxwell-Boltzman distribution and that $E_2 - E_1 = \Delta E >> k_BT$. (a) What fraction of the atoms are in the excited state at a temperature T? (b) What is the average energy of the atoms at that temperature? (c) What is the total energy of the atoms? (d) Find the molar specific heat of the system.

(*Answer*: (a) $\exp(-\Delta E/k_BT)$,
(b) $E_1 + \Delta E \exp(-\Delta E/k_BT)$,
(c) $N[E_1 + \Delta E \exp(-\Delta E/k_BT)]$,
(d) $R(\Delta E/k_BT)^2 \exp(-\Delta E/k_BT)$.)

23.22 Can photoelectrons be emitted thermally? If the work function of a metal surface is 1 eV, calculate the temperature the electron must have to escape in the dark. What fraction of electrons in a solid have enough energy to escape at room temperature assuming that the electrons have the Maxwell-Boltzmann distribution of energies?

CHAPTER 24

Band Theory of Solids

24.1 INTRODUCTION

The free electron models of metals that we have presented in the preceding chapter gives us a good deal of insight into several properties of metals. Yet there are many other important properties that these models do not explain. In particular, they do not tell us why, when chemical elements crystallize to become solids, some are good *conductors*, some are *insulators*, and yet others are *semiconductors* with electrical properties that vary greatly with temperature. These differences are not minor, but rather remarkable. The resistivity may vary from $\rho \sim 10^{-8}$ ohm-m for a good conductor to $\rho \sim 10^{22}$ ohm-m for a good insulator.

We can understand the differences between insulators and conductors by extending the free electron model to take into account the interaction of the electrons with the positive ion lattice. In the quantum mechanical free electron (QMFE) model, we assumed that the potential energy inside the solid was uniform. It would be more realistic to assume that it is a periodic (alternating uniformly) function of x, y, z. This is reasonable because of the periodic distribution of the lattice ions in a *crystalline* solid.

When the interaction between the electrons and the lattice ions is considered, we will find some unusual properties possessed by the electrons in the crystal.

1. In the last chapter we saw that the QMFE model gave rise to a series of discrete energy states, about 10^{-15} eV apart, which ranged from $E \approx 0$ to $E = E_F$, the Fermi level. This range of energy levels can be called a *band* of energies, and the energy levels are so close together that they are referred to as quasicontinuous. In the band of the QMFE model there was an infinite number of unoccupied energy states above the Fermi level that could be occupied by excited electrons. When we introduce the potential of the lattice ions, we will see that bands of this type have upper and lower limits of allowable energies. If a band is not filled with electrons, then the electrons may be excited into the empty states and contribute to electrical or thermal conduction. If, however, a band is filled, then there are no states to be occupied and the electrons cannot be excited. For conductors, we will see that the behavior of the electrons in the occupied band of highest energy is almost identical to that predicted by the QMFE model, but not so for insulators and semiconductors. The band theory model will solve the questions concerning the differences between conductors and insulators.
2. The electrons respond to an externally applied electric or magnetic field as if they were endowed with an *effective mass* m^*, which may be greater or smaller than that of the free electron, or even be negative. By this we mean that the electrons can be treated as free in responding to an external electric and magnetic field provided that we assign to them a mass different from the true mass.

3. There are situations in which instead of conduction by electrons it is convenient to attribute the conduction to charge carriers with a positive charge $+e$, called *holes*.

There are several methods (or models) to show the existence of bands and to find the shape of the band. Some work well for certain materials, some with others. Quantitative band calculations must take into account the particular crystal structure, atomic configuration, and type of bonding: These details belong to the realm of solid-state physics research. Our purpose here is simply to show the existence of bands and the general characteristics. This can be achieved with idealized models and by using qualitative arguments.

24.2 BLOCH'S THEOREM

Before we proceed to study the motion of an electron in a periodic potential[1], we should mention a general property of the wavefunctions in such a periodic potential.

For a free electron with E_p = constant, the space part of the wave function $\psi(x, t)$, called the eigenfunction $\chi(x)$, is written as (see Eq. 20.4)

$$\chi(x) = e^{\pm ikx}$$

If the spacing of the ions in the x direction in a solid is d, then the potential energy of an electron at a point x distance from the origin is equal to the potential energy at a point $x + d$ from the origin. This potential energy is equal in turn to that at point $x + 2d$ from the origin, and so on. Therefore, we can generalize and take any point x in the lattice and state that the potential energy at that point is equal to the potential energy at point $x + d$ or, stated mathematically, $E_p(x) = E_p(x + d)$. This is known as a *periodic potential*. There is a theorem by Bloch which states that for a particle moving in a periodic potential, the eigenfunctions $\chi(x)$ are of the form

$$\chi(x) = u_k(x)\, e^{\pm ikx} \tag{24.1}$$

where

$$u_k(x) = u_k(x + d)$$

These eigenfunctions are plane waves modulated by a function $u_k(x)$, where $u_k(x)$ has the same periodicity as the potential energy. Because the potential energy $E_p(x) = E_p(x + d)$, one expects that the probability of finding a particle at a given x is the same as that of finding it at $x + d$. This is guaranteed by

[1]As indicated earlier, in Chapter 20, in quantum mechanics the *potential energy* is often called the *potential*.

the periodicity of u_k and can be seen in the following expression for the probability density.

$$\chi^*(x)\,\chi(x) = u_k^*(x)\,e^{-ikx}\,u_k(x)\,e^{ikx}$$
$$= u_k^*(x)\,u_k(x)$$

Therefore, when

$$u_k(x) = u_k(x + d)$$

then

$$\chi^*(x)\,\chi(x) = \chi^*(x + d)\,\chi(x + d)$$

The specific form of the function $u_k(x)$ will depend on the form of the function $E_p(x)$. We will now consider an idealized, one-dimensional periodic potential.

24.3 THE KRONIG-PENNEY MODEL

Let us try to understand what the potential energy of an electron in a crystalline solid may look like. Consider a positively charged ion q and an electron e at a distance x from q as shown in Fig. 24-1. The electric potential energy from the coulomb attraction experienced by the electron is (Eq. 14.9).

$$E_p(x) = -\frac{1}{4\pi\epsilon_0}\frac{q|e|}{x}$$

The variation of E_p with x is illustrated in Fig. 24-1. Suppose we now place another charge q at a point d away from the first. The potential energy E_p at any point on the x axis will be equal to the algebraic sum of the potential energies due to each individual charge, as illustrated in Fig. 24-2. The dashed lines represent the potential energy due to the individual q's and the solid lines represent the sum of the dashed lines. If we now place a long array of q's separated by a distance d from each other to form a periodic array, the

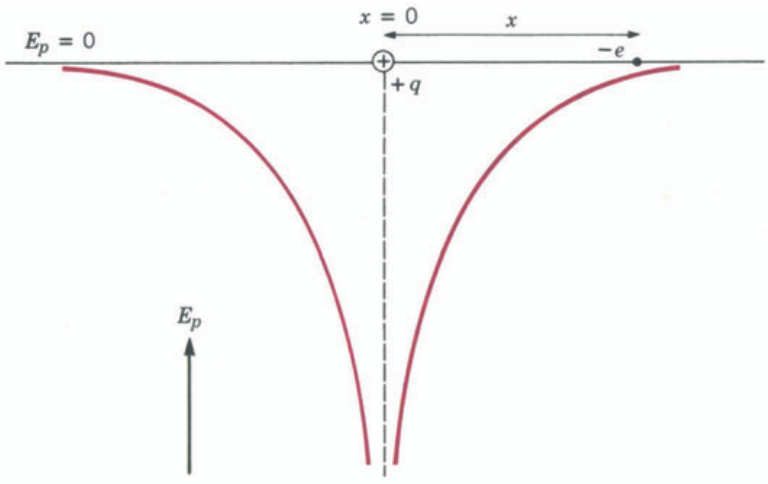


FIGURE 24-1
Potential energy E_p associated with the coulomb interaction of an electron with a positive ion $+q$ as a function of the separation x between the electron and the ion.

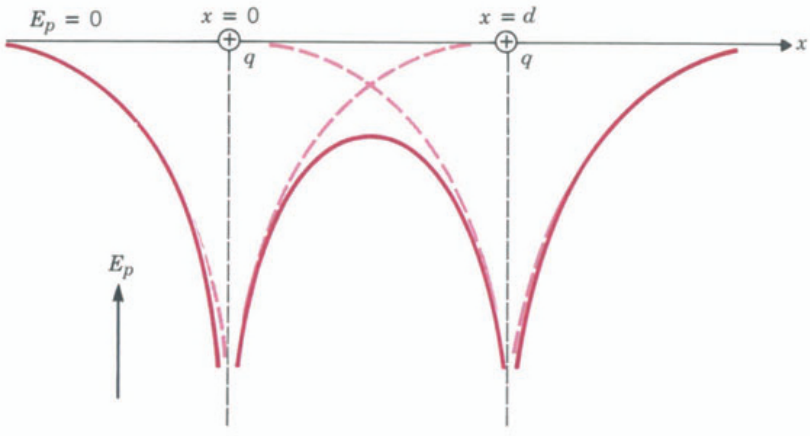


FIGURE 24-2
Potential energy E_p when an electron interacts with two ions (solid lines). The dashed lines correspond to the potential energy associated with the interaction of the electron with the individual ions. The solid lines represent the sum of the dashed lines.

potential energy E_p looks like that shown in Fig. 24-3. The main features of the potential energy in Fig. 24-3 are: (1) it is periodic with a period d, (2) the maxima are halfway between the ions, and (3) the potential energy tends to $-\infty$ as the position of the ions is approached because the electron is bound more strongly to the ion as it comes closer and, because it takes more energy to pull it away, it can be said to lie in a deeper potential energy well the closer it is to the ion.

If one tries to solve the Schrödinger equation for such a potential, one runs into mathematical difficulties that are best solved by a computer. However, we can replace the potential energy of Fig. 24-3 with one that is mathematically simpler to handle while retaining the essential features of the actual one. We replace the potential energy of Fig. 24-3 by one consisting of periodically spaced rectangular wells as shown in Fig. 24-4. The potential energy is a series of rectangular wells of width c, spaced a distance b apart so that the periodicity $d = b + c$. The energy of the wells is $-E_{p0}$. However, it is convenient to shift the zero of potential energy so that the bottoms of the wells are at potential energy $E_p = 0$ and the tops are at $E_p = E_{p0}$. The potential energy of Fig. 24-4 has the same periodicity as the lattice; the potential energy

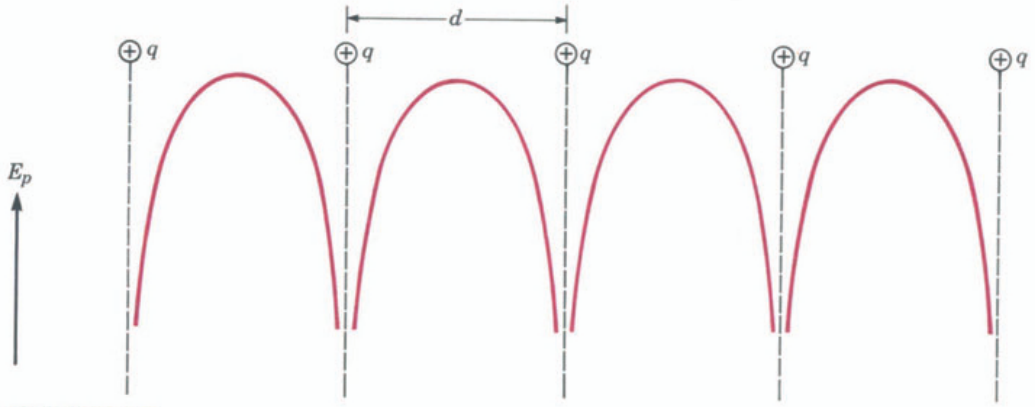


FIGURE 24-3
Potential energy of an electron in a one-dimensional array of periodically spaced ions. The periodicity of the ions is d.

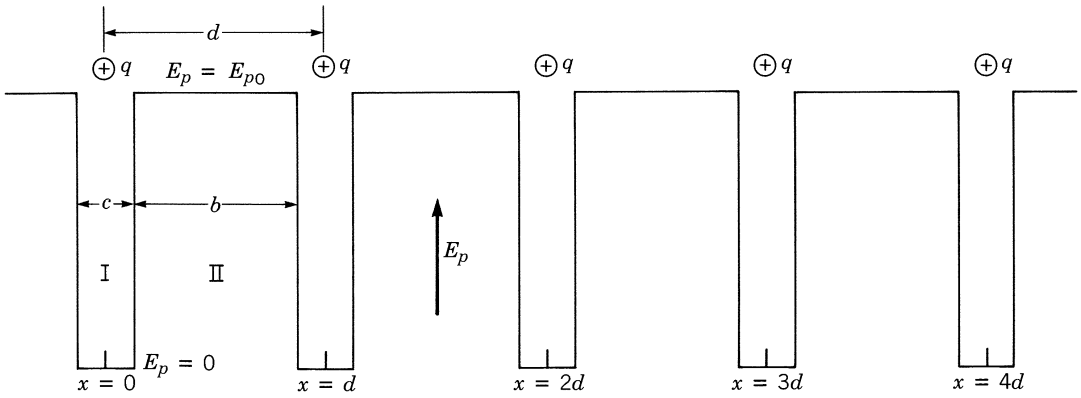


FIGURE 24-4
Periodic array of rectangular potential energy wells of depth E_{po}, width c, separated by a distance b. The periodicity of the wells is $d = b + c$. Such an array of potential energy wells can be used to approximate the potential energy of Fig. 24-3.

is lower in the vicinity of the ions and highest between the ions. This potential energy model is known as the Kronig-Penney model.

To find the behavior of the electrons in such a periodic potential, we have to find the eigenfunction that we can associate with them by solving the Schrödinger equation. We will not show all the mathematical details, but instead we will outline the method in order to understand the origin of the final result.

Because E_p is either 0 or E_{p0}, we solve for χ separately in both regions I and II. We then impose the conditions of continuity for χ and $d\chi/dx$ discussed in Chapter 20, while meeting the periodicity requirements. We will consider the case where $E < E_{p0}$.

Region I

In region I $E_p = 0$ and the Schrödinger equation is written as (Eq. 20.18)

$$-\frac{\hbar^2}{2m}\frac{d^2\chi_I}{dx^2} = E\,\chi_I$$

where χ_I is the eigenfunction in region I. Rearranging terms, we get

$$\frac{d^2\chi_I}{dx^2} + \gamma^2\chi_I = 0 \quad \text{where } \gamma = \sqrt{\frac{2mE}{\hbar^2}} \tag{24.2}$$

From Bloch's theorem, $\chi_I = u_I(x)\,e^{ikx}$ (Eq. 24.1). If we substitute this χ in Eq. 24.2, we get a differential equation for u_I,

$$\frac{d^2u_I}{dx^2} + 2ik\frac{du_I}{dx} + (\gamma^2 - k^2)\,u_I = 0$$

The solution of this equation can be found by the methods that we used

before in Chapter 20 and is

$$u_{\mathrm{I}}(x) = A\, e^{i(\gamma - k)x} + B\, e^{-i(\gamma + k)x}$$

where A and B are arbitrary constants.

Region II

In region II $E_p = E_{p0}$ and the Schrödinger equation is written as

$$-\frac{\hbar^2}{2m}\frac{d^2\chi_{\mathrm{II}}}{dx^2} + E_{p0}\chi_{\mathrm{II}} = E\,\chi_{\mathrm{II}}$$

where χ_{II} is the eigenfunction in region II. Rearranging terms, we get

$$\frac{d^2\chi_{\mathrm{II}}}{dx^2} - \xi^2\chi_{\mathrm{II}} = 0 \quad \text{where } \xi = \sqrt{\frac{2m(E_{p0} - E)}{\hbar^2}}$$

If we substitute $\chi_{\mathrm{II}} = u_{\mathrm{II}}(x)\, e^{ikx}$, we will get a differential equation for u_{II} that can be solved by the same method yielding for u_{II}

$$u_{\mathrm{II}} = C\, e^{(\xi - ik)x} + D\, e^{-(\xi + ik)x}$$

where C and D are arbitrary constants.

The next step is to impose the requirements of continuity and periodicity, between the regions I and II. It is seen in Fig. 24-4 that regions I and II join at $x = c/2$, therefore, following the discussion in Section 20.2e, we recall that both the eigenfunctions and their first derivatives must be continuous across a boundary; mathematically these criteria are

$$\chi_{\mathrm{I}}\left(\frac{c}{2}\right) = \chi_{\mathrm{II}}\left(\frac{c}{2}\right)$$

$$\frac{d\chi_{\mathrm{I}}}{dx}\left(\frac{c}{2}\right) = \frac{d\chi_{\mathrm{II}}}{dx}\left(\frac{c}{2}\right)$$

In addition, the periodicity requirements must be satisfied. This can be done by choosing points separated by the period of the lattice d, such as $x = -c/2$ and $x = b + c/2$. When we substitute Bloch functions for the χ functions, that is, $\chi = u(x)\, e^{+ikx}$, the periodicity requirement on the function $u(x)$ yields

$$u_{\mathrm{I}}\left(-\frac{c}{2}\right) = u_{\mathrm{II}}\left(b + \frac{c}{2}\right)$$

$$\frac{du_{\mathrm{I}}}{dx}\left(-\frac{c}{2}\right) = \frac{du_{\mathrm{II}}}{dx}\left(b + \frac{c}{2}\right)$$

These four conditions on the eigenfunctions lead to four linear algebraic equations for the constant A, B, C, D. In solving these equations, it is found that

a solution exists only if[2]

$$P\frac{\sin \gamma d}{\gamma d} + \cos \gamma d = \cos kd \tag{24.3}$$

$$P\frac{\sin \gamma d}{\gamma d} + \cos \gamma d = \cos kd$$

where

$$P = \frac{mE_{p0}bd}{\hbar^2} \quad \text{and, from Eq. 24.2,} \quad \gamma = \sqrt{\frac{2mE}{\hbar^2}}$$

(Note: This P is a new term with no relation to momentum). If this condition is not satisfied, the boundary conditions on χ cannot be satisfied, and the corresponding χ's are not acceptable solutions. In arriving at Eq. 24.3 we used one of the forms of Bloch functions, namely, $\chi = u(x)e^{+ikx}$. The same result will be obtained if we use the other form, that is, $\chi(x) = u(x)\, e^{-ikx}$.

24.3a Allowed and Forbidden Energy Bands

The main result of the solution of the Schrödinger equation for the periodic potential of Fig. 24-4 is that the only acceptable χ's are those for which Eq. 24.3 holds.

Let us try to understand the significance of Eq. 24.3. Remember that

$$\gamma = \sqrt{\frac{2mE}{\hbar^2}}$$

Therefore γ is a measure of the total energy, E, whereas k is the wave vector $k = 2\pi/\lambda$. From de Broglie's hypothesis, Eq. 19.1, $\lambda = h/p$ and therefore $p = k\hbar$. Therefore k is a measure of the momentum of the particle. Equation 24.3 relates the total energy E of the electron to its momentum p. (The corresponding expression for the free particle is $E = p^2/2m = k^2\hbar^2/2m$.) Finding a direct analytical expression $E(k)$ is not possible because Eq. 24.3 is a transcendental equation that cannot be solved analytically. We can, however, solve it numerically: We choose a γ (an E), insert it into the equation, and solve for k. This is not difficult; in fact, it is a simple exercise in computer programming. When we do this we will find a rather interesting result. There will be ranges of γ for which k will be a real number. Those ranges of γ will be separated by other ranges for which k is imaginary. The momentum of a particle cannot be imaginary, and the conclusion is as follows: A particle in this periodic potential cannot have values of γ for which k is imaginary; therefore, the corresponding values for E for these γ's are not allowed.

We do not have to go through the tedious task of selecting a γ and substituting it into the Eq. 24.3 to show this. We can get the result by qual-

[2]In arriving at Eq. 24.3, an additional simplification is made, namely E_{p0} is assumed to approach infinity and b to tend to zero while their product $E_{p0}b$ remains constant.

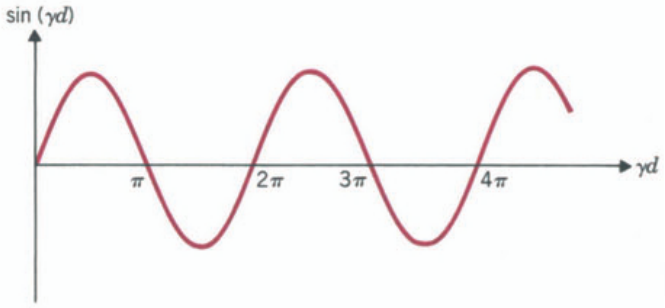


FIGURE 24-5

itatively plotting the left side of Eq. 24.3, calling it $f(\gamma d)$.

$$f(\gamma d) = P\frac{\sin \gamma d}{\gamma d} + \cos \gamma d$$

For example, let $P = 5/2\ \pi$. The sine function is periodic, as in Fig. 24-5, and $P/\gamma d$ behaves as in Fig. 24-6. When we multiply these two functions to get the first term of $f(\gamma d)$, an oscillating function similar to sin γd results, but the amplitude will decrease with increasing γd. Some values are shown in Table 24-1.

FIGURE 24-6

TABLE 24-1

γd	π	$\frac{3}{2}\pi$	2π	$\frac{5}{2}\pi$	3π	$\frac{7}{2}\pi$	4π
$P\frac{\sin \gamma d}{\gamma d}$	0	$-\frac{5}{3}$	0	1	0	$-\frac{5}{7}$	0

Between 0 and π we have to be careful, particularly close to $\gamma d = 0$, because when $\gamma d = 0$, $P \sin \gamma d/\gamma d = P\ 0/0$, which is undetermined. We can, however, use the L'Hopital rule on limits that

$$\lim_{x\to 0}\frac{\sin x}{x} = \lim_{x\to 0}\frac{\frac{d}{dx}\sin x}{\frac{d}{dx}x} = \lim_{x\to 0}\frac{\cos x}{1} = 1$$

Thus, in the limit as $\gamma d \to 0$, sin $\gamma d/\gamma d \to 1$. As γd begins to increase from 0, both sin γd and γd increase; however, the ratio decreases. If we express γd in radians, we can easily show with a calculator that sin $\gamma d/\gamma d$ is a decreasing function of γd, which becomes 0 when $\gamma d = \pi$. We can put all these facts together to obtain the plot of the first terms of $f(\gamma d)$; this is shown in Fig. 24-7. Note that the larger P, the greater the slopes will be, because the zero positions are fixed. To get the entire function $f(\gamma d)$, we must add to Fig. 24-7 the term cos γd, Fig. 24-8. Between 0 and π both functions decrease; therefore, $f(\gamma d)$ decreases and becomes -1 when $\gamma d = \pi$. After π, the first term continues decreasing, while the second begins to increase. Because cos γd changes

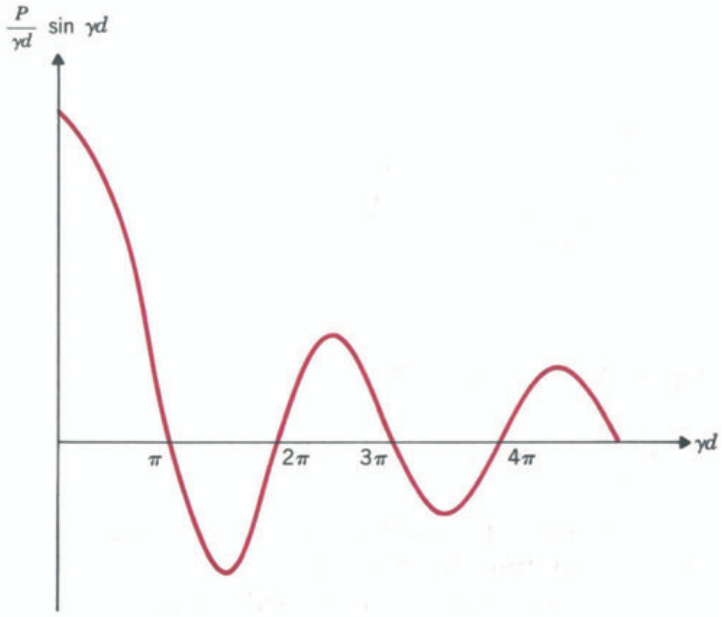


FIGURE 24-7
Plot of the first term of the left side of Eq. 24.3 as a function of γd. The plot, obtained by multiplying the sine function of Fig. 24-5 by the function of Fig. 24-6, is an oscillating function of γd with decreasing amplitude.

slowly near the maximum and minimum, $f(\gamma d)$ continues to decrease below -1. Somewhere between π and 3/2 π the trend reverses itself and $f(\gamma d)$ begins to increase, reaching the value of $+1$ at 2π. After 2π the first term of $f(\gamma d)$ continues increasing while the second decreases. Again, just as before, the first term increases at a faster rate than the rate of decrease of the second term, and as a result $f(\gamma d)$ continues (for a while) to increase past $+1$. Somewhere between 2π and 5/2 π the trend will reverse.

These arguments are reflected in a plot of $f(\gamma d)$ versus γd, Fig. 24-9. The most important fact to note is that there are ranges of γd (shaded regions) for which the values of $f(\gamma d)$ vary between $+1$ and -1. These ranges of γd are separated by others for which $f(\gamma d)$ is either greater than $+1$ or less than -1. The width of the shaded region, the ranges of γd for which $f(\gamma d)$ varies between $+1$ and -1, increases as γd increases. The condition that had to be satisfied for the solutions to the Schrödinger equation to be acceptable was, Eq. 24.3,

$$f(\gamma d) = \cos kd$$

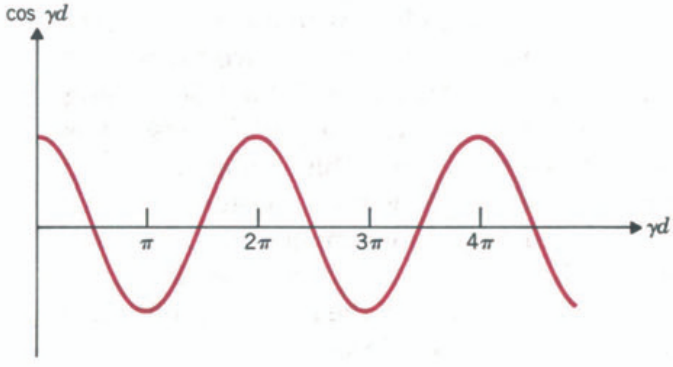


FIGURE 24-8

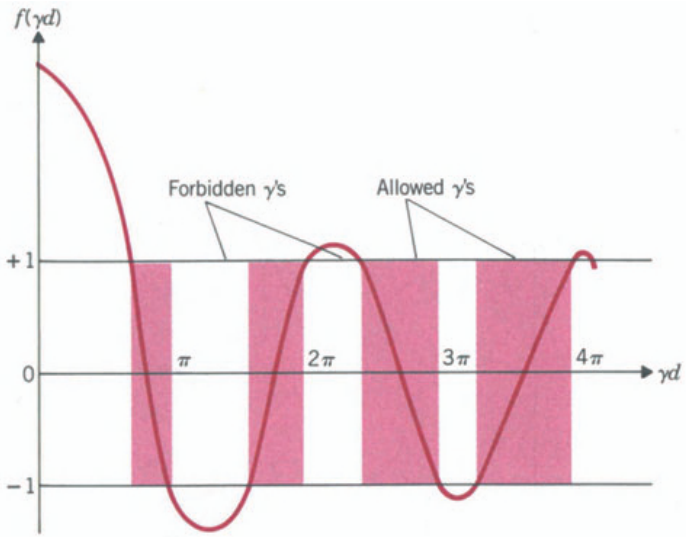


FIGURE 24-9
Plot of the function $f(\gamma d)$, that is, the left side of Eq. 24.3 versus γd (oscillating solid line). The shaded areas represent the values of γd for which $f(\gamma d)$ varies between 1 and -1. These values of γd correspond to allowed energy values. They are separated by ranges of γd for which $f(\gamma d)$ is either greater than 1 of less than -1. These values of γd correspond to forbidden energy values.

Because cos kd takes values ranging from $+1$ to -1, this means that this condition can be satisfied only by those values of γ for which $f(\gamma d)$ lies within those limits. The values of γ for which $f(\gamma d)$ is outside these limits correspond to γ's for which the boundary conditions cannot be satisfied and, therefore, these γ's (and the corresponding E's) are not physically acceptable. We conclude that *the electron may possess energies within certain bands of energy but not outside of them: There are allowed and forbidden bands of energy available to electrons moving in a periodic lattice.*

Another conclusion to be drawn from Fig. 24-9 is that the width of the allowed energy bands increases with increasing γ (increasing energy E). The physical reason for this will become clear when we look at an alternative way of showing how the bands come about in Section 24.4.

24.3b Dispersion Relation

For a free particle the relation between the energy E and the momentum p is $E = p^2/2m$. From de Broglie's relation $p = h/\lambda$ (Eq. 19.1), and because $\lambda = 2\pi/k$ (Eq. 11.12), it follows that

$$p = \frac{h}{\lambda} = \frac{hk}{2\pi} = \hbar k$$

Substitution of this relation for p into the energy relation shows that the energy of the particle can be expressed in terms of the wave vector as

$$E = \frac{\hbar^2 k^2}{2m} \tag{24.4}$$

The relation between the energy of a particle and its wave vector is often referred to as the *dispersion relation*. For a free particle, this relation is parabolic; that is, $E \propto k^2$, Eq. 24.4. This dependence is illustrated by the dashed lines of Fig. 24-10.

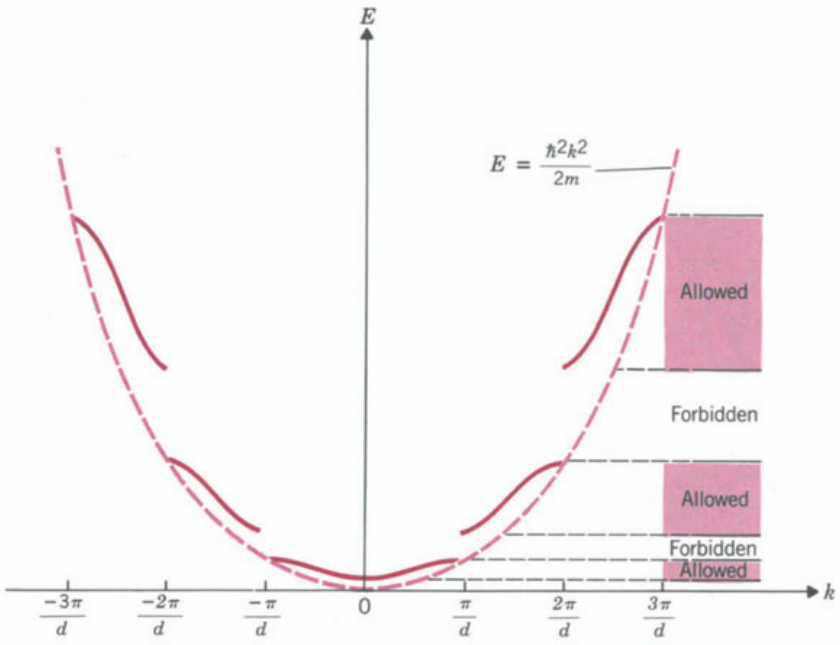


FIGURE 24-10
The solid lines show the dependence of the energy E on the wave vector k (and therefore, the momentum p) for an electron moving in the periodic array of rectangular potential wells of Fig. 24-4. The dashed line represents the relation between the energy and the mementum for the free electron case. The values of E for which Eq. 24.3 yields a real number for k are projected to the right of the figure as allowed energy bands (shaded bands). These are separated by ranges of E for which Eq. 24.3 yields an imaginary solution for k and correspond to forbidden energy values.

When the particle is not free, the dispersion relation is usually more complicated. Thus, as we have seen in Section 24.3a, for an electron moving in one-dimensional array of potential wells the dispersion relation is given by Eq. 24.3

$$P\frac{\sin \gamma d}{\gamma d} + \cos \gamma d = \cos kd \tag{24.3}$$

where

$$\gamma = \sqrt{\frac{2mE}{\hbar^2}}$$

We saw that Eq. 24.3 is transcendental and must be solved numerically. We pick a value of E, substitute it in Eq. 24.3, and obtain the value of k for which the relation holds. If the procedure is repeated systematically for other values of E, we will be able to make a table listing values of E and the corresponding values of k. As we proceed with these numerical calculations, we find that there are energy intervals for which no real solution for k exists. These are the values of E for which the left side of Eq. 24.3 is either greater than $+1$ or less than -1. As indicated in the previous section, this is physically unacceptable and, therefore, these energy values are forbidden. Results from the numerical method just described are illustrated by the solid lines of Fig. 24-10. We note that these solid lines yield the values of E and the corresponding values of k for certain ranges of E. For other energy intervals the value of k is not defined by the solid lines, these are the forbidden energies. Allowed and forbidden energies are projected to the right in Fig. 24-10 to represent the scheme of allowed and forbidden energy bands. Another important result to be noted in Fig. 24-10 is that the curvature of the solid lines is not the same

as that of the dashed line, which represents the dispersion relation for the free particle. This has important implications concerning the effective mass of the electrons and will be discussed in detail in Section 24.6.

Although the Kronig-Penney model presented here shows clearly the existence of allowed and forbidden energy bands and at the same time gives us a mathematical expression with which to find the E versus k curves, it does not give much physical insight for the existence of these bands. Moreover, as presented here, it does not answer a question that will be important in the understanding of the difference between conductors, insulators, and semiconductors. The question is: *How many energy states are allowed within a given band?* The answer so far seems to be an infinite number because within a band γ can vary continuously. It would seem, therefore, that E can take an infinite, continuous range of values within an allowed band. The reason for this result is that in the periodic potential model we have assumed that the periodicity is infinitely long. In a real solid we have boundaries and, although we may have 10^{23} potential wells, it is still a finite number of wells. If one introduces the boundary condition that we used in the infinite potential well model of Chapter 23, namely, $\chi = 0$ at the boundaries of the solid, one finds that the continuous spectrum within a band breaks into a quasicontinuous one. Rather than doing this directly, let us look at an alternative method of showing the existence of bands. This method is less quantitative, less mathematical, and less rigorous (although when used to do actual calculations it can be made rigorous), but it clearly shows the physical reason for the bands. This method will also yield the number of energy states allowed within a given band.

24.4 TIGHT-BINDING APPROXIMATION

One useful way to look at the formation of allowed and forbidden energy bands is to start with the energy levels of the individual neutral atoms when they are very far apart and watch the changes in these levels as the atoms get close together and the charge distributions of adjacent atoms begin to overlap. We can gain some insight into what happens by studying a simple one-dimensional quantum mechanical problem: that of two finite square potential wells.

When we examined the one-dimensional infinite potential well of width a, Chapter 20, we found that the eigenfunctions and corresponding energy values that the particle may have are given by Eqs. 20.25 and 20.26.

$$\chi_n = B \sin n \frac{\pi}{a} x \tag{20.25}$$

$$E_n = n^2 E_0 \qquad n = 1, 2, 3, \ldots \tag{20.26}$$

where

$$E_0 = \frac{\hbar^2 \pi^2}{2ma^2}$$

In Fig. 24-11 we show the first two eigenfunctions for the infinite potential well (see Fig. 20-4). For the finite well the results are similar but with minor differences. One such difference is that the eigenfunctions do not vanish at the boundary but extend a little bit outside the well (see Fig. 24-12).

Let us consider the two finite potential wells B and C shown in Fig. 24-13*a*. If we have an electron with energy E_1 that we know is definitely in well B, the χ describing such an electron will look as in Fig. 24-13*b*; χ_B and therefore χ_B^2 differ from 0 only in the region of the well B. Suppose instead that the electron is definitely in well C with energy E_1. Then the eigenfunction will appear as shown in Fig. 24-13*c*; χ_C and therefore χ_C^2 differ from 0 only in the region of well C. But let us assume that the electron can be found with equal probability in both wells with energy E_1. What eigenfunction do we use to describe the electron? To answer this question, let us put forward the properties that such a wavefunction must have.

1. χ must reflect the fact that the electron can be found with equal probability in both wells. This means that the probability χ^2 must be symmetric with respect to a point halfway between the two wells.
2. The part of χ that reflects the probability that the particle be found in well B with energy E_1 must look like the eigenfunction associated with that energy when the particle is in well B; that is, it must look like Fig. 24-13*b*. And the same applies to the part of the eigenfunction that reflects the probability of finding the particle in well C.

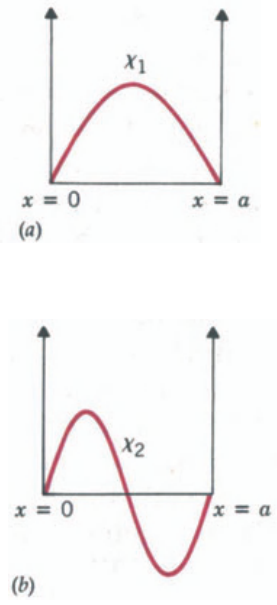


FIGURE 24-11
(*a*) Ground state eigenfunction χ_1 and (*b*) first excited state eigenfunction χ_2 for a particle in an infinite potential well.

Before we answer the question of eigenfunction selection, let us consider the following. We said that χ_C is an eigenfunction that can represent an electron in well C with energy E_1. Actually $-\chi_C$ could have been used instead. It has the same physical properties as χ_C, and it is a solution to the Schrödinger equation for the same value of the energy E_1,[3] and, moreover, $\chi_C^2 = (-\chi_C)^2$. Therefore, $-\chi$ must be considered equally with $+\chi$.

Let us now answer the question: What eigenfunction do we use to describe an electron with energy E_1 that can be found equally in both wells?

There are two possibilities:

$$\chi_S = a\,(\chi_B + \chi_C)$$

$$\chi_A = a\,(\chi_B - \chi_C)$$

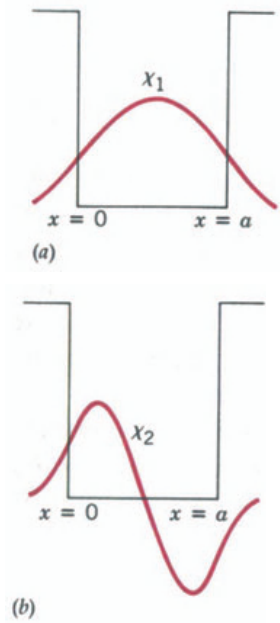


FIGURE 24-12
(*a*) Ground state eigenfunction χ_1 and (*b*) first excited state eigenfunction χ_2 for a particle in a finite potential well.

[3]Does $-\chi_C$ satisfy the Schrödinger equation for E_1? Let us substitute it into the Schrödinger equation,

$$-\frac{\hbar^2}{2m}\frac{d^2(-\chi_C)}{dx^2} + E_p(-\chi_C) = E_1(-\chi_C)$$

Multiplying both sides of this equation by -1, we get

$$-\frac{\hbar^2}{2m}\frac{d^2\chi_C}{dx^2} + E_p\chi_C = E_1\chi_C$$

which demonstrates the same equality as does χ_C; so both χ_C and $-\chi_C$ are equivalent solutions to the Schrödinger equations for the same energy E_1.

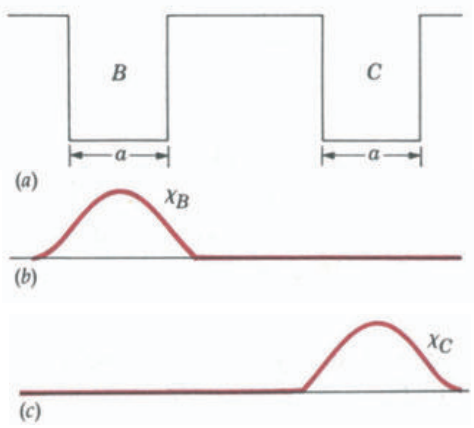


FIGURE 24-13
(*a*) Two finite potential wells *B* and *C* (*b*) Eigenfunction associated with an electron that is definitely in well *B* with the ground state energy. (*c*) Eigenfunction associated with an electron that is definitely in well *C* with the ground state energy.

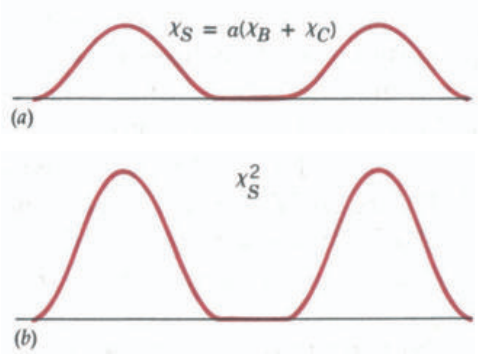


FIGURE 24-14
(*a*) Symmetric eigenfunction representing an electron that can be found with equal probability in the two wells of Fig. 24-13 with the ground state energy. (*b*) Probability density associated with the symmetric eigenfunction in (*a*).

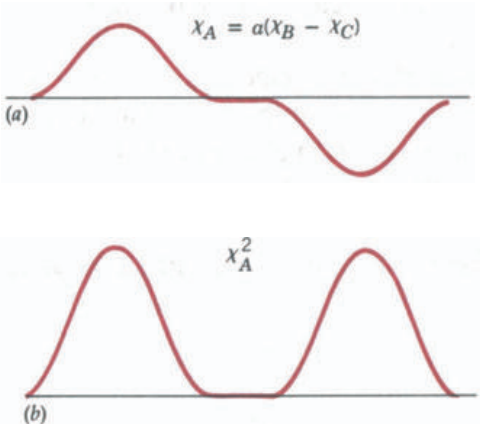


FIGURE 24-15
Antisymmetric eigenfunction representing an electron that can be found with equal probability in the two wells of Fig. 24-13 with the ground state energy. (*b*) Probability density associated with the antisymmetric eigenfunction of (*a*).

The constant a is introduced for normalization purposes; its value is not important to our arguments. It should be clear that χ_S (symmetric) shown in Fig. 24-14*a* satisfies the conditions set forth here. Figure 24-14*b* shows that χ_S^2 is symmetric with respect to the midpoint between the wells, and this reflects the fact that t : particle can be found with equal probability in the two wells. The other possibility, χ_A (antisymmetric), Fig. 24-15*a*, also satisfies the conditions set forth, as can be seen by comparing Fig. 24-15*b* for χ_A^2 with Fig. 24-14*b* for χ_S^2.

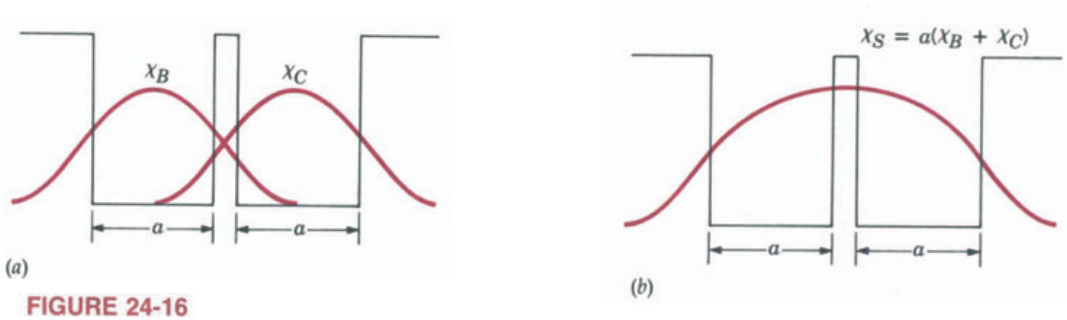


FIGURE 24-16

When the two wells are very close together, the symmetric eigenfunction of Fig. 24-14 looks like the ground state eigenfunction for a finite well of width $2a$ (see Fig. 24-12a).

Let us see what happens to these two eigenfunctions when the separation between the two wells becomes very small. Figure 24-16a illustrates the individual eigenfunctions χ_B and χ_C, and Fig. 24-16b is the sum $\chi_S = a\,(\chi_B + \chi_C)$. The eigenfunction χ_S begins to look like the ground state eigenfunction for a well of width $2a$. In fact, in the limit of no separation it becomes the ground state eigenfunction. On the other hand, the antisymmetric eigenfunction appears as in Fig. 24-17. Figure 24-17a shows the individual eigenfunctions χ_B and $-\chi_C$, and Fig. 24-17b their sum, $\chi_A = a\,(\chi_B - \chi_C)$. The wavefunction χ_A begins to look like the wavefunction of the first excited state for a well of width $2a$.

We conclude that although when the two wells were far apart both χ_S and χ_A were degenerate eigenfunctions (same energy states), the degeneracy begins to disappear as the two wells get close to each other; χ_A corresponds to a state of higher energy than χ_S.

In our discussion we assumed that the electron can be in either well B or well C. Of course, when the two wells are far apart the question is purely academic. However, when they come close together so that the wavefunctions from the two wells overlap, it is possible for the electron that initially was in well C to move into well B, because its wavefunction is not zero at the location of well B, and the same holds for the electron that initially was in well B; it

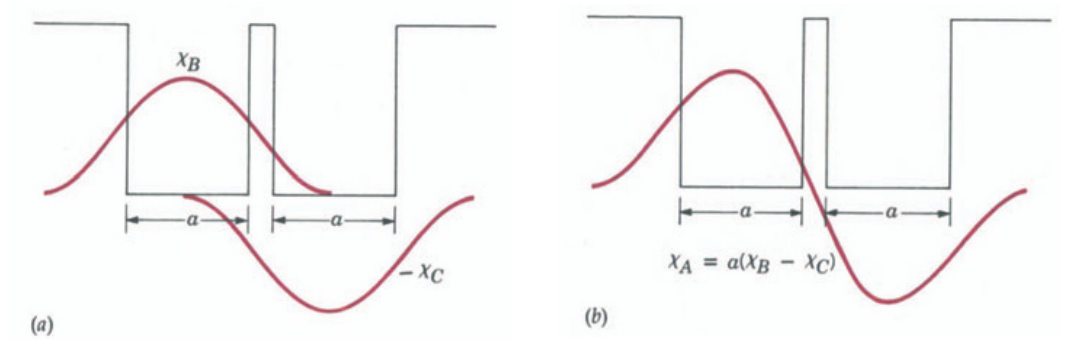


FIGURE 24-17

When the wells are very close together, the antisymmetric eigenfunction of Fig. 24-15 looks like the first excited state of a finite well of width $2a$ (see Fig. 24-12b).

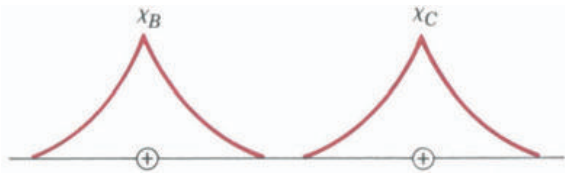


FIGURE 24-18
Radial part of the ground state eigenfunctions of two isolated hydrogen atoms as a function of the distance of the electrons from the respective nuclei.

can move into well C because its wavefunction has a finite value in well C. Thus, even if the electron was initially in one well, after the two wells are brought close together, if we wait long enough, there is no way to predict exactly in which one it will be. Similarly, if we had started with two wells and one electron in each well, after they are brought together, there is no way of saying in which well either of the two electrons will be. The only thing that we can say is that the probability of finding an electron in one well is the same as the probability of finding it in the other. Either χ_A or χ_S will be appropriate to describe its behavior because both χ_A^2 and χ_S^2 are symmetric with respect to the midpoint between the wells.

The important conclusion from this simple artificial example is that *if you start with two identical χ's (same energy) in two identical independent systems, when you bring the two together, the two degenerate χ's break up into two nondegenerate χ's*. However, this example does not give the reason for this result. What is the physical reason for the breakup of the degeneracy?

Let us look at a real example of two hydrogen atoms, each with its electron in the 1*s* ground state. If one solves the radial part of the Schrödinger equation for the hydrogen atom, Eq. 21.6, one obtains the radial part of the eigenfunction for the 1*s* state as $\chi = e^{-r/r_0}$. In the one-dimensional sketch, the eigenfunction decreases exponentially as the distance r from the nucleus increases. The eigenfunctions χ_B and χ_C in Fig. 24-18 are those associated with the two independent atoms B and C. As the two atoms are brought together, the eigenfunctions overlap, and the electrons from B and C can change places. We are led to consider, just as before, the two possible combinations of χ_B and χ_C shown in Fig. 24-19, where the dashed lines represent how the eigenfunctions of each would appear if the other were not present, and the solid line is the sum or combined eigenfunction. The electron distribution between the two protons can be seen by plotting the probability functions $|\chi_S|^2$ and $|\chi_A|^2$, Fig. 24-20. Both distributions are symmetric with respect to the midpoint between the two protons, and therefore the probability of finding an electron at a certain distance from one proton is the same as the probability of finding it at the same distance from the other proton. Both eigenfunctions are thus suitable to represent the behavior of either electron. However, an electron in a state χ_S has a lower energy than one in χ_A. The reason is that an electron in χ_S is more likely to be between the two protons than being near to just one. Figure 24-20*a* shows that the probability function χ_S^2 in the region between the two protons is greater than on either side of the protons.

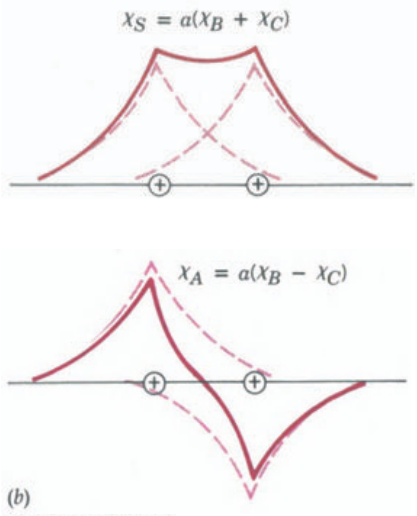


FIGURE 24-19
(*a*) Symmetric combination of the ground state eigenfunctions of the two individual hydrogen atoms of Fig. 24-18. (*b*) Antisymmetric combination of the same two eigenfunctions.

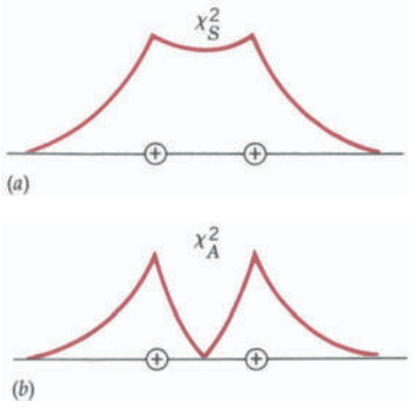


FIGURE 24-20
(*a*) Probability density associated with the symmetric eigenfunction of Fig. 24-19*a*. (*b*) Probability density associated with the antisymmetric eigenfunction of Fig. 24-19*b*. Note that χ_S^2 is large in the region between the two nuclei so therefore the electron represented by χ_S spends considerable time between both nuclei. χ_A^2 does not have this feature.

As a result, the electron spends a considerable amount of time between the two protons. In this region it is under the attractive influence of both protons at once. The binding energy of the electron resulting from the presence of the two protons will be more negative than if it was only under the influence of one of them. On the other hand, an electron in state χ_A spends its time with either one proton or the other. It is hardly ever with both (see Fig. 24-20*b*) and, as a result, this extra contribution to the binding is not there or at least is very small.

Thus, when two atoms are brought together, two separate energy levels are formed from each level of the isolated atom. The physical reason for this effect is the differing ways that the electrons interact with the ions in the symmetric and antisymmetric states.

This is not a band, but two atoms do not make a solid. Suppose that N atoms are brought together to form a solid, then each of the levels of the individual isolated atoms breaks up into N discrete, closely spaced levels and becomes a band of energy levels. This can be illustrated with the 1*s* level of six hydrogen atoms. If we start with six individual 1*s* states and consider all the possible ways of adding the six individual 1*s* states, we get six types of combinations having different energies (see Fig. 24-21). In the first level, the

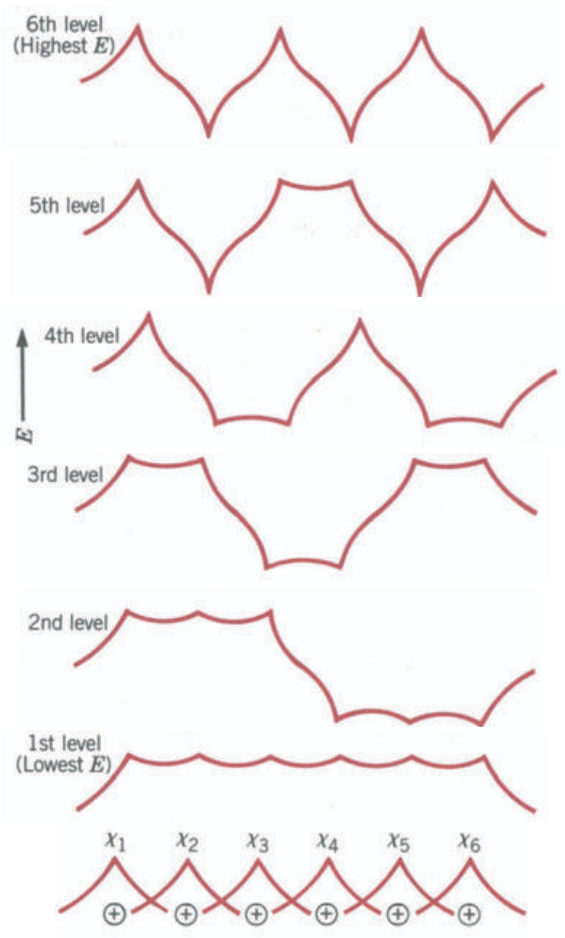


FIGURE 24-21
Six possible combinations of the ground state eigenfunctions of six hydrogen atoms, each corresponding to a different energy state. At the bottom of the figure the six individual eigenfunctions are sketched. The sketches above them represent six possible ways of adding them either symmetrically or antisymmetrically.

six individual eigenfunctions all add symmetrically,

$$\chi_{\text{first level}} = \chi_1 + \chi_2 + \chi_3 + \chi_4 + \chi_5 + \chi_6$$

As a result, there are *five* places along the lattice where the electron will be under the attractive (binding) influence of two nuclei. In the second level, the χ's from the first three atoms add symmetrically among themselves and so do the last three, but the resulting χ from the first three adds antisymmetrically with the resulting χ from the last three:

$$\chi_{\text{second level}} = (\chi_1 + \chi_2 + \chi_3) - (\chi_4 + \chi_5 + \chi_6)$$

As a result, there are only *four* places along the lattice where the electron will be under the simultaneous influence of two nuclei that gives the extra negative contribution to the energy. The energy of an electron in this level will be higher than that of an electron in the first level. The combination of the individual χ's to form the rest of the levels is done in a similar way as the first two examples. The result concerning the energy of each can also be reasoned in a similar way: In the third level, there are only *three* places where the electron would be under the simultaneous influence of two nuclei and, consequently, the energy of the level will be higher than the previous two. This goes on until the sixth level, where this extra contribution to the binding is nowhere in the lattice. Therefore, this state corresponds to the highest energy.

Are these the only possible combinations of the six χ's? The answer is no; but any other combination will be degenerate with one of the preceding as shown in the examples of Figs. 24-22*a* and *b*. The two examples shown in Fig. 24-22 are different combined χ's, but they both have the same energy as the second level of Fig. 24-21. You may play the permutation game to achieve other combinations, but you will find that as far as the energy is concerned, all will fall within one of the six considered in Fig. 24-21.

It should be apparent that when we extend this analysis to a lattice of N atoms, the individual identical states of the atoms will give rise to N *different energy states*. However, regardless of the number, the two extremes will look like the first and sixth level of our example. It should also be apparent that the energy difference between the two extremes and, therefore the width of the band, should not depend appreciably on N. Increasing N increases the number of sites where the extra contribution to the energy can take place. But it simultaneously decreases the amount of time that the electron spends in any one site, making the total time in all such sites constant. What affects the width of the band is how close any two atoms are to each other. The

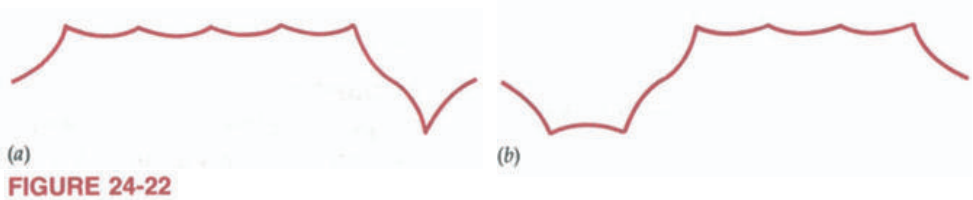


FIGURE 24-22
Two alternative combinations of the six eigenfunctions in Fig. 24-21 that correspond to the same energy as the second level of Fig. 24-21.

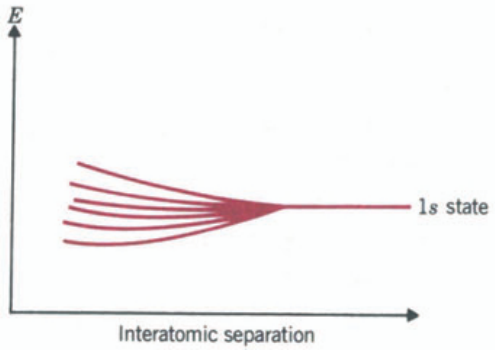


FIGURE 24-23
Splitting of the 1*s* state of six hydrogen atoms into a band of six energy levels as the separation between the atoms decreases. Note the increase in the bandwidth with decreasing interatomic separation.

closer, the greater the overlap of the wavefunctions and therefore the larger χ^2 will be between the two nuclei for symmetric states, with the consequence of stronger binding. Figure 24-23 shows how the wavefunction for six different atoms separates into six different energy levels when the atoms are brought close enough together. And the closer they are brought, the greater is the separation of the energy levels. For the usual interatomic separation found in a solid, the width of the band is typically a few electron volts. For a macroscopic solid with $N \sim 10^{23}$, the separation between adjacent levels will be $\sim 10^{-23}$ eV, an insignificant amount.

We thus have shown that when one brings a large number N of atoms together to form a solid, the individual atomic energy levels of the atom break up into a quasicontinuous energy band. Within the band there are N distinct but very close energy levels.

So far, we have limited ourselves to one-electron atoms where the electron is in the 1*s* state. When the analysis is extended to multielectron atoms where we have electrons in other states, one finds that each of the individual atomic states breaks up into similar bands of quasicontinuous states. Thus, if we consider sodium (Na) with an electronic configuration $1s^2\ 2s^2\ 2p^6\ 3s^1$ (Section 21.7b), we may expect the band structure illustrated schematically in Fig. 24-24. Notice that the bandwidth for the low-lying levels is smaller than for the higher energy ones. The reason is that electrons in the lower levels are electrons in the inner subshells of the atoms; these electrons are not influenced much by the presence of other atoms because their wavefunctions do not

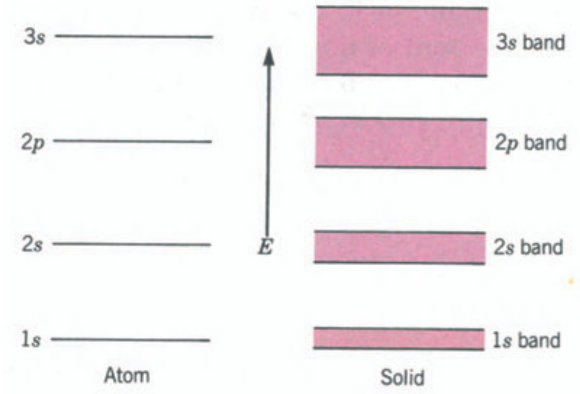


FIGURE 24-24
Expected splitting of the first four atomic levels of sodium into four energy bands in a sodium crystal.

overlap significantly with those of the electrons of other atoms. Therefore, they give rise to narrower bands.

The situation depicted in Fig. 24-24 is what one may expect in general. The situation for real solids is somewhat more complicated. We must realize that we have been using qualitative arguments and one-dimensional models for the sake of mathematical as well as conceptual simplicity. These models have educational value because they bring out the main features of band theory. However, if we want to get theoretical results that can be compared with those of the experiments, we must face the three-dimensional world, and our qualitative arguments must become quantitative. When this is done, we should not be surprised that the simple pictures may have to be modified somewhat. The main features that we have found—namely, that each atomic level breaks up into a band and that in each band there are N energy levels—are retained. These two features will enable us to understand the differences between conductors, insulators, and semiconductors.

24.5 CONDUCTORS, INSULATORS, AND SEMICONDUCTORS

We are now in a position to understand why some solids are good conductors, and some are not. We must keep in mind two facts.

1. As we explained in Section 23.3f, for electrons to experience an acceleration in the presence of an electric field $\mathscr{E}$ and therefore to contribute to the current, they must be able to move into new, slightly higher energy states. This means that the states that are available for the electrons must be both *empty* and *allowed*. For example, if relatively few electrons reside in an otherwise empty band, a large number of unoccupied states are available into which the electrons can move; these electrons can acquire energy from the electric field and contribute to the current. On the other hand, if a band is full, then the electrons in that band cannot contribute to the current because they cannot move into slightly higher energy states. They therefore cannot be accelerated by the electric field.
2. There is a limit to the number of electrons that can be placed in a given band. We know that there are N different energy states in each band. If the band is an s band (one formed from atomic s states) then, the orbital quantum number $l = 0$ and therefore $m_l = 0$ and $m_s = \pm 1/2$. We can place two electrons in each of the N states without violating Pauli's exclusion principle. In a p band $l = 1$, $m_l = 0, 1, -1$ and for each value of m_l, $m_s = \pm 1/2$. In each of the N energy levels we can put six electrons. In general, because for a given l there are $(2l + 1)$ values of m_l and for each m_l there are two values of m_s, we have $2(2l + 1)N$ openings available to the electrons in a given band; for example, in a d band the number is $2(2 \times 2 + 1)N = 10N$.

Let us consider some hypothetical examples on the basis of these two facts. Consider a solid of N atoms with each atom having 11 electrons. Altogether there are $11N$ electrons. $2N$ electrons may be put in the $1s$ band, $2N$ in the $2s$ band, $6N$ in the $2p$ band. There remain N electrons that may be placed in the next available band, the $3s$. But the $3s$ band has room for $2N$ electrons, and therefore it will be only half full (see Fig. 24-25). As a consequence, half the states in the $3s$ band (the unoccupied ones) are available to the N electrons in that band. Because there are available states, the N electrons in this band can be accelerated by an electric field and move into higher energy states. *This solid would be an electrical conductor.* Note that because the N electrons in the $3s$ band obey the Pauli exclusion principle, they obey Fermi-Dirac statistics; that is, at $T = 0$ K they all reside in the lowest energy levels of the band and the highest occupied level becomes the Fermi level. For $T > 0$, a few (as we saw in the previous chapter) can be above the Fermi level. The N electrons in the $3s$ band behave as predicted by the QMFE model with one minor modification that we will discuss later. We can see that the effect of the periodic potential on the motion of electrons in a solid with a partially filled band is unimportant, and that is the reason why the QMFE model was so successful in predicting the properties of conductors. This hypothetical solid is of course Na ($Z = 11$). The only difference between the hypothetical solid and the real Na is that in sodium metal the next band, the $3p$, overlaps the $3s$ band. But all that this does is to provide additional empty energy levels for the N electrons in the $3s$ band.

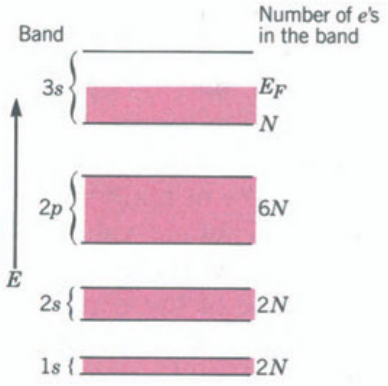


FIGURE 24-25
Schematic of the occupation of the bands by electrons in a sodium crystal of N atoms and having, therefore, $11N$ electrons. The highest energy band with electrons ($3s$ band) is only half full with N electrons, and thus sodium is a monovalent metal. The $3s$ band is the conduction band of sodium.

The same arguments used for Na apply to lithium ($1s^2\ 2s^1$), potassium ($1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^1$), as well as to rubidium and cesium. (See the periodic table in Chapter 21.)

It should be noted that the highest energy band containing electrons is called the *valence band.* If, as in the case of sodium, this band is only partially filled, it is also called the *conduction band* because electrons in that band are responsible for conduction processes.

Next consider magnesium ($Z = 12$) with electronic configuration $1s^2\ 2s^2\ 2p^6\ 3s^2$. In a solid with N atoms, there are $12N$ electrons. Following the previous scheme, $2N$ electrons go into the $1s$ band, $2N$ into the $2s$, $6N$ into the $2p$, and the remaining $2N$ into the $3s$. All $12N$ electrons have been accounted for in the process. They have completely filled the $3s$ band. According to our previous argument, there are no empty energy states available for the electrons in the $3s$ band to move into; therefore, they cannot contribute to conduction. With no empty states through which charged particles may contribute to conduction, Mg should be an insulator. But it is not. The reason is that as in the case of Na, the $3p$ and $3s$ bands overlap (see Fig. 24-26). Because the $3p$ band has $6N$ empty states, the $2N$ electrons have available to them $2N + 6N$ states of which only $2N$ are occupied. Mg is therefore a conductor. Similar arguments apply to beryllium, calcium, zinc and barium, all of which are in the same group in the periodic table.

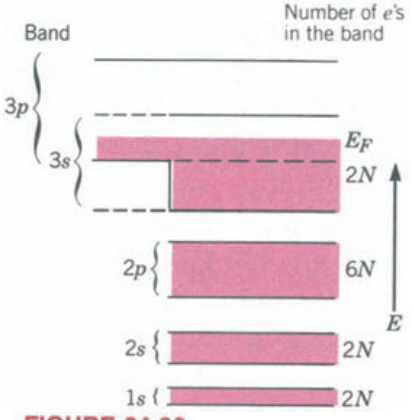


FIGURE 24-26
Occupation of the energy bands in a magnesium crystal of N atoms ($12N$ electrons). The overlap between the $3s$ and the $3p$ bands make magnesium a metal.

Now consider carbon ($1s^2\ 2s^2\ 2p^2$) in its diamond structure. As N atoms of C are brought together, they have $6N$ electrons. $2N$ fill the $1s$ band, $2N$ fill

the 2*s* band, and there are 2*N* electrons left to place in the next available band, the 2*p*, which has room for 6*N* electrons. The 2*p* band would be a partially filled band with plenty of empty states available (see Fig. 24-27). Diamond should therefore be a conductor. But it is not. It is an excellent insulator. As we mentioned before, the qualitative arguments elucidate the main features of band structure. However, when dealing with a specific crystalline material, the arguments must become quantitative. When this is done, interesting features occur, such as band overlap. In the case of diamond, germanium, and silicon, an even more interesting feature is revealed.

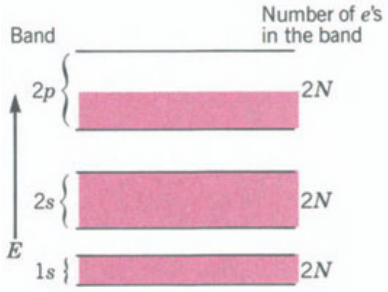


FIGURE 24-27
Expected occupation of the energy bands in a diamond crystal of N atoms.

As the carbon atoms are brought together to form diamond, the energy levels begin to split into bands starting with the outermost shell, $n = 2$ (2*s* and 2*p* levels) (Fig. 24-28). As the interatomic spacing decreases farther, the 2*s* and 2*p* bands begin to overlap and merge into a single '2*s* 2*p*' band with 8*N* states available. As the separation decreases even farther, approaching the interatomic equilibrium spacing r_0, the '2*s* 2*p*' band splits again into two hybrid bands separated by an energy gap E_g, which increases with decreasing separation. The value of E_g is about 6 eV for the equilibrium distance of $r_0 \approx 1.5 \times 10^{-10}$ m. However, each of these two bands now contains 4*N* states. The result: Of the total 6*N* electrons, 2*N* go into the 1*s* band and the remaining 4*N* into the lower hybrid '2*s* 2*p*' band and fill it. Thus, at $T = 0$ K the valence band (the lower '2*s* 2*p*' band) is full (Fig. 24-29), and diamond is an insulator. Note that this is only true for the diamond structure of carbon, not for graphite. The bands of germanium (Ge) and silicon (Si) show a similar behavior. In the case of Si, the mixing and subsequent splitting occurs between the 3*s* and 3*p*, whereas in the case of Ge it occurs between the 4*s* and the 4*p*. There is, however, an important *quantitative* difference between diamond and Si and Ge. The energy gap E_g between the filled valence band and the next empty band for Ge and Si is much smaller than for C: E_g (Ge) = 0.7 eV, E_g (Si) = 1.1 eV. At $T = 0$ K, pure C, Si, and Ge behave identically. They are perfect insulators because the valence band is filled. However, as T increases, some of the electrons in the valence band can be thermally excited across the energy gap into the next band, which now becomes the *conduction band*, and as a result electrical conduction can take place. How many electrons can be excited

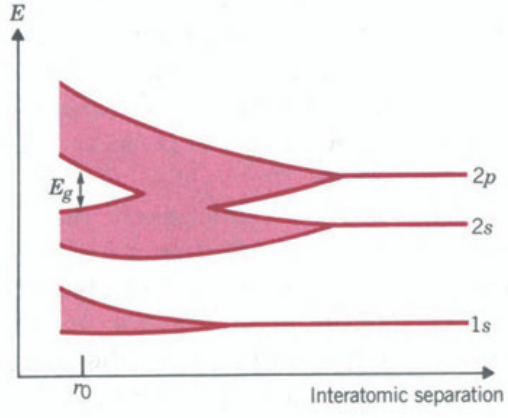


FIGURE 24-28
The splitting of the atomic energy levels of carbon into energy bands is followed by the merging of the 2*s* and 2*p* bands and a subsequent splitting of these bands as the interatomic spacing decreases. At the equlibrium interatomic spacing r_0, an energy gap E_g separates two hybrid '2*s* 2*p*' energy bands in a diamond crystal.

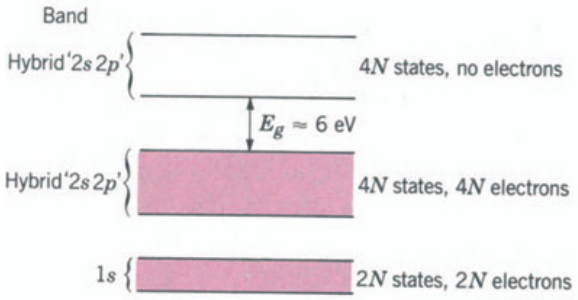


FIGURE 24-29
Actual occupation of the energy bands in a diamond crystal with N atoms ($6N$ electrons). The lower hybrid '$2s$ $2p$' band is full and separated by an energy gap E_g from the higher hybrid '$2s$ $2p$' band, which has no electrons.

depends on how big E_g is and, of course, on T. The higher T, the greater the thermal energy, and therefore the greater the number of electrons that will be able to make the jump across the energy gap and, naturally, the greater the electrical conductivity. This is the reason why the conductivity of an insulator and of a semiconductor increases with T (as opposed to a metallic conductor). We will discuss this further in the next chapter.

The number of electrons in the conduction band at a given temperature will depend on E_g, the magnitude of the energy gap between the valence and the conduction bands. The smaller E_g, the greater the probability that the electrons at the top of the valence band will jump into the conduction band. Thus, at a given temperature, we expect that Si ($E_g \approx 1$ eV) will have more electrons in the conduction band and therefore, be a better conductor than diamond ($E_g = 6$ eV). In fact, as we will show in Chapter 25, the probability of transition across the energy gap is very sensitive to the magnitude of E_g. A doubling of E_g will reduce the number of conduction electrons by several orders of magnitude (powers of 10). It is the magnitude of E_g that determines whether a solid is an insulator (diamond) or a semiconductor (Si, Ge) at ambient temperatures. Detailed calculations of the number of conduction electrons and its dependence on E_g and T are presented in Chapter 25.

24.6 EFFECTIVE MASS

When an electric field $\mathscr{E}$ acts on a *free* electron, it exerts a force $e\mathscr{E}$ that, from Newton's law, will produce an acceleration inversely proportional to its mass, $a = e\mathscr{E}/m$. What happens when the electron to be accelerated is not free but happens to be in a crystal under the influence of the potential of the lattice ions? The answer is that it will still accelerate according to Newton's law; however, the electron responds as if it had some *effective mass*, which is different from its true mass. As we will show, this is because $\mathscr{E}$ is not the only electric field acting on the electron inside the crystal.

We will introduce this concept by using a semiclassical picture: an argument that is half classical and half quantum mechanical. The quantum mechanical part lies in the fact that the motion of an electron is governed by a wave, and that the velocity of the electron is equal to the group velocity v_{group} of the wave, that is, the velocity of the envelope, of the wave packet. In our treatment of matter waves and wave packets (Section 19.7), we saw

that the group velocity is given by Eq. 19.10:

$$v_{\text{group}} = \frac{dE}{dp} \tag{19.10}$$

where E is the energy of the particle and p is its momentum.

In the case of a free particle, we can readily show that the group velocity is equal to the particle velocity. For a free particle, the energy

$$E = \frac{1}{2} m v_{\text{particle}}^2 = \frac{p^2}{2m}$$

Therefore

$$v_{\text{group}} = \frac{dE}{dp} = \frac{d}{dp}\left(\frac{p^2}{2m}\right) = \frac{p}{m} = \frac{m v_{\text{particle}}}{m} = v_{\text{particle}}$$

Although we have shown that $v_{\text{group}} = v_{\text{particle}}$ for the free particle case only, it can be shown that the relation holds even when the particle is not free, such as the case of an electron in a lattice.

Equation 19.10 defines the group velocity in terms of the energy E and the momentum p of the particle. As we have seen in the Kronig-Penney model, the energy is often expressed in terms of the wave vector k. It is convenient, therefore, to define the group velocity in terms of E and k. This can be done by using de Broglie's relation $p = h/\lambda$ (Eq. 19.1) and the fact that $\lambda = 2\pi/k$ (Eq. 11.12). Combining Eqs. 19.1 and 11.12 we have $p = \hbar k$ and hence $dp = \hbar\, dk$. Substituting this result for dp in Eq. 19.10, we obtain

$$v_{\text{group}} = \frac{1}{\hbar}\frac{dE}{dk} \tag{24.5}$$

The classical part of the argument uses the definition from mechanics that if a force does work dW on a particle, the energy of that particle increases by the same amount: $dE = dW$. Applying this to the present case, we have

$$dE = dW = e\mathscr{E}\, dx = e\mathscr{E}\frac{dx}{dt}\, dt = e\mathscr{E}\, v_g\, dt$$

The rate at which the energy of the particle is changing is therefore

$$\frac{dE}{dt} = e\mathscr{E}\, v_g \tag{24.6}$$

We also know from Newton's law that when a force acts on a particle, it will be accelerated. By definition, the acceleration a is

$$a = \frac{dv_{\text{particle}}}{dt} = \frac{dv_g}{dt}$$

Substituting Eq. 24.5 for v_g we obtain

$$a = \frac{1}{\hbar}\frac{d}{dt}\frac{dE}{dk}$$

Interchanging the order of the differentiation of E, we may write

$$a = \frac{1}{\hbar}\frac{d}{dk}\frac{dE}{dt} \tag{24.7}$$

Substituting Eq. 24.6 for dE/dt in Eq. 24.7 yields

$$a = \frac{e\mathscr{E}}{\hbar}\frac{dv_g}{dk}$$

From Eq. 24.5, it follows that

$$a = \frac{1}{\hbar^2}\frac{d^2E}{dk^2}e\mathscr{E}$$

Rearranging terms, we get

$$e\mathscr{E} = \frac{\hbar^2}{\dfrac{d^2E}{dk^2}}\ a \tag{24.8}$$

Noting that $e\mathscr{E}$ is the force of the externally applied electric field, we conclude that Eq. 24.8 has the form $F = m^*a$, where

$$m^* = \frac{\hbar^2}{\dfrac{d^2E}{dk^2}} \tag{24.9}$$

$$m^* = \frac{\hbar^2}{\dfrac{d^2E}{dk^2}}$$

The response of the electron in the solid to an externally applied electric field is as if it had an effective mass m^* given by the expression in Eq. 24.9. Let us see if Eq. 24.9 gives the correct result for the free electron case. From Eq. 24.4, for the free electron

$$E = \frac{\hbar^2k^2}{2m} \tag{24.4}$$

and

$$\frac{dE}{dk} = \frac{\hbar^2k}{m}\ ;\quad \frac{d^2E}{dk^2} = \frac{\hbar^2}{m}$$

Substitute this result into Eq. 24.9 and obtain

$$m^* = \frac{\hbar^2}{\dfrac{\hbar^2}{m}} = m$$

When the electron is free, the effective mass is the true mass, as it should be. However, when the electron is in a crystal, m^* is different from m because the energy is not proportional to k^2, as we saw in the Kronig-Penney model.

The physical reason is the following. The electron in the crystal moves under the influence of *internal forces* exerted by the electric fields of the ions

of the lattice and the *external force* resulting from the externally applied electric field. If we choose to use Eq. 24.8 to describe the motion of the electrons, we describe the motion in terms of the *external force alone*. However, the effect of the internal forces is hidden in m^*.

Let us recall the results that we obtained from the Kronig-Penney model concerning the relation between E and k (Fig. 24-10). The relation for the first allowed band is represented by the solid line in Fig. 24-30*a*. For small k's both curves, the free electron one (the dashed line) and the one obtained from the model, are quite similar. Both the first and the second derivatives are almost the same; therefore, m is about equal to m^*. Notice, however, that in the Kronig-Penney curve the first derivative (the slope) first increases for small values of k and then decreases as k approaches π/d. We illustrate a qualitative plot of the slope dE/dk versus k in Fig. 24-30*b*. We may also draw a qualitative illustration of d^2E/dk^2 from Fig. 24.30*b*. Its appearance is that of Fig. 24-30*c*. The reciprocal of d^2E/dk^2 is proportional to the effective mass, from Eq. 24.9. These reciprocal curves are shown in Fig. 24-30*d*. We see that for small values of k the effective mass m^* is essentially equal to the mass of a free electron m. As dE/dk approaches the maximum, d^2E/dk^2 begins to decrease and m^* increases. When dE/dk reaches the maximum, $d^2E/dk^2 = 0$ and m^* becomes infinite. Subsequently, d^2E/dk^2 becomes negative and m^* is negative.

We may draw the following conclusions about the effective mass m^* of an electron moving in a periodic lattice (see Fig. 24-30*d*).

1. m^* is not always equal to m.
2. m^* can be greater than m and, in fact, infinite.
3. m^* can be less than m or even negative.

We can gain some understanding of the behavior of m^* if we consider the way an electron wave is reflected by the lattice ions. We know from Section 12.7 that a wave is totally Bragg reflected if

$$2d \sin \theta = n\lambda \qquad (12.11)$$

In the case of a one-dimensional lattice that we have been considering, $\theta = 90°$ (Fig. 23-21) and Eq. 12.11 becomes

$$2d = n\lambda$$

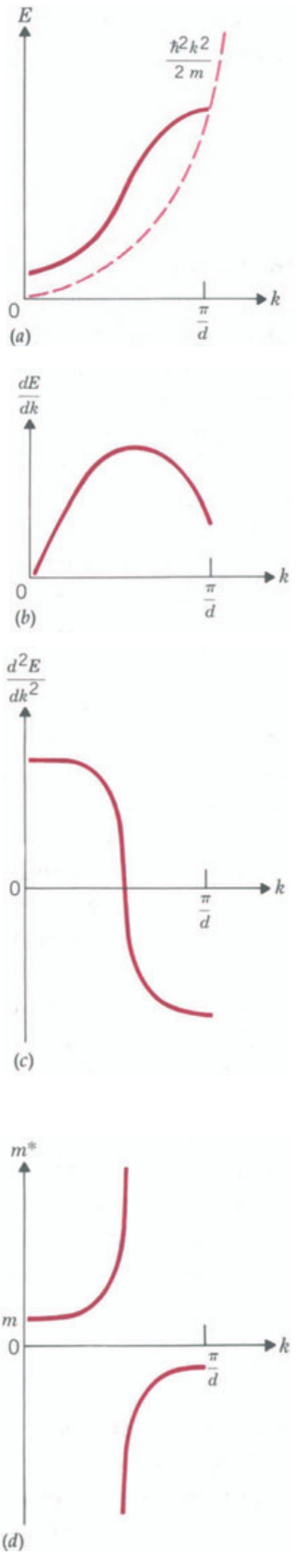


FIGURE 24-30
(*a*) Dependence of the energy E of an electron on its wave vector k, and hence its momentum p (solid line) for the first allowed energy band of the Kronig-Penney model (see Fig. 24-10). The dashed line illustrates the relation between E and k for the free electron case. (*b*) Qualitative plot of the derivative of E with respect to k as a function of k for the solid line of (*a*). (*c*) Qualitative plot of the second derivative of E with respect to k as a function of k. (*d*) Dependence of the effective mass m^* of the electron on the wave vector k for the situation depicted by the solid line of Figure (*a*). Note that the effective mass is negative near the top of the band, that is, as k approaches π/d.

This is sketched in Fig. 24-31. We can express the Bragg equation in terms of the wave vector $k = 2\pi/\lambda$

$$2d = n\lambda = n\frac{2\pi}{k}$$

and

$$k = n\frac{\pi}{d}$$

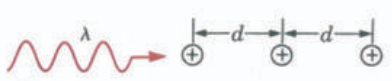


FIGURE 24-31
Bragg scattering of an electron wave by a one-dimensional array of ions.

At the bottom of a band where $k \approx 0$, there is practically no reflection because the Bragg condition is far from being satisfied. The lattice ions will have little effect on the electron when it is accelerated by the external field $\mathscr{E}$. Higher up in the band, k will get closer to the critical value π/d and reflection starts to become appreciable. In this region, as the external field $\mathscr{E}$ accelerates the electron, the momentum increases and gets closer to the critical value. The external field increases the forward momentum, but at the same time enhances reflection, and the reflection corresponds to reversing the sign of the momentum. At the point where $1/m^* = 0$ (that is, $m^* = \infty$), the gain in the forward momentum resulting from the applied $\mathscr{E}$ is exactly compensated by the resulting enhancement in reflection by the lattice ions. The net change in the forward momentum is zero. Thus the overall response of the electron to the field $\mathscr{E}$ is as if it had an infinite mass; that is, it cannot be accelerated. At the top of the band, even closer to the critical value for total reflection, the second effect (enhanced reflection) is more important than the direct action of the applied field. The net result is that the electron responds with a change in momentum that is in the opposite direction to what the free electron would have acquired: the electron responds as if it had a negative mass.

The concept of effective mass has many uses. For example, in the free electron models we derived an expression for the electrical conductivity σ by considering the response of the electrons to an electric field, and we found the conductivity $\sigma \propto 1/m$ (Eq. 23.14). We can easily make the correction to the model, that is, account for the fact that the electrons are not really free but instead they move in the electric potential of the ions. All that has to be done is to replace the true mass m by the effective mass m^*. In many metals this has little effect because $m^* \approx m$ (for example, Cu, Na, Al, and K). But in some metals it has a significant effect. As an example, the average m^* for iron (Fe) is about 10 times the free electron mass m, and this is one reason why iron is not a very good electrical conductor.

What about negative mass? Are there any cases where the electrons travel in the same direction as the electric field? These questions lead us to our next topic.

24.7 HOLES

The concept that electrons near the top of the band have negative effective mass leads to a very interesting feature that has a tremendous importance in the operation of all semiconductor devices.

At $T = 0$ K, the band structure of a semiconductor is characterized by a

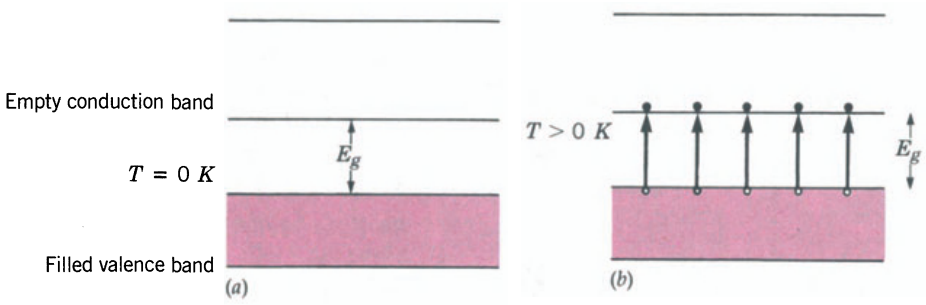


FIGURE 24-32
(*a*) At $T = 0$ K, the valence band of a semiconductor is filled with electrons and separated by an energy gap E_g from an empty conduction band. (*b*) At $T > 0$ K electrons are thermally excited into the conduction band, leaving behind in the valence band unoccupied energy states called *holes*, which behave as mobile positive charge carriers.

fully occupied valence band and a completely empty conduction band (see Fig. 24-32*a*). The semiconductor ideally is an insulator with zero conductivity at $T = 0$ K. As the temperature is raised, some electrons in the valence band can receive enough thermal energy and be excited into the conduction band because the energy gap between the two bands is rather narrow. The result is that there are some electrons in an otherwise empty conduction band and some unoccupied states in an otherwise filled valence band, see Fig. 24-32*b*. An empty state in the valence band is called a *hole*.

The electrons in the conduction band can move under the influence of an external electric field because they have available to them many empty higher energy states, and they can contribute to the current density J. Similarly, the electrons in the valence band can move into the empty states (holes) left by the electrons that were excited into the conduction band. We will assume, based on our previous discussion, that the empty states at the top of the valence band are *negative effective mass states*.

The interesting and important feature that we mentioned before is that *the conduction by the electrons in the valence band as they move into the empty negative mass states is completely equivalent to the conduction by particles of positive charge and positive mass. The number of such $+q$, $+m$ particles is equal to the number of available empty states, that is, the number of "holes."* Basically, what we are saying is that when considering the contribution to the electric current from the valence band, we ignore the electrons, and instead we treat it as if conduction took place via positively charged holes.

There are many analogies to explain the phenomenon of electrical conduction by holes, although one must be careful not to carry them too far. We present now one such analogy given by Rudden and Wilson.[4]

Consider the seats in a theater, and let us assume that the orchestra section is separated by a gap from the mezzanine. Moreover, let us assume that the orchestra is completely empty and that each seat in the mezzanine

[4]M. N. Rudden and J. Wilson, Elements of Solid State Physics (New York: John Wiley & Sons, 1980).

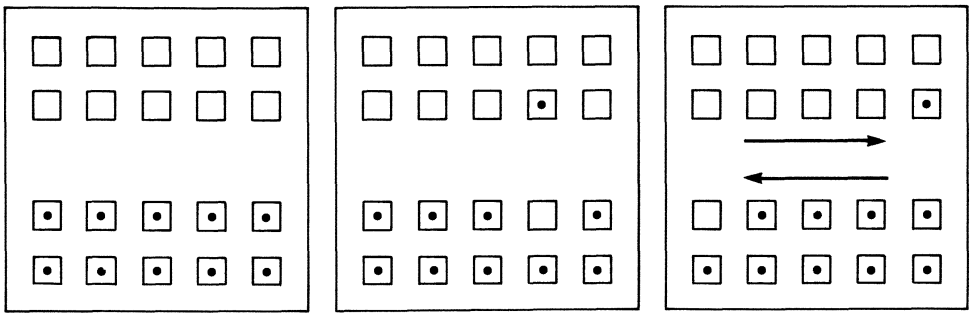


FIGURE 24-33
Theater analogy of the behavior of electrons in the conduction band and holes in the valence band of a semiconductor. (From M.N. Rudden and J. Wilson, *Elements of Solid State Physics*. Copyright © 1980 by John Wiley & Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.)

is occupied. No motion of people from one seat to another can take place as long as everybody stays in the mezzanine. However, suppose that one person moves to the orchestra section, and thus leaves an empty seat behind. This seat can then be occupied by the former neighbor, who having moved, leaves behind another empty seat (see Fig. 24-33). A third person moves now to occupy this seat and so on. The motion could be studied from two viewpoints: the individual motion to the right of each person could be considered (a many-body problem) or the motion of the vacant seat to the left could be investigated (a single-body problem). Clearly, the latter is more desirable.

When we have a completely filled band, its contribution to the current in the presence of an electric field is 0, or

$$J_{\text{full}} = 0$$

Let us now remove electron i from the top of the valence band (remember that it has a negative effective mass); the contribution to the current density J from the remaining electrons will be

$$J_{\text{remaining}} = J_{\text{full}} - J_i$$

where J_i is the contribution from the i electron. But $J_{\text{full}} = 0$, and therefore

$$J_{\text{remaining}} = -J_i$$

But $J_i = -|e|v_i$, where v_i is the drift velocity that the i electron would acquire from $\mathscr{E}$ and $|e|$ is the magnitude of the charge of the electron. We know from our earlier discussion of conduction (see Eq. 23.12) that $v_i = a_i\tau_i$, where a_i is the acceleration of the i electron and τ_i its relaxation time (time between collisions). Therefore

$$J_{\text{remaining}} = |e|\, a_i\tau_i \tag{24.10}$$

Thus, the contribution to the electric current density resulting from the electrons that remain after electron i is removed is equivalent to that of a single charged particle. If we had removed two electrons, the current density would be equivalent to that of two charged particles. We cannot say positive charged particles yet because a_i could be negative (that is, in the direction opposite to $\mathscr{E}$). From Newton's law, the acceleration of the i electron is

$$a_i = \frac{-|e|\mathscr{E}}{m_i^*}$$

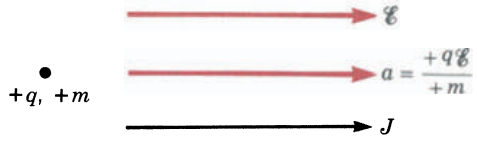


FIGURE 24-34
Response of a positive charge q of positive mass m to an electric field $\mathscr{E}$ resulting in a current density J.

But m_i^* is negative because the i electron is from the top of the valence band, and therefore

$$a_i = \frac{|e|\mathscr{E}}{|m_i^*|} \tag{24.11}$$

The acceleration a_i is in the same direction as the electric field $\mathscr{E}$. We can therefore say that the contribution to the current from the electrons that remain in the valence band when a certain number have been removed by thermal excitation or some other process is equivalent to that of the same number of positively charged holes with positive mass. The expressions in Eqs. 24.10 and 24.11 are what one would expect for a real positively charged particle with positive mass, as illustrated schematically in Fig. 24-34. Therefore, when calculating the current density J resulting from the valence electrons, we ignore them and instead consider only the few positive holes. The total current will be the sum of the contribution of the holes moving in the direction of the field and the electrons opposite to the field direction. A measurement of current with an ammeter cannot distinguish between the two charge carriers.

The reality of holes can be experimentally verified by means of the Hall effect. As we showed in Chapter 16, Section 16.7, the Hall voltage V_H will yield information about the carrier concentration and the sign of their charge. This experimental information is usually given in tables of the Hall coefficient $R_H = 1/Nq$, where N is the number of conduction carriers per unit volume and q is the charge of the carriers. In many solids, both electrons and holes participate in the conduction process, and thus the sign of R_H is determined by the predominant type. When listing R_H, the convention is that it is negative if the predominant type of carrier is the electron, and positive if it is the hole. Regardless of the complexities involved in the calculation of R_H when both types of carriers are present, it should be clear that R_H could never be positive if holes were not a reality. Some examples of the value of the Hall coefficient are listed in Table 24-2.

TABLE 24-2

Solid	R_H (m^3/C)
Lithium	-17×10^{-11}
Sodium	-25×10^{-11}
Beryllium	$+24 \times 10^{-11}$
Zinc	$+3 \times 10^{-11}$
Cadmium	$+6 \times 10^{-11}$

It can be seen in this table that some well-known metals conduct electricity with holes as the predominant charge carrier.

We conclude this chapter with a comment about the behavior of holes. Consider the situation depicted in Fig. 24-32*b*, where a few electrons have been excited into the conduction band of the semiconductor leaving behind an equal number of holes at the top of the valence band. As we indicated earlier in this section, if additional energy is provided to the charge carriers of the semiconductor, such as by the application of an external electric field, the electrons in the valence band will move up into the empty states (the holes). When this happens, the holes move down from their initial position in the valence band. Because we have decided to ignore the electrons in the valence band and instead consider only the holes, we conclude that when energy is provided to the carriers in the valence band, the holes move down, or putting it differently, the energy of the holes *increases downward* from the top of the valence band. This fact will be used in Section 25.2a to calculate the hole concentration in the valence band of a semiconductor.

PROBLEMS

24.1 In the Kronig-Penney model, the relation between the energy E and the momentum p of the electrons is given by the relation

$$P\frac{\sin \gamma d}{\gamma d} + \cos \gamma d = \cos kd$$

where

$$P = \frac{mE_{p0}bd}{\hbar^2}, \quad \gamma = \left(\frac{2mE}{\hbar^2}\right)^{1/2}, \quad \text{and} \quad k = \frac{p}{\hbar}$$

Show that if $E_{p0} = 0$, the energy spectrum becomes continuous and it is that of the free particle,

$$E = \frac{\hbar^2 k^2}{2m} = \frac{p^2}{2m}$$

Use physical arguments to justify this result.

24.2 Sketch three different wavefunctions that are degenerate (that is, correspond to the same energy) with the one of the third level of Fig. 24-21.

24.3 The solution of the Kronig Penney model shows that the width of the allowed bands increases as the energy increases (see Fig. 24-9). Use the arguments presented in Section 24-4 concerning the overlap of the atomic wavefunctions to explain this result.

24.4 Use the arguments presented in Section 24-5 to predict the band structure of (a) Li, and (b) Al. (c) Given the results predicted by these simplified arguments, how do you explain that aluminum is trivalent?

24.5 Use the concept of band theory to explain the following observations: Most insulators are transparent to visible light, semiconductors are transparent to infrared light but opaque to visible light, all metals are opaque to light of all wavelengths.

24.6 The energy gaps of some alkali halides are KCl = 7.6 eV, KBr = 6.3 eV, KI = 5.6 eV. Which of these are transparent to visible light? At what wavelength does each become opaque?

24.7 The experimentally determined value of the Fermi energy in Na is 2.50 eV. In Chapter 23 we showed that

$$E_F = \frac{\hbar^2}{2m}[3N\pi^2]^{2/3}$$

Use the measured value of E_F to calculate the effective mass of the electrons. Na is monovalent. Its atomic weight and density are 22.99 g/mole and 0.97 g/cm^3, respectively.

(*Answer*: 1.26 m, where m is the free electron mass.)

24.8 The density of aluminum is 2.70 g/cm^3 and its molecular weight is 26.98 g/mole. (a) Calculate the Fermi energy. (b) If the experimental value of E_F is 12 eV, what is the electron effective mass in aluminum? Aluminum is trivalent.

(*Answer*: (a) 11.6 eV, (b) 0.97 m, where m is the free electron mass.)

24.9 The experimentally measured specific heat of the conduction electrons in beryllium is $C_v = 0.54 \times 10^{-4}\, T$ calories/mole-K. (a) What is the Fermi energy for beryllium? (b) What is the electron effective mass in beryllium? The density of beryllium is 1.86 g/cm^3 and its molecular weight is 9.01 g/mole.

(*Answer*: (a) 15.7 eV, (b) 0.92 m, where m is the free electron mass.)

24.10 For a free electron, the relation between its energy and its momentum is given by

$$E = \frac{p^2}{2m}$$

Because $p = \hbar k$, the relation can be written as

$$E = \frac{\hbar^2 k^2}{2m}$$

The dependence of E on k is parabolic. For an electron moving in the periodic potential of the lattice ions, the relation between E and k is more complicated. As we have seen, the solution of the Schrödinger equation for an electron moving in a one-dimensional periodic potential consisting of equally spaced potential wells yields the following relation between E and k

$$P\frac{\sin \gamma d}{\gamma d} + \cos \gamma d = \cos kd$$

where

$$P = \frac{mE_{p0}bd}{\hbar^2}$$

is a constant, b is the separation between the wells, d is the periodicity of the wells (see Fig. 24-4), and $\gamma = [2mE/\hbar^2]^{1/2}$. This equation cannot be solved explicitly for $E(k)$. It can be solved numerically. That is, we pick a value of E, plug it into the equation, and find k. The procedure is repeated for other values of E. This numerical solution yields a rather interesting result. There are ranges of E for which k is a real number. These ranges are separated by ranges of E for which k is imaginary. This happens because the left side of the equation is greater than 1 or smaller than -1. Because the right side is cos kd, and the cosine of a real number varies between 1 and -1, k is imaginary for those values of E for which the left side of the equation is greater than 1 or less than -1. The momentum of a particle cannot be imaginary; therefore, the electron cannot possess those values of E for which k is imaginary. This leads to the existence of allowed and forbidden energy bands in solids.

(a) Write a computer program that will yield the allowed values of E and the corresponding values of k as well as the forbidden values of E (the values of E for which k is imaginary). This can be done by letting E vary between 0 and 200 eV in small increments. It is suggested that in the range 0 to 50 eV, $\Delta E = 0.2$ eV, in the range 50 to 200 eV, $\Delta E = 1$ eV.

(b) Plot E versus k for the first allowed energy band. Is the relation between E and k parabolic?

Data

$$P = \frac{5}{2}\pi$$

$$d = 2\ \text{Å} = 2 \times 10^{-10}\ \text{m}$$

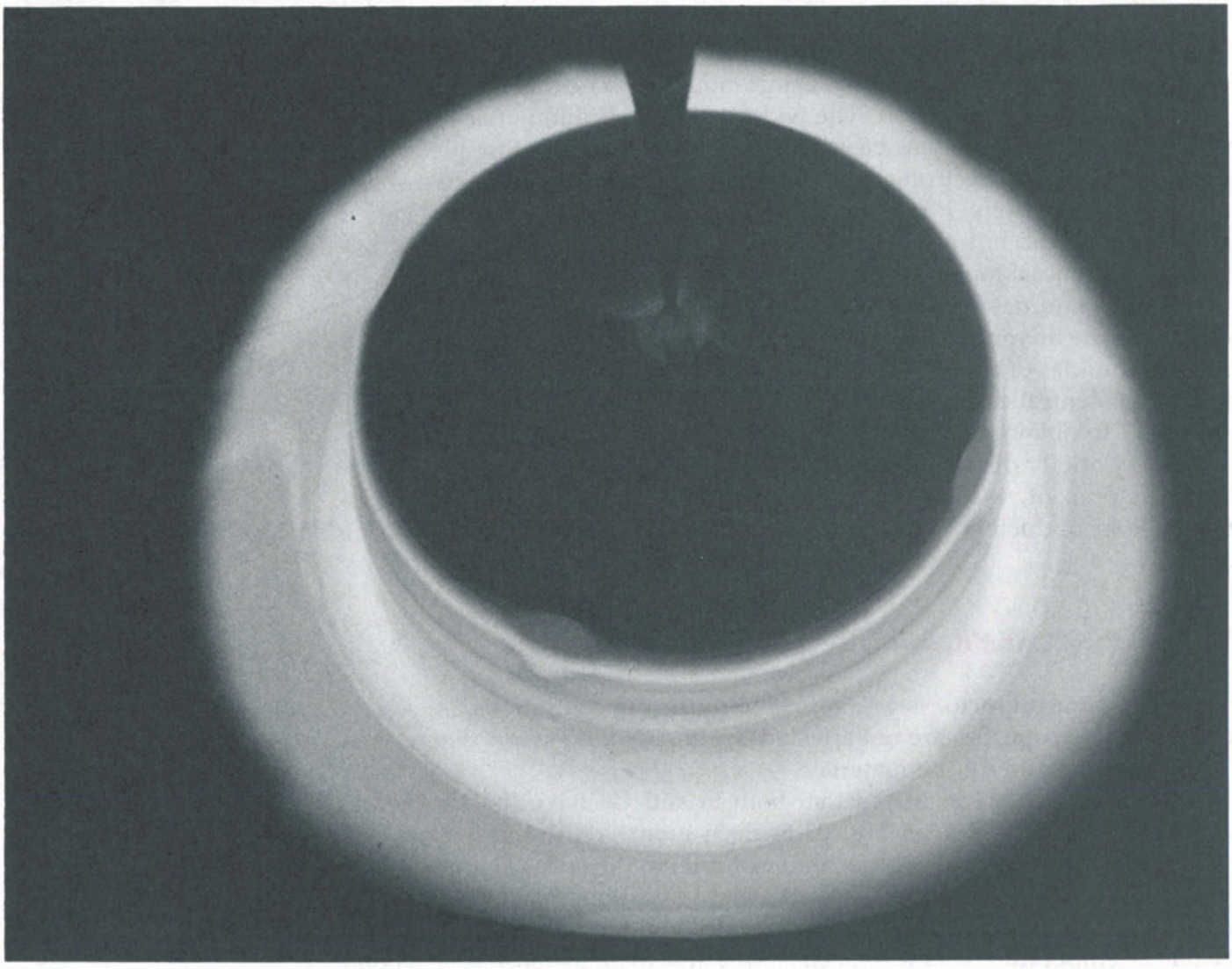
$$m = 9.1 \times 10^{-31}\ \text{kg}$$

$$\hbar = 1.05 \times 10^{-34}\ \text{J-sec}$$

24.11 To see how the width of the allowed bands depends on the interatomic separation, use the pro-

gram of Problem 24.10, with the proper modifications, to find the width of the first allowed energy band for: (a) $d = 1.0$ Å, (b) $d = 1.5$ Å, (c) $d = 2.0$ Å. In all three cases assume that the product $E_{p0}b = 25$ eV-Å. (d) Do the results agree with the qualitative arguments of Problem 24.3?

24.12 The width of the allowed bands depends also on the depth of the potential wells. Verify this with the program of Problem 24.10 by finding the width of the first allowed energy band for: (a) $E_{p0}b = 50$ eV-Å, (b) $E_{p0}b = 75$ eV-Å, (c) $E_{p0}b = 100$ eV-Å. Assume $d = 1$ Å.



CHAPTER 25

Semiconductors

25.1 INTRODUCTION

In the previous chapter we saw how band theory permits the classification of solids into metals, insulators, and semiconductors. The practical use of the first two is as old as civilization. The widespread application of semiconductors, on the other hand, goes back only to the 1950s. The consequence has been a revolution in electronic technology that in turn has had large socioeconomic repercussions. It is interesting to note that the fact which triggered this revolution was not a better understanding of the physics of semiconductors. The general model of a semiconductor was developed principally by A. H. Wilson and dates back to the 1930s. The reason for the technical revolution was the advent of refining techniques that have permitted the preparation of samples of Si and Ge with impurity concentrations of 1 part in 10^{10}. Impurity concentrations as low as a few parts per million in most other solids are difficult to obtain.

We will briefly discuss some of these preparation techniques in Chapter 28. For the present, we will concentrate our attention on the electrical properties of semiconductors, which is why they are such useful and interesting materials.

25.2 INTRINSIC SEMICONDUCTORS

An *intrinsic* semiconductor is one whose impurity concentrations are so low that its electric properties are not affected by them but are solely determined by the band structure of the material.

In Section 22.2 it was stated that both Si and Ge have the tetrahedral diamond structure shown in Fig. 22-6. Each atom is surrounded by four nearest neighbors so that it is the center of a tetrahedron. It shares its four valence electrons equally with its four neighbors, thereby forming strong covalent bonds. The structure is so stable because of the strength of the bonds that at low temperature very few electrons are able to break loose from a bond and migrate in the crystal and participate in conduction processes. In fact, at $T = 0$ K both Ge and Si are insulators.

Having learned about band theory, we can express this fact in another way. A semiconductor is a solid that at $T = 0$ K has a valence band completely filled with electrons, separated by a forbidden energy gap E_g from an empty conduction band. However, the energy gap is small enough that at reasonable temperatures a small number of electrons will be thermally excited across the forbidden gap into the conduction band. Once there, these electrons, as well as the holes they leave behind in the valence band, can be accelerated by an electric field and measurable electrical conduction can take place.

25.2a Electron and Hole Densities

It should be clear that the most important factor determining the electrical conductivity of semiconductors is the number of electrons per unit volume,

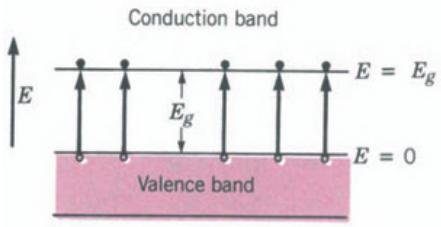


FIGURE 25-1
Schematic of the conduction and valence bands of Si. The top of the valence band is chosen as the zero energy point; the energy of the bottom of the conduction band is E_g.

N_e, that are excited into the conduction band at a given temperature T. How do we determine this number? Let us choose as the zero of energy the top of the valence band as shown in Fig. 25-1. The number of electrons per unit volume in the conduction band, N_e, can be obtained by multiplying the density of states in each incremental energy interval by the probability of occupancy of that interval. The result will be the number of occupied states in that particular energy interval. We then integrate from the bottom to the top of the conduction band.

$$N_e = \int_{\text{bottom of conduction band}}^{\text{top of conduction band}} (\text{density of states}) \times (\text{Fermi function})\, dE \tag{25.1}$$

In Chapter 23 we introduced the concept that the electrons in the conduction band are arranged in energy states according to the Pauli exclusion principle and therefore they obey Fermi-Dirac statistics. We must introduce the modification derived in Chapter 24, that in order to include the interaction of the electrons with the periodic lattice potential, the free electron mass must be replaced by the effective mass m_e^*. We can therefore use for the density of states $g(E)$ the expression that we derived in Chapter 23, Eq. 23.27.

$$g(E) = C_e E^{1/2} \tag{23.27}$$

where

$$C_e = \frac{(2m_e^*)^{3/2}a^3}{2\hbar^3\pi^2} \tag{23.28}$$

Note that Eq. 23.28 contains the term a^3, which is the volume of the solid. We now wish to calculate N_e, defined as the number of electrons per unit volume, and therefore the right side of Eq. 25.1 has to be divided by a^3. This can be done by deleting a^3 from the expression for C_e. In the derivation of Eq. 23.27, E was measured from the bottom of the conduction band. Because we have chosen the top of the valence band as $E = 0$, we must shift the zero point by an amount E_g by subtracting it from E; therefore,

$$g(E) = C_e (E - E_g)^{1/2}$$

We substitute this expression for the density of states and Eq. 23.31 for the Fermi function into Eq. 25.1 and integrate. There are a couple of difficulties, however. Integrals involving the Fermi function are not particularly easy to integrate. Often they cannot be integrated analytically. The second difficulty

is that we need to know the width of the band (the upper limit of the integral). What saves us is the fact that the Fermi level E_F lies near the middle of the energy gap (this will be shown later). In the case of both Si and Ge, $E_g \approx 1$ eV. Therefore, because the bottom of the conduction band is $E_g/2$ above E_F, we estimate that $E - E_F \geq 0.5$ eV. At room temperature ($T = 300$ K) $k_BT = 0.025$ eV. These quantities introduced into the Fermi function permit us to neglect the 1 in the denominator compared with the exponential term, which is much larger than 1. For example, for $E = E_g$, that is, for the energy level at the very bottom of the conduction band,

$$\exp\left(\frac{E - E_F}{k_BT}\right) = \exp\left(\frac{1 \text{ eV} - 0.5 \text{ eV}}{0.025 \text{ eV}}\right) = e^{20} \approx 10^9$$

For higher energy values, the exponential term is even greater; therefore

$$F(E) = \frac{1}{\exp\left(\dfrac{E - E_F}{k_BT}\right) + 1} \approx \exp\left(-\frac{E - E_F}{k_BT}\right) \tag{25.2}$$

The resulting exponential is an easier function to integrate. We see that in the low temperature, high energy approximation Fermi-Dirac statistics can be approximated by Maxwell-Boltzmann statistics (Chapter 9, Supplement 9-1). Moreover, because $F(E)$ decreases exponentially with E, the product $g(E)F(E)$ will be appreciable only near the bottom of the conduction band where E is small. Conversely, the integrand will be negligible above certain energies. If we take $E_g = 1$ eV, $E_F = 0.5$ eV, $T = 300$ K, and we assume a band width of 1 eV (typical value), then

$$F(E_{\text{bottom}}) = \exp\left(-\frac{1 \text{ eV} - 0.5 \text{ eV}}{0.025 \text{ eV}}\right) = \exp(-20) \approx 10^{-9}$$

and

$$F(E_{\text{top}}) = \exp\left(-\frac{2 \text{ eV} - 0.5 \text{ eV}}{0.025 \text{ eV}}\right) = \exp(-60) \approx 10^{-26}$$

Because $F(E)$ at the top of the band is so small, we do not have to know the exact width of the band to integrate Eq. 25.1; we simply use ∞ as the upper limit of integration. Eq. 25.1 is then written as

$$N_e = C_e \int_{E_g}^{\infty} (E - E_g)^{1/2} \exp\left(-\frac{E - E_F}{k_BT}\right) dE \tag{25.1'}$$

This can be integrated readily by using a couple of simple tricks. Let us multiply the right side of Eq. 25-1′ by a form of unity that leaves the equality unchanged, namely,

$$\frac{(k_BT)^{3/2}}{(k_BT)^{3/2}} \exp\left(-\frac{E_g}{k_BT}\right) \exp\left(\frac{E_g}{k_BT}\right) = 1$$

Eq. 25.1′ can then be written as

$$N_e = C_e\,(k_BT)^{3/2} \exp\left(\frac{E_F - E_g}{k_BT}\right) \int_{E_g}^{\infty} \left(\frac{E - E_g}{k_BT}\right)^{1/2} \exp\left(-\frac{E - E_g}{k_BT}\right) \left(\frac{dE}{k_BT}\right)$$

If we let

$$\frac{E - E_g}{k_BT} = x$$

then

$$N_e = C_e\,(k_BT)^{3/2} \exp\left(\frac{E_F - E_g}{k_BT}\right) \int_0^{\infty} x^{1/2}\, e^{-x}\, dx$$

Notice that when $E = E_g$, $x = 0$. The integral can be found in standard tables and is equal to $\sqrt{\pi}/2$. Therefore, the number of electrons per unit volume in the conduction band is

$$N_e = \frac{1}{4}\left(\frac{2m_e^*}{\hbar^2\pi}\right)^{3/2} (k_BT)^{3/2} \exp\left(\frac{E_F - E_g}{k_BT}\right) \tag{25.3}$$

$$N_e = N_c \exp\left(\frac{E_F - E_g}{k_BT}\right)$$

where

$$N_c = \frac{1}{4}\left(\frac{2m_e^*\, k_BT}{\hbar^2\pi}\right)^{3/2}$$

Example 25-1

The energy gap E_g in silicon is 1.1 eV. The average electron effective mass is 0.31 m, where m is the free electron mass. Calculate the electron concentration in the conduction band of silicon at room temperature, $T = 300$ K. Assume $E_F = E_g/2$.

Solution From Eq. 25.3,

$$N_e = \frac{1}{4}\left[\frac{2 \times 0.31 \times 9.1 \times 10^{-31}\ \text{kg} \times 1.38 \times 10^{-23}\ \text{J/K} \times 300\ \text{K}}{(1.05 \times 10^{-34}\ \text{J-sec})^2\,\pi}\right]^{3/2}$$

$$\times \exp\left(\frac{E_F - E_g}{k_BT}\right)$$

$$N_e = 4.36 \times 10^{24}\ \text{m}^{-3} \times \exp\left(-\frac{0.55\ \text{eV} \times 1.6 \times 10^{-19}\ \text{J/eV}}{1.38 \times 10^{-23}\ \text{J/K} \times 300\ \text{K}}\right)$$

$$N_e = 2.6 \times 10^{15}\ \text{per m}^3$$

A typical metal has $\approx 10^{28}$ free electrons per m^3.

In Chapter 24, Section 24.5, we indicated that the magnitude of E_g determines whether a solid is an insulator or a semiconductor. Equation 25.3 shows that, at a given temperature, the number of electrons in the conduction band decreases exponentially with increasing E_g. We can readily see why diamond ($E_g \approx 6$ eV) is an insulator whereas Si ($E_g \approx 1$ eV) is a semiconductor.

From Eq. 25.3, the ratio of N_e in diamond (C) to that in silicon (Si) is (remember that $E_F \approx E_g/2$)

$$\frac{N_e(\text{C})}{N_e(\text{Si})} = \frac{\frac{1}{4}\left(\frac{2m_e^*(\text{C})}{\hbar^2\pi}\right)^{3/2}(k_BT)^{3/2}\exp\left(-\frac{E_g(\text{C})}{2k_BT}\right)}{\frac{1}{4}\left(\frac{2m_e^*(\text{Si})}{\hbar^2\pi}\right)^{3/2}(k_BT)^{3/2}\exp\left(-\frac{E_g(\text{Si})}{2k_BT}\right)}$$

where $m_e^*(\text{C})$ and $m_e^*(\text{Si})$ are the electron effective masses in diamond and silicon respectively, $E_g(\text{C})$ is the energy gap for diamond and $E_g(\text{Si})$ the energy gap for Si. If we neglect differences in the effective masses as being much smaller than the difference in the exponential terms, the ratio becomes

$$\frac{N_e(\text{C})}{N_e(\text{Si})} = \frac{\exp\left(-\frac{6\text{ eV}}{2\,k_BT}\right)}{\exp\left(-\frac{1\text{ eV}}{2\,k_BT}\right)} = \exp\left(-\frac{2.5\text{ eV}}{k_BT}\right)$$

At room temperature $k_BT = 0.025$ eV; therefore

$$\frac{N_e(\text{C})}{N_e(\text{Si})} = \exp\left(-\frac{2.5\text{ eV}}{0.025\text{ eV}}\right) = \exp(-100) \approx 10^{-44}$$

The density of conduction electrons in diamond is 44 orders of magnitude smaller than in Si. In Example 25-1 we found that at room temperature the electron density in Si is $N_e \approx 10^{15}$ per m^3. We conclude that at room temperature there are essentially no electrons in the conduction band of pure diamond, and therefore it is an insulator.

Let us now calculate the density of holes. In order to calculate N_e from Eq. 25.3 we needed to know E_F. We assumed, without proof, that for an intrinsic semiconductor $E_F = E_g/2$. We can obtain an expression for N_e that is independent of E_F and at the same time find E_F by calculating N_h (the number of holes per unit volume in the valence band) in a similar manner to the calculation for N_e. We then set $N_e = N_h$ and we will be able to obtain E_F.

The probability of having a hole at some energy E in the valence band is the same as the probability of the absence of an electron at that same energy; thus if F_h is the hole probability,

$$F_h = 1 - F_e = 1 - \frac{1}{\exp\left[\frac{E - E_F}{k_BT}\right] + 1}$$

Simplifying, we obtain

$$F_h(E) = \frac{\exp\left[\frac{E - E_F}{k_BT}\right]}{\exp\left[\frac{E - E_F}{k_BT}\right] + 1}$$

and, on dividing the numerator and denominator by the numerator and eliminating the resulting negative exponential by reversing the order of E and E_F, we have

$$F_h(E) = \frac{1}{\exp\left[\dfrac{E_F - E}{k_B T}\right] + 1}$$

Notice that because E corresponds to energies in the valence band it is negative. Because $E_F = +E_g/2$, the argument of the exponential is always positive. Moreover, because $E_F - E >> k_B T$, we may neglect the 1 in the denominator because it is much smaller than the exponential term. We may thus approximate F_h by the Boltzmann distribution function as we did for the electrons, that is,

$$F_h(E) \approx \exp\left[\frac{E - E_F}{k_B T}\right] \tag{25.4}$$

We know that the holes behave as positive charged particles with energies that increase *downward* from the top of the valence band just as the electron energies increase upward from the bottom of the conduction band. We can, by analogy to the electron case, write an expression for the density of hole states in the valence band. Note from Fig. 25-1 that the $E = 0$ position is at the top of the valence band and therefore $g(E)$ will not be a function of $(E - E_g)^{1/2}$, as was the case for the electronic states in the conduction band. Instead it will be a function of $(-E)^{1/2}$ (the minus sign is included because E is negative in the valence band and $g(E)$ must be real).

$$g_h(E) = C_h\,(-E)^{1/2} \quad \text{where } C_h = \frac{(2m_h^*)^{3/2}}{2\hbar^3\pi^2} \tag{25.5}$$

Note that since we are calculating the number of holes per unit volume, we have deleted a^3 from the expression for C_h. It should be noted that the effective mass of a hole is not expected to be the same as that of an electron. The number of holes per unit volume will be given by the integral

$$N_h = \int_{-\infty}^{0} g_h(E)\, F_h(E)\, dE \tag{25.6}$$

We have used $-\infty$ as the lower limit of the integration instead of the energy of the bottom of the valence band for the same reason as before: The exponential term decreases rapidly as we move away from $E = 0$. Substituting Eq. 25.5 for g_h and Eq. 25.4 for F_h into Eq. 25.6, we obtain

$$N_h = \int_{-\infty}^{0} C_h(-E)^{1/2} \exp\left[\frac{E - E_F}{k_B T}\right] dE$$

$$N_h = C_h \exp\left[-\frac{E_F}{k_B T}\right] (k_B T)^{3/2} \int_{-\infty}^{0} \left(-\frac{E}{k_B T}\right)^{1/2} \exp\left(\frac{E}{k_B T}\right)\left(\frac{dE}{k_B T}\right)$$

Letting $x = -E/k_BT$, we write

$$N_h = C_h \exp\left[-\frac{E_F}{k_BT}\right](k_BT)^{3/2}\int_{\infty}^{0} x^{1/2}\, e^{-x}\,(-dx)$$

The minus sign in the integrand can be eliminated by inverting the limits of integration. When this is done, we get the same integral that we had before, which was equal to $\sqrt{\pi}/2$. Substituting for C_h from Eq. 25.5 and rearranging terms yields

$$N_h = \frac{1}{4}\left(\frac{2m_h^*}{\hbar^2\pi}\right)^{3/2}(k_BT)^{3/2}\exp\left(-\frac{E_F}{k_BT}\right) \qquad (25.7)$$

$N_h = N_v \exp\left(-\frac{E_F}{k_BT}\right)$ where $N_v = \frac{1}{4}\left(\frac{2m_h^* k_BT}{\hbar^2\pi}\right)^{3/2}$

The expression for N_h, just as that for N_e, depends on E_F, which has not been determined. However, the product N_eN_h is independent of the Fermi energy.

$$N_eN_h = \left(\frac{1}{4}\right)\left(\frac{1}{4}\right)\left(\frac{2m_e^*}{\hbar^2\pi}\right)^{3/2}\left(\frac{2m_h^*}{\hbar^2\pi}\right)^{3/2}(k_BT)^3 \times \exp\left(\frac{E_F - E_g}{k_BT}\right)\exp\left(-\frac{E_F}{k_BT}\right)$$

$$= \frac{1}{16}\left(\frac{2}{\hbar^2\pi}\right)^3(k_BT)^3(m_e^*\, m_h^*)^{3/2}\exp\left(-\frac{E_g}{k_BT}\right)$$

$$N_eN_h = \frac{1}{2}\left(\frac{k_BT}{\hbar^2\pi}\right)^3(m_e^*m_h^*)^{3/2}\exp\left[-\frac{E_g}{k_BT}\right]$$

Note that this product does not depend on the Fermi energy. Because we are talking about intrinsic semiconductors, the only source of electrons for the conduction band is the valence band. Consequently, $N_e = N_h = (N_eN_h)^{1/2}$, and we may write that

$$N_e = N_h = \left(\frac{1}{2}\right)^{1/2}\left(\frac{k_BT}{\hbar^2\pi}\right)^{3/2}(m_e^*m_h^*)^{3/4}\exp\left(-\frac{E_g}{2k_BT}\right) \qquad (25.8)$$

For an intrinsic semiconductor $N_e = N_h$

This gives us N_e and N_h in terms of some universal constants, the temperature and three intrinsic parameters of the semiconductor: m_e^*, m_h^*, and E_g.

25.2b The Fermi Level

We are now ready to locate the Fermi level of an intrinsic semiconductor. As we will see in the next chapter, the position of E_F is crucial to the understanding of semiconductor devices.

We have found two different expressions for N_e, Eq. 25.3 and 25.8. If we equate them,

$$\left(\frac{1}{2}\right)^{1/2}\left(\frac{k_BT}{\hbar^2\pi}\right)^{3/2}(m_h^*m_e^*)^{3/4}\exp\left(-\frac{E_g}{2k_BT}\right)$$

$$= \frac{1}{4}\left(\frac{2m_e^*}{\hbar^2\pi}\right)^{3/2}(k_BT)^{3/2}\exp\left(\frac{E_F - E_g}{k_BT}\right)$$

Upon simplification, we get

$$(m_h^* m_e^*)^{3/4} \exp\left(-\frac{E_g}{2k_B T}\right) = (m_e^*)^{3/2} \exp\left(\frac{E_F - E_g}{k_B T}\right)$$

which can be rewritten as

$$\exp \frac{E_F}{k_B T} = \left(\frac{m_h^* m_e^*}{m_e^{*2}}\right)^{3/4} \exp\left(\frac{E_g}{2k_B T}\right)$$

Taking the natural log of both sides, we obtain

$$E_F = \frac{E_g}{2} + \frac{3}{4} k_B T \ln \frac{m_h^*}{m_e^*} \tag{25.9}$$

$$E_F = \frac{E_g}{2} + \frac{3}{4} k_B T \ln \frac{m_h^*}{m_e^*}$$

It is seen in this equation that at $T = 0$ K, $E_F = E_g/2$, which was our original assumption. For $T > 0$ K, E_F is still equal to $E_g/2$ if $m_e^* = m_h^*$, because $\ln 1 = 0$. But even if $m_e^* \neq m_h^*$, $E_F \approx E_g/2$ at ordinary temperatures unless the two masses differ greatly: This is because $\ln x$ is a slowly varying function of x. In Si and Ge, $m_e^* \approx m_h^*$; thus $E_F \approx E_g/2$.

The concept of the Fermi energy was first introduced in connection with electrons in a three-dimensional potential well; we defined E_F as the demarcation line between filled and empty states. This could be done at $T = 0$ K. For temperatures greater than 0 K such a separation was not clear cut. E_F had to be redefined. We did this (see Section 23.3e) by means of the Fermi-Dirac distribution, and we saw that E_F was that energy for which the probability of occupation was $\frac{1}{2}$. The fact that in the case of a semiconductor E_F lies in the gap, does not imply that we have electrons in the gap. This would contradict the results of band theory. The probability that an energy level in the gap is occupied can be finite without having electrons in it. We have seen that the number of electrons in an energy interval dE is given by the density of available energy states in that dE multiplied by the probability that those states be occupied (i.e., the Fermi-Dirac function). In the gap the Fermi-Dirac function is finite, but the density of states is zero. There are no available states, and hence no electrons.

We can see that it makes sense to have $E_F = E_g/2$. In a metal at $T = 0$ K all the electrons are in energy states below E_F. As the temperature is increased, electrons can be excited to higher energy states. However, we saw in Chapter 23 that only those electrons near E_F have a reasonable chance of being excited. Physically, we can say that E_F is the reference energy level from which charge carriers are more likely to be excited. With this interpretation in mind, let us turn our attention to semiconductors. When an electron is excited from the top of the valence band to the bottom of the conduction band, it leaves behind a hole; an amount of energy equal to E_g is needed for the transition to take place. We could look at this process as a one-particle process (an electron making the transition) or, having agreed to treat holes as proper charged particles, we could look at it as a two-particle process (a simultaneous excitation of an electron into the conduction band and of a hole into the valence

band). We assume that each particle needs an amount of energy $E_g/2$ to make the transition. The excitation can be viewed as an upward transition of an elecctron from the middle of the energy gap and a downward transition of a hole from the same reference point, remembering that the energy of holes increases downward. This is illustrated schematically in Fig. 25-2.

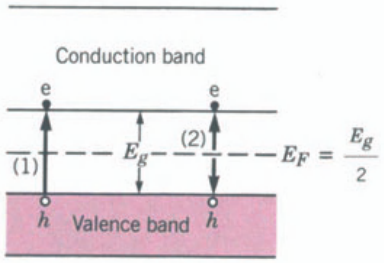


FIGURE 25-2
In an intrinsic semiconductor, the Fermi level is located approximately in the middle of the forbidden energy gap between the valence and the conduction bands. An electron making a transition from the valence band into the conduction band leaves behind a hole(1). The transition can also be looked at as the simultaneous upward transition of an electron into the conduction band and a downward transition of a hole into the valence band, both originating at the Fermi level(2).

25.3 EXTRINSIC OR IMPURITY SEMICONDUCTORS

The electrical properties of a semiconductor can be changed drastically by adding minute amounts of suitable impurities to the pure semiconductor crystal. Thus, the addition of 1 atom of boron to 10^5 atoms of Si raises the conductivity of the sample by 10^3 at room temperature. The most commonly used impurities are elements of group V (P (phosphorous), As (arsenic), Sb (antimony), . . .) and group III (B (boron), Al (aluminum), Ga (gallium), . . .) of the periodic table (see Section 21.7b). The reasons for the choice follow.

25.3a Donor and Acceptor Energy Levels

Because the impurity concentration is usually small, we may suspect that the lattice structure is hardly changed from that of the pure crystal. Our suspicion is confirmed by X-ray studies of the lattice structure. The same studies show that the impurity atoms enter the crystal lattice substitutionally, that is, the impurity atom replaces a lattice atom. Figure 25-3 illustrates a substitution of a phosphorous atom, which has five electrons in its outer shell, into a planar representation of a lattice of silicon atoms, which have four valence electrons each.

Let us now consider what happens when an atom from group V replaces a silicon atom in the lattice structure. We know that each silicon atom is normally bound by four covalent bonds with its four nearest neighbors; that is, it shares its four valence electrons with them. When the group V atom replaces a silicon atom, it will use four of its own electrons for the covalent bonding. There remains, however, an extra electron. This electron (we will show later) will not be very tightly bound to its parent nucleus ($\sim 10^{-2}$ eV), and consequently the impurity atom can be easily ionized and the extra electron will be free to move through the crystal lattice. This loosely bound electron can contribute to conduction processes. We can rephrase this statement in the language of band theory as follows. The extra electron from the impurity atom at $T = 0$ K occupies an energy level E_D that lies $\sim 10^{-2}$ eV below the conduction band; that is, it takes only about 10^{-2} eV to ionize the impurity atom (see Fig. 25-4). As the temperature is raised, this electron can be excited into the conduction band where it can contribute to electrical conduction. Such electrons are said to be donated by the impurity atom, and for

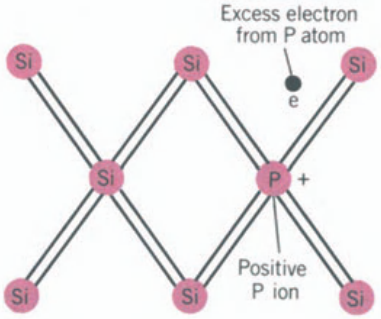


FIGURE 25-3
Planar representation of the silicon lattice structure where a phosphorous atom has taken the place of a silicon atom. After using four of its five valence electrons to complete the four covalent bonds, the phosphorous atom is left with an extra electron that can be easily freed and thus can contribute to electrical conduction.

this reason the energy level E_D is called a *donor* level. Note that when this donor atom is ionized, no hole is created in the valence band.

We should emphasize that these donor electrons provided to the conduction band are in excess of those which may have been excited across the energy gap from the valence band. This implies that we now have more negative charge carriers (electrons in the conduction band) than positive ones (holes in the valence band). For this reason, a semiconductor *doped* (that is, into which impurities have been introduced) with donor impurities is called an *n-type* semiconductor.

Let us backtrack a little and try to calculate the energy with which the extra electron is bound to the impurity atom. Equivalently, we want to see how far below the bottom of the conduction band E_D lies. We can obtain an estimate that is in fairly good agreement with the experimentally measured value by using a rather crude model. According to this model, the excess electron is held to the impurity atom by a net charge of $+e$ (the net charge resulting from the $+Ze$ charge of the nucleus shielded by the remaining $Z-1$ electrons). The problem is reduced to a hydrogen-like atom. There are, however, two important modifications that must be made. In the isolated hydrogen atom, the electron orbits the nucleus in free space. Here the excess electron interacts with the impurity atom while moving in a medium of silicon atoms. This implies two things:

1. The dielectric constant κ (Section 14.7) of the medium must replace the dielectric constant of free space, $\kappa = 1$.
2. We must use the effective electron mass m_e^* rather than the free electron mass.

Keeping these two facts in mind, we can easily evaluate the energy that binds the extra electron. Both the Bohr model and the quantum mechanical model of the hydrogen atom give for the binding energy of the electron (Eqs. 18.16 and 21.9).

$$E = \frac{e^4 m}{8(\kappa\epsilon_0)^2 h^2} = 13.6 \text{ eV}$$

In silicon $\kappa\epsilon_0 = 12\,\epsilon_0$ and $m_e^* = 0.31\,m$. Consequently,

$$E = (13.6 \text{ eV}) \frac{0.31}{(12)^2}$$

$$E = 0.029 \text{ eV}$$

This is the energy necessary to free the excess electron from the impurity atoms, or in the language of band theory, we can say that the impurity energy level E_D lies 0.029 eV below the bottom of the conduction band. For the purpose of comparison, the thermal energy at room temperature is $k_B T = 0.025$ eV. This means that at ordinary temperatures quite a few of these impurities can be ionized; that is, the electrons will be in the conduction band.

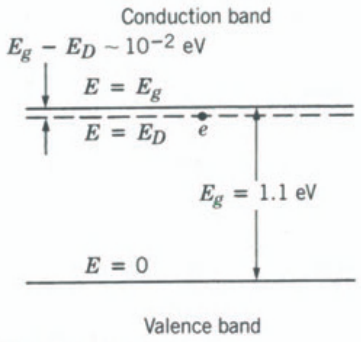


FIGURE 25-4

At $T = 0$ K the extra electron from the phosphorous atom of Fig. 25-3 occupies an energy level E_D (donor level) that lies close to the bottom of the conduction band. This electron can easily jump into the conduction band when provided with thermal energy.

The experimentally measured values of E_D agree quite well with the value found here.

When Si is doped with impurity atoms from group III, it is again found that the impurity atom takes the place of a silicon atom in the crystal lattice. These impurities, being trivalent, cannot complete the tetravalent bond scheme. A vacancy (a hole) is created. At $T = 0$ K this vacancy remains localized; that is, it stays with the impurity atom. However, as T is raised, electrons from adjacent Si atoms may jump into the vacancy of the impurity atom. The vacancy can now migrate through the crystal. Just as before, we can rephrase the situation in band theory terminology: An energy level E_A lies above the valence band. At $T = 0$ K this level is empty, and the valence band is full. As thermal energy becomes available, electrons from the valence band can jump into the empty impurity levels, thus creating mobile holes in the valence band. Calculations similar to the one for pentavalent impurities, as well as experimental results, show that E_A lies relatively close to the top of the valence band (between 0.04 and 0.1 eV in Si and $\approx$ 0.01 eV in Ge) (see Fig. 25-5). Because these impurity atoms may accept electrons, they are called *acceptor* impurities, and E_A is called an *acceptor level*. The holes created by the presence of the acceptor impurities are in addition to those created by thermal excitations across the energy gap. Thus a semiconductor doped with acceptor impurities has more positive charge carriers (holes) than negative ones (electrons). These impurity semiconductors are called *p-type* semiconductors.

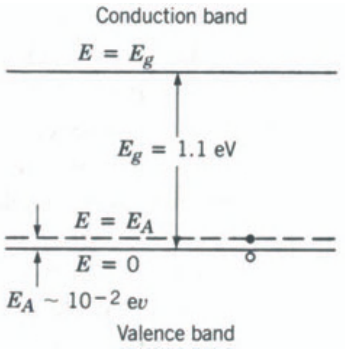


FIGURE 25-5
The introduction of trivalent atoms in the silicon lattice structure of Fig. 25-3 gives rise to an empty energy level E_A (acceptor level) near the top of the valence band. With little thermal energy, electrons from the valence band can jump into this level, thus creating holes in the valence band.

25.3b Carrier Density and Fermi Level in Impurity Semiconductors

We have already indicated in the preceding section that the presence of impurity states will significantly affect the number of electrons in the conduction band or holes in the valence band. They will also very dramatically affect the position of the Fermi level. For simplicity, we will discuss the case of a semiconductor with one type of impurity.

Let us consider an n-type semiconductor and try to answer the same question as before: What is the concentration of electrons in the conduction band N_e and the concentration of holes in the valence band N_h at a given temperature?

When we derived the expressions for N_e and N_h for the intrinsic semiconductor, Eqs. 25.3 and 25.7, we simply multiplied the density of states $g(E)dE$ in an incremental energy interval by the probability of occupancy in that interval, that is, by the Fermi-Dirac function $F(E)$. Neither $g(E)$ nor $F(E)$ are affected by the introduction of impurities; therefore, the expression for N_e and N_h for doped semiconductors will still be given by Eqs. 25.3 and 25.7. However, although the expressions for N_e and N_h are the same as those for the intrinsic semiconductor, their numerical value is not the same. The reason is that the position of the Fermi level E_F in the doped semiconductor is not

the same as in the intrinsic semiconductor. The way we obtained E_F in the intrinsic case was by equating N_e and N_h. This is fine if the only source of electrons for the conduction band is the valence band. However, this is not true for impurity semiconductors. In an n-type semiconductor electrons can be excited into the conduction band from the donor levels as well as from the valence band. How do we determine the Fermi level in this case? The condition we use is that the crystal must be electrically neutral. More explicitly, the number of electrons in the conduction band N_e must be equal to the number of holes in the valence band plus the number of donor impurity atoms per unit volume that have been ionized, N_D^+. This condition is written as

For an n-type semiconductor

$N_e = N_h + N_D^+$

$$N_e = N_h + N_D^+ \tag{25.10}$$

Eq. 25.10 is often referred to as the *neutrality equation*. Our first task is to find N_D^+. This can be done in the following way. Suppose there are N_D donor impurity atoms per unit volume. There will be N_D electrons in the energy level E_D that may make the transition to the conduction band. If we multiply N_D by the probability that an electron not be at E_D, we should have the number of ionized donor impurities N_D^+.

$$N_D^+ = N_D\,(1 - F(E_D)) \tag{25.11}$$

If we substitute Eq. 25.3 for N_e, Eq. 25.7 for N_h, and Eq. 25.11 for N_D^+ in Eq. 25.10, we get

$$N_c \exp\left[-\frac{E_g - E_F}{k_B T}\right] = N_v \exp\left[-\frac{E_F}{k_B T}\right] + N_D\left[1 - \frac{1}{\exp\left(\dfrac{E_D - E_F}{k_B T}\right) + 1}\right] \tag{25.12}$$

where

$$N_c = \frac{1}{4}\left(\frac{2m_e^* k_B T}{\hbar^2 \pi}\right)^{3/2}$$

and (25.13)

$$N_v = \frac{1}{4}\left(\frac{2m_h^* k_B T}{\hbar^2 \pi}\right)^{3/2}$$

Eq. 25.12 must be solved to obtain E_F. If we could solve it analytically, we would obtain E_F in terms of T and the properties of the semiconductor, that is, E_g and N_D. Unfortunately, it cannot be solved analytically, so it is solved numerically. The results from such calculations are shown in Fig. 25-6. The figure shows the variations in E_F with temperature for different impurity

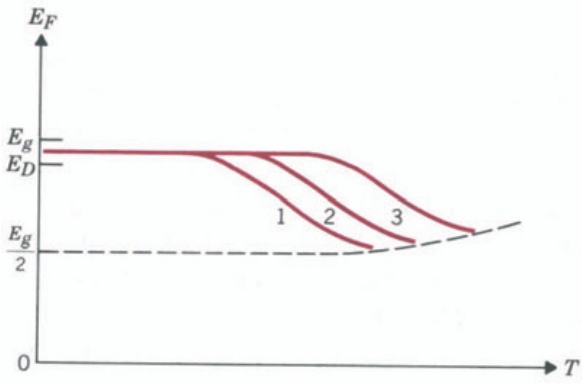


FIGURE 25-6
The Fermi level in an n-type semiconductor as a function of temperature (solid lines). At low temperatures, E_F lies halfway between the donor level E_D and the bottom of the conduction band E_g. At high temperatures, E_F approaches the value of the intrinsic semiconductor (dashed line). The three lines labeled 1, 2, and 3 correspond to increasing values of the impurity concentration.

concentrations. The curves 1, 2, 3 correspond to increasing values of donor impurities N_D. The dashed line represents the value of E_F for the intrinsic semiconductor with the assumption that $m_h^* > m_e^*$. As indicated in Section 25.2b, Eq. 25.9, for the intrinsic semiconductor, at low temperatures $E_F \approx E_g/2$. As it is raised to high temperatures, the second term of Eq. 25.9 becomes significant and E_F increases (because ln m_h^*/m_e^* is positive) from its low temperature value. This is the reason for the upward curvature of the dashed line of Fig. 25-6. The general features of the results are:

1. At low temperatures the Fermi level lies halfway between E_D and the bottom of the conduction band; that is,

$$E_F = \frac{E_g + E_D}{2} \tag{25.14}$$

   As the temperature increases, E_F approaches the value for the intrinsic semiconductor. Note that "low T" is a relative expression; this statement can hold at room temperature, depending on the type of semiconductor.
2. The transition between these two regimes occurs at higher temperatures as the impurity concentration is increased.

$$E_F = \frac{E_g + E_D}{2}$$

We have found the location of E_F with a mathematical argument. Now that we have the answer, we can understand the physical origin of the result. We know that the position of the donor level E_D from the bottom of the conduction band (the ionization energy of the impurities) is ≈ 0.03 eV for Si. We also know that $E_g = 1.1$ eV. By now, we are also familiar with the fact that the probability of a transition across an energy gap E_{gap} is given by (see Eq. 25.8)

$$P \sim \exp\left[-\frac{E_{gap}}{2k_BT}\right]$$

What this means is that the probability that an electron will jump from the donor level to the conduction band across a small energy gap, $E_g - E_D = 0.03$ eV, is many orders of magnitude greater than the probability of making a transition from the valence band to the conduction band across a large

energy gap $E_g = 1.1$ eV. We may write the ratio of these probabilities as

$$\frac{\text{Probability from the impurity level}}{\text{Probability from the valence band}} = \frac{\exp\left[-\dfrac{E_g - E_D}{2k_BT}\right]}{\exp\left[-\dfrac{E_g}{2k_BT}\right]}$$

$$= \frac{\exp\left[-\dfrac{0.03 \text{ eV}}{2k_BT}\right]}{\exp\left[-\dfrac{1.1 \text{ eV}}{2k_BT}\right]} = \exp\left[\frac{1.07 \text{ eV}}{2k_BT}\right]$$

At room temperature $k_BT = 0.025$ eV; this ratio becomes $\exp(21.4) \approx 10^9$. Even though there are many more electrons in the valence band than in the donor levels ($\sim 10^{28}$ valence electrons/m^3 compared with $\sim 10^{22}$ impurity electrons/m^3, assuming an impurity concentration of one part per million), the fact that the probability of transition from E_D is so much greater allows us to conclude that at room temperature practically all the electrons in the conduction band come from the impurity levels. This means that at low temperatures, the semiconductor behaves as an intrinsic semiconductor with an energy gap $E_g - E_D$. Thus we can expect that just as in the case of the true intrinsic semiconductors, E_F will be in the middle of that gap. This is what the solution of the neutrality equation yielded. At high temperature, all the impurities will be ionized and, because we have many more electrons in the valence band than in the impurity levels, the number of electrons in the conduction band will be much greater than N_D. The contribution of the impurities will thus be negligible, and the material will behave like a real intrinsic semiconductor; we may expect that E_F will be the same as that of the intrinsic semiconductor. This is in fact what happens. Clearly, the larger the impurity concentration, the higher the temperature at which the contribution from the electrons excited from the valence band becomes more significant than that of the impurity electrons.

It can be inferred from our discussion what will happen in the p-type semiconductor. Acceptor levels of energy E_A exist very close to the top of the valence band. At low temperatures transitions from the valence band to these levels will dominate the transitions to the conduction band. The impurity levels accept electrons from the valence band, thereby permitting hole conduction. We may therefore expect E_F to be half way between the top of the valence band and E_A. As the temperature increases, these impurity levels will become filled, and transitions to the conduction band become the most important source of charge carriers. The material then behaves like an intrinsic semiconductor. We can argue that E_F will tend toward the intrinsic semiconductor value. This is in fact what happens. A plot of E_F versus T for a p-type semiconductor is shown in Fig. 25-7. Just as in Fig. 25-6, the curves 1, 2, 3 correspond to increasing impurity concentration. The reason for the upward

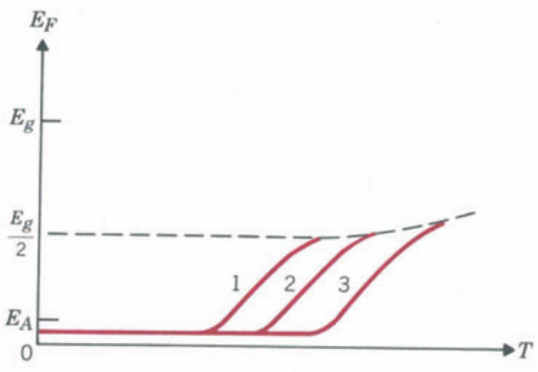


FIGURE 25-7
The Fermi level in a p-type semiconductor as a function of temperature (solid lines). At low temperatures, E_F is halfway between the acceptor level E_A and the top of the valence band. At high temperatures, E_F approaches the value of the intrinsic semiconductor (dashed line). The three lines labeled 1, 2, and 3 correspond to increasing values of the impurity concentration.

curvature of the dashed line is the same as the explanation given in connection with Fig. 25-6.

We conclude this section with one important observation that should be obvious by now. We have indicated that one of the effects of the impurities was to break the equality that existed in the intrinsic semiconductor between electrons and holes. In the n-type semiconductor, owing to the contribution from the donor states, there are more electrons in the conduction band than holes in the valence band. Moreover, we have shown that at low temperatures most of the electrons in the conduction band come from the impurity levels. This in turn means that in the n-type semiconductor electrons are much more abundant than holes (see Example 25-2). For this reason, in an n-type semiconductor, the electrons are called the *majority carriers*, and the holes are named *minority carriers*. Obviously, in the p-type semiconductor the situation is reversed.

Example 25-2

A sample of Si is doped with phosphorous. The donor impurity level lies 0.045 eV below the bottom of the conduction band. At $T = 300$ K, E_F is 0.010 eV above the donor level. Calculate (a) the impurity concentration, (b) the number of ionized impurities, (c) the free electron concentration, and (d) the hole concentration. (For Si, $E_g = 1.100$ eV, $m_e^* = 0.31\ m$, $m_h^* = 0.38\ m$).

Solution

(a) From Eq. 25.12,

$$N_c \exp\left(-\frac{E_g - E_F}{k_B T}\right) = N_v \exp\left(-\frac{E_F}{k_B T}\right)$$

$$+ N_D \left[1 - \frac{1}{\exp\left(\frac{E_D - E_F}{k_B T}\right) + 1}\right] \qquad (25.12)$$

From Eq. 25.13,

$$N_c = \frac{1}{4}\left[\frac{2m_e^* k_B T}{\hbar^2 \pi}\right]^{3/2}$$

$$= \frac{(2 \times 0.31 \times 9.1 \times 10^{-31}\ \text{kg} \times 1.38 \times 10^{-23}\ \text{J/K} \times 300\ \text{K})^{3/2}}{4 \times (1.05^2 \times 10^{-68}\ \text{J}^2\text{-sec}^2 \times \pi)^{3/2}}$$

$$= 4.39 \times 10^{24}\ \text{m}^{-3}$$

and

$$N_v = \frac{1}{4}\left[\frac{2m_h^* k_B T}{\hbar^2 \pi}\right]^{3/2}$$

$$= \frac{(2 \times 0.38 \times 9.1 \times 10^{-31}\ \text{kg} \times 1.38 \times 10^{-23}\ \text{J/K} \times 300\ \text{K})^{3/2}}{4 \times (1.05^2 \times 10^{-68}\ \text{J}^2\text{-sec}^2 \times \pi)^{3/2}}$$

$$= 5.95 \times 10^{24}\ \text{m}^{-3}$$

Substituting for N_c, N_v, E_F, and E_D into Eq. 25-12 (keep in mind that energies are measured from the top of the valence band), we get

$$4.39 \times 10^{24}\ \text{m}^{-3} \exp\left(-\frac{(1.100 - 1.065)\ \text{eV}}{0.025\ \text{eV}}\right)$$

$$= 5.95 \times 10^{24}\ \text{m}^{-3} \exp\left(-\frac{1.065\ \text{eV}}{0.025\ \text{eV}}\right)$$

$$+ N_D\left[1 - \frac{1}{\exp\left(-\frac{0.010\ \text{eV}}{0.025\ \text{eV}}\right) + 1}\right]$$

$$1.08 \times 10^{24}\ \text{m}^{-3} = 1.88 \times 10^{6}\ \text{m}^{-3} + N_D\,(0.40)$$

$$N_D = 2.7 \times 10^{24}\ \text{m}^{-3}$$

(b) The number of ionized impurities is given by the second term of the right side of Eq. 25.12

$$N_D^+ = N_D\left[1 - \frac{1}{\exp\left(\frac{E_D - E_F}{k_B T}\right) + 1}\right]$$

$$N_D^+ = 2.7 \times 10^{24}\ \text{m}^{-3}\left[1 - \frac{1}{\exp\left(-\frac{0.010\ \text{eV}}{0.025\ \text{eV}}\right) + 1}\right]$$

$$= 1.08 \times 10^{24}\ \text{m}^{-3}$$

(c) The free electron concentration is equal to the left side of Eq. 25.12, which from part (a) of this example is

$$N_e = 1.08 \times 10^{24}\ \mathrm{m}^{-3}$$

(d) The hole density is found in the first term of the right side of Eq. 25-12

$$N_h = 1.88 \times 10^{6}\ \mathrm{m}^{-3}$$

25.4 ELECTRICAL CONDUCTIVITY OF SEMICONDUCTORS

In Chapter 23 on free electron theories we saw that for a conductor, except at very low temperatures, the electrical conductivity σ decreased with increasing temperature as $\sigma \propto T^{-1}$. For an intrinsic semiconductor, the conductivity increases exponentially with increasing temperature,

$$\sigma = \sigma_0 \exp(-\alpha/T)$$

where α is a constant. This is illustrated schematically in Fig. 25-8 where a plot of the natural log of σ versus the inverse of the temperature yields a straight-line relationship.

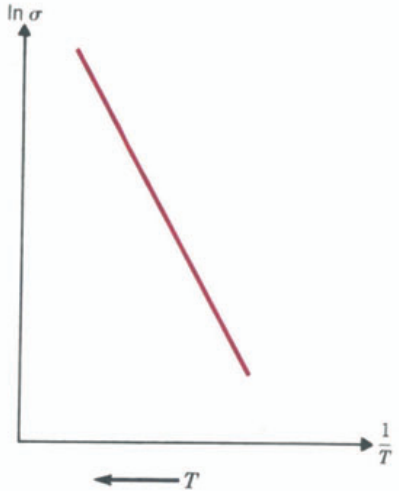


FIGURE 25-8
Experimentally observed temperature dependence of the electrical conductivity of an intrinsic semiconductor.

In the case of a doped semiconductor, the same exponential behavior is observed. However, the constant α has two different values. This is immediately obvious if we look at the graph of Fig. 25-9, which shows the experimentally observed variation of σ with temperature. The graph clearly shows that at low T's and high T's ln σ varies linearly with the inverse of the temperature. The slope is much greater for high T's than for low T's.

We can understand these results if we keep in mind our discussion of the carrier concentration and the position of the Fermi level. In Chapter 23 we found that

$$\sigma = \frac{Nq^2\tau}{m} \tag{23.14}$$

where N was the number of free charge carriers (at that time electrons), q was the charge of the carriers, τ was the time between collisions with ions, and m was the mass of the carriers (at that time the true mass of the electron). Because we now have two types of charge carriers, electrons and holes, which contribute to the electrical conduction, the expression for σ must be modified to

$$\sigma = |e|^2 \left[\frac{N_e\, \tau_e}{m_e^*} + \frac{N_h\, \tau_h}{m_h^*} \right] \tag{25.15}$$

where τ_e and τ_h are the mean times between collisions with ions for electrons and holes, respectively, m_e^* and m_h^* their effective masses, and $|e|$ is the magnitude of the electron charge because conductivity does not distinguish between types of carriers.

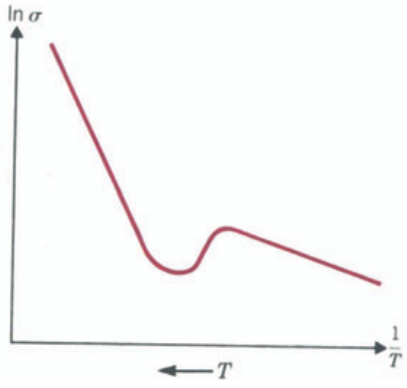


FIGURE 25-9
Experimentally observed temperature dependence of the electrical conductivity of an impurity semiconductor.

Let us first consider the case of intrinsic semiconductors. In this case $N_e = N_h$ and N_e can be factored from Eq. 25.15

$$\sigma = |e|^2 N_e \left[\frac{\tau_e}{m_e^*} + \frac{\tau_h}{m_h^*}\right] \tag{25.16}$$

Unlike the case of a metal where N_e was constant and the temperature dependence of σ was hidden solely in the time between collisions with ions ($\tau \propto T^{-1}$), in semiconductors both N_e and τ vary with T. Calculation of τ for silicon and germanium has shown that $\tau \propto T^{-3/2}$. The carrier concentration N_e, on the other hand, depends exponentially on T, and it is therefore much more sensitive to changes in temperature. If we substitute N_e of Eq. 25.8 in Eq. 25.16, we obtain

$$\sigma = \text{constant } T^{3/2} \exp\left[-\frac{E_g}{2k_BT}\right]\left(\frac{\tau_e}{m_e^*} + \frac{\tau_h}{m_h^*}\right) \tag{25.17}$$

For an intrinsic semiconductor at all T's and for a doped semiconductor at high T's.

Using the $T^{-3/2}$ power law mentioned previously for τ, the multiplicative temperature terms cancel and the conductivity may be written as

$$\sigma = \text{constant} \exp\left[-\frac{E_g}{2k_BT}\right] \tag{25.18}$$

$$\sigma \propto \exp\left(-\frac{E_g}{2k_BT}\right)$$

where the constant is now modified by the sum of the reciprocal effective masses. A plot of the ln σ versus T^{-1} yields a straight line with slope $-E_g/2k_B$. The experiment (see Fig. 25-8) confirms this prediction. Moreover, the value of E_g obtained from conductivity measurements agrees with that found by other means such as optical absorption, which we will discuss later.

Let us now examine impurity semiconductors. We will consider n-type semiconductors. Similar arguments will apply to the p-type. In discussing the position of the Fermi level in n-type semiconductors, we indicated that at high temperatures, when all the impurities have been ionized, their contribution to N_e becomes negligible compared with the contributions from excitations from the valence band. As a result, the crystal behaves like an intrinsic semiconductor. What we have shown about the conductivity of the intrinsic semiconductor holds here too: The ln σ versus T^{-1} will be a straight line with slope $-E_g/2k_B$.

At low temperatures the situation is quite different. There, most of the electrons in the conduction band originate from the impurity levels, and as a result $N_e >> N_h$. We can ignore the contribution of the holes to the conductivity, and Eq. 25.15 can be approximated as

$$\sigma \approx \frac{|e|^2 N_e \tau_e}{m_e^*}$$

If we substitute Eq. 25.3 for N_e and put in the temperature dependence of τ as $T^{-3/2}$, we obtain

$$\sigma = \text{constant} \exp\left[-\frac{(E_g - E_F)}{k_BT}\right]$$

We have shown previously (Fig. 25-6) that at low temperature E_F lies halfway between the donor levels and the bottom of the conduction band; that is, $E_F = (E_g + E_D)/2$, Eq. 25.14. Using this result, we may write the expression for the conductivity as

$$\sigma = \text{constant} \exp\left[-\frac{(E_g - E_D)}{2k_BT}\right] \tag{25.19}$$

Thus, at low temperatures ln σ will vary linearly with T^{-1} and the slope will be $-(E_g - E_D)/2k_B$. We recognize $E_g - E_D$ as the impurity ionization energy that we know is much smaller than the gap energy E_g. Therefore, the slope of the curve at low temperatures should be much smaller than at high T's. The expected behavior of σ is shown in Fig. 25-10. The actual experimental data, Fig. 25-9 shows a dip between the high and the low temperature regimes. At intermediate temperatures, most of the impurities will have been ionized, and consequently there will be no change in their contribution to N_e with variations in T. The temperatures are still too low ($k_BT << E_g$) to have a significant number of electrons excited from the valence band across the energy gap. As a consequence, the temperature dependence of σ will be controlled by the collision time τ. Because τ decreases with increasing T ($\tau \propto T^{-3/2}$), we would expect a decrease in σ as T is raised. This is the explanation for the dip in Fig. 25-9.

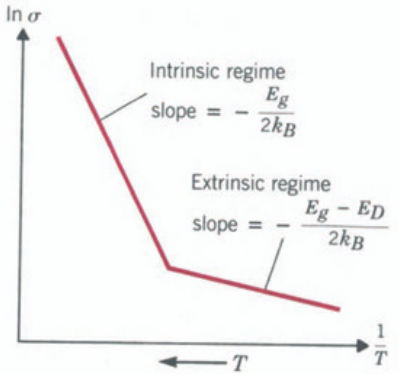


FIGURE 25-10
Theoretical temperature dependence of the electrical conductivity of an impurity semiconductor.

For an n-type semiconductor at low T's

$$\sigma \propto \exp\left(-\frac{E_g - E_D}{k_BT}\right)$$

25.5 PHOTOCONDUCTIVITY

We conclude this chapter with some brief remarks about the behavior of a semiconductor when it is exposed to electromagnetic radiation. Being by now familiar with the concept of photons, one can immediately think of a simple technique to measure the energy gap of a semiconductor: Illuminate the semiconductor with photons of different frequencies and observe the ability of the sample to absorb them.

The experimental arrangement is rather simple. Light from a device in which one can select a given wavelength of light, called a *monochromator*, strikes a thin slab of material under study. The amount of radiation transmitted through the sample is then measured by means of a light meter (Fig. 25-11). The amount absorbed is found by subtracting the intensity of the transmitted beam from that of the incident beam. The experiment is repeated for different wavelengths to find the dependence of the absorption on the energy of the photons. The experimental results for insulators and semiconductors show negligible absorption for long wavelengths (small ν) and then

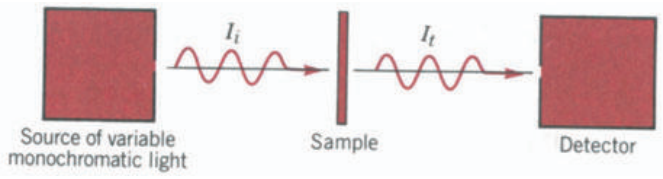


FIGURE 25-11
Schematic of the setup used to measure the absorbtion of electromagnetic radiation by a semiconductor.

a sudden rise, which is called the *absorption edge* (see Fig. 25-12). The wavelength λ_c at which the abrupt change takes place depends on the material under consideration. For example, for Si, $\lambda_c \approx 1.1 \times 10^{-6}$ m (infrared).

The results can be easily explained in terms of the band model of the semiconductor. If the energy of the incident photons $h\nu$ is less than the energy of the gap E_g, the electrons will not be able to be excited from the valence band into the conduction band; the photons will not be absorbed by the electrons, and the radiation will go through the sample. If, however, $h\nu \geq E_g$, electrons will be excited across the gap as a result of the absorption of a photon. Notice that unlike the case of absorption by isolated atoms, where $h\nu$ must be exactly equal to the energy difference between atomic energy levels, in this situation $h\nu$ can be greater than E_g. The reason is simple: There are plenty of empty energy levels in the conduction band. This explanation can be easily checked by equating $h\nu$ to E_g for the wavelength of the absorption edge in silicon.

$$E_g = h\nu = \frac{hc}{\lambda_c} = \frac{(6.63 \times 10^{-34}\ \text{J-sec})\ (3 \times 10^{8}\ \text{m/sec})}{1.1 \times 10^{-6}\ \text{m}}$$

$$= 1.8 \times 10^{-19}\ \text{J} = 1.1\ \text{eV}$$

This agrees well with the value obtained from the temperature dependence of the electrical conductivity.

A useful application of the property of transmission at wavelengths above the absorption edge is the examination of the perfection of the interior of large crystals of silicon. While visible light is blocked, infrared is transmitted and can be observed with a suitable detector.

Because optical absorption creates extra electrons and holes, the electrical conductivity of the semiconductor will be enhanced when illuminated with radiation of the proper frequency. This effect is called *photoconductivity*. It has a large number of practical applications, such as automatic controls for nightlights and light meters in cameras. When a light beam is aimed at the photoconductor, current can be made to flow in a circuit. Moving objects between the light source and the photoconductor will interrupt the current and thereby be detected. These systems are called *photoelectric cells*.

One of the most familiar applications of photoconductivity is the process of *xerography*. A high resistance semiconductor, usually selenium, is applied as a coating to a metal surface. The metal serves as the plate, which is often in a cylindrical form. Ionization of the air near the surface by a high intensity light of suitable wavelength puts a charge on the exposed surface of the selenium coating. Selenium has a resistivity in the dark of the order of 10^{14} ohm-m, so the charge stays on the surface. However, when a portion of the surface is illuminated, the photoconductivity causes the charge at that part to leak through to the metal base plate, leaving that region of the selenium uncharged. Thus the image of a printed page projected on the selenium surface would discharge all parts except those on which the black printing was projected. A dark resin powder is then dusted on to the selenium surface,

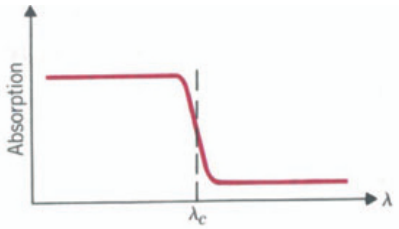


FIGURE 25-12
Experimentally observed absorbtion of electromagnetic radiation by a semiconductor as a function of the wavelength of the radiation. Note the sharp change in absorbtion at λ_c.

and it clings to the charged region, that is, that onto which the printing was projected. This powder is then transferred to a piece of paper and fused by the application of heat.

PROBLEMS

25.1 The band gap in pure germanium is $E_g = 0.67$ eV. (a) Calculate the number of electrons per unit volume in the conduction band at 250 K, 300 K, and at 350 K. (b) Do the same for silicon assuming $E_g = 1.1$ eV. The effective mass of the electrons in germanium is 0.12 m and in silicon 0.31 m, where m is the free electron mass.

25.2 Suppose that the effective mass of holes in a material is four times that of electrons. At what temperature would the Fermi level be shifted by 10% from the middle of the forbidden energy gap? Let $E_g = 1$ eV.

(*Answer*: 557 K.)

25.3 The energy gap in germanium is 0.67 eV. The electron and the hole effective masses are 0.12 m and 0.23 m, respectively, where m is the free electron mass. Calculate (a) the Fermi energy, (b) the electron density, and (c) the hole density, at $T = 300$ K.

(*Answer*: (a) 0.348 eV, (b) 4.1×10^{18} m^{-3}, (c) 4.1×10^{18} m^{-3}.)

25.4 In the crude model used in Section 25.3a to calculate the binding energy of the excess electron to the donor impurity atom, we assumed that the electron moves in a medium of silicon atoms that screen the donor impurity ions; that is, we use the dielectric constant $\kappa = 12$ of silicon. (a) Justify this assumption by calculating the radius of the Bohr orbit of the excess electron (see Eq. 18.14), (b) Estimate the number of silicon atoms in a sphere with that radius. The nearest neighbor separation in silicon is 2.34 Å.

(*Answer*: (a) 20.4 Å, (b) 2700 atoms.)

25.5 (a) Starting with Eq. 25.3 and 25.7, show that the product N_eN_h is a constant; that is, it is not affected by the introduction of impurities in the pure semiconductor. (b) The number of electrons in the conduction band can be greatly increased by the introduction of donor impurities. These two facts imply that the number of holes is greatly reduced when the semiconductor is doped with donor impurities. Explain the physical reason for the reduction in hole concentration.

25.6 Consider a sample of germanium doped with phosphorous. Assume that the excess electron revolves around the impurity ion P^+ in a hydrogen-like orbit. Calculate (a) the ionization energy of the excess electron, (b) the radius of the orbit (see Eq. 18.14). The dielectric constant κ in germanium is 16 and the electron effective mass is 0.12 m.

(*Answer*: (a) 6.5×10^{-3} eV, (b) 70 Å.)

25.7 Silicon is doped with an acceptor type impurity such as boron or gallium. The impurity concentration is N_A atoms per unit volume. Write the neutrality equation, similar to Eq. 25.12, from which the Fermi energy can be determined.

25.8 Germanium ($E_g = 0.67$ eV) is doped with gallium. The acceptor impurity level lies 0.011 eV above the top of the valence band. (a) What is the impurity level concentration if E_F, at room temperature ($T = 300$ K), coincides with the acceptor level? (b) Calculate the fraction of ionized impurities ($m_e^* = 0.12\ m$, $m_h^* = 0.23\ m$).

(*Answer*: (a) 3.6×10^{24} m^{-3}, (b) 50%.)

25.9 Referring to the situation described in Problem 25.8, (a) calculate the concentration of holes and electrons, (b) show that the product N_eN_h is the same as the one found in Problem 25.3.

(*Answer*: (a) $N_h = 1.8 \times 10^{24}$ m^{-3}, $N_e = 9.0 \times 10^{12}$ m^{-3}.)

25.10 (a) Show that the most probable value of the energy for an electron in the conduction band of an intrinsic semiconductor is $\frac{1}{2}k_BT$ above the bottom of the conduction band. (b) What is the average energy

of an electron in the conduction band of an intrinsic semiconductor?

25.11 In Table 25-1 we list the conductivity of pure germanium at different temperatures. (a) Make a plot of ln σ versus $1/T$. (b) Determine E_g for germanium.

TABLE 25-1
Problem 25-11

σ ($\Omega^{-1}m^{-1}$)	T(K)
2	300
13	350
52	400
153	450
362	500

25.12 The conductivity of pure germanium increases by 50% when the temperature is increased from 20°C to 30°C. (a) What is the energy gap E_g between the conduction and the valence bands of germanium? (b) For silicon $E_g = 1.1$ eV, what is the percentage charge in the conductivity for the same temperature change?

(*Answer*: (a) 0.62 eV, (b) 150%.)

25.13 A certain intrinsic semiconductor has a band gap $E_g = 0.2$ eV. Measurement shows that it has a resistivity at room temperature (300 K) of 0.3 Ω-m. What would you predict its resistivity to be at 350 K?

(*Answer*: 0.17 Ω-m.)

25.14 (a) What is the relative number of electrons N_e and holes N_h in a doped semiconductor for which the electrical conductivity σ (see Eq. 25.15) is a minimum at a given temperature? (b) What is the ratio N_e/N_h if the scattering time τ is the same for electrons and holes and $m_e^*/m_h^* = 1.5$? (*Hint*: The product N_eN_h at a given temperature is a constant, see Problem 25.5).

(*Answer*: (a) $N_e/N_h = \tau_h\, m_e^*/\tau_e\, m_h^*$, (b) 1.5.)

25.15 The resistivity of a sample of n-type silicon at 300 K is 9×10^{-3} Ω-m and its Hall coefficient is 3.9×10^{-4} m^3/C. (a) Assuming that electrical conduction is due solely to electrons, what is the free electron concentration? (b) How does the answer to (a) compare with the result of Example 25-1? Is the assumption in part (a) justified? (c) The effective mass of the electrons $m_e^* = 0.31\ m$. What is the electron scattering time τ_e?

(*Answer*: (a) 1.6×10^{22} m^{-3}, (c) 7.6×10^{-14} sec.)

25.16 The energy gap in silicon is 1.1 eV, whereas in diamond it is 6 eV. What conclusion can you draw about the transparency of the two materials to visible light (4000 Å to 7000 Å)?

25.17 A cadmium sulfide ($E_g = 2.4$ eV) photodetector is illuminated with light of wavelength 3000 Å. The intensity of the radiation falling on the detector is 30 W/m^2. The area of the detector is 9 mm^2. (a) Show that electron-hole pairs will be generated by the light. (b) Assuming that each photon produces an electron-hole pair, how many pairs will be generated every second?

25.18 When a photon of energy $E >> E_g$ enters a semiconductor, it can produce several electron-hole pairs; that is, it can excite several electrons from the top of the valence band to the bottom of the conduction band. A germanium crystal is used as a gamma ray (high energy photon) detector. (a) What is the maximum number of electron-hole pairs created by a 1.5 MeV gamma ray? (b) If the resolution of the detector is $\pm 4 \times 10^3$ electron-hole pairs, what is the optimal energy resolution of the detector?

(*Answer*: (a) 2.24×10^6, (b) 2.7×10^3 eV.)

25.19 When a semiconductor is doped with a donor type of impurity, the number of electrons in the conduction band is equal to the number of electrons excited from the valence band into the conduction band (the number of holes) plus the number of impurity atoms that have been ionized. This leads to the equation (See Eq. 25.12)

$$N_c \exp\left[-\frac{E_g - E_F}{k_BT}\right] = N_v \exp\left[-\frac{E_F}{k_BT}\right] + N_D\left[1 - \frac{1}{\exp\left(\frac{E_D - E_F}{k_BT}\right) + 1}\right]$$

where

$$N_c = \frac{1}{4}\left[\frac{2m_e^* k_B T}{\hbar^2 \pi}\right]^{3/2}$$

and

$$N_v = \frac{1}{4}\left[\frac{2m_h^* k_B T}{\hbar^2 \pi}\right]^{3/2}$$

Equation 25.12 is an equation for the Fermi energy E_F (dependent variable) as a function of the temperature T (independent variable). The equation cannot be solved analytically. It can be solved by numerical methods with the aid of a computer. This is done as follows: We choose a value for T, and then we find the value of E_F that makes both sides of the equation equal. Actually, because the problem is solved by numerical methods, it is practically impossible to find the value of E_F that will make both sides of the equation *exactly* equal. We can approximate the problem by requiring that E_F be such, that both sides differ from each other by a small amount, for example, 1%.

Write a computer program that will yield E_F for different T's. Do this for $T = 10\ \text{K} - 2500\ \text{K}$. After finding E_F, substitute it in the expressions for N_e and N_h to obtain the free electron and hole densities.

The following data and hints may be helpful:

N_D is the impurity concentration. Take it to be 10^{24} atoms/m^3.

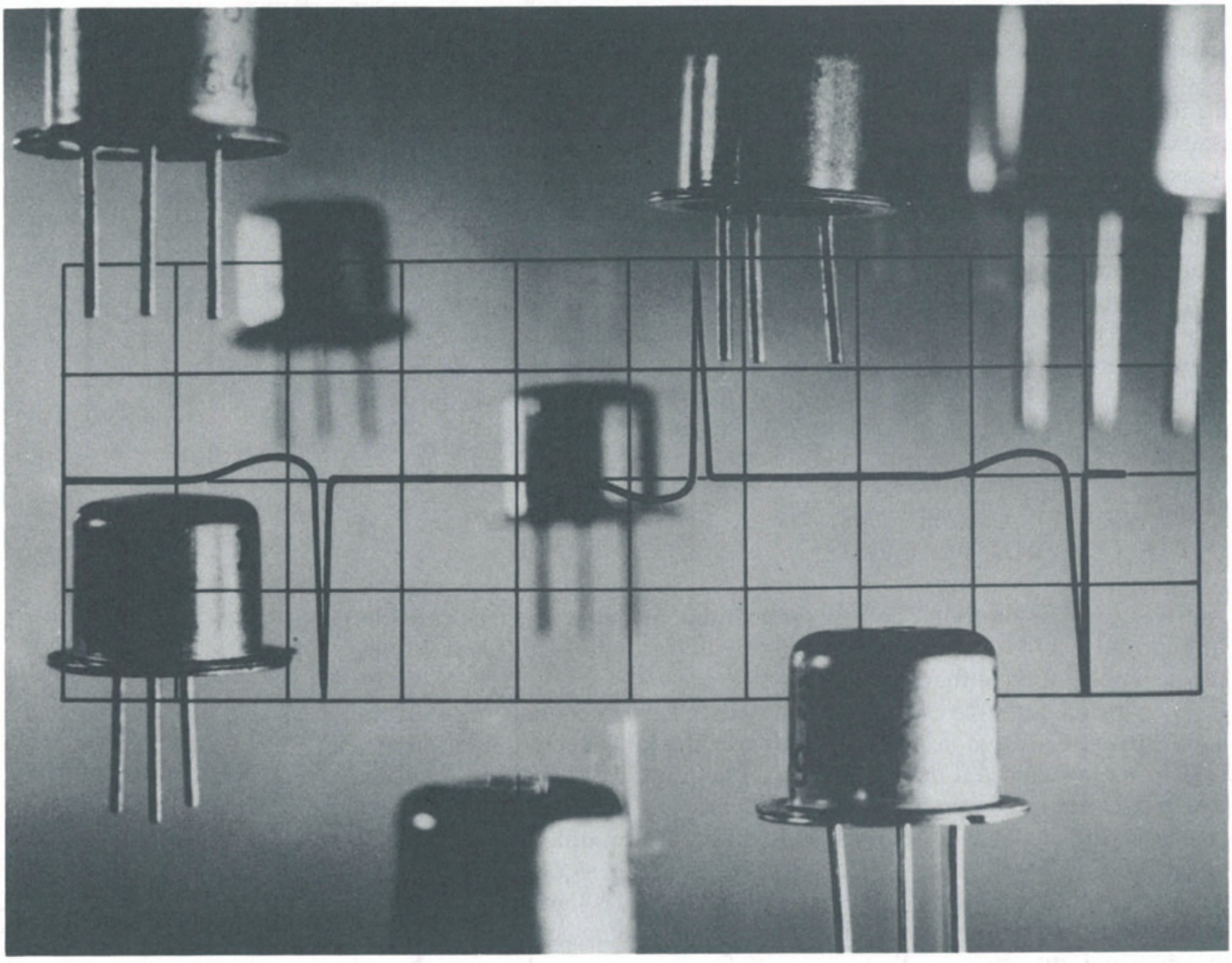
E_g for silicon is 1.10 eV.

E_D is the impurity energy level. Assume it is 0.04 eV below the bottom of the conduction band, that is, $E_g - E_D = 0.04$ eV.

E_F lies between 1.1 eV and 0.55 eV.

$m_e^* = 0.31\,m$; $m_h^* = 0.38\,m$

$k_B = 1.38 \times 10^{-23}\ \text{J/K} = 8.63 \times 10^{-5}\ \text{eV/K}$



CHAPTER 26

Semiconductor Devices

26.1 INTRODUCTION

In Chapter 25 we discussed the electrical properties of semiconductors and how these properties are affected by the introduction of different types of impurities in a pure silicon sample. We showed that the free electron concentration is significantly increased by doping pure silicon with a small amount of phosphorous or other pentavalent element. We call these semiconductors *n*-type semiconductors. Similarly, the hole concentration is greatly enhanced when silicon is doped with trivalent impurities such as boron. This results in the formation of a *p*-type semiconductor.

In this chapter we will consider how these two types of semiconductors are used to create the semiconductor devices—diodes and transistors—that are fundamental to the performance of such functions as rectification, amplification, and switching in electronic circuits. These devices are the basic building blocks of computers and other digital instruments.

26.2 METAL-METAL JUNCTION: THE CONTACT POTENTIAL

Before we consider the operation of semiconductor devices, let us consider what happens when two dissimilar metals are joined together. This will bring out two important facts that we will use later.

Let us recall a couple of facts that we discussed in connection with the photoelectric effect (Section 17.3). We saw that the kinetic energy with which the electrons were ejected from a metal was not uniform. The electrons came out with a continuous spread of energies. The explanation given was that electrons in the metal were bound with different amounts of energy. The smallest binding energy was called the work function ϕ of the metal. The other fact, which is pertinent to our present discussion, is that different metals have different work functions. For example, the photoelectric effect occurs with lower energy photons, that is, lower frequency, in potassium than it does in copper.

We can explain these results rather simply in terms of the quantum mechanical model of electrons in a solid. The electrons occupying the highest energy levels in a solid (those most loosely bound) are the electrons in the conduction band. These electrons, as we have indicated in Chapter 24, behave according to the predictions of the quantum mechanical free electron model; that is, they occupy quasicontinuously all the energy levels available to them from the bottom of the band up to the Fermi level E_F. One reasonable modification to the quantum mechanical free electron model is to change the infinite potential well to a finite one. An infinite potential well would imply that an infinite amount of energy is needed to liberate these electrons, which is experimentally not true. If we take as our zero energy the vacuum, that is, when the electrons are outside the metal, the energy diagram for the conduction band can be represented as in Fig. 26-1. It is now clear that the

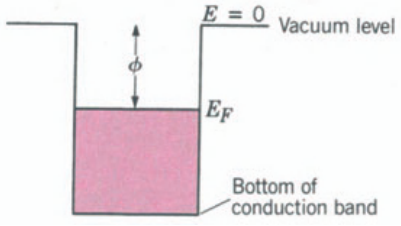


FIGURE 26-1
Energy diagram for the electrons in the conduction band of a metal. The zero energy level is chosen to be the energy outside the metal (vacuum level). The electrons at E_F need the least amount of energy to come out of the metal. This minimum energy was first introduced in Chapter 17 as the work function ϕ of the metal.

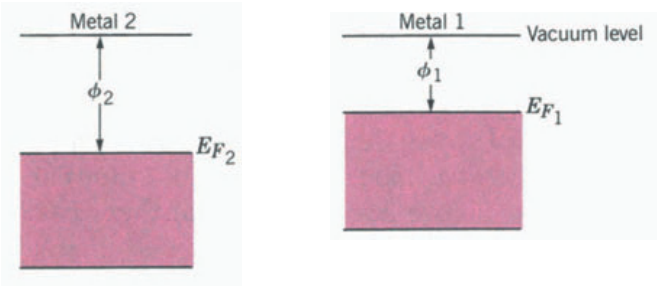


FIGURE 26-2
Energy diagram for the electrons in the conduction band of two dissimilar isolated metals with different work functions. Note that E_{F1} is closer to the vacuum level than is E_{F2}, and therefore $\phi_1 < \phi_2$

meaning of the work function is simply the energy difference between the vacuum level and the Fermi level. The electrons at E_F require the least photon energy to be ejected to the outside. An electron below the Fermi level requires a photon of correspondingly higher energy to be ejected. If this higher energy photon is absorbed by an electron at the Fermi level, however, it will have more energy than ϕ, the required energy to leave the metal. This extra energy is the kinetic energy of the ejected electron. Because kinetic energy is a function of the velocity, when photons with energy $h\nu > \phi$ are used, the emitted electrons will have a distribution of velocities. The fact that the work function is not the same for all metals implies that the position of the Fermi level relative to the vacuum level is different for different metals.

Let us now consider two metals, 1 and 2, with work functions ϕ_1 and ϕ_2, respectively, and let $\phi_1 < \phi_2$. When the two metals are far apart, their conduction bands will look as in Fig. 26-2. If these metals are placed in contact, electrons are free to flow from one to the other. Because the electrons near the Fermi level of metal 1 have higher energy than those in metal 2, there will be a net flow of electrons from metal 1 to metal 2. This flow will continue until the highest occupied energy level is the same in both metals, that is, until both metals have a common Fermi level. Figure 26-3*a* illustrates that the Fermi levels of the two metals in contact are both changed from their values in isolation to a common level. This is analogous to the situation where two containers filled with water up to different levels are interconnected. Water will flow from the container where the level is higher (and therefore the potential energy of the water is higher) into the other container, until the water level is the same in both containers.

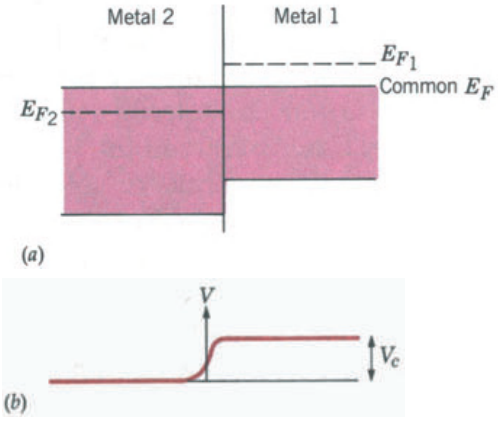


FIGURE 26-3
(*a*) The two metals of Fig. 26-2 are brought in contact. Electrons near the top of the Fermi level of metal 1 will move into the lower energy levels above the Fermi level of metal 2 until the maximum energy of the electrons in both metals is the same, that is, until the Fermi level is the same in both metals. (*b*) The flow of electrons from metal 1 into metal 2 leaves the surface of the latter negatively charged whereas that of metal 1 remains positively charged. This results in a potential difference V_c between the two metals.

As a result of the flow of electrons, metal 2 will gain some electrons and become negatively charged whereas metal 1 will become correspondingly positively charged because of the loss of electrons. The net result is that *an electrical potential difference will be established across the junction. Metal 1, which is positively charged, will be at a higher potential than metal 2.* This potential difference is called the *contact potential* V_c. Figure 26-3*b* shows that there has been a shift in voltage between the two metals in contact as a result of the shift in electron concentration from one to the other.

There is an important principle here: *The Fermi levels of two conducting solids in contact must be equal.* Although we have shown this for two metals, the conclusion holds for semiconductors and is the result of more general thermodynamic arguments. It should be noted, however, that this is an equilibrium condition that can be varied by the application of an external voltage. As will be shown later, such controlled variation is basic to the operation of the diode and the transistor.

26.3 THE SEMICONDUCTOR DIODE

26.3a Contact Potential: Band Scheme of a *p-n* Junction

The simplest semiconductor device is the *p-n* junction, or semiconductor *diode*, a device that permits current flow in only one direction. It consists of a *p*-type semiconductor placed in contact with an *n*-type semiconductor (see Fig. 26-4*a*). The electronic symbol of the diode is shown in Fig. 26-4*b*. Although in practice the junction is not formed by physically putting side by side a piece of *n*-type semiconductor and a piece of *p*-type, we will assume this for the present time. Some of the methods used to make *p-n* junctions are discussed in Chapter 28.

Before contact is made, the *n* region has a large number of free electrons (electrons in the conduction band) provided by the donor impurities. The charge of these mobile electrons in the *n*-type semiconductor is neutralized by the space charge of the positive donor ions, thus guaranteeing charge neutrality. Similarly, in the *p* type semiconductor, there is a large number of mobile holes whose charge is neutralized by the space charge of the negative acceptor ions. When the two types of semiconductors come in contact there will be a net diffusion of electrons from the *n* region into the *p* region. This is analogous to what happens when two gas containers are interconnected. If the density of molecules (and therefore the pressure) is not the same in both containers, there will be a net diffusion of molecules from the high density container to the low density one. Having crossed the junction, these electrons find themselves in a region where there are lots of holes with which they can and do recombine. While this is going on, the holes in the *p* region are diffusing across the junction into the *n* region; once there, they recombine with the electrons of the conduction band of that region. The net effect of this double diffusion and recombination is to deplete the region near the junction, on both sides, of its majority mobile carriers, thus making it highly resistive. At the same time, the immobile donor ions on the *n* side and the acceptor ions on the *p* side are left charged, thus creating charged layers on

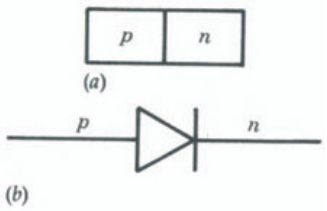


FIGURE 26-4
(*a*) Physical composition of a semiconductor diode. (*b*) Electronic symbol of a semiconductor diode.

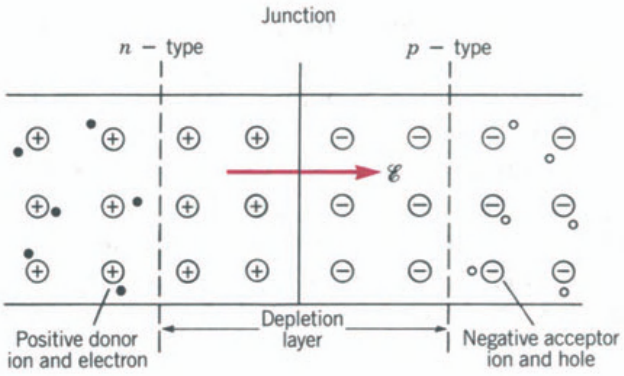


FIGURE 26-5
Formation of a *depletion layer* (region without mobile charge carriers) at the *p-n* junction resulting from the diffusion of electrons from the *n* region into the *p* region and diffusion of holes from the *p* region into the *n* region.

both sides of the junction, a positive layer in the *n*-type region and a negative layer in the *p*-type region. The situation is illustrated in Fig. 26-5. The diffusion of majority carriers across the junction in each direction ends quickly. The reason is that the charge layers at the junction create an internal electric field $\mathscr{E}$ that prevents further diffusion, see Fig. 26-5. Thus, equilibrium is established and no net current across the junction takes place.

This picture of how equilibrium is established is a kinetic one. The situation can also be analyzed from an energy point of view. The charge layers create a potential difference V_c between both sides of the junction, the *n* side being positively charged will be at a higher potential (voltage) than the *p* side, which is negatively charged. We may expect that the potential will vary monotonically in the depletion region (see Fig. 26-6*a*). This potential difference

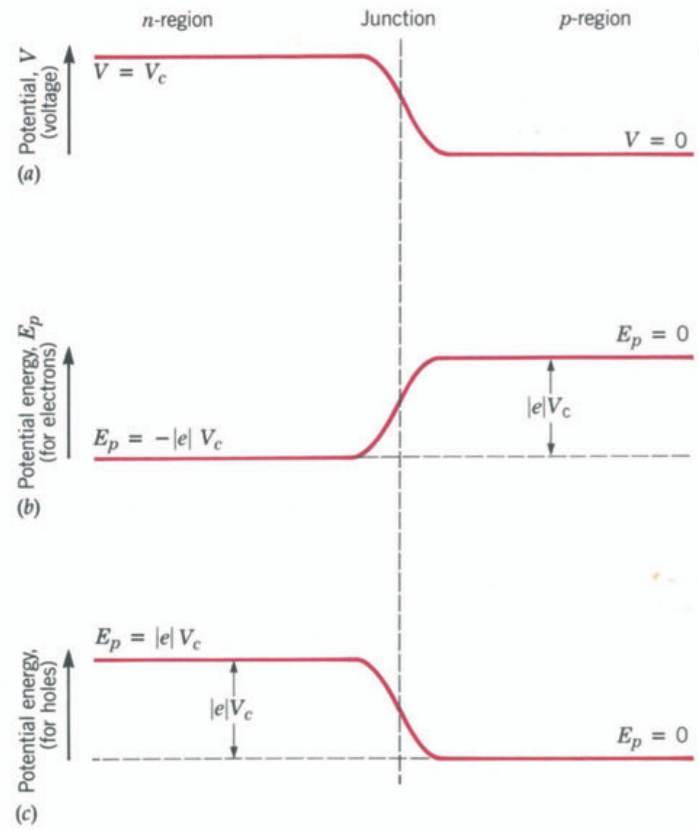


FIGURE 26-6
(*a*) Potential difference V_c resulting from the positive donor ions in the *n* side of the depletion layer and the negative acceptor ions in the *p* side of the depletion layer. (*b*) Potential energy barrier faced by the majority charge carriers (electrons) in the *n* side of the diode as they attempt to cross the junction. (*c*) Potential energy barrier faced by the majority charge carriers (holes) in the *p* side of the diode as they attempt to cross the junction.

is analogous to the contact potential that appeared in the metal-metal junction discussed in the previous section.

We know that when a charge q is placed at a point of potential V, the potential energy of the charge $E_p = qV$ (Eq. 14.16). Therefore, when an electron is in the n-region its potential energy $E_p = -|e|V_c$, whereas in the p-region the potential energy $E_p = 0$, that is, the potential energy of the electrons in the p side of the junction is shifted upward relative to the n side by an amount $|e|V_c$; this is illustrated in Fig. 26-6*b*. For the holes, the opposite is true. In the n side, the potential energy of the holes is $E_p = +|e|V_c$, whereas in the p side $E_p = 0$, see Fig. 26-6*c*. These potential energy barriers stop the further flow of majority carriers across the junction. Thus, one of the consequences of the contact potential is the upward shift of the electronic energy bands in the p side, relative to the bands in the n side, by an amount $|e|V_c$. This shift is associated with the requirement that the Fermi level be the same on both sides of the junction. The shifting of the bands is illustrated in Fig. 26-7. The fact that E_F must be the same on both sides determines the value of the contact potential. Because the position of the Fermi level in each side relative to its band structure (for example, relative to the top of its conduction band) does not change on contact, the energy shift in the bands $|e|V_c$ must be equal to the difference that existed between the two Fermi levels before contact, that is,

$$|e|V_c = E_{Fn} - E_{Fp} \tag{26.1}$$

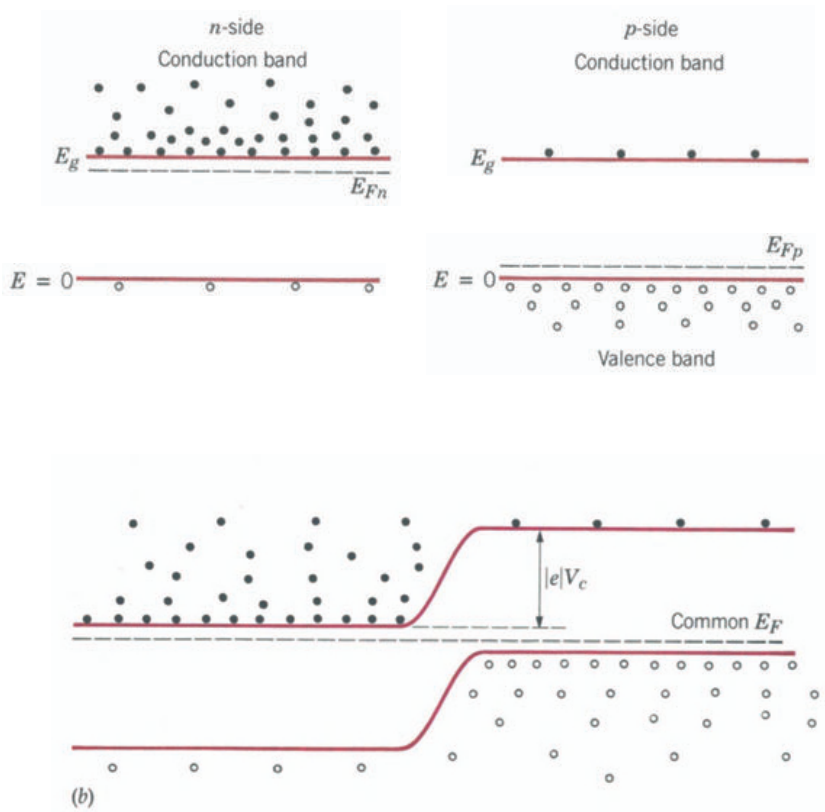


FIGURE 26-7
(*a*) Electronic energy bands of the n-type and the p-type semiconductors before contact is made. (*b*) Upward shifting of the electronic energy bands in the p side relative to the n side after the junction is formed. The shifting is due to the increase in the potential energy of the electrons in the p region relative to the n region (see Fig. 26-6*b*).

where E_{Fn} and E_{Fp} are the Fermi levels in the n side and in the p side, respectively. In a doped semiconductor the position of the Fermi level depends on the impurity concentration and on the temperature (see Section 25-3b); the contact potential V_c will be determined by these two parameters. Typically, for a silicon diode V_c has values that range between 0.6 and 0.9 V at ambient temperatures.

26-3b Equilibrium Currents Across the p-n Junction

In the preceding section we saw that when the p-n junction is formed a potential energy barrier $|e|V_c$ is formed that stops the flow of majority carriers across the junction. This is a dynamic equilibrium. Electrons and holes continuously flow across the junction; the *net* flow is zero because equal amounts flow in opposite directions. Let us consider only the flow of electrons as an illustration. There are electrons in the conduction band of both the n-type and the p-type semiconductors. In Chapter 25, Section 25-2a, we showed that the number of electrons N_e in the conduction band is (Eq. 25.3),

$$N_e = N_c \exp\left(-\frac{E_g - E_F}{k_B T}\right) \tag{25.3}$$

Although Eq. 25.3 gives the free electron concentration for both types of semiconductors, the numerical value of N_e, at ambient temperatures, is several orders of magnitude smaller for the p-type than for the n-type. The reason is that $E_g - E_F$ is much greater for the p-type than for the n-type (see Fig. 26-8).

The electrons (minority carriers) in the conduction band of the p region are not impeded by the potential energy barrier from crossing the junction from the p side to the n side. The electron current from p to n, $i(p \rightarrow n)$, will be proportional to the total number of electrons in the p region. We may write the proportionality of this current to N_e as an equality by introducing A as

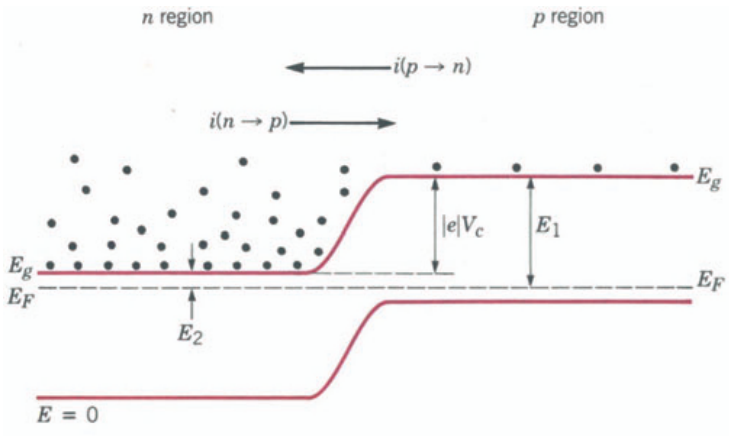


FIGURE 26-8
After the junction is formed, an equilibrium is established where there is no net flow of electrons (or holes) across the junction. Electrons do flow across the junction. The relatively few (minority carriers) electrons in the p side are not impeded by the energy barrier from crossing the junction. This flow is compensated by the flow (in the opposite direction) of those electrons in the n side with energies $E > |e|V_c$.

the proportionality constant:

$$i(p \rightarrow n) = A \exp\left(-\frac{E_1}{k_B T}\right) \tag{26.2}$$

where $E_1 = E_g - E_F$ of the p side (see Fig. 26-8).

In the n side there are a large number of electrons (majority carriers) in the conduction band. However, only those having energy equal to or greater than the barrier energy $|e|V_c$ will be able to cross the junction from the n side to the p side. The electron current, $i(n \rightarrow p)$, associated with this flow will be proportional to the number of electrons in the n region with energies greater than or equal to $|e|V_c$,

$$i(n \rightarrow p) = A\, N_e\, f(E \geq |e|V_c) \tag{26.3}$$

where N_e is the total number of electrons in the conduction band of the n side and $f(E \geq |e|V_c)$ is the fraction of these electrons with energies greater than or equal to $|e|V_c$. In Chapter 25, Eq. 25.2, we indicated that for the electrons in the conduction band of silicon, at ambient temperatures, the Fermi-Dirac distribution can be approximated by the Maxwell-Boltzmann distribution. Moreover, in Supplement 9-1 of Chapter 9, we showed that for a system of particles having the Maxwell-Boltzmann distribution of energies, the fraction of particles with energies greater than or equal to a given value E_i is given by the Boltzmann factor, $\exp(-E_i/k_B T)$, Eq. 9.41. Therefore, in the present case, the fraction of electrons having energies greater than the barrier $|e|V_c$ is

$$f(E \geq |e|V_c) = \exp\left(-\frac{|e|V_c}{k_B T}\right) \tag{26.4}$$

Substitution of Eq. 25.3 for N_e and Eq. 26.4 for $f(E \geq |e|V_c)$ in Eq. 26.3, yields

$$i(n \rightarrow p) = A \exp\left(-\frac{E_2 + |e|V_c}{k_B T}\right) \tag{26.5}$$

where $E_2 = E_g - E_F$ in the n region (see Fig. 26-8). From Fig. 26-8, it is seen that $E_2 + |e|V_c = E_1$, and it follows that

$$i(n \rightarrow p) = i(p \rightarrow n)$$

The flow of electrons from n to p is equal to the flow from p to n. Therefore, the net current is zero. Although the preceding calculations have not yielded any startling results, we can draw an important conclusion from them. The electron current from p to n (that is, associated with minority carriers) is not affected by the height of the potential energy barrier because once an electron is in the conduction band there is no barrier in the p to n direction. On the other hand, the electron current from n to p (associated with majority carriers) depends by a negative exponential on the height of the barrier, Eq. 26.5. It is this property that makes the p-n junction a rectifier; that is, current flows easily in one direction but not in the other. This is considered next.

26-3c Voltage-Current Characteristics of a Diode: Rectification

Let us now consider what happens to the potential energy barrier at the *p-n* junction when an external voltage V_0 is applied across the diode. The two possible ways this voltage can be applied are illustrated in Figs. 26-9*a* and 26-9*b*. In the first case (Fig. 26-9*a*) the diode is said to be *forward biased*, and in the second case (Fig. 26-9*b*) it is said to be *reverse biased*. We have seen in Fig. 26-5 that the contact region between the *n* and the *p* sides is depleted of its majority carriers. This depletion region therefore has a high resistance and, as a consequence, the potential difference V_0 will appear almost entirely across it (see Eq. 15.15). In the forward biased case (Fig. 26-9*a*) the battery lowers the potential of the *n* side of the junction relative to the *p* side by V_0. Because initially (before V_0 was applied) the potential of the *n* side was higher than the *p* side by V_c (see Fig. 26-6*a*), the difference in potential between the *n* side and the *p* side is now *reduced* to $V_c - V_0$. This is shown in Fig. 26-10*a*, where the solid line represents the new potential difference and the dashed line represents the potential difference before V_0 is applied.

Concomitant with the reduction in the potential difference between both sides of the junction, there is a lowering of the potential energy barrier encountered by the majority carriers as they try to cross the junction. In the *n* region, the potential energy of the electrons is now $E_p = -|e|(V_c - V_0)$,

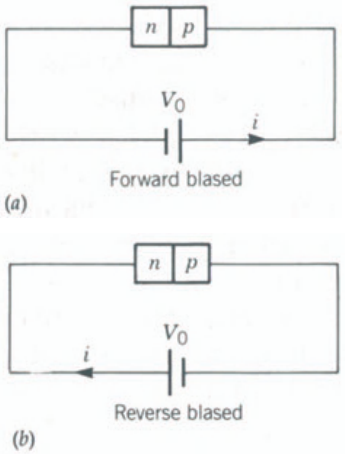


FIGURE 26-9
Two possible ways of connecting a diode to an external source of potential difference V_0. (*a*) Forward biased diode. (*b*) Reverse biased diode.

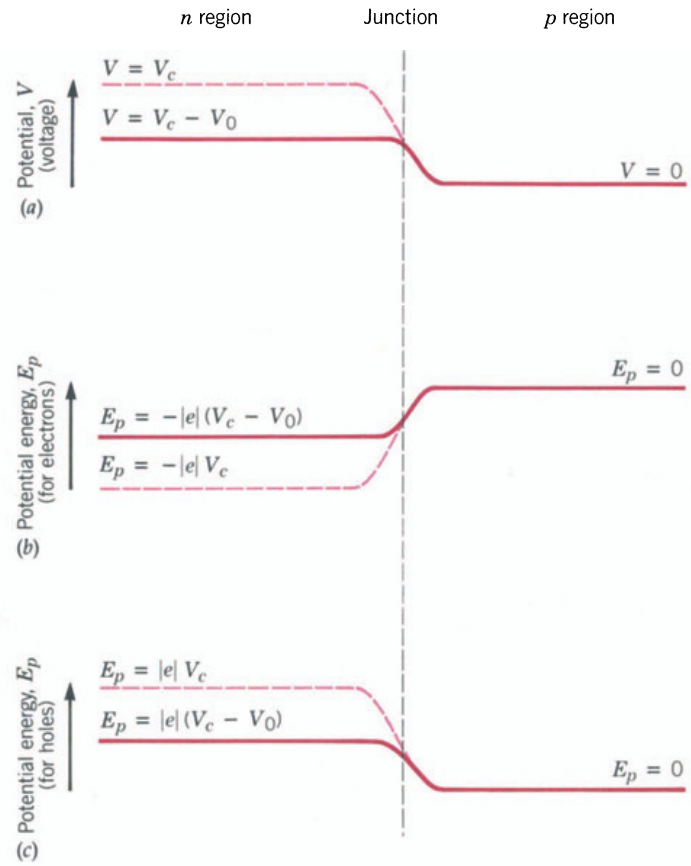


FIGURE 26-10
Forward biased diode. (*a*) Potential difference between the *n* and the *p* sides of the junction in the forward biased case of Fig. 26-9*a*. The battery has reduced the potential difference from the value V_c (dashed line) in the unbiased case (see Fig. 26-6*a*) to $V_c - V_0$ (solid line). (*b*) New (solid line) and old (dashed line) potential energy barrier for the electrons in the *n* side of the junction. (*c*) New (solid line) and old (dashed line) potential energy barrier for the holes in the *p* side of the junction.

whereas in the p region it is $E_p = 0$. The new energy barrier $|e|(V_c - V_0)$ encountered by the electrons as they attempt to cross from the n side to the p side is represented schematically by the solid lines of Fig. 26-10*b*. The dashed lines in Fig. 26-10*b* represent the old energy barrier $|e|V_c$. Similar remarks apply for the holes. In the n region, their potential energy $E_p = |e|(V_c - V_0)$ whereas in the p region $E_p = 0$. The new (solid lines) and the old (dashed lines) energy barriers encountered by the holes as they attempt to cross the junction from the p side to the n side are shown in Fig. 26-10*c*. The lowering of the energy barrier will greatly increase the flow of majority carriers crossing the junction, thus giving rise to a large current through the diode. We will return to this shortly. First, however, let us consider what happens when the diode is reverse biased (Fig. 26-9*b*). In this case, the battery will raise the potential of the n side relative to the p side by V_0. Because the potential of the n side was already higher than that of the p side by V_c, the potential difference between the two sides of the junction will now *increase* to $V_c + V_0$. As a result, the potential energy barrier encountered by the majority carriers will also increase from $|e|V_c$ (unbiased condition) to $|e|(V_c + V_0)$ (reverse biased condition). Figure 26-11*a* shows the new (solid lines) and the old (dashed lines) potential differences between the two sides of the junction.

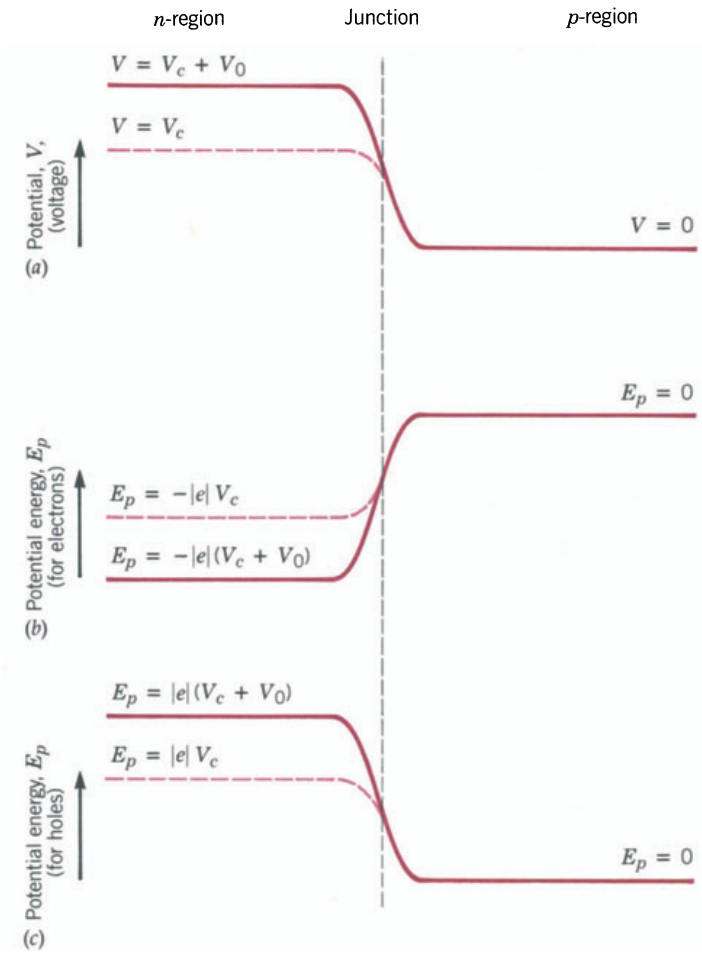


FIGURE 26-11
Reverse biased diode. (*a*) Potential difference between the n and the p sides of the junction in the reverse biased case shown in Fig. 26-9*b*. The battery increases the potential difference from V_c (dashed line) in the unbiased case to $V_c + V_0$ (solid line). (*b*) and (*c*) The potential energy barriers faced by electrons in the n side of the junction and by the holes in the p side increase from $|e|V_c$ to $|e|(V_c + V_0)$.

The associated potential energy barriers for electrons and holes are shown in Figs. 26-11*b* and 26-11*c*, respectively. Again the solid lines represent the energy barriers in the reverse biased condition and the dashed lines in the unbiased condition. In sum, we have shown that the height of the energy barrier is $|e|(V_c - V_0)$ when the diode is forward biased and $|e|(V_c + V_0)$ when it is reverse biased. If we adopt the sign convention that V_0 *is positive in the forward biased case and negative in the reverse biased case,* we can make the following statement that applies to both bias conditions: *When an external voltage* V_0 *is applied across the diode, the height of the potential energy barrier at the* p-n *junction is* $|e|(V_c - V_0)$.

Let us now consider the crucial question: What is the net flow of charge carriers across the junction under forward and reverse bias conditions? Again for the sake of simplicity, we will limit our discussion to the flow of electrons. The *net* electron current from the n side to the p side will be

$$i = i(n \rightarrow p) - i(p \rightarrow n) \tag{26.6}$$

The first term $i(n \rightarrow p)$, which represents the flow of electrons (majority carriers) from n to p, is proportional to the number of electrons in the conduction band of the n region having energies greater than the height of the energy barrier, which now is $|e|(V_c - V_0)$. This number, as we saw earlier, Eq. 26.3, is equal to the total electron density N_e multiplied by the Boltzmann factor $\exp(-E_i/k_BT)$, where in this case E_i is the barrier energy $|e|(V_c - V_0)$; that is,

$$i(n \rightarrow p) = A \exp\left(-\frac{E_2 + |e|(V_c - V_0)}{k_BT}\right) \tag{26.7}$$

where $E_2 = E_g - E_F$ in the n side (see Fig. 26-8). The second term of Eq. 26.6, $i(p \rightarrow n)$, which represents the flow of electrons (minority carriers) from p to n, is the same as in the unbiased condition, because there is no barrier for the electrons in the p to n direction. Thus $i(p \rightarrow n)$ is still given by Eq. 26.2. Substituting Eq. 26.7 for $i(n \rightarrow p)$ and Eq. 26.2 for $i(p \rightarrow n)$ in Eq. 26.6, we obtain

$$i = A \exp\left(-\frac{E_2 + |e|(V_c - V_0)}{k_BT}\right) - A \exp\left(-\frac{E_1}{k_BT}\right)$$

Recalling that $E_2 + |e|V_c = E_1$ (see Fig. 26-8), we write

$$i = A \exp\left(-\frac{E_1 - |e|V_0}{k_BT}\right) - A \exp\left(-\frac{E_1}{k_BT}\right)$$

$$= A \exp\left(-\frac{E_1}{k_BT}\right) \exp\left(+\frac{|e|V_0}{k_BT}\right) - A \exp\left(-\frac{E_1}{k_BT}\right)$$

Letting

$$i_0 = A \exp\left(-\frac{E_1}{k_BT}\right)$$

we obtain

$$i = i_0\left[\exp\left(+\frac{|e|V_0}{k_BT}\right) - 1\right] \tag{26.8}$$

$$i = i_0\left[\exp\left(+\frac{|e|V_0}{k_BT}\right) - 1\right]$$

We note that i_0 is the current associated with the flow of minority carriers (electrons) from the p side to the n side (Fig. 26-8) because E_1 is unchanged by the applied voltage. It is called the *reverse saturation current.* In a silicon diode at room temperature, the magnitude of i_0 is typically $\sim 0.1\ \mu$A. This is a very small value, and the reason is that the supply of electrons in the conduction band of a p-type semiconductor is very small because E_1 is large compared with the room temperature thermal energy k_BT.

Equation 26.8 is called the *diode equation.* Let us see what it tells us. Recall that in the derivation of Eq. 26.8, we used the convention that V_0 is taken as positive in the forward biased case and negative in the reverse biased case. Because V_0 is negative when the diode is reverse biased, the exponential term of the equation becomes negligible for relatively small voltages, and a constant *small* electron current $i \approx -i_0$ remains. (The negative sign means that the net electron flow is from p side to n side.) We do not need a very large voltage to reach this situation. At room temperature, $k_BT = 0.025$ eV. If we take $V_0 = 0.1$ V, then $i = i_0(e^{-4} - 1) = -0.98\, i_0$. On the other hand, if the diode is forward biased, the current increases very rapidly with increasing voltage. For example, if

$$V_0 = 0.1\text{ V}, \quad i = i_0(e^4 - 1) = 54\, i_0$$

whereas if

$$V_0 = 0.5\text{ V}, \quad i = i_0(e^{20} - 1) = 4.8 \times 10^8\, i_0$$

In practice, the current does not rise quite so rapidly because of the resistance of the body of the semiconductor material as well as other resistances present in the circuit. In most circuits used in computers and other digital instruments, the current is limited by resistors in the circuit to a few milliamps. As a result, *the voltage across a forward biased diode is almost never greater than a few tenths of a volt; 0.6 V or less* is a fairly typical value for a silicon diode. The behavior described by Eq. 26.8 is illustrated in Fig. 26-12.

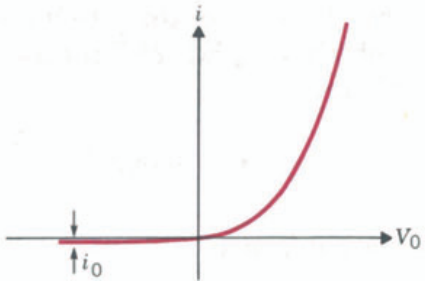


FIGURE 26-12
Current-voltage characteristics for a biased diode. Positive values of V_0 correspond to the forward biased case and negative values of V_0 to the reverse biased one. The value of i_0 in the figure is highly exagerated compared with the value of the current in the forward biased condition.

The discussion of the direction of the electron current referred only to the direction of motion of the electrons. As indicated in Chapter 15, Section 15.3, the direction of the conventional current is that of the hypothetical motion of positive charges, which is, of course, opposite to the direction of motion of the electrons. This conventional concept, that in a circuit the current is from the positive terminal of the battery to the negative one, is a holdover from the early days of electricity before the discovery of the electron, and it is still in use. Thus, in the forward bias case, the net electron flow is from the n region to the p region, whereas the conventional current would be from the p to the n (see Fig. 26-9*a*).

Holes have not been discussed for the sake of simplicity. It is not difficult to go through arguments for the holes similar to those for electrons and to

show that their behavior, under forward and reverse bias, is identical to that of the electrons. Of course, the flow of holes is in the opposite direction to that of the electrons, and so is the potential barrier. The expression for the hole current is the same as that obtained for the electrons.

Two final comments should be made about the width of the depletion layer.

1. The width of the depletion layer increases as the potential difference across the junction increases. In Section 26-3a, we saw that a contact potential V_c is created by the ions on both sides of the junction when the majority carriers diffuse across the junction. Clearly, the greater the number of charged ions (and therefore, the wider the depletion layer), the greater the potential difference across the junction will be. Conversely, if the potential difference across the junction is increased, such as by the application of a reverse bias voltage V_0 (see Fig. 26-11*a*), the number of charged ions on both sides of the junction will also increase.
2. For a given junction potential, the depletion layer is wider for the semiconductor that is the least doped. The reason for this is that a greater depth has to be depleted to match the number of carriers that can be obtained from a small depth of a highly doped semiconductor. These two facts about the width of the depletion layer will be important in understanding the behavior of field effect transistors (FET), which will be discussed later.

26.4 THE BIPOLAR JUNCTION TRANSISTOR (BJT)

When a layer of *n*-type silicon is sandwiched between two layers of *p*-type silicon, we obtain a *pnp transistor*. If a layer of *p*-type is sandwiched between two layers of *n*-type, we have an *npn transistor*. The filling of the sandwich is called the *base* (B) and the two ends are called the *emitter* (E) and the *collector* (C). The two types of transistors and their electronic symbols are shown in Fig. 26-13. One very important fact is that in both cases *the base is very narrow* ($\sim 10^{-6}$ *m) and very lightly doped compared with the emitter.*

An *npn* transistor can be considered as an *np* diode followed by a *pn* diode, and a similar analogy applies to the *pnp* transistor. Consequently, the ideas presented in the previous discussion of the diode will be used to explain the physical behavior of the transistor. We will limit our discussion to the *npn* transistor. Arguments similar to ones to be presented for the *npn* transistor can be made for the *pnp* transistor.

When the three semiconductor layers are put together, majority carriers will flow across the two junctions, creating contact potential differences between the emitter and the base and between the collector and the base. If we assume completely symmetrical junctions (we assume that the emitter and the collector have identical doping concentrations), the contact potential V_c at the emitter-base junction and at the collector-base junction will be equal. Recall that in the absence of an external bias voltage, the potential of the *n*

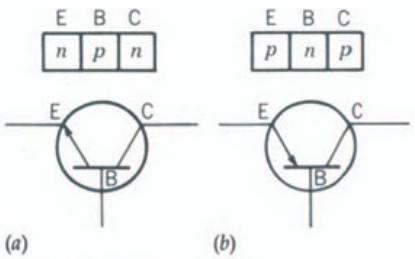


FIGURE 26-13
(*a*) Physical composition and electronic symbol of an *npn* transistor. (*b*) Physical composition and electronic symbol of a *pnp* transistor.

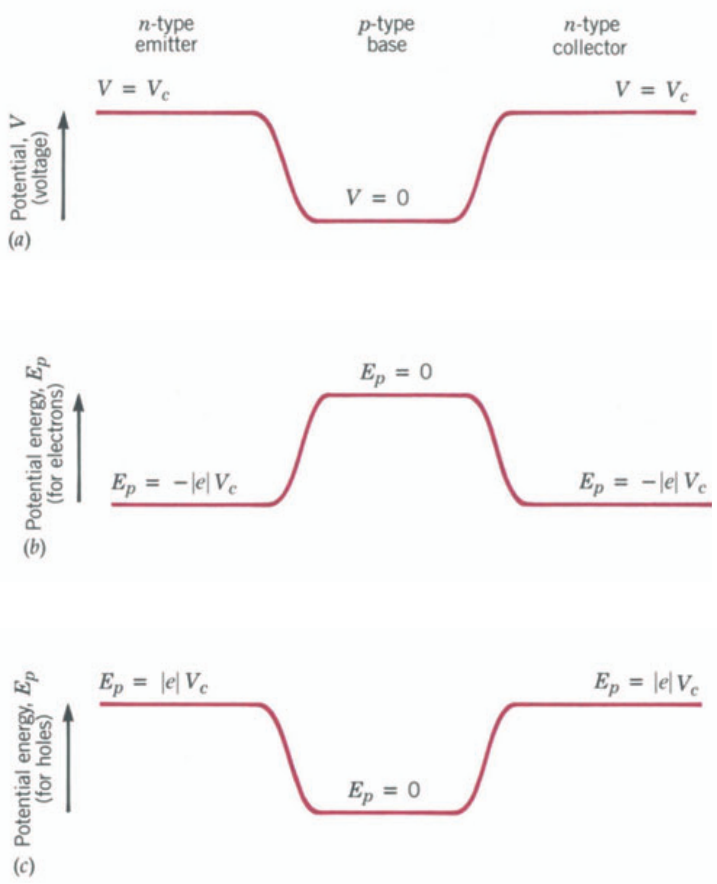


FIGURE 26-14
(*a*) Potential differences between the three parts (emitter, base, collector) of the transistor associated with the depletion layers at the emitter-base junction and at the base-collector junction. (*b*) and (*c*) Potential energy barriers for electrons and holes, respectively, associated with the potential differences of (*a*).

side of a *pn* junction is higher than that of the *p* side by the contact potential V_c (Section 26-3a). The variations in potential in the three sections of the *npn* transistor are shown in Fig. 26-14a. The associated potential energy barriers for electrons and holes are illustrated in Figs. 26-14*b* and 26-14*c*, respectively.

26-4a Common Base Configuration

To appreciate the basic features of a transistor as an active circuit element, let us consider the circuit shown in Fig. 26-15. The left side of the circuit, that is, the section containing V_1, the base, and the emitter, is identical to the forward biased diode circuit of Fig. 26-9*a*. On the other hand, the right side, the section containing V_2, the collector, and the base, is identical to the reverse biased diode circuit of Fig. 26-9*b*. Because the base is common to both sections, Fig. 26-15 is referred to as a *common base configuration circuit*. For the reason given earlier (high resistance of the depletion layer at the junctions), the externally applied voltages will appear almost entirely across the junctions. The battery V_1, on the left-hand circuit, will *raise* the potential of the *p* type base relative to the *n* type emitter by an amount that we will call V_{BE}. As a result, the potential difference between the emitter and base is *reduced* to $V_c - V_{BE}$. On the right-hand circuit, the opposite happens. The battery V_2 *lowers* the potential of the base relative to the collector by an amount that we

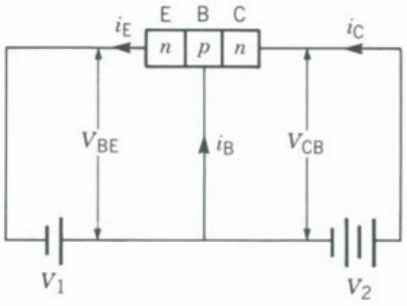
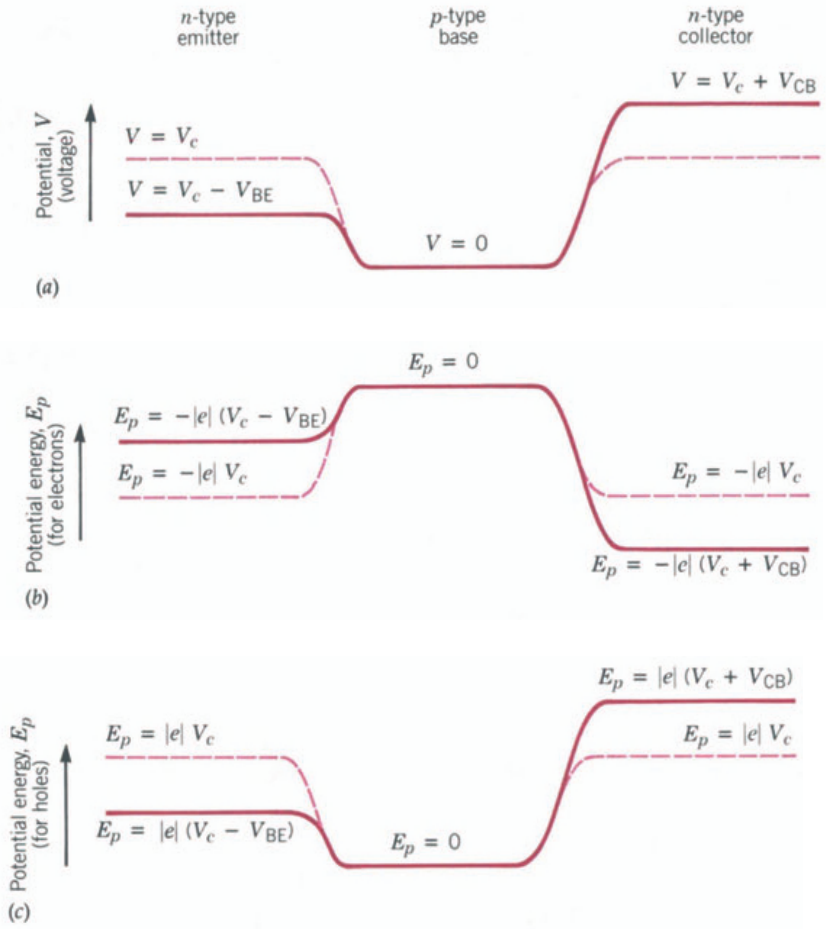


FIGURE 26-15
Common base configuration transistor circuit.



(*a*) Reduction in the potential difference between the emitter and the base by the battery V_1 of Fig. 26-15 and increase in the potential difference between the collector and the base by the battery V_2 of the same figure. The dashed lines represent the potential differences in the unconnected transistor. (*b*) and (*c*) Potential energy barriers for the electrons and holes, respectively, associated with the new potential differences of (*a*). The dashed lines are the energy barriers in the unconnected transistor.

will call V_{CB}; as a result the potential difference between the collector and the base is *increased* to $V_c + V_{CB}$. Just as in the case of the diode, this leads to a reduction in the potential energy barrier at the emitter-base junction and an increase in the barrier height at the collector-base junction. The new potential differences between the junctions are sketched in Fig. 26-16*a*. The associated potential energy barriers for electrons and holes are shown in Figs. 26-16*b* and 26-16*c*, respectively. The dashed lines in the three figures represent the values of these quantities before the application of the biasing voltages.

The lowering of the barrier at the emitter base junction will permit electrons to be injected from the emitter into the base, as well as holes from the base into the emitter. The electrons that neither recombine with holes in the base region nor flow through the contact lead on the base will diffuse across the base and reach the collector junction (see Fig. 26-17). Once there, they see the positive potential difference $V_c + V_{CB}$ (negative potential energy), they are accelerated across the junction, and are "collected" by the collector.

Let us analyze the various components contributing to the emitter current i_E, the collector current i_C, and the base current i_B shown in Fig. 26-15. First, however, we should note that any charge carrier crossing the emitter-base junction will contribute to the emitter current. For example, if an electron

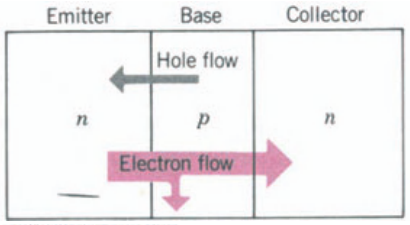


FIGURE 26-17
Schematic representation of the flow of electrons and holes across the energy barriers at the two junctions of the transistor for the common base configuration circuit of Fig. 26-15.

from the emitter crosses this junction into the base, an electron from the wire to the left of the emitter will enter the emitter region to preserve the charge neutrality of this region. When a hole crosses the junction from the base to the emitter, it will recombine with an electron in the emitter. As a result, the emitter region will have a net positive charge $+|e|$ that will be neutralized by the entrance of an electron from the wire to the left of the emitter. For similar reasons, any charge carrier crossing the collector-base junction will contribute to the collector current i_C. The difference between the carriers crossing the two junctions accounts for the base current i_B.

The emitter current i_E consists of an electron current i_{eE} (majority electrons crossing from the emitter into the base) and a hole current i_{pE} (majority holes crossing from the base into the emitter),

$$i_E = i_{eE} + i_{pE} \tag{26.9}$$

In a commercial transistor, the doping of the emitter is much greater than that of the base. Because the current across the junction depends not only on the height of the barrier but also on the number of majority carriers in each region, this implies that $i_{pE} << i_{eE}$. Eq. 26.9 can be approximated as

$$i_E \approx i_{eE} \tag{26.10}$$

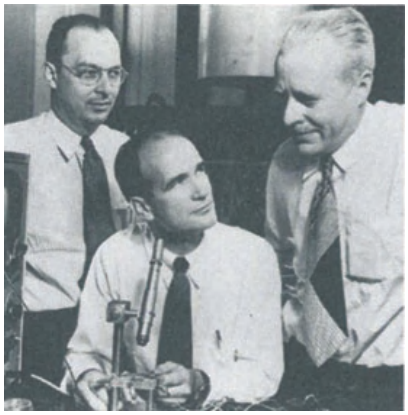
Let us now look at the collector current i_C. What are the components of i_C? The main contribution comes from the electrons emitted into the base that are able to diffuse across the base without recombining with holes. We can designate this contribution as $\alpha\, i_E$, where α is the fraction of the electrons that are able to diffuse across the base. How large is α? Its value depends on two things: (1) the so-called minority carrier lifetime τ_e (the time that an electron can survive without recombining in a hole-rich region), and (2) the time that the electron remains in the base as it diffuses toward the collector junction, τ_D. A typical value for τ_e is 10^{-4} sec and can be made greater by reducing the doping of the base. By making the base region very thin, the diffusion time τ_D can be made much smaller than the lifetime τ_e. For a base 0.1 mm wide, the diffusion time turns out to be roughly 10^{-6} sec. Thus, practically all the electrons will be able to get to the collector. Typical values for α range from 0.900 to 0.998.

Because the base-collector junction is reverse biased (see Fig. 26-16), there is no contribution from the majority carriers (holes in the base and electrons in the collector) to i_C. The only additional contribution is that from the small reverse saturation current i_0 due to the flow of minority carriers across the collector-base junction. This contribution, as shown in Section 26.3c, is extremely small. Thus,

$$\begin{aligned} i_C &= \alpha i_E + i_0 \\ &\approx \alpha i_E \end{aligned} \tag{26.11}$$

$i_C \approx \alpha i_E$

The important conclusion is that *the collector current is essentially determined by the emitter current*, which is determined by the potential difference between



In 1956 John Bardeen, William Shockley and Walter Brattain shared the Nobel Prize in Physics for their discovery of the transistor.

the base and the emitter V_{BE}. It is essentially independent of the potential difference between the collector and the base V_{CB}. Actually, the previous statement is not entirely correct because α can vary by 0.1% as V_{CB} changes from 0 V to 10 V. The reason is that the width of the depletion layer at the junction depends on the potential difference across the junction, as indicated in Section 26.3c. Thus, as the reverse bias voltage at the collector-base junction is increased, the depletion layer (region depleted of its majority carriers) penetrates deeper into both the collector and the base. This means that the "effective base" (the region where free holes exist) decreases. When this happens, the chances of recombination within the base decrease and thus α increases.

26-4b The Transistor as a Voltage Amplifier

Figure 26-18 illustrates a modified version of the common base configuration circuit that we have been considering, Fig. 26-15. A small resistor R_1 has been introduced in the emitter-base circuit and a larger resistor R_2 in the collector-base circuit. The reason for the introduction of these two resistors and for choosing R_2 greater than R_1 will become clear soon.

We should note that although the introduction of these two resistors will reduce the potential differences across the junctions (recall that when a current i flows through a resistor R, there is a potential drop iR across the resistor), it does not affect the type of bias of the junctions. The emitter-base junction remains forward biased because the p-type base is connected directly to the positive side of V_1 whereas the n-type emitter is connected through the resistor R_1 to the negative side of V_1. Similarly, the collector-base junction remains reverse biased because the p-type base is connected directly to the negative side of V_2 whereas the n-type collector is connected through the resistor R_2 to the positive terminal of V_2. We should also point out that the reduction of V_{CB} resulting from the presence of a large R_2 will not significantly affect the value of i_C because, as we have shown in the preceding section, i_C is essentially determined by i_E, not V_{CB}.

One of the main uses of the transistor is as an amplifier. In the common base configuration shown in Fig. 26-18, the input, as a voltage difference, is fed into the emitter-base circuit by applying a voltage ΔV_1 as shown in Fig. 26-18.

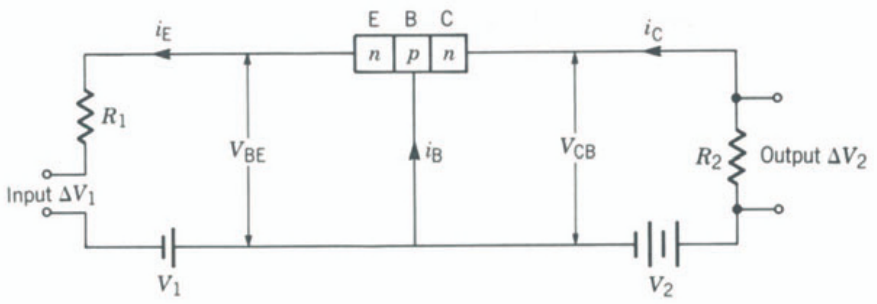


FIGURE 26-18
Modified version of the common base configuration circuit of Fig. 26-15. A small resistor R_1 and a large resistor R_2 have been introduced for voltage amplification purposes.

The output is taken out of the collector-base circuit as a voltage drop across resistor R_2. The circuit in Fig. 26-18 is an amplifier because *the voltage is amplified* by a large factor. The emitter current i_E is limited by R_1, which is usually chosen to be small (for example, 100 Ω), and the effective resistance R_e of the forward biased emitter-base junction. This current reappears ($i_C = \alpha i_E$) in the output circuit and flows through a large load resistor R_2 (for example, 50 kΩ), called output resistance, and causes a large voltage drop across it. Let us assume that a small voltage ΔV_1 is applied to the input terminals; this will lead to a change in the emitter current i_E so that

$$\Delta i_E = \frac{\Delta V_1}{R_1 + R_e}$$

and to a corresponding change in i_C

$$\Delta i_C = \alpha \, \Delta i_E = \alpha \frac{\Delta V_1}{R_1 + R_e} \tag{26.12}$$

The output voltage across the load resistor R_2 will increase by

$$\Delta V_2 = R_2 \, \Delta i_C$$

Substituting Eq. 26.12 for Δi_C

$$\Delta V_2 = \alpha \frac{R_2}{R_1 + R_e} \Delta V_1$$

and, because $\alpha \approx 1$, we obtain

$$\Delta V_2 \approx \frac{R_2}{R_1 + R_e} \Delta V_1 \tag{26.13}$$

$$\Delta V_2 \approx \frac{R_2}{R_1 + R_e} \Delta V_1$$

The input voltage will be amplified at the output by a factor equal to the ratio of the output to the input resistances.

The effective resistance R_e of the forward biased emitter-base junction is not constant but depends on the voltage across the junction. This is because in a *p-n* junction the current does not increase linearly with voltage; that is, a *p-n* junction does not obey Ohm's law (see Fig. 26-12). As we saw in Section 26.3c, for a forward biased voltage greater than a few tenths of a volt, the current increases very rapidly as the voltage across the junction increases; this means that R_e is very small, typically a few ohms. If we take $R_e = 10\ \Omega$, $R_1 = 100\ \Omega$, and $R_2 = 50\ \text{k}\Omega$,

$$\Delta V_2 = \frac{5 \times 10^4\ \Omega}{100\ \Omega + 10\ \Omega} \Delta V_1$$

$$\Delta V_2 = 455\ \Delta V_1$$

The input voltage is amplified by a factor of 455.

26.4c Common Emitter Configuration

The most common configuration, particularly in switching circuits, is the one in which the emitter is common to both the input and the output circuits. Figure 26-19 is an example of a common emitter configuration circuit. The input branch, consisting of V_1, R_1, and the base-emitter junction is identical to the input section of the common base configuration circuit of Fig. 26-18. The p-type base is connected directly to the positive terminal of V_1 whereas the n-type emitter is connected through the resistor R_1 to the negative terminal of V_1. The emitter-base junction is therefore forward biased. The output branch is different from any of the circuits that we have considered thus far. In addition to the battery V_2 and the load resistor R_2, it includes two junctions, the collector-base junction and the emitter-base junction. Because the emitter is connected directly to the negative terminal of V_2, whereas the collector is connected through R_2 to the positive terminal of V_2, the potential of the collector is higher than that of the emitter by an amount $V_{CE} = V_2 - i_C R_2$. The nature of the bias at the collector-base junction will depend on the relative values of the voltage between the collector and the emitter V_{CE} and the voltage between the base and the emitter V_{BE}. This will be considered in more detail below.

In the common emitter configuration, the input current i_B is the independent variable and the output current i_C is the dependent variable. Putting it differently, the base current i_B, which is determined by V_1 and R_1 because the emitter-base junction is forward biased, controls the collector current i_C. Figure 26-20 shows the characteristic curves for a standard 2N2222A *npn* transistor. The horizontal axis is the collector-emitter voltage V_{CE}, and the vertical axis is the collector current i_C. The curves correspond to different values of i_B.

The output characteristics can be divided into two regions: the *active region* and the *saturation region*. The active region is to the right of the knee of the curves; at the knee $V_{CE} \approx$ a few tenths of a volt. In this region i_C is very

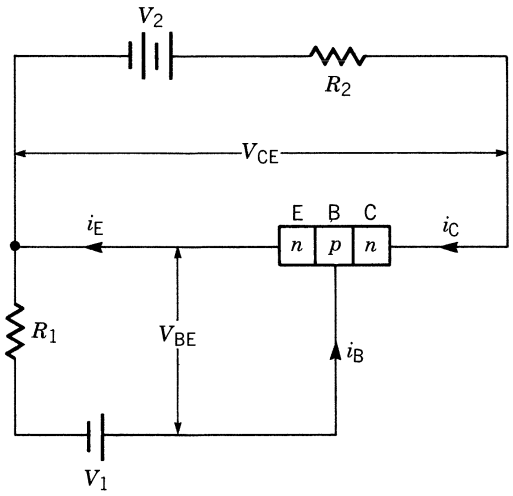


FIGURE 26-19
Common emitter configuration transistor circuit.

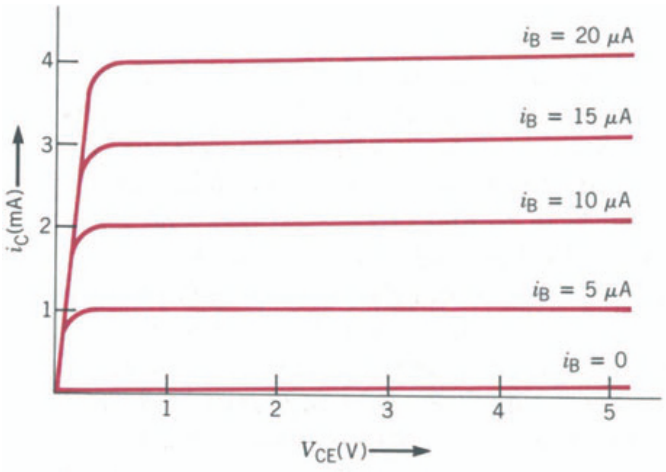


FIGURE 26-20
Output characteristic curves—i_C versus V_{CE} for different values of i_B, for the common emitter configuration circuit of Fig. 26-19.

sensitive to i_B but almost independent of V_{CE}. The saturation region is to the left of the knee. In the saturation region i_C increases rapidly with V_{CE}. In the active region, the collector-base junction is reverse biased. Let us see how this comes about. The emitter-base junction is forward biased. As indicated in Section 26.3c, the voltage across a forward biased junction is usually limited to a few tenths of a volt, 0.6 V or less is a fairly typical value for a silicon junction. Therefore, V_{BE} will be 0.6 V or less. If V_{CE} is greater than 0.6 V, then the collector is more positive than the base relative to the emitter, and this in turn means that the collector is at a higher potential than the base: Consequently, the collector-base junction will be reverse biased. Thus, in the active region the emitter-base junction is forward biased whereas the collector-base junction is reverse biased, that is, the situation is analogous to the previously discussed common base configuration. We can explain the variation of i_C by making use of our previous result, Eq. 26.11, $i_C = \alpha\, i_E$, and Kirchhoff's current rule,

$$i_E = i_B + i_C \tag{26.14}$$

Substituting for $i_E = i_C/\alpha$ in Eq. 26.14,

$$\frac{i_C}{\alpha} = i_B + i_C$$

$$i_C\left(\frac{1}{\alpha} - 1\right) = i_B$$

$$i_C\left(\frac{1-\alpha}{\alpha}\right) = i_B$$

$$i_C = \beta\, i_B$$

where $\beta = \alpha/(1-\alpha)$ is called the *current gain parameter*. We mentioned before (Section 26.4a) that α varies slightly with V_{CB} (and in this case with V_{CE}). The

$$i_C = \beta\, i_B$$

where

$$\beta = \frac{\alpha}{1-\alpha}$$

change in α as V_{CE} changed by 10 V was $\sim 0.1\%$. The variations in β can be significantly greater. For example, if we take $\alpha = 0.990$, then $\beta = 99$; if $\alpha = 0.991$, $\beta = 110$; a 0.1% change in α causes an 11% change in β. Other than this, the magnitude of i_C is essentially determined by i_B, and i_C is very sensitive to i_B. For example, in the case when $\alpha = 0.99$, we saw that $\beta = 99$; therefore, a change of 10 μA in i_B will give rise to a change of 990 μA in i_C. Thus, *when the common emitter circuit is used in the active region, it acts as a current amplifier.*

Let us now try to understand the rapid decrease in i_C when V_{CE} drops below a few tenths of a volt (see Fig. 26-20). In this region, $V_{CE} < V_{BE}$; that is, the base is more positive than the collector relative to the emitter, and therefore the base is at a higher potential than the collector. As a consequence, the collector-base junction will become *forward biased*. In our previous discussion of the common base configuration, we mentioned that the only significant contribution to i_C was due to that fraction α of the emitter-injected electrons that diffused across the base (that is, $i_C = \alpha i_E$). This is fine if the collector-base junction is reverse biased. In this case the only additional contribution to i_C is the small reverse saturation current i_0 associated with the flow of minority carriers across the CB junction; this is very small. If, however, the CB junction is forward biased, holes (majority carriers) in the base can flow into the collector (we will call this component i_{pC}) and electrons (majority carriers) in the collector can flow into the base (we will call this component i_{eC}) (see Fig. 26-21). This double flow of charge carriers corresponds to a large current that is in the opposite direction as that of αi_E. Thus the net collector current will be given by

$$i_C = \alpha i_E - (i_{pC} + i_{eC})$$

The net collector current will decrease rapidly as the CB junction becomes increasingly forward biased (that is, as V_{CE} decreases below 0.6 V), until $V_{CE} = 0$. At this point, the emitter and the collector are at the same potential. This in turn implies that the CB and the EB junctions are equally forward biased, and therefore the net current is zero.

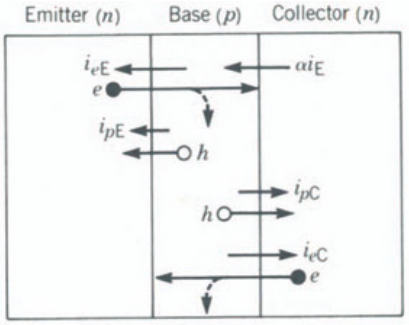


FIGURE 26-21
Schematic representation of the flow of electrons and holes across the two junctions of a transistor, connected in the common emitter configuration, when V_{CE} is less than a few tenths of a volt.

26.5 FIELD-EFFECT TRANSISTORS (FET)

One disadvantage of the bipolar junction transistor is the low resistance of the input circuit; in many cases this is an undesirable feature. For example, if the voltage drop across a resistor is used for V_1 in Fig. 26-19, the current through the resistor will change and so will the magnitude of the voltage across the resistor. The disadvantage of low resistance in the input circuit is illustrated in Problem 27.8. This difficulty is remedied by another type of transistor whose general name is *Field-Effect Transistor* (FET). There are two basic types of FET's: the *Junction Field-Effect Transistor* (JFET) and the *Metal-Oxide-Semiconductor Field-Effect Transistor* (MOSFET).

FET's have two salient features:

1. They are *voltage controlled* devices. The current through the device is controlled by an electric field associated with a voltage placed at an electrode called the *gate*. This feature gives them the name FET.
2. The current is transported by carriers of one polarity only (majority carriers). For this reason, this type of transistor is also called a *unipolar* transistor.

The resistance of a wire filament is given by $R = \rho\, d/A$ (Section 23.1), where $\rho = 1/\sigma = m/q^2N\tau$ (Eq. 23.14), N is the carrier concentration, d is the length of the filament, A is the cross-sectional area of the filament, and τ is the time between collisions. In the FET, the resistance, and therefore the current for a given voltage, is controlled by an input voltage applied at the gate. The input voltage determines either the area A or the concentration N or both. In the JFET, A varies; in the MOSFET, both A and N vary.

26.5a The Junction Field-Effect Transistor (JFET)

A schematic of a JFET and how it can be used in a circuit is shown in Fig. 26-22. The JFET consists of a rod of a semiconductor (for example, n type) to which two metallic ohmic (as opposed to rectifying) contacts, called the *source* and the *drain*, are made. A *heavily doped* p-type semiconductor region, called the *gate*, surrounds the n-type rod.

In its normal mode of operation, the drain is maintained at a positive potential V_d with respect to the source. The gate is maintained at a negative potential V_g with respect to the source. The potential difference between the drain and the source will give rise to a current i_d through the n-type rod. This current will produce an i_dR drop along the rod. As a consequence, the potential along the rod varies from nearly V_d for that part of the rod close to the drain to nearly zero for the part close to the source. The p-n junction formed where

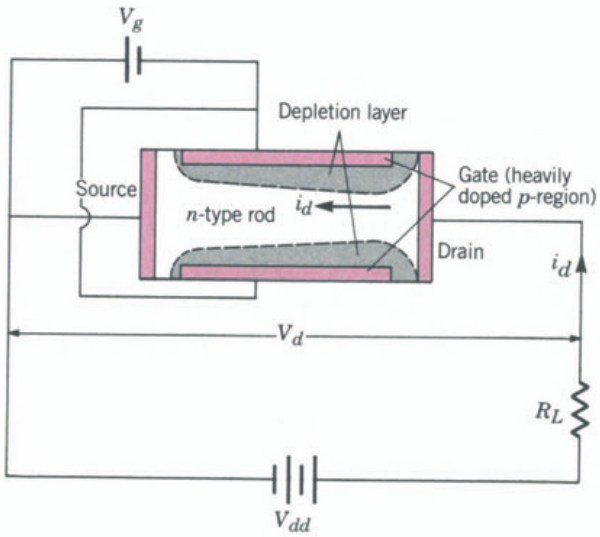


FIGURE 26-22
Structure of a Junction Field-Effect Transistor (JFET). In the figure the JFET is connected in series with a load resistor R_L and the battery V_{dd}. The potential of the gate is the same as (when $V_g = 0$) or lower than that of the source (when $V_g \neq 0$).

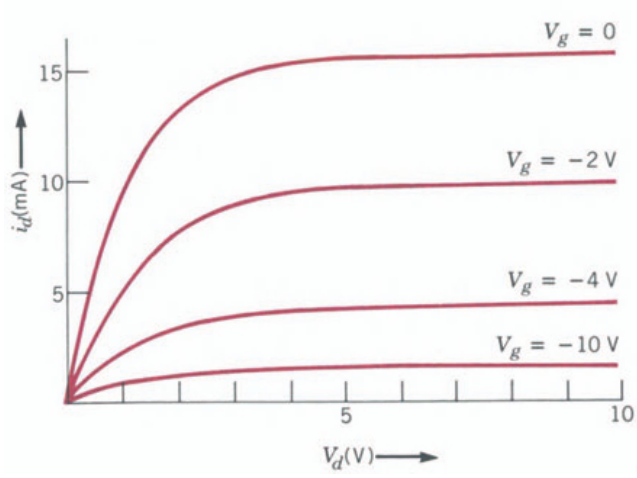


FIGURE 26-23
Output characteristic curves—i_d versus V_d—for several values of the gate voltage V_g for the JFET circuit of Fig. 26-22.

the gate comes in contact with the rod is reverse biased, becoming progressively more so toward the drain. As a result, near the p-n junction there is a depletion layer. Moreover, because the gates are very heavily doped compared to the rod, this depletion layer lies primarily in the n region (in the rod). Because the width of this layer depends on the potential difference across the reverse biased junction, it will be wider close to the drain than close to the source; that is, it will be wedge-shaped. The depletion layer effectively acts as an insulator; this means that i_d is forced to move through a funnel-shaped channel in the n-type rod. Figure 26-23 shows the dependence of i_d versus V_d for several values of the gate voltage, V_g.

Let us consider the case where $V_g = 0$. As V_d begins to increase from zero, the n-type rod behaves as an ohmic device; that is, i_d is proportional to V_d. For small V_ds, the depletion layer is very small and the rod acts simply as a resistor of constant value. However, as V_d increases, the p-n junction at the gate becomes more and more reverse biased; this means that the depletion layer spreads increasingly into the body of the n-type rod and forces the current to flow through a channel that is becoming progressively narrower. This has the effect of increasing the resistance of the rod (because of the decrease of the effective cross-section A). Thus the current i_d does not increase as fast as it did initially when V_d was small. Eventually, the channel becomes so small that the expected increase in i_d due to the increase in V_d is inhibited by the associated increase in R due to the reduction in size of the conducting channel: The current levels off. When this situation occurs, the FET is said to be in the "pinch off" region.

If V_g is at some negative potential to begin with, the width of the depletion layer for a given V_d will be greater than when V_g was zero. This has two effects: (1) the initial resistance of the rod (when V_d was small) will be greater, which explains why the slopes of the curves decrease as V_g becomes more and more negative; and (2) the pinch off will occur at smaller V_d and i_d.

The electronic symbol for an n-channel JFET is shown in Fig. 26-24*a*. A p-channel JFET can be constructed by making the rod of a p-type material and the gates of heavily doped n-type material. The polarity of V_d and V_g has to

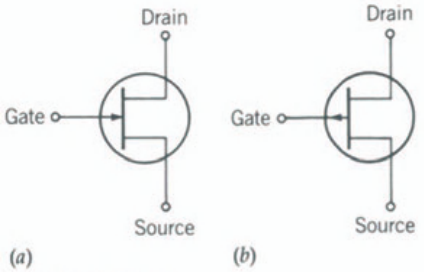


FIGURE 26-24
(*a*) Electronic symbol for an n-channel JFET like the one in Fig. 26-22. (*b*) Electronic symbol for a p-channel JFET.

be reversed. The electronic symbol for a *p*-channel JFET is shown in Fig. 26-24*b*.

At the beginning of our discussion, we indicated that the FET's solved the problem of low input resistance. The input to the FET, whether used as an amplifier element or as a switch, is fed in the gate circuit. In this circuit there is a reverse biased *p-n* junction; therefore, the resistance is very large, typically on the order of $10^{12}\ \Omega$.

26.5b The Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET)

Another type of FET, the metal-oxide-semiconductor field-effect transistor, is illustrated in Fig. 26-25. It consists of two heavily doped *n*-type regions diffused into a *p*-type substrate. The two *n* regions serve as the drain and the source. The gate electrode is a metallic layer deposited on top of an insulating layer of SiO_2. When $V_g = 0$, no current can flow between the drain and the source. The reason is that where the drain contacts the *p*-type substrate we have a reverse biased junction. If a positive potential is applied at the gate, electrons from the *p* substrate are attracted toward the SiO_2-substrate inter-

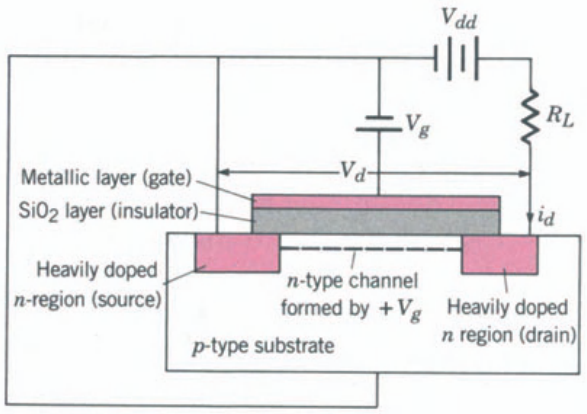


FIGURE 26-25
Structure of a Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET).

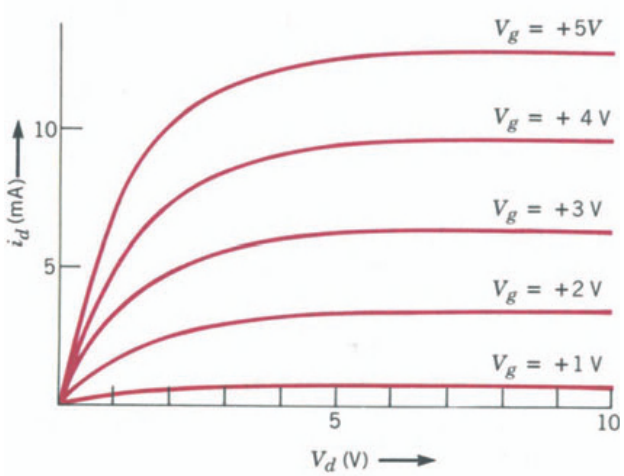


FIGURE 26-26
Output characteristic curves—i_d versus V_d—for several values of the gate voltage V_g for the MOSFET circuit of Fig. 26-25.

face. These electrons will recombine with the holes in the region of the p type substrate near the interface. Eventually a small layer near the interface will become an n-type layer. This provides a channel for the current i_d to flow from the drain to the source when a voltage V_d is applied between them. The greater V_g, the greater the width of the channel and consequently, the greater i_d will be for a given V_d. This is shown in Fig. 26-26, which illustrates the dependence of i_d on V_d for different values of V_g.

In addition to the feature already mentioned (the increase of i_d with V_g), there is also a leveling off of i_d after a certain V_d. The reason is the same as in the case of the JFET. When the n-type channel is formed, we have a p-n junction between the p-type substrate and the n-type channel. A depletion layer is formed that increases as V_d becomes more and more positive because the n-type channel becomes positive (more so near the drain than the source) with respect to the p-type substrate, which is connected directly to the negative terminal of V_{dd}. This depletion layer increases the resistance of the channel and eventually leads to the leveling off of the current i_d.

The electronic symbol for an n channel (p-type substrate) MOSFET and for a p-channel (n-type substrate) MOSFET are shown in Figs. 26-27a and 26-27b, respectively.

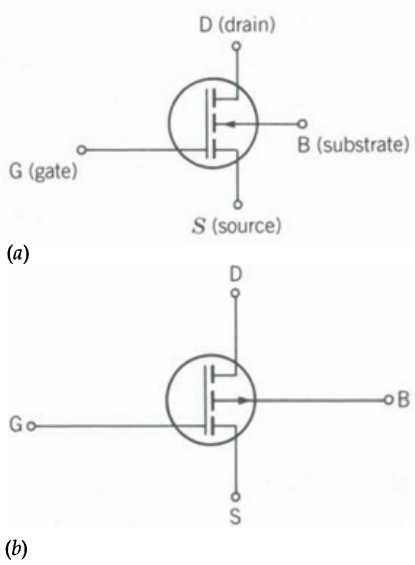


FIGURE 26-27
(a) Electronic symbol for an n-channel (p-type substrate) MOSFET. (b) Electronic symbol for a p-channel (n-type substrate) MOSFET.

PROBLEMS

26.1 The current through a p-n junction is 1×10^{-8} A when a reverse bias voltage of 10 V is applied across the junction at $T = 300$ K. What will be the current through the diode when a forward bias voltage of (a) 0.1 V, (b) 0.3 V, and (c) 0.5 V is applied?

26.2 In the ideal diode the reverse saturation current should be as small as possible. Considering the fact that E_g for Si is 1.1 eV and E_g for Ge is 0.67 eV, which material is better suited for the fabrication of p-n junction diodes?

26.3 The reverse saturation current of a silicon diode is $i_0 = 5 \times 10^{-9}$ A. The voltage across that diode when forward biased is 0.45 V. (a) What is the current through the diode at $T = 27°C$? (b) If the voltage across the diode is held constant, and we assume that i_0 does not change with temperature, what is the current through the diode at $T = 47°C$?

(*Answer*: (a) 0.18 A, (b) 0.06 A.)

26.4 In Problem 26.3 we assumed that the reverse saturation current i_0 remains constant when the temperature changes. (a) Show that this assumption is highly incorrect by calculating i_0 at $T = 47°C$ when $i_0 = 5 \times 10^{-9}$ A at 27°C. Assume that the Fermi level on the p side of the junction is 1 eV below the bottom of the conduction band. (b) If the voltage across the forward biased diode is 0.45 V, as in Problem 26.3, what is the current through the diode at $T = 47°C$?

(*Answer*: (a) 5.6×10^{-8} A, (b) 0.67 A.)

26.5 The reverse saturation current of a silicon diode doubles when the temperature changes from 27°C to 33°C. What is the position of the Fermi level on the p side of the junction?

26.6 In the circuit of Fig. 26-28, the voltage across the resistor R is 4.6 V. What is the reverse saturation current of the silicon diode if the temperature is 27°C?

(*Answer*: 1.8×10^{-10} A.)

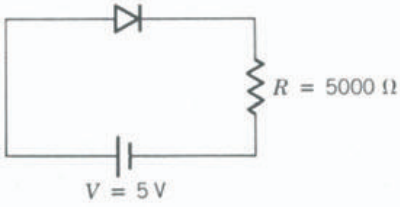


FIGURE 26-28
Problem 26.6.

26.7 (a) What is the voltage across the resistor R of Problem 26.6 if the diode is reversed, as shown in Fig. 26-29? (b) What is the effective resistance of the diode?

(*Answer*: (a) 9×10^{-7} V, (b) 2.8×10^{10} Ω.)

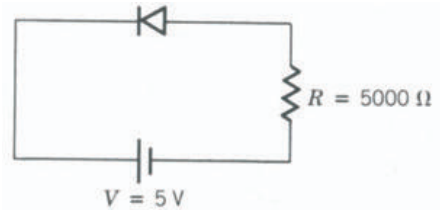


FIGURE 26-29
Problem 26.7.

26.8 A diode is connected in series with a resistor $R = 1000\ \Omega$ and a battery $V = 10$ V (see Fig. 26-30). The reverse saturation current of the diode is $i_0 = 10^{-8}$ A. (a) Assume that the voltage V_0 across the forward biased diode is 0.6 V, what is the current i in the circuit? (b) Use the diode equation, Eq. 26.8, and the answer found in part (a) to evaluate V_0 (assume $T = 300$ K). Does the answer to (b) agree with the assumption in part (a)? (c) Repeat (a) and (b) for the following assumed values of V_0: 0.5 V, 0.4 V, and 0.3 V. (d) What is the approximate value of V_0?

(*Answer*: (a) 9.4×10^{-3} A, (b) 0.356 V, (c) 9.5×10^{-3} A, 0.356 V; 9.6×10^{-3} A, 0.357 V; 9.7×10^{-3} A, 0.357 V, (d) 0.357 V.)

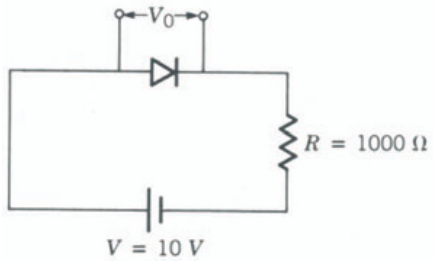


FIGURE 26-30
Problem 26.8.

26.9 In Problem 26.8, by trying different values for the voltage V_0 across the diode, we were able to get an estimate of the actual value. The exact value of V_0 can be found by applying Kirchhoff's loop rule (see Section 15-6) to the circuit of Fig. 26-30,

$$V = V_0 + iR$$

or

$$10\ \text{V} = V_0 + 10^3\ \Omega\ i$$

because the current through the resistor is the same as the current through the diode, i is given by Eq. 26.8, therefore

$$10\ \text{V} = V_0 + 10^3\ \Omega\ i_0 \left[\exp\left(\frac{|e|V_0}{k_B T}\right) - 1\right]$$

This equation must be solved to find V_0 at a given temperature. Unfortunately, this is a transcendental equation that cannot be solved analytically. It can be solved numerically with the aid of a computer. Different values of V_0 are systematically inserted into the equation until a value is found for which the equation holds. Assume that $T = 300$ K and that the reverse saturation current is $i_0 = 10^{-8}$ A. Write a computer program that will yield a value for V_0 within 0.1% of the actual value.

26.10 Use the program of Problem 26.9 to find the voltage across the forward biased diode of Fig. 26-30 when the temperature rises to 320 K. The Fermi level in the p side of the diode is 1 eV below the bottom of the conduction band. Recall (see Problem 26.4) that i_0 varies with temperature.

26.11 An alternating voltage $V = 120 \sin \omega t$ V is applied to the input terminals of the circuit shown in Fig. 26-31. Assume that the forward voltage across the diode is 0.5 V. (a) Sketch the input voltage as a function of time. (b) Sketch the output voltage (voltage across R) as a function of time.

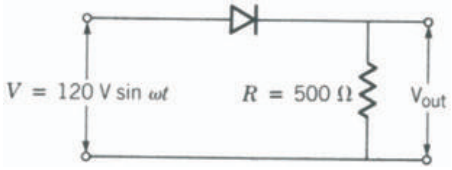


FIGURE 26-31
Problem 26.11.

26.12 The alternating voltage of Problem 26.11 is applied to the input terminals of the circuit shown in Fig. 26-32. Sketch the output voltage V_{out} as a function of time t.

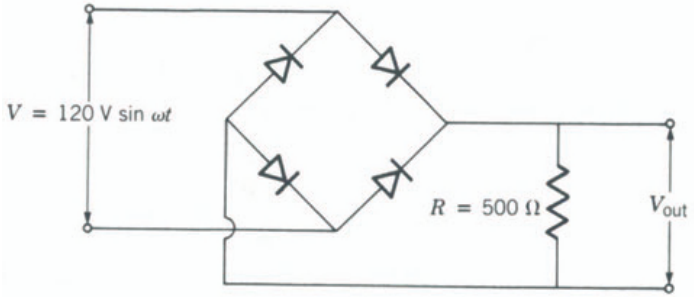


FIGURE 26-32 Problem 26.12.

26.13 Sketch the unbiased and the biased energy diagrams for a *pnp* transistor in the common base configuration.

26.14 Draw a common emitter circuit similar to the one shown in Fig. 26-19 for a *pnp* transistor. Indicate in a figure similar to Fig. 26-21 the different current components crossing the two junctions in the saturation region and in the active region. Assume that the base is very lightly doped compared with the emitter and the collector.

26.15 An *npn* transistor with a gain parameter $\beta = 10$ is connected in the common base circuit shown in Fig. 26-33. An alternating voltage $V_{in} = 2 \sin \omega t$ V is applied to the input terminals. What is the output voltage V_{out}? Assume that resistance of the emitter-base junction is constant and equal to 10 Ω.

(*Answer*: 71 sin ωt V.)

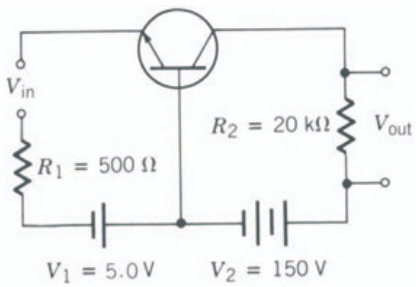


FIGURE 26-33
Problem 26.15.

26.16 Consider the data presented in Fig. 26-20. (a) What is the approximate value of the current gain parameter? (b) What fraction of the electrons injected from the emitter into the base reach the collector?

(*Answer*: (a) 200, (b) 0.995.)

26.17 A 2N2222A transistor having the characteristics shown in Fig. 26-20 is used in a common emitter circuit (see Fig. 26-19). The base resistor $R_1 = 20\ k\Omega$, $V_1 = 3$ V, and the voltage across the emitter-base junction is 0.4 V. (a) Calculate the base current i_B. (b) Use the current gain parameter found in Problem 26.16 to obtain the collector current i_C. (c) If $R_2 = 1\ k\Omega$ and $V_2 = 35$ V, what is the collector-emitter voltage V_{CE}?

(*Answer*: (a) 130 μA, (b) 26 mA, (c) 9 V.)

26.18 The width of the depletion layer on the *p* side of a *p-n* junction is given by

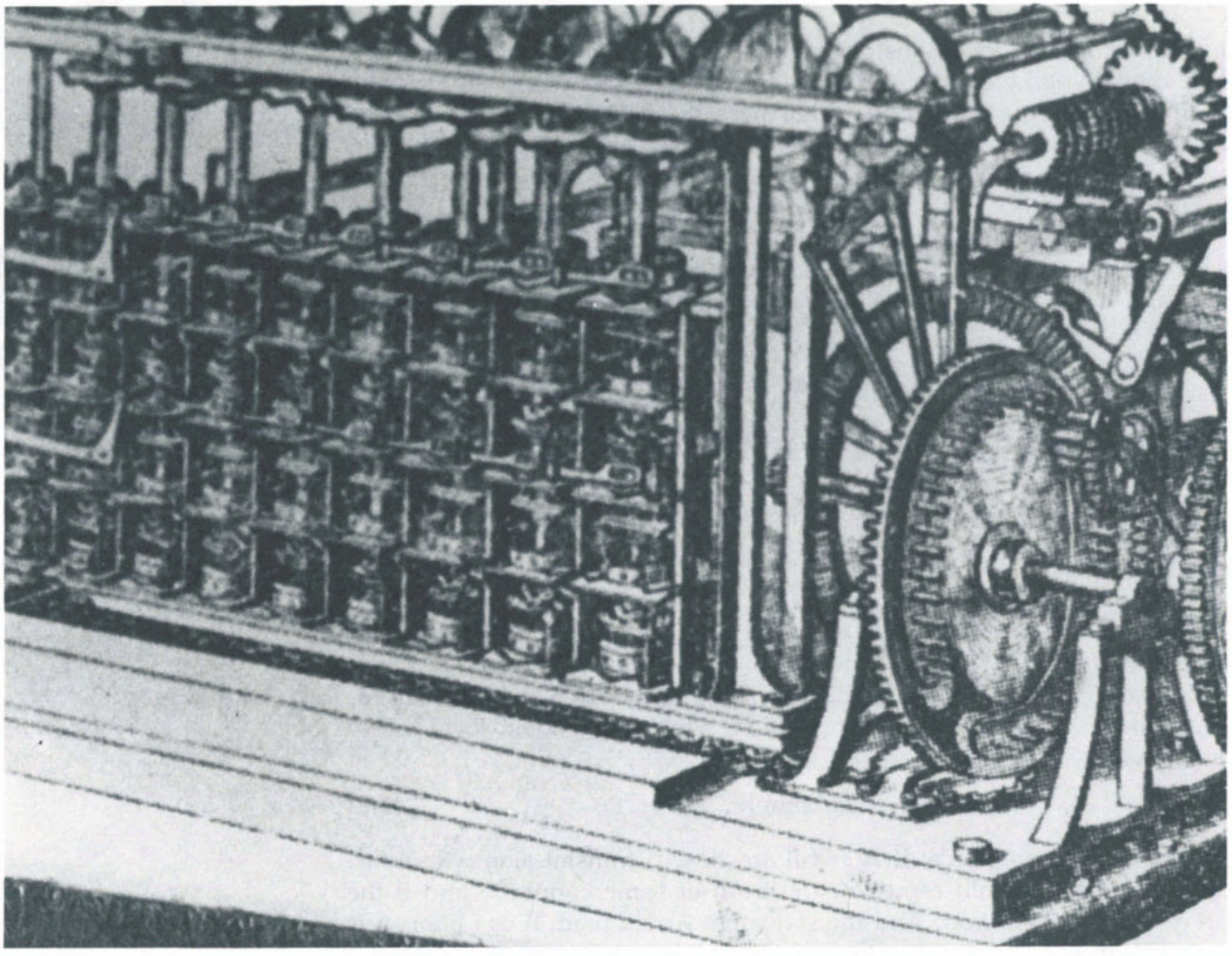
$$x_p = \left[\frac{2\kappa\epsilon_0 (V_c - V_0)}{e N_A \left(1 + \frac{N_A}{N_D}\right)} \right]^{1/2}$$

where κ is the dielectric constant of the semiconductor (for Si, $\kappa = 12$), $\epsilon_0 = 8.85 \times 10^{-12}$ farad/m is the permitivity of free space, V_c is the equilibrium contact potential, V_0 is the external bias voltage of the junction (taken as positive when forward biased and negative when reverse biased), and N_A and N_D are the impurity concentrations in the *p* region and in the *n* region, respectively. Calculate the width of the depletion layer on the *p* side of a junction when (a) $V_0 = 0$ V, and (b) $V_0 = -10$ V. Assume $V_c = 0.6$ V, $N_A = 10^{20}\ m^{-3}$, and $N_D = 10^{24}\ m^{-3}$.

(*Answer*: (a) 2.82×10^{-6} m, (b) 11.86×10^{-6} m.)

26.19 An *npn* silicon transistor with an $\alpha = 0.990$ when unbiased is used in a common emitter circuit similar to the one in Fig. 26-19. The width of the base is 30 μm. Let us assume that the number of electrons injected from the emitter into the base that recombine with holes in the base is proportional to the width of the "effective base" (the nondepleted region). (a) What will be the value of α when $V_{CE} = 10$ V? (b) What will be the change in β as V_{CE} varies from 0 V to 10 V? Use the doping levels and the equilibrium contact potential of Problem 26.18.

(*Answer*: (a) 0.993, (b) β (0 V) = 99, β (10 V) = 142.)



CHAPTER 27

Some Basic Logic Circuits of Computers

27.1 INTRODUCTION

The complex logical operations performed by any digital instrument, be it a simple digital voltmeter or a sophisticated computer, are based on a system of symbolic concepts known as *boolean algebra*, named after George Boole, who began publishing the algebra of logic in 1847. In a digital instrument, these symbolic concepts are implemented by means of electronic circuits called *logic circuits* or *gates*.

In this chapter we will present a brief introduction to some of the principles of boolean algebra. We will then show how the semiconductor devices, diodes and transistors discussed in the preceding chapter, are used to construct the gates that implement this algebra.

George Boole (1815–1864).

27.2 RUDIMENTS OF BOOLEAN ALGEBRA

The three basic logic operations of boolean algebra are named AND, OR, and NOT.

AND Operation

Consider two switches A and B wired in series with a battery and a load resistor R as in Fig. 27-1. Current will flow if both switches A *and* B are closed. These switches perform the AND operation, which in symbolic logic is represented as

$$A \cdot B = T \quad \text{or simply} \quad AB = T \tag{27.1}$$

The statement above reads: If A *and* B are closed, transmission will occur. More generally, A could represent the truth of some condition and B the truth of another condition. And the statement would read: If condition A is satisfied (is true) *and at the same time* condition B is satisfied, then the desired result will be obtained and the output will be true (T).

It is customary to represent a condition that is TRUE by 1 and a condition that is NOT TRUE by 0. A way to represent all possible combinations of the states for all the variables involved is by means of a *truth table*. The truth table for the AND operation with two variables A and B and the result T is shown in Table 27-1. In this table the symbol 0 and 1 under A and B mean open and

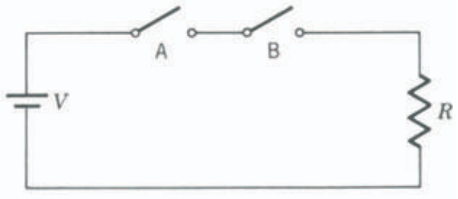


FIGURE 27-1
Implementation of the AND logic function with two manual switches.

closed, respectively, while these symbols under T mean not true and true, respectively.

TABLE 27-1
Truth Table for the AND Operation

A	B	T
0	0	0
0	1	0
1	0	0
1	1	1

OR Operation

If two switches A and B are wired in parallel with each other, and this parallel combination is then connected in series with a battery and a load resistor R, as in Fig. 27-2, transmission will occur if either A or B is closed. The circuit will perform the OR operation, which in symbolic logic is represented as

$$A + B = T \tag{27.2}$$

This statement means that the desired result is true if either condition A *or* condition B *or* both are true. The truth table for the OR operation with two variables A and B is shown in Table 27-2.

TABLE 27-2
Truth Table for the OR Operation

A	B	T
0	0	0
0	1	1
1	0	1
1	1	1

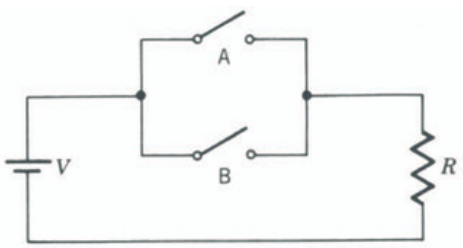


FIGURE 27-2
Implementation of the OR logic function with two manual switches.

NOT Operation

The NOT operation simply inverts the logic condition. It is written as $\overline{A}$, which reads NOT A. Thus if A stands for "the sun will shine tomorrow," then $\overline{A}$ stands for "the sun will not shine tomorrow." As an example of a truth table using the NOT concept, consider the following boolean equation.

$$A \cdot \overline{B} = T \tag{27.3}$$

Equation 27.3 means that if condition A is satisfied and at the same time condition B is *not* satisfied, the output will be true. The truth table for such an equation is given in Table 27-3.

TABLE 27-3
Truth Table for $A \cdot \overline{B} = T$

A	B	T
0	0	0
0	1	0
1	0	1
1	1	0

27.2a Some Theorems in Boolean Algebra

Boolean algebra allows us to manipulate these three basic concepts in order to simplify statements of logic and the circuitry necessary to implement them. We list here some important theorems. As before, + means *or* and a product means *and*

1. Commutation theorems

$$A + B = B + A$$

$$AB = BA$$

2. Association theorems

$$A + (B + C) = (A + B) + C$$

$$A(BC) = (AB)C$$

3. Distribution theorems

$$A + (BC) = (A + B)(A + C)$$

$$A(B + C) = AB + AC$$

4. Absorption theorems

$$A + AB = A$$

$$A(A + B) = A$$

...

5. DeMorgan theorems

$$\overline{A + B} = \overline{A} \cdot \overline{B}$$

$$\overline{AB} = \overline{A} + \overline{B}$$

These theorems can be verified by simple intuition or by writing the appropriate truth tables. As an example, let us consider the second of the DeMorgan theorems, $\overline{AB} = \overline{A} + \overline{B}$.

Because $\overline{AB}$ stands for the inverse of AB, and the output of AB is 1 only when A is 1 and B is 1, the output of $\overline{AB}$ will be 0 only when A = 1 and B = 1. The truth table for $\overline{AB}$ = T is shown in Table 27-4.

TABLE 27-4
Truth Table for $\overline{AB}$ = T

A	B	T
0	0	1
0	1	1
1	0	1
1	1	0

A comparison of the truth tables for $\overline{AB}$ (Table 27-4) and for AB (Table 27-1) shows that the outputs are inverted.

By the definition of the OR concept, $\overline{A} + \overline{B}$ will have an output 1 if either A = 0 or B = 0 or both A and B are 0. The truth table for $\overline{A} + \overline{B}$ = T is therefore identical to that for $\overline{AB}$ = T (Table 27-4), and we conclude that $\overline{AB} = \overline{A} + \overline{B}$.

27.3 ELECTRONIC LOGIC CIRCUITS

The logic concepts introduced in Section 27.2 can be implemented with electronic circuits where the signals have two different values. For example, state 1 could be represented by a signal of 5 V and state 0 by a signal of 0 V. The most commonly used signal levels are 5 V and 0 V or 10 V and 0 V, although the exact value of the input and output voltages are not critical, as long as they are clearly distinguishable. A *logic circuit* is one with an output signal that is the logical function of the inputs. These circuits are called *gates*.

27.3a AND Gate

The AND gate is one with inputs and output levels corresponding to the truth table 27-1 for the AND function. It may have several input terminals, one for each condition of A, B, and so on, and one output terminal for the result T. Only when all the inputs are at logic level 1 (5 V) will the output be a logic level 1. The electronic symbol for an AND gate with two inputs A and B is shown in Fig. 27-3.

FIGURE 27-3
Electronic symbol for a two-input AND gate.

Before we consider how an AND gate can be built, let us introduce a convention that will be used often in this chapter. When we specify the potential of a point in a circuit, we must always specify the reference point, that is, the point whose potential we arbitrarily call zero. It is common practice to call this reference point the *ground*. In household wiring, the ground is obtained by making a tight connection to a water pipe made of copper. In an electronic circuit the ground is the metal chassis of the instrument that is assumed to be equipotential throughout, that is, to have no resistance. The electronic symbol used to represent the ground is shown in Fig. 27-4. When we say that an input or an output signal is 5 V, we mean that the input or output voltage is 5 V above ground. Similarly, an input of 0 V means that the input connection is at the same potential as the chassis ground; we often use the expression "it is grounded."

FIGURE 27-4
Electronic symbol for the ground.

In today's computers, the logic circuits are built with semiconductor diodes and transistors. In the early days of the computer era, relay switches and vacuum tubes were widely used in the construction of computers. Because they are easier to understand, we will first show how the gates can be built with relay switches.

A relay switch is a remotely controlled switch. A schematic representation of one type of relay switch is shown in Fig. 27-5. A current through a coil produces a magnetic field that attracts one end of a pivoted lever toward the coil. A switch contact mounted on the lever provides a movable contact that can connect to either of two stationary contacts. When no current flows through the coil, a spring attached to the other end of the lever keeps the lever in the position shown in Fig. 27-5 and, as a result, the movable contact is connected to the upper stationary contact. If one end of the coil is grounded and an input voltage is applied to the other end, a current will be set through the coil. The end of the lever above the top of the coil will be attracted downward, and the movable contact will be connected to the lower stationary contact. The electronic symbol used to represent a relay switch is shown in Figs. 27-6*a* and 27-6*b*. In Fig. 27-6*a*, the coil is not activated (no current); as a result, the movable contact is connected to the upper stationary contact. In Fig. 27-6*b*, the coil is activated; consequently, the movable contact is connected to the lower stationary contact. We will assume that a 5-V input will produce the necessary current to have the coil attract the lever.

An AND gate built with relay switches is shown in Fig. 27-7. It has two inputs, V_A and V_B, which represent two logic conditions A and B. The output

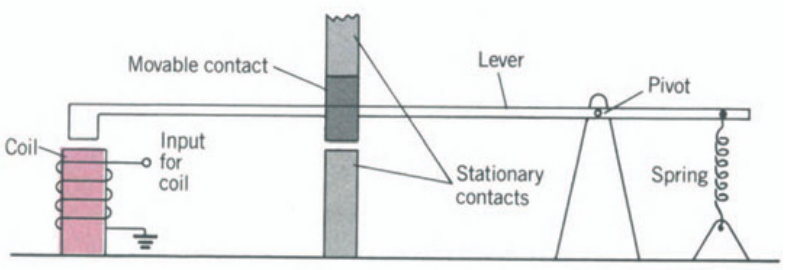


FIGURE 27-5
Relay switch.

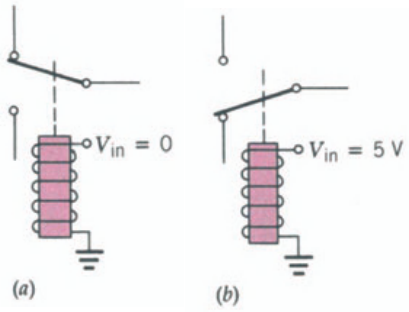


FIGURE 27-6
Electronic symbol for a relay switch. (*a*) The coil is not activated (no current through the coil) resulting in the connection of the movable contact with the upper stationary contact. (*b*) The coil is activated, the movable contact is connected to the lower stationary contact.

is the potential of the movable contact of switch 2. The two upper stationary contacts are grounded. The lower stationary contact of switch 1 is connected to a 5-V source (note that this 5-V signal does not represent an input condition), while the lower stationary contact of switch 2 is connected to the movable contact of switch 1. Suppose $V_A = V_B = 5$ V. Both coils will be activated, the movable contacts of both switches will be connected to their lower stationary contact, and thus $V_{out} = 5$ V. If $V_B = 0$ V, then the movable contact of switch 2 will be connected to its upper stationary contact, which is grounded; thus, regardless of whether $V_A = 5$ V or 0 V, $V_{out} = 0$ V. When $V_A = 0$ V and $V_B = 5$ V, the output through the movable contact of switch 2 is connected to the lower contact of switch 2, which in turn is connected to the movable contact of switch 1, and because switch 1 is not activated, its movable contact is connected to its upper contact, and therefore grounded, $V_{out} = 0$ V. Therefore, only when both V_A and V_B are 5 V (logic level 1) will the output be 5 V. The circuit in Fig. 27-7 implements the AND logic function.

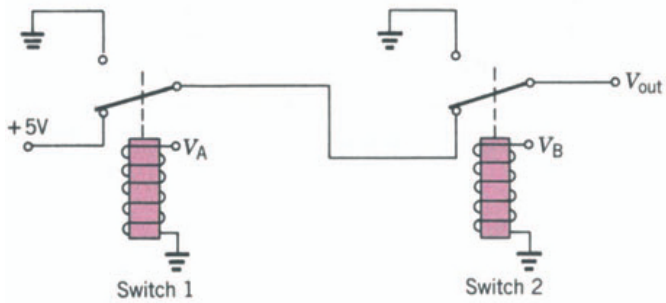


FIGURE 27-7
A two-input AND gate made with relay switches.

27.3b OR Gate

The OR gate has a logic level 1 output when any one input or all the inputs are at logic level 1 (see Table 27-2). The electronic symbol used to represent the OR gate is shown in Fig. 27-8. An OR gate built with relay switches with two inputs V_A and V_B is schematically represented in Fig. 27-9. The output is again the potential of the movable contact of switch 2. If both V_A and V_B are 5 V, then the output is connected through the lower contact of switch 2 to the 5-V source; that is, $V_{out} = 5$ V. The same is true if $V_A = 0$ V

FIGURE 27-8
Electronic symbol for a two-input OR gate.

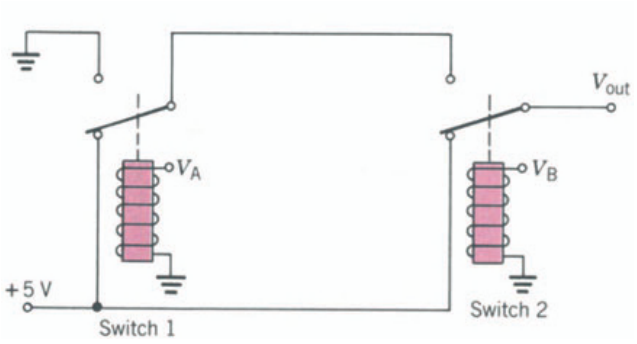


FIGURE 27-9
A two-input OR gate constructed with relay switches.

and $V_B = 5$ V. If $V_A = 5$ V and $V_B = 0$ V, the output is connected through the upper contact of switch 2 and the movable and lower contacts of switch 1 to the 5-V source, V_{out} is still 5 V. Finally, if $V_A = V_B = 0$ V, the output will be connected, through the two movable contacts and the two upper contacts, to ground and $V_{out} = 0$ V. Thus, if either V_A or V_B or both are 5 V (logic level 1), the output will also be 5 V. Only when both inputs are 0 V (logic level 0) will the output be 0 V. The circuit in Fig. 27-9 implements the OR logic function illustrated in Table 27-2.

27.3c The Inverter

To complete the list of circuits needed to perform the basic logic operations, the NOT function is implemented by an *inverter*. An inverter is a circuit with an output logic level opposite that of the input. The electronic symbol of an inverter is shown in Fig. 27-10.

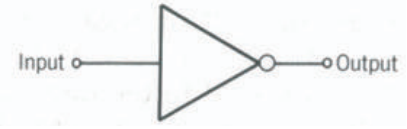


FIGURE 27-10
Electronic symbol for the inverter, a circuit that performs the NOT logic function.

A relay switch circuit that implements the NOT function is shown in Fig. 27-11. It is seen that if the relay is not activated ($V_A = 0$ V) the movable contact will be connected to the upper contact and as a consequence the output, that is, the potential of the movable contact will be 5 V. An input of 5 V in the relay switch will ground the output, thereby inverting the logic signal.

27.4 SEMICONDUCTOR GATES

We have implemented the different logic functions with relay switches because they are easy to understand. The early electronic computers were built with relay switches. However, in modern technology various semiconductor devices are used for switches. Several examples follow.

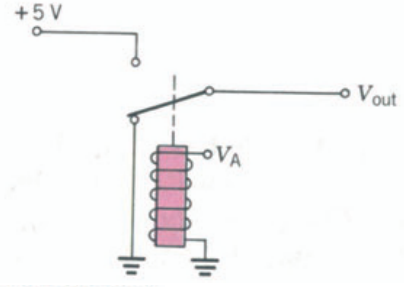


FIGURE 27-11
An inverter constructed with a relay switch.

27.4a Diode Switch

A diode may be used as an electronic switch in the circuit shown in Fig. 27-12*a*. The circuit has been redrawn in more conventional form in Fig. 27-12*b*. Because, as indicated earlier, potential differences are measured with

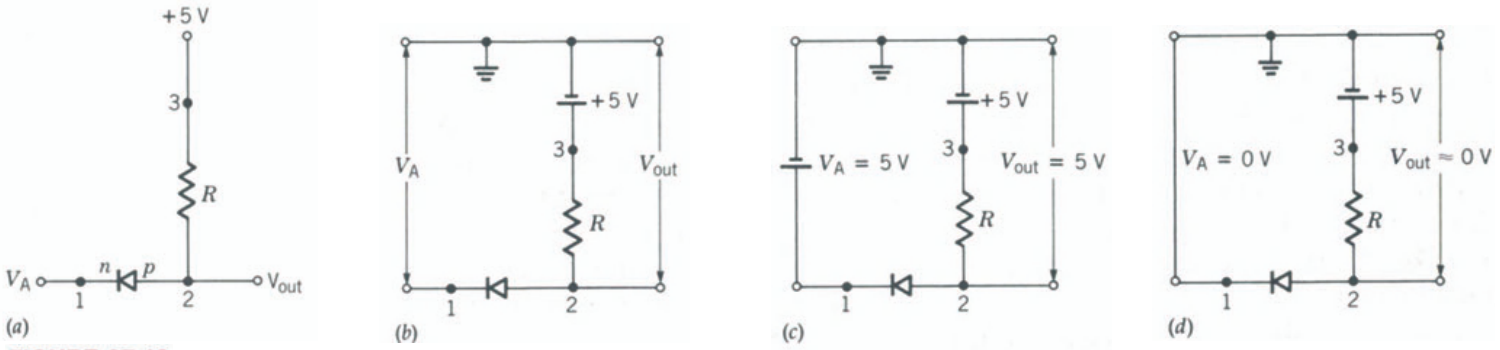


FIGURE 27-12

An electronic diode switch. (*a*) Abbreviated diagram of the circuit. (*b*) Circuit redrawn in more conventional form. (*c*) The input V_A is 5 V; this is done by connecting the input terminal to the positive terminal of a 5-V battery. (*d*) The input V_A is 0 V; this is done by connecting the input terminal to ground.

respect to ground, in Fig. 27-12*b* we have represented the + 5-V source of Fig. 27-12*a* as a battery with its positive terminal connected to point 3 and its negative terminal connected to ground. The output voltage will be the potential difference between the *p* side of the diode and ground. The input signal will be a voltage applied between the *n* side of the diode and ground. An input of 5 V can be represented by a battery with the positive terminal connected to the *n* side of the diode and the negative terminal to ground, as in Fig. 27-12*c*. An input of 0 V places the *n* side of the diode at ground potential; therefore, we can represent this by connecting the *n* side to ground, as in Fig. 27-12*d*.

Let us consider the situation shown in Fig. 27-12*c*, $V_A = 5$ V. We note that the potential of points 1 and 3 is the same, 5 V. Consequently, there will be no current from point 3 to point 1. If we assume that the voltmeter used to measure V_{out} has a very high input resistance (ideally infinite), there will be no current through the resistor *R*. This in turn means that both sides of the resistor, points 2 and 3, are at the same potential (recall that the potential drop across a resistor is *iR*). Because the potential of point 3 is 5 V, V_{out}, the potential of point 2, will also be 5 V.

Let us now consider the case $V_A = 0$ V, as in Fig. 27-12*d*. In this case the diode is forward biased because the *n* side of the diode is connected directly to ground whereas the *p* side is connected through the resistor *R* to the 5-V source. The forward biased diode will conduct and, as we saw in Section 26.3c, the voltage across the diode (between point 2 and the grounded point 1) will be a few tenths of a volt, 0.6 V or less is a typical value for a silicon diode. Thus, when $V_A = 0$ V, V_{out} (the potential of point 2) will be approximately 0. Note that now most of the 5-V drop between points 3 and 1 occurs across resistor *R*, not across the diode.

The circuit of Fig. 27-12*a* is an electronic switch because an input signal of 5 V turns the output "on" (5 V), whereas an input of 0 V turns the output "off" (approximately 0 V).

27.4b Diode AND Gate

A diode AND gate with two inputs V_A and V_B to represent two logic conditions can be constructed with two diode switches like the one discussed in the previous section, see Fig. 27-13*a*. The circuit consists of two diode switches connected in parallel with the resistor R and the 5-V source common to both. This is readily seen if we redraw the circuit in more conventional form as in Fig. 27-13*b*. As we showed in the preceding section, a particular diode will conduct (will be forward biased) if its corresponding input is grounded, that is, 0 V, and will not conduct if the input is 5 V. Thus, when $V_A = V_B =$ 5 V (Fig. 27-13*c*), neither diode will conduct; as a consequence there will be no current through the resistor R and this in turn means that both sides of R, points 3 and 2, are at the same potential. Because the potential of point 3 is 5 V, the output V_{out}, which is the potential of point 2, will also be 5 V, that is, $V_{out} = 5$ V. Thus when both inputs are at logic level 1, the output will also be at logic level 1. If $V_A = 0$ V and $V_B = 5$ V (Fig. 27-13*d*), diode A will be forward biased and will conduct. As a result, the potential of the p side

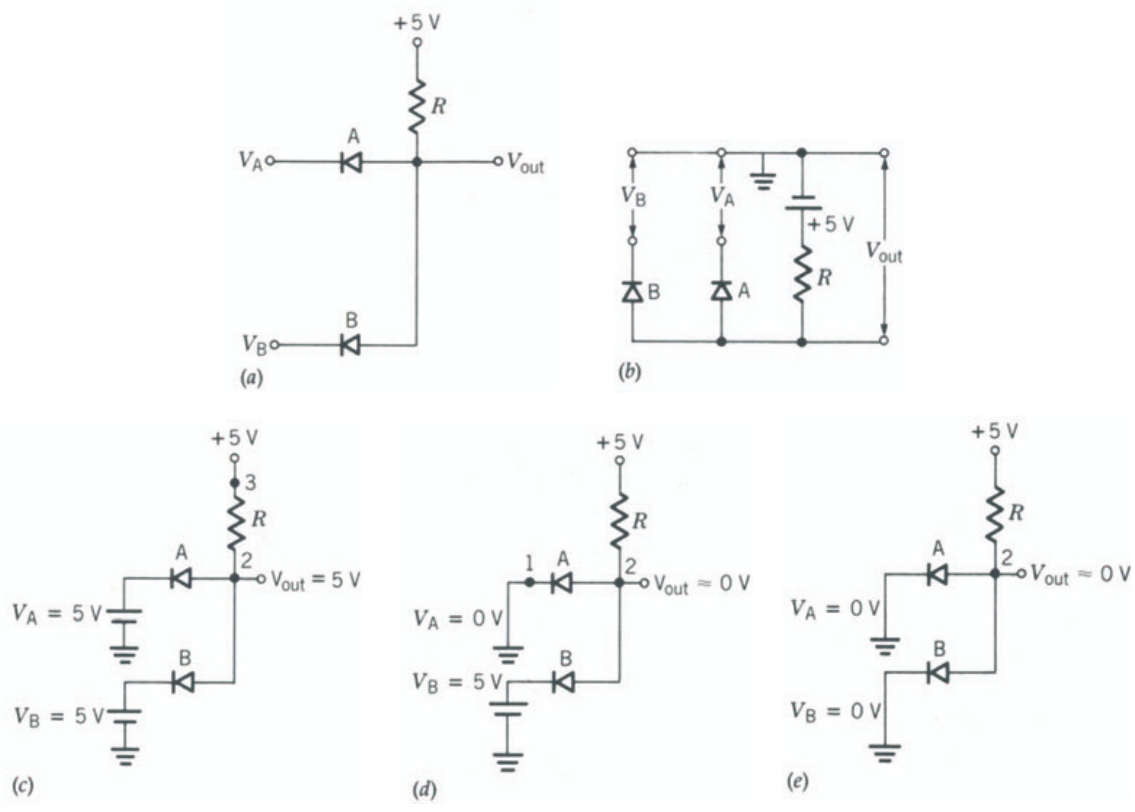


FIGURE 27-13
Diode AND gate. (*a*) Abbreviated diagram of the circuit. (*b*) Circuit redrawn in more conventional form. (*c*) The inputs to both diodes are connected to the positive terminal of 5-V batteries to represent the condition $V_A = 5$ V, $V_B = 5$ V. (*d*) The input to diode B is connected to the positive terminal of a 5-V battery, whereas the input to diode A is grounded to represent the condition $V_B = 5$ V, $V_A = 0$ V. (*e*) Both inputs are grounded, $V_A = 0$ V, $V_B = 0$ V.

of diode A (point 2) will be 0.6 V or less higher than the n side of diode A, which is grounded. V_{out} will be approximately 0 V. Note that the p side of diode B is connected to the p side of diode A; therefore its potential will also be 0.6 V above ground; because the potential of the n side of diode B is 5 V, we see that diode B is reverse biased and therefore does not conduct. Similar conclusions can be drawn if $V_A = 5$ V and $V_B = 0$ V. Diodes A and B reverse roles, but the output voltage will still be the voltage of the forward biased diode, diode B in this case—that is, $V_{out} \approx 0$ V. Finally, if $V_A = V_B = 0$ V (Fig. 27-13*e*), both diodes will be forward biased and $V_{out} \approx 0.6$ V. Thus, if either or both inputs are at logic level 0, the output will be at logic level 0. The circuit in Fig. 27-13 implements the AND logic function.

27.4c Diode OR Gate

The circuit of two diodes in Fig. 27-13 can be rearranged to create an OR gate. Consider the circuit in Fig. 27-14. The circuit has two inputs, V_A and V_B, to represent two logic conditions. Note that now the input signals are applied to the p side of the diodes and the output is the potential of the n side of the diodes. The 5-V source has been removed, and the load resistor is connected directly to ground.

If $V_A = V_B = 5$ V (Fig. 27-14*b*), the diodes will be forward biased because their p side will be directly connected to the positive terminal of the input sources and therefore will be at a higher potential than the n side. Being forward biased, the potential difference across the diodes will be approximately 0.6 V. Because the potential of the p side is 5 V, the potential of the n side will be 5 V − 0.6 V = 4.4 V. That is, $V_{out} \approx 5$ V or logic level 1. If $V_A = 5$ V and $V_B = 0$ V (Fig. 27-14*c*), diode A will be forward biased and for the same reason as before the potential of its n side will be approximately 5 V (4.4 V); that is, the output will be at logic level 1. We should note that diode B is now reverse biased because the potential of the n side is 4.4 V whereas the p side is grounded. Similar conclusions can be drawn if

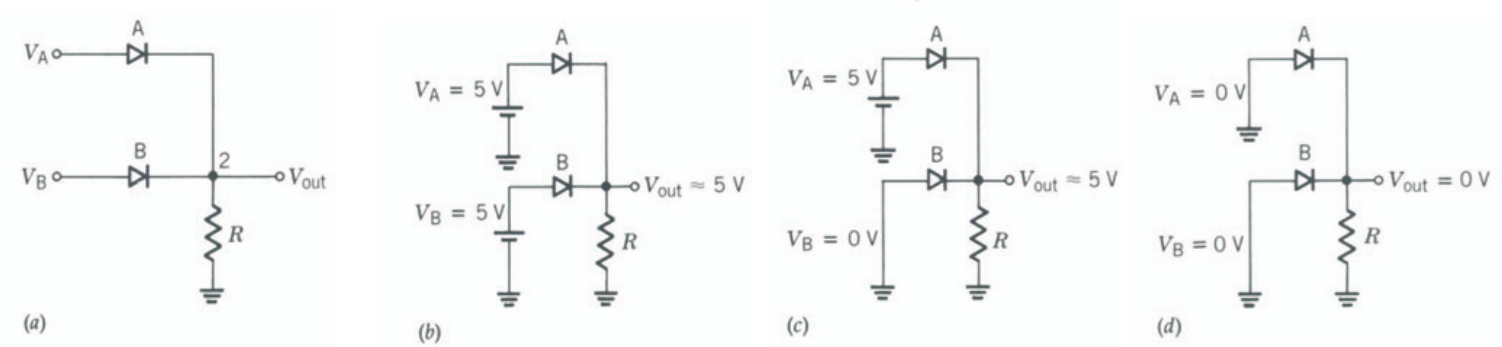


FIGURE 27-14
Diode OR gate. (*a*) Abbreviated diagram of the circuit. (*b*) The inputs to both diodes have been connected to the positive terminal of 5-V batteries to represent the condition $V_A = 5$ V, $V_B = 5$ V. (*c*) The input to diode A is connected to the positive terminal of a 5 V battery, whereas the input to diode B is grounded to represent the condition $V_A = 5$ V, $V_B = 0$ V. (*d*) Both inputs are grounded, and therefore $V_A = V_B = 0$ V.

$V_A = 0$ V and $V_B = 5$ V. Thus if either or both inputs are at logic level 1, the output will also be at logic level 1. If $V_A = V_B = 0$ V (Fig. 27-14*d*), there will be no current through any part of the circuit because there are no sources of potential differences in the circuit. If there is no current through the resistor, both sides of R must be at the same potential. One side of R is grounded; therefore the other side that is connected to the output will also be at ground potential, $V_{out} = 0$ V. The circuit in Fig. 27-14 implements the OR logic function.

27.4d The Inverter

As discussed in Section 27.3*c* and shown in Fig. 27-11, an inverter can be used as a NOT circuit. The transistor can replace the switch to make such a circuit. Figure 27-15*a* shows a diagram in which the transistor is wired in its common emitter configuration considered in Section 26.4c. V_{CE} represents the voltage between the collector and the emitter, that is, the potential drop across the transistor. Figure 27-15*b* gives the current-voltage characteristics for that configuration. The curves in Fig. 27-15*b* give the current through the collector, i_C, as a function of the voltage between the collector and the emitter, V_{CE}, for different values of the base current i_B.

Suppose that V_2 and R_C are fixed. Using Kirchhoff's second rule (Section 15.6), we see that

$$V_2 = R_C i_C + V_{CE}$$

which can be written as

$$i_C = \frac{V_2}{R_C} - \frac{V_{CE}}{R_C} \tag{27.4}$$

A plot of i_C versus V_{CE} gives the straight line in Fig. 27-15*b*, called the *load line*, whose slope is determined by R_C. We can plot the load line by locating two points, for example, the intercepts of the line with vertical and horizontal axes. The vertical axis intercept is obtained by setting $V_{CE} = 0$ in Eq. 27.4; it corresponds to $i_C = V_2/R_C$. The horizontal intercept corresponds to $i_C = 0$, and from Eq. 27.4 occurs when $V_{CE} = V_2$. In Fig. 27-15*b* we have chosen $V_2 = 5$ V. For a given V_2 and R_C the values of i_C and V_{CE} must satisfy the

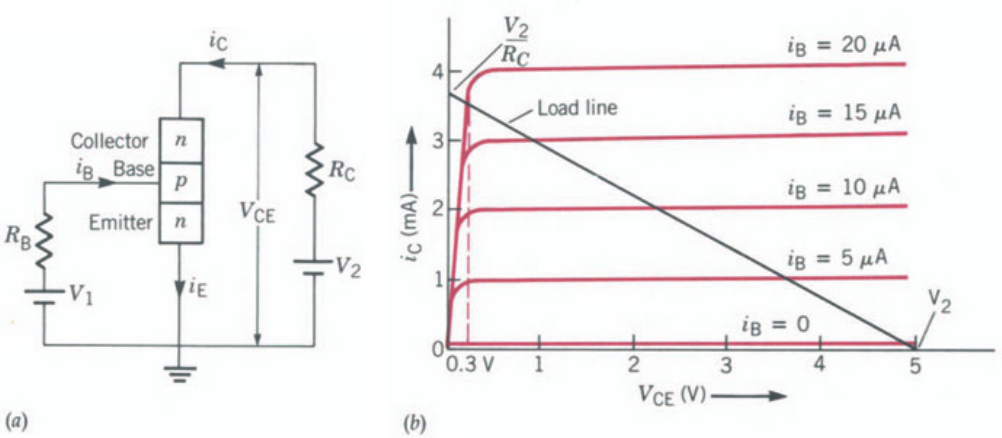


FIGURE 27-15
(*a*) An *npn* transistor in a common-emitter configuration circuit. (*b*) The output characteristic curves—i_C versus V_{CE}—for different values of i_B for the common-emitter circuit of (*a*). The load line corresponds to $V_2 = 5$ V and $R_C = 1350\ \Omega$.

load line equation, that is, the value of i_C and corresponding value of V_{CE} must be such that the point lies on the load line. At the same time, for a given i_B the relation between i_C and V_{CE} is given by the characteristic curve corresponding to that value of i_B. This means that in a given situation (a given V_2, R_C, and i_B), the actual value of i_C and V_{CE} is given by the intercept of the load line and the appropriate characteristic curve. Thus, by changing the value of i_B (which is controlled by V_1 and R_B), from 0 μA to $\geq$ 20 μA, V_{CE} can be made to switch from V_2 (5 V) to $\approx$ 0 V. This is seen in Fig. 27-15*b*, where the load line intersects the $i_C - V_{CE}$ curve at 5 V when $i_B = 0$ and at $V_{CE} = 0.3$ V ($\approx$ 0 V) when $i_B = 20$ μA. In Fig. 27-16, the circuit of Fig. 27-15*a* has been redrawn symbolically. V_{out} is the collector-emitter voltage difference V_{CE}, whereas V_{in} corresponds to V_1 in the preceding diagram. Let us assume that R_B is such that when $V_{in} = 5$ V, $i_B \geq 20$ μA. If $V_{in} = 0$ V, $i_B = 0$ μA and the intercept of the load line of Fig. 27-15*b* and the appropriate characteristic curve corresponds to V_{out} (V_{CE}) $\approx$ 5 V. Thus, if the input is at logic level 0, the output will be at logic level 1. If $V_{in} = 5$ V, $i_B \geq 20$ μA and the intercept of the load line and the characteristic curve corresponds to $V_{out} \approx 0$ V. Thus, if the input logic level is 1, the output logic level is 0. This circuit performs the NOT logic function. It should be noted that because R_B controls the base current and R_C determines the slope of the load line, they must be properly chosen so that the two logic levels be properly distinguishable. That is, the intercept of the load line and the appropriate characteristic curve when $V_{in} = 5$ V must correspond to a V_{CE} close to 0 V.

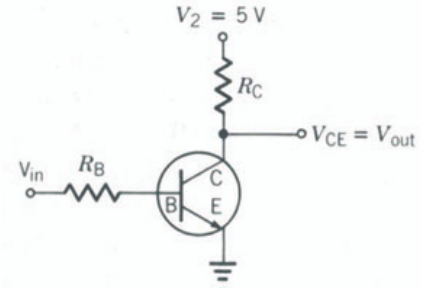
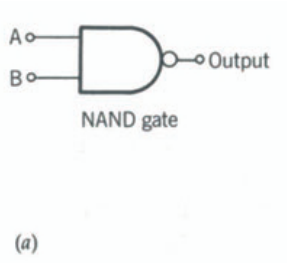


FIGURE 27-16
Symbolic schematic of the circuit in Fig. 27-15*a*. The output voltage V_{out} is the inverse of the input voltage V_{in}; that is, when $V_{in} = 0$ V, $V_{out} = 5$ V and when $V_{in} = 5$ V, $V_{out} = 0$ V. The circuit implements the NOT logic function.

27.5 NAND AND NOR GATES

Once we have the three basic logic circuits, we can combine them to form other more sophisticated ones. Two gates that are often used in *integrated circuits* (IC) are the NAND and the NOR gates.

If we take the output signal from the diode AND gate of Fig. 27-13*a* and connect it to the input of the transistor inverter circuit of Fig. 27-16, we will have a final output that is the inverse of the simple AND gate output. The resulting circuit is one of the family of *diode transistor logic* (DTL) gates. This type of gate will implement the function $T = \overline{AB}$, that is, NOT-AND (NAND). The electronic symbol for a NAND gate is shown in Fig. 27-17*a*; the actual



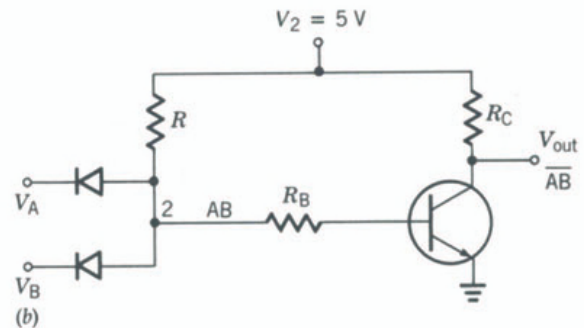


FIGURE 27-17
(*a*) Electronic symbol of a NOT-AND or NAND gate. (*b*) A DTL circuit that implements the NAND logic function.

circuit is illustrated in Fig. 27-17*b*. It is seen that the left side of the circuit in Fig. 27-17*b*, the section containing V_2, the resistor R, the two diodes, and the two inputs V_A and V_B, is the diode AND circuit of Fig. 27-13*a*. The output of that section is AB and appears at point 2. This output becomes the input for the right side of the circuit, the section consisting of V_2, the resistors R_C and R_B, and the transistor. This section is the transistor inverter of Fig. 27-16. Thus, whatever input (in this case AB) is fed into the base of the transistor through the base resistor R_B appears inverted at the collector of the transistor; that is, the final output V_{out} (V_{CE}) = $\overline{AB}$.

A DTL gate that performs the NOT-OR, or NOR operation, can be constructed by connecting the output of the diode OR gate of Fig. 27-14*a* to the input of the transistor inverter, Fig. 27-16. The final output will be $T = \overline{A + B}$ (see Problem 27.3).

27.6 OTHER GATES, RTL AND TTL

Another family of gates often found in an IC is the *resistor transistor logic* (RTL) gate. Figure 27-18*a* is an example of such a gate, which implements the NOT-OR or NOR function. The electronic symbol for a NOR gate is shown in Fig. 27-18*b*. The circuit of Fig. 27-18*a* is essentially two inverters (Fig. 27-16) wired in parallel with V_2 and the collector resistor R_C common to both. In Fig. 27-16 we had only one input, which we called V_{in}. We now have two inputs V_A and V_B, and each affects the output of its transistor in the same way as V_{in} did in the inverter circuit. Suppose the $V_A = 5$ V and $V_B = 0$ V. Transistor A will conduct, that is, i_{B1} will be high and therefore i_C will be high (see Fig. 27-15*b*). Consequently, the collector of A will be at approximately 0 V with respect to ground ($V_{CE} \approx 0$ V) and V_{out} will be 0 V. A similar output will be obtained if $V_A = 0$ V and $V_B = 5$ V or if both $V_A = V_B = 5$ V. Thus, if either or both inputs are at logic level 1, the output will be at logic level 0. If, however, inputs $V_A = V_B = 0$ V, then neither of the two transistors will conduct ($i_C = 0$) and both collectors will be at 5 V with respect to their emitters, which are

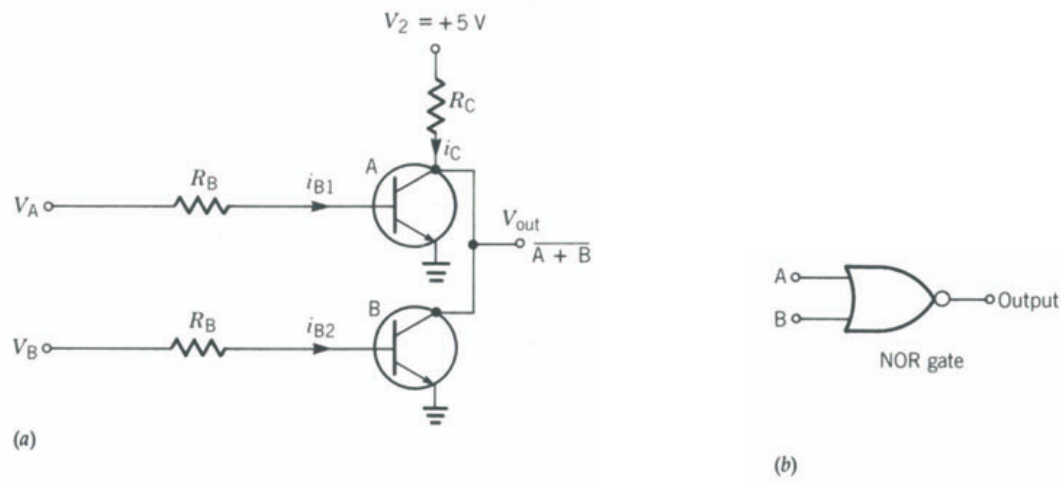


FIGURE 27-18
(*a*) An RTL circuit that implements the NOR logic function. (*b*) Electronic symbol for the NOT-OR or NOR gate.

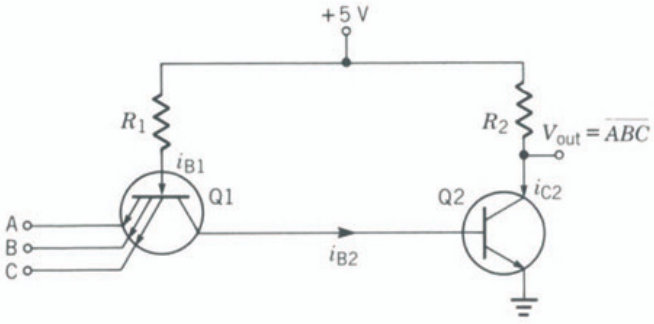


FIGURE 27-19
Circuit diagram of an integrated circuit (IC) TTL gate that implements the NAND logic function. Transistor Q1 is the input transistor that is built with several emitters all having the same base and collector. The inputs are applied at the emitters of Q1.

grounded; $V_{out} = 5$ V. Thus, the circuit in Fig. 27-18*a* implements the NOR logic function.

The advent of new and sophisticated fabrication techniques has allowed the construction and interconnection of hundreds of diodes, transistors, resistors, and such in a single silicon wafer a few square millimeters in area. The advantages of these integrated circuits are many; reduction in size and cost, higher speed (due in part to shorter leads), and improved circuitry; many more components per function can be used than would be practical with discrete components. One such example is the *transistor-transistor logic* (TTL) gate shown in Fig. 27-19. The input transistor Q1 is an *npn* transistor with several emitters. Transistors with eight or more input emitters are not unusual nor difficult to manufacture in IC form. If any of the inputs of Q1 is grounded (logic level 0), the base-emitter junction of Q1 will be forward biased because the *n*-type emitter is at ground potential whereas the *p*-type base is connected via R_1 to the + 5-V source and, therefore, it is more positive than the emitter. This produces a large base current i_{B1} that turns Q1 *on* (see Fig. 27-15*b*). As a result, the collector of Q1 becomes a virtual ground, that is, the voltage between the collector and the emitter of Q1 will be ≈ 0 V. This in turn will cut off the output transistor Q2 because no current flows in its emitter base junction. With Q2 cut off, its collector emitter voltage becomes 5 V, that is, $V_{out} = 5$ V. Thus, if any of the inputs are at logic level 0, the output will be at logic level 1.

If all the inputs of Q1 are at 5 V, its base emitter will be reverse biased, and consequently no current can flow from the base of Q1 to its emitter. Current can flow through R_1 and through the forward biased base collector junction of Q1 into the base of Q2. This will turn Q2 *on* and make its collector emitter voltage ≈ 0 V and, therefore, $V_{out} \approx 0$ V. Thus, when all the inputs are at logic level 1, the output will be a logic level 0. This is logically a NAND gate.

TTL gates have many advantages over RTL and DTL gates. With some simple modifications the *fan-out*[1] of this TTL gate can be as great as 15. (For RTL gates it is typically 5 and for DTL gates it is 8). The *propagation delay*[2] for

[1]*Fan-out:* The number of identical logic gates that can be connected to the output of a gate without impairing the output logic level.
[2]*Propagation delay:* The time interval between the introduction of a change in the logic level at the input and the appearance of the change at the output.

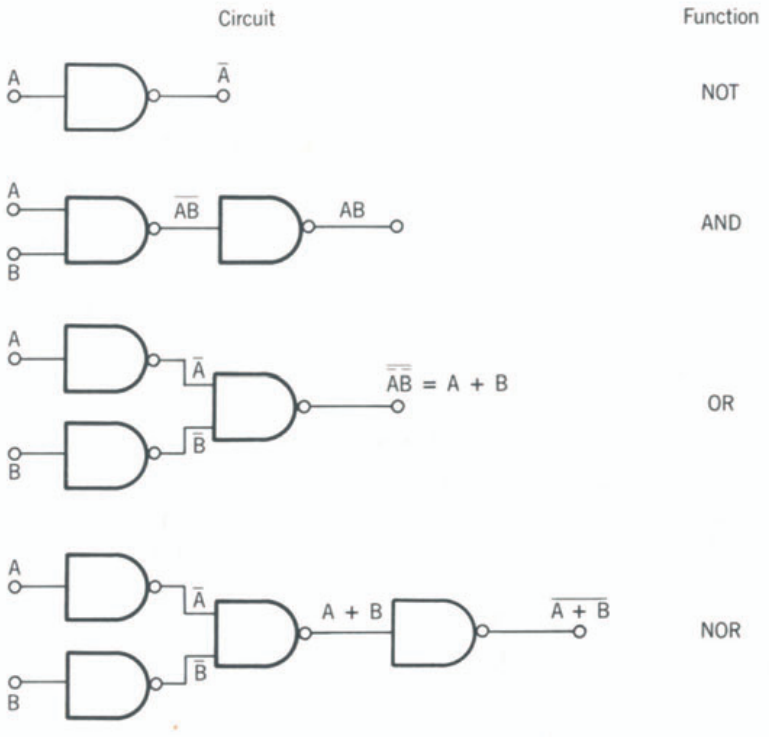


FIGURE 27-20
Implementation of several logic functions with NAND gates.

a TTL gate is about 10 nsec (10×10^{-9} sec) as compared with 25 nsec for the DTL and the RTL gates. The *noise immunity*[3] is also better for the TTL gates. A detailed explanation of these characteristics for the gates discussed here can be found in any digital electronics textbook.

In the preceding sections, we have introduced gates that permit the implementation of the various logic functions of boolean algebra. Although it may seem unnecessary, the needed logic functions are usually obtained with just one type of gate. The reason is that by doing so, the speed of signal propagation, the fan-out, and the noise immunity are well known and uniform throughout the entire digital system; Fig. 27-20 shows how the different logic functions can be implemented with NAND gates. We recall that the NAND logic function is the inverse of the AND, and therefore the truth table has an output that is the opposite of the one in Table 27-1. In particular, when both inputs are 0, the output will be 1 and, when they are both 1, the output will be 0. Thus, if we take a two-input NAND gate and interconnect the two inputs to make a single input it becomes an inverter: a 0 applied at the common input will yield an output of 1, whereas a 1 input will yield a 0 output.

---

[3]*Noise immunity:* A parameter that refers to the ability of a gate to stay in a given logic state in the presence of fluctuations (noise) in the input signal. These fluctuations can be caused by variations in the DC power supply, thermal voltages, or pick-up from adjacent lines where the current may be varying rapidly. A large noise immunity (also called *noise margin*) is a desired property for a gate, for it means that the output logic level will not change in the presence of large fluctuations in the input logic level.

27.7 MEMORY CIRCUITS

The gates that we have discussed can be used to implement the desired boolean functions. Two other elements are needed in the operation of a computer; memory circuits to store information, and clocks to drive the computer, that is, to tell the different parts when to perform a given operation. Both these elements can be constructed with the basic gates using *feedback;* in other words, part of the output is fed back into the input. The general term used for these feedback circuits is *multivibrator.* We will first discuss memory circuits.

The basic memory circuit is called the *flip-flop.* A positive signal feedback is used, resulting in two stable output states that can represent the two logic levels 1 and 0. The circuit is placed in either state and remains in that state indefinitely until the input is changed; it stores a *binary bit* (a 1 or a 0) indefinitely. There are many different flip-flops. We will discuss two basic ones, the *reset-set flip-flop* or simply the *RS flip-flop*, also called the *latch*, and the *data flip-flop* or simply the *D flip-flop.*

27.7a RS Flip-Flop or Latch

Consider two NAND gates connected as shown in Fig. 27-21. Keep in mind that a NAND gate gives a logic level 1 output if any or all inputs are at logic level 0. It gives a 0 level output only when all inputs are at level 1.

The latch has the peculiarity that if a signal 1 is applied to both the "set" and the "reset" inputs, the resulting outputs are ambiguous. This is of no consequence to its operation as a memory circuit, as we will see.

Suppose that initially both inputs are at level 1 and then the "set" input is changed to level 0. This will result in $Q = 1$ and $\overline{Q} = 0$. If the "set" input is changed back to 1, the outputs will not change. The reason is that the other input to gate 1, Fig. 27-21, is at level 0 when the change in the "set" input is made. The output of gate 1 (Q) will remain at level 1. If now a 0 V pulse is applied to the "reset," $Q = 0$ and $\overline{Q} = 1$. The outputs will remain unchanged when the "reset" input is changed back to 1 because the other input to gate 2 is at level 0 when the change is made. Thus the circuit "remembers" which input had the latest momentary 0-V pulse applied to its input.

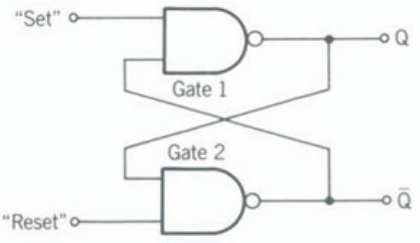


FIGURE 27-21
A memory circuit called the RS (reset-set) flip-flop (sometimes referred to as the latch) constructed with NAND gates.

27.7b The D Flip-Flop

A type of memory circuit without a need to keep both inputs activated, and thereby a more flexible one, can be obtained with a slight modification of the latch. Figure 27-22 shows the data or D flip-flop. Note that the output stage, gates 3 and 4, is the latch discussed in Section 27.7a.

In the D flip-flop, the data to be stored is fed only to the D input. The T input controls the receptiveness of the circuit to the data being fed into the

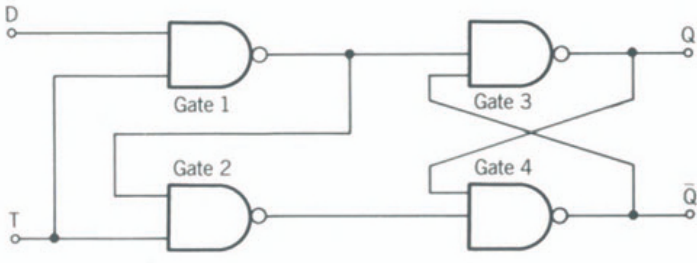


FIGURE 27-22
A memory circuit called the data flip-flop constructed with four NAND gates.

D input. If T = 1, the circuit stores in the Q output the data fed at D; if T = 0, the circuit ignores the input at D and remains in its previous state.

Let us assume that a 1 input is fed at D while the level input of T is 1. The output of gate 1 will be 0; because this output becomes the input for gate 2 and gate 3, their outputs will become 1. As we saw when we discussed the latch, the Q output will be a logic level 1. Moreover, because the outputs of gate 2 and gate 3 become inputs for gate 4, the output $\overline{Q}$ will be 0. If the T input is changed to 0 and then the D input is removed (that is, changed to 0), the outputs will remain unchanged. The only way to change the output Q to logic level 0 is to feed a 0 input at D while T is a logic level 1.

Thus the D flip-flop stores at Q the data fed into the D input when T = 1 and keeps it until new data is fed.

27.8 CLOCK CIRCUITS

A clock is a circuit that produces periodic pulses. These pulses can be used to drive the different gates in the digital instrument. A clock can be constructed with a multivibrator whose output states are unstable and as a consequence has an output signal that changes periodically between high and low. The instability is achieved by using negative feedback. We will now discuss two such clocks.

27.8a Astable Transistor Multivibrator

Consider the circuit shown in Fig. 27-23. Initially, when the power is turned on, both transistors, Q1 and Q2, will tend to conduct. If however, there is

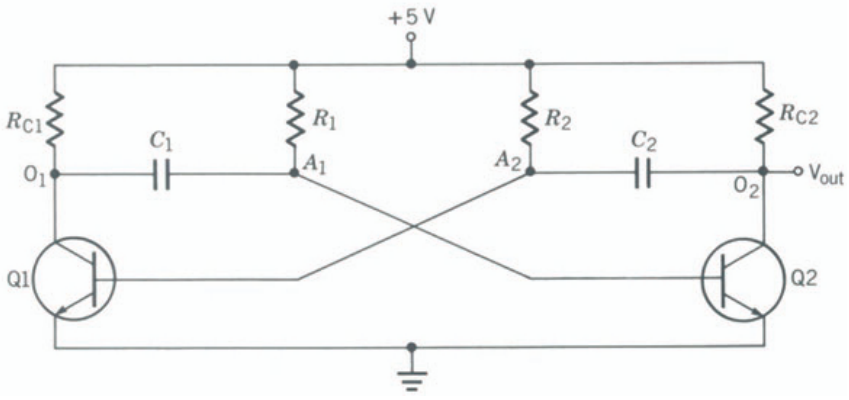


FIGURE 27-23
Circuit diagram of the astable transistor multivibrator. The output V_{out} is a voltage that alternates between high (5 V) and low (0 V). The circuit is therefore a clock.

some type of unbalance, one will be turned on first, and as a consequence the other will be turned off. Suppose that Q1 is turned on first. While this is taking place, capacitor C_1 was being charged to some voltage from the +5 V source (via R_{C1} and Q2). As soon as Q1 is fully turned on, point O_1 becomes an effective ground (its voltage is about 0 V). This means that point A_1 is at a negative voltage with respect to ground. As a result the *p*-type base of Q2, which is connected to point A_1, will be at a lower potential than the *n*-type emitter of Q2, which is grounded. The base-emitter junction of Q2 is reverse biased, and therefore i_B for Q2 will be zero. This will turn Q2 off. Although Q2 is being turned off, V_{out} is not yet at 5 V because current can flow through R_{C2}, C_2, and Q1; eventually C_2 will be charged and V_{out} will become 5 V.

The negative voltage at A_1 kept Q2 off; however, this does not last. While Q2 was being turned off, C_1 was discharging (via R_1 and Q1) as a result of point O_1 having become a virtual ground. When C_1 is fully discharged, it begins to charge in the opposite direction (via the same path), and when A_1 becomes positive Q2 begins to conduct because now the base-emitter junction of Q2 becomes forward biased. As a result, V_{out} drops immediately to 0 V. Because C_2 has been charged and point O_2 is now a virtual ground, A_2 is now at a negative potential with respect to ground. This turns Q1 off, although the potential of point O_1 does not rise immediately to 5 V; it does so exponentially until C_1 is charged. At the same time, while C_1 was being charged, C_2 was discharging (via R_2 and Q2). When C_2 is fully discharged, it begins to charge in the opposite direction, and when A_2 becomes positive, Q1 is turned on, making O_1 a virtual ground; A_1 becomes now negative relative to ground and turns Q2 off. V_{out} then rises exponentially to 5 V while C_2 is being charged. The cycle is completed.

Johann von Neumann (1903-1957) is considered the father of the modern computer for his development in 1946 of the architectural concepts of the computer mainframe.

27.8b IC Free-Running Oscillator Clock

A circuit called the free-running oscillator clock made with two inverters is shown in Fig. 27-24. Suppose that initially $V_a = 0$ V, V_b will then be 5 V, and $V_d = 0$ V. As a consequence, the capacitor C will begin to charge (via R), see Fig. 27-25. This charging will continue until the voltage at point a (the input

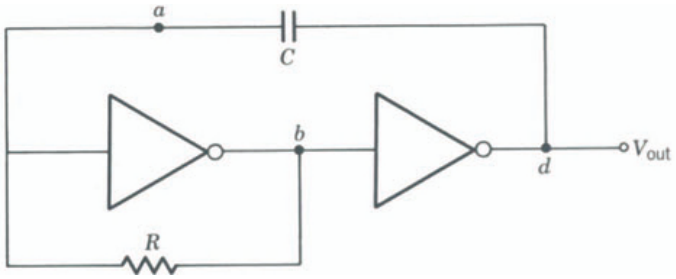


FIGURE 27-24
A clock, called the free-running oscillator, constructed with two inverters, a capacitor, and a resistor. The output voltage oscillates between high and low. The frequency of oscillation is determined by the values of the capacitor C and the resistor R.

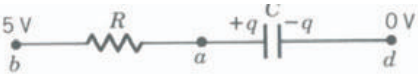


FIGURE 27-25
Charging of the capacitor C in the circuit of Fig. 27-24.

of the first inverter) has risen sufficiently to cause the output of that inverter to change to low, that is, 0 V. Now $V_b = 0$ V and $V_d = 5$ V. The capacitor will now discharge (see Fig. 27-26) until the voltage at point a has dropped enough to cause the output of the first inverter to change back to high, that is, $V_b = 5$ V and consequently $V_d = 0$ V. The cycle begins anew. The RC circuit was derived in Section 15.9, and it is seen by the result that the time to achieve a predetermined voltage depends on the values of R and C. Therefore, the cycle of the clock can be adjusted by these choices.

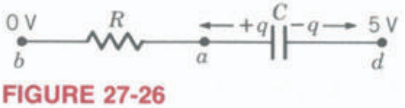


FIGURE 27-26
When the capacitor in Fig. 27-25 has been charged, the potential of point a becomes 5 V. This causes the potential of point b in Fig. 27-24 to become 0 V and that of point d 5 V. The capacitor now discharges and then begins to charge in the opposite direction.

PROBLEMS

27.1 Make the appropriate truth tables to prove the following distributive law of boolean algebra:

$$A\,(B + C) = AB + AC$$

27.2 Prove DeMorgan theorem $\overline{A + B} = \overline{A}\,\overline{B}$ by making the truth table for both sides of the equation; that is, make a truth table for $\overline{A + B} = T$ and for $\overline{A}\,\overline{B} = T$ and show that the result T is the same in both cases.

27.3 Combine the circuits shown in Figs. 27-14*a* and 27-16 to create a DTL gate that will perform the logic function NOR. Make up a truth table for the possible combinations of the two inputs V_A and V_B and the resulting output. Present the reasons for your results.

27.4 The inputs shown in Fig. 27-27 are fed into the circuit of Fig. 27-18. Draw the output waveform.

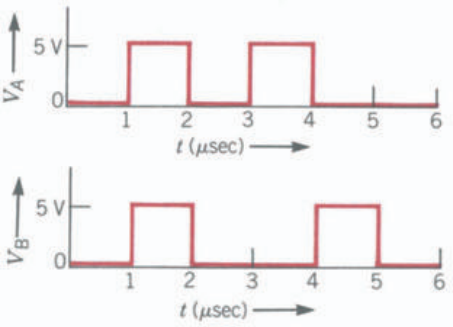


FIGURE 27-27
Problem 27.4.

27.5 Figure 27-28 shows the MC846, a commercial DTL gate. Analyze the circuit and make a truth table for the possible combinations of the two inputs V_A and V_B. What logic function does the circuit perform?

(*Answer*: NAND.)

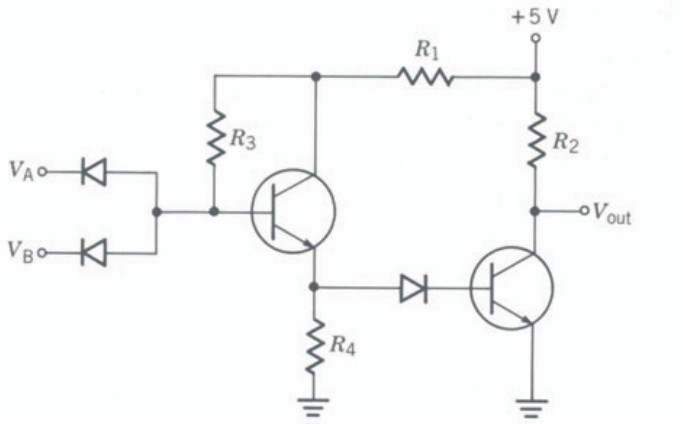


FIGURE 27-28 Problem 27.5.

27.6 Analyze the circuit shown in Fig. 27-29. Determine the logic function performed by the circuit by making and justifying the appropriate truth table.

(*Answer*: AND.)

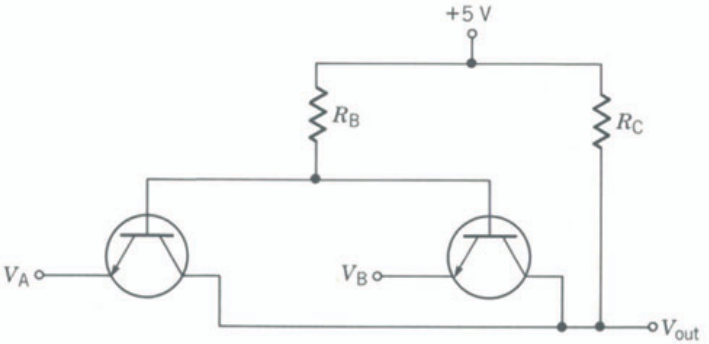


FIGURE 27-29 Problem 27.6.

27.7 In Section 27-6 we showed that the different logic functions could be implemented with NAND gates only. Show, using NOR gates only, how to implement the AND, the OR, the NOT, and the NAND logic functions.

27.8 An important parameter of a gate is its fan-out, that is, the number of identical gates that can be connected to the output of the gate without imparing the output logic level. Consider the inverter of Fig. 27-16. We saw that if $V_{in} = 0$ V, i_B will be zero, and this in turn leads to $i_C = 0$; because no current flows through R_C, there is no potential drop across R_C; therefore $V_{out} = 5$ V. Suppose that the output of such an inverter is connected to the input of a similar inverter as in Fig. 27-30. (a) What is the output voltage of the first inverter (potential of point A) when its input $V_{in} = 0$ V? Take the potential across the forward biased junction between the base and emitter of Q2 to be 0.4 V. (b) Suppose a third inverter, identical to the two in Fig. 27-30, is connected to the output of the first (point A), what will be the potential of point A? (c) What is the fan-out of the inverters in Fig. 27-30?

(*Answer*: (a) 4.08 V, (b) 3.47 V, (c) 3.)

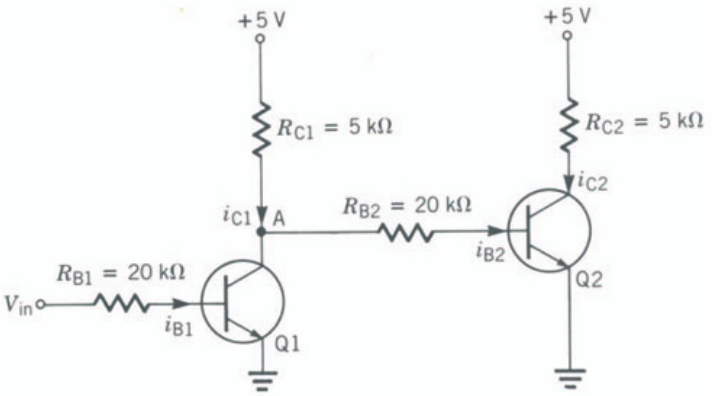


FIGURE 27-30 Problem 27.8.

27.9 (a) Find the truth table for the circuit shown in Fig. 27-31. What logic function does the circuit perform? (b) What logic function will the circuit perform if the constant +5 V input to the first two gates is changed to ground potential?

(*Answer*: (a) OR, (b) None.)

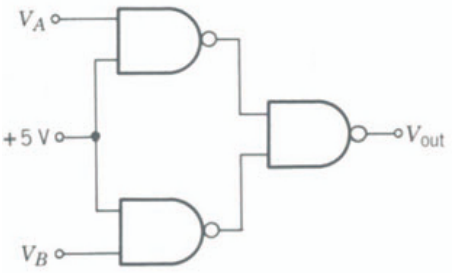


FIGURE 27-31 Problem 27.9.

27.10 (a) Find the truth table for the circuit of Fig. 27-32. What logic function does the circuit perform? (b) What logic function will the circuit perform if the common grounded input to the first two NOR gates is changed to +5 V?

(*Answer*: (a) AND, (b) None.)

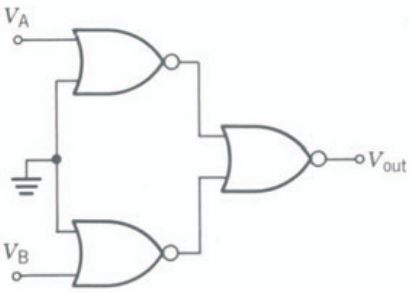


FIGURE 27-32 Problem 27.10.

27.11 The circuit of Fig. 27-33 is a latch memory circuit made with NOR gates. When the circuit is not active, both inputs, the S (set) and R (reset), are kept at logic level 0. Show that when a 1 input is fed at S, the Q output becomes 0 and remains 0 when the S input returns to 0. Similarly, when a 1 input is fed at R, the $\overline{Q}$ output becomes 0 and remains 0 when the R input returns to 0. Thus, the circuit remembers which input was last at level 1.

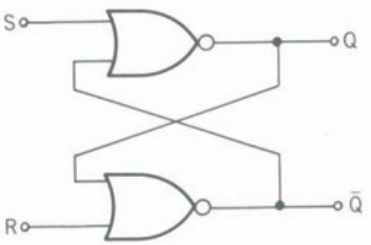


FIGURE 27-33 Problem 27.11.

27.12 The circuit of Fig. 27-34 is a data flip-flop made with NOR gates. Show that the circuit stores at Q the input fed at the D input if the T input is 0 and ignores the input at D if the T input is 1.

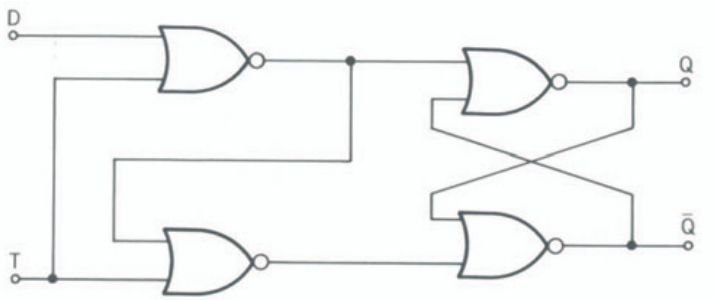


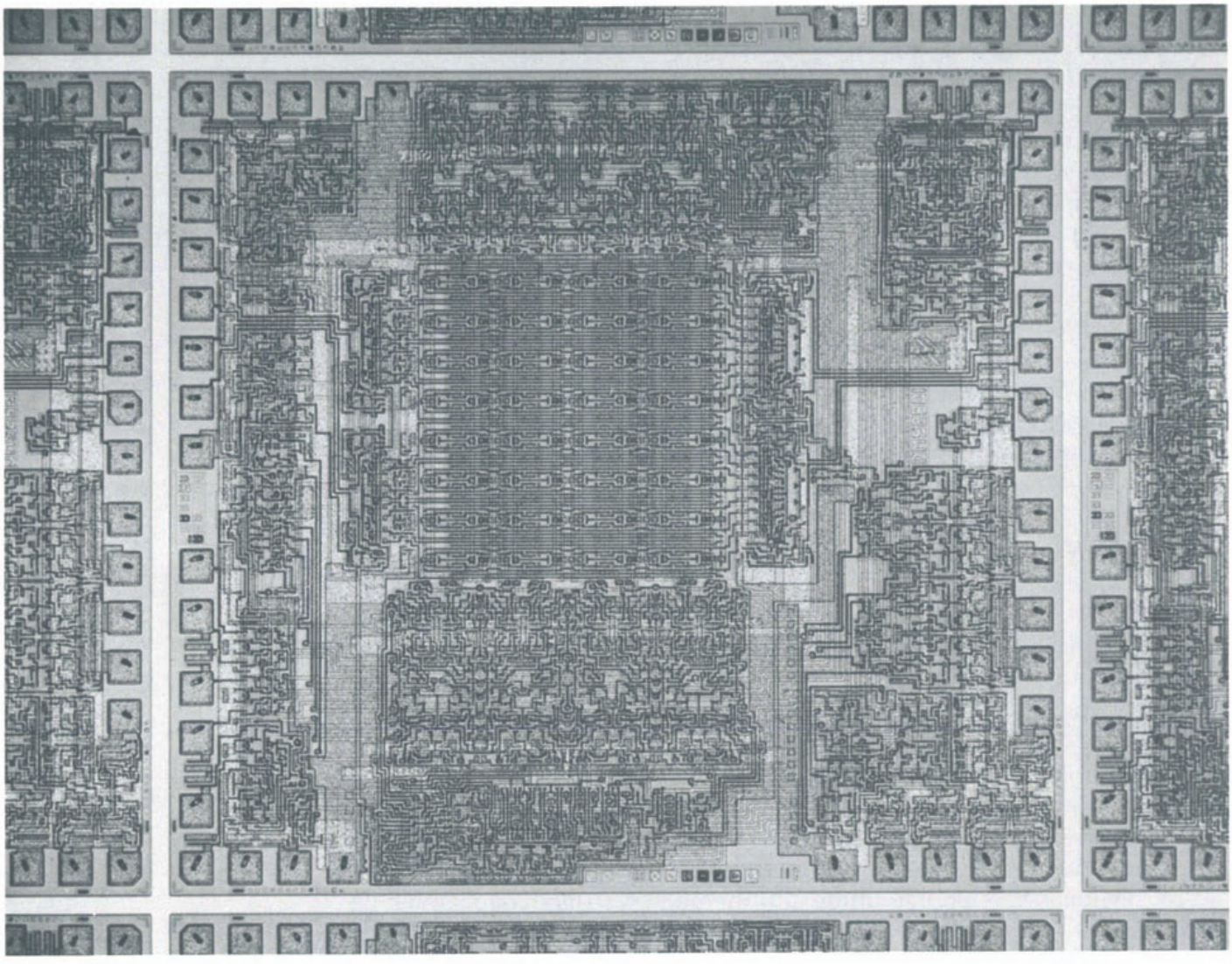
FIGURE 27-34 Problem 27.12.

27.13 In the binary system of counting, only two digits, 0 and 1, are used as opposed to 10 digits in the decimal system. The addition of two binary numbers is performed in a manner analogous to that of the decimal system. Thus, the sum of 0 and 1 is 1; the sum of 1 and 1 is 0 and we carry 1, just as the sum of 9 and 1 in the decimal system is 0 and we carry 1. The truth table for the addition of two binary numbers A and B with the possible combinations of A and B and the resulting sum S and carry C is

A	B	S	C
0	0	0	0
1	0	1	0
0	1	1	0
1	1	0	1

From the truth table, we can see that logically the sum $S = A\bar{B} + \bar{A}B$ and the carry $C = AB$. Draw a circuit, made up exclusively of NAND gates, that will perform the addition of two binary numbers A and B. Indicate the output at each step of the circuit.

27.14 Repeat Problem 27.13 with NOR gates.



CHAPTER 28

The Technology of Manufacturing Integrated Circuits

28.1 INTRODUCTION

Many of the technological achievements of the past two or three decades have been based on microelectronics. Microelectronic devices are at the heart of new products ranging from communication satellites to computers and space shuttles, to name but a few. The revolution in the electronic technology brought about by microelectronics has its roots in metallurgical techniques whose development started in the early 1950s.

In this chapter, we will review the methods used to produce the high purity, single crystal wafers (slices) of silicon on which *integrated circuits* (ICs) are made. We will also consider the techniques employed to fabricate ICs. The use of these techniques will be illustrated at the end of the chapter when we describe the step-by-step production of a simple integrated NOR gate. Because silicon is the basic material for computers at the present time, we will confine our discussion to the purification and preparation of silicon semiconductors, although other semiconductor materials are being developed.

28.2 SEMICONDUCTOR PURIFICATION: ZONE REFINING

The basic element silicon used in the fabrication of semiconductor devices is obtained from the chemical decomposition of compounds such as SiO_2 (the main constitutent of ordinary sand), $SiHCl_3$, and $SiCl_4$. By means of different chemical reactions the Si is chemically prepared with impurity concentrations of about one part per million. The chemically purified silicon is then melted and cast into ingots. The resulting ingot is polycrystalline in nature, that is, it consists of a large number of small (typically a few microns) single crystals having random orientations with respect to one another. To obtain device-grade semiconductor single crystals, the impurity concentration must be reduced by several orders of magnitude and the polycrystalline ingots must be transformed into large single crystals. The stringent purity requirement is achieved by the method known as *zone refining*. It consists of moving a molten zone through the ingots. The molten zone sweeps the impurities along with it and thus, after several sweeps of the molten zone in one direction, most of the impurities are brought to one end of the ingot.

One familiar example illustrates the principle on which zone refining is based. Let us consider a solution of salt in water. As the temperature of the solution is lowered and we reach the freezing point, ice crystals begin to form. The first ice crystals formed contain less salt (in fact, almost no salt) than the initial solution. As a consequence, the salt concentration of the remaining solution increases and concomitantly its freezing point is lowered. These facts are not unique to salt and water. The behavior of solutions, their temperature-concentration characteristics, are presented in phase diagrams. One hypothetical diagram is shown in Fig. 28-1.

Referring to Fig. 28-1, let us start with liquid silicon at a temperature T_1 with an impurity concentration C_1. The temperature is lowered to T_2. At this

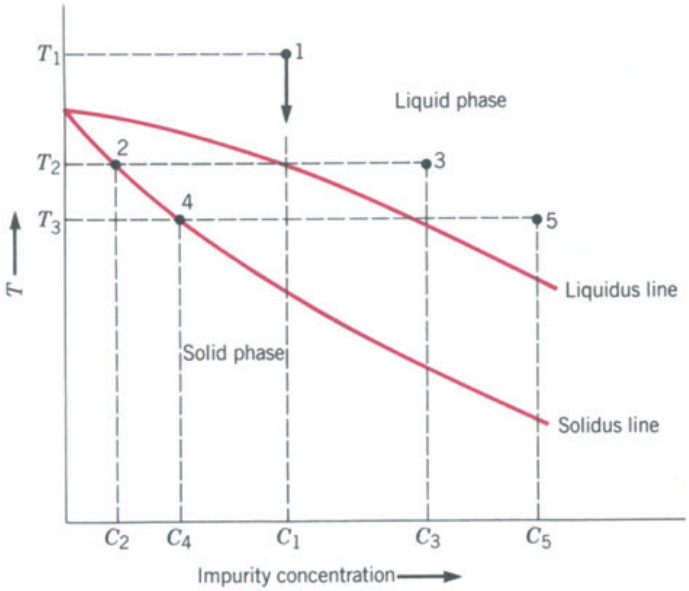


FIGURE 28-1
Phase diagram showing why impurities segregate from a solution on freezing.

temperature, silicon can exist as a solid with a maximum impurity concentration of C_2, shown by the intersection of the dashed temperature line with the solidus line. Consequently, the silicon that crystallizes first will have an impurity concentration C_2 whereas the remaining liquid, because the impurities are still present in the total sample, will have a higher impurity concentration than before, such as C_3. No further crystallization will take place until the temperature of the liquid is lowered to T_3, when crystals with impurity concentration C_4 are formed, and so on. Zone refining exploits this redistribution of impurities between the solid and liquid phases at the freezing point. Figure 28-2 shows one of the zone refining process arrangements.

The Si ingot is placed in a chemically stable, nonmelting boat and is pulled slowly through a furnace. A narrow region of the furnace is maintained at a temperature above the melting point of Si by means of a heating coil. Thus, a narrow molten zone is created in the ingot and, as the boat moves along the furnace, the molten zone moves along the ingot. The moving zone leaves behind it relatively pure solid silicon, whereas the impurities remain in the melted region, or zone. Repeated passes of the molten zone in one direction will result in an ingot with one end highly purified. The impure end of the ingot is then cut off. In today's commercial zone refiners, the boat is kept stationary inside the furnace and moving molten zones are produced by moving a series of heating coils along the walls of the furnace.

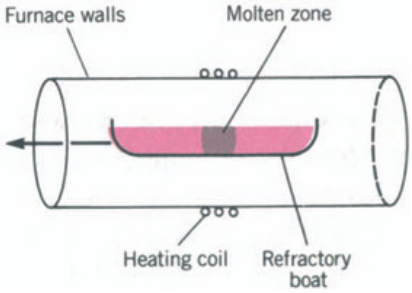


FIGURE 28-2
Schematic of the arrangement for zone refining. As the boat is moved to the left, the molten zone sweeps with it the impurities to the right end of the boat.

28.3 SINGLE-CRYSTAL GROWTH

The production of even the simplest of ICs is complex, time consuming, and prone to defects. Despite this, highly sophisticated ICs can be produced at reasonable cost. One of the main reasons is that methods have been developed in the last few decades for growing large single crystals of Si, thereby per-

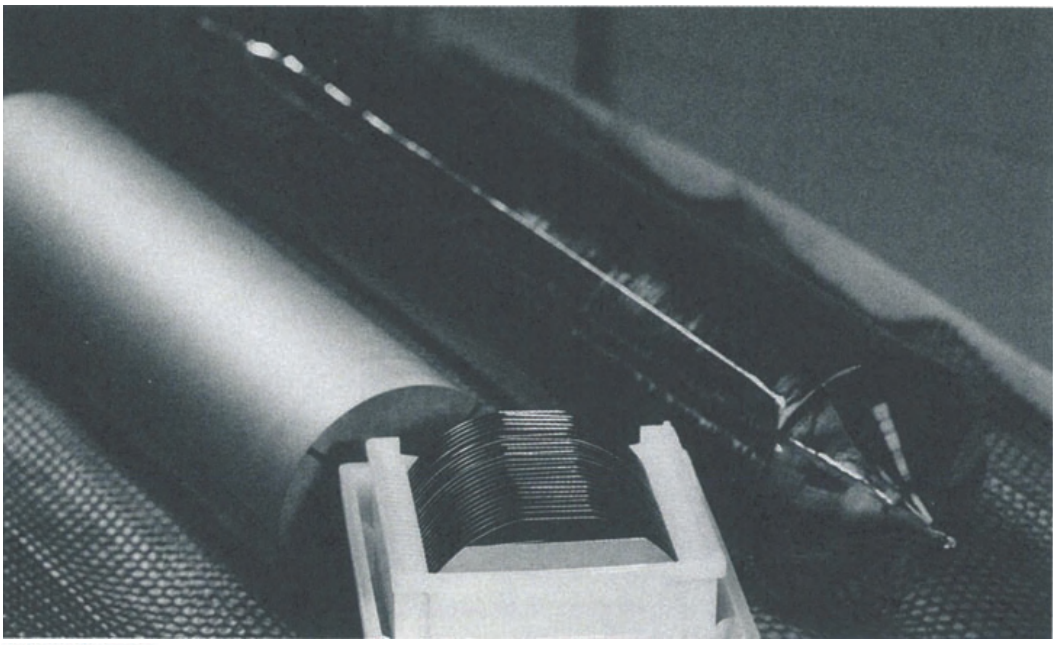


FIGURE 28-3
Large single crystals of silicon grown by the Czochralski method of Fig. 28-4.

mitting mass production. Samples 50 cm long and 10 cm in diameter are not unusual. Figure 28-3 shows two single-crystal ingots approximately 70 cm long and 12 cm in diameter. From such ingots, hundreds of wafers can be cut. In each of these wafers, 10 cm in diameter, hundreds of identical integrated circuits are formed using the techniques described later in this chapter. By processing several wafers simultaneously, the cost of the individual IC is minimized. We will now discuss some of the methods used for single-crystal growth.

28.3a The Czochralski Method

One of the techniques commonly used to grow single-crystal silicon consists in dipping a small seed crystal of Si into a crucible containing the molten silicon. The seed is then slowly raised from the melt. As the seed is lifted away from the melt a single crystal grows continuously onto the seed. A schematic of this method, known as the Czochralski method (named for the inventor), is shown in Fig. 28-4.

To average any variations in the temperature of the melt and thus ensure uniformity in the growing crystal, the crystal is rotated slowly (a few revolutions per minute) as it is being raised. The conditions needed for good single-crystal growth are initially determined by trial and error. For example, the rate at which the growing crystal is raised is a crucial factor, and a few

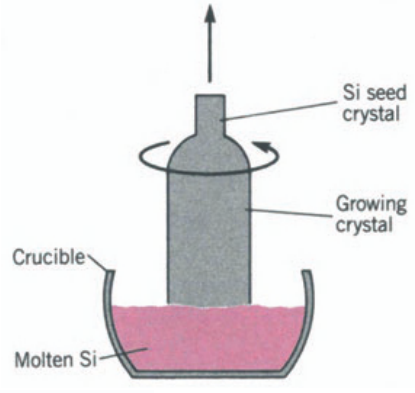


FIGURE 28-4
Schematic diagram of the Czochralski method for growing single crystals.

millimeters per hour is a fairly typical rate. Once these growing conditions are determined, they can be readily reproduced.

28.3b The Bridgman-Stockbarger Method

Another method used for growing single crystals, called the Bridgman-Stockbarger method, is shown in Fig. 28-5. The material is placed in a crucible having a conical tip. The sample, which is initially polycrystalline, is first melted in an upper furnace that is maintained a few degrees above the melting point. The crucible is then slowly lowered into a second furnace with a temperature a few degrees below the melting point. As the crucible enters this lower furnace, a small crystal is formed at the tip of the crucible. This crystal acts then as the seed for the rest of the melt and, as the crucible continues to be lowered, a single crystal is grown.

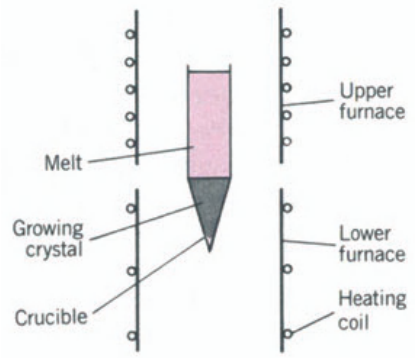


FIGURE 28-5
Schematic diagram of the Bridgman-Stockbarger method for growing single crystals.

28.3c Floating Zone Method

The two methods just described have the disadvantage that the melt tends to dissolve some of the oxygen from the walls of the crucible (usually made of silica, SiO_2, which has a higher melting point than silicon). The solution to this problem is found in the *floating zone* method, which dispenses with the crucible altogether. A schematic of the arrangement used in the floating zone method is shown in Fig. 28-6.

A polycrystalline rod of the crystal to be grown is held between two support posts inside a furnace filled with an inert gas. A small seed crystal is placed between the lower support post and the rod. A small molten zone is created at the end of the rod in contact with the seed by means of a movable external heating coil. As the coil is slowly raised, the molten zone moves upward, while the crystal solidifying behind it grows onto the seed. The molten zone is held between the unmelted ends by surface tension. The floating zone technique is also used for zone refining.

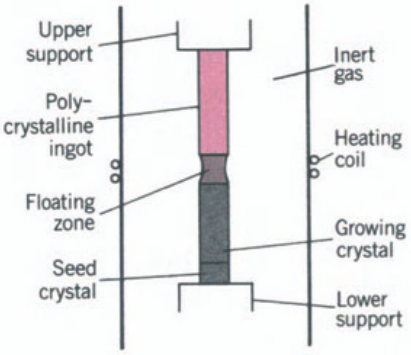


FIGURE 28-6
Schematic diagram of the floating zone method for single-crystal growth.

28.3d Vapor-Phase Epitaxy

A crystal-growing method that plays an important role in the fabrication of ICs is *vapor-phase epitaxy*. Very often the whole integrated circuit is made on a layer of Si 10 or 20 μm thick that is grown *epitaxially* onto a Si substrate (wafer). The word "epitaxial" originates from the Greek word meaning "arranged on." In this method, atoms of Si from a vapor are deposited on the substrate in a layer that has the same crystal structure and orientation as the substrate. Thus, the substrate serves as the seed crystal onto which the epitaxial layer grows. This first layer in turn serves as the substrate for the second layer, and so on. A schematic of the apparatus used for vapor epitaxy is shown in Fig. 28-7.

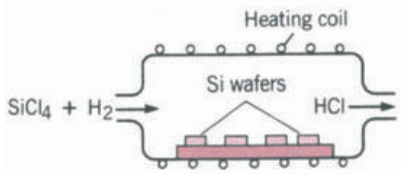


FIGURE 28-7
Arrangement for growing Si layers on silicon wafers by epitaxial growth.

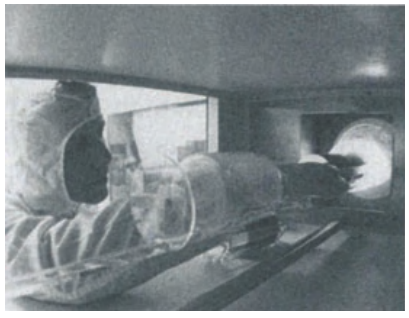
Single crystal wafers of Si are placed in a heated chamber called the "reactor." Gaseous compounds of silicon (for example, silicon tetrachloride, $SiCl_4$), together with the appropriate reactant gas, are introduced into the reactor chamber. The temperature of the reactor is adjusted to produce the reaction that will liberate the silicon by decomposition of the compound. Thus, for example, at 1250°C the reaction

$$SiCl_4 + H_2 \rightarrow Si + HCl$$

occurs. Some of the Si atoms released in the reaction are deposited onto the silicon substrates, thereby forming epitaxial layers. If the chemicals used for the reaction are of high purity, the epitaxial layer of Si will be highly pure. Alternatively, the layer can be deliberately doped, making it either *p*-type or *n*-type, by passing the hydrogen through a solution containing boron or phosphorous atoms (for example, boron trichloride or phosphorous trichloride), before it is introduced into the reactor.

28.4 THE PROCESSES OF IC PRODUCTION

The processes involved in the fabrication of integrated circuits include epitaxial growth, oxidation, oxide removal and pattern definition, doping (introduction of selective impurities in the Si), and interconnection of components. Epitaxial growth has been discussed in Section 28.3d. We will now explain the additional steps in the manufacturing of integrated circuits.



Si wafers being unloaded from an oxide growth furnace.

28.4a Oxidation

A key step in the production of an IC is the formation of a silicon dioxide (SiO_2) layer on the surface of the silicon. This oxide layer permits the opening of windows on the silicon surface by the method described in the next section. The oxide layer also protects *p-n* junctions from contamination. Finally, because SiO_2 is nearly an insulator, the oxide allows the interconnection of the circuit components by means of thin aluminum strips without short-circuiting sections of the IC.

The oxide layer is grown by heating the silicon wafer to temperatures ranging between 1000°C and 1200°C in an atmosphere of either pure oxygen or steam. The thickness of the oxide layer depends on the oxidation time and the temperature and the composition of the atmosphere in which the oxidation is performed. By careful selection of these three parameters, the exact thickness of the layer can be controlled. A layer 0.1 μm thick can be grown in one hour, at $T = 1000$°C, in pure oxygen. In the same time, a layer 0.5 μm thick grows in a steam environment.

28.4b Pattern Definition

To make an integrated circuit, a method of creating accurate patterns on the silicon wafer is needed. *Photolithography* (or *masking*) allows the removal of the silicon dioxide in the desired sections of the wafer. Once these "windows"

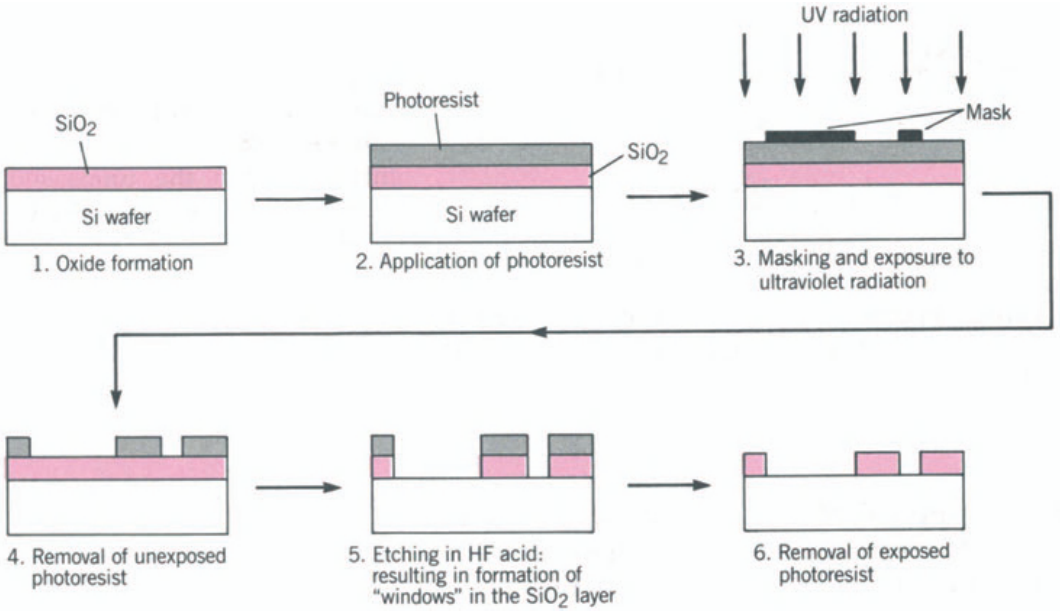


FIGURE 28-8
Steps involved in the masking process (photolithography) used to open "windows" selectively on the silicon wafer.

have been opened, diffusion of dopants or deposition of metallic contacts can be performed.

The process of photolithography is carried out as follows. The oxidized wafer is coated with a thin layer of a photosensitive material called *photoresist*. This is done by placing a drop of a solution containing the photoresist on the wafer. A thin film is then formed by spinning the wafer very rapidly. Finally, the wafer is heated to speed up the evaporation of the solvent and to enhance the adhesion of the photoresist film to the oxide layer. A mask with the desired pattern is then placed on the photoresist film. The wafer is exposed to ultraviolet light that changes the structure of the exposed photoresist so that the exposed and unexposed parts have different solubilities in certain chemical solutions. Thus, exposure to the ultraviolet light, followed by development in the appropriate chemical solution, allows the removal of the unexposed section of the photoresist. The sections of the silicon dioxide layer not protected by the photoresist are then etched away in a solution of hydrofluoric acid (HF), which selectively attacks the SiO_2 while leaving the photoresist and the silicon intact. After the window pattern has been opened in the SiO_2, the remaining photoresist is washed away with the appropriate solvent. The wafer is now ready for the introduction of the dopants or for the evaporation of metallic contacts. The various steps involved in the formation of these "windows" are illustrated in Fig. 28-8.

28.4c Doping

An integrated circuit has the colloquial name *chip*, which we will now begin to use. The formation of circuit components in a chip is achieved by the

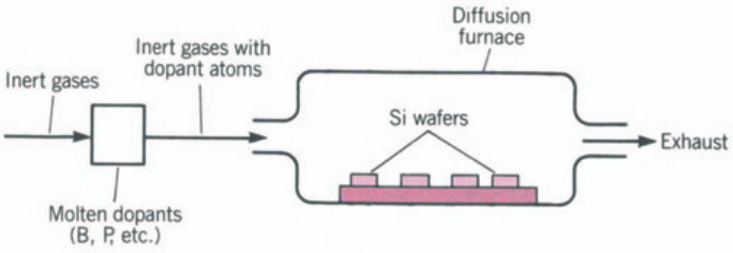


FIGURE 28-9
Schematic diagram of the apparatus used for the introduction of impurities into the unmasked sections of the silicon wafer by diffusion.

selective introduction of donor and acceptor impurities into the Si wafer to create localized *n*-type and *p*-type regions. The two most commonly used techniques are *diffusion* and *ion implantation*.

Diffusion

When Si is heated to temperatures in the neighborhood of 1000°C, some of the semiconductor atoms move out of their lattice sites, leaving behind empty lattice sites that can migrate through the sample. If the heating is done in an atmosphere of either phosphorous or boron atoms, these impurity atoms move into the vacant lattice sites at the surface of the silicon and subsequently migrate slowly into the bulk of the Si with the assistance of the vacant lattice sites formed at high temperature. The diffusion of the dopant impurities can be stopped by cooling down the wafer. Because this solid state diffusion of impurities is time and temperature dependent, the depth of the diffusion layer can be controlled by varying these two parameters. Phosphorous atoms can be diffused up to a depth of 1 μm by heating the Si crystal to 1000°C for 1 h. Another important aspect of the diffusion of either boron or phosphorous impurities is that at the same temperature they move much more slowly in SiO_2 than in pure Si. Thus, the oxide pattern, formed by the photolithographic method described earlier, acts as a mask that permits the diffusion of the impurities only in specific regions of the wafer.

Figure 28-9 shows a schematic of a typical diffusion furnace used in the fabrication of Si chips. The dopants to be diffused into the silicon wafers are introduced into the heated furnace by means of an inert carrier gas that is bubbled through the molten dopant (B, P, or such). Alternatively, gaseous compounds of B or P, such as diborane (B_2H_6) or phosphorous trifluride (PF_3), can be introduced directly into the furnace.

Ion Implantation

An alternative method for introducing impurities into the semiconductor is *ion implantation*. After being ionized, atoms of the required dopant are accelerated in a vacuum through a potential difference of several thousand volts. The accelerated ions are directed onto the masked silicon substrate (see Fig. 28-10). As the impurity ions enter the Si, they make multiple collisions with the lattice ions and eventually come to rest. Because these collisions are random processes, not all the ions penetrate the same depth into the semiconductor. In Fig. 28-11 we show a typical distribution of the dopant atoms as

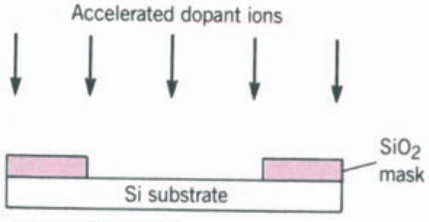


FIGURE 28-10
Schematic of the ion implantation method for doping the unmasked sections of the silicon wafer.

a function of the distance from the surface of the Si wafer. The distribution is approximately gaussian about an average penetration $\bar{d}$.

The average penetration $\bar{d}$ depends on the type of impurity and on the energy (the voltage through which the ions are accelerated) with which they strike the surface of the Si. This latter fact can be exploited to produce a fairly uniform profile of impurities. Figure 28-12 shows the result of varying the accelerating voltage several times during the implantation process. Each of the gaussian distributions (dashed lines) corresponds to a different accelerating potential. The overall distribution (solid line) is obtained by summing up the individual gaussian distributions, and the sum, as can be seen from the graph, is relatively flat.

Ion implantation has many advantages over diffusion. The process is performed at room temperature. This permits the implantation of doped layers without disturbing previously diffused or implanted layers. Because the impurities are ionized, they represent a current that can be measured very accurately. This permits accurate control of the impurity concentration. Ion implantation can be used with impurities that do not diffuse easily in Si. The relatively recent use of arsenic as a dopant in MOS (metal-oxide-semiconductors) devices is due to the advent of ion implantation.

One problem encountered with ion implantation is that the very energetic ions cause considerable lattice damage as a result of the collisions. This difficulty is corrected by annealing, that is, heating the wafer moderately to about 800°C, after implantation.

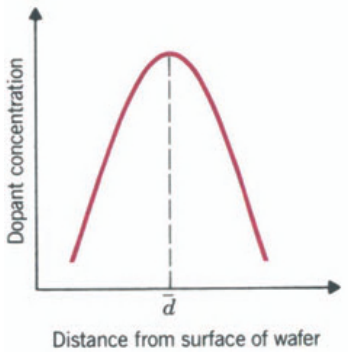


FIGURE 28-11
Profile of impurities introduced by the ion implantation method. The distribution of impurities is approximately a gaussian curve centered about some average penetration $\bar{d}$ from the surface of the wafer.

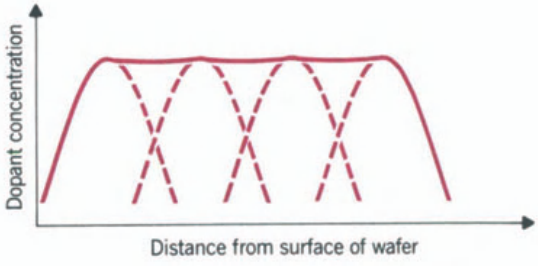


FIGURE 28-12
Profile of impurities obtained by varying the accelerating voltage four times during the implantation process. Each of the four gaussian curves (dashed lines) represents the distribution of impurities for a particular value of the accelerating voltage. The solid line, which gives the total distribution of impurities, is the sum of the four individual gaussians.

28.4d Connection of Components in a Chip

An IC chip consists of many superimposed doped layers. To complete the circuit, the electronic components within a layer as well as the layers themselves must be electrically connected. This can be done either by forming (by diffusion or ion implantation) heavily doped (and therefore conductive) regions of silicon or by metal electrodes. This latter method is performed by metallic *thin film evaporation*. A schematic of the setup used for the evaporation of thin metallic films is shown in Fig. 28-13. A charge of the metal to be evaporated (usually aluminum or gold) is placed in a crucible. The silicon wafers are placed above the crucible in a device called the *planetary*. Vapor of the metal is produced by heating the crucible with a heater coil wrapped around it, or by electron bombardment. As the metal vapor hits the cooled,

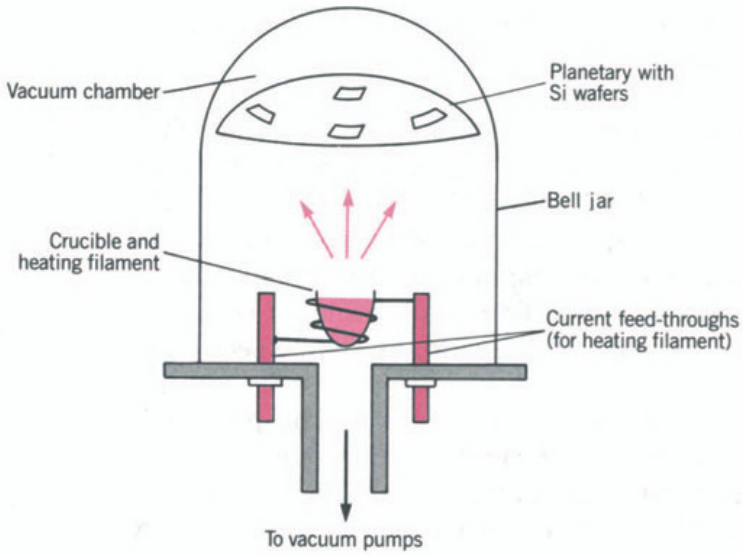


FIGURE 28-13
Schematic representation of a vacuum chamber used for vapor deposition of metallic interconnecting strips.

masked wafer, it condenses on it, thus forming a thin metallic layer in a desired pattern that connects the different sections of the IC chip. The whole evaporation process is performed in a vacuum to avoid contamination of the metal vapor with the oxygen in the air.

28.5 ELECTRONIC COMPONENT FABRICATION ON A CHIP

In this section we will illustrate how the techniques just described can be used to fabricate the basic components of a circuit: transistors, diodes, resistors, and capacitors. For simplicity, we will discuss the formation of each component separately. In the actual fabrication of an IC, parts of several components are formed simultaneously. This will be illustrated at the end of this section, where we discuss how a simple gate can be made.

28.5a Transistors and Diodes

The steps needed to build a bipolar transistor on a Si wafer are illustrated in Fig. 28-14.

1. An epitaxial (called an *epilayer*) *n*-type layer is grown onto a *p*-type wafer (Fig. 28-14*a*). Part of this layer will serve as the collector.
2. The *n*-type epilayer is then oxidized, masked, exposed to UV light, and so on, resulting in the formation of a window in the oxide layer (Fig. 28-14*b*).
3. Acceptor-type impurities are diffused to convert part of the exposed *n*-layer to *p*-type (Fig. 28-14*c*). A part of this *p*-type island will be the base of the transistor.

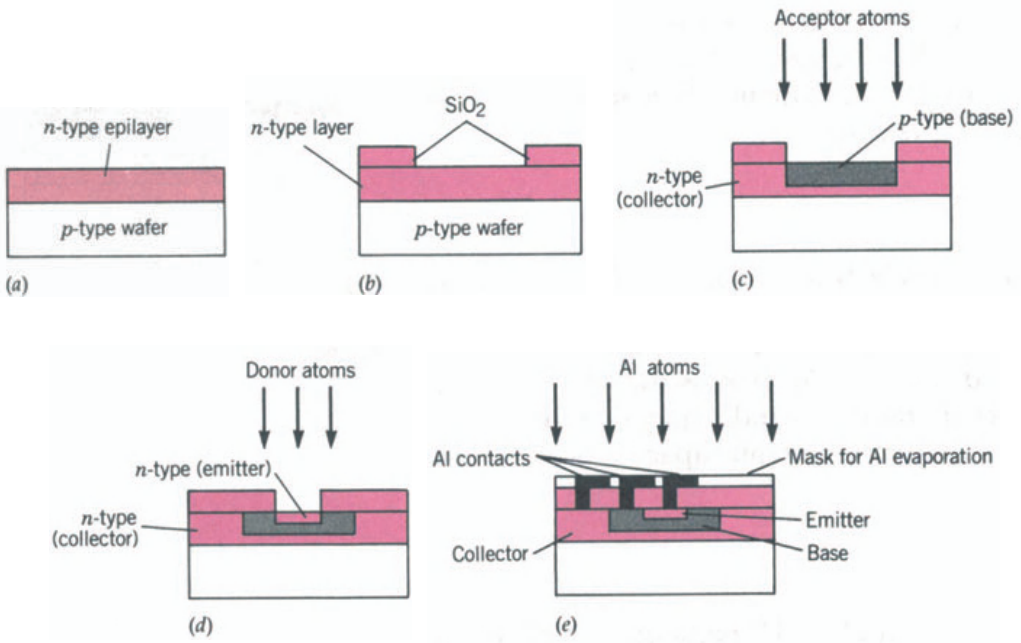


FIGURE 28-14
Steps involved in the fabrication of a transistor on a silicon chip.

4. The wafer is again oxidized, a window is opened in the new oxide layer, and a diffusion of donor impurities is performed (Fig. 28-14*d*), reconverting part of the *p*-type island to *n*-type. The latter region is the emitter of the transistor.
5. In the final step the wafer is reoxidized and three windows are opened: one into the collector, one into the base, and one into the emitter. Aluminum is evaporated to connect the three elements of the transistor (emitter, base, and collector) to other components of the circuit (Fig. 28-14*e*).

The steps needed to build a diode are identical to those used in the fabrication of a transistor except that the last diffusion of donor impurities (step 4) is omitted.

28.5b Resistors

An integrated circuit resistor can be made by the shallow diffusion of a *p*-type channel into an *n*-type region or vice versa. The current is forced to flow through the channel by maintaining the channel at a negative voltage with respect to the surrounding *n*-type region. The resistance of the channel will be determined by its length, its cross section, as well as the doping concentration. Because of its relatively high conductivity, Si is not a useful material for a resistor. As a result, it is difficult to obtain large resistance values in ICs without using too much space in the chip. In cases where large resistors are needed, one standard approach is direct substitution. As we saw in Chapter 26, a transistor used in the common emitter configuration can be considered as a base current-controlled resistor. Thus, a transistor can be

introduced in a circuit where a resistor is needed. The effective resistance between the collector and the emitter will be determined by the base current. Circuit designers often introduce transistors in a circuit where resistors might have been employed to save space on the chip.

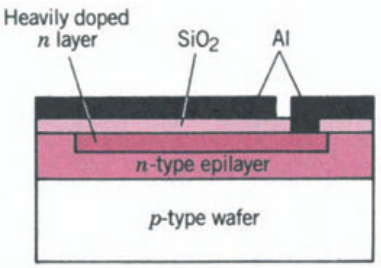


FIGURE 28-15
Microelectronic capacitor on a silicon chip.

28.5c Capacitors

A capacitor is essentially two conducting electrodes separated by a very thin insulator. The first electrode of the microelectronic capacitor is usually made by doping very heavily (thus making it highly conductive) a region of the epitaxial layer. This region is covered with a SiO_2 layer as the insulator and the second electrode is formed by evaporating a conducting aluminum film on the oxide layer. A schematic of an integrated circuit capacitor is shown in Fig. 28-15.

28.5d Fabrication of a Simple IC

As we mentioned earlier, in the fabrication of an IC parts of several circuit components are often formed simultaneously. We will show how the simple NOR gate of Fig. 28-16 can be made (see Problem 27.3).

The first step consists in growing three *n*-type islands on a *p*-type substrate. One of these islands will serve to construct the transistor, another the diodes, and the third the three resistors. These three islands are isolated electrically from one another by maintaining them at a positive potential with respect to the *p*-type substrate; that is, the substrate-islands junctions are reverse biased. To form these three islands, an *n*-type epilayer is grown onto the *p* substrate (Fig. 28-17*a*). The epilayer is then oxidized and coated with photoresist (Fig. 28-17*b*). The wafer is exposed to UV radiation through the mask shown in Fig. 28-17*c*. After developing the photoresist (removing the unexposed parts) and etching the wafer with HF (removing the unprotected SiO_2), the wafer has the form that is illustrated schematically in Fig. 28-17*d*. The exposed photoresist is dissolved away, and acceptor type impurities are diffused into the *n*-type epilayer. The diffusion is allowed to continue until the entire epilayer is converted to *p*-type, except for the three regions protected by the SiO_2 coating. A top view of the wafer at this stage is shown in Fig. 28-17*e*. The wafer is now reoxidized. Using the mask shown in Fig. 28-17*f*

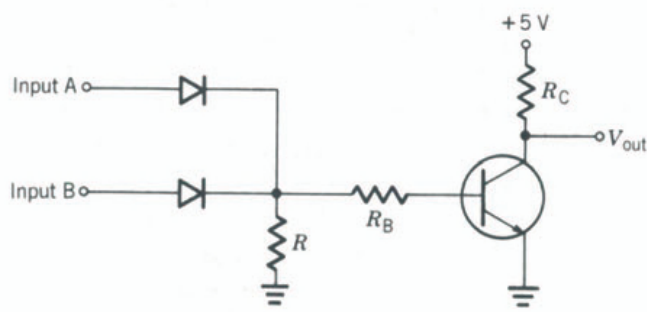
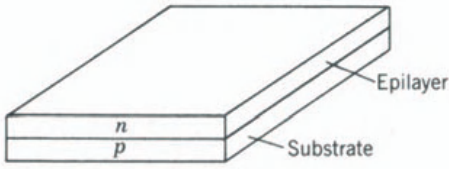
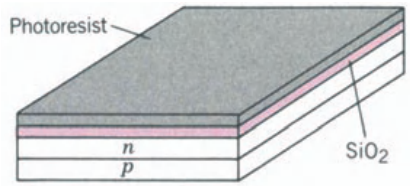


FIGURE 28-16
Electronic diagram of a DTL NOR gate.

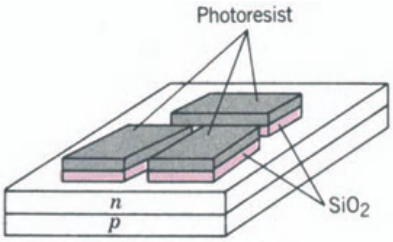


(a) n-type epitaxial layer grown on p-type wafer

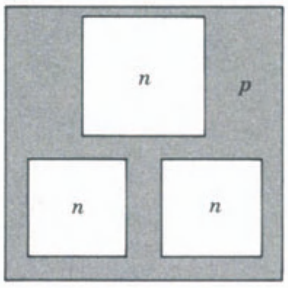


(b) Oxidation of n-type layer followed by photoresist coating

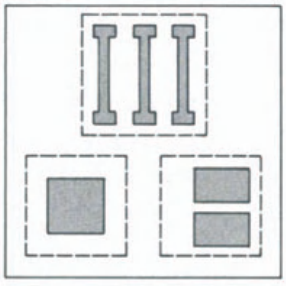
(c) Mask for n-type island formation



(d) Formation of oxide island with exposed photoresist on top

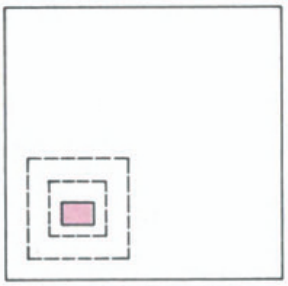


(e) Top view of wafer after n-type island formation

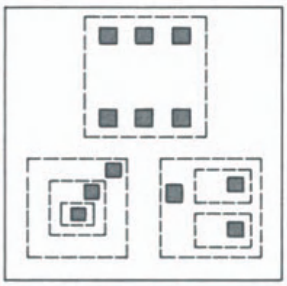


(f) Mask for p dopant diffusion

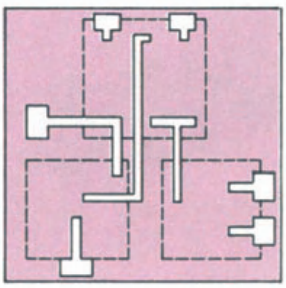
(g) State of the chip after p-type diffusion; shaded areas are p-type, white areas are n-type



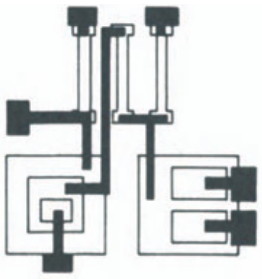
(h) Mask for emitter diffusion



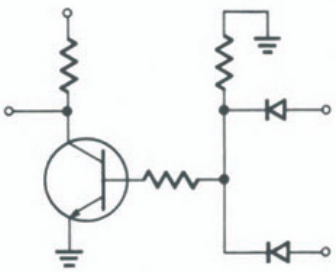
(i) Mask for electrical contact window-opening



(j) Mask for aluminum evaporation



(k) Complete and interconnected circuit



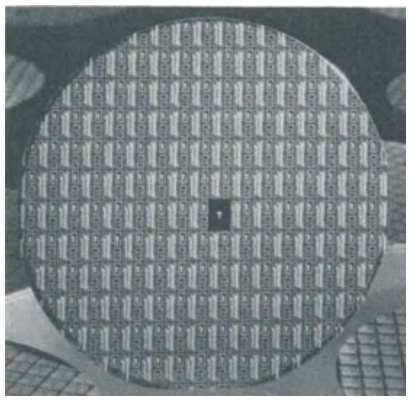
(l) Gate using conventional electronic symbols

FIGURE 28-17

Steps involved in the fabrication of the NOR gate of Fig. 28-16 in integrated form.

and following the procedure outlined earlier for window opening in the oxide layer (photoresist coating, masking, UV radiation, and so on), windows with the shape of the shaded areas of mask (*f*) are opened in the *n*-type islands. Acceptor-type impurities are diffused through these windows into the three *n*-type islands. In this case, the *p*-type dopant is not allowed to penetrate all the way through the *n*-type layers. When this step is completed, there are three resistors in the upper island, the collector and the base of an *npn* transistor in the lower left island, and two diodes with their *n* side common in the lower right island. The state of the chip is shown in Fig. 28-17*g*. The wafer is oxidized again, and, using the mask shown in Fig. 28-17*h*, a window is opened onto the base of the transistor for the diffusion of *n*-type dopants to form the emitter.

All the circuit components are now in place, and they must now be interconnected. To achieve this, the wafer is reoxidized again, and, with the mask shown in Fig. 28-17*i*, small openings are made in the ends of the resistors, the three elements of the transistors, and both sides of the diodes. The wafer is then covered with the mask of Fig. 28-17*j*, and an aluminum evaporation is performed. The final structure is shown in Fig. 28-17*k*. The circuit is redrawn, using electronic symbols, in Fig. 28-17*l*.



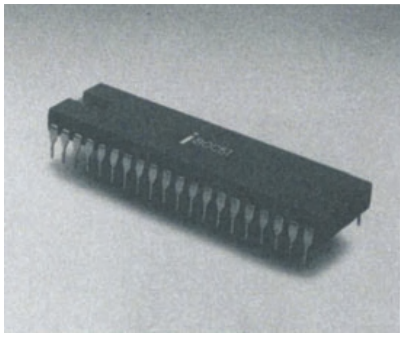
A silicon wafer on which over one hundred identical integrated circuits have been simultaneously fabricated. The wafer is subsequently scribed and broken into individual chips.

28.6 CONCLUSION

In this chapter, we have reviewed the technological processes used in the manufacture of integrated circuits. All these techniques became available by 1960, when the era of integrated circuits began. Since then, the progress has been truly amazing. A modern-day integrated circuit 5 mm square contains more electronic components than the most complex piece of electronic equipment built before 1950. The $100 hand-held calculator has more computing capability than the ENIAC, the first large electronic computer. It is 20 times faster and thousands of times more reliable, has a larger memory, occupies 1/30,000 the volume, consumes the power of a hearing aid rather than that of an electric heating system, and its price tag is 1/100,000 that of the ENIAC.

The main reason for this rapid development has been the steady increase in *component density*, that is, the implementation of more and more components in a given area of the silicon wafer. The evolution of component density is illustrated in Fig. 28-18, known as Moore's curve (Dr. Gordon Moore predicted this evolution in 1964, and his prediction has been proven true up to now). Moore's curve simply states that the component density on a chip quadruples every two years.

The increase in component density has been achieved by reducing the dimensions of the circuit elements. This has been made possible by improving the resolution of the photolithographic process. The line widths of the masks used to expose the photoresist to the UV radiation is limited, because of diffraction, to a few wavelengths. With ultraviolet radiation, this is of the order of a micrometer, a limit that has now been reached. In the search for



After mounting the IC on a substrate and bonding the necessary wire leads for external connections, the chip is encapsulated in a protective plastic casing.

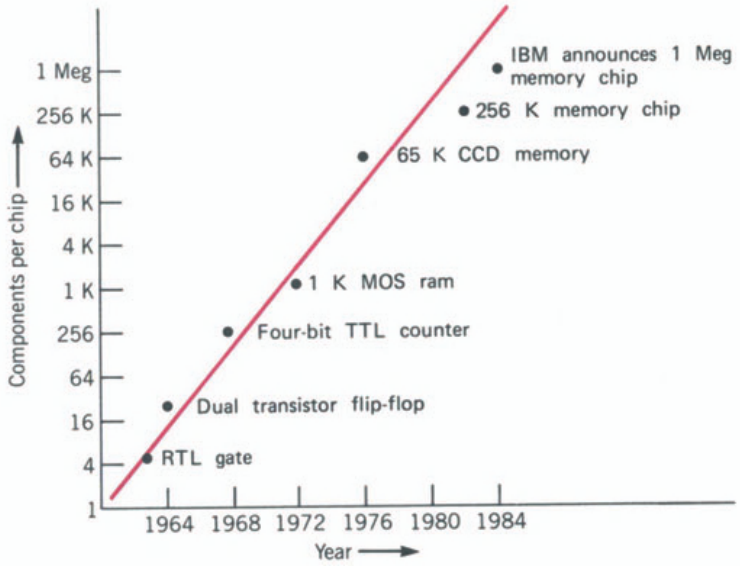
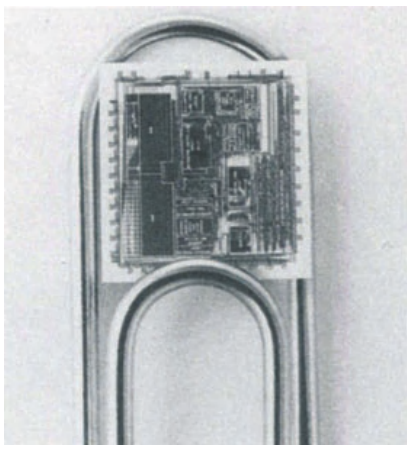


FIGURE 28-18
Moore's curve showing that since 1964 the number of circuit components on a chip has quadrupled every two years.

further reduction of the component size, methods are being developed in which ultraviolet radiation is being replaced by X-rays and electron beams. We may reasonably expect that component density of Si chips will continue to increase, although probably not as rapidly as Moore's curve predicts. Current research on organic molecules as circuit components may someday create a new era of component density.

Concommitant with the reduction in the size of the circuit elements, there has been a steady improvement in the performance of the device and a reduction in production costs. Because delay times are directly proportional to the dimensions of the circuit, the circuit becomes faster as it becomes smaller. Moreover, as circuits become more integrated, the number of external interconnections decreases, leading to an increase in the reliability of the device and a reduction in maintenance costs. Further reductions in cost are associated with the fact that power dissipation decreases as the size of the device decreases. In addition to the energy savings, this leads to savings in such accessories as power transformers, cooling fans, racks, and cabinets.

Microelectronics has now gone through MSI (Medium-Scale Integration, a few hundred components per chip), LSI (Large-Scale Integration, a few thousand components per chip) and VLSI (Very-Large-Scale Integration). It is rapidly progressing into the SLSI (Super Large-Scale Integration with several hundred thousand components per chip). We are now technically able to implement a complete computer on a single chip selling for a few dollars.



The BELLMAC-4 a one chip computer developed for telecommunications is compared to a standard size paper clip.

Photo Credits

CHAPTER 1
Opener: Photo, NASA; quote Copyright © 1985 by the New York Times Company. Reprinted with permission. Page 3: National Bureau of Standards.

CHAPTER 2
Opener: Fredrick D. Bodon.

CHAPTER 3
Opener: Ira Kirschenbaum/Stock Boston. Page 28: Educational Development Center. Page 30: Educational Development Center. Page 31: From Artillerie Zutphen by Diego Ufano, 1621, plate 22.
New York Public Library Rare Book Collection.

CHAPTER 4
Opener: Jim Harrison/Stock Boston. Page 38: Courtesy Yerkes Observatory, University of Chicago. Page 39: (top) New York Public Library Picture Collection; (bottom) Utrecht Museum. Page 41: The U.S. National Standard of Mass.

CHAPTER 5
Opener: Peter Menzel/Stock Boston. Page 60: St. Ruan Robertson/Magnum.

CHAPTER 6
Opener: Michael Philip Manheim/Photo Researchers. Page 75: From *PSSC Physics*, 2nd ed., DC Heath & Co., with Educational Development Center, Newton, MA.

CHAPTER 7
Opener: Photo Courtesy Kings Island. Page 91: (bottom) Bettmann Archive. Page 92: Drawing from Coulomb's Memoire to the French Academy of Sciences.

CHAPTER 8
Opener: Bettmann Archive. Page 108: Steve E. Sutton, © 1985 duomo.

CHAPTER 9
Opener: Sacramento Peak Observatory. Page 112: Edgar Fahs Smith Collection. University of Pennsylvania. Page 113: Bazzechi, Florence. Page 114: New York Public Library Picture Collection. Page 116: Culver Pictures. Page 126: University of Vienna, Courtesy Neils Bohr Institute.

CHAPTER 10
Opener: R. D. Ullmann/Taurus photos.

CHAPTER 11
Opener: Florida State Department of Tourism. Page 151: Henry Groskinsky/© 1965 Silver Burdett Company.

CHAPTER 12
Opener: Dravo Corp. Pittsburgh, Pennsylvania. Page 159: From *PSSC Physics*, 2nd ed., DC Heath & Co. with Educational Development Center, Inc., Newton, MA. Page 162: (top) Culver Pictures; (bottom) From *PSSC Physics*, 2nd ed., DC Heath & Co., with Educational Development Center, Newton, MA. Page 163: (top) Culver Pictures; (bottom) *Atlas of Optical Phenomena* by M. Cagnet, M. Francon & J. C. Thrierr, Springer Verlag, Berlin 1962. Page 165: Educational Development Center, Newton, MA. Page 166: Reprinted from *University Physics* by Sears & Zemansky, 3rd, Addison-Wesley 1964. Page 167: *Atlas of Optical Phenomena*, M. Cagnet et al. Springer-Verlag, Berlin, 1962. Page 168: *Atlas of Optical Phenomena*, M. Cagnet et al. Springer-Verlag, Berlin 1962. Page 170: Culver Pictures.

CHAPTER 13
Opener: Bettmann Archive. Page 178: Carolina Biological Supply Co. Page 179: Bettman Archive. Page 182: AIP Neils Bohr Library.

CHAPTER 14
Opener: Fredrick D. Bodin/Stock Boston. Page 194: Bettmann Archive. Page 197: Photo by Jack Kohl.

CHAPTER 15
Opener: Jean-Claude Le Jeune/Stock Boston. Page 204: Bettmann Archive. Page 208: Bettmann Archive. Page 214: Brown Brothers.

CHAPTER 16
Opener: Grant White/Monkmeyer. Page 228: Brown Brothers. Page 237: University of California. Page 238: Bettmann Archive.

CHAPTER 17
Opener: Frank Siteman/Taurus. Page 246: Drawing by William Numeroff. Page 251: American Institute of Physics, Neils Bohr Library. Page 255: American Institute of Physics, Neils Bohr Library.

CHAPTER 18
Opener: Courtesy Neils Bohr Archive. Page 264: (top) Bettmann Archive; (bottom) Brown Brothers. Page 268: Reprinted with permission from *Introduction to Atomic & Nuclear Physics* by Harvey E. White D. Van Nostrand Co., Inc. 1964. Page 269: American Institute of Physics, Neils Bohr Library, Margrethe Bohr Collection.

CHAPTER 19
Opener: Courtesy AT&T Bell Labs. Page 280: American Institute of Physics, Neils Bohr Library. Page 283: Courtesy Sumio Iijima, Arizona State University. Page 284: Bettmann Archive. Page 286: American Institute of Physics, Neils Bohr Library. Marshak Collection. Page 288: Bettmann Archive. Page 294: From *Fundamentals of Physics* 2nd Ed. Extended, by Halliday and Resnick © 1981 John Wiley & Sons, Inc., New York.

CHAPTER 20
Opener: Educational Development Center. Page 298: American Institute of Physics, Neils Bohr Library.

CHAPTER 21
Opener: From "Highly Excited Atoms," by Daniel Kleppner, Michael G. Littman and Myron L. Zimmerman. Copyright © 1981 by Scientific American, Inc. All rights reserved. Page 329: Brown Brothers. Page 336: American Institute of Physics, Neils Bohr Library. Page 337: New York Public Library Picture Collection.

CHAPTER 22
Opener: American Museum of Natural History.

CHAPTER 23
Opener: © Paul Robert Perry. Page 363: (top) American Institute of Physics, Neils Bohr Library, E. Scott Barr Collection; (bottom) Bettmann Archive. Page 370: American Institute of Physics, Neils Bohr Library, E. Scott Barr Collection. Page 390: (top) Los Alamos Scientific Laboratory; (bottom) Bettmann Archive.

CHAPTER 24
Opener: Owen Franken/Stock Boston.

CHAPTER 25
Opener: SEH America, Inc.

CHAPTER 26
Opener: AT&T Bell Labs. Page 468: UPI.

CHAPTER 27
Opener: The Science Museum, London. Page 482: Culver Pictures. Page 499: UPI.

CHAPTER 28
Opener: Courtesy Texas Instruments. Page 506: Compliments Monsanto Electronic Materials Company. Page 508: Courtesy Bell Labs. Page 516: (top) Courtesy GE Research & Development Center; (bottom) Courtesy AT&T Bell Labs. Page 517: Reprinted by permission of Intel Corp.

Index

Absolute temperature scale, 115
Absorption edge, 449
Acceleration, 24
  angular, 85
  average, 24
  instantaneous, 24
  gravitational, 28
  motion with constant, 25
  radial, 88
  rotational, 85
  tangential, 84
Acceptor energy level, 438, 440
Acceptor impurities, 440
Accuracy of numbers, 6
Action and reaction, 39
Active region, 471
Alkali elements, 341
Alpha particles, 264
Ammeter, 219
Ampere, 205
Amplitude, 134, 147
AND gate, *see* Logic circuits
AND operation, 482
Angstrom, 163
Angular momentum, 106
  conservation of, 107
  quantization of, 269, 326
  spin, 331
Antisymmetric eigenfunction, 409, 413
Atmospheric pressure, 114
Atomic mass unit, 112
Atomic models, 264, 321
  Bohr model, 269
  Rutherford model, 265
  Thompson model, 264
  Quantum mechanical model, 321
Atomic number, 323
Atomic weight, 113
Avogadro, A., 112
Avogadro's number, 113

Band theory, 396
  of diamond, 417
  of germanium, 417
  of silicon, 417, 431
Band width, 413, 427, 428
Base, transistor, 465
Basis, 349
Battery, 197
Binary bit, 497
Blackbody radiation, 244
Bloch's theorem, 397
Bohr, N., 269
Bohr model of the atom, 269
Bohr's postulates, 269
Boltzmann, L., 126
Boltzmann constant, 120
Boltzmann factor, 128
Bonding, *see* Crystal bonding
Boole, G., 482
Boolean algebra, 482
Born, M., 284
Boundary conditions, 135
Boyle, R., 114
Boyle's law, 114
Bragg, W., 170
Bragg condition, 171
Bragg scattering, 170, 281, 385, 422
Bravais lattices, 349
Bremsstrahlung, 254
British thermal unit (BTU), 121

Calorie, 121
Capacitance, unit of, 198. *See also* Capacitor
Capacitor, 198
  charging, 223
Carrier density, 440
Cavendish, H., 91
Celsius scale, 113
Center of gravity, 68
Center of mass, 68
  motion of, 71
Centigrade scale, 113
Centripetal force, 89
Charge, 178
  force between charges, 178
  unit of, 181
Chip, 509
  component density in, 516
Circular motion, 88
Clock circuit, 498
  astable transistor multivibrator, 498
  free running oscillator clock, 499
Collector, transistor, 465
Collimator, 264
Collision time, 367
Collisions, 74
  elastic, 75, 276
  inelastic, 77, 276
Complementarity principle, 293
Complex Conjugate, 302
Compton, A., 255
Compton effect, 255
Conduction band, 416, 417
Conduction electrons, 363
Conductivity:
  electrical, 209, 362, 369, 382, 446
  thermal, 387

Conductors, 179, 415
Conservation:
  of angular momentum, 107
  of energy, 59, 61, 141, 196
  of linear momentum, 72
Contact potential, 456
Conventional current, 198
Conversion of units, 3
Coulomb, C., 179
Coulomb's law, 179
Covalent bond, 354
Cross product, 16
Cross-section, scattering, 386
Crystal bonding, 352
  covalent, 354
  ionic, 352
  metallic, 356
  van der Waals, 356
Crystal growth, 505
  Bridgman-Stockbarger method, 507
  Czochralski method, 506
  floating zone method, 507
  vapor-phase epitaxy, 507
Crystal structures, 348
Current, *see* Electric current
Current density, 207, 387
  heat, 387
Current gain parameter, 472

Davisson and Germer experiment, 281
De Broglie's hypothesis, 280
Degenerate states, 325, 372, 410, 413
DeMorgan theorems, 485
Density of states, 375, 431
Depletion layer, 457, 465, 475, 477, 479
Descartes, R., 39
Diamond structure, 351, 355, 416
Dielectric constant, 199, 439
Diffraction, electron, 281
  single slit, 166
  X-ray, 170
Diffraction grating, 166, 170
Diffusion, 456, 510
  across *pn* junctions, 456
  of impurities, 510
Diode, 461
Diode equation, 464
Diode transistor logic (DTL), 493
Dipole:
  electric, 185, 202
  magnetic, 232, 327
Dipole-dipole interaction, 357
Dispersion relation, 405
Donor energy level, 439
Donor impurities, 438
Doping, 509
Dot product, 15
Double slit interference, 162
Drain, 474
Drift velocity, 205, 365
Drude, P., 363
d states, 326

Effective mass, 396, 418, 431
Effective mass states, 423
  negative, 423
Eigenfunctions, 309
Eigenstates, 309
Eigenvalues, 309
Einstein, A., 251
Elastic limit, 133
Electric current, 205
  unit of, 205
Electric dipole, *see* Dipole
Electric field, 188
Electric potential, 194
  unit of, 195
Electric potential energy, 191
Electrical conductivity, *see* Conductivity
Electrical resistivity, 208, 362
Electrode, 197
Electrolyte, 197
Electromagnetic spectrum, 170, 238
Electromagnetic waves, 238
Electromotive force, 197
Electron, 179, 378, 430
  charge of, 182
  conduction, 363
  density, 430, 433, 440
  energy distribution, 377
  specific heat, 368, 380
  valence, 363
Electron volt, 197
Electronic specific heat, 368, 380
Electroscope, 178
emf, 198
Emitter, transistor, 465
Energy bands, 396, 402, 405, 416
Energy conservation, *see* Conservation
Energy gap, 417, 430, 433, 448
Energy quantization, 246, 270, 314, 325, 372
Epilayer, 512
Equilibrium, 43
Evaporation, thin film, 511
Excited states, 271
Exclusion principle, 336, 373, 391, 415, 431
Expectation values, 303, 316

Fahrenheit scale, 113
Fan-out, 495, 501
Farad, 198
Fermi-Dirac distribution, *see* Fermi-Dirac statistics
Fermi-Dirac statistics, 373, 378, 390, 431, 432
Fermi energy, 373, 375, 379, 391
  in semiconductors, 436, 437, 442, 443
  in metal-metal junctions, 454, 455
  in *pn* junctions, 458
Fermi level, *see* Fermi energy
Fermi velocity, 383
FET, 473
Fine structure, 331
First law of thermodynamics, 124
Flip-flop, 497
  reset-set (RS), 497
  data (D), 497

Forward bias, 461
Force, 38, 41, 42
  between charges, 178, 181
  centripetal, 89
  elastic, 132
  gravitational, 91
  magnetic, 229, 234
  unit of, 42
Franck-Hertz experiment, 274
Free electron models, 363
  classical, 363
  quantum-mechanical, 370
Frequency, 133, 146
  unit of, 133
Friction, 48, 121
  coefficient of, 48
f states, 326

Galileo, G., 39
Galvanometer, 219
Gamma ray detector, 451
Gas constant, 115
Gases, kinetic theory of, 111
Gate, in FETs, 474
Gates, *see* Logic circuits
Gauss, 230
Gay-Lussac's law, 115
Glasses, 348
Gradient, magnetic field, 329
Gram molecular weight, 113
Grating, diffraction, 166
Gravitation, law of universal, 91
Gravitational acceleration, 28
Gravitational constant, 91
Ground, 486
Ground state, 271, 315
Group velocity, 292, 419

Hall coefficient, 236, 425
Hall effect, 235, 425
Hall voltage, 235, 425
Halogens, 343
Harmonic oscillator, 322
Heat, 112, 121, 122
  mechanical equivalent of, 122
Heat capacity, 122
Heisenberg, W., 285
Hertz, H., 133, 238
Holes, 397, 422
  density of, 430, 436, 440
Hooke's law, 132
Horsepower, 63
Huygens, C., 162
Huygens' principle, 163
Hydrogen, 265, 323
  electromagnetic spectrum of, 267
  energy spectrum of, 270, 325

Ideal gas law, 114
Impulse, 40
Integrated circuits (IC), 493, 504
IC production, 508
  connection of components, 511
  diffusion, 510
  doping, 508
  fabrication, 512
  ion implantation, 510
  oxidation, 508
  photolithography, 508
Inertia, moment of, 98, 100
Inertial mass, 38
Insulator, 179, 415, 434
Intensity, 153, 284
Interference, 158
  constructive, 158
  destructive, 158
  double slit, 162
  from two sources, 159
Ionic bonding, 352
Ion implantation, 510
Isotopes, 112

JFET, 474
Joule, 54
Junctions,
  metal-metal, 454
  pn, 456

Kelvin, Lord, 115
Kelvin scale, 115
Kepler, J., 91
Kilomole, 112
Kinetic energy, 58, 101, 299
  rotational, 101
Kinetic theory of gases, 112
Kirchhoff, G., 214
Kirchhoff rules, 214, 472, 492
Kronig-Penney model, 398

Lattice, periodic, 349
Laue, M. von, 170, 253
Light:
  photon theory of, 251
  speed of, 237
  wavelength of, 238
  wave theory of, 237
Limit of resolution, 170
Linear motion, 25
Logic circuits, 485
  AND gate, 485
  diode AND gate, 490
  diode OR gate, 491
  diode switch, 488
  DTL gates, 493
  inverter, 488, 492
  NAND gates, 493, 495
  NOR gates, 493, 494
  OR gate, 487
  parameters, 495, 496
    fan-out, 495
    noise immunity, 496
    propagation time, 495
  RTL gates, 494
  TTL gates, 494
London force, 357
Lorentz, H. A., 363

Lorentz number, 387
LSI, 517

Madelung constant, 353
Magnetic dipole moment, 232
Magnetic field, 228
  unit of, 230
Magnetic force, 228
  on current carrying wire, 229
  on moving charge, 234
Magnetic moment, 232
  of hydrogen, 233, 327
Magnetic quantum number, 326
Majority carriers, 444
Mass, 38, 41
  atomic unit of, 112
  effective, 396, 418, 431
  unit of, 41
Matter waves, 280, 283, 289
Maxwell, J. C., 237
Maxwell-Boltzmann distribution, *see* Maxwell-Boltzmann statistics
Maxwell-Boltzmann statistics, 126, 364, 390, 432
Mean free path, 367, 369, 384
Mechanical equivalent of heat, 122
Memory circuit, 497
  D flip-flop, 497
  RS flip-flop, 497
Metallic bond, 356
Millikan, R., 182
Minority carriers, 444
Molar specific heat, 122
Mole, 112
Molecular weight, 112
Mole fraction, 113
Moment of inertia, 98, 100
Momentum, 39, 68, 106
  angular, 106
  conservation of, 72, 107
  linear, 72
  of photon, 260
Monochromator, 448
Moore's curve, 516
MOSFET, 476
MSI, 517
Multivibrator, 497

NAND gate, *see* Logic circuits
Neutrality equation, 441
Neutron star, 392
Newton, I., 38, 42
Newtonian body, 38
Newton's laws, 38
Newton's second law, 39, 204, 366
Nodes, 174
Noise immunity, 496
Nonohmic conductor, 208
NOR gate, *see* Logic circuits
Normal force, 43
Normalization, 285, 304, 315
NOT operation, 484

npn transistor, 465
n-type semiconductor, 439

Oersted, H., 228
Ohm, G., 208
Ohmic conductor, 208
Ohm's law, 208, 362, 364, 367
Oil drop experiments, 182
Operator, 299, 306, 307
Orbital motion, 91
Orbital quantum number, 326
OR gate *see* Logic circuits
OR operation, 483
Oscillatory motion, 131
Oxidation, 508

Parallel, resistors in, 210
Path difference, 160, 171
Pauli, W., 336
Pauli exclusion principle, 336, 373, 391, 415, 431
Period, 133, 146
Periodic potential, 397
Periodic table, 337
Phase angle, 134
Phase shift, 149
Photoconductivity, 448
Photoelectric cells, 449
Photoelectric effect, 247
  cut-off frequency, 248, 252
  retarding voltage, 248
  stopping potential, 248
Photoemission, 252
Photolithography, 508
Photon, 251, 283
  momentum of, 260
  propagation of, 283
Pinch-off region, 475
Planck, M., 246
Planck's constant, 246, 251
pn junction, 456
pnp transistor, 465
Polar molecule, 357
Position vector, 22
Potential, 194
  contact, 456
Potential energy, 56, 140, 191
  gravitational, 56
  electrical, 191
  of spring, 140
Potential energy well, 311, 370
  one-dimensional, 311
  three-dimensional, 370
Pound, 43
Power, 62, 105, 152
  rotational, 105
Powers of 10, 5
Principal quantum number, 326
Pressure, 114, 117
  kinetic theory of, 117
Probability density, 285
Projectile motion, 30
Propagation constant, 148

Propagation delay, 495
p states, 326
p-type semiconductor, 440

Quantum, 246, 272
Quantum mechanics, 298
Quantum number, 325, 372
  magnetic, 326
  orbital, 326
  principal, 326
  spin, 332
Quasicontinuous energy spectrum, 376

Radial acceleration, 88
Radian, 82
Radiancy, spectral, 245
Radiation, electromagnetic, 237
Radius vector, 133
Rare gases, 340
Rayleigh criterion, 170
RC circuits, 222
Rectification, 461
Relay switch, 486
Resistance, 209
  in parallel, 210
  in series, 210
  unit of, 209
Resistivity, 208, 362
Resistor, 210
  power dissipation in, 221
Resistor transistor logic (RTL), 494
Resolving power, 168
Resultant, 10, 14
Reverse bias, 461
Reverse saturation current, 464
Right-hand rule, 17, 228
Root-mean-square velocity, 120
Rotational dynamics, 98
Rotational motion, 82, 86
Rutherford, E., 264
Rydberg constant, 268, 273
Rydberg-Ritz formula, 267

Saturation region, 471
Scalar, 10
Schrödinger, E., 298
Schrödinger equation, 298, 300, 309, 323
  free-particle, 300
  for the H atom, 323
  time-independent, 309
Semiconductor, 429
  electrical conductivity of, 446
  diode, 456
  doped, 439
  extrinsic, 438
    n-type, 439
    n-type, Fermi level, 441
    p-type, 440
    p-type Fermi level, 443
  impurity, *see* Semiconductor, extrinsic
  intrinsic, 430, 436
    Fermi level, 436
Semiconductor devices, 453
Separation of variables, 308, 323
Series, resistors in, 210
Shunt, 219
SI units, 3
  basic units, 3
  units of:
    capacitance, 198
    charge, 181
    current, 205
    electric potential, 195
    energy, 56, 58
    force, 42
    frequency, 133
    length, 3
    magnetic field, 230
    mass, 3
    power, 62
    pressure, 114
    resistance, 209
    time, 3
    work, 54
Significant figures, 7
Silicon:
  band structure, 417
  crystal structure, 351, 438
  purification, 504
  wafers, 504, 516
Simple harmonic motion, 134
Single slit diffraction, 166
SLSI, 517
Slug, 43
Sommerfeld, A., 370
Source, 474
Space lattice, 349
Space quantization, 326
Specific heat, 121, 368, 380
  electronic, 368, 380
  molar, 122
Spectral radiancy, 245
Spectrum:
  electromagnetic, 238
  of hydrogen, 267
Speed, 22
Spin, 331
Spring constant, 132
s states, 326
Standing waves, 173, 315
Statistical distribution, 126, 390
  Fermi-Dirac, 390
  Maxwell-Boltzmann, 126
Stefan-Boltzmann law, 245
Stern-Gerlach experiment, 329
Subshell, 338
Superposition principle, 158, 183
Symmetric eigenfunction, 409, 413

Tangential acceleration, 84
Tangential velocity, 83
Temperature, 112, 119
  kinetic theory of, 119
Tension, 41
Tesla, 230
Thermal conductivity, 387

Thermodynamics, first law of, 124
Thermometers, 113
Thompson, J. J., 264
Tight-binding approximation, 407
Time, 38
  collision, 367
  constant in RC circuit, 223
  unit of, 3
Torque, 98
  on current loop, 230
Transistor, 465
  base, 465
  bipolar junction (BJT), 465
  collector, 465
  common base configuration, 465
  common emitter configuration, 471
  current amplifier, 473
  current gain parameter, 472
  emitter, 465
  field effect (FET), 473
  inverter, 492
  junction field-effect (JFET), 474
  load line, 492
  metal-oxide-semiconductor (MOSFET), 476
  transistor logic (TTL), 495
  unipolar, 474
  voltage amplifier, 469
Traveling wave, 147
Truth table, 482
  for AND operation, 483
  for OR operation, 483

Uncertainty principle, 285
Unit vectors, 14

Valence band, 416
Valence electrons, 344, 363, 370
Van der Waals bond, 356
Vector, 10
  components, 10, 14
  cross product, 16
  displacement, 22
  dot product, 15
  position, 22
  radius, 133
  resultant, 10, 14
Velocity, 22
  angular, 83
  average, 22
  drift, 205, 365
  distribution of electrons, 382
  group, 292, 419
  instantaneous, 23
  root-mean-square, 120
  rotational, 83
  tangential, 83
  wave, 147
VLSI, 517
Volt, 195
Volta, A., 194
Voltage, 195
Voltage-current characteristics
  of diode, 461, 464
  of JFET, 475
  of MOSFET, 476
  of transistor, 471
Voltage difference, 195
Voltage drop, 210
Voltmeter, 210, 219

Water molecule, 357
Watt, 62, 221
Wave front, 162
Wavefunction, 285
  of hydrogen, 333
  required properties of, 309
  significance of, 285, 302
Wavelength, 147
Wave number, 148
Wave packet, 290
  group velocity, 292
Wave-particle duality, 293
Wave(s), 146, 147
  amplitude of, 146
  electromagnetic, 237
  frequency of, 146
  matter, 280, 283, 289
  intensity of, 153
  standing, 173, 315
  traveling, 147
  velocity of, 146, 147
  wavelength of, 146
Weight, 42
Well, infinite potential, 311, 370
Wheatstone bridge, 225
White dwarf star, 392
Wiedemann-Franz law, 387
Wien's displacement law, 260
Work, 54, 123
  by constant force, 54
  by gas, 123
  by variable force, 57
  unit of, 54
Work-energy theorem, 59, 102, 141
Work function, 251, 454

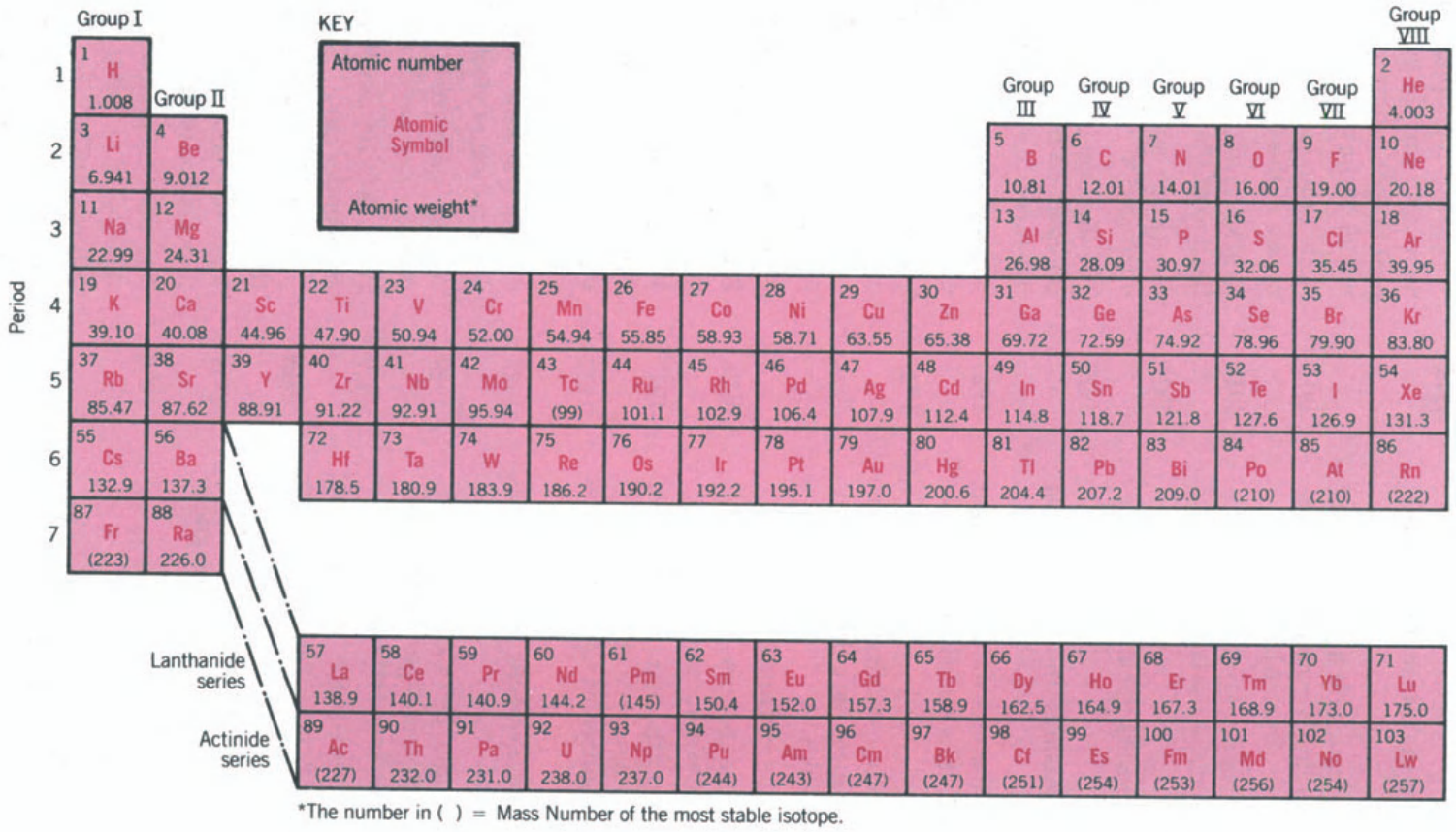
X-rays, 253
  characteristic, 254
  crystal structure and, 348, 438
  cut-off wavelength, 254
  diffraction of, 170
  production of, 253
  spectrum, 254
Xerography, 449

Young, T., 163
Young's double slit experiment, 162

Zeeman, P., 326
Zeeman effect, 326
Zone refining, 504

Milestones in the Development of the Concepts Presented in this Book

1634	Galileo Galilei publishes his *Dialoge* in which he presents his ideas on the motion of the solar system.
1660	Robert Boyle enunciates the pressure-volume law of gases.
1664	The work of René Descartes on his concepts of physical science, including momentum, is published posthumously in Amsterdam.
1687	Isaac Newton publishes his *Principia* enunciating three laws of motion and the law of universal gravitation.
1786–89	Charles Coulomb discovers the law of force between charges.
1798	Henry Cavendish measures the gravitational constant.
1801	Thomas Young demonstrates the phenomenon of interference of light, thus establishing its wave nature.
1811	Amadeo Avogadro proposes that gram-molecular weights of substances have the same number of atoms or molecules.
1820	Hans Oersted discovers electromagnetism.
1827	Georg Ohm states the relationship between voltage, current, and resistance.
1823–71	Charles Babbage designs the first programmable digital calculator with memory. Although it was never built, he is recognized as the father of the modern computer.
1847–54	George Boole pioneers symbolic logic and develops boolean algebra.
1850s	Julius Mayer, James Joule, Hermann von Helmholtz, and L. A. Colding independently establish the equivalence of heat and energy.
1873	James Maxwell synthesizes the laws of electromagnetism and predicts the existence of electromagnetic waves.
1885	Heinrich Hertz verifies experimentally Maxwell's theory of the existence of electromagnetic waves.
1900	Max Planck introduces quantum theory to explain blackbody radiation.
1905	Albert Einstein proposes the quantization of electromagnetic radiation (photons) to explain the photoelectric effect.
1909	Robert Millikan performs oil drop experiments that determine the charge of the electron.
1911	To explain the alpha particle scattering experiments of Hans Geiger and Ernest Madden, Ernest Rutherford proposes the nuclear planetary model of the atom.
1913	Neils Bohr proposes the quantum mechanical model of the hydrogen atom.
1914	James Franck and Gustav Hertz verify the quantization of the internal energy of atoms.
1921	Otto Stern and Walter Gerlach discover space quantization.
1923	The increase in wavelength of scattered X rays is explained by Arthur Compton with the photon theory.
1924	Louis de Broglie proposes the wave nature of particles.
1925	Wolfgang Pauli enunciates the exclusion principle.
1925	To explain the fine structure in the spectrum of atoms, Samuel Goudsmit and George Uhlenbeck introduce the concept of the electron spin.
1926	Erwin Schrödinger presents his theory of quantum mechanics.
1927	Max Born suggests the probabilistic interpretation of the wavefunction.
1927	Werner Heisenberg presents the Uncertainty Principle.
1927	Clinton Davisson and Lester Germer verify the wave nature of electrons.
1928	Arnold Sommerfeld presents the quantum mechanical free electron theory of solids.
1939	A. H. Wilson develops the general model of semiconductors.
1946	Johann von Neumann develops the concepts of the modern computer mainframe.
1948	John Bardeen, Walter Brattain, and William Shockley construct the first transistor.
1958	The first integrated circuit on a silicon chip is developed at Texas Instrument and Fairchild Semiconductor.
1976	Intel announces the production of an eight-bit computer on a single silicon chip.



*The number in ( ) = Mass Number of the most stable isotope.

Atomic and Semiconductor Data

Electronic charge	1.6×10^{-19} C
Mass of the electron	9.11×10^{-31} kg
Mass of the proton	1.67×10^{-27} kg
Mass of the neutron	1.67×10^{-27} kg
Bohr radius	5.3×10^{-11}m
Ionization energy of hydrogen	13.6 eV
Effective mass of electrons in silicon	$0.31 \times 9.11 \times 10^{-31}$ kg
Effective mass of holes in silicon	$0.38 \times 9.11 \times 10^{-31}$ kg
Energy gap (E_g) in silicon	1.1 eV
Effective mass of electrons in germanium	$0.12 \times 9.11 \times 10^{-31}$ kg
Effective mass of holes in germanium	$0.23 \times 9.11 \times 10^{-31}$ kg
Energy gap (E_g) in germanium	0.67 eV

Conversion Factors

1 eV = 1.6×10^{-19} J
1 Å = 10^{-10} m
1 calorie = 4.19 J
1 u = 1.66×10^{-27} kg

Some General Constants

Avogadro's number	$N_A = 6.02 \times 10^{23}$ molecules/mole
Boltzmann constant	$k_B = 1.38 \times 10^{-23}$ J/K
	$= 8.63 \times 10^{-5}$ eV/K
Coulomb constant	$1/4\pi\varepsilon_O = 8.99 \times 10^9$ N-m^2/C^2
Gravitational constant	$G = 6.67 \times 10^{-11}$ N-m^2/kg^2
Permittivity of free space	$\varepsilon_O = 8.85 \times 10^{-12}$ C^2/N-m^2
Planck constant	$h = 6.63 \times 10^{-34}$ J-sec
	$= 4.14 \times 10^{-15}$ eV-sec
Speed of light	$c = 3.00 \times 10^8$ m/sec
Universal gas constant	$R = 8.31$ J/mole-K

Mathematical Symbols

$\neq$ is not equal to
$\approx$ is approximately equal to
$\sim$ is of the order of
$\propto$ is proportional to
$>$ is greater than
$\geq$ is greater than or equal to
$<$ is less than
$\gg$ is much greater than
$\ll$ is much less than