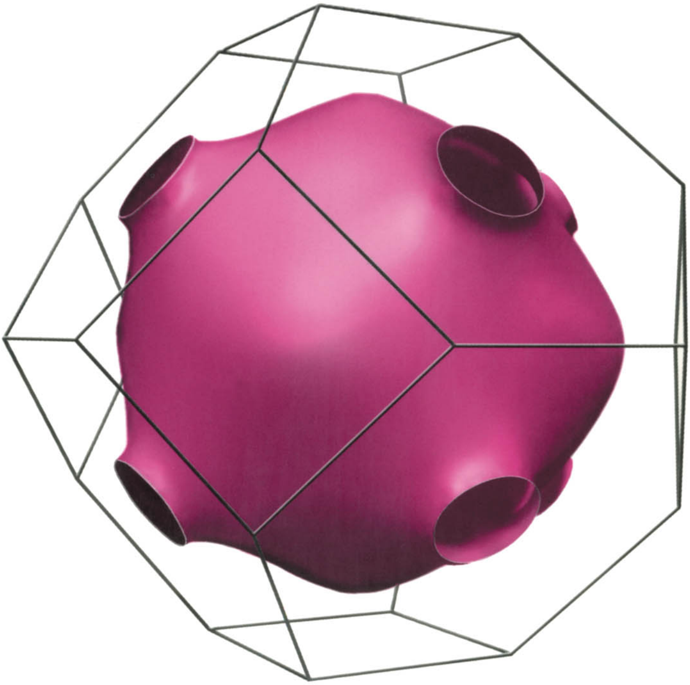
Introduction to the

ELECTRONIC PROPERTIES of MATERIALS

DAVID JILES



SECOND EDITION

nelson thornes

Introduction to the Electronic Properties of Materials

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Introduction to the Electronic Properties of Materials

SECOND EDITION

David Jiles
Ames Laboratory, US Department of Energy
and
Department of Materials Science and Engineering
and
Department of Electrical and Computer Engineering
Iowa State University

First published in 1994 by:
Chapman & Hall

Second edition published in 2001 by:

Nelson Thornes Ltd
Delta Place
27 Bath Road
CHELTENHAM
GL53 7TH
United Kingdom

Transferred to Digital Printing 2010

A catalogue record for this book is available from the British Library

ISBN 0 7487 6042 3

Page make-up by Aarontype Limited, Easton, Bristol

The cover illustration represents the three-dimensional Fermi surface of copper.

Publisher's Note
The publisher has gone to great lengths to ensure the quality of this reprint but points out that some imperfections in the original may be apparent.

CONTENTS

PART ONE: *FUNDAMENTALS OF ELECTRONS IN MATERIALS*

PART TWO: *PROPERTIES OF MATERIALS*

PART THREE: *APPLICATIONS OF ELECTRONIC MATERIALS*

PREFACE TO THE FIRST EDITION

The subject of electronics, and in particular the electronic properties of materials, is one which has experienced unprecedented growth in the last thirty years. The discovery of the transistor and the subsequent development of integrated circuits has enabled us to manipulate and control the electronic properties of materials to such an extent that the entire telecommunications and computer industries are dependent on the electronic properties of a few semiconducting materials. The subject area is now so important that no modern physics, materials science or electrical engineering degree programme can be considered complete without a significant lecture course in electronic materials. Ultimately the course requirements of these three groups of students may be quite different, but at the initial stages of the discussion of electronic properties of materials, the course requirements are broadly identical for each of these groups. Furthermore, as the subject continues to grow in importance, the initial teaching of this vital subject needs to occur earlier in the curriculum in order to give the students sufficient time later to cover the increasing amount of material.

It is with these objectives in mind that the present book has been written. It is aimed at undergraduates who have only an introductory knowledge of quantum mechanics. The simplified approach used here enables the subject to be introduced earlier in the curriculum. The goal at each stage has been to present the principles of the behaviour of electrons in materials and to develop a basic understanding with a minimum of technical detail. This has resulted in a discussion in breadth rather than depth, which touches all of the key issues and which provides a secure foundation for further development in more specialized courses at a later stage. The presentation here should be of interest to two groups of students: those who have a primary interest in electronic materials and who need an introductory text as a stepping-stone to more advanced texts; and those whose primary interest lies elsewhere but who would nevertheless benefit from a broad, passing knowledge of the subject.

As with the earlier textbook, *Introduction to Magnetism and Magnetic Materials* (1991) the subject area under discussion here is truly multidisciplinary, spanning the traditional subject areas of physics, electrical engineering and materials science. In writing this book I have striven to keep this in mind in order to maintain the interest of a wider audience. Therefore some of the treatment will seem relatively easy for one group of students while relatively hard for another. Over the entire book, however, I think that the general mix of subject areas leads to a text that is equally difficult for these three groups of students. Chapters 1–5 could easily be included in a traditional solid-state physics course and should be very familiar to physicists. However Chapters 6–10 will appeal more to materials scientists since they will be more familiar with dealing with meso- and macroscopic properties. Finally Chapters 11–15 discuss the functional performance of these materials in technological applications which are likely to be of most interest to electrical engineers. These chapters provide a rapid introduction to five important applications of electronic materials, each of which could be further developed in a separate advanced course. Also, as in *Introduction to Magnetism and Magnetic Materials*, the early chapters of this book contain a number of key exercises for the student to attempt. Completed worked solutions are given at the back of the book. It has been my experience that this is much more useful than simply giving a numerical answer at the back, since if you do not get the problem exactly right under those conditions, you cannot easily find out where you went wrong!

On completion of the text the reader should have gained an understanding of the behaviour of electrons within materials, an appreciation of how the electrons determine the magnetic, thermal, optical and electrical properties of materials and an awareness of how these electronic properties are controlled for use in a number of important technological applications. I hope that the text will provide a useful introduction to more detailed books on the subject and that it will also provide the background for developing the interest of students in this fascinating subject at an early stage in their careers.

Finally, I would like to acknowledge the assistance of several friends and colleagues who have helped me in writing this book. In particular thanks go to M. F. Berard, F. J. Friedlaender, R. D. Greenough, R. L. Gunshor, J. Mallinson, R. W. McCallum, R. E. Newnham, S. B. Palmer and A. H. Silver.

DJ, Ames, Iowa

ACKNOWLEDGEMENTS

I am grateful to those publishers credited in captions for permission to reproduce some of the figures in this book.

PREFACE TO THE SECOND EDITION

Electronic materials provide the basis for many of our 'high tech' industries such as computers, semiconductors, data storage, electronic devices, sensors and actuators. In particular, the range of available materials and their technological applications, have made enormous progress in the seven years since the first edition of this book was published. So it is timely to bring the book up to date. I have chosen to maintain the same basic layout as in the first edition. The early chapters contain the basic concepts and are, in places, rather abstract and mathematical. The later chapters describe applications and are more descriptive and practical. In this way, I have deliberately sought to maintain a blend and balance between a need for basic understanding of ideas and a description of how these are incorporated into applications.

The underlying physics and physical descriptions of these materials change only slowly with time, so that the information contained in the early chapters of the first edition of the book has remained as relevant to today as it was seven years ago. Therefore in Chapters 1–10, I have chosen simply to expand the number of exercises with complete worked solutions to offer a wider range of examples that will deepen knowledge of the underlying physical basis for understanding these materials. These examples have been used over the last few years by students at Iowa State University to consolidate concepts presented in lecture classes. In a few instances additional information on topics such as anharmonicity and Gruneisen parameters, Bragg reflection at Brillouin zone boundaries and more detailed descriptions of charge carriers in the conduction and valence bands at semiconductor junctions have been added to increase the scope of the chapters where it seemed appropriate.

The technological applications in the areas of semiconductor materials and devices, computer technology, data storage and magnetics have all seen remarkable progress since the first edition. Therefore in the later chapters of the book that deal with technological applications, Chapters 11–15, it was essential to provide more up-to-date information. In microelectronics there has been an expansion of the number of materials that are available and now being used in technological applications. The continual reduction of device and feature sizes in microelectronics has allowed a dramatic increase in the number of components on a single semiconductor chip. In optoelectronics, particularly the applications to data communication, there have been important developments both on the small scale (local computer communications) and the large scale (intercontinental telecommunications). Once again, the introduction of new materials with improved performance over previously available materials has been the enabling technology. In some cases, optoelectronic materials with completely new characteristics (such as the optical amplifier materials based on rare-earth doped silicon) have been introduced. In superconductivity, after the surge of interest in the high-temperature ceramic superconducting materials prior to the first edition, the major advances subsequently have been in applications, rather than in the identification of new materials. So new sections on superconducting wires, superconducting energy storage devices, and superconducting transformers, motors and generators have been added. In magnetic recording the rate of progress has accelerated since 1994 so that today, as a result of the availability of new multi-layered giant magneto-resistive materials, data storage densities are doubling every 9 months instead of every two years as they were in the early 1990s. In 1994 a typical personal computer came with a hard disk drive that had 500 Mbytes of capacity. In the year 2000 a typical computer came with a hard disk drive that had about 80 Gbytes of memory. The availability of new magnetic materials has made these advances possible. Data storage technology has moved so far since the first edition that the whole chapter on magnetic recording had to be rewritten. In the area of transducers, sensors and actuators new ferroelectric thin film materials have been developed, and applications have been found in dynamic random access memory (DRAM) and nonvolatile ferroelectric random access memory (FRAM) and microelectronic mechanical machines (MEMs).

In selecting material for this new edition I have attempted to focus on the advances in major technological areas. Clearly, in order to keep any book on such an important and diverse subject as electronic materials to a reasonable size, many interesting areas must necessarily be left to others. However, I believe that the information contained in this second edition will provide a sound introduction to this subject. The essential concepts that allow understanding of these materials is included, together with a description of the most important materials and their technological applications.

DJ, Ames, Iowa

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank various friends and colleagues who have helped to make this second edition possible, either through helpful discussions, indicating corrections to the original book or through suggestions for topics to include. I particularly would like to thank M. J. Sablik, J. E. Snyder, F. J. Friedlaender, H. Hauser, D. L. Atherton, R. D. Greenough, F. Salas and D. P. Cann.

GLOSSARY OF SYMBOLS

A	Area
a	Distance
$\mathbf{a}$	Lattice spacing
α	Mean field constant
	Optical attenuation coefficient
$\boldsymbol{B}$	Magnetic induction (magnetic flux density)
B_{R}	Remanent magnetic induction
B_{s}	Saturation magnetic induction
C	Capacitance
	Curie constant
	Specific heat or heat capacity
C^{e}	Electronic specific heat
C^{l}	Lattice specific heat
C_{v}	Specific heat capacity at constant volume
C_{P}	Specific heat capacity at constant pressure
c	Velocity of light
χ	Magnetic susceptibility
χ_{P}	Pauli paramagnetic susceptibility
$\boldsymbol{D}$	Electric displacement (electric flux density)
$D(\omega)$	Vibrational density of states
	Phonon density of states
$D(E)$	Density of available energy states
d	Diameter
	Distance
δ	Optical penetration depth
E_{F}	Fermi energy
ξ	Electric field strength
E_{B}	Elastic (bulk) modulus
	Binding energy
E_{Y}	Elastic (Young's) modulus
E_{S}	Elastic (shear) modulus
E	Energy and electric field
E_{c}	Cohesive energy
e	Electronic charge
	Strain
E_{a}	Anisotropy energy
E_{ex}	Exchange energy

E_{H}	Magnetic field energy (Zeeman energy)
ξ_{Hall}	Hall field
E_{K}	Kinetic energy
E_{loss}	Energy loss
E_{P}	Potential energy
$E_{\mathrm{P}}(x)$	Potential energy at location x
E_σ	Stress energy
ε	Permittivity (dielectric constant)
ε_1	Real component of dielectric constant (polarization)
ε_2	Imaginary component of dielectric constant (absorption)
ε_0	Permittivity of free space
F_{int}	Internal, or interactive, force
F	Force
F_{app}	Applied force
f	Fermi function
g	Transducer generation coefficient
	Spectroscopic splitting factor
	Lande splitting factor
	Rate of generation of charge carriers
γ	Gyromagnetic ratio
H	Magnetic field strength
h	Planck's constant
$\hbar$	Planck's constant divided by 2π
H_{c}	Coercivity
	Critical field
H_{cr}	Remanent coercivity
H_{d}	Demagnetizing field
H_{e}	Weiss mean field
H_{eff}	Effective magnetic field
I	Magnetic polarization (intensity of magnetization)
I	Intensity of light
	Electric current
J_{Q}	Thermal current density
J	Electric current density
	Atomic angular momentum
J	Total atomic angular momentum quantum number
	Exchange constant
j	Total electronic angular momentum quantum number
J_{atom}	Exchange integral for an electron on an atom with electrons on several nearest-neighbours
J_{ex}	Exchange integral; exchange interaction between two electrons

K	Anisotropy constant
	Thermal conductivity
k	Optical extinction coefficient
	Interatomic force constant
	Coupling coefficient of transducer
	Wave vector
k_B	Boltzmann's constant
K_{u1}	First anisotropy constant for uniaxial system
K_{u2}	Second anisotropy constant for uniaxial system
K_1	First anisotropy constant for cubic system
K_2	Second anisotropy constant for cubic system
L	Inductance
	Length
	Electronic orbit length
	Macroscopic length of lattice chain
	Length of side of cubic specimen
$\mathbf{L}$	Atomic orbital angular momentum
l	Length
l_0	Unstrained length
Δl	Change in length
l	Orbital angular momentum quantum number
$\mathscr{L}(x)$	Langevin function of x
λ	Wavelength
	Magnetostriction
	Penetration depth in superconductor
λ_d	Penetration depth
λ_t	Transverse magnetostriction
λ_s	Saturation magnetostriction
λ_0	Spontaneous bulk magnetostriction
M	Magnetization
m	Magnetic moment
	Momentum
M	Mass
M_{an}	Anhysteretic magnetization
m_e	Electronic mass
m_ℓ	Orbital magnetic quantum number
m_0	Orbital magnetic moment of electron
M_R	Remanent magnetization
M_0	Saturation magnetization (spontaneous magnetization at 0 K)
M_s	Spontaneous magnetization within a domain
m_s	Spin magnetic moment of electron
m_s	Spin magnetic quantum number

m_{tot}	Total magnetic moment of atom
m^*	Effective mass of electrons in bands
μ	Permeability
	Mobility of charge carriers
μ_{B}	Bohr magneton
μ_0	Permeability of free space
$N(E)$	Density of occupied energy states ($=2D(E)f(E)$), electron population density
N	Number of atoms per unit volume
	Number of electrons per unit volume
	Number of turns on coil or solenoid
n	Refractive index
	Principal quantum number
	Number of atoms
N_0	Avogadro's number
$N_0(E)$	Total number of energy states between zero energy and energy E
N^*	Effective number of conduction electrons
ν	Frequency ($\omega/2\pi$)
ω	Angular frequency ($2\pi\nu$)
P	Pressure
$\mathbf{P}$	Polarization
$P(E)$	Probability of occupancy of state with energy E
$P(x)$	Probability of electron being at location x
	Angular momentum operator
P_{o}	Orbital angular momentum of electron
P_{s}	Spin angular momentum of electron
P_{tot}	Total angular momentum of electron
$\mathbf{p}$	Angular momentum
π	Peltier coefficient
Φ	Magnetic flux
ϕ	Angle
	Work function
	Spin wave function
Ψ	Total wave function
ψ	Electron wave function
q	Electric charge
Q	Quantity of heat
R	Resistance
	Reflectance
r	Radius vector

r	Interatomic separation
	Radius
	Radius of ionic cores of atoms in lattice
	Electronic orbit radius
	Rate of recombination of charge carriers
ρ	Density
	Resistivity
ρ_mag	Magnetoresistivity
S	Atomic spin angular momentum
S	Entropy
s	Electronic spin angular momentum quantum number
σ	Conductivity
	Stress
T	Temperature
t	Time
	Thickness
T_c	Curie temperature
	Critical temperature
t_0	Orbital period of electron
θ	Angle
τ	Torque
τ	Time constant
	Relaxation time
	Charge carrier lifetime
τ_max	Maximum torque
U	Internal energy
u	Unit vector
u	Displacement of an atom from equilibrium
v_t	Final velocity (terminal velocity) of electrons in Drude model
v	Velocity
v_F	Velocity of electrons at Fermi surface
V	Electrical potential
V	Voltage (potential difference)
	Volume
W	Power
W_a	Atomic weight
W_H	Hysteresis loss
x	Distance along x-axis

y	Distance along y-axis
Z	Impedance Atomic number
z	Distance along z-axis
Σ	Number of nearest-neighbour atoms

SI UNITS, SYMBOLS AND DIMENSIONS

	Unit		MKSA	
Quantity	Symbol	Name	Base units	Dimensions
Length	m	metre	m	L
Mass	kg	kilogram	kg	M
Time	s	second	s	T
Frequency	Hz	hertz	$\mathrm{s^{-1}}$	T^{-1}
Force	N	newton	$\mathrm{kg\,m\,s^{-2}}$	MLT^{-2}
Pressure	Pa	pascal	$\mathrm{kg\,m^{-1}\,s^{-2}}$	$ML^{-1}T^{-2}$
Energy	J	joule	$\mathrm{kg\,m^{2}\,s^{-2}}$	$ML^{2}T^{-2}$
Power	W	watt	$\mathrm{kg\,m^{2}\,s^{-2}}$	$ML^{2}T^{-3}$
Electric charge	C	coulomb	$\mathrm{A\,s}$	CT
Electric current	A	ampere	A	C
Electric potential	V	volt	$\mathrm{kg\,m^{2}\,A^{-1}\,s^{-3}}$	$ML^{2}C^{-1}T^{-3}$
Resistance	Ω	ohm	$\mathrm{kg\,m^{2}\,A^{-2}\,s^{-3}}$	$ML^{2}C^{-2}T^{-3}$
Resistivity	Ω m	ohm metre	$\mathrm{kg\,m^{3}\,A^{-2}\,s^{-3}}$	$ML^{3}C^{-2}T^{-3}$
Conductance	S	siemens	$\mathrm{kg^{-1}\,m^{-2}\,A^{2}\,s^{3}}$	$M^{-1}L^{-2}C^{2}T^{3}$
Conductivity	$\mathrm{S\,m^{-1}}$	siemens/metre	$\mathrm{A^{2}\,s^{3}\,kg^{-1}\,m^{-3}}$	$M^{-1}L^{-3}C^{2}T^{3}$
Capacitance	F	farad	$\mathrm{A^{2}\,s^{4}\,kg^{-1}\,m^{-2}}$	$M^{-1}L^{-2}C^{2}T^{4}$
Electric displacement (flux density)	$\mathrm{C\,m^{-2}}$	coulomb metre^{-2}	$\mathrm{A\,s\,m^{-2}}$	$CL^{-2}T$
Electric field	$\mathrm{V\,m^{-1}}$	volt metre^{-1}	$\mathrm{kg\,m\,A^{-1}\,s^{-3}}$	$MLC^{-1}T^{-3}$
Electric polarization	$\mathrm{C\,m^{-2}}$	coulomb metre^{-2}	$\mathrm{A^{2}\,s\,m^{-2}}$	$CL^{-2}T$
Permittivity	$\mathrm{F\,m^{-1}}$	farad/metre	$\mathrm{A\,s^{4}\,kg^{-1}\,m^{-3}}$	$M^{-1}L^{-3}C^{2}T^{4}$
Inductance	H	henry	$\mathrm{kg\,m^{2}\,A^{-2}\,s^{-2}}$	$ML^{2}C^{-2}T^{-2}$
Magnetic flux	Wb	weber	$\mathrm{kg\,m^{2}\,A^{-1}\,s^{-2}}$	$ML^{2}C^{-1}T^{-2}$
Magnetic induction (flux density)	T	tesla	$\mathrm{kg\,A^{-1}\,s^{-2}}$	$MC^{-1}T^{-2}$
Magnetic field	$\mathrm{A\,m^{-1}}$	ampere/metre	$\mathrm{A\,m^{-1}}$	CL^{-1}
Magnetization	$\mathrm{A\,m^{-1}}$	ampere metre^{-1}	$\mathrm{A\,m^{-1}}$	CL^{-1}
Permeability	$\mathrm{H\,m^{-1}}$	henry/metre	$\mathrm{kg\,m\,A^{-2}\,s^{-2}}$	$MLC^{-2}T^{-2}$

VALUES OF SELECTED PHYSICAL CONSTANTS

Avogadro's number	$N_0 = 6.022 \times 10^{26}$ atoms kg mol^{-1}
Boltzmann's constant	$k_B = 1.381 \times 10^{-23}\,\mathrm{J\,K^{-1}}$
Gas constant	$R = 8.314\,\mathrm{J\,mol^{-1}\,K^{-1}}$
Planck's constant	$h = 6.626 \times 10^{-34}\,\mathrm{J\,s}$
	$h/2\pi = 1.054 \times 10^{-34}\,\mathrm{J\,s}$
Velocity of light in empty space	$c = 2.998 \times 10^{8}\,\mathrm{m\,s^{-1}}$
Permittivity of empty space	$\varepsilon_0 = 8.854 \times 10^{-12}\,\mathrm{F\,m^{-1}}$
Permeability of empty space	$\mu_0 = 1.257 \times 10^{-6}\,\mathrm{H\,m^{-1}}$
Atomic mass unit	$\mathrm{amu} = 1.661 \times 10^{-27}\,\mathrm{kg}$
Properties of electrons	
Electronic charge	$e = -1.602 \times 10^{-19}\,\mathrm{C}$
Electronic rest mass	$m_e = 9.109 \times 10^{-31}\,\mathrm{kg}$
Charge to mass ratio	$e/m_e = 1.759 \times 10^{11}\,\mathrm{C\,kg^{-1}}$
Electron volt	$\mathrm{eV} = 1.602 \times 10^{-19}\,\mathrm{J}$
Properties of protons	
Proton charge	$e_P = 1.602 \times 10^{-19}\,\mathrm{C}$
Rest mass	$m_P = 1.673 \times 10^{-27}\,\mathrm{kg}$
Gyromagnetic ratio of proton	$\gamma_P = 2.675 \times 10^{8}\,\mathrm{Hz\,T^{-1}}$
Magnetic constants	
Bohr magneton	$\mu_B = 9.274 \times 10^{-24}\,\mathrm{A\,m^2}\ (=\mathrm{J\,T^{-1}})$
	$= 1.165 \times 10^{-29}\,\mathrm{J\,m\,A^{-1}}$
Nuclear magneton	$\mu_N = 5.051 \times 10^{-27}\,\mathrm{A\,m^2}\,(=\mathrm{J\,T^{-1}})$
magnetic flux quantum	$\Phi_0 = 2.067 \times 10^{-15}\,\mathrm{Wb}\ (=\mathrm{V\,s})$

FOREWORD FOR THE STUDENT

The objective of this book is to present an introduction to the electronic properties of materials that is broad in its coverage but not exhaustive. The book focuses on the understanding of a few basic principles of the behaviour of electrons in materials and uses them to provide a description of a wide range of phenomena including magnetic, electrical, thermal and optical properties of materials. I have also given a number of historical references in the text, particularly in the early chapters. It seems to me that an appreciation of the historical development of a subject helps the overall understanding, apart from which it is interesting to know who originally developed the underlying ideas and even to re-read some of these landmark papers.

It has been my experience that, with the possible exception of the prospective specialist in solid-state physics, the majority of students do not benefit greatly from being confronted with a mass of detailed results arising from the theory of electrons in solids. This can come later for the intending specialist. In introducing this subject it seems more useful to present a few key results based on relatively simple models, which give a general feel for the behaviour of electrons in materials and how they contribute to the observed properties. These models themselves need not be particulary complex to be useful. For example, the basic premises of both the classical Drude model and the Sommerfeld model are quite far from reality. Yet the predictions that they make about the properties of the material contain some of the essential known results, for example the Wiedemann–Franz law and the electronic contribution to the heat capacity.

Therefore, the general approach taken here has been to introduce and discuss the consequences of such simple models which can be used to guide our thinking. We begin on the level of a few electrons subjected to an electrostatic potential due to the rest of the material. Subsequently the bulk properties of materials are considered and the phenomena are related to the earlier discussion of the behaviour of electrons. Finally several key applications are discussed, in which the electronic properties of materials play the central role in determining the suitability of materials for these applications. In particular the areas of microelectronics, optoelectronics, superconductivity, magnetism and piezoelectricity are examined.

PART ONE

FUNDAMENTALS OF ELECTRONS IN MATERIALS

1 PROPERTIES OF A MATERIAL CONTINUUM

OBJECTIVE

The objective of this chapter is simply to remind ourselves of the macroscopic properties of materials and to point out that in uses of electronic materials we are mostly interested in these bulk properties, which are the ones that we usually measure. In order to measure these properties it is necessary that we also give exact definitions of the various quantities. The microscopic properties are of interest because they help us to explain the variation of the macroscopic properties with external conditions, including any interrelationships which exist between the macroscopic properties. Once we have achieved an understanding of the relationship between macroscopic properties and the microscopic structure of a material it becomes possible to control the structure in order to produce materials with specific desired properties.

1.1 RELATIONSHIPS BETWEEN MACROSCOPIC PROPERTIES OF MATERIALS

How do the electrical and thermal properties of a material relate to its optical properties?

On the everyday scale our means of interacting with a material generally rely on its macroscopic properties, and these, as a rule, are based on the assumption that the material is a classical continuum. For example, it is well known that, generally, materials with highly reflecting surfaces are both good electrical and thermal conductors, and conversely materials with dull surfaces are not. Furthermore, highly reflecting materials, which are usually metals, generally have similar mechanical properties such as high ductility. These observations are so familiar from our everyday experience that they do not even give us pause for thought, but we should ask why there is a correlation between these apparently unrelated properties of materials. However, the macroscopic continuum model of materials gives no indication what the underlying common mechanism might be.

The unexplained relationships between the macroscopic properties of materials form the starting point for our investigation of the electronic properties of materials. It seems that there must be some common underlying mechanism that is responsible for all three properties, optical, electrical and thermal, and that this causes the close relationship between them. In fact the correlation between the behaviour of the various properties of the materials can not be explained without some understanding of the structure of the materials and this involves the development of microscopic theories of the atomic and electronic structure inside the materials. The relationship between the structure of matter and its physical properties has been treated in detail using only classical physics to describe the materials in the excellent work of Landau *et al.* [1].

We will begin with some simple definitions of macroscopic properties and then consider some of the well-known macroscopic laws obeyed by materials. Our goal will then be to provide a conceptual framework for understanding these properties and relationships.

1.1.1 Measurable properties of materials

How do we characterize materials in terms of measurable quantities?

In order to measure the properties of a material we do not need to know anything about its internal structure. The properties of interest depend, of course, on the application under consideration, but broadly we are usually interested in one or more of the following categories: mechanical, electrical, optical, thermal and magnetic properties. In most cases the materials properties are obtained as a result of measurements of two quantities, which by themselves do not represent materials properties.

Often, a measurement is made of the response of a material, in terms of a state parameter (e.g. strain, change in temperature or current density), to the influence of an external effect or field parameter (e.g. stress, amount of heat input or electric field strength). The quotient of these two measurements is then the material property (e.g. elastic modulus, specific heat capacity or electrical conductivity). Compilations of the various macroscopic properties of materials have been made by many authors, of which the most comprehensive is that by Lide [2].

1.1.2 Bulk properties of materials

How can these macroscopic properties be explained?

These bulk, continuum properties of materials are almost exclusively what we are really concerned with in using the materials, because these are the properties which can be directly measured. However, explanation of the behaviour of these properties as a function of external conditions, such as temperature, field or frequency of incident electromagnetic radiation, for example, requires a deeper insight into the underlying physical mechanisms.

Although these properties are often documented in great detail for materials, the macroscopic continuum picture gives no explanation of why, for example, copper is a better conductor than glass; why iron is ferromagnetic but manganese is not; why aluminium conducts heat better than sulphur and so on. In order to explain these properties of materials we must look inside the material and try to develop a better understanding of what is happening. These explanations are founded on a description of microscopic rather than macroscopic effects.

1.1.3 Dependence of properties on the environment

Are the material 'constants' really invariant when the external conditions change?

The macroscopic properties of materials, such as Young's modulus, thermal conductivity and electrical conductivity and magnetic permeability, do not remain constant, however. The optical parameters k and n are dependent on the wavelength of incident electromagnetic radiation, permeability is dependent on temperature, and so is electrical conductivity.

The elastic modulus of gadolinium for example shows unusual behaviour close to 293 K. The variation of the reflectivity of silver with energy of incident electromagnetic radiation reveals a drastic change at about 4 eV. The specific heat of nickel reveals anomalous behaviour at around 600 K, and the magnetic susceptibility of manganese fluoride MnF_2, shows an anomaly in the vicinity of 70 K. All of these show variations in bulk properties that lie beyond explanation on the basis of the continuum theory of matter.

These examples show interesting features in some of the bulk properties of these materials. In order to explain these observations it is necessary to consider the properties of the elementary constituents of these materials, that is the atoms and particularly the electrons. Before doing this, however, we will look briefly at a few definitions. These are used to quantify the material properties in which we will be interested and which we will need to refer to throughout this book.

1.2 MECHANICAL PROPERTIES

How do we quantify the mechanical behaviour of materials?

The mechanical properties broadly encompass the elastic, plastic and acoustic properties of a material. These may be quantified by the following: the bulk modulus E_B, Young's modulus E_Y and the shear modulus E_S. (We use these symbols to avoid possible confusion between the elastic moduli, particularly Young's modulus which is often given the symbol E, and the energy which we use extensively later, and which also takes the symbol E.)

1.2.1 Elastic moduli

How does a material respond to stress?

In a material that is isotropic the elastic properties can be completely specified in terms of two elastic moduli, the longitudinal (or Young's) modulus and the transverse (or shear) modulus. Other elastic properties, such as Poisson's ratio, can be completely defined in terms of a combination of these two moduli.

Young's modulus, is a material property obtained from measurement of two quantities; the applied longitudinal stress σ and the resulting strain e in the same direction. Since by Hooke's law stress is proportional to strain for small displacements,

$$E_Y = \frac{\text{stress}}{\text{strain}} = \frac{\sigma_\parallel}{e_\parallel}. \tag{1.1}$$

The shear modulus is a material property obtained from measurement of two quantities, the applied shear stress and the resulting shear strain,

$$E_S = \frac{\text{shear stress}}{\text{shear strain}} = \frac{\sigma_\perp}{\mathrm{e}_\perp}. \tag{1.2}$$

The following table gives values of the elastic Young's modulus for various materials.

Table 1.1 Elastic moduli of various materials.

Material	Elastic modulus E_Y (Pa)	Shear modulus E_S (Pa)
Lead	0.16×10^{11}	0.06×10^{11}
Glass	0.55×10^{11}	0.23×10^{11}
Iron	0.91×10^{11}	0.70×10^{11}
Aluminium	0.70×10^{11}	0.24×10^{11}
Nickel	2.1×10^{11}	0.77×10^{11}
Steel	2.0×10^{11}	0.84×10^{11}
Tungsten	3.6×10^{11}	1.5×10^{11}

These elastic 'constants' are not actually constant but can vary quite markedly with temperature, for instance. An example of interesting temperature dependence of the elastic modulus in gadolinium is shown in Fig. 1.1. Although the elastic moduli are usually determined principally by the lattice potential between atoms we will see later that this anomalous behaviour of the elastic modulus of gadolinium at 293 K and 225 K is the result of reorganization of the electronic magnetic moments during magnetic phase transitions. The electronic interactions perturb the interatomic potential, and at magnetic phase transitions this can cause unusual behaviour of the elastic moduli.

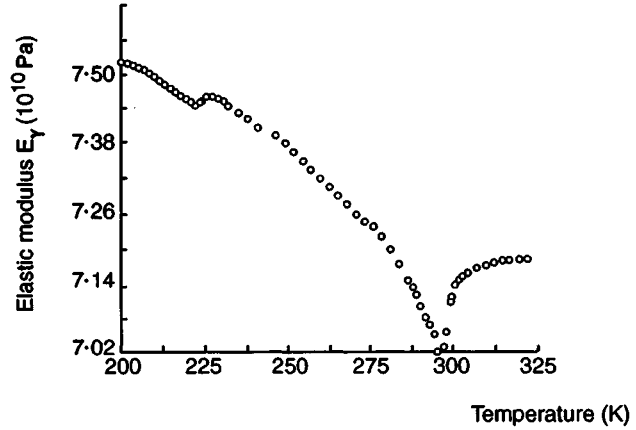


Figure 1.1 Variation of the elastic modulus of the metal gadolinium with temperature [3].

1.3 ELECTRICAL PROPERTIES

How do we quantify the electrical behaviour of materials?

In the case of the electrical properties, we are often concerned with the conductivity. This determines, for example, whether we are dealing with an electrical conductor or insulator. In some cases we may be concerned with the electrical polarization as determined by the dielectric constant or permittivity, and in others with the dissipation of electrical energy under ac conditions (eddy currents). The electrical properties of principal interest are: the electrical conductivity σ and the dielectric constant ε.

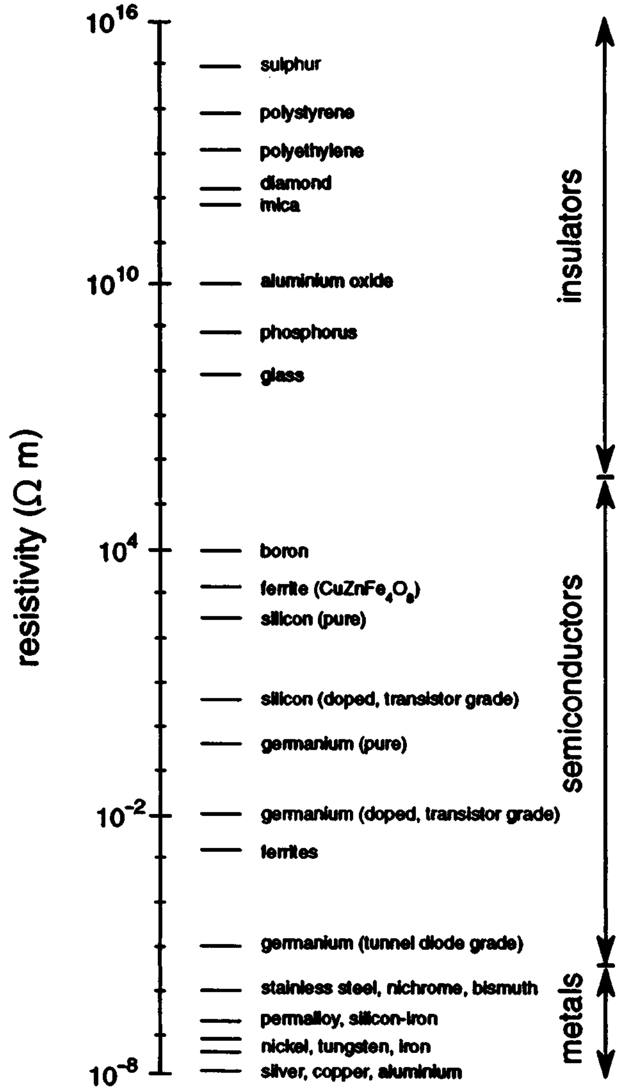


Figure 1.2 Range of resistivities for various materials, including metals, semiconductors and insulators.

1.3.1 Ohm's law

Is there a relationship between electric current and electric field strength in a material?

The well-known Ohm's law deserves a mention because it provides a test of how well the electronic models of materials perform. It states that the current density J is related to the electric field ξ in a material by the relationship

$$J = \sigma\xi. \tag{1.3}$$

We use ξ here to distinguish the electric field from energy which is denoted later by E and σ is the electrical conductivity. Alternatively, if the voltage across a material is V V and the current passing is I A then,

$$V = IR, \tag{1.4}$$

where R is the resistance of the material. Figure 1.2 shows a range of resistivities for various materials.

1.3.2 Electrical conductivity

How is electric charge transmitted in a material?

The electrical conductivity is the amount of electric charge transferred per unit time dq/dt across unit cross-sectional area A under the action of unit potential gradient dV/dx

$$\sigma = \frac{(dq/dt)}{A(dV/dx)}, \tag{1.5}$$

Table 1.2 Numerical values of electrical resistivities and conductivities of various materials at room temperature.

Material	ρ (Ω m)	σ (Ω^{-1} m^{-1})
Sulphur	2×10^{-15}	0.5×10^{-15}
Diamond	1×10^{12}	1×10^{-12}
Glass	0.1×10^{9}	10×10^{-9}
Boron	17×10^{3}	5.9×10^{-6}
Silicon	0.35×10^{3}	2.9×10^{-3}
Germanium	10×10^{-6}	0.1×10^{6}
Gadolinium	1.4×10^{-6}	0.7×10^{6}
Mercury	1.0×10^{-6}	1.0×10^{6}
Stainless steel	0.75×10^{-6}	1.3×10^{6}
Iron	100×10^{-9}	10×10^{6}
Platinum	100×10^{-9}	10×10^{6}
Nickel	83×10^{-9}	12×10^{6}
Zinc	59×10^{-9}	16.9×10^{6}
Tungsten	55×10^{-9}	18×10^{6}
Aluminium	28.5×10^{-9}	35×10^{6}
Gold	24.4×10^{-9}	41×10^{6}
Copper	17×10^{-9}	59×10^{6}
Silver	17×10^{-9}	63×10^{6}

and from Ohm's law

$$\sigma = \frac{J}{\xi} = \frac{\text{current density}}{\text{electric field}} \tag{1.6}$$

The electrical conductivities of materials exhibit probably the widest range of variations of all material properties: 23 orders of magnitude between the conductivities of copper and sulphur, as shown in Table 1.2. Macroscopic continuum theory gives no reason for this variation.

1.3.3 Dielectric properties

How does a nonconducting material respond to the presence of an external electric field?

The dielectric constant or permittivity ε is a material property which relates the amount of electric polarization (charge displacement) P of a material under the action of an electric field ξ

$$\varepsilon = \varepsilon_0 + \frac{P}{\xi} = \varepsilon_0\left(1 + \frac{P}{\varepsilon_0 \xi}\right) = \varepsilon\varepsilon_r. \tag{1.7}$$

The term $P/\varepsilon_0\xi$ is known as the electric susceptibility χ_e. Materials with high permittivity, and hence high electric susceptibility, give a large electric polarization for a given field strength. Values of the relative permittivity ε_r can be as high as 7000 in barium titanate, but in most cases are much lower, for example the relative permittivity of water is 80. The relative permittivity of a material can also be determined from the capacitance C of a condensor with the material as dielectric, compared with that of the same condensor C_0 with a vacuum in place of the material:

$$\varepsilon_r = \frac{C}{C_0}. \tag{1.8}$$

Table 1.3 Dielectric properties of various materials.

Material	ε_r
Copper	∞
Barium titanate	7000
Distilled water	80
Inorganic glasses	6–20
Alumina	9
Polyester	4
Polystyrene	2
Air	1
Vacuum	1

The dielectric strength is a material property which represents the resistance of a material to electrical breakdown (i.e. spontaneous electrical conduction) under the action of a strong electric field. It is sometimes called the breakdown potential, and is measured in volts per metre. Below this field strength the material is an insulator, and above it is a conductor. Unfortunately this property varies widely even among materials which are nominally identical. Therefore, it is not reliable to quote values for particular materials, although it is typically in the range of $10^6\,\mathrm{V\,m^{-1}}$ for dielectrics.

1.4 OPTICAL PROPERTIES

How do we quantify the optical behaviour of materials?

The optical properties of a material tell us how the material interacts with incident electromagnetic waves. These properties can be expressed in terms of two optical constants. Often, the refractive index n and the extinction coefficient k are used, both of which change with the wavelength of the incident light. Alternatively, we can define the optical properties using the reflectance R together with one of the above. We can also use instead the real and imaginary components of the dielectric constant ε. These five quantities are the principal optical properties of interest, and all five change with the frequency of the incident electromagnetic waves.

1.4.1 Refractive index and Snell's law

How does the speed of light in a material determine its change of direction at an interface?

The refractive index of a material is the ratio of wavelength, or phase velocity, of light in a vacuum to that in the material. It is a material property which can be obtained, in principle, solely from the measurement of the speed of light in a material, although this is never attempted in practice:

$$n = \frac{\text{speed of light in vacuum}}{\text{speed of light in material}}. \tag{1.9}$$

The refractive index of a transparent material is usually determined on the basis of the measurements of two angles. θ_i is the angle of incidence of a light beam at the surface of the material and θ_r is the angle of refraction of the light beam inside the material:

$$n = \frac{\sin\theta_\mathrm{i}}{\sin\theta_\mathrm{r}}. \tag{1.10}$$

In fact, the refractive index is frequency dependent, which is why a prism can be used to split white light into different colours (dispersion on the basis of frequency).

1.4.2 Extinction coefficient *k* and the Lambert–Beer law

How is light energy absorbed by a material?

The optical extinction coefficient k is defined as the fractional rate of decrease of light intensity $\mathrm{d}I/I$ in a material per unit path length multiplied by $\lambda/4\pi$ where λ is the wavelength

$$k = \frac{-\lambda}{4\pi I}\left(\frac{\mathrm{d}I}{\mathrm{d}x}\right). \tag{1.11}$$

This is a dimensionless material property, but it is also dependent on the frequency of light. We can also define an attenuation coefficient α which represents the rate at which the intensity of light decays with depth in a material,

$$\alpha = -\frac{1}{I}\frac{dI}{dx} = -\frac{1}{x}\log_e\left(\frac{I}{I_0}\right), \tag{1.12}$$

and the attenuation coefficient in units of m^{-1} is then given by $\alpha = \frac{4\pi k}{\lambda}$. We arrive therefore at the equation

$$I = I_0 \exp(-\alpha x), \tag{1.13}$$

which is the law attributed variously to either Lambert or Beer.

1.4.3 Reflectance

How do we quantify the amount of light reflected at an interface?

The optical reflectance R is the fraction of incident light that is reflected from a surface. The value of R is dependent on both the frequency of the light and the angle of incidence

$$R = \frac{\text{reflected intensity}}{\text{incident intensity}}.$$

It is usually measured using normal incidence of light.

The optical constants shown in Table 1.4 are valid at an energy of 1 eV or equivalently at a wavelength of 1240×10^{-9} m.

Table 1.4 Optical properties of various materials at an energy of 1 eV ($\lambda = 1240 \times 10^{-9}$ m).

Material	***n***	***k***	***R***
Aluminium	1.212	12.464	0.9697
Cobalt	4.46	5.86	0.722
Copper	0.44	8.48	0.976
Gold	0.13	8.03	0.992
Iron	3.43	4.79	0.678
Nickel	3.06	5.74	0.753

The optical reflectances of metals and semiconductors have very characteristic features. Metals have high reflectance at long wavelengths but at shorter wavelengths the reflectance declines. On the other hand semiconductors have low reflectance at long wavelengths but beyond a threshold wavelength known as the band edge or absorption edge, the reflectance increases rapidly as the wavelength decreases. This frequency or energy dependence of optical properties is demonstrated in Figs. 1.3. and 1.4. The continuum model gives no reasons for this characteristic behaviour of the materials.

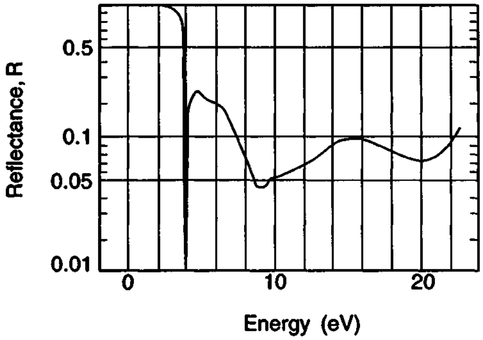


Figure 1.3 Reflectance spectrum of polycrystalline silver [4].

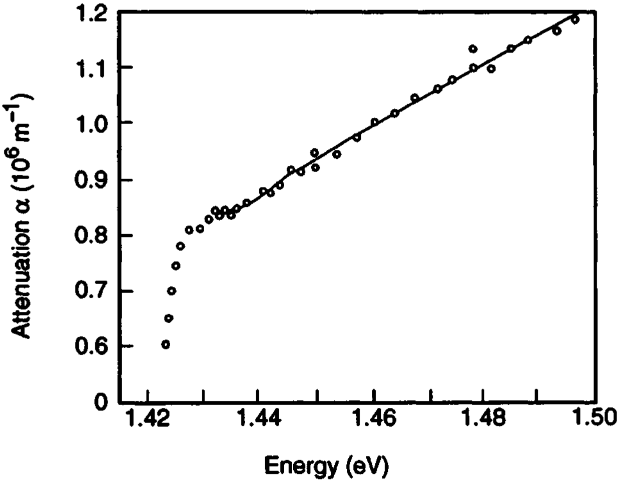


Figure 1.4 Optical absorption spectrum $\alpha(E)$ of gallium arsenide [5]. Reproduced with permission from M. R. Sturge, *Phys. Rev.* 1962.

1.4.4 The Hagen–Rubens law

Is there a relationship between the electrical and the optical properties of a metal? The optical reflectivity and the electrical conductivity of metals at 'low' frequencies ($\nu < 1 \times 10^{14}\,\mathrm{s}^{-1}$) or long wavelengths ($\lambda > 3\,\mu\mathrm{m}$) are also related by an equation of the form,

$$R = 1 - 2\sqrt{\frac{4\pi\varepsilon_0 \nu}{\sigma_0}}, \tag{1.14}$$

where σ_0 is the dc electrical conductivity and R is the reflectance. This is known as the Hagen–Rubens relation. Therefore the mechanisms underlying conductivity and reflectivity seem to be related. The prediction of the reflectance on the basis of the Hagen–Rubens law is shown in Fig.1.5.

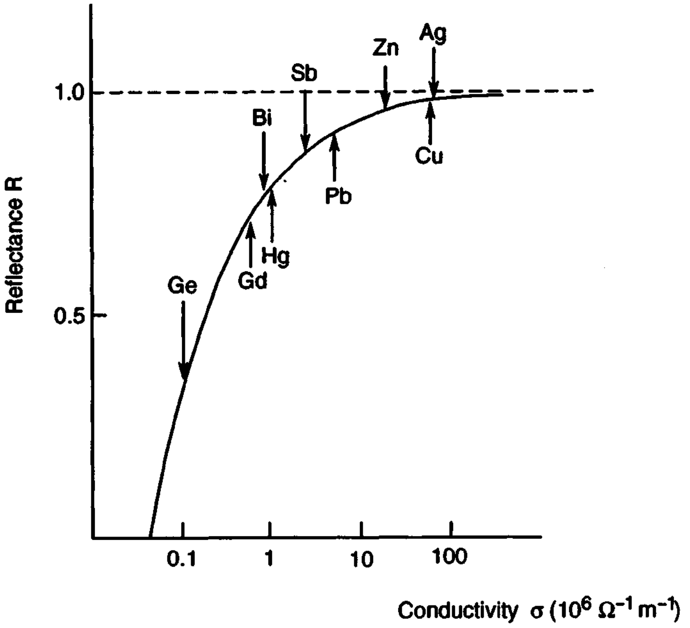


Figure 1.5 Variation of the optical reflectance with electrical conductivity at a wavelength of 3 μm ($\nu = 1 \times 10^{14}$ s^{-1}).

1.5 THERMAL PROPERTIES

How do we quantify the thermal properties of materials?

In the case of thermal properties we are often concerned with the rate of flow of heat through the material as measured by the thermal conductivity K. This determines whether the material is a thermal conductor or insulator. Another quantity of interest is the amount of heat which must be supplied to raise the temperature of unit mass by one degree, that is the specific heat or heat capacity C.

1.5.1 Thermal conductivity

How is the thermal conductivity defined?

The thermal conductivity K of a material is the rate of transfer of heat per unit time, per unit cross sectional area, per unit distance, per unit temperature gradient

$$K = \frac{Qx}{(T_2 - T_1)At} = \frac{J_Q}{(\mathrm{d}T/\mathrm{d}x)}, \tag{1.15}$$

where A is the cross-sectional area through which the heat passes, Q is the heat energy transferred in time t between two locations a distance x apart, where T_2 and T_1 are the temperatures at the two locations. An alternative, but equivalent definition is that K is the quotient of the thermal flux density J_Q with respect to the temperature gradient dT/dx. This equation only applies under steady-state conditions.

Table 1.5 Thermal conductivities of various materials.

Material	**K (W m^{-1} K^{-1})**
Silver	428
Copper	398
Gold	315
Aluminium	237
Nickel	158
Tungsten	182
Zinc	115
Iron	80
Silicon	83
Platinum	73
Glass	0.2

1.5.2 The Wiedemann–Franz law

Is there a relationship between the electrical and thermal properties of a metal? In most cases good electrical conductors are also good thermal conductors. Quantitative investigation by Wiedemann and Franz revealed that for most metals the relationship between electrical conductivity and thermal conductivity K obeyed the following law

$$\frac{K}{\sigma T} = 2.4 \times 10^{-8}\,\mathrm{J}\,\Omega\,\mathrm{K}^{-2}\,\mathrm{s}^{-1}. \tag{1.16}$$

This seems to imply that the underlying mechanisms behind electrical and thermal conductivity are related in some way. The continuum model offers no explanations.

1.5.3 Specific heat capacity

What determines the increase in temperature of a material when it is heated? The specific heat C_m of a material is the amount of heat required to raise unit mass of the substance by one degree of temperature, while the heat capacity C is the amount of heat required to raise the temperature of an unspecified mass by one degree of temperature.

$$C_m = \frac{1}{M}\frac{dQ}{dT}, \tag{1.17}$$

$$C = \frac{dQ}{dT} = \frac{dU}{dT}. \tag{1.18}$$

Here, M is the mass, dT is the change in temperature, U is the internal energy and dQ is the heat energy absorbed. The specific heat is itself dependent on temperature. It is also dependent on whether the measurement is made under constant-volume or constant-pressure conditions.

The heat capacity of some materials varies in a very characteristic way. For example, the temperature dependence of the heat capacity of iron which is shown in Fig. 1.6 has an anomaly at 1040 K. As we shall see later this also corresponds to a magnetic phase transition.

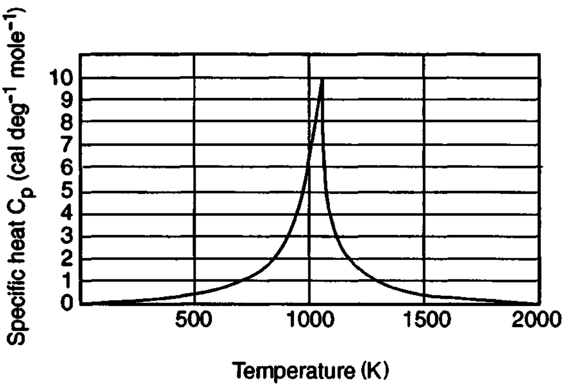


Figure 1.6 Heat capacity of iron showing anomalous behaviour at about 1040 K [6]. Reproduced from *J. Phys. Chem. Solids*, **1**, J. A. Hoffmann *et al.*, p. 52, copyright 1956, with kind permission from Elsevier Science.

1.5.4 The Dulong–Petit law

Is there a relationship between the heat capacities of various materials?

The heat capacities of many materials are found to be linearly dependent on the molecular or atomic weight of the substance, at least at higher temperatures. This can be expressed as the 'molar heat capacity'. This is the heat capacity of a fixed number ($N_0 = 6.02 \times 10^{23}$) of atoms or molecules of a substance. For many materials this has a value close to $25\,\mathrm{J\,mol^{-1}\,K^{-1}}$, a result discovered by Dulong and Petit.

Table 1.6 Heat capacities of various materials.

Material	**C ($\mathbf{J\,mol^{-1}\,K^{-1}}$)**
Aluminium	24.3
Iron	25.7
Nickel	26.8
Copper	24.4
Lead	26.9
Gold	25.5

This seems to imply that the heat capacity is dependent only on the number of elementary entities, either atoms or molecules depending on the material. However, even this law only applies at high temperatures, since the heat capacity varies with temperature as we shall see in the next chapter. The molar heat capacity deviates significantly from the value predicted by the Dulong–Petit law at lower temperatures (e.g. below 100 K in lead, below 400 K in aluminium and copper and even at higher temperatures in carbon).

1.6 MAGNETIC PROPERTIES

How do we quantify the magnetic behaviour of materials?
When dealing with magnetic properties, we are usually concerned either with the permeability μ of the material, which describes its response to an external field, or its magnetization M which is the magnetic moment per unit volume.

1.6.1 Magnetic moment and magnetization

Which properties determine the response of a material to an applied magnetic field?
The magnetic moment m of a material is the maximum torque $\tau_{\max}$ experienced by the material under the action of a magnetic field in free space, divided by the strength of the magnetic field H:

$$m = \frac{\tau_{\max}}{\mu_0 H}, \tag{1.19}$$

where μ_0 is the permeability of free space which is a universal constant. The magnetization M of a magnetic material is the magnetic moment per unit volume:

$$M = \frac{m}{V}. \tag{1.20}$$

This is a function of magnetic field strength H, and so is not a material constant.

1.6.2 Magnetic susceptibility

How does a material respond to an external magnetic field?
The magnetic susceptibility χ is the ratio of magnetization M to magnetic field strength H:

$$\chi = \frac{M}{H}. \tag{1.21}$$

Table 1.7 Magnetic permeabilities and susceptibilities of various materials.

Material	**Relative permeability** μ_r	**Susceptibility** χ
Permalloy	$\sim 10^4$	$\sim 10^4$
Iron	$\sim 10^2 - 10^3$	$\sim 10^2 - 10^3$
Samarium-cobalt	~ 1	-
Aluminium	1.000 02	2×10^{-5}
Manganese	1.000 83	8.3×10^{-4}
Copper	0.999 99	-1×10^{-5}
Bismuth	0.999 83	-1.7×10^{-4}

It is not a material constant for strongly magnetic materials such as iron, cobalt and nickel because in these materials χ is dependent on field strength $\boldsymbol{H}$. However, it can be approximately constant for weakly magnetic materials such as aluminium and sodium. Values of the susceptibility for various materials are shown in Table 1.7.

1.6.3 Magnetic permeability

How is the magnetic induction in a material related to the magnetic field?
The magnetic permeability of a material is the ratio of magnetic induction $\boldsymbol{B}$ to the magnetic field strength $\boldsymbol{H}$:

$$\mu = \frac{\boldsymbol{B}}{\boldsymbol{H}}, \tag{1.22}$$

and since $\boldsymbol{B} = \mu_0(H + M)$, there is an exact relationship between permeability and susceptibility:

$$\mu = \mu_0(\chi + 1) = \mu_0\mu_r, \tag{1.23}$$

where μ_r is the relative permeability and μ_0 is the permeability of free space.

This quantity is strongly dependent on magnetic field strength applied to ferromagnetic materials and so is not a materials constant. In weakly magnetic materials such as paramagnets or diamagnets it is close to the value μ_0 since χ is approximately zero.

1.6.4 The Curie–Weiss law

How does the susceptibility of weakly magnetic materials vary with temperature?
Some magnetic materials undergo a magnetic phase transition at a temperature T_c from a high temperature state (with low susceptibility $\chi \approx 10^{-3}$) to a low temperature state (with a significantly higher susceptibility). In many cases the variation of the susceptibility of these materials with temperature T in the region $T > T_c$ can be described by an equation known as the Curie–Weiss law:

$$\chi(T) = \frac{C}{T - T_c}. \tag{1.24}$$

In other cases the value of T_c is zero, which means that there is no transition temperature, and in this case we have the more restricted Curie law:

$$\chi(T) = \frac{C}{T}. \tag{1.25}$$

In both cases the susceptibility varies inversely with temperature. An example of the variations of χ^{-1} with temperature for some manganese compounds, which obey the Curie or Curie–Weiss laws, are shown in Fig. 1.7.

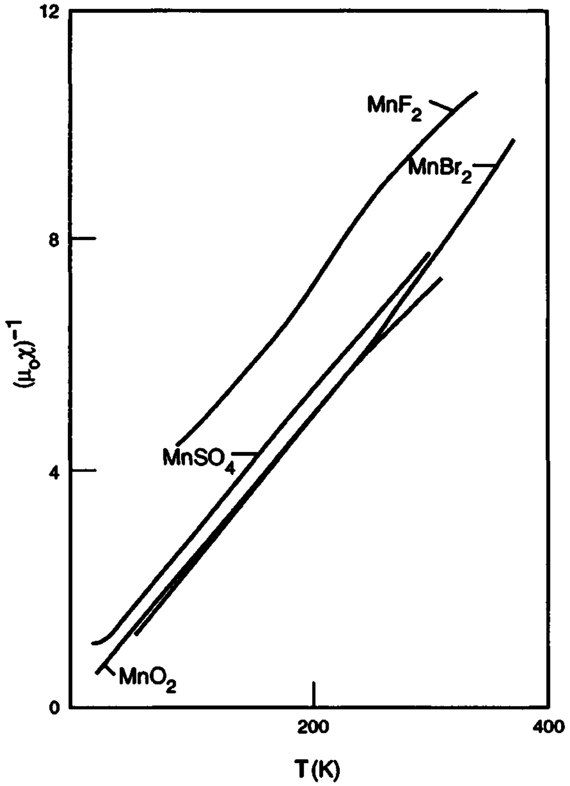


Figure 1.7 Magnetic susceptibility of some manganese compounds as a function of temperature showing agreement with the Curie law and the Curie–Weiss law.

1.6.5 Magnetoresistance

How is the electrical resistance of a material affected by application of a magnetic field?

The electrical resistance of some magnetic materials can change quite dramatically when a magnetic field is applied. Normally this effect is measured in terms of the fractional change in resistance $\Delta R/R$ versus magnetic field. In the case of materials such as permalloy (an alloy of nickel and iron) $\Delta R/R$ can be about 2% at magnetic saturation. However, the effect can be much larger in other materials, such as bismuth in which $\Delta R/R$ increases by 150% in a magnetic field of $9.5 \times 10^5\,\mathrm{A\,m^{-1}}$ (1.2 T). More recently, materials in which the resistance is rather more sensitive to magnetic field have been found such as InSb-NiSb, which exhibits a 300% change in resistance in a field of $2.3 \times 10^5\,\mathrm{A\,m^{-1}}$. However, in this case the magnetoresistance of the material is also strongly affected by temperature, which therefore limits its applications.

Magnetoresistance can be very large in some thin film magnetoresistive materials. The first layered magnetic material to exhibit this 'giant magnetoresistance'

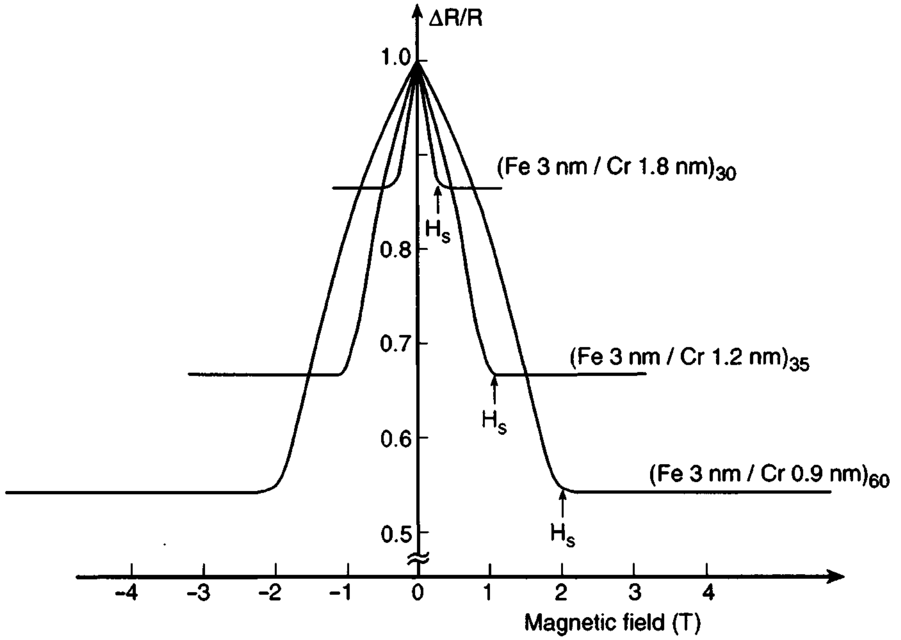


Figure 1.8 Giant magnetoresistance in Fe/Cr/Fe showing 45% decrease in resistance in a field of 2T (20kG) [7].

was Fe/Cr/Fe [7] and this was soon followed by other materials including Fe/Au/Fe [8], Co/Au/Co [9] and a range of other layered structures including Co/Cu/Co [10]. The magnitude of the effect depends on the thickness of the layers as shown in Fig. 1.8. The physical effect that causes this change in resistance with field is the spin-dependent scattering of electrons. This means that electrons with spins in one direction are scattered more than electrons with spins in other directions. In order to explain why this happens it will be necessary to gain an understanding of the electron band structure of the materials. The phenomenon is exploited in magnetic recording where the magnetic read heads are made from giant magnetoresistive materials.

1.7 RELATIONSHIPS BETWEEN VARIOUS BULK PROPERTIES

Are there any relationships between the bulk properties of materials which are not obvious on the basis of the continuum model?

In addition to the variation of the bulk properties of materials with temperature and applied fields, empirical relationships have been noticed between various macroscopic properties. We have considered five well-known empirical laws determined from macroscopic properties which require an explanation based on a better understanding of the structure of materials. We have observed, for example, that a material with high optical reflectance is usually a good electrical conductor and is also usually a good thermal conductor. This, and other property relationships, are largely unexplained by the continuum model. These relationships can not be simply fortuitous. They therefore provide the starting point for the development of a theory of the material properties because they challenge us to provide an explanation.

We need to understand the underlying physical reasons why such relationships occur. It is common knowledge that good optical reflectors are good electrical conductors. Despite the everyday familiarity of this relationship, it is at first a strange result because it is difficult to see why there should be any relation between these two properties at all. We need to consider the problem at a deeper level because our continuum models provide no clue to the explanation.

1.8 CONCLUSIONS

Does the continuum model help us to understand what takes place inside materials? The main objective of this chapter is to indicate the types of macroscopic measurements that are usually made for property determination of materials. These measurements are almost always macroscopic measurements which do not depend on any assumptions about the underlying mechanisms, nor do they depend on a more fundamental interpretation. They are merely defined in terms of an individual measurement or sometimes in terms of a combination of measurements. For example, current density J and electric field ξ are both measurable quantities which are not material properties, but their ratio $\sigma = J/\xi$ is a material property known as the electrical conductivity. However, the inability to explain the behaviour of these properties and their interrelationships through the use of the continuum model forces us to develop more sophisticated descriptions of materials based on an atomistic approach, and ultimately on an electronic approach.

REFERENCES

1. L. D. Landau, A. I. Akhiezer and E. M. Lifshitz (1967) *General Physics: Mechanics and Molecular Physics*, Pergamon Press, Oxford.
2. D. R. Lide (1985) *Handbook of Chemistry and Physics*, 82nd edn, CRC Press, Boca Raton, Florida.
3. D. C. Jiles and S. B. Palmer (1980) *J. Phys. F.* 10, 2857.
4. H. Ehrenreich *et al.* (1965) *IEEE Spectrum* 2, 162.
5. M. D. Sturge (1962) *Phys. Rev.* 127, 768.
6. J. A. Hofmann, A. Paskin, K. J. Tauer, and R. J. Weiss (1956) *J. Phys. Chem. Solids* 1, 45.
7. M. N. Baibich, J. M. Broto, A. Fert, F. Nguyen, F. Petroff, P. Etienne, G. Creuzet, A. Frederick and J. Chazelas (1988) *Phys. Rev. Lett.* 61, 2472.
8. F. Saurenbach, J. Barnas, G. Binasch, M. Vohl, P. Gruenberg and W. Zinn (1989) *Thin Solid Films* 175, 317.

9. T. Takahata, S. Araki and T. Shinjo (1989) *J. Mag. Mag. Mater.* 82, 287.
10. S. S. P. Parkin, R. Bhadra and K. P. Roche (1991) *Phys. Rev. Lett.* 66, 2152.

FURTHER READING

The following is an excellent series on the physical properties of materials:

A. H. Cottrell (1967) The nature of metals. *Scientific American* 217, 90.
J. M. Ziman (1967) The thermal properties of materials. *Scientific American* 217, 180.
H. Ehrenreich (1967) The electrical properties of materials. *Scientific American* 217, 194.
F. Keffer (1967) The magnetic properties of materials. *Scientific American* 217, 222.
A. Javan (1967) The optical properties of materials. *Scientific American* 217, 238.
J. Fraden (1993) *Handbook of Modern Sensors: Physics, Designs and Applications*, American Institute of Physics, New York.

EXERCISES

Explain each of the following well-known laws. You should describe the relations involved and try to explain the underlying reason for each on a classical basis.

Exercise 1.1 The Wiedemann–Franz law

Exercise 1.2 The Hagen–Rubens relation

Exercise 1.3 The Dulong–Petit law

Exercise 1.4 Macroscopic properties
Explain what is meant by

(i) Permeability
(ii) Conductivity
(iii) Elastic modulus
(iv) Extinction coefficient
(v) Heat capacity

What is the physical significance of each of these properties and how are they measured?

Exercise 1.5 Empirical laws
Certain observed relations, for which there seems to be no obvious reason, were in the past called ‘laws’. There are a number of empirical ‘laws’ which relate different properties of materials: the Wiedemann–Franz law, the Hagen–Rubens law, and the Dulong–Petit law. Other laws describe the behaviour of certain material properties as a function of external conditions, such as the Curie and Curie–Weiss laws. Describe each of these laws and try to provide a simple continuum explanation for each.

Exercise 1.6 Electronic and lattice contributions to properties
Which of the following properties are determined primarily by the electrons and which by the atomic lattice (or maybe both?) in a material?

(a) Heat capacity
(b) Thermal conductivity
(c) Permeability

How is the elastic modulus of a material related to the strength of the interaction between the atoms in a material?

2 PROPERTIES OF ATOMS IN MATERIALS

OBJECTIVE

This chapter provides a background for the rest of the book since clearly the electrons within materials are contained within a volume defined by the atomic cores of the material, and it is important to realise that the properties of those electrons are determined largely by their interactions with these atomic cores. In fact, we should go even further and state that the electronic properties of interest in this book are determined exclusively by the energy 'landscape' provided by the ionic cores. Therefore, before going further we should look at some of the properties of this ionic background. In the case of crystalline materials this ionic background forms an ionic 'lattice', but in other materials, such as amorphous solids, it forms a random aggregate.

2.1 THE ROLE OF ATOMS WITHIN A MATERIAL

What is the next level of sophistication beyond the continuum model?
Moving on from the continuum model our next question must be to ask what lies beneath the surface of the material. Before proceeding with a discussion of the electronic properties of materials, we shall pause briefly to consider the arrangement of the atomic cores within a solid. The simplest case occurs when the atoms are arranged in a regular crystal lattice with a well-defined spatial periodicity or symmetry.

Many authors like to consider the crystal structure to be the combination of a crystal lattice plus a 'basis' (which is an identical configuration of atoms attached to each lattice point) [1, p. 4]. We shall go further and consider our solids to be composed of a periodic array of atomic cores immersed in a sea of electrons:

$$\text{solid} = \text{ionic lattice} + \text{electronic 'sea'}.$$

The lattice here represents the arrangement of atoms in a periodic structure and the electron sea is a random arrangement of high-energy electrons without obvious periodicity, although we shall see later that the lattice imposes periodicity on the electrons. This separation may seem almost trivial at first sight, but it does have important implications for our subsequent description of materials. Some of the properties of materials can be attributed principally to the lattice and some principally to the electrons. It will be our purpose to separate and identify these, where possible, and of course we will later pay more attention to those properties which depend on the electrons.

Since all electrons are identical we will eventually see that the differences in the electronic properties of materials are due more to the ions and their lattice symmetry than to the electrons themselves, which is an interesting conclusion.

2.1.1 Types of lattice symmetry

How do the atoms arrange themselves inside a material?

The classification of the various types of lattice is merely a mathematical abstraction of the forms of symmetry exhibited by materials [2,3]. Figure 2.1 shows the various forms of lattice or crystal structures. These have been verified by X-ray diffraction results [4].

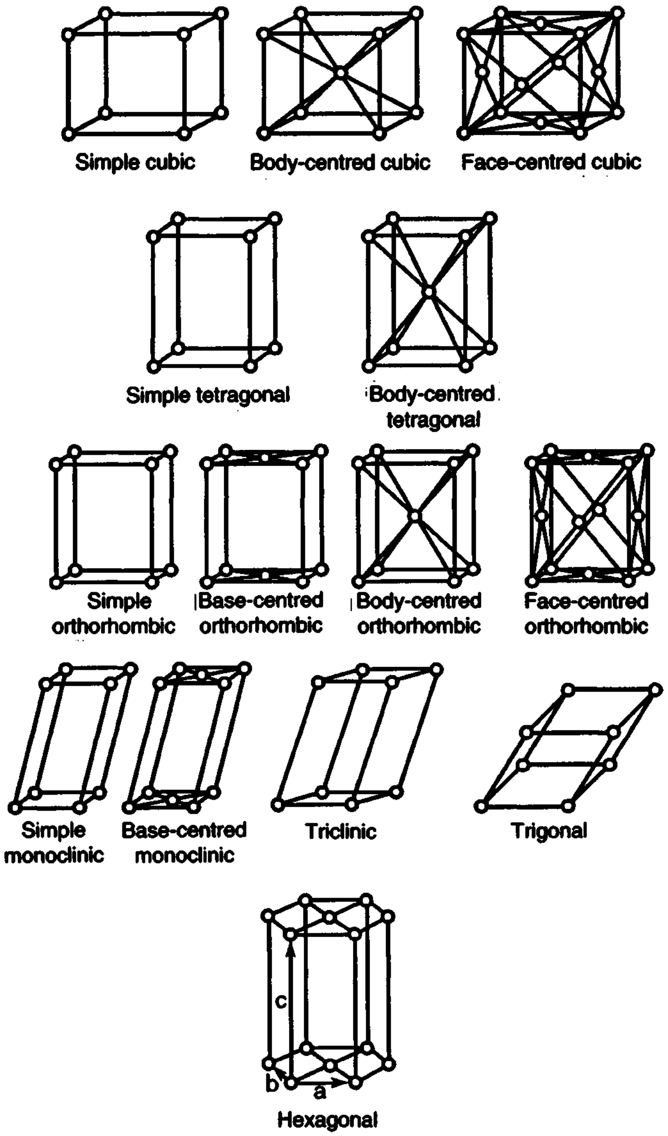


Figure 2.1 Different forms of lattice symmetry exhibited by crystalline materials.

The principal classes of lattice, with the number of elements exhibiting each form of symmetry in the solid state, are as follows: 35 cubic (19 face-centred cubic, 14 body-centred cubic, 2 simple cubic); 29 hexagonal (mostly hexagonal close packed); 5 trigonal; 2 tetragonal; 21 orthorhombic, monoclinic or triclinic.

2.1.2 Cohesive energy of the lattice

What forces hold the lattice together?

When considering the structure of a solid two questions arise: (i) what holds the lattice together? and (ii) what determines the symmetry of the lattice? The cohesive energy of a crystal is the amount of energy which must be supplied in order to separate it into free neutral atoms. The melting temperatures of crystals are approximately proportional to the cohesive energy. Typical values of the cohesive energy of solids are shown in Table 2.1.

Table 2.1 Values of cohesive energies for various materials.

	Cohesive energy **E_c (kJ mol^{-1})**	**Melting point** **T_c (K)**	**T_c (°C)**
Ruthenium	650	2723	2450
Silicon	446	1687	1410
Iron	413	1811	1575
Germanium	372	1211	985
Cadmium	112	594	321
Mercury	65	234	−39
Xenon	16	161	−112

The cohesive energy of a solid is determined exclusively by electrostatic interactions between the electrons, which have negative charge, and the ionic cores, which have positive charge. The ionic cores are located on the lattice sites, whereas the electrons may be located on the lattice sites or may be free to move throughout the solid.

The exact form of the electrostatic interactions holding the solid together may be different in different cases. We can identify three main types: (i) ionic interactions, for example NaCl; (ii) covalent bonds, for example carbon; and (iii) Van der Waals interactions in inert crystals, for example xenon. There is also the question of the interaction in metals in which all ionic cores have a net positive charge and are immersed in a sea of negatively charged electrons. In these cases the negatively charged sea of electrons is composed of some of the outer electrons of the atoms, while the inner electrons remain localized at the ionic sites. The subject of the cohesive forces in materials has been treated in great detail by Maitland *et al.* [5].

In cases where there is no separation of electrical charges within a material the electrostatic interactions cannot be very large. In these cases the atoms are held together by the 'fluctuating dipole' forces also known as Van der Waals forces.

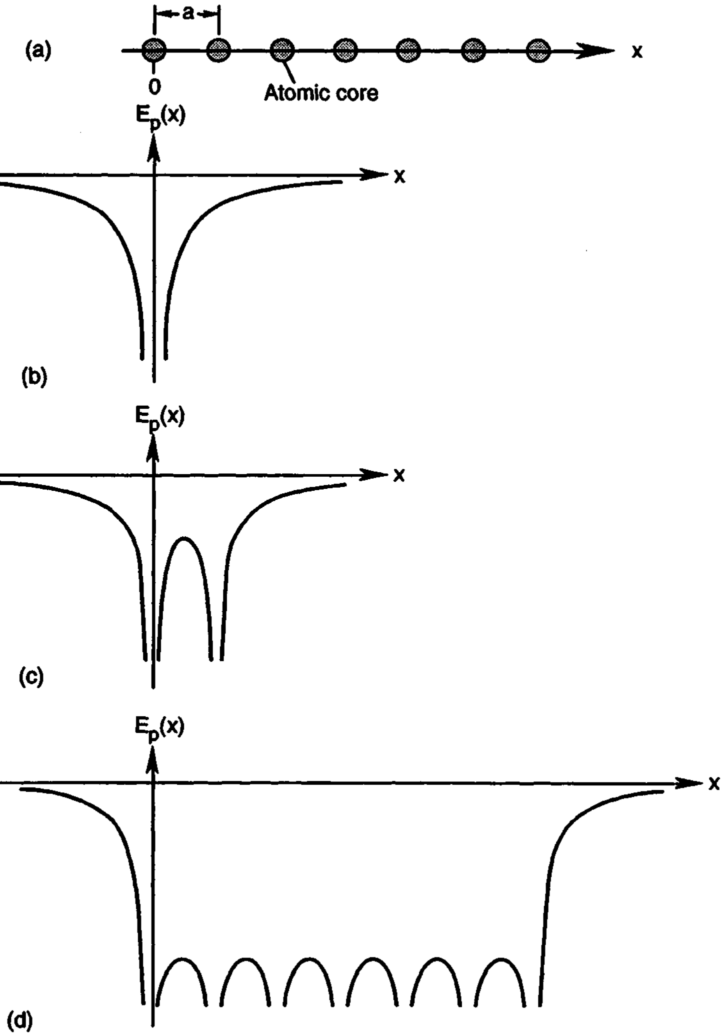


Figure 2.2 Variation of the interatomic potential $E_p(x)$ of the ions with position in a regular crystal lattice. (a) Linear lattice of ions, (b) potential well of isolated ion, (c) potential wells of two ions in proximity, (d) periodic potential of a linear array.

These forces are dominant in the inert solids such as Ne, Ar, Kr, Xe, and Rn. They form the weakest type of bond and only play the most important role if none of the other cohesive mechanisms occurs.

In these solids, although the time average of the charge distribution is always zero for a given atom, at any given instant there will be a net electric dipole moment. The dipole moments on neighbouring lattice sites lead to a weak attractive interaction between the dipoles. This force then holds the atoms together.

2.2 THE HARMONIC POTENTIAL MODEL

How can the atomic forces be modelled in the simplest way?

We now need to consider the form of bonding between the atoms in order to provide an adequate model for the behaviour of the lattice. The interactions between the individual ions can be represented by a single force. Over a large range of deformations the force between the neighbouring lattice sites is proportional to their displacement. Ultimately for large deformations this proportionality no longer holds of course, but we should first explore the consequences of this simplest of models.

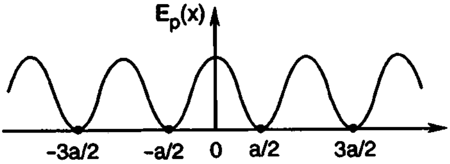


Figure 2.3 Ionic cores in a one-dimensional crystal lattice subject to a harmonic potential.

In this case the harmonic potential leads to an internal restoring force F on each atom which is proportional to the displacement u of the individual atom from its unstrained equilibrium position with respect to its two nearest-neighbours,

$$F = -2ku, \tag{2.1}$$

where k is a force constant, sometimes known as the stiffness coefficient. Since the force $F = -dE_p/du$ where E_p is the potential energy, this means that each lattice site reaches its equilibrium position at the bottom of a parabolic, or harmonic, potential of the form

$$E = -\int(-2ku)du = ku^2. \tag{2.2}$$

Such a potential can arise also from a mechanical model in which the atomic cores are connected by linear springs, as shown in Fig. 2.4.

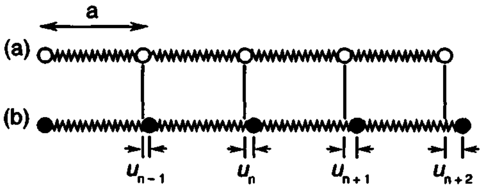


Figure 2.4 Linear lattice of atoms in which the interaction force is represented by the spring model.

In this case the force on any given atom in the lattice is determined solely by its displacement u relative to its two nearest-neighbours.

$$F = \underset{\substack{\text{Force due}\\ \text{to atom}\\ \text{on left}}}{-ku} + \underset{\substack{\text{Force due}\\ \text{to atom}\\ \text{on right}}}{k(-u)} = -2ku. \tag{2.3}$$

Considering now, for simplicity, a transverse vibration of the linear lattice with interactions only between nearest-neighbours. For the nth atom in the lattice the force is,

$$F = m\frac{\mathrm{d}^2u_n}{\mathrm{d}t^2} = k(u_{n-1} - 2u_n + u_{n+1}), \tag{2.4}$$

where u_n is the displacement of the nth atom and m is its mass. Using a Taylor series expansion for the displacement,

$$u_{n+1} = u_n + \left(\frac{\mathrm{d}u_n}{\mathrm{d}x}\right)a + \left(\frac{\mathrm{d}^2u_n}{\mathrm{d}x^2}\right)\frac{a^2}{2} + \cdots, \tag{2.5}$$

where a is the lattice spacing. Substituting this back into the force equation gives

$$F = m\frac{\mathrm{d}^2u_n}{\mathrm{d}t^2} = ka^2\frac{\mathrm{d}^2u_n}{\mathrm{d}x^2}, \tag{2.6}$$

which is the equation of a wave with velocity v given by

$$v = \sqrt{\frac{ka^2}{m}}, \tag{2.7}$$

and a frequency of $\omega = \sqrt{(2k/m)}$. This gives a direct relationship between the form of the interatomic potential through the force constant k, the lattice spacing a, the mass of the atoms and the velocity and frequency of vibrations passing through the lattice. We will see in the next section that the elastic modulus is also determined by these quantities.

2.2.1 Elastic modulus

Can a simple expression for the elastic modulus be derived on the basis of the spring model?

We can use the above simple linear lattice model, in which each atomic core lies within its harmonic potential, to calculate the elastic modulus E_Y. Suppose the lattice is subjected to a constant applied force F_{app} along its length. Under equilibrium conditions the lattice spacing will be equal throughout the chain, ensuring no net force on each atom, in accordance with the above equations. Nevertheless, this lattice spacing will be different from the unstrained spacing a. This equilibrium lattice spacing under the applied force will be $a + u_n - u_{n-1}$.

Considering the forces on the last atom in the lattice, since this is the only one acted upon directly by the applied force, it has only one nearest-neighbour. This leads to a restoring force of $-k(u_n - u_{n-1})$ which, under equilibrium conditions, must balance the externally applied force. Also under equilibrium conditions we can assume that $u_n - u_{n-1}$ is the same for all pairs of atoms. Let this be Δu. For the purposes of this derivation we will assume that the cross-sectional area per linear lattice chain is a^2, which is equivalent to assuming a simple cubic lattice. This gives a stress of $\sigma = F/a^2$:

$$\sigma = \frac{k\Delta u}{a^2}. \tag{2.8}$$

Let the macroscopic deformation of the lattice be $\Delta\ell$. The strain e is then

$$e = \frac{\Delta\ell}{\ell_0} = \frac{N\Delta u}{Na} = \frac{\Delta u}{a}, \tag{2.9}$$

where N is the total number of atoms in the linear lattice and ℓ_0 is the undeformed macroscopic length. The elastic modulus E_Y is the ratio of stress to strain, and therefore

$$E_Y = \frac{\sigma}{e} = \frac{k}{a}. \tag{2.10}$$

This means that the elastic modulus of the linear chain lattice is defined by the unstrained interatomic spacing a and the force constant k. The elastic modulus of the lattice can then be related to the wave velocity in the lattice by the equation,

$$v = \sqrt{\frac{E_Y a^3}{m}} = \sqrt{\frac{E_Y}{\rho}}, \tag{2.11}$$

where ρ is the density of the material.

We should note that under large deformations this simple parabolic potential is not sufficient to correctly describe the elastic behaviour, and in this case higher-order terms become significant, particularly third-order contributions to the elastic modulus. This leads to anharmonicity of the lattice potential which is essential in the explanation of such phenomena as thermal expansion. This is described in Sections 2.2.3 and 2.2.4.

2.2.2 Quantization of lattice vibrations

If the material consists of a discrete lattice are all frequencies of vibration allowed? We consider again the simplified problem of the linear lattice. The expression for the force on an individual atom can be transformed into an equation in terms of the positions of the atoms x instead of their displacements from equilibrium u.

Consider a longitudinal vibration of the lattice. Let x_n be the position of the nth atom. Then the dispacement of the nth atom from unstrained equilibrium can be written $u_n = x_n - (x_{n-1} + a)$ or alternatively $u_n = x_n - (x_{n+1} - a)$. The force on the nth atom is then,

$$F_n = -k(x_n - (x_{n-1} + a)) + k(-x_n + (x_{n+1} - a)). \tag{2.12}$$

At any lattice spacing the nth atom will be at equilibrium, according to this model, provided it is located midway between its nearest-neighbours. Therefore the net force on the atom is not dependent on the equilibrium spacing a. Eliminating a leads to,

$$F = k(x_{n-1} - x_n) - k(x_n - x_{n+1}). \tag{2.13}$$

Then the force and acceleration are related by the equation,

$$F = m\frac{\mathrm{d}^2 x_n}{\mathrm{d}t^2} = k(x_{n-1} - 2x_n + x_{n+1}). \tag{2.14}$$

This again leads to the following equation for the velocity of propagation of vibrations,

$$v = \sqrt{\frac{ka^2}{m}} = \sqrt{\frac{E_Y}{\rho}}. \tag{2.15}$$

For the position of the nth atom in the chain this equation has solutions of the form

$$x_n = A\exp(\mathrm{i}(qna - \omega t)). \tag{2.16}$$

This is a valid solution of the equation of motion for any value of A provided the following condition is satisfied for the frequency of vibration ω,

$$\omega^2 = \frac{2k}{m}(1 - \cos qa), \tag{2.17}$$

and when $qa = \pi/2, 3\pi/2, 5\pi/2$ and so on, this will give the 'classical' single atom vibration frequency $\omega = \sqrt{(2k/m)}$. However, in general, the solution for the vibration of the discrete lattice is different from that of the single atom.

From eqn (2.17) we may take the positive root

$$\omega = \sqrt{\frac{4k}{m}}\sin\left(\frac{qa}{2}\right). \tag{2.18}$$

If we have a total of N atoms in the chain and impose periodic boundary conditions (the Born–von Karman condition), then there are just N allowed values of q which independently satisfy all the conditions [6].

The proof of this is fairly simple. If we have a wave passing along the lattice we must have $x_N = x_1$ for periodic boundary conditions. Therefore the allowed values of the product qNa must be an integral multiple of 2π in order to satisfy the periodic boundary condition,

$$qNa = 2\pi, 4\pi, 6\pi, \ldots 2N\pi. \tag{2.19}$$

Since a is the interatomic spacing and N is the total number of atoms, then $Na = L$ the length of the lattice. Therefore,

$$q = 2\pi/L, 4\pi/L, \ldots 2N\pi/L. \tag{2.20}$$

Let us try a larger value of q which can still meet the boundary conditions, for example $q = 2(N+1)\pi/L$, in order to see what happens when we go outside this specified range of q:

$$x_n = A \exp \mathrm{i}\left\{\frac{2(N+1)\pi}{L}na - \omega t\right\} \tag{2.21}$$

$$= A \exp \mathrm{i}\left\{\frac{2\pi}{L}na - \omega t\right\} \exp \mathrm{i}\left\{\frac{2N\pi}{L}na\right\}, \tag{2.22}$$

and since $L/N = a$, this gives the following relation for all n,

$$x_n(2(N+1)\pi) = x_n(2\pi) \exp \mathrm{i}\{2n\pi\}, \tag{2.23}$$

which is not an independent solution. Higher values of q will simply lead to wave motion that is identical to one of the values of q in the range $(2\pi/L, 4\pi/L, \ldots 2N/L)$. In the linear lattice of N atoms there are consequently only N allowed distinct vibrational modes [7]. This is demonstrated graphically in Fig. 2.5.

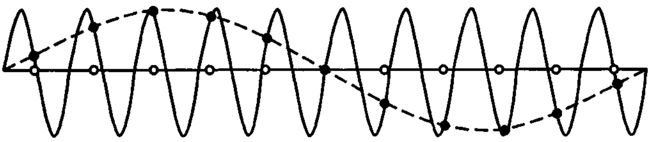


Figure 2.5 Waves in a discrete lattice showing that two different values of wave vector give identical displacements of the lattice. The longer wavelength vibration is the only meaningful interpretation of the wave on the basis of the discrete lattice. This demonstrates that there are only N independent solutions of the wave equation in a lattice with N atoms.

This can easily be extended to three dimensions. Consider an array of N^3 atoms in a cubic lattice: the allowed values of q are

$$q = q_x + q_y + q_z, \tag{2.24}$$

where now the constraint is

$$q_x;\ q_y;\ q_z = 2\pi/L,\ 4\pi/L, \ldots 2N\pi/L. \tag{2.25}$$

Again only certain vibrational modes are possible and there are N^3 distinct modes. We have therefore shown by this simple example that the allowed vibrations of a periodic lattice are restricted and discrete, that is they are quantized. This is an important result in which we can understand the quantization of the allowed lattice vibrations on the basis of a simple discrete classical model of the material.

2.2.3 Anharmonicity

What are the immediate and obvious drawbacks of the simple spring model? Although the harmonic potential, or spring model, works quite well for small displacements of the atoms, it is quite easy to demonstrate that it must fail for large displacements. If we simply consider two atoms, the energy of the system when the atoms are moved closer together will be larger than the energy of the system if they are moved apart. Put another way, the atoms cannot be displaced relative to one another so that they occupy the same location, but they can have the distance between them doubled.

This argument demonstrates that the lattice potential must ultimately be anharmonic [8]. It is this anharmonicity which gives rise to thermal expansion and is also responsible for third-order elastic constants, that is the variation of elastic moduli with strain. The anharmonic potential can most easily be described as a perturbation from our simple harmonic potential

$$E_{\mathrm{p}} = ku^2 \tag{2.26}$$

by the addition of third- and fourth-order terms

$$E_{\mathrm{p}} = ku^2 - fu^3 - gu^4, \tag{2.27}$$

where all the coefficients k, f and g are positive.

Alternatively, you will often find the anharmonic potential expressed as a Lennard-Jones 6–12 potential which has the form shown in Fig. 2.6:

$$E_{\mathrm{p}} = \frac{A}{a^6} + \frac{B}{a^{12}}, \tag{2.28}$$

where a is the lattice spacing. However, this is also just a simple method of obtaining an approximate form of energy as a function of interatomic spacing.

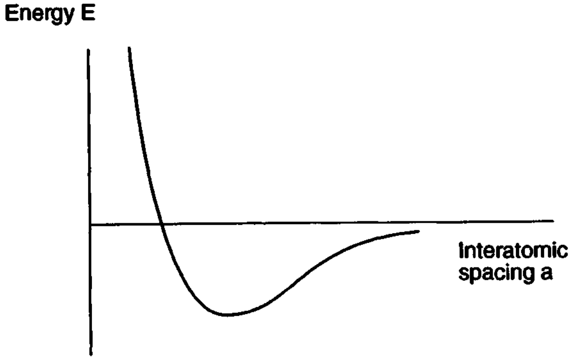


Figure 2.6 Anharmonic potential as a function of distance between two ionic cores.

2.2.4 Gruneisen parameters

How can anharmonicity in the lattice be described?

If the characteristic vibration frequency of the atoms within a crystal lattice changes with volume due to variation of the stiffness coefficient with atomic displacement, then the simplest assumption to make is that the fractional change in frequency is the same for all vibrational modes and that this is proportional to the fractional changes in volume

$$\frac{\Delta\omega}{\omega} = \gamma\frac{\Delta v}{v}, \tag{2.29}$$

where the coefficient of proportionality γ is the so-called Grüneisen parameter.

The fractional change in volume can be directly related to a change in pressure ΔP that causes the volume change

$$\Delta P = K\frac{\Delta v}{v}, \tag{2.30}$$

where K is the bulk elastic modulus. Therefore,

$$\frac{\Delta\omega}{\omega} = \frac{\gamma}{K}\Delta P. \tag{2.31}$$

Considering that the vibrational frequency is related to the stiffness coefficient as shown in Section 2.2.2, it becomes possible to relate the Grüneisen parameter directly to the change in stiffness for a given change in volume:

$$\omega = \left(\frac{2k}{m}\right)^{1/2} \tag{2.32}$$

$$\mathrm{d}\omega = \frac{1}{m}\left(\frac{2k}{m}\right)^{-1/2}\mathrm{d}k, \tag{2.33}$$

and, therefore,

$$\frac{\mathrm{d}\omega}{\omega} = \frac{1}{2k}\mathrm{d}k = \gamma\frac{\mathrm{d}v}{v}. \tag{2.34}$$

This equation shows how the vibrational modes of the lattice change with strain and how this relates to changes in the stiffness of the lattice. It can be shown, for example, that the Grüneisen parameter γ and the thermal expansion coefficient α are related via the specific heat capacity C_v and the bulk elastic modulus K

$$\alpha = \gamma\frac{C_v}{3K}. \tag{2.35}$$

2.3 SPECIFIC HEAT CAPACITY

Can the observed specific heat capacity be interpreted in terms of the lattice properties?

As we consider the physical properties of materials and try to explain their behaviour we will find that these properties often can be attributed primarily to either the lattice or the electron sea. The first of these that we will consider is the heat capacity or the specific heat, which is the heat capacity per unit mass. This is determined largely by the lattice, although there is a contribution due to the electrons, which we shall discuss later.

The vibrations of a crystal lattice are related to the heat content of the system and hence to the thermodynamic temperature. We have already noted the Dulong and Petit law which simply states that the molar heat capacity is a constant. Let us see how this can be obtained from a consideration of lattice vibrations.

2.3.1 Classical theory of heat capacity

How can the heat capacity of a material be explained in terms of the vibration of the atoms?

From classical statistical thermodynamics we expect a thermal energy k_BT to be associated with each mode of vibration at any given temperature T. If there are N atoms each with three degrees of freedom, then there are $3N$ modes each with energy k_BT. The internal energy is therefore,

$$U = 3Nk_BT, \tag{2.36}$$

and since the heat capacity at constant volume is simply $C_v = dU/dT$, then

$$C_v = 3Nk_B. \tag{2.37}$$

If we consider a mole of material, for which $N = 6.02 \times 10^{23}$ atoms, and $k_B = 1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ we obtain

$$C_v = 24.94\,\mathrm{J\,mole^{-1}\,K^{-1}}, \tag{2.38}$$

which is the classical value of the molar heat capacity predicted by the Dulong–Petit law.

Notice, however, from Fig. 2.7 that the heat capacity is not constant as temperature changes. Therefore, a more sophisticated theory is needed, particularly at lower temperatures.

In the quantum theory of heat, atoms can only take up energy in discrete amounts (quanta) rather than continuously, as we have assumed so far. At room temperature the magnitudes of the energy quanta that most atoms can absorb are small enough that these can be provided by the thermal vibrations of the surroundings. Exchange of heat with the surroundings can occur almost continuously leading to equipartition of energy and this means that these atoms have the classically expected heat capacity. Examples of materials that come into this category at room temperature are lead, copper, aluminium and the other materials given in Table 1.6.

However, in the case of materials which have light strongly bound atoms, such as diamond, the energy quanta that must be absorbed are much larger. This means

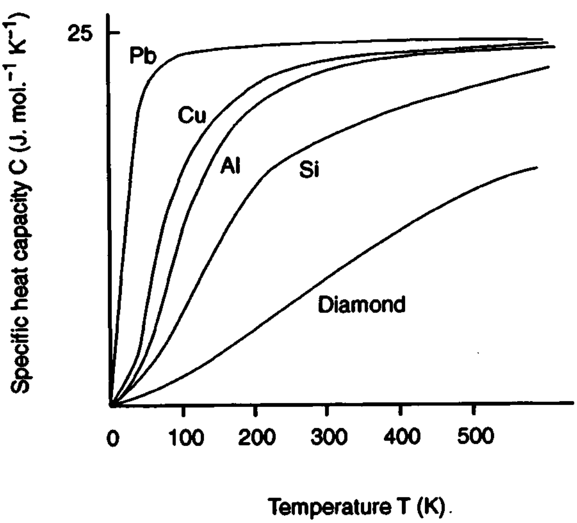


Figure 2.7 Variation of specific heat C_v of various materials as a function of temperature.

that exchange of heat with the surroundings is no longer continuous and so equipartition of energy is not established at room temperature. The heat capacity is therefore lower than the classically expected value.

For sufficiently low temperatures equipartition of energy breaks down for all materials. This occurs when the available thermal energy quanta become comparable to the lattice vibration quanta. Therefore at sufficiently low temperatures the heat capacities of lead, copper, aluminium and other metals are reduced as shown in Fig. 2.7 and they no longer obey the classical Dulong–Petit law. In fact, at low temperatures the heat capacity obeys a relation of the form,

$$C_v = 233.8 N_0 k_B \left(\frac{T}{\theta_D} \right)^3 \, \mathrm{J\,mol^{-1}\,K^{-1}}, \tag{2.39}$$

which is known as Debye's law. In this law, θ_D is a characteristic temperature known as the Debye temperature which varies from material to material.

2.3.2 Quantum corrections to the theory of heat capacity

How can we account for the deviations from the Dulong–Petit law at low temperatures?

We will now consider a correction to the theory of heat capacity which incorporates quantum effects. This we will term the Einstein–Debye theory in recognition of the two main contributors. The original ideas of the single-frequency Einstein theory were extended and generalized by Debye to give an equation which applies across the entire temperature range.

Einstein showed that the lattice vibrations of crystalline solids should be quantized and that therefore only certain vibrational modes were allowed. These vibrational quanta are called phonons. Raising the temperature increases the amount of lattice vibrations and increases the number of phonons. If we consider a single oscillator, that is an atomic core, oscillating at frequency ω, the average number of phonons N at a given temperature T obeys the following equation [9],

$$N = \frac{1}{\exp(\hbar\omega/k_B T) - 1}, \tag{2.40}$$

where $\hbar$ is Planck's constant divided by 2π (which has the value 1.054×10^{-34} J s), $\hbar\omega$ is the phonon energy, and $k_B T$ is the thermal energy available to each vibrational mode at a temperature T, including both kinetic and potential energy. For a single frequency of oscillation, therefore, the thermal energy of the material is

$$U = N\hbar\omega, \tag{2.41}$$

$$U = \left(\frac{1}{\exp(\hbar\omega/k_B T) - 1}\right)\hbar\omega. \tag{2.42}$$

Considering the thermal energy of a solid, and remembering that a mole of material contains N_0 atoms and hence $3N_0$ oscillators, the thermal energy per mole is

$$U = 3N_0 \frac{\hbar\omega}{\exp(\hbar\omega/k_B T) - 1}. \tag{2.43}$$

Consequently, according to the Einstein theory the molar heat capacity should be

$$C_v = 3N_0 k_B \left(\frac{\hbar\omega}{k_B T}\right)^2 \frac{\exp(\hbar\omega/k_B T)}{(\exp(\hbar\omega/k_B T) - 1)^2}. \tag{2.44}$$

A graph of the prediction of the Einstein theory of heat capacity is given in Fig. 2.8. Note that for high temperatures we obtain $\exp(\hbar\omega/k_B T) \cong 1 + (\hbar\omega/k_B T)$ giving $C_v \cong 3N_0 k_B$ which is the classical limit expressed by the Dulong–Petit law.

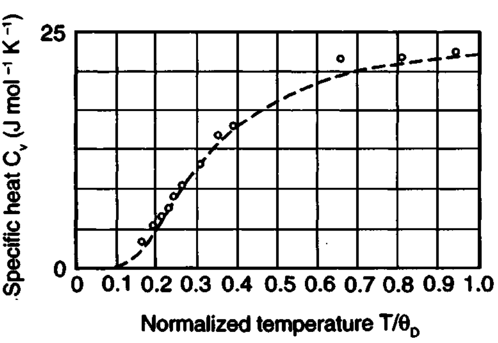


Figure 2.8 Comparison of the prediction of the Einstein model with observed values of heat capacity. Reproduced with permission from C. Kittel, *Introduction to Solid State Physics*, 6th edn, published by John Wiley & Sons, 1986.

The Einstein model of heat capacity works well at high and intermediate temperatures. It predicts a heat capacity which decreases as the temperature is reduced, which is in accordance with experimental observation. But the decrease that is observed in practice is not quite as rapid as the Einstein model suggests. The theory therefore does not work well at low temperatures and so a further correction to the theory is needed at these temperatures.

2.3.3 Extended theory of lattice vibrations

Can the single frequency Einstein theory be improved to give better agreement with observations at low temperature?

In the Einstein model only one frequency of vibration ω was considered. If we allow interactions between the atoms, many more frequencies can exist which range from the Einstein frequency down to frequencies in the acoustic range. This extension to the theory was developed by Debye.

It is then necessary to know the number of atoms vibrating with any given frequency ω. This is the vibrational density of states $D(\omega)$ and in three dimensions this is given by, [1, p. 106]

$$D(\omega) = \frac{3V\omega^2}{2\pi^2 v^3}, \tag{2.45}$$

where V is the volume of the specimen and v is the velocity of sound which is assumed constant. The total energy of vibration of the solid is then simply

$$U = \int U_{\text{osc}}(\omega) D(\omega) \mathrm{d}\omega, \tag{2.46}$$

where $U_{\text{osc}}(\omega)$ is the energy of one oscillator at frequency ω which is given by the same expression as before

$$U_{\text{osc}}(\omega) = \frac{\hbar\omega}{\exp(\hbar\omega/k_B T) - 1}. \tag{2.47}$$

We now need to integrate this energy over the range of allowed vibrational modes. Substituting the expressions for U_{osc} and $D(\omega)$ into the above equation for the total energy of vibration gives

$$U = \frac{3V\hbar}{2\pi^2 v^3} \int_0^{\omega_D} \frac{\omega^3}{\exp(\hbar\omega/k_B T) - 1} \mathrm{d}\omega \tag{2.48}$$

where ω_D is the so-called 'Debye frequency' above which the oscillators behave classically. If we wish, we can define a 'Debye temperature' θ_D in terms of this Debye frequency

$$\theta_D = \frac{\hbar\omega_D}{k_B}. \tag{2.49}$$

2.3.4 Relationship of Debye temperature to lattice properties

How can we interpret the Debye temperature in terms of the properties of the lattice and the interactions between the atoms?

If the potential seen by any individual atom has the form $E_p = ku^2$, then the equation of motion of the atom is

$$F = m\frac{d^2u}{dt^2} = -2ku, \tag{2.50}$$

and by solving this equation the result is simple harmonic motion of frequency ω_0 given by

$$\omega_0^2 = \frac{2k}{m}, \tag{2.51}$$

and if the phonon energy is equated with the thermal energy at the Debye temperature θ_D, then

$$\hbar\omega_0 = k_B\theta_D \tag{2.52}$$

$$= \hbar\sqrt{\frac{2k}{m}}, \tag{2.53}$$

and consequently,

$$\theta_D = \frac{\hbar}{k_B}\sqrt{\frac{2k}{m}}. \tag{2.54}$$

The consequence of this is that stiff materials (high k) with light atoms (small m) have high Debye temperatures, and vice versa.

2.3.5 Significance of the Debye temperature

Can the Debye temperature be related to macroscopic properties?

The Debye temperature marks the boundary between the high-temperature classical behaviour which approximately follows Maxwell–Boltzmann statistics (and hence leads to the Dulong–Petit law) and a low-temperature region in which quantum statistics must be used.

Since the Debye temperature marks the boundary between the quantum and classical regimes, its value tells us something about the atomic bonding in the material we are dealing with. A high value of θ_D implies a lattice with light atoms and strong interactions between the atoms. For example, diamond has $\theta_D \cong 2000\,K$. On the other hand, lead, which has heavy atoms that are weakly bound together, has $\theta_D \cong 100\,K$.

The Debye temperature is, in many ways, analogous to the Fermi level in the theory of electron states which we will discuss in Chapter 4. It plays the same role in the theory of lattice vibrations as the Fermi temperature plays in the theory of electrons in metals.

2.3.6 Quantum theory of heat capacity

What is the heat capacity expected on the basis of the quantum theory of lattice vibrations?

For N atoms or ions the total phonon energy is,

$$U = 9Nk_BT\left(\frac{T}{\theta_D}\right)^3 \int_0^{(\hbar\omega_D/k_BT)} \left(\frac{x^3}{e^x - 1}\right) dx, \tag{2.55}$$

where

$$x = \frac{\hbar\omega}{k_BT}. \tag{2.56}$$

The Debye heat capacity is the derivative of the total energy of vibration U with respect to the temperature T

$$C_v = \frac{dU}{dT}. \tag{2.57}$$

The molar heat capacity is then

$$C_v = \frac{3V_0\hbar^2}{2\pi^2 v^3 k_B T^2} \int_0^{\omega_D} \frac{\omega^4 \exp(\hbar\omega/k_BT)}{[\exp(\hbar\omega/k_BT) - 1]^2} d\omega, \tag{2.58}$$

$$C_v = 9k_BN_0\left(\frac{T}{\theta_D}\right)^3 \int_0^{\theta_D/T} \frac{x^4 e^x}{(e^x - 1)^2} dx, \tag{2.59}$$

where V_0 is the molar volume and N_0 is the number of atoms or ions in a mole. A plot of the Debye heat capacity C_v against temperature is given in Fig. 2.9.

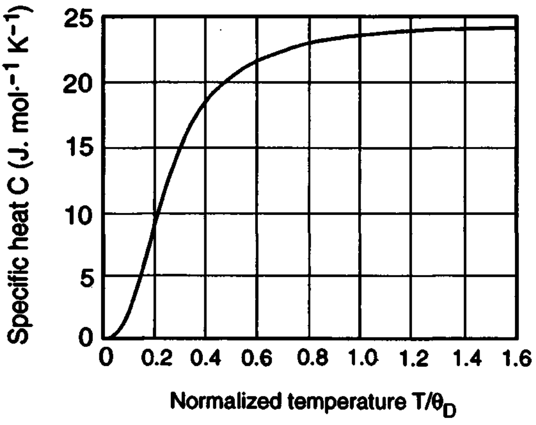


Figure 2.9 Variation of the heat capacity C_v with temperature according to the Debye theory. Reproduced with permission from C. Kittel, *Introduction to Solid State Physics*, 6th edn, published by John Wiley & Sons, 1986.

2.3.7 Heat capacity at low temperatures: the Debye law

Can we obtain a simple expression for the heat capacity at low temperatures?
The heat capacity can be found from the above expression by evaluating the integral on the right-hand side. At low temperatures this results in the following expression for the heat capacity

$$C_v = \frac{12\pi^4}{5} N_0 k_B \left(\frac{T}{\theta_D}\right)^3, \qquad (2.60)$$

which gives the experimentally observed dependence of the heat capacity on T^3 at low temperatures known as the Debye T^3 law which was mentioned in Section 2.3.1.

2.3.8 Heat capacity at high temperatures: the classical limit

What is the predicted heat capacity at high temperatures?
At high temperatures the integral has the value $x^3/3$ where $x = \theta_D/T$. So that

$$C_v = 3N_0 k_B, \qquad (2.61)$$

which, of course, is the expected classical Dulong–Petit result.

2.4 CONCLUSIONS

What important results did we find that went beyond the simple continuum description of materials?
In this chapter we have discussed some of the properties of materials that arise principally from the behaviour of the atoms. We have also introduced the concept of quantization of allowed vibrational energies in discrete lattices. This serves as a precursor to the discussion of quantization of electron energy levels in Chapters 4, 5 and 6. It has given us a relatively simple introduction to the concept of quantized energy states because it can be quite easily visualized, even from a classical argument, why the quantization of lattice vibration occurs.

When we come to discussing the quantization of electron energies rather more abstraction will be required. However, having understood the reasons for quantization of lattice vibrations in this chapter, it should be easier to follow the discussion of quantization of electron energies in which the concept of imposed boundary conditions again plays a crucial role.

REFERENCES

1. C. Kittel (1986) *Introduction to Solid State Physics*, 6th edn, Wiley, New York.
2. W. B. Pearson (1972) *Crystal Chemistry and Physics of Metals and Alloys*, Wiley Interscience, New York.
3. H. D. Megaw (1973) *Crystal Structures: a Working Approach*, W. B. Saunders, Philadelphia.
4. B. D. Cullity (1978) *X-ray Diffraction*, 2nd edn, Addison-Wesley, Reading, Massachusetts.

5. G. C. Maitland, M. Rigby, E. B. Smith and W. A. Wakeham (1981) *Intermolecular Forces: Their Origin and Determination*, Oxford.
6. M. Born and K. Huang (1954) *Dynamical Theory of Crystal Lattices*, Oxford.
7. W. Cochran (1973) *Dynamics of Atoms in Crystals*, Crane & Russak.
8. R. A. Cowley (1968) *Anharmonic Crystals, Rep. Prog. Phys.* 31, 123.
9. M. W. Zemansky (1968) *Heat and Thermodynamics*, 5th edn, McGraw-Hill, New York, p. 312.

FURTHER READING

J. C. Anderson and K. D. Leaver (1969) *Materials Science*, Van Nostrand Reinhold, New York.

H. M. Rosenberg (1978) *The Solid State*, 2nd edn, Oxford University Press, Chapters 5 and 6.

B. K. Tanner (1995) *Introduction to the Physics of Electrons in Solids*, Chapman and Hall, London.

EXERCISES

Exercise 2.1 Elastic modulus of a linear atomic lattice

The potential energy of a pair of atoms in a crystal is of the form

$$E_{\mathrm{p}}(r) = \alpha_1 a^{-9} - \alpha_2 a^{-1},$$

where a is the interatomic separation. The equilibrium separation is 0.3 nm and the cohesive energy is 4 eV (equivalent to 386 kJ mol^{-1}). Find the effective modulus of elasticity for the pair of atoms and determine the force which would be needed to reduce the spacing by 1%.

Exercise 2.2 Lattice stabilized by electrostatic repulsion

Consider a one-dimensional lattice made up of atoms with a charge of 1.6×10^{-19} C and a mass of 107×10^{-27} kg each. The array is held together by electrostatic forces with neighbouring atoms at a distance of 0.5 nm. Estimate the velocity of a long wavelength lattice vibration and calculate the elastic Young's modulus. (You may assume the force between the ions is given by Coulomb's law $F = q_1 q_2 / 4\pi\varepsilon_0 x^2$, where x is the separation between the charges q_1 and q_2.)

Exercise 2.3 Classical and Debye theories of specific heat

Use the Dulong–Petit relation to calculate the classically expected thermal energy of 1 g mole of material at 300 K. Aluminium has a Debye temperature of 430 K. Prove that the Debye theory and the Dulong–Petit law give the same results at high temperature. Estimate the thermal energy of 1 g mole of aluminium at 300 K by using Fig. 2.9. Explain why the results are different.

Exercise 2.4 Lattice vibrations

What is the difference between a damped and an undamped vibration? Write the appropriate equations describing each of these types of vibration in three dimensions and show how these equations can be applied to a discrete lattice.

What is meant by a boundary condition, and how are the solutions of a wave equation affected by the boundary conditions?

Exercise 2.5 Interatomic potential

Explain how the elastic modulus of a material is determined by the interatomic potential, particularly the harmonic force constant (stiffness coefficient) of the potential. If this interatomic potential is not quite harmonic, what are the consequences of the 'anharmonicity' for the elastic modulus and thermal expansion? How can we quantify the anharmonicity in terms of measurable quantitites?

Exercise 2.6 Heat capacity

Find the value of the heat capacity C_v at the temperature $T_D = \hbar\omega/k_B$ as a fraction of the classically expected value, $3Nk_B$, according to Einstein's theory of heat capacity. The value of this temperature T_D is given below for some materials. Find the corresponding values of the frequency ω and explain why high values of this temperature T_D occur in stiffer, lighter materials.

Material	**T_D (K)**
Pb	95
Au	170
NaCl	280
Fe	360
Se	650
C (diamond)	1850

3 CONDUCTION ELECTRONS IN MATERIALS – CLASSICAL APPROACH

OBJECTIVE

In this chapter we shall approach the description of electrons in solids using one of the simplest models possible, that of electrons as classical particles moving almost freely within the material experiencing minimal interactions with the ionic potential. In fact, the model assumes that the ionic potential is completely flat, and that the only constraints on electron motion, apart from electron–electron collisions, are provided by the physical boundary of the material. At first, such a model seems so far from reality as to be probably of little use; however, quite surprisingly the model can give some useful insights and provides an initial description of electrical and thermal conductivity, the Hall effect and the Wiedemann–Franz law. Nevertheless, this classical model has no way of distinguishing between conductors and insulators and gives an incorrect prediction of the heat capacity of the electrons, so that ultimately a more comprehensive model is needed.

3.1 ELECTRONS AS CLASSICAL PARTICLES IN MATERIALS

What models can be developed for the behaviour of electrons in materials?

So far we have paid no attention to the effects that are due to the electrons in a solid. However, it was realized many years ago that, for example, the electrical conductivity of a metal was due exclusively to the motion of electrons inside the material.

The classical free electron model was developed by Drude and Lorentz [1–4], in which it was assumed that electrons were classical particles with a kinetic energy $\frac{3}{2}k_BT$ where T is the thermodynamic temperature and k_B is Boltzmann's constant. These free electrons, which comprised only a small fraction of the total number of electrons in each atom, could account for both the electrical and thermal conductivity of a metal and also the optical reflectance at low frequencies.

Given the simplicity of this model, its successes are quite impressive. The most quoted success was the explanation of the Wiedemann–Franz law by the model. One notable failure, however, was its prediction of the electronic contribution to the heat capacity. According to this model each electron should contribute $\frac{3}{2}k_B$ to the total heat capacity. In practice, this is not observed and, in fact, the electronic heat capacity is smaller by two orders of magnitude than expected on the basis of this theory.

Before going on to look at the model in detail let us consider how much of the volume inside a typical material is actually occupied by the atomic cores, and how much is 'empty space', so to speak. Consider, for example, a material in which the lattice spacing is a and the radius of the ionic cores is r and which forms a body-centred cubic structure. Since there are two atoms or ions per unit cell, the fraction of space occupied by the atoms or ions is $\frac{8}{3}\pi(\frac{r}{a})^3$. In sodium, for example, $r = 0.98 \times 10^{-10}$ m and $a = 4.2 \times 10^{-10}$ m. Therefore, the fraction of space occupied by the atoms or ions is about 11%, meaning that the majority of the volume of the material is simply empty space for the electrons to move through.

3.1.1 Basis of the classical model

How might electrons behave inside a material?

The Drude model assumes that the metal behaves like an empty box containing a free electron 'gas'. The free electron gas consists of the outer conduction electrons of the individual atoms. So a monovalent metal such as sodium contributes one electron per atom.

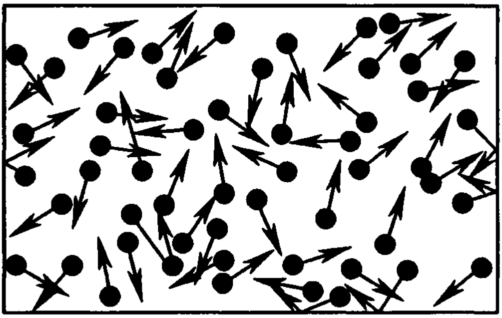


Figure 3.1 Classical free electron model of a solid consisting of a box containing a gas of free particles which obey the kinetic theory.

Therefore, the number of free electrons per unit volume will be

$$N = \frac{N_0 \rho}{W_A} Z, \tag{3.1}$$

where Z is the valence of the atom, ρ is the density and W_A the atomic mass. These electrons have energies that are due to the thermal energy of the material and according to the model behave like the atoms of a gas in the kinetic theory of gases. The number density of conduction electrons in a solid is typically 10^{28} m^{-3} (10^{22} cm^{-3}) which is, of course, about three orders of magnitude greater than for a typical gas at normal temperatures and pressures.

This should cause us some concern over the viability of the model. Another concern is, of course, that the material can hardly be considered to be an empty box since there are ions located on the lattice sites and these are electrically charged. Despite this, the Drude model rigidly follows the methods of the kinetic theory of a dilute, neutral gas, using Maxwell–Boltzmann statistics.

The assumptions of the model are as follows:

1 Collisions between electrons are instantaneous and lead to scattering.
2 Between these collisions, other interactions of the electrons with each other and with the ions are neglected in detail (although the interactions with the lattice are incorporated implicitly through an averaged resistive term in the equation of motion).
3 The mean free time of the electron between collisions is τ, and this time is independent of the electron's position and velocity.
4 Electrons achieve thermal equilibrium with their surroundings only through collisions with other electrons.

We will now develop the theory to help explain some of the well-known electrical and thermal properties of metals. According to the kinetic theory the kinetic energy of each electron which is due to thermal energy will be $3k_BT/2$. In the absence of an applied field the direction of motion is random.

3.2 ELECTRICAL PROPERTIES AND THE CLASSICAL FREE ELECTRON MODEL

What effect does an applied electric field have on the electrons inside the material? In the absence of a field, the electrons move randomly in all directions and consequently the net drift velocity is zero. When an electric field ξ is applied the electrons are accelerated with force $F = e\xi$, so that

$$m \frac{\mathrm{d}v}{\mathrm{d}t} = e\xi, \tag{3.2}$$

where m is the mass of the electrons (9.109×10^{-31} kg), e is the charge on the electrons (-1.602×10^{-19} C), and v is the velocity of the electrons.

There are also interactions of the drifting electrons with some of the lattice ions. This leads to resistance in the metal and ensures that the electrons are not accelerated indefinitely when a field is applied. Therefore an additional resistive term is needed. The equation of motion then becomes

$$m \frac{\mathrm{d}v}{\mathrm{d}t} = e\xi - \gamma v, \tag{3.3}$$

where γ is a constant which represents a resistive force which is proportional to electron velocity, preventing the electrons from being accelerated to infinite velocity.

The above equation is analogous to that encountered in viscosity, in which the resistive force is proportional to the velocity v. The electrons therefore reach an equilibrium velocity under the action of an accelerating field and this equilibrium velocity is determined by the interactions with the lattice.

3.2.1 Electrical conductivity

Can we obtain a first principles expression for the electrical conductivity from the motion of the free electrons?

In the steady state, when the electrons have reached a final velocity v_f the force due to the electric field is equal to the resistive force, and opposite in direction

$$m\frac{dv}{dt} = e\xi - \gamma v_f = 0, \tag{3.4}$$

and so the coefficient γ can be obtained from

$$\gamma = \frac{e\xi}{v_f} = \frac{e}{\mu}, \tag{3.5}$$

where v_f/ξ is known as the mobility of the electrons and is usually denoted μ. If this is used in the equation of motion,

$$m\frac{dv}{dt} + \frac{e\xi}{v_f}v = e\xi, \tag{3.6}$$

which leads to the solution

$$v = v_f\left\{1 - \exp\left(-\frac{e\xi}{mv_f}t\right)\right\}. \tag{3.7}$$

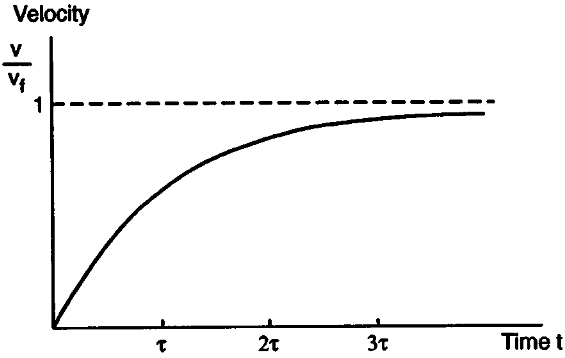


Figure 3.2 Variation of the velocity of free electrons with time. The electrons reach a terminal velocity which is dependent on the resistive force caused by interactions with the lattice.

We can define a relaxation time τ as

$$\tau = \frac{mv_f}{e\xi} = \frac{m}{\gamma}, \tag{3.8}$$

so that

$$v = v_f\left\{1 - \exp\left(\frac{-t}{\tau}\right)\right\}. \tag{3.9}$$

The current density J is the product of the number of charge carriers per unit volume N, the charge per carrier e and the velocity v,

$$J = Nev. \tag{3.10}$$

Since N should really be the number density of free electrons, rather than the total number of electrons per unit volume, we will use N_f instead of N, and the final velocity v_f under steady-state conditions,

$$J = N_f v_f e, \tag{3.11}$$

and remembering that the final drift velocity is $v_f = e\xi\tau/m = e\xi/\gamma$

$$J = \frac{N_f e^2 \tau}{m}\xi, \tag{3.12}$$

or, since $J = \sigma\xi$,

$$\sigma = \frac{N_f e^2 \tau}{m} = \frac{N_f e^2}{\gamma}, \tag{3.13}$$

therefore we have derived a direct relationship between the electrical conductivity σ and the resistive coefficient γ from the equation of motion of the electrons. Typically for a metal τ is 10^{-14} to 10^{-15} s, N_f is $10^{28}\,\mathrm{m}^{-3}$, e is 1.6×10^{-19} C and m is 9.1×10^{-31} kg, giving,

$$\sigma = 0.28 \times 10^6 \text{ to } 2.8 \times 10^6\,\Omega^{-1}\,\mathrm{m}^{-1}. \tag{3.14}$$

3.2.2 Ohm's law

Can we go even further and derive the well-known expression of Ohm's law from this free electron model?

If we rearrange the above equations putting the current density $J = I/A$ where I is the current and A is the cross-sectional area, and the electric field $\xi = V/l$ where V is the voltage and l the length

$$\frac{I}{A} = \frac{N_f e^2 \tau}{m}\frac{V}{l}, \tag{3.15}$$

$$I = \frac{N_f e^2 \tau A}{ml} V. \tag{3.16}$$

This then is an expression for Ohm's law which gives the relationship between current in an electrically conducting sample and the voltage across the sample:

$$I = \frac{V}{R}, \tag{3.17}$$

so that

$$R = \frac{m\ell}{N_f e^2 \tau A} = \frac{\ell}{\sigma A}. \tag{3.18}$$

We can see therefore that the familiar macroscopic relationship known as Ohm's law can be derived on the basis of the classical free electron model. This gives some confidence in the model because it enables us to predict a known law on the basis of the model, and to relate model parameters to measurable properties.

3.3 THERMAL PROPERTIES AND THE CLASSICAL FREE ELECTRON MODEL

How can the thermal properties be explained on the basis of the free electron model?

We now look at the explanation of thermal conductivity and the Wiedemann–Franz law provided by the classical free electron theory. In these, the theory gives a satisfactory description.

3.3.1 Thermal conductivity

Can we obtain an expression for the thermal conductivity from the classical electron model equations?

The thermal conductivity K is the rate of transfer of thermal energy per unit thickness per unit area per unit temperature gradient (see Section 1.5.1),

$$K = -\frac{1}{A}\frac{\mathrm{d}Q}{\mathrm{d}t}\frac{\mathrm{d}x}{\mathrm{d}T}. \tag{3.19}$$

We can define $(\mathrm{d}Q/\mathrm{d}t)/A$ as the thermal current density, J_Q, so that in one dimension

$$J_Q = -K\frac{\mathrm{d}T}{\mathrm{d}x}. \tag{3.20}$$

Consider a one-dimensional temperature gradient in a material which contains free electrons,

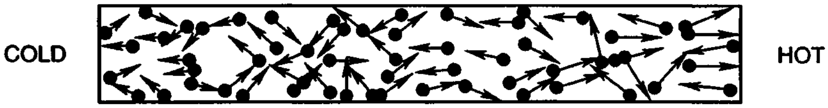


Figure 3.3 Free electrons moving in a material under the action of a one-dimensional temperature gradient.

Let $E(T)$ be the thermal energy per electron at temperature T, and the temperature at a point x be $T(x)$. An electron whose last collision was at another point x' will have an energy $E(T(x'))$ because the electrons can only gain or lose energy by colliding with other electrons according to the theory. Assuming that the time between collisions is τ, those electrons from the high-temperature end will have energy,

$$E = E\{T(x - v\tau)\}. \tag{3.21}$$

Their contribution to the thermal current density will be the product of the number of electrons moving towards the cold end (←), let us say $N_f/2$, their velocity v and their energy E:

$$\Delta J_Q^{\leftarrow} = (N_f v/2)E\{T(x - v\tau)\}, \tag{3.22}$$

similarly for electrons moving towards the hot end of the material (→),

$$\Delta J_Q^{\rightarrow} = (N_f v/2) E\{T(x + v\tau)\}, \tag{3.23}$$

the net contribution to the thermal current density is then,

$$J_Q = (N_f v/2)[E\{T(x - v\tau)\} - E\{T(x + v\tau)\}], \tag{3.24}$$

and if the temperature difference over the range $v\tau$ is small then

$$E\{T(x - v\tau)\} - E\{T(x + v\tau)\} = -2v\tau(\mathrm{d}E/\mathrm{d}x), \tag{3.25}$$

and therefore

$$J_Q = N_f v^2 \tau \frac{\mathrm{d}E}{\mathrm{d}T}\left(-\frac{\mathrm{d}T}{\mathrm{d}x}\right). \tag{3.26}$$

In fact, to be exact, since we have a distribution of velocities we should use the mean square value of velocity for the electron gas, so the equation becomes

$$J_Q = N_f \langle v^2 \rangle \tau \frac{\mathrm{d}E}{\mathrm{d}T}\left(-\frac{\mathrm{d}T}{\mathrm{d}x}\right), \tag{3.27}$$

where E is the thermal energy per electron, and $N_f(\mathrm{d}E/\mathrm{d}T) = (N_{\mathrm{Tot}}/V)(\mathrm{d}E/\mathrm{d}T) = C_V^e$ is the electronic specific heat

$$J_Q = \langle v^2 \rangle \tau C_V^e \left(-\frac{\mathrm{d}T}{\mathrm{d}x}\right), \tag{3.28}$$

and by comparison with the above equations, we see that we now have an expression for the thermal conductivity K

$$K = \langle v^2 \rangle \tau C_V^e, \tag{3.29}$$

or in three dimensions

$$K = \tfrac{1}{3} \langle v^2 \rangle \tau C_V^e, \tag{3.30}$$

where $\langle v^2 \rangle$ is the mean square electronic velocity.

This shows that the free electron theory is able to account for the thermal conductivity of a metal. Typical values of $\langle v^2 \rangle$, τ and the electronic specific heat C_V^e are

$$\langle v^2 \rangle = 5 \times 10^{10}\,\mathrm{m^2\,s^{-2}} \tag{3.31}$$

$$\tau = 2 \times 10^{-14}\,\mathrm{s} \tag{3.32}$$

and using

$$N_f = 5 \times 10^{28}\,\mathrm{m^{-3}} \tag{3.33}$$

$$C_V^e = 1 \times 10^6\,\mathrm{J\,m^{-3}\,K^{-1}}, \tag{3.34}$$

this gives the thermal conductivity as

$$K = 300\,\mathrm{J\,s^{-1}\,m^{-1}\,K^{-1}}, \tag{3.35}$$

which compares well with the values of thermal conductivity K for copper, silver and gold which are 398, 428 and $315\,\mathrm{J\,s^{-1}\,m^{-1}\,K^{-1}}$, respectively.

3.3.2 The Wiedemann–Franz law

Can the relationship between electrical and thermal conductivities be explained by the classical electron model?

We have mentioned the Wiedemann–Franz law in the previous chapter. The law can be explained on the basis of the classical free electron model. Expressions have been obtained for both the electrical and thermal conductivities, and therefore the ratio can be expressed in terms of the electronic properties,

$$\frac{K}{\sigma} = \frac{1}{3}\frac{\langle v^2\rangle C_{\mathrm{V}}^{\mathrm{e}} m\tau}{N_{\mathrm{f}} e^2 \tau}, \tag{3.36}$$

$$= \frac{1}{3}\frac{m C_{\mathrm{V}}^{\mathrm{e}}\langle v^2\rangle}{N_{\mathrm{f}} e^2}. \tag{3.37}$$

Using the relation between the heat capacity and the number of free electrons

$$C_{\mathrm{V}}^{\mathrm{e}} = \tfrac{3}{2} N_{\mathrm{f}} k_{\mathrm{B}}, \tag{3.38}$$

and relating the kinetic energy of the electrons to the thermal energy

$$\tfrac{1}{2} m\langle v^2\rangle = \tfrac{3}{2} k_{\mathrm{B}} T, \tag{3.39}$$

gives the following ratio between thermal and electrical conductivity,

$$\frac{K}{\sigma} = \frac{3}{2}\left(\frac{k_{\mathrm{B}}}{e}\right)^2 T, \tag{3.40}$$

and therefore,

$$\frac{K}{\sigma T} = \frac{3}{2}\left(\frac{k_{\mathrm{B}}}{e}\right)^2, \tag{3.41}$$

which is a constant with value $1.11 \times 10^{-8}\,\mathrm{W\,\Omega\,K^{-1}}$ which is in reasonable agreement with the Wiedemann–Franz law, giving a value about half of that observed experimentally. The value of $\frac{3}{2}(k_{\mathrm{B}}/e)^2$ is known as the Lorentz number.

So the theory does not quite predict the correct result. However, it does remarkably well for such a simple model, and if this were the only problem it could be overcome with some corrections. As we shall see there are more serious problems with the model which ultimately require a new approach to the description of electrons in metals.

In fact, Drude's original calculation used $E = 3k_{\mathrm{B}}T$ as the energy per electron (assuming all thermal energy was kinetic energy) leading to a much better agreement with the Wiedemann–Franz law with a value of $K/\sigma T = 2.2 \times 10^{-8}\,\mathrm{W}\,\Omega\,\mathrm{K}^{-1}$ which is exact to within experimental error. It was later shown by Lorentz that the correct expression is $E = \frac{3}{2}k_{\mathrm{B}}T$ because of the equipartition between potential and kinetic energy.

3.4 OPTICAL PROPERTIES OF METALS

How can the classical electron model explain the interaction of light with electrons in materials?

We have already given the equation of motion of free electrons in a solid as

$$m\frac{\mathrm{d}v}{\mathrm{d}t} + \gamma v = F, \tag{3.42}$$

where F is the force on the electrons due to any stimulus. If we consider the excitation due to an incident light beam, then the electric field ξ of the incident light is

$$\xi = \xi_0 \exp(\mathrm{i}\omega t), \tag{3.43}$$

and since the force on a charge e is given by $F = e\xi$, the force on the electrons due to the incident light is

$$F = e\xi_0 \exp(\mathrm{i}\omega t). \tag{3.44}$$

It has been shown above that the coefficient γ is equal to

$$\gamma = \frac{N_{\mathrm{f}}e^2}{\sigma_0}, \tag{3.45}$$

where N_{f} is the number of free electrons per unit volume, and where we have written σ_0 to distinguish the dc electrical conductivity instead of σ, and e is the electronic charge. Therefore, the equation of motion of the electrons is

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \frac{N_{\mathrm{f}}e^2}{\sigma_0}\frac{\mathrm{d}x}{\mathrm{d}t} = e\xi_0 \exp(\mathrm{i}\omega t). \tag{3.46}$$

We must expect a sinusoidal solution of this equation of motion for the electrons of the form $x = x_0 \exp(\mathrm{i}\omega t)$. Inserting this into the differential equation yields the following expression for the amplitude of oscillation, x_0, of the electrons,

$$x_0 = \frac{\xi_0}{(N_{\mathrm{f}}e\omega/\sigma_0)\mathrm{i} - (m\omega^2/e)} \tag{3.47}$$

3.4.1 Dielectric polarization and absorption

How does the classical free electron model account for the dielectric constant of a metal?

Since the electric polarization P is given by $P = eN_f x_0$, and the relative dielectric constant by $\varepsilon_r = 1 + (P/\varepsilon_0\xi)$ the classical free electron model necessarily leads to the following equation for the dielectric constant,

$$\varepsilon_r = 1 + \frac{1}{\varepsilon_0} \frac{N_f e}{(N_f e\omega/\sigma_0)\mathrm{i} - (m\omega^2/e)} \tag{3.48}$$

$$= 1 + \frac{1}{(\varepsilon_0\omega/\sigma_0)\mathrm{i} - (\varepsilon_0 m\omega^2/N_f e^2)}. \tag{3.49}$$

If we then substitute for the two characteristic frequency terms,

$$\omega_1 = \sqrt{\frac{e^2 N_f}{m\varepsilon_0}} = 2\pi\nu_1 \tag{3.50}$$

and

$$\omega_2 = \frac{\varepsilon_0\omega_1^2}{\sigma_0} = 2\pi\nu_2, \tag{3.51}$$

this gives

$$\varepsilon_r = 1 + \frac{\omega_1^2}{\mathrm{i}\omega\omega_2 - \omega^2} = 1 + \frac{\nu_1^2}{\mathrm{i}\nu\nu_2 - \nu^2}. \tag{3.52}$$

ν_1 is called the 'plasma frequency', and ν_2 is called the 'damping frequency'. If we separate the dielectric constant into real and imaginary components, then on the basis of the classical free electron model the real component of the relative dielectric constant, the polarization, is

$$\varepsilon_1 = n^2 - k^2 = 1 - \frac{\nu_1^2}{\nu^2 + \nu_2^2}, \tag{3.53}$$

and the imaginary component, the absorption, is

$$\varepsilon_2 = 2nk = \frac{\nu_2}{\nu}\left(\frac{\nu_1^2}{\nu^2 + \nu_2^2}\right). \tag{3.54}$$

This shows that the two independent optical constants n and k (or alternatively ε_1 and ε_2) can be derived theoretically from the Drude free electron model.

Table 3.1 Plasma and damping frequencies of various materials for the Drude free electron model.

Material	**Plasma frequency** ν_1 (s^{-1})	**Damping frequency** ν_2 (s^{-1})
Lithium	1.7×10^{15}	10×10^{12}
Sodium	1.4×10^{15}	4.8×10^{12}
Potassium	1.0×10^{15}	3.1×10^{12}
Rubidium	0.9×10^{15}	4.8×10^{12}
Caesium	0.75×10^{15}	5.15×10^{12}
Copper	$\sim 10^{15}$	4.7×10^{12}
Silver	$\sim 10^{15}$	4.4×10^{12}
Gold	$\sim 10^{15}$	5.9×10^{12}

The variations of the polarization ε_1 and the absorption ε_2 with frequency according to the classical Drude free electron model are shown in the following figures.

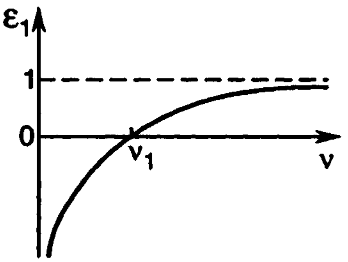


Figure 3.4 Polarization ε_1 as a function of frequency according to the classical free electron theory of metals.

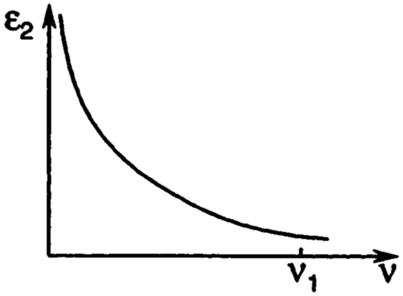


Figure 3.5 Absorption ε_2 as a function of frequency on the basis of the classical free electron theory of metals.

The term ε_1 is known as the 'polarization', the real component of the dielectric constant, and ε_2 is the 'absorption', the imaginary component of the dielectric constant. The 'plasma frequency' ν_1 marks the boundary between the low frequency reflecting regime and the high frequency transparent regime.

3.4.2 Optical reflectance

Can the classical free electron model explain the dependence of optical reflectance on wavelength of light in metals?

The variation of reflectance on the basis of the free electron model can now be derived. The reflectance at normal incidence is related to the two optical constants by the relation

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2}, \tag{3.55}$$

and consequently,

$$R = \frac{(\varepsilon_1^2 + \varepsilon_2^2)^{1/2} + 1 - \{2[(\varepsilon_1^2 + \varepsilon_2^2)^{1/2} + \varepsilon_1]\}^{1/2}}{(\varepsilon_1^2 + \varepsilon_2^2)^{1/2} + 1 + \{2[(\varepsilon_1^2 + \varepsilon_2^2)^{1/2} + \varepsilon_1]\}^{1/2}}, \tag{3.56}$$

when the above values of ε_1 and ε_2 can be substituted into the equation. The actual expression remains cumbersome; however, the limits as $\nu \rightarrow 0$ and $\nu \rightarrow \infty$ can be found relatively easily, see Example 3.2. The variation of reflectance with frequency of incident light is shown in Fig. 3.6.

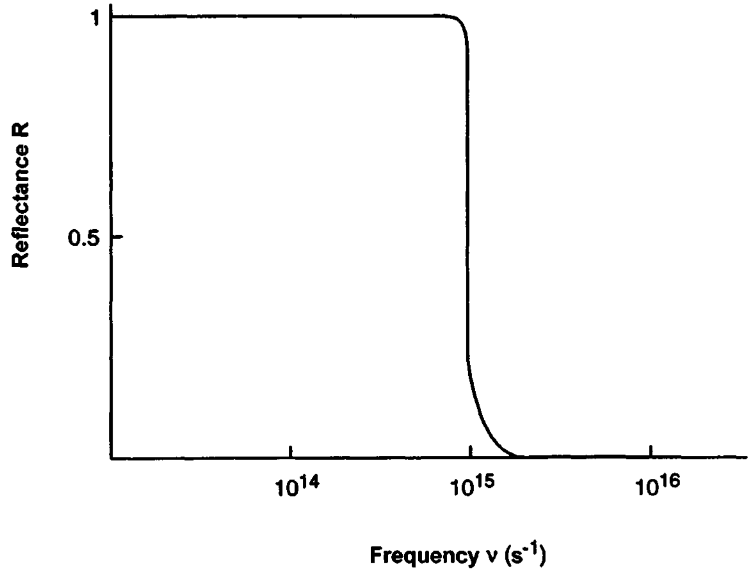


Figure 3.6 Variation of reflectance with frequency of incident light according to the Drude free electron model.

3.4.3 The Hagen–Rubens law

How can the well-known relationship between electrical conductivity and optical reflectance be explained?

From Maxwell's electromagnetic equations a relationship can be derived between the optical constants n and k, the dielectric coefficients ε_1 and ε_2, and the electrical conductivity σ. The relative dielectric constant ε_r is related to n and k by the equation,

$$\varepsilon_r = \varepsilon_1 + i\varepsilon_2 = n^2 - k^2 + i2nk \tag{3.57}$$

and the expression for ε_2 in terms of conductivity and frequency is, in SI units,

$$\varepsilon_2 + \frac{\sigma(\omega)}{\omega\varepsilon_0} = \frac{\sigma(\nu)}{2\pi\nu\varepsilon_0}, \tag{3.58}$$

where ν is the frequency of the electric field vector in the incident electromagnetic wave and ω is the angular frequency, $\omega = 2\pi\nu$.

In a metal the absorption ε_2 is much higher than the polarization ε_1, so that

$$\varepsilon_r \approx i\varepsilon_2, \tag{3.59}$$

and the refractive index is related to the dielectric constants by

$$n^2 = \frac{\varepsilon_1}{2} + \frac{\sqrt{\varepsilon_1^2 + \varepsilon_2^2}}{2} \tag{3.60}$$

$$= \frac{\varepsilon_1}{2} + \frac{\sqrt{\varepsilon_1^2 + (\sigma^2/\omega^2\varepsilon_0^2)}}{2}. \tag{3.61}$$

Since $\sigma^2/\omega^2\varepsilon_0^2 \gg \varepsilon_1^2$, this gives

$$n^2 \cong \frac{\sigma}{2\omega\varepsilon_0}. \tag{3.62}$$

Furthermore, since $\varepsilon_2 = 2nk = \sigma/\omega\varepsilon_0$ it is clear that we must also have

$$k^2 \approx \frac{\sigma}{2\omega\varepsilon_0} = n^2 \tag{3.63}$$

At longer wavelengths, and hence lower frequencies, ν, the high conductivity σ leads to values of both k and n which are much greater than unity, and are typically of the order of 10^4.

Now, considering the expression for the reflectance

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2} \tag{3.64}$$

$$= 1 - \frac{4n}{(n+1)^2 + k^2}, \tag{3.65}$$

and since $n \approx k \gg 1$ this leads to

$$R \approx 1 - \frac{2}{n} \tag{3.66}$$

$$R \approx 1 - 2\sqrt{4\pi\varepsilon_0\nu/\sigma}, \tag{3.67}$$

$$= 1 - 2\sqrt{2\omega\varepsilon_0/\sigma} \tag{3.68}$$

which is the Hagen–Rubens relation [5], showing that high-conductivity materials have high reflectance at long wavelengths. This provides the physical justification for our observation at the outset in Chapter 1 that good electrical conductors are also good optical reflectors.

3.4.4 Extensions of classical electron theory to optical properties at high frequencies

How can the higher-frequency absorption bands be explained by the model?

At higher frequencies it is known from experimental observation that the reflectance does not necessarily remain low, but can show some localized peaks. These can be explained by an extension of the classical free electron theory due to Lorentz, in which some electrons behave as classical bound oscillators. These electrons are more tightly bound to the atoms and can therefore only respond to higher-energy excitations. This leads to the following equation of motion for the electrons,

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+\gamma\frac{\mathrm{d}x}{\mathrm{d}t}+kx=e\xi_0\exp(\mathrm{i}\omega t), \tag{3.69}$$

where now the additional term kx represents a binding force between electrons and ionic sites. This equation describes the motion of these bound oscillators and gives absorption at higher frequencies in the form of bound oscillator resonances.

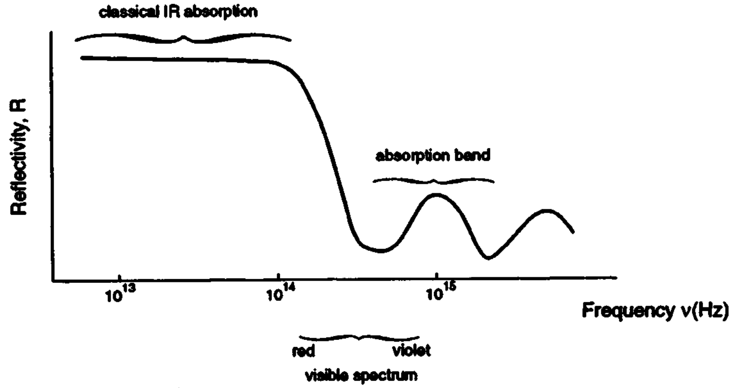


Figure 3.7 Optical reflectance of metals beyond the infrared range, in which 'resonances' at higher energies are observed. These can be attributed to bound oscillators rather than free electrons. Reproduced with permission from R. E. Hummel, *Electronic Properties Materials*, 2nd edn, published by Springer Verlag, 1993.

3.4.5 The photoelectric effect

What happens to the conduction electrons when high-energy light impinges on certain metals?

It was shown by Hertz that a metallic surface emits electrons when illuminated by light of a very short wavelength. The emission of electrons from the surface is dependent on the wavelength of the light and not on the total energy incident on the surface.

The emission of electrons does not occur when the surface is irradiated with longer wavelength light over a longer time period, if the wavelength is below a certain critical value. In other words, if the frequency of the incident light is below a certain threshold value, exposure for longer periods will not lead to the emission of electrons, even though the total energy absorbed by the surface can be increased indefinitely in this way.

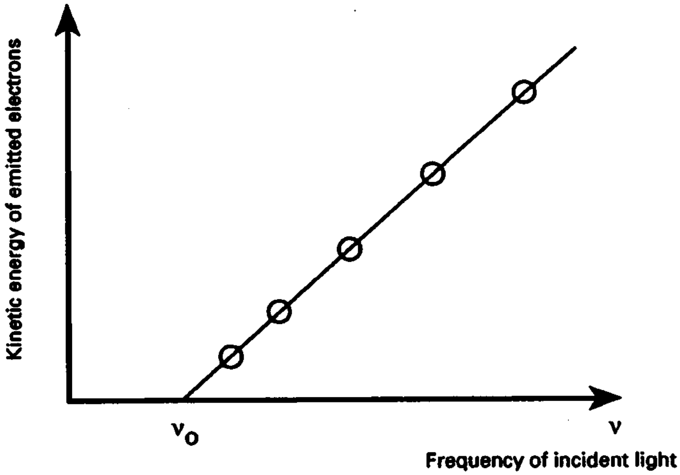


Figure 3.8 Kinetic energy of emitted electrons in the photoelectric effect as a function of frequency of incident light.

Furthermore, the kinetic energy E_K of the emitted electrons is dependent on the frequency of the light, but not on the intensity of the light:

$$E_K = \text{constant}(\nu - \nu_0), \tag{3.70}$$

where ν is the frequency of incident light and ν_0 is the threshold frequency which just enables electrons to escape from the material.

Table 3.2 Values of the work function and threshold frequencies for the photoelectric effect in various materials.

Material	**Threshold energy or work function ϕ (eV)**	**Threshold frequency ν_0 ($10^{14}\,s^{-1}$)**
Caesium	1.91	4.62
Rubidium	2.17	5.25
Potassium	2.24	5.42
Lithium	2.28	5.51
Sodium	2.46	5.95
Zinc	3.57	8.63
Copper	4.16	10.06
Tungsten	4.54	10.98
Silver	4.74	11.46
Platinum	6.30	15.23

An explanation of these observations was given by Einstein [6]. If ω is the angular frequency of the incident radiation and $\hbar$ is Planck's constant the energy of an incident light photon is

$$E(\omega) = \hbar\omega. \tag{3.71}$$

Now, considering the electrons as classical particles trapped in a finite square-well potential of height ϕ, and assuming one light quantum interacts with one electron, the energy imparted contributes to the energy needed to overcome the binding energy E_B of the electron to the solid ϕ and to the final kinetic energy of the electron,

$$\hbar\omega = E_K + E_B \tag{3.72}$$

$$= \tfrac{1}{2}mv^2 + \phi \tag{3.73}$$

$$= \tfrac{1}{2}mv^2 + \hbar\omega_0, \tag{3.74}$$

where ϕ is the symbol used for the work function of the metal. This is identical to the threshold energy needed to liberate the electrons. Rearranging the equation gives,

$$E_K = h\nu = h\nu_0, \tag{3.75}$$

which agrees with experimental observations. This model of the photoelectric effect we may call semiclassical. It relies on the quantum nature of the incident light, but still treats the electrons as classical particles in a finite potential well. The work function is the energy needed to extract one electron from the box, or alternatively is the depth of the potential well.

3.5 CONCLUSIONS

In the final analysis, what are the advantages and disadvantages of the free electron model?

The classical free electron theory provides a simple model of the behaviour of electrons within a solid which seems to work reasonably well in certain cases for metals, for example the alkali metals, but not for insulators. The model takes no account of the properties of the lattice and conceptually this is a shortcoming. Nevertheless, there are a number of physical properties of solids that are dominated by the electrons rather than the lattice, such as electrical and thermal conductivities.

The greatest success of the model was its ability to predict the relation between electrical and thermal conductivity known as the Wiedemann–Franz law. It also allows for a phenomenological explanation for electrical resistance and Ohm's law can be derived on this basis. Even the Hall effect can be described in the alkali and noble metals using this model. When combined with the quantum theory of light it gives an explanation of the photoelectric effect. The Drude model is also able to account for the optical properties of metals in the infrared range and when combined with the Lorentz theory of bound oscillators it can account for the optical properties at higher frequencies.

However, despite all these successes the simple Drude model has several shortcomings. The greatest of these arises in the calculation of the electronic specific heat capacity, and also its prediction of the optical properties of metals in the visible and ultraviolet ranges of the spectrum. It is unable to account for the temperature dependence of the dc conductivity (aside from the use of an ad hoc dependence of τ on temperature). Also, it has not accounted for the temperature and field dependence of the Hall coefficient.

Consequently, an improved theory of the electronic structure of materials is needed. This should show first of all that the heat capacity of the electrons is 100 times smaller than is predicted by the classical electron theory. In order to address this problem it is quite clear that a simple modification of the classical particle model, such as replacing the 'empty box' of the solid by a periodic lattice with Coulomb-type interactions, will not be sufficient. Even though such an approach would, by its very nature, bring the model closer to reality, it would be unable to resolve the critical problem of the electronic heat capacity. Therefore a more radical approach is needed. This can be achieved by the inclusion of a quantum-mechanical description of the free electrons.

REFERENCES

1. P. Drude (1900) *Ann. der Physik* **1**, 566.
2. P. Drude (1900) *Ann. der Physik* **3**, 369.
3. H. A. Lorentz (1905) *Proc. Amst. Acad.* **7**, 438.
4. H. A. Lorentz (1909) *The Theory of Electrons*, Leipzig.
5. E. Hagen and H. Rubens (1903) *Ann. der Physik* **11**, 873.
6. A. Einstein (1905) *Ann. der Physik* **17**, 132.

FURTHER READING

J. S. Dugdale (1977) *The Electrical Properties of Metals and Alloys*, Edward Arnold, London, Chapter 2.

R. G. Chambers (1990) *Electrons in Metals and Semiconductors*, Chapman and Hall, London, Chapter 1.

B. K. Tanner (1995) *Introduction to the Physics of Electrons in Solids*, Chapman and Hall, London.

EXERCISES

Exercise 3.1 Drude free electron theory of metals

Outline the basic assumptions of the classical (Drude) free electron theory of metals. Describe the extension by Lorentz of the original theory to include the effects of bound electrons and describe how these were represented in the model. Calculate the expected instantaneous velocity of electrons at 300 K.

Exercise 3.2 Reflectivity based on Drude theory

State the relationship between the dielectric constants of a material and the Drude parameters ν_1 and ν_2 from the classical free electron model. Using the relationship between the reflectance R and ε_1 and ε_2 show the limiting values of R as $\nu \rightarrow 0$ and $\nu \rightarrow \infty$. Under what frequency conditions does the Drude prediction of R work and under what conditions is it inadequate?

Exercise 3.3 Electrical and optical properties of a classical free electron metal

Write a short discussion of the principal successes and failures of the classical free electron theory. The mobility of electrons is defined as the ratio of electric field to velocity ($\mu = \nu/\xi$). In a piece of copper the mobility was found to be $3.5 \times 10^{-9}\ \mathrm{m^2\,V^{-1}\,s^{-1}}$. Assuming that the classical free electron model can be used calculate the resistive coefficient γ, the electrical conductivity σ and the mean free time between collisions. (Copper has a density of $8940\ \mathrm{kg\,m^{-3}}$, an atomic weight of 63 and each atom donates one conduction electron).

Exercise 3.4 Classical free electron description of resistivity

The density of copper is $8.95 \times 10^3\ \mathrm{kg\,m^{-3}}$. Calculate the number of free electrons per cubic metre and hence their drift velocity when a current is flowing in the metal with density $10\ \mathrm{kA\,m^{-2}}$. When 1 atomic per cent of a monovalent impurity is added to copper, the mean free time between collisions of electrons with the impurities is 5×10^{-14} s. Calculate the resistivity of the impure metal.

Exercise 3.5 Mobility of classical free electrons

From the data given below calculate the mobility of an electron in each metal at room temperature.

Element	**Atomic wt.**	**Resistivity (Ω m)**	**Valence electrons (per atom)**	**Density at 20°C ($\mathrm{kg\,m^{-3}}$)**
Copper	63	1.8×10^{-8}	1	8950
Silver	108	1.6×10^{-8}	1	10490
Gold	197	2.4×10^{-8}	1	19302
Cadmium	112	7.5×10^{-8}	2	8650
Zinc	65	6.0×10^{-8}	2	7130
Aluminium	27	2.7×10^{-8}	3	2700

Describe the changes in a metal that result from plastic deformation and suggest what changes you would expect in the electric properties of copper as a result of cold working. How could you restore the electrical properties of the metal to their original value after cold working?

Exercise 3.6 Absorption of light by a metal

Calculate the attenuation coefficient, alpha, in reciprocal metres of (a) a metal foil 50-nm thick which reflects 40% and transmits 20% of the light incident upon it, and (b) a glass which absorbs 90% of the incident light in a thickness of 0.2 m. If the energy of the incident light is 1.5 eV calculate the extinction coefficient *k*.

Explain how the real and imaginary components of the dielectric coefficient, known as the polarization and absorption, respectively, are related to the refractive index and extinction coefficient via the free electron model. Calculate the absorption and polarization for the metal if the refractive index is $n = 0.5$.

4 CONDUCTION ELECTRONS IN MATERIALS - QUANTUM CORRECTIONS

OBJECTIVE

In this chapter we look at another 'free' electron model in which the electrons are described by wavefunctions contained within a material boundary. Once again the model assumes that the potential inside the material is completely flat, which amounts to ignoring the presence of the lattice. The electrons are therefore only constrained by the limits of the material. The main difference between this and the classical free particle model is that only certain energy levels are allowed. This means that with constraints on the numbers of electrons that can occupy these energy levels, only those electrons with energies close to the top of the electron 'sea' can contribute to the heat capacity. This results in a reduction by two orders of magnitude of the expected electronic contribution to the heat capacity, bringing it into agreement with measured values. However, ultimately the model is insufficient because of its failure to take the lattice potential into account.

4.1 ELECTRONIC CONTRIBUTION TO SPECIFIC HEAT

How can the free electron model be developed further to allow prediction of a wider range of properties?

In view of the apparent failures of the classical free electron model of metals, we must look at the next level of sophistication to see whether an improved theory can be derived. The next procedure is to try a quantum-mechanical approach to the free electron model. The most critical problem in the Drude theory was the contribution to the heat capacity from the electrons. The theory predicted a contribution of $\frac{3}{2}k_B$ per electron, whereas in practice this contribution was known to be much smaller. The classically expected electronic contribution to the specific heat is

$$C_V^e = \tfrac{3}{2}N_0 k_B. \tag{4.1}$$

With $N_0 = 6.02 \times 10^{26}$ atoms per kg mole, and assuming one conduction electron per atom,

$$C_V^e = 12.5\,\mathrm{J\,mol^{-1}\,K^{-1}}. \tag{4.2}$$

The observed value is typically,

$$C_V^e = 0.2\,\mathrm{J\,mol^{-1}\,K^{-1}}. \tag{4.3}$$

The discrepancy arises from the treatment of the conduction electrons as a classical free gas. The fact that $N_f = 10^{28}\,\mathrm{m^{-3}}$, which is 1000 times greater than for a classical free electron gas, and also the fact that the particles are not electrically neutral (and therefore interact through Coulomb repulsion) should immediately have caused suspicion about the adoption of the classical kinetic theory for the model. It was found that the whole concept of classical statistical thermodynamics was inapplicable to this situation and that a new form of statistics was needed to describe electrons in solids.

4.2 WAVE EQUATION FOR FREE ELECTRONS

What is the equation of motion of the electrons in this case?
If quantum-mechanical principles are used to describe the electrons the general expression for the time-independent Schrödinger equation [1] is,

$$-\frac{\hbar^2}{2m}\nabla^2\psi(x) + V(x)\psi(x) = E\psi(x), \tag{4.4}$$

where, in our case, m is the mass of an electron, $\hbar$ is Planck's constant divided by 2π and $V(x)$ is the potential energy at the point in space defined by the vector x. E is the energy of the electron, which we may consider as a wave or particle. Note that the Schrödinger equation here is simply an expression of the conservation of energy. The sum of kinetic energy and potential energy on the left-hand side equals the total energy on the right-hand side.

We can find solutions of the wave equation in particular cases. Generally, the solution has the form

$$\psi(x) = A_1 \cos kx + A_2 \sin kx, \tag{4.5}$$

in one dimension, where x is the spatial dimension, and k is the wave vector which may be real or imaginary. The probability of finding the electron at any given point x in space is

$$P(x) = \psi^*(x)\psi(x) = |\psi(x)|^2, \tag{4.6}$$

where ψ^* is the complex conjugate of the wavefunction ψ.

The intention in this book is to solve the Schrödinger equation in one dimension only. This gives the greatest clarity of explanation by presenting the essential concepts behind the quantum-mechanical free electron model. Nothing new is learned by generalizing to three dimensions at this stage, other than additional detail.

4.2.1 Consequences of the quantum theory of free electrons

How does the energy of a free electron depend on its wave vector?
It is found, by substituting the solution of the wavefunction into the Schrödinger equation, that when the electrons are completely free, that is providing that $V(x)$ is everywhere zero, the energy E and the wave vector k are related by

$$k = \sqrt{\frac{2mE}{\hbar^2}}. \tag{4.7}$$

We can rearrange this so that,

$$E = \frac{\hbar^2 k^2}{2m}. \tag{4.8}$$

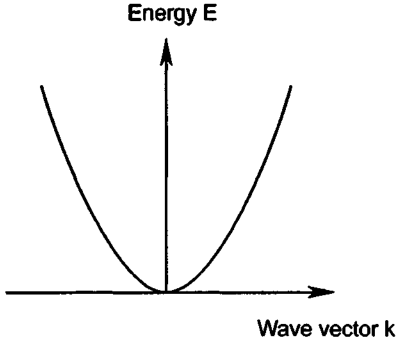


Figure 4.1 Free electron parabola showing the dependence of energy *E* on wave vector *k* in which all energies are possible.

The relationship between energy E and wave vector k is depicted by the free electron parabola shown in Fig. 4.1. This applies only to completely free electrons.

The important point here is that since the electrons are completely free they can be considered to be in the presence of a completely flat potential, $V = 0$ over all space. Therefore, the electron wavefunction ψ can extend to infinity and there are no boundary conditions to be applied to the wavefunction.

With no further constraints to be applied, we see that all values of k are allowed solutions of the Schrödinger wave equation, and therefore all energies are allowed for free electrons. Stated in alternative words, the allowed energy values E form a continuum for free electrons.

If we extend the above one-dimensional model to three dimensions the components of the wave vector along the x, y, and z directions are k_x, k_y, and k_z. Consequently the energy of a free electron, with wave vector components in three dimensions of k_x, k_y, and k_z, is

$$E = \frac{\hbar^2}{2m}\left(k_x^2 + k_y^2 + k_z^2\right). \tag{4.9}$$

This means that if the electrons are completely free all electron states with the same energy form a sphere in k-space. Therefore, if we have free electrons occupying all energy levels up to a maximum energy E_F, the electrons are all contained in a sphere in k-space. We shall return to this idea later in discussing the Fermi surface for free electrons.

4.3 BOUNDARY CONDITIONS: THE SOMMERFELD MODEL

How can a quantum-mechanical description of the electrons be used while maintaining most of the classical description from the previous model?

The first application of quantum mechanics to the problem of the electronic properties of metals was made by Sommerfeld [2,3]. The initial motivation was to resolve some of the discrepancies between the classical Drude free electron model and experimental observation, in particular those problems which arose over the electronic contribution to the specific heat.

The Sommerfeld model is still very much a free electron model. It assumes that the conduction electrons reside within a potential which is everywhere constant inside the metal. This of course is only an approximation, but it is a useful first step because it leads to a relatively simple formulation of the problem.

Sommerfeld applied the Pauli exclusion principle [4] (see Section 4.4.1) to the free electron model of metals. This resulted in the resolution of the most serious anomalies in the classical electron theory of metals. In particular, the difficulty over the electronic contribution to the specific heat capacity was explained.

In its simplest form, the Sommerfeld model involves only this single modification to Drude theory. Later we shall look at further modifications such as the solution of the Schrödinger equation with more realistic assumptions such as periodic potentials. For the time being, however, we will just look at the solid as a simple, flat-bottomed potential box containing electrons. In this the solid is represented as a square-well potential. In the most elementary calculation an infinite square well is assumed, but a finite square-well potential also gives relatively simple solutions.

We will find that there are only certain allowed or accessible energy states for the electrons under these conditions. Note here that it is the boundary conditions that are crucial in determining the allowed energy states, not just the wave equation itself.

4.3.1 Wave equation for bound electrons in an infinite square-well potential

What happens to the allowed electronic states, that is to say solutions of the wave equation, if boundary conditions are imposed?

Let us now go to the other extreme and consider bound electrons. This would seem to be a quite reasonable approach to the problem of electrons in a solid since mostly the electrons are constrained to remain within the solid. The simplest possible model then is to suppose that the electrons move freely within the confines of the volume of the solid, but encounter an infinite potential at the boundary of the solid which prevents them from leaving the solid. We may represent this under the simplest conditions in one dimension as an infinite square-well potential.

The flat potential inside, which is everywhere constant, may seem like a gross approximation to a solid, but remember that most of the solid is 'empty space' anyway and the outer electrons are screened from the ionic cores by the localized electrons. If we allow the potential outside the range $-a \leq x \leq a$ to go to infinity, then the probability of the electron appearing in the ranges $x < -a$ or $x > a$ is zero. This means that outside the solid

$$\psi^*(x)\psi(x) = 0. \tag{4.10}$$

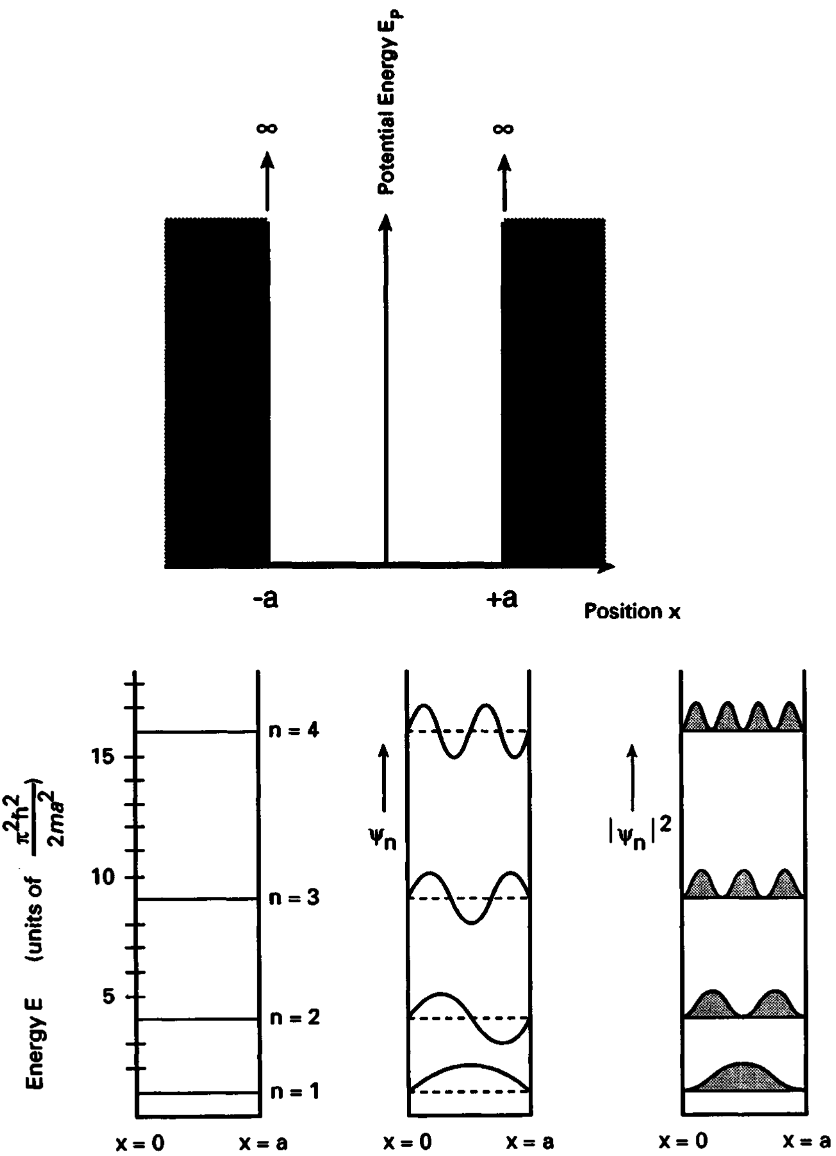


Figure 4.2 Infinite square-well potential which will be used to represent a material in which electrons are constrained only by the physical limits of the material.

Our boundary conditions therefore must be

$$\psi(x = a) = 0 \tag{4.11}$$

$$\psi(x = -a) = 0. \tag{4.12}$$

If we then solve the one-dimensional wave equation in the time-independent case,

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + V(x)\psi = E\psi \tag{4.13}$$

we again obtain the solutions

$$\psi(x) = A_1 \cos kx + A_2 \sin kx. \tag{4.14}$$

The difference from the previous example of the completely free electron is that now we need to apply the known boundary conditions. It is these boundary conditions which determine the allowed solutions.

At $x = a$, $\psi(a) = 0$, and so

$$\psi(a) = A_1 \cos ka + A_2 \sin ka = 0. \tag{4.15}$$

At $x = -a$, $\psi(-a) = 0$ and so

$$\psi(-a) = A_1 \cos(-ka) + A_2 \sin(-ka) = 0, \tag{4.16}$$

which leads to the solution, $A_1 = 0$ and

$$ka = n\pi, \tag{4.17}$$

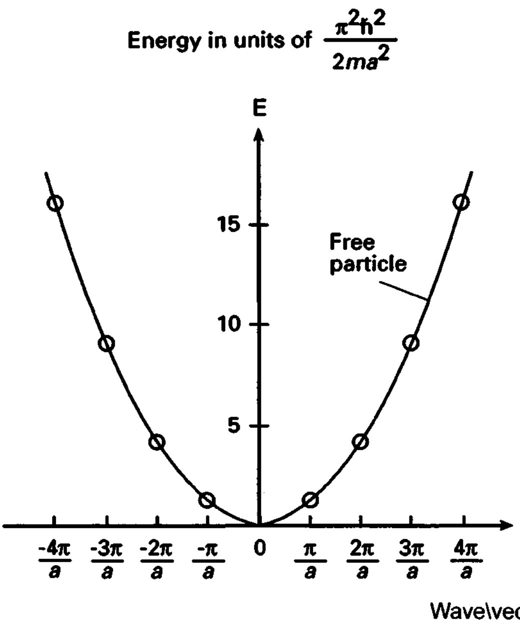


Figure 4.3 The dependence of energy on wave vector for electrons confined in an infinite square-well potential. The parabolic relation still holds, but only certain discrete energy levels are allowed.

so that

$$\psi_n(x) = A_2 \sin\left(\frac{n\pi}{a}x\right). \tag{4.18}$$

This is the solution of the wave equation under the given boundary conditions. Note that only integer values of n satisfy the boundary conditions and so the allowed energy levels shown in Fig. 4.3 are now discrete:

$$E_n = \frac{\hbar^2}{2m}\frac{n^2\pi^2}{a^2} \qquad n = 0, 1, 2, 3\dots\ . \tag{4.19}$$

This means that only wavefunctions with certain isolated energies can fit into the solid and still meet the boundary conditions. The extension of this model to three dimensions is simple and introduces no new concepts.

4.3.2 Wave equation for electrons in a finite potential well

What are the allowed electronic states, i.e. solutions of the wave equation, when the potential box is finite?

The next level of complexity, which brings us closer to reality, is to consider the possible electron states in the presence of a finite potential well. Clearly in a real material the potential barriers marking the end of the solid will be finite. This now brings us close to the idea of the original Drude free electron model which was a box, consisting of a finite potential barrier represented by the boundary of the solid, containing classical particles which we called electrons. Now we will use the same idealized box to represent the solid boundary but instead fill it with wave-like representations of electrons instead of classical particles. The depth of the potential well must, of course, be finite because of the evidence supplied by the photoelectric effect and thermionic emission. In fact these two phenomena can be used to determine the depth V_0 of the potential well, or more correctly the depth of the most energetic electrons below the top of the potential well.

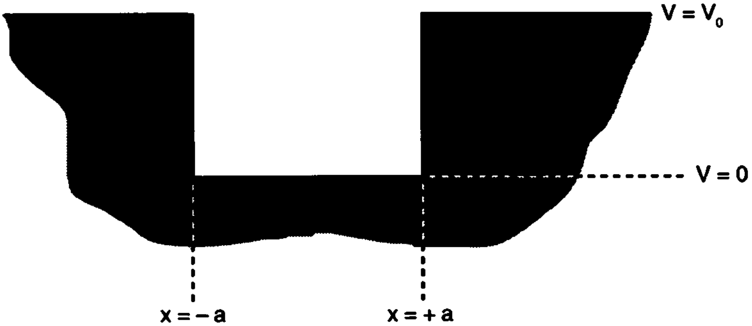


Figure 4.4 Finite square-well potential representation of the energy potential experienced by an electron in a solid.

Again we solve the Schrödinger equation. Inside the solid we assign for simplicity $V = 0$. The wave equation is then

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} = E\psi, \tag{4.20}$$

where E is the energy of the electron.

Solutions of the wave equation for positions inside the box, that is for $|x| < a$, give the following wavefunction

$$\psi = A_1 \cos kx + A_2 \sin kx, \tag{4.21}$$

where $k = (2mE)^{1/2}/\hbar = 2\pi/\lambda$.

For lower energies ($E < V_0$) λ is the wavelength of the standing waves represented by the wavefunction inside the potential box. Note that larger energies correspond to shorter wavelengths and vice versa. For higher energies ($E > V_0$) the wavefunction will extend well beyond the boundaries of the box and therefore does not correspond to a standing wave even in the region of the potential box.

Outside the material, that is for $|x| > a$, we now have a finite potential $V = V_0$, so in this region the wave equation is,

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}(E - V_0)\psi = 0. \tag{4.22}$$

The solutions of the wave equation in this region depend on the energy E. If $E < V_0$ then the second term on the left-hand side of the equation becomes negative and solutions for $|x| > a$ have the form,

$$\psi(x) = B_1 e^{k_1 x} + B_2 e^{-k_1 x}, \tag{4.23}$$

where $k_1^2 = -2m(E - V_0)/\hbar^2$.

These solutions for $E < V_0$ in the region outside the box are not periodic. They represent an exponentially decaying waveform which means that for $E < V_0$ the wavefunction decays with distance beyond the limits of the material.

It is easy to see that since the wavefunction cannot diverge outside the potential well, we must have $B_1 = 0$ for $x > a$ and $B_2 = 0$ for $x < -a$ so we are left with

$$\psi(x) = B_2 e^{-k_1 x} \text{ for } x > a \text{ and } E < V_0 \tag{4.24}$$

$$\psi(x) = B_1 e^{-k_1 x} \text{ for } x < -a \text{ and } E < V_0. \tag{4.25}$$

In order to find the coefficients in those cases with $E < V_0$ we must again use the boundary conditions at $x = -a$ and $x = a$. Since we already have the form of solution above for both inside and outside the box, we merely need to ensure that these match at the boundaries of the box.

We will now proceed using only the antisymmetric component of the wavefunction, $\sin kx$. A similar argument can be applied to the symmetric wavefunction, cos kx. The amplitude of ψ must be continuous across the boundary, so that, for the odd parity (sine) solutions

at $x = a$,

$$A_2 \sin ka = B_2 \mathrm{e}^{-k_1 a}, \tag{4.26}$$

at $x = -a$,

$$A_2 \sin(-ka) = B_1 \mathrm{e}^{-k_1 a}, \tag{4.27}$$

and the derivative $\mathrm{d}\psi/\mathrm{d}x$ must also be continuous across the boundary

at $x = a$,

$$A_2 k \cos ka = -k_1 B_2 \mathrm{e}^{-k_1 a}, \tag{4.28}$$

at $x = -a$,

$$-A_2 k \cos(-ka) = k_1 B_1 \mathrm{e}^{-k_1 a}. \tag{4.29}$$

We therefore simply equate the solutions both inside and outside the box at the boundaries and these lead to the condition,

$$k \cot(ka) = -k_1. \tag{4.30}$$

Similarly, for the even parity (symmetric or cosine) solutions it can be shown by the above argument that,

$$k \tan(ka) = k_1. \tag{4.31}$$

These last two equations may be solved graphically or numerically for the allowed energies E_n, remembering that $E = \hbar^2 k^2/2m$.

The important result here is that for energies $E < V_0$, which represent electrons 'contained' within the potential well, the allowed energies lead to a discrete set of energy levels. We say 'contained' because it is also clear from the wavefunction that the electron has a finite probability of being just outside the potential well, that is $\psi(x) \neq 0$ for $|x| > a$ (or more precisely $|\psi(x)|^2 \neq 0$ in this range, since the probability function is dependent on $|\psi|^2$ rather than ψ).

This is a new result emerging from the quantum-mechanical treatment of the problem, that does not arise in the classical description of the electrons in materials.

For energies $E > V_0$ the electron wavefunctions extend outside the limits of the potential well:

$$\psi(x) = B_1 \mathrm{e}^{k_1 x} + B_2 \mathrm{e}^{-k_1 x}, \tag{4.32}$$

where $k_1^2 = -2m(E - V_0)/\hbar^2$, which gives an imaginary value of k_1, and hence the solutions are periodic for $E > V_0$.

The electrons with energies greater than V_0, the depth of the potential, are not constrained by the presence of the potential and can therefore have a continuous spectrum of energies and extend spatially beyond the box. We should note, however, that the wavefunctions of these higher-energy electrons in the locations $|x| < a$ are perturbed by the presence of the potential well.

4.3.3 The harmonic oscillator

Are there other examples where there is a particularly simple relationship between energy and frequency of an electron?
The available energy states of a particle such as an electron or atom in a harmonic (i.e. parabolic) potential can also be determined in a similar way. We have already seen that to a first approximation atoms in a lattice experience a potential of the form $E_p = ku^2$, and have an oscillation frequency given by $\omega_0 = \sqrt{2k/m}$ as shown in Section 2.2. Using the quantum-mechanical description, the allowable energy states of a particle in a parabolic potential are found to be simple multiples of the ground state [5],

$$E_n = (n + \tfrac{1}{2})\hbar\omega_0. \tag{4.33}$$

Therefore, a discrete lattice of particles interacting via a harmonic potential can be considered as a collection of oscillators, each able to absorb or emit energy at frequencies that are linear multiples of $\hbar\omega_0$, and similarly an electron in a parabolic potential will have allowed energy states that are linear multiples of the ground frequency. Compare this result with eqn (4.19).

4.4 DISTRIBUTION OF ELECTRONS AMONG ALLOWED ENERGY LEVELS

How does quantum mechanics affect the distribution of electron energies?
Unlike classical statistical mechanics, which allows any energy state to be occupied by any number of electrons, quantum mechanics imposes restrictions on the number of electrons which can occupy a given energy level.

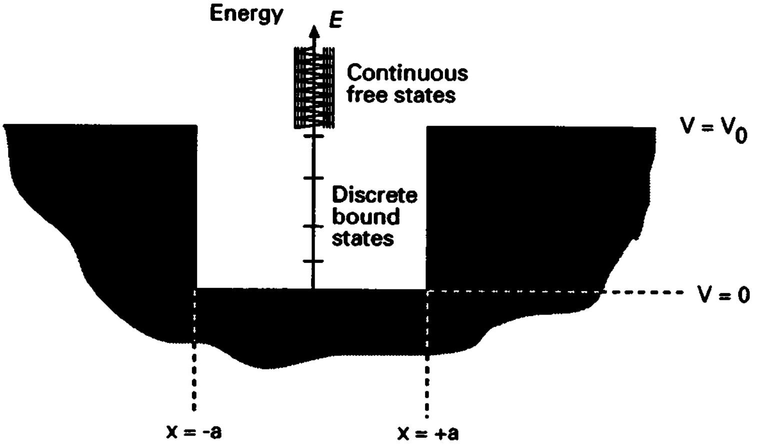


Figure 4.5 Allowed energy states in a finite square-well potential obtained by solution of the Schrödinger equation.

4.4.1 The Pauli exclusion principle

How many electrons can we fit into a given energy state?

The Pauli principle states that no two electrons can have the same set of quantum numbers, and therefore cannot occupy identically the same energy level in our solid. We are ignoring electron spin for the moment. This leads to a radical change in our understanding of the ground energy state of a solid containing a number of electrons, since no longer can all of the electrons reach the nominal lowest energy level that they would be allowed to occupy in classical particle physics.

If we have a certain number N electrons in the solid, then we must begin filling the lowest energy level with an electron. When this is filled we move up to the next energy level and so on. We do this until each electron has been assigned to the lowest remaining energy state available. This then is the ground state of our solid and corresponds to the electronic occupancy at absolute zero temperature.

If we proceed in this way until all electrons have been assigned to an available energy state, there will exist a highest occupied energy level. This is known as the Fermi level. It separates the occupied from the unoccupied states only at 0 K (i.e. in the ground state of the solid). When we consider electrons occupying available energy states, we can define a function $f(E)$ which describes the probability that a given energy state is occupied.

Clearly because of the constraints of a limited discrete set of allowed energies, and the limitation of one electron per energy state (assuming we treat the spin-up and spin-down as distinct states), the probability of the lowest-energy states being occupied is 1 and providing we have enough electrons available, then, at least in the ground state 0 K, the probability of any state being occupied remains 1 until we literally run out of electrons to occupy the allowed states. The probability of occupancy $f(E)$ at 0 K therefore must have the form, shown in Fig. 4.6, where E_F is the highest occupied level, which is called the Fermi level.

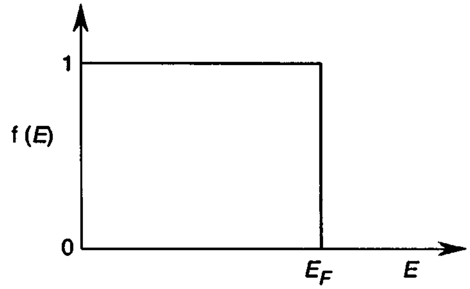


Figure 4.6 Probability of occupancy of energy states in the ground ($T = 0$ K) state. All levels up to E_F are occupied; above that the energy levels are unoccupied.

4.4.2 The failure of classical Maxwell–Boltzmann statistics

Can Maxwell–Boltzmann statistics apply once we have restricted the number of allowed states for an electron to occupy?

In classical statistical mechanics, all energies are available and any number of particles can have an identical energy E, or equivalently have the same state. Under these conditions, the probability of a particle having an energy E at temperature T is,

$$P(E) = P_0 \exp\left(-\frac{E}{k_B T}\right). \tag{4.34}$$

The normalization constant P_0 is found simply from the condition that the integral of the probabilities over all possible states must be unity:

$$\int_{\text{all } E} P(E) \mathrm{d}E = 1. \tag{4.35}$$

The Maxwell–Boltzmann distribution function has the following form, illustrated in Fig. 4.7.

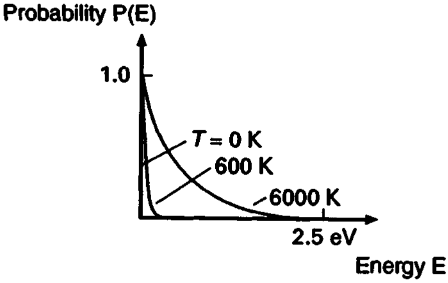


Figure 4.7 The Maxwell–Boltzmann distribution function which represents the probability of a classical particle being found with energy E.

Once we have accepted the idea of a finite number of discrete energy levels and applied the Pauli exclusion principle which limits occupancy of these available states, it is immediately apparent that classical Maxwell–Boltzmann statistics can no longer apply, since classical statistics is based on the concept that any energy can be taken by any number of particles.

4.4.3 Probability of occupancy

With a finite number of allowed electron states in a material, how do the electrons arrange themselves?

Since the electrons can no longer follow classical continuum statistics a new description of the energy spectrum of electrons is needed. The Fermi function $f(E)$ which describes the probability of occupancy of electrons as a function of energy E [6, 7] has the form

$$f(E) = \frac{1}{1 + \exp((E - E_F)/k_B T)}. \tag{4.36}$$

When $E = E_F$ it can easily be seen that $f(E) = 0.5$, and this corresponds to the Fermi energy, the highest occupied energy level when the electron configuration is in its ground state.

At temperatures above 0 K the step function form of this probability distribution becomes distorted with a range of intermediate energies close to E_F for which $f(E)$ is neither 1 or 0. The reason for this is that the thermal energy can excite electrons from some lower-energy states to higher-energy states. Therefore, some lower-energy states which were occupied at 0 K are now unoccupied and some higher-energy states that were unoccupied at 0 K are now occupied.

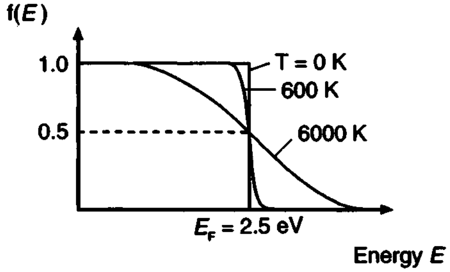


Figure 4.8 The Fermi–Dirac function $f(E)$, which gives the probability of occupancy, varies with temperature. At absolute zero temperature only the lowest energy states up to the Fermi level are occupied. At higher temperatures energy states above the Fermi level can be occupied.

4.4.4 Fermi–Dirac statistics

What statistics do the electrons obey?

When Drude was developing the classical free electron model of a metal, it was natural to assume that the electron energy distribution was like that of a classical gas and obeyed Maxwell–Boltzmann statistics because no other statistics for describing assemblies of particles was available at the time. This gives the number of electrons per unit volume with velocities in the range $v \pm dv$ as

$$f_B(v) = N\left(\frac{m}{2\pi k_B T}\right)^{3/2} \exp\left(-\frac{mv^2}{2k_B T}\right). \tag{4.37}$$

However, since electrons need to obey the Pauli exclusion principle the distribution of electrons with velocities in the range of $v \pm dv$ must be described by Fermi–Dirac rather than Maxwell–Boltzmann statistics. In this case the expression according to Fermi–Dirac statistics becomes,

$$f(v) = \frac{1}{4}\frac{m^3}{h^3}\frac{1}{\exp((mv^2/2 - k_B T_0)/k_B T) + 1}. \tag{4.38}$$

4.4.5 Electron energy distributions

What other information is needed to describe the electron distribution?

We can describe the number of available electron states as a function of energy E by the density of available states $D(E)$. This density of states is independent of the available electrons to fill the states; it is simply an expression of what energy values are allowed.

The occupational density of states $N(E)$ describes the number of electron states per unit energy interval as a function of energy. This is related to the density of available states $D(E)$ through the probability of occupancy $f(E)$ by the equation

$$N(E) = 2f(E)D(E), \tag{4.39}$$

where the factor 2 arises because electrons can have spin-up and spin-down [8, 9], and therefore each available energy state can be occupied by two electrons, one with spin-up the other with spin-down. This simple expression allows the electron distribution to be described in terms of the available levels $D(E)$ as determined by the ionic potential, and the distribution of electrons among these levels $f(E)$ as determined by temperature, for example.

4.4.6 Density of states

How does the number of allowed energy states vary with energy level?

We have talked at length about the allowed energy levels, how they are discrete states at low levels near to the atomic cores, and how they form almost continuous bands of allowed energies at higher levels. We have also shown that the electrons occupy these states beginning at the lowest energy level and working upwards to the Fermi level. In doing so they obey the Pauli exclusion principle. So far, however, we have made no mention of exactly how many allowed levels $D(E)$ there are for electrons at any given energy level or how this can be determined. We must, therefore, address this question of an equation for the density of states, at least in some simplified cases.

4.4.7 Model density of states in square-well potential

How does the density of states vary with energy in the simplest case of a three-dimensional square-well potential?

Returning to our familiar model, the square-well potential, we can investigate mathematically the density of available states. If we generalize the earlier result for the infinite square-well potential (Section 4.3.1) to three dimensions, the energy of a given state is

$$E = \hbar^2k^2/2m \tag{4.40}$$

$$E_n = \frac{\hbar^2\pi^2}{2ma^2}(n_x^2 + n_y^2 + n_z^2), \tag{4.41}$$

where n_x, n_y, n_z are quantum numbers representing allowed solutions of the wave equation, such that

$$n^2 = n_x^2 + n_y^2 + n_z^2. \tag{4.42}$$

A given energy state, which represents a particular combination of n_x, n_y, n_z, can be represented as a point in quantum number space. All states with equal energy will lie on the surface of a sphere. Since only positive value of n_x, n_y and n_z are allowed, the available states are restricted to $\frac{1}{8}$ of the volume of a sphere in quantum number space.

The number of distinct allowed states $N_0(E_n)$ with energies equal to or less than an energy E_n is

$$N_0(E_n) = \int_0^{E_n} D(E)\mathrm{d}E. \tag{4.43}$$

This is equal to the total volume of quantum number space divided by the volume occupied by one quantum state. Since each quantum state occupies unit volume in quantum number space this gives,

$$N_0(E_n) = \frac{1}{8}\frac{4}{3}\pi n^3. \tag{4.44}$$

Using the free electron approximation, and remembering that there can be two electrons per energy state,

$$2N_0(E_n) = \frac{\pi}{3}\left(\frac{2ma^2}{\pi^2\hbar^2}\right)^{3/2} E^{3/2}. \tag{4.45}$$

We will use the $N_0(E)$ notation to remind us that it represents the number of states from $E = 0$ to $E = E$. If we differentiate this expression, we obtain the number of energy states between E and $E + \mathrm{d}E$, which we will denote $D(E)$

$$D(E) = \frac{\mathrm{d}}{\mathrm{d}E}(N_0(E)) = \frac{\pi}{4}\left(\frac{2ma^2}{\pi^2\hbar^2}\right)^{3/2} E^{1/2}, \tag{4.46}$$

and putting $V = a^3$, the volume of a cube of material of side a, gives the density of states

$$D(E) = \frac{V}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2} = \frac{V}{4\pi^2}\frac{2mk}{\hbar^2} \tag{4.47}$$

and since we can have two electrons per energy state, one with spin-up and the other with spin-down, the actual number density of electron states is twice $D(E)$.

Remember, this is an approximate expression, which is only valid for free electrons. In a real material the dependence of the density of states on the energy is modified and therefore not truly parabolic.

In order to adequately describe the allowed electron states in a solid we not only need the available energies as determined by the boundary conditions imposed by the lattice but also the density of states at each of the allowed energies. Generally, the number of distinct available states $D(E)$ increases with energy E.

4.5 MATERIALS PROPERTIES PREDICTED BY THE QUANTUM FREE ELECTRON MODEL

Do the predictions of the quantum free electron model correlate with observations? In this section we look at how the quantum free electron model overcomes some of the failures of the classical free electron model. In particular, how it explains the low electronic contribution to the heat capacity, the temperature-independent paramagnetic susceptibility in metals, and thermionic emission.

4.5.1 Heat capacity

How does the electron heat capacity predicted by the quantum free electron model vary from that predicted by the classical free electron model?

We now address directly the main question which proved the downfall of classical Drude theory – the discrepancy between the observed electronic heat capacity and the predicted value based on classical theory. The application of the Pauli principle and quantum mechanics to the problem of electrons in a solid leads to a new distribution of energy states given by Fermi–Dirac statistics.

When a solid is heated from absolute zero temperature not every electron is able to gain an energy $k_B T$ because the energy state which it would need to occupy in order to absorb the energy is either already occupied by another electron, or in some cases does not exist. Only those electrons in energy levels within $k_B T$ of the Fermi level can be thermally excited to higher available energy levels. This leads to a change in the density of occupied states as shown in Fig. 4.9. The result is a blurring of the distinction between occupied and unoccupied states, which at 0 K was a very sharp distinction.

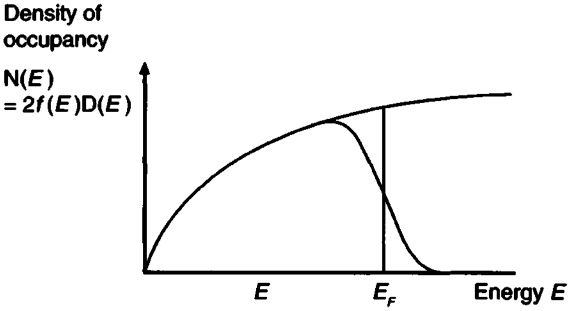


Figure 4.9 Density of states $D(E)$ as a function of energy E from quantum free electron theory. Only those electrons within $k_B T$ of the top of the distribution can absorb thermal energy. This reduces the expected electronic heat capacity compared with classical theory.

The kinetic energy dE_k of the electrons above the Fermi level due to thermal energy is,

$$dE_k = \tfrac{3}{2} k_B T \, dN, \tag{4.48}$$

where dN is simply the number of electrons above the Fermi level. This depends on the density of states at the Fermi level and the thermal energy. Assuming that the density of states close to the Fermi level does not change drastically as a function of energy E we can write,

$$dN = N(E_F) k_B T, \tag{4.49}$$

because as shown in Fig. 4.9 only those electrons within k_BT of the Fermi level can absorb thermal energy and contribute to the specific heat:

$$dE_k = \tfrac{3}{2}k_B^2T^2N(E_F). \tag{4.50}$$

We now immediately have a qualitative solution to the problem of the small electronic contribution to the heat capacity. If the total number of electrons is N then only a fraction of these can be thermally excited at a temperature T. This fraction is typically T/T_F where T_F is the temperature corresponding to the Fermi energy $k_BT_F = E_F$. The electronic heat capacity is then

$$C_V^e = \frac{dE}{dT} \tag{4.51}$$

$$= 3k_B^2TN(E_F). \tag{4.52}$$

For free electrons, the density of states at the Fermi level is given by

$$N(E_F) = \frac{3}{2}\left(\frac{N}{E_F}\right), \tag{4.53}$$

so that

$$C_V^e = \frac{9}{2}Nk_B\frac{T}{T_F}. \tag{4.54}$$

If we take quantum mechanics fully into account then the equation for the heat capacity is slightly modified to

$$C_V^e = \frac{\pi^2}{2}Nk_B\frac{T}{T_F} = \frac{\pi^2}{2}\frac{Nk_B^2T}{E_F}, \tag{4.55}$$

where N is the total number of conduction electrons per unit volume. At room temperature this value of C_V^e is about 1% of the classically expected electronic heat capacity of $(\frac{3}{2})N_0k_B$, because typically $T_F = 30\,000$ K, and hence $T/T_F = 0.01$.

4.5.2 Pauli spin paramagnetism

What predictions does the model make about other bulk properties, such as magnetic susceptibility?

Electrons have a magnetic moment associated with both their spin and orbital angular momentum. All metals exhibit a weak paramagnetism which is independent of temperature and this may be explained with the free electron model of metals developed in this chapter [10]. When an external magnetic field is applied to a material the orientations of the magnetic moments of the electrons are constrained to lie either parallel or antiparallel to the field. This leads to a splitting of the energy of the parallel and antiparallel states.

Those electrons with magnetic moments m parallel to the field direction have energies reduced by $\Delta E = -\mu_0 mH$, while those with magnetic moments antiparallel to the field have energies increased by $\Delta E = \mu_0 mH$. Some of these

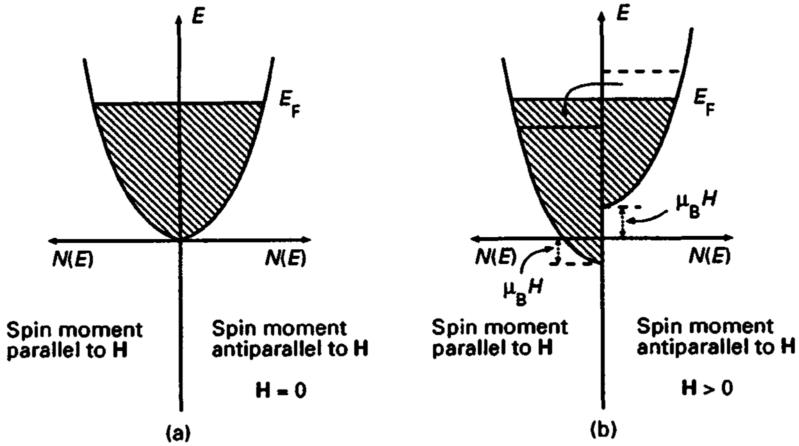


Figure 4.10 Separation of spin-up and spin-down energy states in Pauli paramagnetism.

antiparallel electrons can reduce the energy of the system by occupying parallel states of lower energy.

The number of electrons which can change orientation and still reduce the total energy are those which were within $\mu_0 mH$ of the Fermi level in the absence of the field. If $N(E)$ is the occupancy density of states, then as we have shown in Section 4.4.5,

$$N(E) = 2D(E)f(E), \tag{4.56}$$

where $f(E)$ is the probability of a given state of energy E being occupied, and $D(E)$ is the density of the states (i.e. number of states with the same energy E). Clearly $0 \leq f(E) \leq 1$ where f is the Fermi function.

The Pauli paramagnetic susceptibility $\chi_P = M/H$ is therefore given by

$$\chi_P = 2m\left(\frac{\mathrm{d}N_0(E_F)}{\mathrm{d}E}\right)\frac{\Delta E}{H}, \tag{4.57}$$

and since $\Delta E/H = \mu_0 m$

$$\chi_P = 2D(E_F)\mu_0 m^2, \tag{4.58}$$

where $D(E_F)$ is the density of states at the Fermi level and m is the magnetic moment of an electron.

It is clear that the Pauli paramagnetic susceptibility is dependent entirely on the small fraction of electrons which reside close to the Fermi level. Typically it is found that $\chi_P \approx 10^{-10}$ (dimensionless).

4.5.3 Thermionic emission

What happens when a metal is heated to very high temperatures?

Finally, we will mention in passing the phenomenon of thermionic emission, which again confirms our basic understanding of electrons in metals. Electrons do not escape from a metal very easily. So their energies outside the metal must be considerably higher than their energies inside. We can represent the metal as a finite potential well for the electrons as described above.

We define the work function ϕ as the minimum extra energy, measured above the Fermi level E_F, which an electron must obtain in order for it to escape from the solid:

$$\phi = E_0 - E_F. \tag{4.59}$$

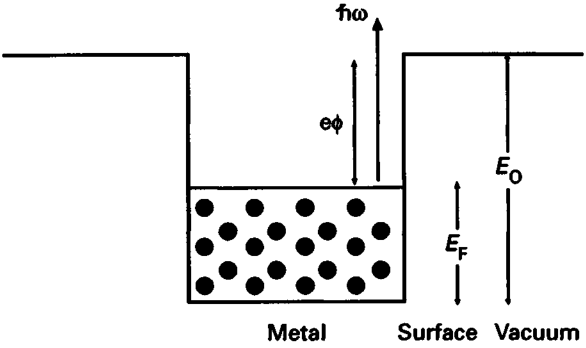


Figure 4.11 Square-well potential model of a metal. The work function ϕ is equal to the difference in energy between the Fermi level and the energy of an electron outside the material.

Once the temperature of the metal is raised above absolute zero the sharp distinction between occupied and unoccupied states at the Fermi level becomes blurred as more electrons are thermally excited to higher-energy states. For most metals a typical value of the work function is $\phi = 4\,\text{eV}$, as shown in Table 3.2. but for some oxide-coated filaments it is as low as $2\,\text{eV}$. It is clear that we must have $k_B T > \phi = E_0 - E_F$ to observe thermionic emission.

4.6 CONCLUSIONS

In what ways does the model succeed or fail to give good predictions of the properties of materials?

The Sommerfeld quantum-mechanical model had some notable successes, particularly the explanation of the electronic contribution to the heat capacity, but also ultimately it had some failures. The model failed to explain the distinction between metals and insulators or semiconductors. It also failed to explain the number of conduction electrons by making no distinction between the bound and free electrons in the solid.

In addition, it did not adequately explain the following:

- The value of the Hall coefficient
- The Wiedemann–Franz law at intermediate temperatures
- The temperature dependence of dc conductivity

- The anisotropy of dc conductivity
- The number of conduction electrons

In order to resolve some of these difficulties it is necessary to consider the electrons in the presence of a periodic potential instead of a flat square-well potential. Certain electrons are then trapped within the periodic potential wells around the ionic cores while other higher-energy electrons ride above the local potential wells and are constrained only by the boundaries of the material.

The periodic potential within the material is much more realistic, of course, because the electrostatic interaction between the atomic cores at the lattice sites and the electrons, which the model neglects, will lead to periodic spatial fluctuations in the potential that are not allowed for in the flat internal potential of the Sommerfeld model. These have been already described classically in Chapter 2.

Metals are those materials in which there are a number of high-energy electrons riding above the periodic potential in quasi-free states. Insulators are materials in which the highest-energy states are constrained within the localized ionic potentials. This means that the insulators do not have 'free' electrons inside the material and therefore the highest-energy electrons in insulators can not contribute to the conductivity in their ground state.

REFERENCES

1. E. Schrödinger (1926) *Ann. Physik* 79, 361.
2. A. Sommerfeld (1928) *Z. Physik* 47, 1.
3. A. Sommerfeld and H. Bethe (1933) *Handbuch der Physik* 24, II.
4. W. Pauli (1925) *Z. Physik* 31, 765.
5. E. H. Wichman (1971) *Quantum Physics*, Addison-Wesley, Reading, Massachusetts.
6. E. Fermi (1926) *Z. Physik* 36, 902.
7. P. A. M. Dirac (1926) *Proc. Roy. Soc. Lond.* 112, 661.
8. S. Goudsmit and G. E. Uhlenbeck (1926) *Nature* 117, 264.
9. W. Gerlach and O. Stern (1924) *Ann. Physik* 74, 673.
10. W. Pauli (1926) *Z. Physik* 41, 81.

FURTHER READING

R. H. Bube (1992) *Electrons in Solids: an Introductory Survey*, 3rd edn, Academic Press, San Diego.

B. K. Tanner (1995) *Introduction to the Physics of Electrons in Solids*, Chapman & Hall, London.

EXERCISES

Exercise 4.1 Fermi energy for a free electron metal

Discuss the principal differences between the classical free electron model and the quantum free electron model. If the electrons described by the wave equation are completely free, determine the number density of electron states per unit volume with wave vector between k and $k + dk$, assuming that there are states per unit volume of k-space. From this expression determine the number of states per unit volume of material with energy between E and $E + \mathrm{d}E$ and calculate the Fermi energy for a free electron distribution with n electrons per unit volume.

Exercise 4.2 Solution of wave equation in a finite square well

Derive the solution for the wave equation in a one-dimensional finite square-well potential of height V. What happens to the solutions if the electrons encounter a periodic potential within the square well? Describe the different types of energy states permissible under these conditions. Why are only certain energy levels allowed for 'quasi-free' electrons in the Sommerfeld model?

Exercise 4.3 Electronic specific heat of copper at 300 K

Using the quantum free electron model calculate the electronic specific heat of copper at 300 K. At what temperatures are the electronic and lattice specific heats of copper equal to one another (assume $\theta_D = 348$ K)?

Exercise 4.4 Fermi energy of electrons

Calculate the root mean square velocity of electrons at the Fermi surfaces of aluminium, copper, and gold if their Fermi energies are respectively 11.7 eV, 7.0 eV, 5.5 eV.

Using the resistivities and densities given in Exercise 3.2 on p. 58, determine the mean free time between collisions of these electrons and find the electron mean free path length in all three metals. The electron rest mass is 9.1×10^{-31} kg.

Exercise 4.5 Diffraction of electrons at the Fermi energy by the crystal lattice

The Fermi energies and interatomic spacings of sodium, copper and silver are given in the following table.

Calculate the wavelength of the electrons at the Fermi energy and determine whether these electrons are diffracted by the lattice.

	Fermi energy (eV)	**Lattice spacing (nm)**
Na	3.12	0.30
Cu	7.04	0.21
Ag	5.51	0.24

Exercise 4.6 Diffraction of electrons and validity of the quantum free electron model

Calculate the energy and velocity of an electron which has a wavelength just short enough to be diffracted by the lattice of each of the materials given in the previous example. From this result what do you deduce about the validity of the quantum free electron model, which assumes no scattering of electrons by the lattice?

5 BOUND ELECTRONS AND THE PERIODIC POTENTIAL

OBJECTIVE

We now look at the situation in which the wave equation is solved in the presence of a periodic potential. The result is a qualitative change in the form of solutions. In this chapter we focus exclusively on the electron properties as defined by the allowable solutions of the wave equation under these conditions. The lattice is present only to the extent that it provides a background for finding the allowed energy levels for the electrons. An important development of the model that arises as a result of the periodic potential is that the electrons inside the material are separated into two types: low energy 'bound' electrons which are spatially constrained to occupy the localized energy wells, and 'free' electrons which have higher energy and can migrate throughout the material. The result is allowed energy 'bands' separated by unallowed energy 'gaps'. The higher energy 'free' electrons are, in general terms, the same as the electrons described in the Sommerfeld model. The 'bound' electrons here represent a new addition to the quantum-mechanical description of the electrons in the material.

5.1 MODELS FOR DESCRIBING ELECTRONS IN MATERIALS

What conceptual models do we have for describing the behaviour of electrons in materials?

So far we have dealt with the description of the electrons in solids in extremely simple terms. We have looked at

(i) Classical particles in a box.
(ii) Waves in free space.
(iii) Waves in an infinite potential.
(iv) Waves in a finite potential.

Although each of these models is useful in as much as it gives us a general understanding of the possible behaviour of electrons in a solid, it must be admitted that all of the above are approximations that are likely to be far removed from the situation inside a real material. A real electron in a real solid will see the periodic potential formed by the atoms on the lattice sites. Ultimately this can hardly be described by the flat potential within the solid which is used in the square-well potential calculations.

5.1.1 Electrons in a periodic potential

What is the next extension of the model to bring it closer to the description of a real material?

The next level of complexity, which brings us surprisingly close to the real conditions in a solid, is to invoke a periodic potential within the solid. This was first attempted by Bethe [1] and Brillouin [2] who expanded the wavefunctions for an electron in terms of a series of plane waves subjected to the periodic potential. For calculation purposes, these can be square wells as shown, although in practice they will be more like parabolic potentials. A useful observation is that the solutions of the wave equation under these conditions are the Fourier transform of the periodic potential.

We have already established the methods for finding the allowed wavefunctions in this situation:

(i) identify the boundary conditions
(remember that both ψ and $d\psi/dx$ must be continuous across the boundaries)
(ii) solve the wave equation
(in accordance with the boundary conditions).

In one dimension the simplest representation of a periodic potential has a square potential well as shown in Fig. 5.1. In this case there are three different energy ranges of interest

(i) $E < V_1$
(ii) $V_1 < E < V_2$
(iii) $E > V_2$.

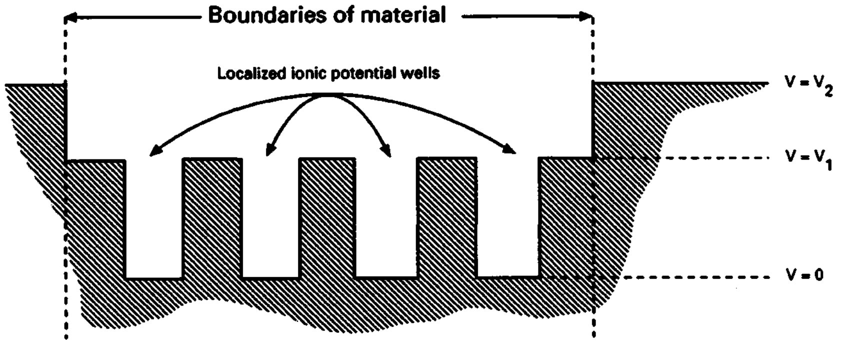


Figure 5.1 One-dimensional representation of a periodic potential using simplified square wells.

The result of subjecting the electrons to a periodic potential is that the energy levels become separated into allowed energy bands, where electron states exist, and forbidden band gaps where no electron states exist.

5.1.2 High energy, free electrons

What are the solutions of the equation of state for very high-energy electrons in the volume occupied by the material?

Clearly for $E > V_2$ we have the case of free electrons so that the wave equation can be solved without the electrons being constrained within the potential well. This means that all values of k are allowed and the energy spectrum is continuous. This is similar to the solution for the finite square-well potential in Section 4.3.2 when $E > V_0$. In this case the electrons have escaped from the material.

5.1.3 Low energy, bound electrons

What form do the solutions of the equation of state take for low energy electrons?
For $E < V_1$ we have a number of finite square-well potentials of width a. Electrons with these energies occupy bound states within these local periodic potentials. This is similar to the situation described above for the single finite square well potential with $E < V_0$. The electron states in this case correspond to those of electrons localized at the ionic cores on the lattice sites in a solid and consequently these electrons can not take part in conduction. These are discrete energy levels.

5.1.4 Intermediate energy, conduction electrons

What form do the solutions of the equation of state take for electrons of intermediate energy?
For $V_1 < E < V_2$ we have a different situation. Here the boundary conditions are provided by $x = \pm b$. Since b is much larger than a this ensures that once the energy is above V_1 the difference in energy between successively higher allowed energy states is less than for the energy levels below V_1. The solutions here are similar to those of a finite square-well potential of depth $V_2 - V_1$ and for electrons with energy $E > V_1$. These are the conduction electrons. However, the periodic potential does perturb the solution of the equations even at these higher energies. The degree of perturbation depends on the depth, width, and number of periodic potentials within the box.

The electrons in this energy range we should correctly call 'quasi-free'. That is, they are constrained only by the physical limitations of the material. They are equivalent to the electrons described by the Sommerfeld model.

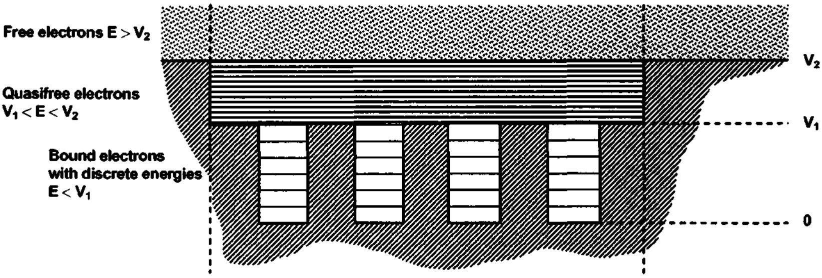


Figure 5.2 Allowed electron energy levels in a one-dimensional 'square-well' lattice obtained by solving the Schrödinger equation.

5.1.5 Comparison with Sommerfeld free electron model

How do the solutions differ from the Sommerfeld model predictions?
This means that there are now three groups of electron states with qualitatively different properties. The totally free electrons are those whose energies are large enough for them to have completely escaped from the metal. These will be of little intrinsic interest, however. Next, the quasi-free conduction electrons, which are constrained only by the boundaries of the solid and form an almost continuous spectrum of k values or allowed energy states. Finally, there are the bound electrons whose energies are low enough for them to be trapped in the potential wells close to the atomic cores on the lattice sites. For these electrons the allowed energy levels are more widely separated.

It is the second group, the quasi-free electrons, which contribute to the electrical and thermal conductivity of metals. These electrons can move throughout the entire solid. We can see now from this model why a free electron theory such as the Sommerfeld model seems to account so well for the bulk electrical and thermal properties of a metal. It is because the solutions of the wavefunction in the Sommerfeld model are quite close to the wavefunctions of the higher-energy electrons in a metal, which are the main contributors to the electrical and thermal properties of metals. We shall find that the difference between a metal and an insulator is that the metal has some of these quasifree electrons whereas the insulator does not.

5.2 SOLUTION OF THE WAVE EQUATION IN A ONE-DIMENSIONAL PERIODIC SQUARE-WELL POTENTIAL

What changes occur in the solution of the wave equation in the presence of a periodic square-well potential?
We now introduce the periodicity of the crystal lattice in one dimension and solve the wave equation under these conditions, as in the paper by Kramers [3]. Consider square-well potentials of height V_0, width c, periodicity a and distance between wells b. Then $b = a - c$ and the potential has the form

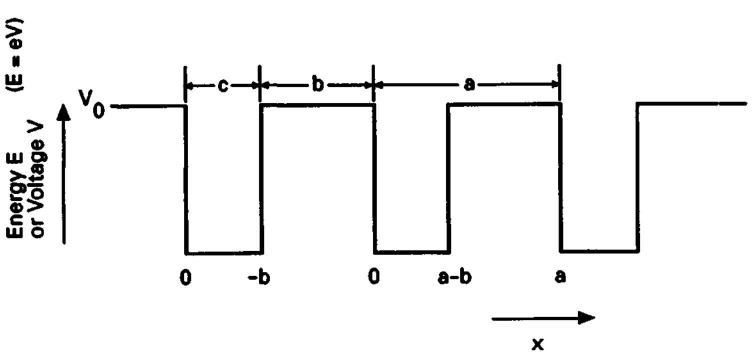


Figure 5.3 A simplified square-well potential representation of a one-dimensional lattice.

As we have shown before, the allowed energies can be found by solving the Schrödinger equation for the regions where $V = 0$ and $V = V_0$ separately. Spatially we separate the regions $V = V_0$ for $-b \leq x \leq 0$, and $V = 0$ for $0 \leq x \leq a - b$, and furthermore, $V(x + a) = V(x)$. The Schrödinger equation for the one-dimensional, time-independent case is,

$$-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\psi(x) + V(x)\psi(x) = E\psi(x), \qquad (5.1)$$

which we now solve in the two regions of interest.

(i) Where $V = 0$,

$$\psi_0(x) = A\exp(i\beta x) + B\exp(-i\beta x), \qquad (5.2)$$

and since $V = 0$, then the wave vector β is given by

$$\beta = \frac{(2mE)^{1/2}}{\hbar}. \qquad (5.3)$$

(ii) Where $V = V_0$

$$\psi_v(x) = C\exp(\alpha x) + D\exp(-\alpha x), \qquad (5.4)$$

and since $V = V_0$, and assuming that $E < V_0$ we obtain the relation

$$\alpha = \frac{(2m(V_0 - E))^{1/2}}{\hbar}. \qquad (5.5)$$

So, if $E > V_0$ this leads to oscillatory (wave-like) solutions, but if $E < V_0$ the solutions are simply decaying exponentials.

Applying boundary conditions, it must be remembered that two continuity conditions must be satisfied:

(i) $\psi(x)$ must be continuous,
(ii) $\frac{d\psi(x)}{dx}$ must be continuous.

So, for example, at $x = 0$ at the boundary of one potential well, our boundary conditions require that the wavefunction ψ_0 in the $V = 0$ region matches the wavefunction ψ_v in the $V + V_0$ region,

$$\psi_0(0) = \psi_v(0), \qquad (5.6)$$

and also the derivatives must match at the boundary,

$$\frac{d}{dx}\psi_0(0) = \frac{d}{dx}\psi_v(0). \qquad (5.7)$$

There is also a periodicity condition on the wavefunction. In order that the wavefunction should have the same periodicity as the lattice we also require that the wavefunction at $x+a$ is the same as at x. This periodicity condition can be expressed in terms of a wavefunction with vector k, as was shown by Bloch [4]:

$$\psi(x+a) = \psi(x)\exp(\mathrm{i}ka). \tag{5.8}$$

Therefore, at the boundaries $x = -b$ and $x = a - b$ we have

$$\psi_v(-b) = \exp(-\mathrm{i}ka)\psi_0(a-b) \tag{5.9}$$

$$\frac{\mathrm{d}\psi_v(-b)}{\mathrm{d}x} = \exp(-\mathrm{i}ka)\,\frac{\mathrm{d}\psi_0(a-b)}{\mathrm{d}x}. \tag{5.10}$$

The above four conditions lead to four simultaneous equations in the unknowns A, B, C and D:

$$\psi(0){:}\quad A+B = C+D \tag{5.11}$$

$$\frac{\mathrm{d}\psi(0)}{\mathrm{d}x}{:}\quad \mathrm{i}\beta(A-B) = \alpha(C-D) \tag{5.12}$$

$$\psi(-b){:}\quad \begin{matrix} C\exp(-\alpha b) \\ D\exp(\alpha b) \end{matrix} = \exp(-\mathrm{i}ka)\begin{bmatrix} A\exp(\mathrm{i}\beta(a-b)) \\ +B\exp(-\mathrm{i}\beta(a-b)) \end{bmatrix} \tag{5.13}$$

$$\frac{\mathrm{d}\psi(-b)}{\mathrm{d}x}{:}\quad \begin{matrix} \alpha C\exp(-\alpha b) \\ -\alpha D\exp(\alpha b) \end{matrix} = \exp(-\mathrm{i}ka)\mathrm{i}\beta\begin{bmatrix} A\exp(\mathrm{i}\beta(a-b)) \\ -B\exp(-\mathrm{i}\beta(a-b)) \end{bmatrix}. \tag{5.14}$$

These four simultaneous equations can be solved if we require the determinant of the coefficients A, B, C, D to vanish. In that case the following energy restriction is obtained

$$\cos(ka) = \left(\frac{\alpha^2-\beta^2}{2\alpha\beta}\right)\sinh(\alpha b)\sin(\beta(a-b)) + \cosh(\alpha b)\cos(\beta(a-b)). \tag{5.15}$$

The limitation imposed by this condition is that the expression on the right-hand side of the equation must lie between ±1 for allowed solutions. Mathematically it is possible for the expression on the right-hand side to lie outside the range 1, but then this does not correspond to a physically allowed solution.

5.2.1 Kronig–Penney approximation

What simple approximations can be made to the periodic potential in order to demonstrate the form of the solutions?

A simplification of the above constraint can be obtained by a mathematical model known as the Kronig–Penney approximation [5]. In this the width of the potential barriers, b, is allowed to decrease to zero, while the height V_0 of the barriers is allowed to increase to infinity under the condition that the product abV_0 remains constant. In this case then we have the following limits:

$$\lim_{b\to 0} \sin\beta(a-b) = \sin\beta a \tag{5.16}$$

$$\lim_{b\to 0} \cos\beta(a-b) = \cos\beta a \tag{5.17}$$

$$\lim_{b\to 0} \sinh\alpha b = \alpha b \tag{5.18}$$

$$\lim_{b\to 0} \cosh\alpha b = 1. \tag{5.19}$$

The following relations for the wave vectors still hold,

$$\beta = \frac{(2mE)^{1/2}}{\hbar}, \tag{5.20}$$

$$\alpha = \frac{(2m(V_0 - E))^{1/2}}{\hbar}, \tag{5.21}$$

and again for electrons in the regions with potential $V = V_0$ and with energies greater than V_0 solutions are wave-like, while for those in the same regions with energies less than V_0, α is real and the solutions are exponentials.

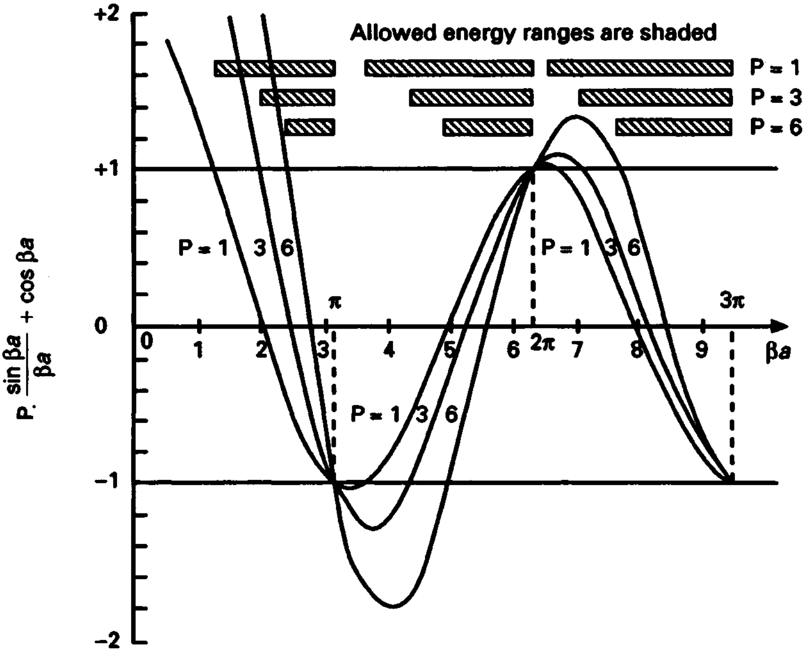


Fig. 5.4 Allowed energy levels of electrons according to the Kronig–Penney model. Reproduced with permission of R. H. Bube, *Electronics in Solids: An Introductory Survey*, published by Academic Press, 1971.

Now the previous constraint becomes,

$$\cos(ka) = \frac{P}{b\alpha}\sin(\beta a) + \cos(\beta a), \tag{5.22}$$

where,

$$P = \frac{mabV_0}{\hbar^2}. \tag{5.23}$$

Allowed energy bands according to this model are shown in Fig. 5.4.

The only energies allowed are those for which

$$\left|\frac{P}{b\alpha}\sin(\beta a) + \cos(\beta a)\right| \leq 1. \tag{5.24}$$

As βa increases, for example with the lattice parameter a, the allowed energy bands become wider because the solutions become more like those of free electrons. However, as the potential V_0 increases, so $P(=abmV_0/\hbar^2)$ increases, the electrons become less free and the allowed energy bands become narrower. As V_0 increases further the solutions ultimately will resemble those of an electron trapped in an infinite square-well potential.

5.2.2 The nearly free approximation

What values of energy are allowed for a wave in a one-dimensional discrete periodic lattice?

Another way of looking at the behaviour of electrons in a periodic potential is to consider the allowed solutions of the wave equation in a periodic potential as a perturbation from free electron behaviour. Using a one-dimensional lattice of length L and with n atoms it can be shown that the allowed values of k are governed by

$$k_n^2 = \frac{\pi^2 n^2}{L^2} \tag{5.25}$$

and since for free electrons $E = \hbar^2 k^2/2m$, it is immediately apparent that only certain restricted values of energy are allowed by the lattice: those corresponding to integer values of n:

$$E_n = \frac{\hbar^2}{2m}\frac{n^2\pi^2}{L^2} = \frac{n^2 h^2}{8mL^2}. \tag{5.26}$$

Extending this to three dimensions gives

$$k_n^2 = k_x^2 + k_y^2 + k_z^2 \tag{5.27}$$

$$= \frac{\pi^2}{L^2}(n_x^2 + n_y^2 + n_z^2), \tag{5.28}$$

and

$$E_n = \frac{\hbar^2}{2m}(k_x^2 + k_y^2 + k_z^2). \tag{5.29}$$

The idea of solving the wave equation in an 'empty lattice' has been discussed by Shockley [6] and provides a limiting case for very weak potentials.

5.2.3 Density of states

How does the number of available energy states for electrons vary with energy of the electrons in the free electron approximation?

The number of different k states that is possible up to a given value k_n can be calculated easily in the free electron limit by noting that the values n_x, n_y and n_z must all be positive. Therefore the number of available states is simply the number of unit cubes in the positive quadrant of n-space of radius n. In other words it is one eighth of the volume of a sphere in n-space:

$$\mathrm{N}_0(E_n) = \frac{1}{8}\frac{4}{3}\pi n^3, \tag{5.30}$$

and since $n = \dfrac{Lk_n}{\pi}$,

$$N_0(E_n) = \frac{1}{6\pi^2} L^3 k_n^3. \tag{5.31}$$

Here we must remember that the electrons have spin, and therefore each energy level can be occupied by two electrons. Therefore the number of electrons from $k = 0$ to $k = k_n$ will be twice $N_0(E_n)$

$$2N_0(E_n) = \frac{1}{3\pi^2} L^3 k_n^3. \tag{5.32}$$

If we wish to express this number of electron states in terms of energy E_n instead of k_n, we need to know the relationship between E_n and k_n. In general, this is not known, but if we make the free electron approximation, $E(k) = \hbar^2 k^2/2m$, then the number of available states between $E = 0$ and $E = E_n$ is

$$2\mathrm{N}_0(E_n) = \frac{1}{3\pi^2} L^3 \left(\frac{2mE_n}{\hbar^2}\right)^{3/2} \tag{5.33}$$

This function can then be differentiated with respect to E to obtain the density of states per unit energy interval

$$D(E) = \frac{\mathrm{d}}{\mathrm{d}E} \mathrm{N}_0(E) \tag{5.34}$$

$$= \frac{L^3}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2} \tag{5.35}$$

$$= \frac{V}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2}, \tag{5.36}$$

where V is the volume of the material. This is the same result that was obtained in Section 4.4.7 beginning with a different condition, the square-well potential.

5.3 THE ORIGIN OF ENERGY BANDS IN SOLIDS: THE TIGHT-BINDING APPROXIMATION

How can we view the electron energy levels in a material as if they had evolved from the energy levels in single atoms?

We have already seen in the previous discussion that there is a qualitative difference in the properties of the high-energy, quasi-free electrons which are responsible for electrical and thermal conduction; and the low-energy bound electrons which are trapped in the atomic potential wells located on the lattice sites.

Now we shall take another approach originated by Bloch [7] which also contributes to our understanding of the properties of electrons in solids. It is well known that the energy levels of electrons within a single isolated atom are highly discrete. This was shown, for example, in the Bohr model of the atom, in which these energies were given by

$$E_n = -\left(\frac{me^4}{8h^2\varepsilon_0^2}\right)\frac{1}{n^2}, \tag{5.37}$$

where n is an integer, m is the mass of the electron, e is the charge of the electron, h is Planck's constant and ε_0 is the permittivity of free space.

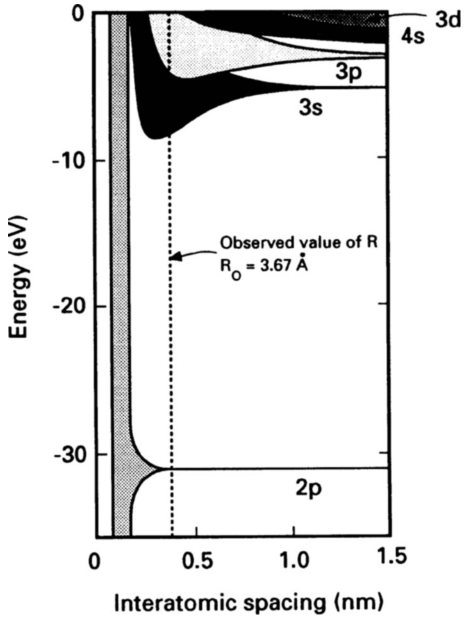


Figure 5.5 Broadening of the allowed electron energy states into electron 'bands' as the interatomic spacing a is reduced. Reproduced with permission of R. H. Bube, *Electronics in Solids: An Introductory Survey*, published by Academic Press, 1971.

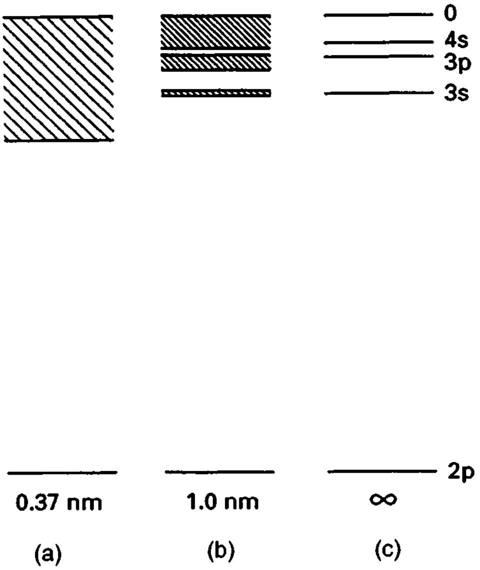


Figure 5.6 Energy 'spectrum' of allowed states for (a) free electrons outside a solid, (b) band electrons within a solid and (c) bound electrons in isolated atoms. Reproduced with permission of R. H. Bube, *Electronics in Solids: An Introductory Survey*, published by Academic Press, 1971.

If we consider what happens to these sharply defined electronic energy levels as the atoms of a solid are brought together we find that they become broadened. This occurs because the Pauli exclusion principle does not allow two electrons to have completely identical states. Therefore, the energies which were identical in the isolated atoms must shift relative to one another, and the closer the atoms are together the more marked is the shift in available energy states. Furthermore, the number of different energy states is dependent on the number of atoms.

This means that a discrete energy level in a particular type of atom broadens into an allowed band of energies when a large number of identical such atoms are brought together in a solid. The higher-energy states broaden first as the atoms are brought closer together. The broadening of the lower-energy states, which are closer to the atomic cores, is less marked.

If we look more at the three cases consisting of the two extreme cases of bound electrons and free electrons and the intermediate case of electrons in a solid, we find the following types of energy spectrum.

5.3.1 The tight-binding approximation versus the free electron approximation

How do the extreme viewpoints of single atom versus collective electron compare? The approach which we have just taken, considering the energy levels in the isolated atom and then bringing the atoms together and modifying the energy levels, is known as the 'tight-binding approximation' because we start from tightly bound electrons in the atoms. The previous approach, beginning with the free electron model and progressing through to a periodic potential model is known as the free electron approximation. Both should lead to the same result if their conditions are relaxed sufficiently.

Both are valid methods of looking at the electronic properties of materials but approach the problem from different directions. In fact, an intermediate model by Herring [8] has also been used to calculate the energy states of electrons in solids by combining the best aspects of both models. This is known as the orthogonalized plane wave method.

In metals, of course, as we have already noted, the higher-energy electrons behave as if almost free. Therefore, the free electron approximation is more relevant in these cases. In semiconductors and insulators, the electrons do not become free in the same sense because the highest-occupied energy states in the ground state are localized at the atomic lattice sites, and so in this case the tight-binding approximation is more relevant. However, in either model we are led to the conclusion that when large numbers of atoms are brought together, energy bands arise and these bands are separated by forbidden energy regions or band gaps.

5.3.2 Transition from insulator to metal under pressure

What happens to the electronic properties of an insulator if it is subjected to extremely high pressure?

Since in their ground states electrons will always occupy the lowest available energy state, it is possible to convert an insulator to a metal under very intense pressure as a result of the broadening of the energy bands which occurs when the atomic cores are moved closer together.

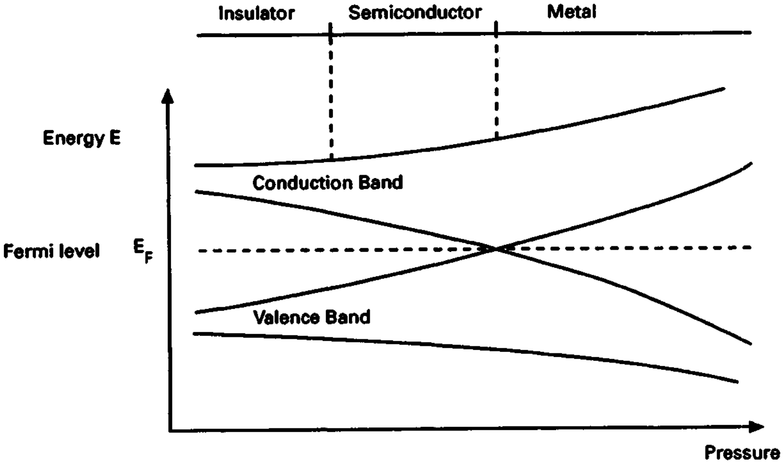


Figure 5.7 Effects of pressure on the allowed electron energy levels of atoms. At low pressure the atoms are isolated. At intermediate pressure they form an insulator and at very high pressures they form a metal. Compare with Figs. 5.5 and 5.6.

If we make the assumption that the Fermi level does not change, then the material will undergo a transition from insulator to metal at the point where the two previously separated valence and conduction bands begin to overlap.

We should note that the pressures needed to cause this type of transition are very high. For example, germanium, which is a semiconductor under normal conditions, becomes a metal under a pressure of 120 kbars (12 GPa). Recent results also suggest that solid hydrogen becomes metallic under a pressure of 2.5 Mbar (250 GPa).

5.4 ENERGY BANDS IN A SOLID

How are the properties of the electrons in the different energy bands qualitatively different from each other?

We have seen how the discrete energy levels in an atom are broadened into allowed energy bands when the atoms are brought together in a solid. The upper energy bands, which correspond to electrons with energies above the periodic potential wells of the atoms in the solid, contain the so-called 'free electrons'. Below this there may be other free-electron-type bands. Finally, below the free electron bands are the bound states. The free electron bands are constrained only by the boundaries of the solid. The bound electrons are constrained by the local potential wells around the atomic cores. The local ionic potentials cause a perturbation of the wavefunctions of the free electron bands so that they are not identical to the solutions obtained in the Sommerfeld model.

5.4.1 Width of energy band gaps

How are the widths of the energy gaps related to the periodic potential?

Clearly since the energy gaps are caused by the presence of the periodic potential there must be some relationship between them. In fact the relationship is quite simple. The band gaps in the electron energy levels are equal to the Fourier components of the crystal potential. This can be understood because the electrons behave as waves under the influence of the periodic potential.

5.4.2 Electron band structure in conventional space

How can we view the various electron bands in real space?

In real space the electrons are confined within the solid. Low-energy 'bound' electrons cannot participate in conduction unless they are thermally excited and escape from the atomic core. High-energy 'quasi-free electrons' migrate throughout the solid, being constrained by the physical boundary of the solid only. The movement of these higher-energy electrons is almost unaffected by the periodic potential of the atomic cores.

So far we have looked only at the distribution of electrons in real space. We will introduce later the concept of reciprocal space which can be used to describe the electronic states of a solid in a particularly economical and elegant manner.

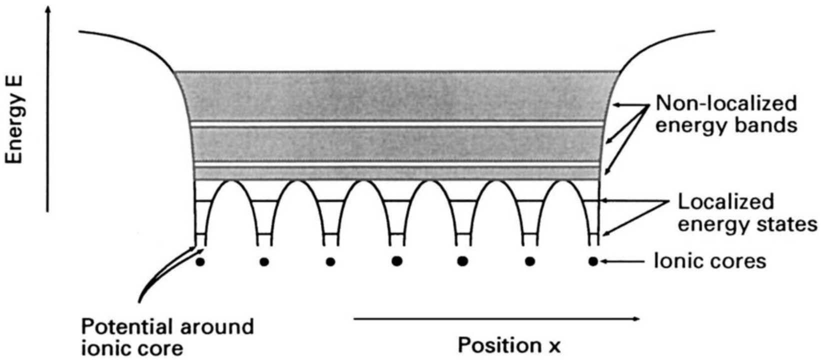


Figure 5.8 Schematic diagram of energy levels in a one-dimensional lattice, shown in real space.

5.4.3 The Fermi energy

What is the highest energy level occupied by an electron when all electrons are in their lowest available state?

We have stated that at the absolute zero of temperature, when the electrons all occupy the lowest available energy state, the energy of the highest occupied state is the Fermi level. This energy level separates the occupied from the unoccupied electron levels only when the electron configuration is in its ground state, that is only at 0 K.

The location of the Fermi level in relation to the allowed energy states is crucial in determining the electrical properties of a solid. Metals always have a partially filled free electron band, so that the Fermi level corresponds to a level in the middle of the band and this makes the metals electrical conductors. Semiconductors always have completely filled or completely empty electron bands. This means that the Fermi energy lies between the bands, and consequently they are poor electrical conductors at ambient temperatures.

5.4.4 Nomenclature of electron bands

How can the most important energy bands be described in a distinct and self-consistent manner?

We shall define the highest-energy electron band containing electrons when the material is in its ground state as the valence band. We shall define the lowest-energy band containing unoccupied electron states when the material is in its ground state as the conduction band, since it is through this band that electrical conduction can take place. In a semiconductor or insulator, the distinction is clear and we can represent the bands as shown in Fig. 5.9.

In a metal which contains a partially filled band, this band satisfies both criteria and so is both a valence band and a conduction band. This explains some of the confusing nomenclature of the bands in metals which occurs in the literature, in which sometimes the free electrons are described as 'conduction' electrons and sometimes they are described as 'valence' electrons. In a metal they are both.

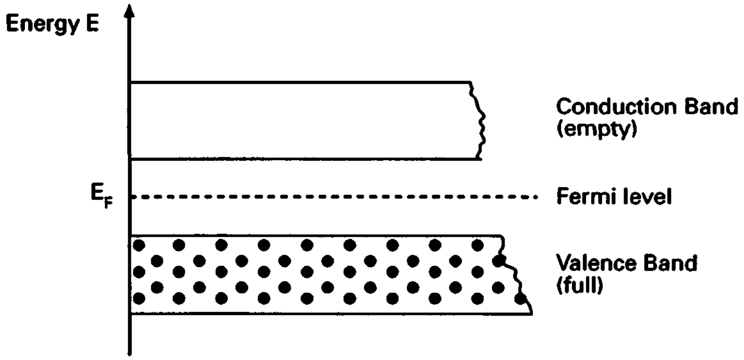


Figure 5.9 Schematic diagram of electron energy bands in a semiconductor or insulator.

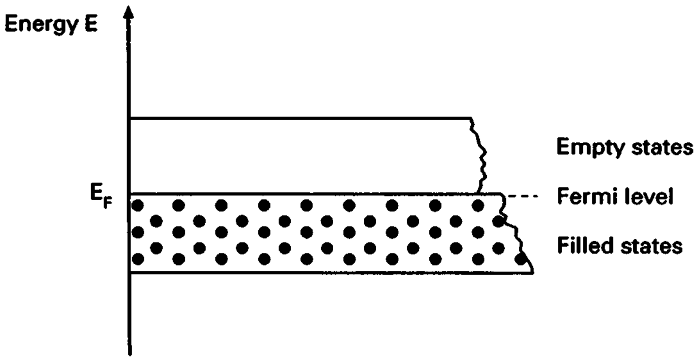


Figure 5.10 Schematic diagram of electron energy bands in a metal.

5.4.5 Effective mass of electrons in bands

How can the motions of electrons within an energy band be described in a simple way?

It is found experimentally that the mobility of electrons in the conduction band is affected by how full the band is. In the case of free electrons we have shown that

$$E(k) = \frac{\hbar^2 k^2}{2m}, \tag{5.38}$$

where m is the mass of the electrons.

As we shall see shortly this relationship between E and k breaks down in a solid. However, we can maintain the form of this relationship by using the relation

$$E(k) = \frac{\hbar^2 k^2}{2m^*}, \tag{5.39}$$

where now m^* is an adjustable parameter which we call the effective mass. This means that any deviations from a parabolic relationship between E and k can be expressed as a change in the effective mass of the electrons at that point in k-space. Remember that, of course, this is merely a convenient artifice which allows us to describe the behaviour of the electrons in bands. The electron does not actually change its mass at these energies, it is simply an expression of the changed relationship between E and k.

The effective mass m^* can be smaller or larger than the free electron mass m. The cause of this apparent change in the mass of the electrons is the interaction between the electrons and the lattice in the material. Collisions between drifting electrons and atomic sites will slow down the acceleration of an electron which will lead to an increase in its effective, or apparent, mass.

Another way to view the situation is in terms of the curvature of the energy bands in k-space. This can be interpreted in the following way: using the relationship between energy E and wave vector k given in Section 4.2 and taking the second derivative of the energy with respect to the wave vector gives $\hbar^2/m$. This relationship only holds exactly for a free electron parabola. When the relationship between E and k is no longer parabolic the deviation can be expressed in terms of a change in the effective mass so that

$$\frac{d^2E}{dk^2} = \frac{\hbar^2}{m^*}, \tag{5.40}$$

and consequently the effective mass can be defined by

$$m^* = \frac{\hbar^2}{(d^2E/dk^2)}, \tag{5.41}$$

where d^2E/dk^2 is the curvature of the electron levels or electron band in k-space.

For electron bands with high curvature m^* is small, while for bands with small curvature, that is flat electron bands, m^* is large. It is also worth noting that since d^2E/dk^2 can be negative m^* can be negative. This simply means that when an electron goes from state k to state $k + \delta k$ the momentum transfer to the lattice is greater than the momentum transfer to the electron. The electron therefore appears to have a negative mass.

As a simple example, consider the following energy states, as shown in Fig. 5.11. Small effective masses occur at low k in this case, large effective masses occur at intermediate k. Since the effective mass is determined by the curvature of the energy band in k-space, this means that narrow bands necessarily contain electrons with high effective mass. Conversely, wide bands can contain electrons of low effective mass or high effective mass.

5.5 RECIPROCAL SPACE OR WAVE VECTOR k-SPACE

Is there an economical way of describing all of the allowed energy states in a solid? Earlier, in Section 4.2, we introduced the idea of the wave vector k. This arose when we made a simple calculation for the solution of the wave equation for free electrons and then for electrons in a square-well potential. A plot of energy E against wave vector k was given first in Fig. 4.1. The dimensions of k are reciprocal

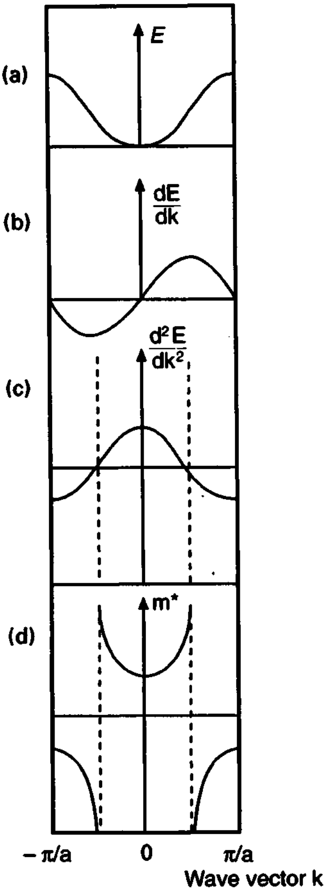


Figure 5.11 Idealized variation of electron energy *E* with wave vector *k*. The lower diagram shows the curvature of the electron energy band giving large effective mass at intermediate *k*, but small effective mass at low *k* as shown in the diagram. At *k* values close to $\pm\pi/a$ the effective mass of the electrons is negative.

length. It tells us the spatial periodicity of the wavefunction, or if you prefer, the number of cycles of the wave which occur in a given distance of 2π metres.

We found that for free electrons the energy depended on k^2, and all values of k were allowed. When the electrons are trapped, as in the square-well potential, only certain values of k are allowed in order that the wavefunctions can meet their boundary conditions.

In the last section we began to plot energy against k because this was useful in determining the effective mass of the electrons. We shall find that when it comes to describing electrons in solids plotting E against k is a very useful way of representing the electronic properties. The plot of E against k is known as a reciprocal space plot because the dimensions of k are m^{-1}.

When an electron is confined within a solid and experiences the periodicity of the lattice this periodicity affects the relationships between E and k. Another way of looking at this is that a wavefunction described by the wave vector k will have different energies depending on the presence and type of the crystal lattice it encounters. We have already noticed, for example, that the interactions between an electron and the lattice alter its effective mass and so distort the relationship between E and k.

5.5.1 Brillouin zones

How can periodicity of the lattice be introduced into reciprocal space?

We may think of the Brillouin zones [9] as a method for introducing the periodicity of the crystal lattice into our model of the electronic structure of materials. The Brillouin zone is a region in reciprocal space. Before saying more about exactly what this Brillouin zone is, let us consider the effects of a periodic potential on the energy levels of 'free electrons'.

We know $E = \hbar^2 k^2/2m$ for free electrons, but when we add the presence of a periodic potential all of this changes. We know from the previous chapter that if the energies of the electrons are below the level of the periodic potential barriers the electrons will penetrate spatially into these potential barriers, but their wavefunctions will be attenuated the further they penetrate. If the potential barriers are much higher than the electron energy, then the wavefunction will be reflected at these barriers and the electron will be contained entirely within the potential box formed by the barriers.

5.5.2 Bragg reflection at the Brillouin zone boundaries

What effect does the lattice have on the allowed electron wavefunctions and the relationship between energy E and wave vector k ?

The solutions of the Schrödinger equation in the case of a periodic potential with infinite energy barriers is simply

$$E_n = \frac{\hbar^2}{2m}\left(\frac{n\pi}{a}\right)^2, \tag{5.42}$$

with integer values of n, as shown in Fig. 4.3. This means that only exact discrete values of the energy are allowed under these conditions.

If a finite periodic potential exists at the lattice sites $x = \pm na$, then this results in a perturbation of the solutions of the wave equation compared with the free electron situation with the amount of perturbation increasing as the height of the potential barriers increases. In the simplest form of periodic potential $v(x) = 0$ for $-na < x < na$ and $V(x) = V_0$ for $x = \pm na$. The locations where the distortion from the free electron parabola is greatest will be the locations where the potential is greatest. These occur at $k = \pm n\pi/a$.

Now it is necessary to find what form the solutions take close to these 'zone boundaries' where the potential is nonzero. Both the wavefunction Ψ and its derivative Ψ' must be continuous in real space. This is one of the conditions of quantum mechanics. Therefore, the potential $V(k)$ is not simply restricted to finite values at the points $k = \pm n\pi/a$. Therefore, even a 'point potential' in real space will have a finite extent in k-space. In fact, $V(k)$ is the Fourier transform of $V(x)$:

$$V(k) = \frac{1}{2a}\int_{-a}^{a} V(x)\exp(ikx)\,dx \tag{5.43}$$

$$= V\left(\frac{2n\pi}{a}\right), \tag{5.44}$$

and this results in the addition of an amount of energy $\pm V(n\pi/a)$ to the allowed solutions of the wave equation at the 'zone boundaries.' The energies are therefore,

$$\lim_{k\to\pi/a} E(k) = \frac{\hbar^2 k^2}{2m} - V(k), \tag{5.45}$$

(from below)

$$\lim_{k\to\pi/a} E(k) = \frac{\hbar^2 k^2}{2m} + V(k). \tag{5.46}$$

(from above)

So at the zone boundary, the energy states in the first zone are $2V(k = \pi/a)$ below the energy of the states in the second zone. This is the so-called 'band gap', whereby no electron states are available with energies in the range $(\hbar^2/2m) \times (\pi/a)^2 \pm V(k = \pi/a)$ because of the presence of the periodic potential. The magnitude of the energy gap $2V(\pi/a)$ can, therefore, be found from the Fourier transform of the periodic potential of the lattice.

At locations away from the zone boundary, the Fourier transform $V(k)$ of the potential also has a finite value, which decreases with distance in k-space from the zone boundary. This means that the E versus k relationship is also distorted or perturbed from the parabolic free-electron-like behaviour at values of k away from the zone boundary, but the farther from the zone boundary, the less the perturbation. This is shown in Fig. 5.12.

The distortion or perturbation of the electron wavefunction that occurs at zone boundaries of the periodic potential is known as Bragg reflection. The reason for this terminology becomes clear if we consider what happens to the wavefunction. In the limiting case of an infinite potential at the atomic or ionic sites the electrons cannot stray into this region. Therefore, the electron wavefunctions must be completely contained within the local potential well between $x = \pm a$ and the wavefunction must meet exact boundary conditions of $\Psi(a) = \Psi(-a) = 0$. This is equivalent to total reflection of the wave.

It is clear, therefore, that an infinite periodic potential will give total reflection at the zone boundaries. If the height of the potential is reduced then the reflection will

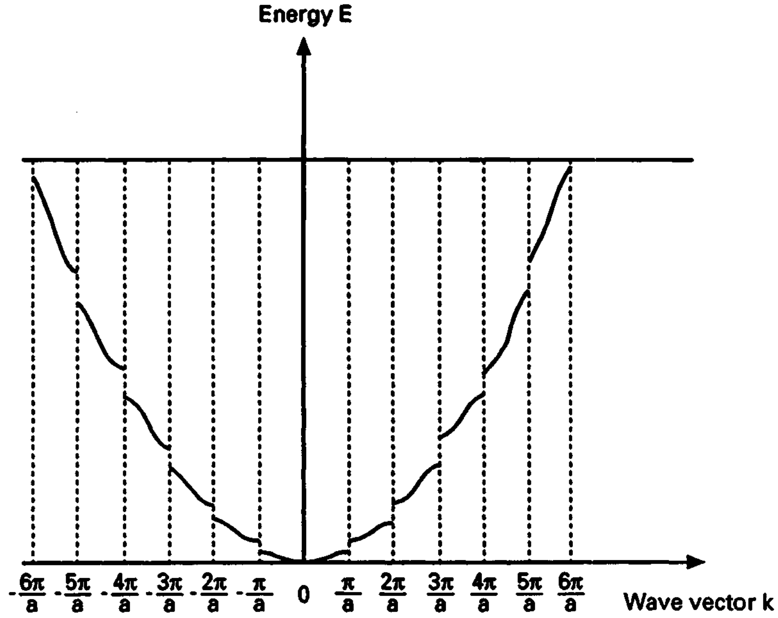


Figure 5.12 Deformation of a free electron parabola due to a weak periodic potential.

not be perfect at the boundary and this will result in partial reflection and partial transmission of the electron wavefunction. The ratio of reflection to transmission depends on the height and width of the potential barriers.

Now that we have introduced the idea of the electron wavefunction being reflected by the energy barriers of the periodic potential, let us find the condition for this. For Bragg reflection we know that,

$$2a\sin\theta = n\lambda, \tag{5.47}$$

where a is the lattice parameter, θ is the angle of reflection, λ is the wavelength of the electrons and n is an integer. For simplicity let us again look at the one-dimensional lattice. In this case $\sin\theta = 1$ and the Bragg reflection condition is,

$$2a = n\lambda \tag{5.48}$$

and $\lambda = 2\pi/k$ for an electron. So that,

$$k = \frac{n\pi}{a}. \tag{5.49}$$

This then is the condition for reflection of the electron wavefunction in a one-dimensional lattice of parameter a. This can easily be generalized to three dimensions and the concept remains the same. We now have a division of reciprocal or k-space into a number of zones, known as Brillouin zones, at the boundaries of which reflection of the electron wavefunctions takes place.

The lattice can be divided into a number of Brillouin zones in reciprocal space,

First Brillouin zone $\dfrac{-\pi}{a}$ to $\dfrac{+\pi}{a}$.

Second Brillouin zone $\dfrac{-2\pi}{a}$ to $\dfrac{-\pi}{a}$ and $\dfrac{\pi}{a}$ to $\dfrac{2\pi}{a}$.

nth Brillouin zone $\dfrac{-n\pi}{a}$ to $\dfrac{-(n-1)\pi}{a}$ and $\dfrac{(n-1)\pi}{a}$ to $\dfrac{n\pi}{a}$.

Since free electron wavefunctions are not reflected, and yet in a solid reflection occurs at the Brillouin zone boundaries, we may reasonably expect that the most severe deviation of the electron energies from free-electron-like behaviour will occur at the Brillouin zone boundaries. This can be seen in Fig. 5.12 in which the E versus k relationship is distorted from the free electron parabola at these boundaries.

In the plot of E versus k for a one-dimensional weak periodic potential shown in Fig. 5.12, the free electron nature of the electron states is clearly apparent but with some distortion at the zone boundaries. At the k values corresponding to the zone boundaries transmission of an electron through the solid is prevented. The incident and reflected electron wavefunction form a standing wave. Certain energies are therefore forbidden since there are values of E for which there is no corresponding value of the wave vector k.

5.5.3 The reduced zone scheme

Can the representation of the electrons in reciprocal space be made more compact by making use of the periodicity condition?

We can take our one-dimensional plot of energy versus wave vector and map all sections into the range $-\pi/a \leq k \leq \pi/a$. This is done by using the periodicity constraint so that,

$$k_1 = k_n + G, \tag{5.50}$$

where k_n is the wave vector in the nth Brillouin zone, k_1 is the wave vector in the first Brillouin zone and G is a suitable translation vector. Now any point in k-space can be mapped by symmetry considerations to an equivalent point in the first Brillouin zone, but notice that many points from the extended zone representation can be mapped to the same point in the reduced zone representation.

This reduced zone scheme representation allows the entire electron band structure to be displayed within the first Brillouin zone. This is a very compact representation and has distinct advantages because the electronic states can be displayed in the most economical way in a single diagram of the first zone.

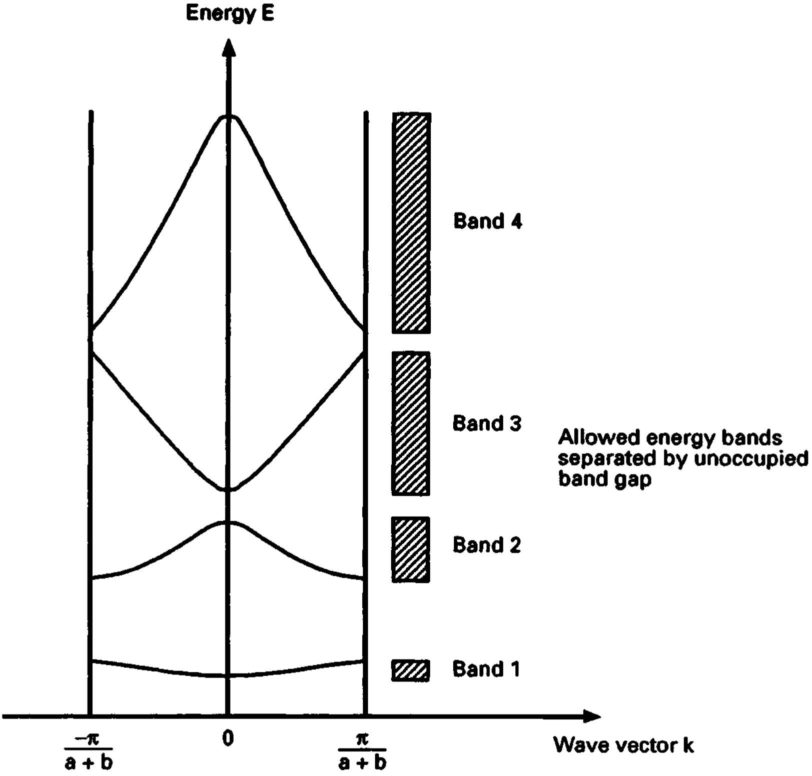


Figure 5.13 The distorted free electron parabola due to a weak periodic potential mapped back onto the first Brillouin zone.

5.5.4 Band structures in three dimensions

How can information about the energy levels of a three-dimensional solid be represented in a compact manner on a two-dimensional diagram?

We have shown in a previous section the importance of plotting the electronic energy E against the wave vector k because this gives immediate information about the electronic properties of a material. We have also mentioned the reduced zone representation which is useful because it presents this information in its most compact form. However, we have only done this so far for a one-dimensional lattice and a one-dimensional reciprocal lattice whose first Brillouin zone extends from $-\pi/a$ to $+\pi/a$.

In three-dimensional crystals with three-dimensional reciprocal lattices, the use of a compact representation is no longer merely a convenience, it is essential. Otherwise the representation of the electronic states becomes very complex to envisage. How then can we display the band structure information from a three-dimensional crystal, which needs, of course, four dimensions (E, k_x, k_y and k_z) to describe it?

The answer is to make representations of certain symmetry directions in the 3-D Brillouin zone as one-dimensional *E* versus *k* plots, and to do this for several important symmetry directions in the Brillouin zone. Only by doing this can we get all of our information onto a two-dimensional page. You can think of this process as cutting the Brillouin zone along certain symmetry directions. Therefore when looking at an *E* versus *k* diagram you are looking at several slices through different directions of *k*-space.

5.5.5 Brillouin zone of an fcc lattice

What does the 'unit cell' of an fcc lattice look like when transformed into reciprocal space?

The face-centred cubic lattice space group has been shown above in Fig. 2.1. Now we will look at the Brillouin zone of such a lattice in *k*-space. This is shown in Fig. 5.14.

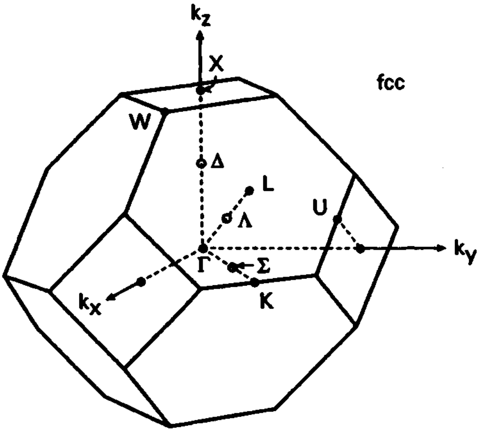


Figure 5.14 Brillouin zone of an fcc lattice in reciprocal space.

Certain symmetry points of the Brillouin zone are marked. Specifically, the Γ, X, W, K and L points and the directions Δ, Λ and Σ. Roman letters are used mostly for symmetry points and Greek letters for symmetry directions. The following is a summary of the standard symbols and their locations in *k*-space:

Γ	<0, 0, 0>
X	<1, 0, 0>
W	<1, 1/2, 0>
K	<3/4, 3/4, 0>
L	<1/2, 1/2, 1/2>

5.5.6 Brillouin zone of a bcc lattice

What does the 'unit cell' of an bcc lattice look like when transformed into reciprocal space?

Similarly, the Brillouin zone of a bcc lattice can be described in terms of its principal symmetry directions. The zone is shown in Fig. 5.15. The symmetry points are conventionally represented as Γ, H, P and N and the symmetry directions as Δ, Λ, D, Σ and G. The various symmetry points are,

Γ	<0, 0, 0>
H	<1, 0, 0>
P	<1, 1, 1>
N	<1, 1, 0 >

Notice that a bcc lattice in real space has a Brillouin zone in reciprocal space that has fcc symmetry, while an fcc lattice in real space has a Brillouin zone in reciprocal space that has bcc symmetry.

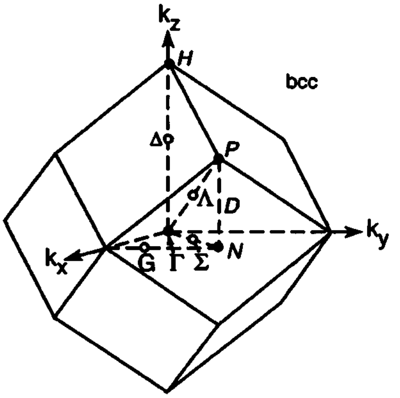


Figure 5.15 Brillouin zone of a bcc lattice in reciprocal space.

5.6 EXAMPLES OF BAND STRUCTURE DIAGRAMS

What does the electron band structure diagram of the first Brillouin zone look like when represented in two dimensions?

Figures 5.16 and 5.17 show the electron band structures of two real materials, copper and aluminium. Aluminium is seen to contain electrons which in their energy versus wave vector relationships are very close to free electrons.

5.7 CONCLUSIONS

How does the presence of a periodic potential affect the energies and behaviour of electrons?

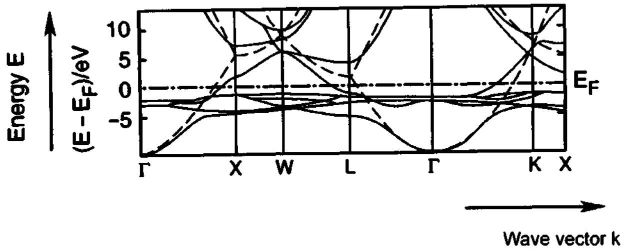


Figure 5.16 Electron band structure diagram of copper.

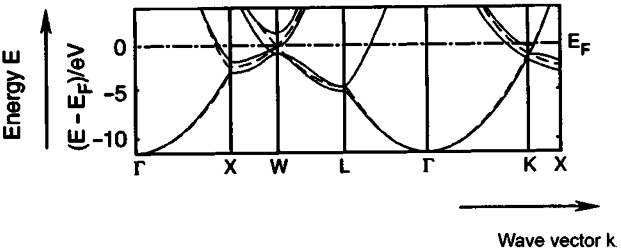


Figure 5.17 Electron band structure diagram of aluminium. Reproduced with permission from B. Segal, *Phys. Rev.*, **125**, 1962, p. 109.

In this chapter we have looked at the behaviour of electrons under the influence of a periodic potential due to the lattice ions. The calculations, although still oversimplified, produce some interesting results such as the existence of localized low energy states, and higher-energy itinerant electron states with the two groups of states separated by a region of forbidden energy known as the band gap. Two different approaches to describing the electrons have been discussed, one beginning from modifying the energy states in isolated atoms, the other beginning with free electrons and modifying their behaviour with the effects of a periodic potential. It is possible to distinguish between conductors and insulators on the basis of this model, the conductors have electrons occupying the higher-energy itinerant states, while the insulators have electrons confined to the localized energy states only.

We should now be in a position to understand the comment made in Section 2.1, that the differences in the electronic properties of materials are due more to the ionic lattice than to the electrons themselves. Since all electrons are identical the collective properties of the electrons are determined by the boundary conditions imposed on the electron wavefunction. The boundary conditions are periodic and are dictated by the ionic lattice. Without the lattice the electrons could have any energy value.

REFERENCES

1. H. Bethe (1928) *Ann. Physik* 87, 55.
2. L. Brillouin (1930) *J. Phys. Radium* 1, 377.

3. H. A. Kramers (1935) *Physica* 2, 483.
4. F. Bloch (1930) *Z. Physik* 59, 208.
5. R. Kronig and W. G. Penney (1931) *Proc. Roy. Soc.* A130, 499.
6. W. Shockley (1937) *Phys. Rev.* 52, 866.
7. F. Bloch (1928) *Z. Physik* 52, 555.
8. C. Herring (1940) *Phys. Rev.* 57, 1169.
9. L. Brillouin (1931) *Quantenstatistik*, Springer, Berlin.

FURTHER READING

R. H. Bube (1992) *Electrons in Solids: An Introductory Survey* 3rd edn, Academic Press, San Diego.

B. R. Coles and A. D. Caplin (1976) *The Electronic Structures of Solids*, Edward Arnold, London.

EXERCISES

Exercise 5.1 Effective mass of electrons in bands

Explain the physical reasons why electrons in energy bands can behave as if they have different masses. Derive an expression for the effective mass and explain its significance in terms of the curvature of the electron bands. Do the electrons actually have a different mass?

Exercise 5.2 Origin of electron bands in materials

Explain how electron energy bands arise in materials. First, beginning your discussion from the free electron approximation and secondly beginning your discussion from the tight-binding approximation. The periodic potential in a one-dimensional lattice of spacing a can be approximated by a square wave which has the value $V = -2eV$ at each atom and which changes to zero at a distance of $0.1a$ on either side of each atom. Estimate the width of the first energy gap in the electron energy spectrum.

Exercise 5.3 Number of conduction electrons in a Fermi sphere of known radius

In a simple cubic, quasi-free electron metal the spherical Fermi surface just touches the first Brillouin zone. Calculate the number of conduction electrons per atom in this metal.

Exercise 5.4. Boundary conditions and solutions of the wave equation

If an electron is confined inside a material, which is represented as a region with a constant potential V_0, show that a solution to the Schrödinger equation is

$$\Psi = A\exp(ikx) + B\exp(-ikx),$$

where $k^2 = 2m(E - V_0)\hbar^2$. What happens to the solutions to the wave equation when boundary conditions are applied?

Explain why the observable quantity associated with the wavefunction is the intensity $|\Psi|^2 (= \Psi^*\Psi)$ and not simply Ψ the wavefunction itself. What does $|\Psi|^2$ actually represent?

Exercise 5.5 Electrons in a periodic potential

Suppose that now the electron inside a material is subjected to a periodic potential with spatial periodicity (i.e. lattice parameter) a. Show that under these conditions the value of the wavefunction Ψ at any point x in real space must be equal to its value at $(x + na)$ where n is an integer (with the limitation on n that $x + na$ must lie within the periodic potential if it is of finite extent).

Furthermore, show that under these conditions the allowed values of k must necessarily be $k = 2n\pi/Na$ where n is an integer (in this case with the limitation that n cannot exceed the number of potential wells N in the lattice, or equivalently the number of lattice sites – compare, for example, with the allowed modes of vibration of a one-dimensional lattice). Explain the significance of this, and prove that $k = 0$ is not an allowed solution, except in the trivial case where A and B are both zero.

Exercise 5.6 Electron energy bands

Explain what is meant by: (i) energy band, (ii) band gap, (iii) conduction band, (iv) valence band. What causes electron energy bands to occur in materials and how do metals, semiconductors and insulators differ in terms of their electron band structures?

Band structure diagrams of two unknown materials are shown below. What can you tell about the nature of these materials from the diagrams? In the second material calculate the optical wavelengths at which you would expect the material to be transparent and the optical wavelengths at which you would expect it to be absorbent.

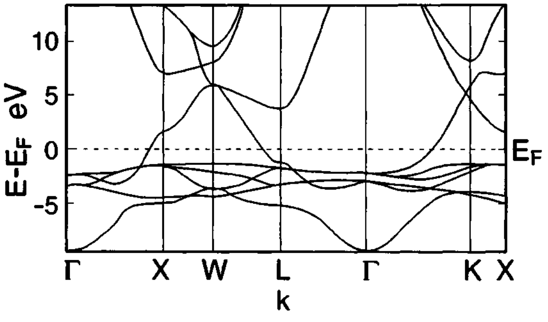


Figure 5.18 Band structure of unknown material.

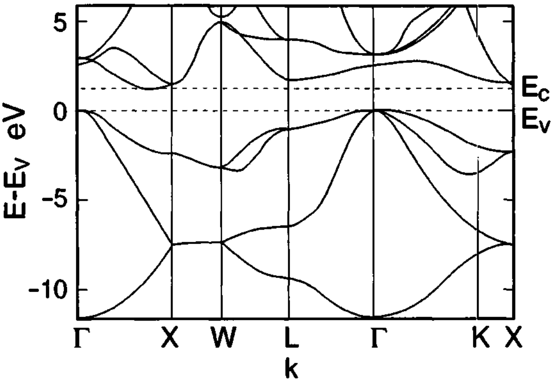


Figure 5.19. Band structure of unknown material.

PART TWO

PROPERTIES OF MATERIALS

6 ELECTRONIC PROPERTIES OF METALS

OBJECTIVE

In this chapter we bring together the basic concepts discussed in earlier chapters to provide a broad description of the electronic properties of metals. The main idea is that metals have some electrons which occupy the higher-energy 'free' electron levels, and can therefore migrate throughout the material. These are the so-called conduction electrons which contribute to both electrical and thermal conduction. These electrons also enable us to explain the optical properties of metals, in particular the high reflectance of metals in the visual range of the spectrum. Since it is the electrons which are close to the upper surface of the electron 'sea' which are most important in defining the electronic properties, we look at this 'Fermi surface' in greater detail here and discuss some of the methods of representing this important electronic characteristic of a metal.

6.1 ELECTRICAL CONDUCTIVITY OF METALS

How do we account for the range of conductivities of materials?

We have mentioned the wide range of observed conductivities in materials ranging from $10^{-15}\,\Omega^{-1}\,\mathrm{m}^{-1}$ in sulphur to $10^{8}\,\Omega^{-1}\,\mathrm{m}^{-1}$ in copper. In order that we can discuss this range of properties systematically, we shall divide the materials into conductors and insulators, with the semiconductors such as germanium and silicon being classed with the insulators.

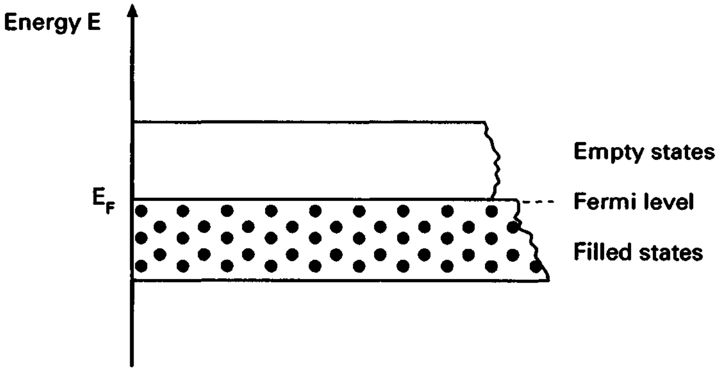


Figure 6.1 Schematic band structure diagram of a metal.

In this chapter we look at the metals. Metals are typically good electrical and thermal conductors and good reflectors of light in the visible spectrum. These properties are all due to the free, or more precisely quasi-free, electrons in the material. Metals contain a partially filled electron band through which the electrons can move relatively freely in order to conduct heat and electrical current, and the electrons in this band can absorb photons of low energies simply by being excited to slightly higher, unoccupied electron levels within the same energy band, later returning to lower unoccupied levels with the emission of a photon.

Figure 6.1 shows a very simple schematic diagram of the highest occupied energy band in a metal. This schematic tells us nothing about the dependence of energy E on the wave vector k in this material. However, we should note that the Fermi energy lies in the middle of an allowed band in a metal and this means that a metal has a Fermi surface, that is a connected set of highest occupied energy levels in k-space.

6.2 REFLECTANCE AND ABSORPTION

How are the various optical properties of materials related to each other and to the electrical properties?

We have mentioned in Section 1.4 that the optical properties of materials can be represented by two constants, either n and k, or alternatively ε_1 and ε_2, as discussed in Section 3.4.1. The dielectric 'constant' of the material is $\varepsilon = \varepsilon_0(\varepsilon_1 + i\varepsilon_2)$, where ε_0 is the permittivity of free space. The two components of the complex relative dielectric constant $\varepsilon_r = \varepsilon_1 + i\varepsilon_2$ are related to n and k by

$$\varepsilon_1 = n^2 - k^2, \tag{6.1}$$

$$\varepsilon_2 = 2nk. \tag{6.2}$$

Both ε_1 and ε_2 are dimensionless. The term ε_2 is known as the absorption. It is also related to the electrical conductivity $\sigma(\omega)$ by

$$\varepsilon_2(\omega) = \frac{\sigma(\omega)}{\omega\varepsilon_0} \tag{6.3}$$

at a given frequency of excitation ω. This expression was given by Drude in the classical theory of electrons in metals.

This means that a good electrical conductor, which has a high value of $\sigma(\omega)$, is a good absorber of light. A good electrical conductor is also known to be a good reflector of light (i.e. high R). This at first seems to be a contradiction. How can a good absorber also be a good reflector? We need to resolve this seeming contradiction immediately. The reflectance R is related to the components ε_1 and ε_2 of the dielectric constant (see Section 3.4.2) by the relation,

$$R = \frac{\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + 1 - \sqrt{2(\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + \varepsilon_1)}}{\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + 1 + \sqrt{2(\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + \varepsilon_1)}}. \tag{6.4}$$

We know that high k leads to high ε_2 and high R. What do we mean by absorption in this context? We really mean that the light does not penetrate the solid very far. Clearly if the light is either reflected or transmitted, then with high absorption the light is not transmitted and so by conservation of energy the light must eventually be reflected back. The mechanism for this process is absorption of light by free electrons, excitation of the electrons to higher-energy states, and then de-excitation of the electrons with the emission of photons. In short, absorption followed by immediate re-emission of the light gives reflection.

6.2.1 Optical properties and electron band structure

How can the optical properties be related to the electronic structure of the material?

Having stated this relation between absorption and reflectance we need to give a brief explanation in terms of the electron theory. In order to absorb incident light the electrons must have available states which they can move to at an energy ΔE above their present energy state, where

$$\Delta E = \hbar\omega, \tag{6.5}$$

ω is the frequency of the light and $\hbar$ is Planck's constant divided by 2π.

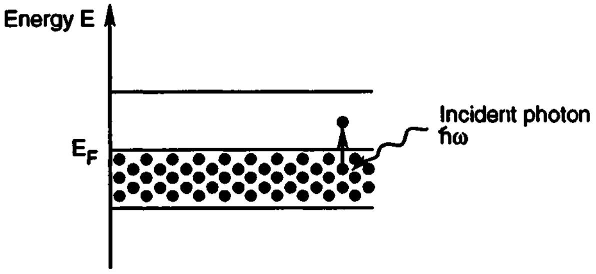
Immediately then it is apparent that an insulator with a filled valence band, empty conduction band and band gap $E_g > \hbar\omega$ cannot absorb the photon with energy $\hbar\omega$, because there are no available energy states into which the electrons can be excited which correspond to the correct gain in energy.

Metals, however, do have such energy states available at low energies because electrons can be stimulated to slightly higher available levels within the same band (intraband absorption) as shown in Fig. 6.2. So the metals have a high absorption at low frequencies. Once the electrons in a metal have absorbed the light they can return to their available lower energy state with the emission of a photon of identical energy, or in some cases of lower energy. This leads to a high reflection coefficient R in metals.

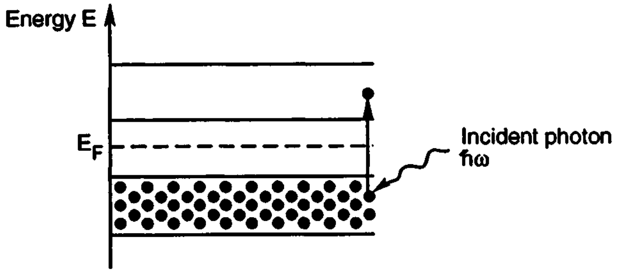
Insulators on the other hand have low absorption at low frequencies, but as the energy of the incident radiation increases a frequency ω_0 is reached at which the energy of the incident photons equals the energy of the band gap,

$$\hbar\omega_0 = E_g. \tag{6.6}$$

Beyond this frequency an insulator has a high absorption. The mechanism by which the electrons absorb energy is shown in Fig. 6.2 where the electrons are stimulated from the valence band to the conduction band. Each semiconductor or insulator has its own characteristic energy gap E_g. This means that if we look at the reflection or absorption spectrum of a material we can soon tell whether it is a metal, semiconductor, or insulator, because high reflectance at low energies is a property of metals, whereas a low reflectance at low energies combined with high reflectance at high energies is a property of insulators.



Metal



Insulator

Figure 6.2 Schematic metal and insulator band structure diagrams showing absorption of photons.

6.3 THE FERMI SURFACE

How can we describe the Fermi level in three dimensions?

We have discussed the concept of a Fermi energy E_F and a Fermi level which is the highest occupied energy state in a metal in its ground state. Now we will generalize this idea further to the Fermi surface [1]. Excellent introductions to the Fermi surface have been given by Ziman [2] and by Mackintosh [3]. The Fermi surface is the plot of the Fermi level in three-dimensional k-space. The volume contained within the Fermi surface represents all the occupied energy levels when the material is in its ground state.

In the simplest case the Fermi surface for free electrons is a sphere since the energy is given by an expression which depends on the sum of the squares of the k vectors in each direction:

$$E = \frac{\hbar^2}{2m}\,(k_x^2 + k_y^2 + k_z^2). \tag{6.7}$$

Since this expression for energy is not directionally dependent, the Fermi surface for free electrons will be spherical in k-space.

In other cases, in particular in real metals, the Fermi surface is not exactly spherical in k-space, but for most metals which have upper electron states which are quasi-free, the Fermi surface is more or less spherical in shape when plotted in the extended zone representation. There will, of course, be local perturbations due to the lattice, because of Bragg reflection at the Brillouin zone boundaries, as discussed in Section 5.5.2. This results in the Fermi surface having bulging contours close to the zone boundaries, as in the case of copper, silver and gold, for example. The forms of these Fermi surfaces have been discussed in detail by Cracknell and Wong [4].

6.3.1 The Fermi surface in the reduced zone scheme

Can the entire description of the allowed energy levels be represented in the first Brillouin zone alone?

As a consequence of the periodicity of the lattice we have shown in the previous chapter that the reciprocal lattice in k-space can be subdivided into Brillouin zones. In the one-dimensional case, the first Brillouin zone extends from $-(\pi/a)$ to where $+(\pi/a)$ is the lattice parameter. A symmetry principle comes into play here whereby the energy of a particular band $E(k)$ is a periodic function of the reciprocal lattice

$$E(k) = E\left(k \pm \frac{\pi}{a}\right). \tag{6.8}$$

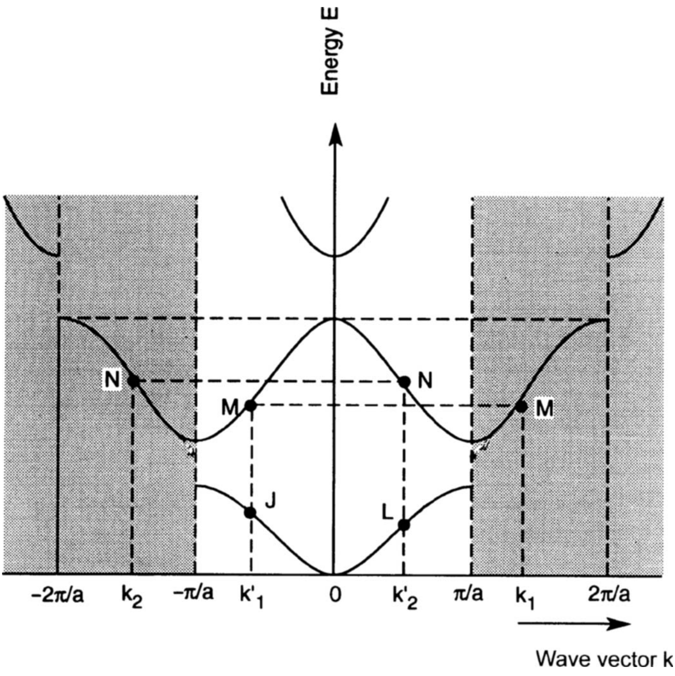


Figure 6.3 Extended- and reduced-zone representations of electron energy levels.

This leads to the periodic-zone scheme shown in Fig. 6.3 in which each energy state in a higher Brillouin zone can be mapped back to an equivalent point in the first zone.

The important result here is that the entire energy band structure of a material can be represented in the first Brillouin zone by transforming all the allowed energy states $E(k + n\pi/a)$ to the equivalent point in the first zone using the reduced-zone scheme:

$$E\left(k + \frac{(2n-1)\pi}{a}\right) \rightarrow E\left(\frac{\pi}{a} - k\right) \tag{6.9}$$

$$E\left(k + \frac{2n\pi}{a}\right) \rightarrow E(k). \tag{6.10}$$

Therefore every point in every zone has an equivalent point in the first Brillouin zone.

6.3.2 Advantages and disadvantages of the reduced zone scheme

Does the compact representation of the reduced zone scheme lead to any disadvantages in visualizing the electronic structure of a material?

An obvious advantage of the reduced zone representation is that the entire electron band structure of the material can be plotted within one Brillouin zone. This allows for compact representation of the electronic properties, particularly when we are dealing with three dimensions. A disadvantage is that the electron energy levels, which may appear to be relatively simple in the extended zone scheme, can become very complicated to visualize in the reduced zone scheme. A good example of this is

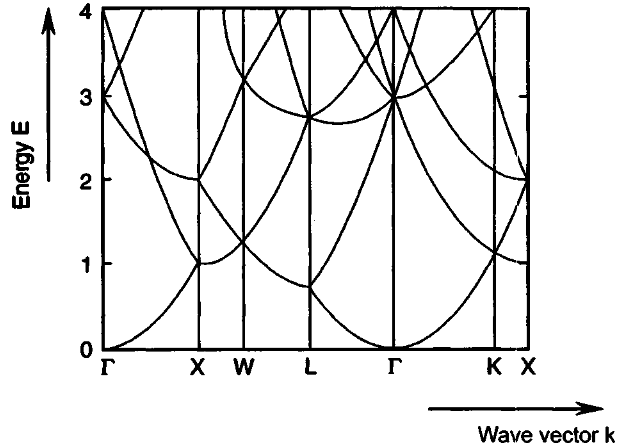


Figure 6.4 Free electron parabola represented in the reduced zone scheme in the first zone of an fcc lattice, which shows how a very simple situation can appear to be superficially complicated under this transformation. Reproduced with permission from R. G. Chambers, *Electronics in Metals and Semiconductors*, published by Chapman & Hall, 1990.

the free electron parabola which, when mapped onto the first zone of an fcc lattice (in the absence of any periodic potential), is transformed as shown in Fig. 6.4. This corresponds to the 'empty lattice' approximation discussed by Shockley [5]. Despite the extremely simple nature of the relationship between E and k in the case of free electrons, the representation in Fig. 6.4 looks superficially quite complex. Even weak Bragg reflection at the Brillouin zone boundaries can make this diagram appear much more complicated.

A Fermi surface which also may have a relatively simple distorted spherical shape in the extended zone scheme can become an extremely complex shape in the reduced zone scheme. An example of this is the Fermi surface of aluminium which is shown later (see Section 6.3.5).

6.3.3 Free electron Fermi surface

What does the free electron Fermi surface look like in the extended zone scheme in two dimensions?

For simplicity we will begin by considering Fermi surfaces in a hypothetical two-dimensional solid. This is easier to represent and discuss than in three dimensions. Once the general idea has been expounded we will go on to consider examples in three dimensions.

We have shown above that the Fermi surface of free electrons in k-space is spherical, and hence in two dimensions it is circular, if we have no Bragg reflection (which is equivalent to an empty lattice). When the Fermi surface is entirely contained within the first Brillouin zone, the Fermi surface in the reduced zone scheme is also circular. For simplicity we consider this situation in relation to a square lattice as shown in Fig. 6.5.

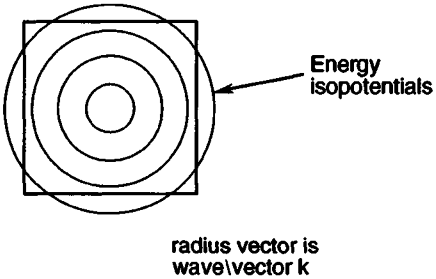


Figure 6.5 Free electron Fermi 'sphere' in two-dimensional k-space.

Now let us consider a metal with more electrons so that not all electron states can be contained in the first Brillouin zone. In this case in the extended zone scheme the Fermi surface is again circular, as shown in Fig. 6.6, but the Fermi surface extends beyond the first zone. Now in the reduced zone scheme the situation appears more complicated, as shown in Fig. 6.7.

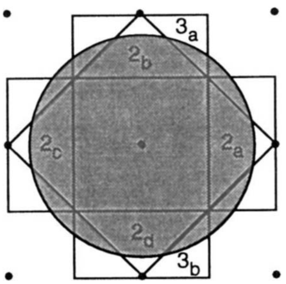


Figure 6.6 Free electron Fermi 'sphere' in two-dimensional *k*-space extending beyond the first zone.

When the free electron Fermi sphere is folded back into the first zone in the reduced zone scheme, the appearance of the surface is complicated by the transformation.

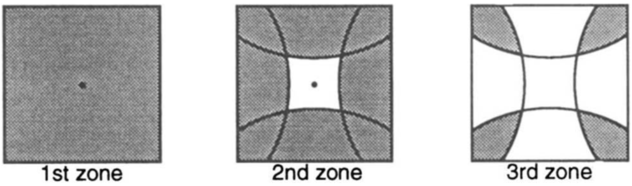


Figure 6.7 Free electron Fermi 'sphere' mapped back onto the first zone in two-dimensional *k*-space.

The important point to note here is that for free electrons:

1 The Fermi surface in *k*-space in the extended scheme is perfectly spherical.
2 When projected back into the first zone using the reduced zone scheme the spherical surface can look very different.

6.3.4 Fermi surface in a periodic potential

How does the presence of a periodic potential change the form of the Fermi surface? Once we have introduced a weak periodic potential the Fermi surface in *k*-space becomes distorted by Bragg reflection even in the extended zone scheme. In our two-dimensional example the distortion from circular appears as shown. The amount of distortion from circular energy levels depends on the strength of the Bragg reflection at the zone boundaries, which alters the relationship between E and k.

It can be seen from the one-dimensional section of the E versus k diagram that an energy gap arises at the Brillouin zone boundary. This is a direct result of the periodic potential and the size of the gap depends on the strength of the periodic potential wells. The energy gap is caused by Bragg reflection by the lattice of electrons with wave vectors close to the zone boundary. Clearly then not all the electrons are reflected in this way.

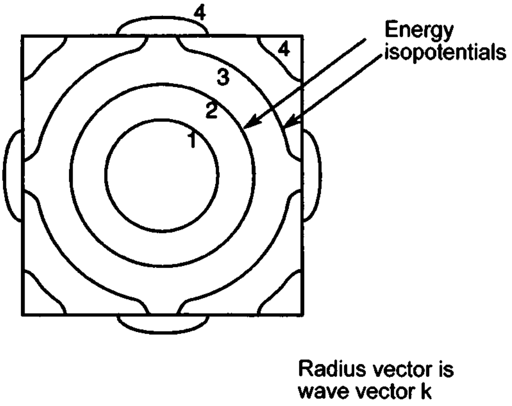


Figure 6.8 Distorted two-dimensional Fermi 'sphere' in the first Brillouin zone resulting from the effect of a periodic potential.

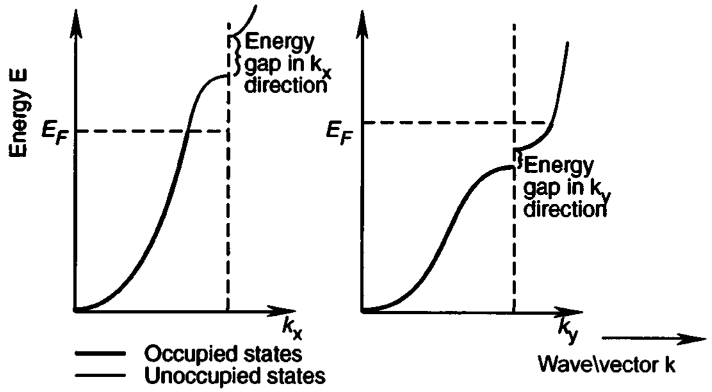


Figure 6.9 One-dimensional electron energy levels in a rectangular first Brillouin zone showing band gaps at the zone boundary. If the reciprocal lattice is rectangular instead of square the energy gaps occur at different energy levels in the different directions, as shown here. This means that the Fermi surface can touch the zone boundary in one direction (in this case the k_y direction) but not in the other direction (in this case the k_z direction).

If we only have a small number of valence electrons, for example one electron per atom in the monovalent metals Cu, Au, Ag, then the Fermi surface is still relatively simple. Of course, if the Fermi surface lies well inside the first zone (as in level 1 in Fig. 6.8) then the deviation from the circular will be insignificant. However, as the band expands to fill the first zone and the Fermi surface gets closer to the zone boundary it will distort to meet the zone faces (e.g. level 3). Larger numbers of electrons will push the Fermi level into the second zone (e.g. level 4).

Once we have parts of the Fermi surface in the second zone, we usually fold these back onto the first zone to obtain a more compact reduced zone representation. It is at this stage that the representation of an essentially rather simple

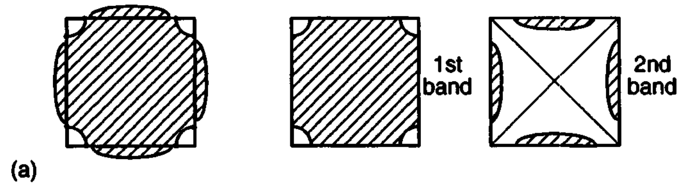


Figure 6.10 Extended zone representation of two-dimensional distorted Fermi 'sphere'.

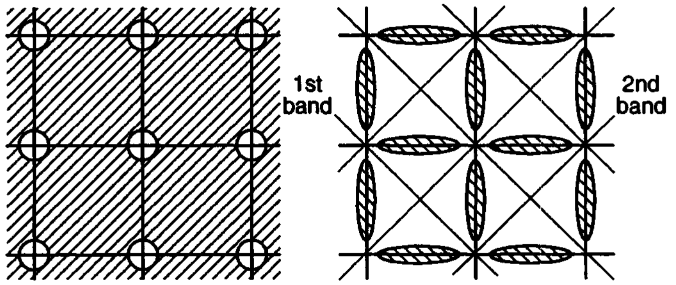


Figure 6.11 Reduced zone representation of 2-D distorted Fermi 'sphere'.

Fermi surface can take on a very complicated appearance as shown in Figs. 6.10 and 6.11.

6.3.5 Three-dimensional Fermi surfaces of metals

How does the Fermi surface appear in three dimensions?

Having looked briefly at the problem of mapping the Fermi surface in two-dimensional k-space we now have a much better idea of what to expect in three dimensions when we look at the Fermi surfaces of real metals.

The first metal for which the Fermi surface was completely mapped in k-space was copper. This work was performed by Pippard [6, 7] using measurements of the anomalous skin effect. The Fermi surface of copper is particularly easy to visualize because it all lies in the first Brillouin zone. From our discussion of the two-dimensional examples above, in which the Fermi surface that was contained entirely within the first zone was circular, we may expect the Fermi surface of copper to be approximately spherical. This is indeed the case: the surface does, however, have necks extending towards the Brillouin zone boundaries. This is caused by Bragg reflection where the Fermi surface comes close to the zone boundary. The shape of the Fermi surface resembles a 'diving sphere' with portholes. This is shown in Fig. 6.12. Both silver and gold have Fermi surfaces with similar shapes to that of copper.

In divalent calcium, which has of course two outer electrons per atom, the Fermi surface extends into the second Brillouin zone. Again in the extended zone scheme, the Fermi surface is approximately spherical. However, when folded back onto the first zone in the reduced zone scheme it looks quite different, resembling

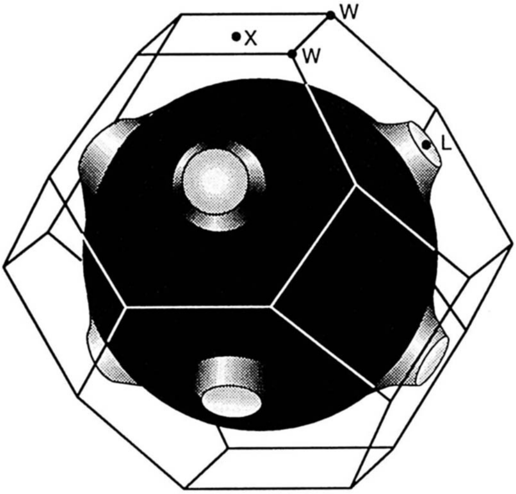


Figure 6.12 Three-dimensional Fermi surface of copper.

a 'coronet'. Divalent beryllium also has a coronet-shaped Fermi surface; however, since it solidifies with hexagonal symmetry the surface in reciprocal space has sixfold symmetry.

Aluminium, being trivalent, has three outer electrons. Its electron bands are very close to free electron parabolae, and its Fermi surface if plotted in the extended zone scheme is fairly simple in shape, being almost spherical and extending into the third Brillouin zone. Again once it is folded back onto the first zone it takes a very different form, which has come to be known as the 'monster'. This is shown in Fig. 6.13.

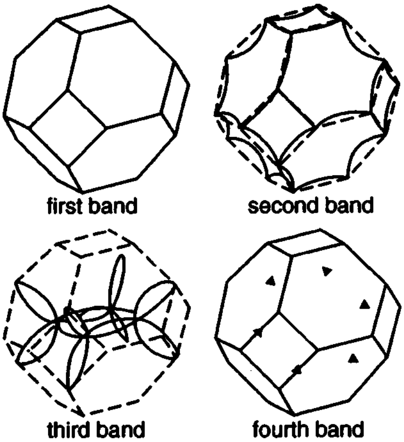


Figure 6.13 Three-dimensional Fermi surface of aluminium.

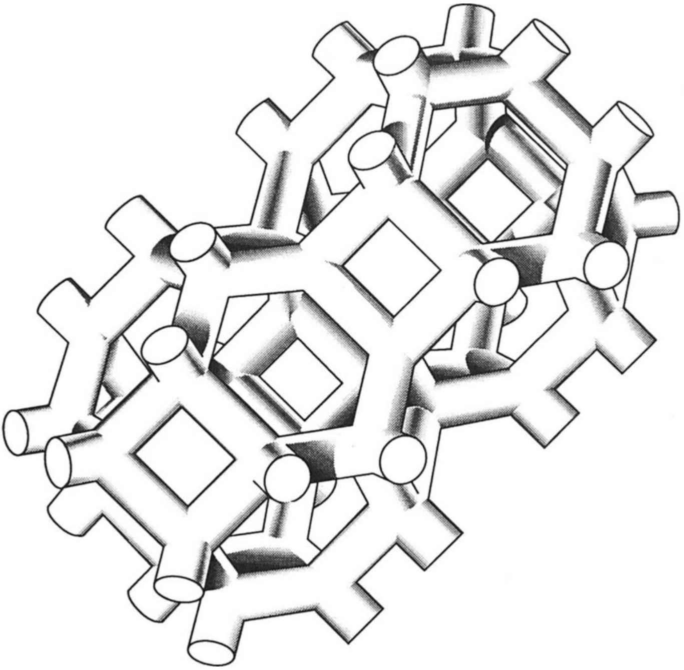


Figure 6.14 Three-dimensional Fermi surface of lead.

Lead has four valence electrons and its Fermi surface extends into the fourth Brillouin zone. In this case the Fermi surface in the first zone is extremely complicated forming a 'pipeline maze', as shown in Fig. 6.14.

6.3.6 Methods of determining the Fermi surface

How can we examine the shape of the Fermi surface of a metal?

First we need to understand why measurements of the Fermi surface are important. Recall that the classical Drude metal of electrons in solids failed over the prediction of the specific heat capacity of metals. This was because even the free electrons in the conduction band of a metal are not able to absorb thermal energy unless they are within an energy k_BT of the Fermi surface. This means that most of the electronic properties of a metal are determined by electrons lying at, or just below, the Fermi surface. Clearly the electrons close to the Fermi surface are most important in determining those properties of a metal which depend on the conduction electrons. We conclude therefore that by knowing the details of the Fermi surface we can make predictions about many of the properties of a metal.

There are a number of different measurements that can be made which give information about the Fermi surface. For a detailed description of these consult Ashcroft and Mermin [8] who have devoted an entire chapter to methods of measuring the Fermi surface. We list here only the most important techniques used:

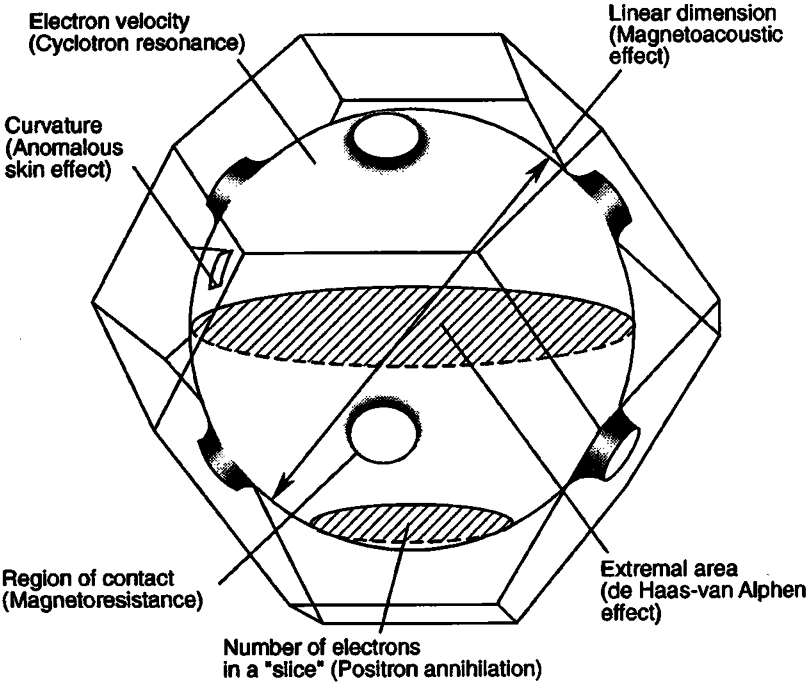


Figure 6.15 Information that can be obtained about the Fermi surface from various physical measurements as described by Mackintosh [3].

(i) de Haas–van Alphen effect
(ii) Magnetoacoustic effect
(iii) Ultrasonic attenuation
(iv) Anomalous skin effect
(v) Magnetoresistance
(vi) Cyclotron resonance
(vii) Positron annihilation.

Each of these measurements gives different information about the Fermi surface. These are depicted in Fig. 6.15.

The de Haas–van Alphen effect [9] is the most important technique which is used to probe the Fermi surface. From these measurements the extremal area (i.e. largest cross-sectional area) of the Fermi 'sphere' can be found along different directions. The magnetoacoustic effect enables the linear dimension (i.e. largest diameter) of the Fermi 'sphere' to be calculated. In the absence of a magnetic field the conventional ultrasonic attenuation also gives information about the Fermi surface; however, in this case the interpretation in terms of Fermi surface geometry is more complicated.

The anomalous skin effect, which can be used to measure the curvature of the Fermi surface in *k*-space, is one of the oldest techniques used for Fermi surface measurements. The penetration of the magnetic field into a solid at higher frequencies deviates from the classical skin effect equation and it can be shown that the field penetration becomes dependent entirely on certain features of the Fermi surface geometry (e.g. the curvature) at sufficiently high frequencies.

Magnetoresistance measurements, that is the dependence of electrical resistance on magnetic field, can be used to find the region of contact of the Fermi surface with the Brillouin zone boundary since the magnitude of this contact affects the conductivity. Cyclotron resonance, the circular motion of a charged particle moving in a plane normal to a magnetic field, can also be used to investigate the Fermi surface. Specifically, it can be used to find the electron velocity on the Fermi surface.

Finally, positron annihilation can be used to find the number of electrons in a two-dimensional slice through the Fermi 'sphere.' When the material is bombarded with positrons the electrons annihilate the positrons yielding two photons. The momentum of the emitted photons can be used to determine the momentum distribution of the electrons in the metal. This can then be used to indicate the number of electrons in a given slice through *k*-space.

6.3.7 The de Haas–van Alphen effect

How does the differential susceptibility of a metal depend on the applied field strength?

The de Haas–van Alphen effect is the most important technique used for obtaining information about the Fermi surfaces of metals. At low temperatures and under high magnetic fields (typically $H > 5\,\text{kOe}$ or $400\,\text{kA}\,\text{m}^{-1}$) it was found that the differential susceptibility dM/dH of metals was dependent on the field strength in an oscillatory manner.

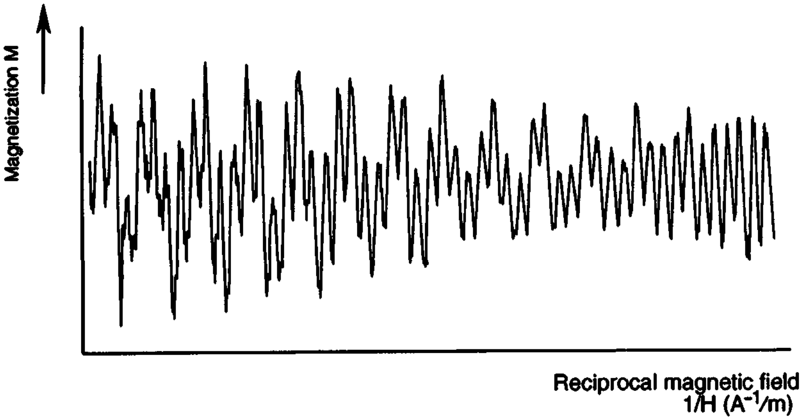


Figure 6.16 de Haas–van Alphen oscillations. Reproduced with permission from R. G. Chambers, *Electrons in Metals and Semiconductors*, published by Chapman & Hall, 1990.

When the differential susceptibility is plotted against $1/H$ this periodic dependence is shown most clearly, although often two or more periods are superposed, as shown in Fig. 6.16. Similar behaviour has been observed in the conductivity and the magnetostriction. The former is known as the Shubnikov–de Haas effect. Two methods are widely employed to measure these de Haas–van Alphen oscillations. One uses a torque magnetometer and simply measures the oscillations in angular position of a sample of the metal as the field strength H and magnetization M are increased.

The second method uses field pulses and measures the voltage induced in a flux coil wound on the sample. Since the voltage from the flux coil will be $V = NA\,\mathrm{d}B/\mathrm{d}t$ where N is the number of turns and A is the cross-sectional area it follows that

$$V = \mu_0 NA\left(1 + \frac{\mathrm{d}M}{\mathrm{d}H}\right)\frac{\mathrm{d}H}{\mathrm{d}t}, \tag{6.11}$$

and the rate of change of magnetic field $\mathrm{d}H/\mathrm{d}t$ is known, so that $\mathrm{d}M/\mathrm{d}H$ can be calculated.

6.3.8 Mechanism of the de Haas–van Alphen effect

Why do the electrons cause these oscillations in the sucsceptibility?

The energy of free electrons at the Fermi surface can be equated to the cyclotron resonance frequency ω_c,

$$E_f = \frac{\hbar^2 k_F^2}{2m} \tag{6.12}$$

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega_c \tag{6.13}$$

and knowing that the cyclotron frequency depends on the magnetic field according to the relation $\omega_c = eB/m$, it follows that the energy levels are

$$E_n = \left(n + \frac{1}{2}\right)\frac{\hbar eB}{m}. \tag{6.14}$$

The effect of the applied magnetic field is to cause the free electron energy levels to group together with energies around $(n + 1/2)\hbar\omega_c$. Considering the free electrons to be restricted to a plane, the density of states at $\hbar\omega_c$ is

$$D(\hbar\omega_c) = \frac{\pi k^2 m\omega_c}{\hbar} = \frac{\pi k^2 eB}{\hbar}. \tag{6.15}$$

Consequently, the density of states at the Fermi level will reach a maximum each time the energy levels grouped around $(n + 1/2)\hbar\omega_c$ have an energy equal to that of the Fermi energy

$$\left(n + \frac{1}{2}\right)\frac{\hbar eB}{m} = \frac{\hbar^2 k_F^2}{2m} \tag{6.16}$$

and this will occur periodically as the strength of the magnetic field is changed.

Both magnetic susceptibility and electrical conductivity in metals depend on the density of states at the Fermi level. It, therefore, follows that both of these quantities will vary periodically as the applied field is changed. Specifically, maxima in both quantities may be expected when

$$B = \frac{\hbar k_F^2}{2e} \frac{1}{(n + 1/2)}. \tag{6.17}$$

The free electron Fermi surface will have an 'extremal area' in k-space of

$$A_{\text{ext}} = \pi k_F^2. \tag{6.18}$$

This 'area' is in units of m^{-2}. Therefore if B_n and B_{n+1} are values of the magnetic induction at which successive peaks occur in differential susceptibility (de Haas–van Alphen) or conductivity (de Haas–Shubnikov) then

$$\frac{1}{B_{n+1}} - \frac{1}{B_n} = \frac{2\pi e}{\hbar A_{\text{ext}}}, \tag{6.19}$$

and this provides a simple method of determining the 'extremal area' πk_F^2 at the Fermi surface along the direction of interest. The measurement can be repeated in different directions to gain more complete information about the form of the Fermi surface in three dimensions.

REFERENCES

1. E. Fermi (1928) *Zeit. für. Physik* 48, 73.
2. J. M. Ziman (1963) *Electrons in Metals – A Short Guide to the Fermi Surface*, Taylor and Francis, London.
3. A. Mackintosh (1963) *Scientific American* 209, 110.
4. A. P. Cracknell and K. C. Wong (1973) *The Fermi Surface*, Clarendon Press, Oxford.
5. W. Shockley (1937) *Phys. Rev.* 52, 866.
6. A. B. Pippard (1957) *Phil. Trans. Roy. Soc.* A250, 325.
7. A. B. Pippard (1960) *Rep. Prog. Phys.* 23.
8. N. W. Ashcroft and N. D. Mermin (1976) *Solid State Physics*, Holt, Rinehart & Winston, New York, p. 76.
9. W. J. de Haas and P. M. van Alphen (1930) *Proc. Neth. Roy. Acad. Sci.* 33, 1106.

FURTHER READING

R. G. Chambers (1990) *Electrons in Metals and Semiconductors*, Chapman & Hall, London.

G. Lehmann and P. Ziesche (1990) *Electronic Properties of Metals*, Elsevier, Amsterdam.

N. F. Mott and E. A. Davis (1971) *Electronic Processes in Non-crystalline Materials*, Clarendon Press, Oxford.

V. L. Moruzzi, J. F. Janak and A. R. Williams (1978) *Calculated Electronic Properties of Metals*, Pergamon, Oxford.

J. S. Dugdale (1976) *The Electrical Properties of Metals and Alloys*, Edward Arnold and Sons, London.

EXERCISES

Exercise 6.1 Brillouin zones in a two-dimensional lattice
Make a plot of the first two Brillouin zones of a rectangular two-dimensional lattice with unit vectors along the x and y directions of $a = 0.2\,\text{nm}$, and $b = 0.4\,\text{nm}$. Give its dimensions in m^{-1}; calculate the radius of the free electron Fermi sphere if the atom has valence 1; draw this sphere on the first Brillouin zone; and show the electron band structure for both the first and second energy bands, assuming there is a small gap at the zone boundary.

Exercise 6.2 Number of k states in reciprocal space
Show that the number of different k states in the reciprocal space of a simple cubic lattice is equal to the number of lattice sites.

Exercise 6.3 Fermi energy of sodium and aluminium
Assuming that the free electron model applies, calculate the Fermi energy of body-centred cubic Na and face-centred cubic Al. The dimensions of the cubic unit cells in the crystal lattices are 0.43 nm and 0.40 nm respectively.

Exercise 6.4 Brillouin zones
Assuming idealized free electron behaviour in a three-dimensional simple cubic lattice, calculate the ratio of the energy of an electron at the corner of the first Brillouin zone to the energy of an electron at the midpoint of a face of the first Brillouin zone.

If a metal has a cubic lattice with parameter of 0.3 nm and its Fermi surface just touches the Brillouin zone boundary, determine the Fermi energy, the number of conduction electron states $N_0(E)$ per unit volume, and the density of states $D(E)$ per unit volume at the Fermi level.

Exercise 6.5 Electron density of states at the Fermi level
If copper has a Fermi energy of 7 eV, calculate the density of states $D(E)$ (i.e. the number of states in either a 1 J or a 1 eV interval of energy) in $1\,\text{m}^3$ of the material at the Fermi level, assuming an idealized free electron model.

Again assuming conduction electrons behave as idealized free electrons, calculate the Fermi energy for a material such as silver which has 6×10^{28} conduction electrons per cubic metre.

Exercise 6.6 The de Haas–van Alphen effect
If the susceptibility of gold and aluminium exhibit de Haas–van Alphen periodicities in a field of $2 \times 10^{-5}\,\text{T}^{-1}$ and $1 \times 10^{-5}\,\text{T}^{-1}$, respectively, calculate the extremal area of the Fermi surface normal to the field, the wave vector k_F in m^{-1} at the Fermi surface and the Fermi energy E_F in electron volts, assuming that both metals can be treated as 'free-electron-like'.

7 ELECTRONIC PROPERTIES OF SEMICONDUCTORS

OBJECTIVE

This chapter discusses the electron band structure of semiconductors and shows how the occupancy of the electron energy levels in these materials is fundamentally different from metals. The reason for this is that in semiconductors and insulators in their lowest-energy state, the electron bands are either filled or empty. This means that it is very difficult for the electrons to move under the action of an electric field because it would result in an increase in energy, and such energy states are not immediately available. Hence the conductivity is low. A vast range of semiconductor devices based on the properties of semiconductor junctions exist. These include devices which are designed for particular current–voltage characteristics, such as diodes and transistors, and optoelectronic devices such as photodetectors and photoemitters. The principles of their operation remain broadly the same. We will touch on the main concepts required for understanding them in this chapter. We will look at some of these semiconductor devices in more detail when we deal with specific applications in Chapters 11–15.

7.1 ELECTRON BAND STRUCTURES OF SEMICONDUCTORS

How does the electron band structure of a semiconductor differ from that of a metal? The electronic band structures of semiconductors and insulators are fundamentally different from those of metals because of the existence of the band gap which lies between a filled valence band and an empty conduction band. Although the energy of the electrons varies with the wave vector k, a more simplified band structure representation is often used for semiconductors. This is the 'flat band' approach which merely represents the allowed energy levels without reference to the corresponding values of k. This approach, which is adequate for most purposes, is used widely here and in subsequent discussion of the electronic structure of semiconductors. One aspect which the flat band model does not represent is the difference between direct and indirect band gap semiconductors. The direct band gap materials, in which the top of the valence band and the bottom of the conduction band are located at the same point in k-space, are very important in optoelectronic applications.

The electronic theory of semiconductors was first worked out by Wilson [1, 2]. The initial concept was that the semiconductors and insulators have highest occupied electron states which are localized in the ionic potentials, as shown for example in Fig. 5.1. These cannot therefore contribute to electrical conduction throughout the material. A better concept is simply that even if the top-most occupied band extends throughout the material the electrons cannot conduct if it is completely filled. In very simple terms then the band structures of metals, semiconductors, and insulators can be represented as shown in Fig. 7.1.

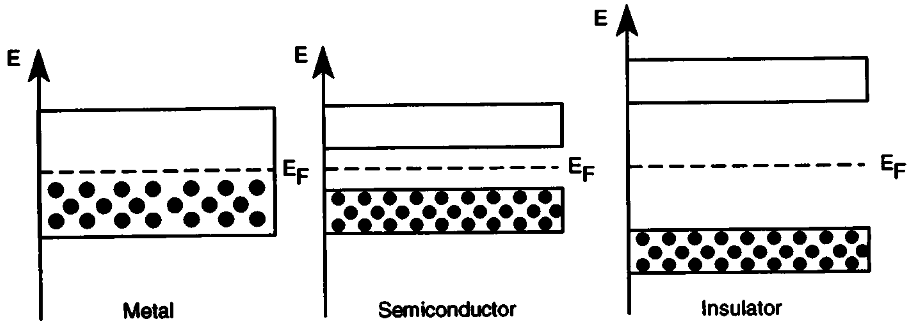


Figure 7.1 Simplified band structure diagrams of a metal, semiconductor and insulator. Typical values of the band gap are 0 eV in metals, 0.5–5.0 eV in semiconductors and 5 eV or greater in insulators.

Table 7.1 Typical band gaps of semiconductors.

Material	Band gap E_g (eV) at 0 K	Direct/indirect gap
Ge	0.75	Indirect
Si	1.17	Indirect
GaAs	1.5	Direct
GaP	2.32	Indirect
InAs	0.43	Direct
GaSb	0.81	Direct
InSb	0.23	Direct
SiC	3.0	Indirect
Te	0.33	Direct

1 eV = 1.602×10^{19} J, visible light range is 2–3 eV.

Typical values of the band gap in various semiconductors and insulators are shown in Table 7.1.

7.1.1 Band structure diagrams

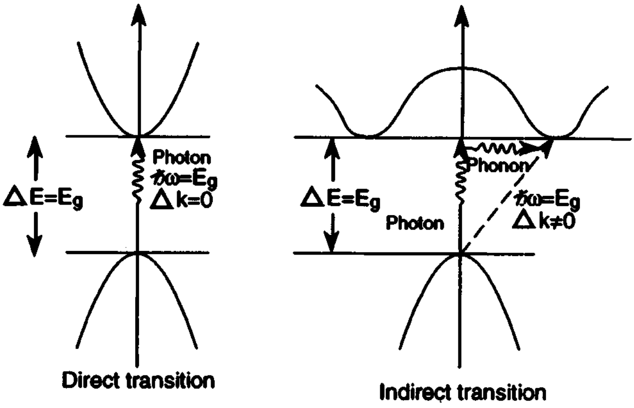
What band structure representation is used to interpret and predict the electronic properties of semiconductors?

Many of the electronic properties of semiconductors can be described by reference to the above simplified energy band diagrams. In fact you will find most textbooks base their whole discussion on these diagrams. However, the true band structure diagrams (energy versus wave vector plots) are much more complicated [3–5].

7.1.2 Direct and indirect band gaps

What do we mean by direct and indirect band gap semiconductors?

In some cases the top of the valence band and the bottom of the conduction band of a semiconductor lie at different points in *k*-space. This is called an indirect band gap. In other cases they lie at the same point in *k*-space. This is called a direct band gap. These two cases are shown schematically in Fig. 7.2. The presence of a direct band gap has important consequences for optical applications of a semiconductor as we shall see later, because the probability of electronic transitions across the band gap is higher in materials with a direct band gap.



Direct transition

Indirect transition

Figure 7.2 Direct and indirect band gaps in a semiconductor.

Since we have already learned how to interpret these diagrams in the case of metals, we shall look at the band structures of germanium, silicon, and gallium arsenide in *k*-space. These are shown in Figs. 7.3, 7.4, and 7.5 respectively. In these diagrams, the band gaps are clearly shown and have the values 0.7 eV in germanium, 1.1 eV in silicon and 1.5 eV in gallium arsenide. Notice that both

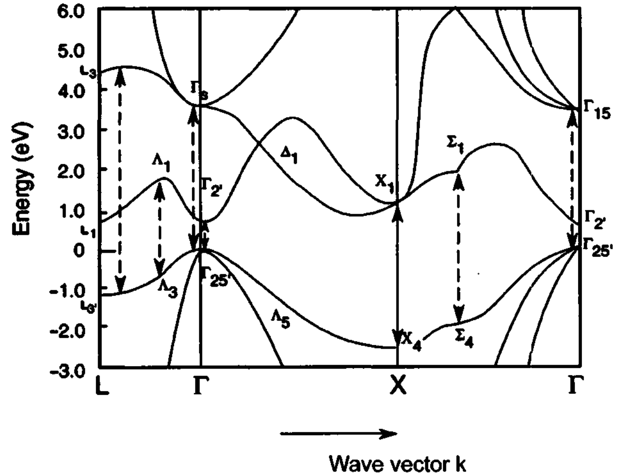


Figure 7.3 Band structure diagram of germanium. Reproduced with permission from D. Brust, *Phys. Rev.*, **A134**, 1964, p. 1337.

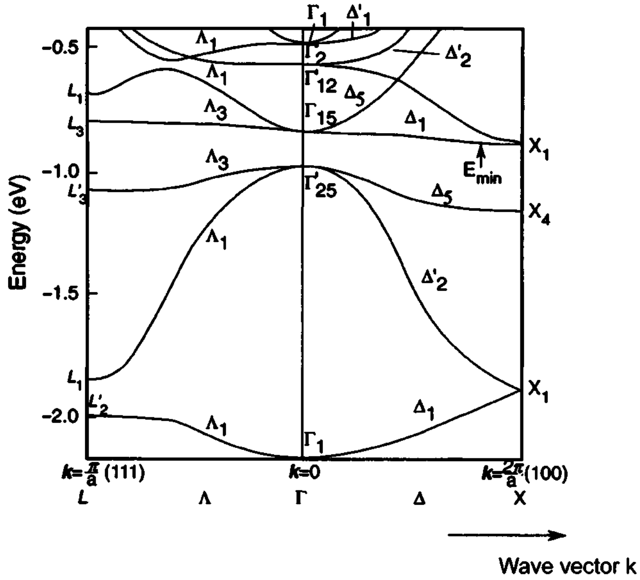


Figure 7.4 Band structure diagram of silicon.

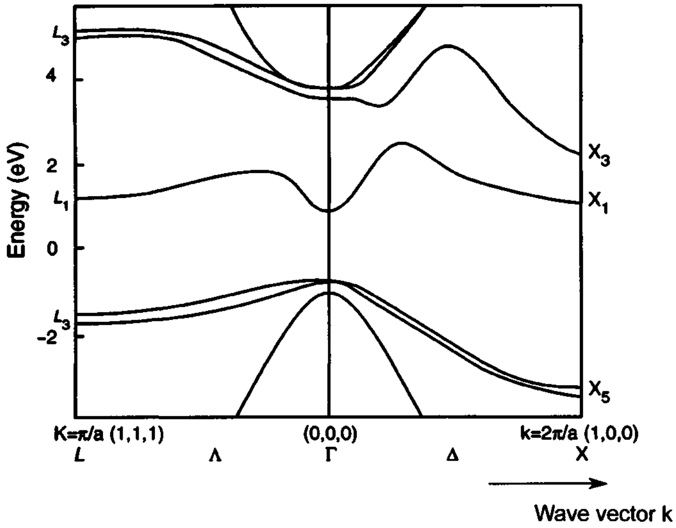


Figure 7.5 Band structure diagram of gallium arsenide.

germanium and silicon have an indirect band gap, whereas gallium arsenide has a direct band gap.

In transforming from this type of band structure diagram to the simplified flat band structure diagrams the connected energy levels above the gap are represented as one continuous band while the connected energy levels below the gap are represented as one continuous band. However, all information about their locations in k-space is lost as part of the simplification.

7.1.3 Position of Fermi level in semiconductors

Where does the Fermi level in semiconductors lie relative to the conduction and valence bands?

In the ground state of a semiconductor there are no partially filled bands, just a filled valence band and an empty conduction band. Therefore the Fermi level, which separates the filled from the empty states, lies in the band gap. Consequently, semiconductors do not have a well-defined Fermi surface, in fact we can argue that they do not have a Fermi surface in any meaningful sense.

7.1.4 Variation of electron bands with interatomic spacing

If the interatomic spacing is changed what happens to the electron band structure?

As the interatomic spacing decreases so the electron energy bands broaden. This means that the band gap in a semiconductor should be reduced under hydrostatic pressure. This is found to occur in practice. So, for example, germanium which is known to be a semiconductor under normal conditions becomes a metal under 12 GPa (120 kbar) of hydrostatic pressure.

7.2 INTRINSIC SEMICONDUCTORS

What is an intrinsic semiconductor?

Intrinsic semiconductors are those materials with relatively small band gaps. In these cases, a number of electrons can be thermally stimulated across the band

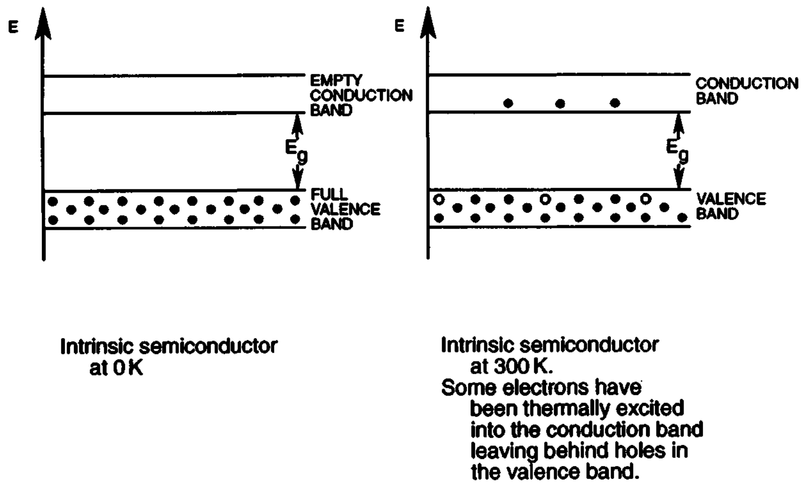


Figure 7.6 Schematic band structure diagram of an intrinsic semiconductor at absolute zero of temperature, when the conduction band is empty, and at 300 K when some electrons have been thermally stimulated into the conduction band. The numbers in the conduction band depend on the size of the band gap and the temperature.

gap at room temperature (300 K), into the conduction band as shown in Fig. 7.6. Once in the conduction band these electrons contribute to the electrical conductivity, as do the 'holes' which are left behind in the valence band.

7.2.1 Thermal excitation of electrons into the conduction band

How is the number of electrons in the conduction band affected by temperature? The electrical properties of intrinsic semiconductors are not greatly affected by the presence of impurities, at least not at room temperature. This is because the number of electrons in the conduction band is determined principally by thermal excitation of electrons from the valence band as a result of the narrow band gap. The Fermi distribution function $f(E)$ is given by

$$f(E) = \frac{1}{1 + \exp((E - E_F)/k_B T)}. \tag{7.1}$$

If we take the band gap E_g to be typically 1 eV with the Fermi level in the middle of the gap, and the ambient temperature to be 300 K, with $k_B = 1.38 \times 10^{-23}\,\mathrm{J\,K^{-1}}$ and consequently $k_B T = 0.05$ eV, we can make the approximation,

$$f(E) \cong \exp\left(-\frac{(E - E_F)}{k_B T}\right) = \exp\left(\frac{-E_g}{2k_B T}\right), \tag{7.2}$$

where $E - E_F \cong E_g/2$.

This gives us the probability of an electron being thermally stimulated from the top of the valence band to the bottom of the conduction band. The values of this probability are given in Table 7.2.

T (K)	$k_B T$ (eV)	$E_g/2k_B T$	$\exp(-E_g/2k_B T)$
0	0	∞	0
100	0.0086	58	0.06×10^{-24}
200	0.0172	29	0.25×10^{-12}
300	0.0258	19.3	4.0×10^{-9}
400	0.0344	14.5	0.5×10^{-6}

Table 7.2 Probabilities of an electron being excited across a 1 eV band gap at various temperatures.

Whereas $f(E)$ gives the probability of any state being occupied, the number of electrons at any given energy level, $N(E)$ is the product of the density of available electron states $D(E)$ and the probability of occupancy $f(E)$

$$N(E) = 2D(E)f(E), \tag{7.3}$$

where the factor of two is introduced because electrons can have spin-up or spin-down. This doubles the number of electrons that can occupy any energy level without violating the Pauli exclusion principle. The density of states $D(E)$ for free electrons as shown earlier in Section 4.4.7 is given by

$$N_0(E) = \frac{V}{6\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{3/2}, \tag{7.4}$$

which is the number of energy levels which exists below an energy E. If this expression is differentiated it gives the number of energy states within unit energy interval $D(E)$

$$\frac{dN_0(E)}{dE} = D(E) = \frac{V}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2}, \tag{7.5}$$

while the density of electronic states is twice this number, and the density of occupied states is given by,

$$N(E) = \frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2} f(E). \tag{7.6}$$

At room temperature an intrinsic semiconductor has about 10^{15}–10^{20} conduction electrons per m^3 (depending on the size of the band gap) caused by thermal stimulation alone. This contribution to the electrical conductivity is known as the 'dark current' simply because it arises in the absence of incident light.

7.2.2 Number density of electrons in the conduction band

The number of conduction electrons per unit volume can be calculated using the equations in Section 7.2.1 for specific materials. At a room temperature of 300 K, the probability of any given electron being excited across the band gap in silicon is 0.15×10^{-9} ($E_g = 1.17\,\mathrm{eV}$) and in germanium is 0.5×10^{-6} ($E_g = 0.75\,\mathrm{eV}$). This probability, when multiplied by the number of available electrons per unit volume at the top of the valence band N_v, gives the number of conduction electrons per unit volume N_e

$$N_e = N_v f(E) \tag{7.7}$$

$$= N_v \exp\left(-\frac{E_g}{2k_BT}\right). \tag{7.8}$$

In practice, the number of electrons per unit volume at the top of the valence band is nearly equal to the number of electron states per unit volume at the bottom of the conduction band.

In an intrinsic semiconductor such as silicon, the electrical charge will be carried by both electrons in the conduction band and by holes in the valence band. The number of holes per unit volume in the valence band is given by an equation similar to eqn (7.8),

$$N_h = N_c \exp\left(-\frac{E_g}{2k_BT}\right), \tag{7.9}$$

where N_c is the number density of holes available at the bottom of the conduction band. In an intrinsic semiconductor N_h should equal N_e. The product N_cN_v gives a number which is the square of the intrinsic electron/hole pair density.

These last two equations are useful because the values of N_c and N_v do not vary much with temperature compared, for example, with the exponential term.

Therefore they are often assumed, to a first approximation, to be constant. Also, interestingly, the values of N_c and N_v do not change by much among the commonly encountered semiconductor materials, for which they have values of typically $2.5 \times 10^{25}\,m^{-3}$. However, it must be remembered that, for example, N_v is a measure of the number of available electrons per unit volume at the top of the valence band which is significantly less than the number density of all valence electrons which is typically 10^{28}–$10^{29}\,m^{-3}$.

For silicon at room temperature, therefore, $N_e = N_h = 3.8 \times 10^{15}\,m^{-3}$, while for germanium $N_e = N_h = 1.25 \times 10^{19}\,m^{-3}$. Table 7.3 shows a comparison of the electronic properties of an insulator such as diamond, two intrinsic semiconductors such as silicon and germanium, and a metal such as copper.

Table 7.3 Comparison of the band gap, conduction electron densities and conductivities of carbon (diamond), silicon, germanium, and copper.

	C (diamond)	**Si**	**Ge**	**Cu**
E_g (eV)	5.3	1.17	0.75	0
N_e at 300 K (m^{-3})	1×10^{-20}	4×10^{15}	1.25×10^{19}	8.5×10^{28}
σ ($\Omega^{-1}\,m^{-1}$)	$\sim 10^{12}$	5×10^{-4}	2.2	5×10^{7}

7.2.3 Conductivity of intrinsic semiconductors

Which factors determine the conductivity of intrinsic semiconductors?

In the intrinsic semiconductors electrons are stimulated across the energy gap from the valence band to the conduction band. The higher the temperature, the more electrons are found in the conduction band. Each electron that undergoes such a transition leaves behind a hole in the valence band. Both electrons and holes can contribute to the electrical conductivity of the material.

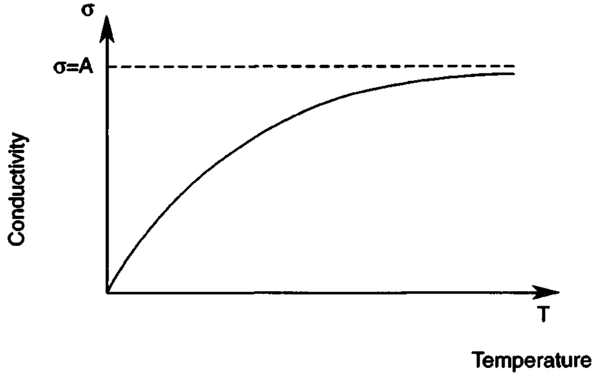


Figure 7.7 Variation of the conductivity of an intrinsic semiconductor as a function of temperature.

The conductivity σ is determined by the sum of contributions due to electrons and holes:

$$\sigma = eN_e\mu_e + eN_h\mu_h, \tag{7.10}$$

where e is the electronic charge, N_e is the number density of conduction electrons, N_h is the number density of conduction holes, μ_e is the mobility of electrons and μ_h is the mobility of holes. The mobility is the average drift velocity of charge carriers per unit electric field strength:

$$\mu = \frac{v}{\xi} = \frac{e\tau}{m} = \frac{e}{\gamma}. \tag{7.11}$$

If we assume that $N(E) = 2D(E)f(E)$ then using the above free electron expression for $D(E)$, it can be shown that the number density of electrons and holes is

$$N_e = 4.82 \times 10^{21}\left(\frac{m^*}{m_0}\right)^{3/2} T^{3/2} \exp\left(\frac{-E_g}{2k_BT}\right) \tag{7.12}$$

$$= N_h. \tag{7.13}$$

Typically $N_e = N_h = 10^{15}$ to 10^{19} carriers per m^3 for an intrinsic semiconductor at room temperature. The electrical conductivity of an intrinsic semiconductor with band gap E_g is

$$\sigma = 4.82 \times 10^{21}\left(\frac{m^*}{m_0}\right)^{3/2} T^{3/2} e(\mu_e + \mu_h) \exp\left(\frac{-E_g}{2k_BT}\right), \tag{7.14}$$

and m^* is the effective mass of the electrons and holes, which is assumed to be the same for these purposes. In fact this is rarely the case. Equation (7.14) has the form

$$\sigma = A \exp\left(\frac{-E_g}{2k_BT}\right). \tag{7.15}$$

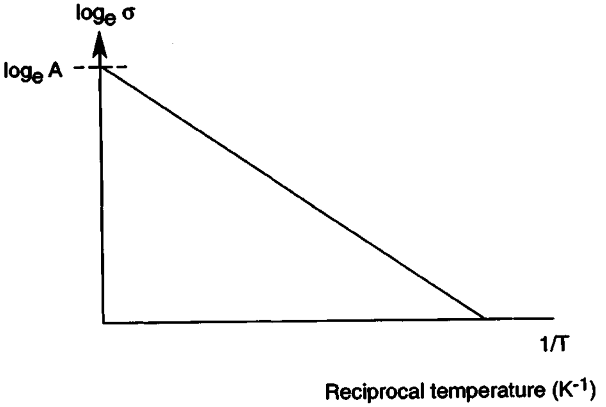


Figure 7.8 Variation of log σ with $1/T$ for an intrinsic semiconductor.

If we take logarithms of both sides

$$\log_e \sigma = \log_e A - \frac{E_g}{2k_B T}, \tag{7.16}$$

which means that if intrinsic conductivity is the only mechanism taking place, we would expect the conductivity σ to vary with temperature T as shown in Fig. 7.8.

7.3 EXTRINSIC (OR IMPURITY) SEMICONDUCTORS

What is an extrinsic semiconductor?

An extrinsic semiconductor is a material with a band gap in which a certain concentration of defect sites is deliberately introduced leading to additional electron or hole states (or both) in the band gap. The electrons can be thermally stimulated from the 'donor' levels in the band gap into the conduction band leading to electron conduction, or alternatively electrons can be thermally stimulated from the valence band into the 'acceptor' levels in the band gap leading to hole conduction in the valence band.

The electrical properties of extrinsic semiconductors can be carefully controlled by the addition of acceptor or donor atoms. This means that the materials can be designed for specific technological applications, and therefore the extrinsic semiconductors remain the most important materials for electronics applications [6].

The donor levels reside at an energy ΔE_d below the conduction band and the acceptor levels reside at an energy ΔE_a above the valence band. Typically these energy differences are about 1% of the energy gap

$$\Delta E_a \approx \Delta E_d \approx 0.01 E_g. \tag{7.17}$$

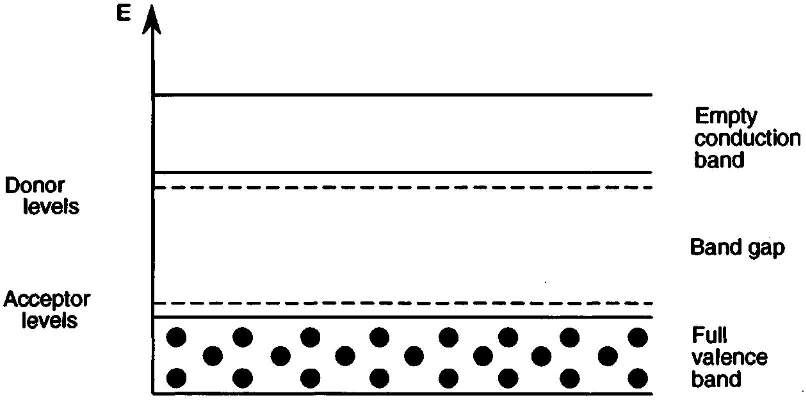


Figure 7.9 Schematic band structure diagram of an impurity semiconductor, showing both acceptor and donor sites in the band gap.

7.3.1 Donor elements

How can the number of electrons available for conduction in the conduction band be altered by the addition of impurities?

Typical donor elements in impurity semiconductors are phosphorus, arsenic, and antimony. When added in small amounts (a few parts in 10^6) they contribute an extra electron which populates the conduction band. The conduction mechanism in this case is via electrons, and the material is called an 'n-type' semiconductor.

7.3.2 Acceptor elements

How can the number of holes available for conduction in the valence band be altered by the addition of impurities?

Typical acceptor elements in impurity semiconductors are boron, aluminium, gallium, and indium. These have one electron less than silicon or germanium. They therefore take an electron from the valence band leaving a 'hole,' which enables conduction to take place in the valence band through effective migration of these 'holes.' Since the charge carrier in this case appears to be positive, such semiconductors are known as 'p-type' semiconductors.

7.3.3 Number density and type of charge carriers

How can the number density and type of charge carriers be determined?

Unlike metals, semiconductors can have either positive or negative charge carriers. The two standard measurements which are used to determine the number and type of charge carriers are the electrical resistivity (or conductivity) and the Hall effect. We will now consider how the number density of charge carriers is altered by temperature and furthermore how the number density of charge carriers and details of the electronic band structure of the semiconductor determine the conductivity.

7.3.4 Temperature dependence of electrical properties

How do the electrical properties of extrinsic semiconductors change with temperature?

Since the populations of electrons in the conduction band, and of holes in the valence band, increase with temperature, so the electrical conductivity of extrinsic

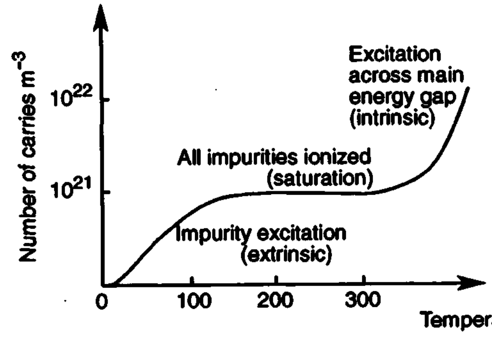


Figure 7.10 Variation of conductivity of an extrinsic semiconductor with temperature.

semiconductors increases with temperature. The presence of impurities provides acceptor and/or donor levels which increase the population levels of charge carriers in the valence band and conduction band, respectively. Therefore, the electrical conductivity increases with impurity content. There is also a contribution to the conductivity arising from absorption of photons, the so-called photoconductivity, providing the photon energy is large enough to excite electrons into the conduction band from the valence band.

7.3.5 Conductivity of extrinsic semiconductors

What factors determine the conductivity of extrinsic semiconductors?

In extrinsic semiconductors we usually have a predominance of donor impurities, leading to an n-type semiconductor or a predominance of acceptor impurities, leading to a p-type semiconductor. Therefore, in n-type semiconductors,

$$\sigma = N_e \mu_e e, \tag{7.18}$$

and in p-type semiconductors,

$$\sigma = N_h \mu_h e. \tag{7.19}$$

Typically the number density of charge carriers in an extrinsic semiconductor is $N = 10^{21}\,\mathrm{m}^{-3}$ at room temperature.

The extrinsic or impurity contribution is for most of these semiconductors the only significant component at room temperature. At higher temperatures thermal stimulation of electrons directly across the band gap may also occur. So there are two contributions to the conductivity: intrinsic and extrinsic conduction whose relative contributions to the total conductivity are dependent on temperature.

If we plot $\log_e \sigma$ against $1/T$ as we did above for the intrinsic semiconductor, then for the extrinsic semiconductor we obtain the following type of plot, shown in Fig. 7.11

We see from this that the extrinsic contribution is more important at low temperatures (high $1/T$) and the intrinsic contribution is more important at high temperatures (low $1/T$).

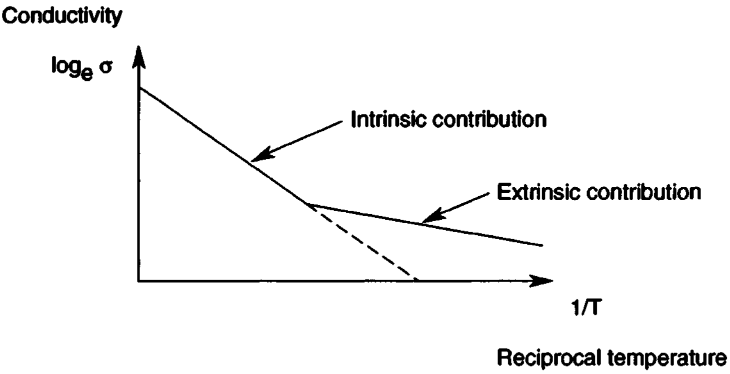


Figure 7.11 Variation of $\log \sigma$ with $1/T$ for an extrinsic semiconductor.

7.3.6 Minority and majority charge carriers

How does the presence of both types of charge carrier affect the electrical properties of a semiconductor?

In discussion of extrinsic semiconductors we have so far only considered situations in which the charge carriers are either exclusively electrons (n-type semiconductors with donors in the band gap) or exclusively holes (p-type semiconductors with acceptors in the band gap). In practice real semiconductor materials will have combinations of both acceptors and donors.

In a material containing predominantly donors (which is known as n-type) the number density of electrons in the conduction band will be high, but there will always be a small, but significant number density of holes in the valence band. A similar situation occurs in p-type semiconductors. The electrons in an n-type material and the holes in a p-type material are known as the majority carriers, while the holes in an n-type and the electrons in a p-type material are known as minority carriers.

In either case the product of the concentration of electrons and holes is a constant for a given material and at a given temperature as shown by the equation

$$N_e N_h = N_i^2, \tag{7.20}$$

where N_i is the intrinsic electron/hole pair density. The charge carrier densities are therefore mutually dependent. Any defect sites that cause N_e to increase must necessarily cause N_h to decrease and vice versa. It follows that a fractional increase in the number density of one type of charge carrier results in a reduction of the number density of the other type of charge carrier in the same ratio. The minority carriers are therefore important in the operation of semiconductor electronic devices because they affect the majority charge carrier concentration.

It is possible to define a difference in number density between the two types of charge carriers by the equation

$$N_r = N_e - N_h, \tag{7.21}$$

where N_e and N_h depend directly on the number density of defect sites, donors and acceptors, respectively. In the case of an n-type semiconductor the majority and minority charge carrier densities are therefore,

$$N_e = \frac{N_r}{2}\left[\sqrt{\left(1 + \frac{4N_i^2}{N_r^2}\right)} + 1\right], \tag{7.22}$$

$$N_h = \frac{N_r}{2}\left[\sqrt{\left(1 + \frac{4N_i^2}{N_r^2}\right)} - 1\right]. \tag{7.23}$$

From the above two equations it is easy to obtain explicit expressions for both N_e and N_h in terms of N_r and N_i. When $N_r \ll N_i$ the behaviour of the material is like an intrinsic semiconductor ($N_e \approx N_h \approx N_i$) and when $N_r \gg N_i$ the behaviour is strongly extrinsic, so that for an n-type semiconductor ($N_e \approx N_r$, $N_h \approx 0$).

7.3.7 Number density of charge carriers: compensation

If both donor and acceptor states are present together in a semiconductor do they compensate for the effects of each other?

We have seen in Section 7.2.2 how the number densities of electrons in the conduction band, and holes in the valence band can be calculated for intrinsic semiconductors. The presence of impurities in extrinsic semiconductors creates donor or acceptor states which strongly affect the number density of charge carriers.

As the number density of electrons in the conduction band is increased by the addition of donors so the density of holes in the valence band is decreased because each hole has a greater probability of encountering an electron with which it can recombine. The presence of acceptors decreases the number of electrons in the conduction band for the same reason. This effect is known as 'compensation,' so that a given number density of acceptors negates the effect of the same number density of donors and vice versa.

The result of this is that the product $N_e N_h$ of the number densities of conduction electrons and conduction holes is independent of the amount of doping or the type of doping, whether n-type or p-type. This leads to the following useful result for extrinsic semiconductors

$$N_e N_h = N_c N_v \exp\left(-\frac{E_g}{k_B T}\right). \tag{7.24}$$

As discussed in Section 7.2.2, the number density of states at the bottom of the conduction band N_c and the number density of states at the top of the valence band N_v are, to a first approximation, independent of temperature and have the same value of typically $2.5 \times 10^{25}\,\mathrm{m^{-3}}$ for the commonly encountered semiconductors. Therefore, with $N_c N_v \simeq 6.25 \times 10^{50}\,\mathrm{m^{-3}}$ the values of the product $N_e N_h$ can readily be estimated for a number of semiconductors simply from a knowledge of the band gap E_g and the temperature. For example, based on the data in Section 7.2.2, $N_e N_h \simeq 1.5 \times 10^{31}\,\mathrm{m^{-6}}$ for silicon and $\simeq 1.6 \times 10^{38}\,\mathrm{m^{-6}}$ for germanium at room temperature, independent of the level of doping.

7.3.8 Charge carrier lifetimes

How long do charge carriers, whether electrons or holes, remain in their excited states?

In a semiconductor when a charge carrier, whether an electron or a hole, is excited into a higher energy state it has a finite life expectancy in that excited state before it recombines in an electron/hole pair. This is known as the charge carrier lifetime. The equilibrium number density of charge carriers will be proportional to the product of the generation rate g and the lifetime τ. For electrons for example,

$$N_e = g\tau, \tag{7.25}$$

and comparing with eqn (7.8) it follows that,

$$g\tau = N_v \exp\left(-\frac{E_g}{2k_B T}\right). \tag{7.26}$$

If additional charge carriers are injected into a device, for example by current injection, then there will be a number of charge carriers $N_{ex} = N_{tot} - N_e$, or $N_{ex} = N_{tot} - N_h$, in excess of the equilibrium charge carrier density N_e or N_h. The number of excess charge carriers will again depend on the charge carrier lifetime τ and the effective recombination rate for charge carriers $(r - g)$ according to the equation

$$\tau(r - g) = N_{ex}, \tag{7.27}$$

where r is the rate of recombination of charge carriers in electron/hole pairs and g is the rate of generation of charge carriers. The larger the lifetime τ the greater is the difference between the actual charge carrier concentration and the equilibrium charge carrier concentration.

Additionally, the longer the lifetime of charge carriers before they recombine, the larger is the distance that they can move as a result of diffusion. The diffusion length L_d is determined by the diffusion coefficient D and the charge carrier lifetime τ according to the following equation,

$$L_d = \sqrt{D\tau}, \tag{7.28}$$

which applies to both electrons and holes. The diffusion length and the charge carrier lifetime affect the operational efficiency of devices such as pn junctions and transistors.

7.4 OPTICAL PROPERTIES OF SEMICONDUCTORS

How are the optical properties of semiconductors determined by the electron band structure?

Due to the band gap energy E_g, semiconductors are unable to absorb and reflect lower frequencies. Absorption and reflection start to occur at a frequency ω_0 given by $\hbar\omega_0 = E_g$ and at this frequency we notice the so-called absorption edge. At higher frequencies absorption and reflection can be relatively high. The absorption edge of a semiconductor is shown in Fig. 7.12.

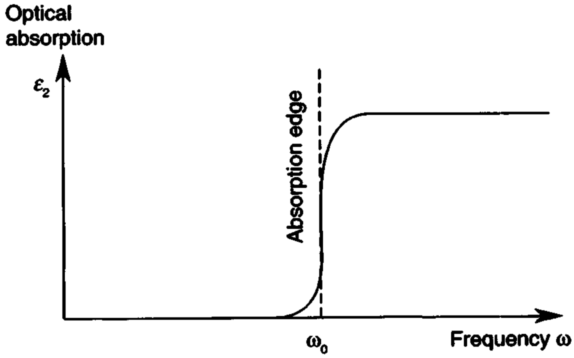


Figure 7.12. Idealized variation of optical reflectance or absorption in a semiconductor with frequency of incident light.

7.5 PHOTOCONDUCTIVITY

What happens when electrons are excited across the band gap by absorption of photons?

In addition to thermal stimulation of electrons into the conduction band, there are other methods of achieving the same result. Illumination of a semiconductor with light of sufficient frequency (i.e. photon energy) also leads to increased numbers of charge carriers in the conduction band and hence to increased conductivity. This arises from the excitation of electrons from the valence band to the conduction band by photons. The dependence of conductivity on the absorption of light is known as photoconductivity.

The frequency of light necessary to increase the conductivity of an intrinsic semiconductor is determined by the band gap energy E_g. In an extrinsic semiconductor it is determined by the energy displacement of the donor or acceptor sites from the conduction or valence bands, respectively.

The variation of conductivity with frequency of incident radiation is shown in Fig. 7.13.

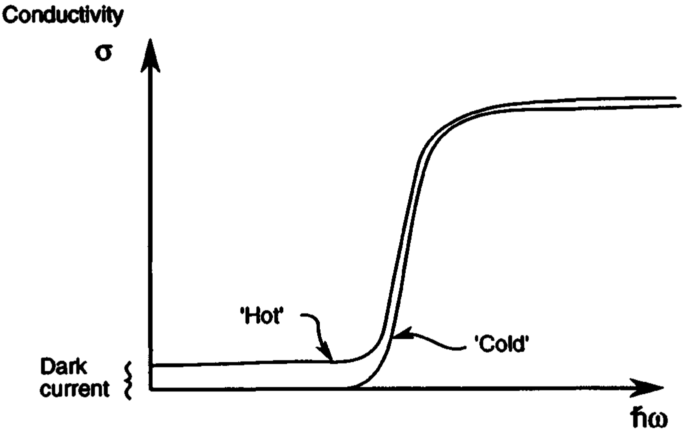


Figure 7.13 Idealized variation of electrical conductivity in a semiconductor with frequency of incident light at constant intensity and at different temperatures.

For an intrinsic semiconductor with band gap 0.7 eV, the frequency ω_0 necessary to stimulate electrons across the band gap E_g is

$$\hbar\omega_0 = E_g \tag{7.29}$$

$$\omega_0 = 1.06 \times 10^{15}\,\mathrm{s}^{-1}, \tag{7.30}$$

which corresponds to a wavelength of 1770 nm which is well into the infrared region of the electromagnetic spectrum. Therefore optical wavelengths carry sufficient energy to stimulate electrons across the gap in these intrinsic semiconductors.

The dependence of photocurrent on the intensity of incident light above the threshold frequency ω_0, means that the change in electrical conductivity of a semiconductor with incident photon flux can be used as a method of detecting infrared radiation.

The component of conduction due solely to the thermodynamic temperature is known as the 'dark current', since this is the current which would be obtained if the semiconductor was shielded from all incoming radiation (i.e. was literally kept in the dark). The variation of conductivity with frequency of light at different temperatures is also shown in Fig. 7.13. It is known that if the charge carriers are prevented from recombining immediately the photoconductivity will persist even after the incident light has been removed [7].

Extrinsic semiconductors can be used for infrared detection providing the frequency ω of incident infrared radiation satisfies the condition,

$$\hbar\omega \geq \Delta E, \tag{7.31}$$

where ΔE is the energy difference between the donor sites and the conduction band, for example. The threshold frequency is $\omega_0 \approx 10^{13}\,\mathrm{s}^{-1}$. This involves photoconductivity by stimulation of electrons from donor levels or to acceptor levels. In the case of infrared detection the semiconductor is usually cooled to 4.2 K to return thermally excited carriers to their donor or acceptor sites and so reduce the dark current, thereby improving the signal-to-noise ratio of the photocurrent.

7.6 THE HALL EFFECT

How do the electrons and holes in a semiconductor behave under the combined action of an electric field and a magnetic field?

Conductivity measurements alone are not sufficient to find the total number of charge carriers, their signs and their mobilities. Measurement of the Hall effect gives the necessary additional information.

When a current flows in a conductor or semiconductor and a magnetic field is applied perpendicular to the current, then a voltage is generated across the material in a direction perpendicular to planes containing the current and the magnetic field, as shown in Fig. 7.14. This voltage is known as the Hall voltage and the phenomenon is known as the Hall effect.

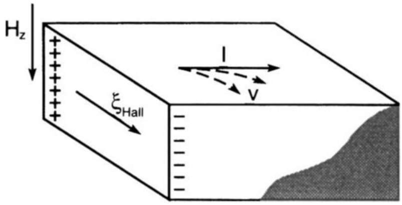


Figure 7.14 Hall emf resulting from the action of an external magnetic field on charge in motion. ***H*** is the magnetic field, *I* the conventional current, ξ_{Hall} is the Hall field, and ***v*** is velocity of electrons.

The explanation of this phenomenon is quite simple. It arises as a result of the Lorentz force on a moving charge in a magnetic field. If a charge e is moving with velocity v in a magnetic field H, then the Lorentz force F_L is given by,

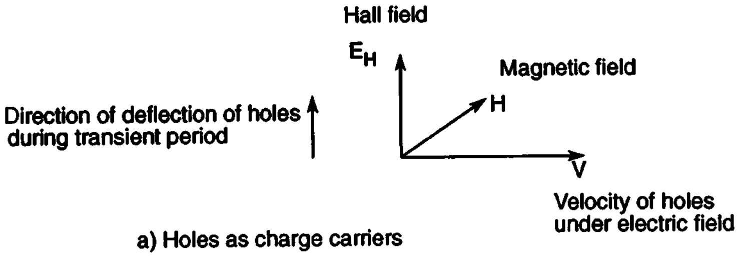
$$F_L = \mu_0 e\boldsymbol{v} \times \boldsymbol{H}. \tag{7.32}$$

If we wish to express this as an equivalent electric field ξ_{Hall}, then remembering that $F_L = e\xi_{\mathrm{Hall}}$, we have simply

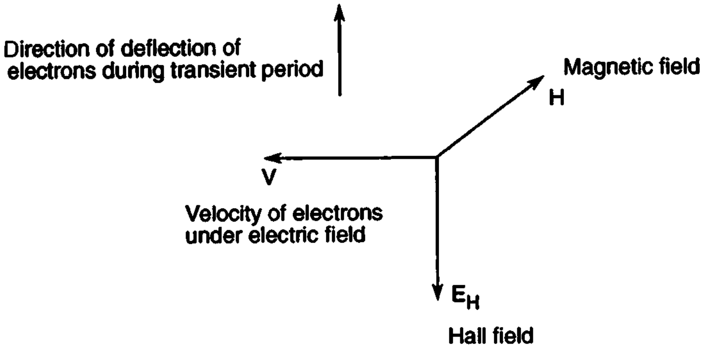
$$\xi_{\mathrm{Hall}} = \mu_0 \boldsymbol{v} \times \boldsymbol{H}. \tag{7.33}$$

We also know that the current density J is given by the product of N the number density of charge carriers, e the electronic charge and v the velocity of the charge carriers

$$J = Nev. \tag{7.34}$$



a) Holes as charge carriers



b) Electrons as charge carriers

Figure 7.15 Direction of Hall field for charge carriers of different sign. In both cases the direction of the magnetic field ***H*** is into the plane of the paper, and in both cases the conventional current is from left to right.

Substituting this into the equation for the Hall emf leaves,

$$\xi_{\text{Hall}} = \mu_0 \frac{J \times H}{Ne}. \tag{7.35}$$

The Hall coefficient R_H is defined such that,

$$R_H = \frac{\xi_{\text{Hall}}}{\mu_0 JH} \tag{7.36}$$

$$= \frac{1}{Ne}. \tag{7.37}$$

This means that we can make measurements of ξ_{Hall}, J and H and from these determine the product Ne. Note that the sign of the Hall coefficient depends on the sign of the charge carriers. Typical values of the Hall coefficient are $R_H \cong 10^{-9}\,\text{m}^3\,\text{C}$.

We note that if the Hall field ξ_{Hall} is in a certain direction for a flow of negative charge carriers, then it will be in the opposite direction for the same current when it is produced by a flow of positive charges in the reverse direction, as shown in Fig. 7.15.

The Hall coefficient is inversely proportional to the number density of charge carriers N. Hall measurements are easy to make on semiconductors because N is relatively low, being typically 10^{15} to $10^{21}\,\text{m}^{-3}$. In metals, N is large being typically $10^{28}\,\text{m}^{-3}$ and the measurement of R_H is consequently more difficult. Typical values are given in Table 7.4. The Hall effect can also exhibit quantum effects as shown by, among others, Von Klitzing *et al.* [8].

Table 7.4 Hall coefficients for various materials.

Material	**R_H ($\text{m}^3\,\text{C}^{-1}$)**
Li	-1.7×10^{-10}
In	$+1.59 \times 10^{-10}$
Sb	-1.98×10^{-9}
Bi	-5.9×10^{-7}

7.7 EFFECTIVE MASS AND MOBILITY OF CHARGE CARRIERS

If the relationship between energy and k vector in a semiconductor is no longer quadratic, how can this be expressed in terms of mobility or effective mass of the charge carriers?

We have stated in Section 5.4.5 that an effective mass can be defined for electrons in a particular location in the band structure, in accordance with the relation,

$$m^* = \frac{\hbar^2}{(\mathrm{d}^2E/\mathrm{d}k^2)}, \tag{7.38}$$

which means that energy bands with high curvature lead to low effective mass, and flat energy bands lead to high effective mass. For an electron near the top of a band d^2E/dk^2 is negative. This means that such an electron decelerates in the presence of a field as it exchanges momentum with the lattice.

The idea of a negative effective mass of an electron is conceptually difficult. In fact we find that the charge carriers near the top of a band are not electrons but 'holes' with a positive charge and a positive mass. The net acceleration produced is the same as that on a negative charge with a negative mass. The 'hole' moves in the opposite direction from the electronic current.

The conduction mechanism in semiconductors is therefore more complicated than in metals. Two types of charge carrier are possible, and in addition the number of charge carriers is temperature dependent.

We have defined the mobility in Section 7.2.3 as the velocity per unit field strength, $\mu = v/\xi$. Remember that inside the semiconductor the electrons or holes cannot accelerate indefinitely under the action of a field (as they would in free space) because of their interaction with the rest of the solid. The mobility is proportional to the inverse of the scattering probability, and we know that scattering of electrons is caused by phonons and impurity atoms. In metals, the scattering probability remains almost constant as a function of temperature, but in semiconductors this is not true because the energy distribution of the carriers in a semiconductor varies with temperature.

7.8 SEMICONDUCTOR JUNCTIONS

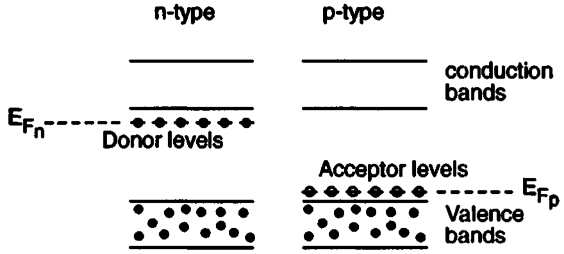
Are there any interesting electronic effects at the boundary between two different semiconducting materials?

So far we have talked at length about the conductivity mechanisms in intrinsic and extrinsic semiconductors. We have not yet looked at the electronic properties of junctions between semiconductors, and yet from the applications viewpoint semiconductor junctions are of crucial importance in devices.

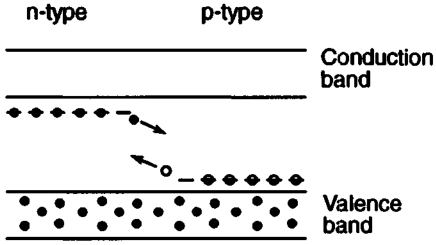
Suppose a piece of p-type semiconductor is in direct contact with a piece of n-type semiconductor. This junction has some very interesting properties which have had a direct bearing on the development of semiconductor technology. In order to understand these properties, however, we need to consider two principles:

(i) When two solids are in contact, charge transfer occurs until their Fermi energies are the same.
(ii) In n- and p-type semiconductors, the Fermi level lies approximately at the donor and acceptor levels, respectively.

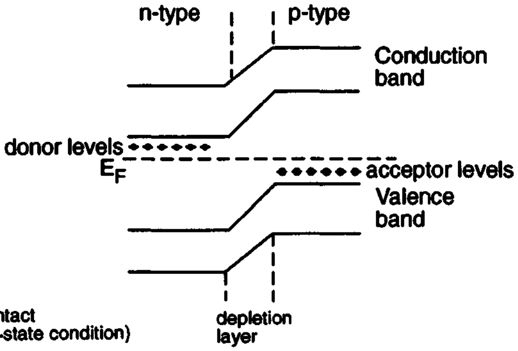
When the n-type and p-type semiconductors are placed in contact, an unstable situation arises temporarily because of the step change in electron and hole densities across the interface. Equalization of the Fermi levels occurs as electrons diffuse from the n-type material into the p-type material. This charges the p-type material negatively and sets up an electric field which opposes further diffusion of electrons. A dynamic equilibrium arises when the Fermi levels on either side are equal. However as a result of the net electrical charge on each side of the junction



a) before contact



b) immediately after contact (transient condition)



c) After contact (steady-state condition)

Figure 7.16 Electron band structure diagram for a pn junction: (a) energy levels of p-type and n-type semiconductor before contact, (b) transient condition immediately after contact, (c) steady-state condition after contact.

the conduction and valence bands of the two sides of the junction are displaced relative to each other, as shown in Fig. 7.16.

7.8.1 Depletion layer

How is the region in the vicinity of a semiconductor junction different from the bulk of the material?

As the electrons diffuse from the n-type to the p-type material, the number of charge carriers in the n-type material decreases, at least in the volume close to the junction. A similar effect occurs as the holes diffuse from the p-type to the n-type material. This means that in the vicinity of the interface there is a 'depletion layer' on each side of the interface containing fixed but opposite charges and a reduced number of charge carriers. This depletion layer is sometimes called the space charge region.

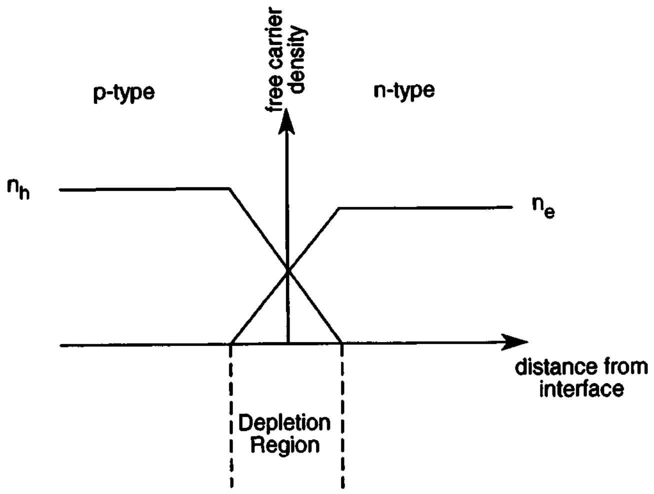


Figure 7.17 Electron and hole densities on either side of a pn junction.

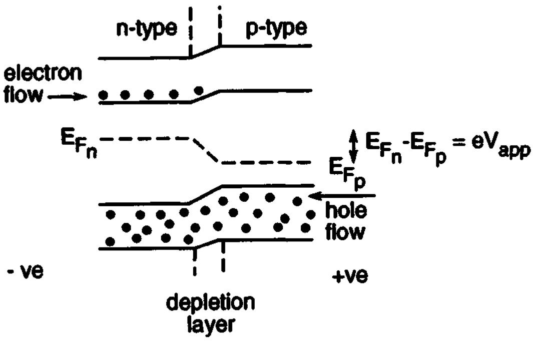
Now, if we consider this junction, it is found to have some useful properties. For example, electrons in the conduction band of the n-type material which may be trying to diffuse into the p-type material encounter a potential barrier at the junction. This makes it difficult for electrons to pass from the n-type to the p-type material, but relatively easy for them to pass in the opposite direction. This is represented in the figure by the energy gradient for electrons passing from n-type to p-type.

On the other hand, there are only a few electrons in the conduction band of the p-type material whereas there are many in the conduction band of the n-type material. Under equilibrium conditions, the diffusion rate of electrons in each direction is equal. Similar arguments apply to the holes in the valence band, whereby the p-type material has many holes in the valence band but the n-type material has only a few holes in its valence band.

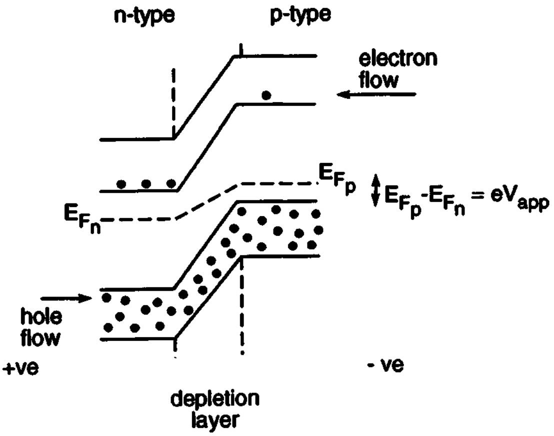
7.8.2 Forward biasing the pn junction

What happens if a voltage is applied across the junction to reduce the potential difference across the junction?

If we place an electric voltage across the junction, for example by connecting the positive terminal of a dc power supply to the p-type material and the negative terminal to the n-type, the voltage difference, and hence the energy difference, between the two conduction bands is lowered. This is called positive or forward biasing. The height of the potential barrier across the conduction band is reduced from E_g to $\Delta E = E_g - eV$ where V is the applied voltage, as shown in Fig. 7.18. As a result of this shift, more of the n-type conduction band is exposed to the p-type material and electrons cross more easily from the n-type material into the p-type material. These can then recombine with holes emitting light. The forward-biased pn junction can therefore be used as a light source known as a light emitting diode (LED), or under certain special conditions as a semiconductor laser.



Forward bias



Reverse bias

Figure 7.18 Electron band structure of reverse-biased and forward-biased pn junctions.

7.8.3 Reverse biasing the pn junction

What happens if a voltage is applied to increase the potential difference across the junction?

If the negative terminal of the dc power supply is connected to the p-type material, the energy separation of the bands on either side of the junction is increased. This process is called negative or reverse biasing. In this case, because the height of the barrier is increased by $\Delta E = E_g + eV$, as shown in Fig. 7.18, it becomes very difficult to drive electrons from the n-type into the p-type material. However, electrons can move easily from the p-type to the n-type material down the potential gradient of the conduction band in the vicinity of the interface. An electrical current pulse occurs for example if electrons are stimulated across the band gap by the absorption of photons in the junction itself. This means that the reverse-biased pn junction can be used as a light detector (photodetector or photocell).

The current in a pn junction diode therefore behaves nonlinearly with the bias voltage as shown in Fig. 7.19. We see that the simple pn junction which we have described can also act as a rectifier or diode by allowing current to flow in one direction only.

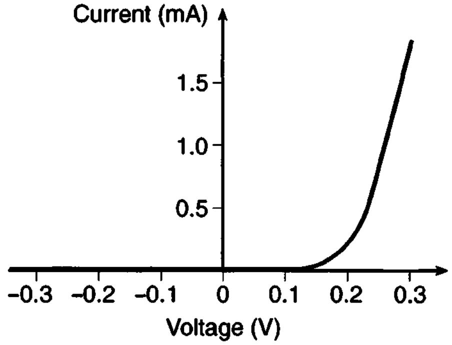


Figure 7.19 Voltage–current characteristic of a pn junction, showing conduction in the forward direction but no conduction in the reverse direction (until electrical breakdown occurs at much higher reverse voltages).

7.8.4 Semiconductor devices

How can we explain the movement of charge carriers in a pn junction?

We have looked at the simplest case of a semiconductor device, the pn junction, and shown that this can be used as a diode. We need to look at the operation of the pn junction in more detail so that the current–voltage characteristics shown in Fig. 7.19 can be understood in terms of the electronic properties. Consider first the band structure diagram of the pn junction shown in Fig. 7.16c.

Electrons in the conduction band of the n-type material cannot easily reach the p-type material because of the potential energy ramp which causes an internal electric field ξ. However, the higher density of electrons in the conduction band of the n-type material will cause a diffusion of electrons from this region of high concentration to the region of low concentration in the conduction band of the p-type material.

In diffusion, the total number of electrons passing through cross-sectional area A in unit time, $\mathrm{d}N/\mathrm{d}t$, is dependent on the rate of change of the number density of electrons with position, $\mathrm{d}N/\mathrm{d}x$, according to the equation,

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -DA\frac{\mathrm{d}N}{\mathrm{d}x}. \tag{7.39}$$

Here, D is the diffusion coefficient in units of $\mathrm{m}^2\mathrm{s}^{-1}$ which in this case is given by

$$D = \frac{\mu k_\mathrm{B} T}{e}, \tag{7.40}$$

where μ is the mobility of the electrons (see Section 7.2.3), k_B is Boltzmann's constant, e is the charge on the electron and T is the temperature in Kelvin. Substituting for D in the diffusion equation,

$$\left(\frac{\mathrm{d}N}{\mathrm{d}t}\right) = \frac{-\mu k_\mathrm{B} TA}{e}\left(\frac{\mathrm{d}N}{\mathrm{d}x}\right). \tag{7.41}$$

We can therefore define a diffusion current density J_d for the charge carriers in terms of the rate of change of the number density N with time,

$$J_\mathrm{d} = \frac{e}{A}\left(\frac{\mathrm{d}N}{\mathrm{d}t}\right) \tag{7.42}$$

$$= -\mu k_\mathrm{B} T\left(\frac{\mathrm{d}N}{\mathrm{d}x}\right). \tag{7.43}$$

Under equilibrium conditions this diffusion current density must be balanced by a conventional current density J_v due to the voltage gradient at the pn junction:

$$J_\mathrm{v} = \sigma\xi = Ne\mu\xi, \tag{7.44}$$

and equating the current densities, $J_\mathrm{d} = J_\mathrm{v}$

$$Ne\mu\xi = -\mu k_\mathrm{B} T\frac{\mathrm{d}N}{\mathrm{d}x}. \tag{7.45}$$

Rearranging and integrating gives

$$V = \int \xi\,\mathrm{d}x = \frac{-k_\mathrm{B} T}{e}\int\frac{\mathrm{d}N}{N}, \tag{7.46}$$

and noting that when $V = 0$, $N = N(0)$, and when $V = V$, $N = N(V)$, this gives

$$V = \frac{-k_\mathrm{B} T}{e}\log_e\left(\frac{N(V)}{N(0)}\right). \tag{7.47}$$

This relates the number density of electrons $N(V)$ at the top of the potential ramp of height V (in the p-type material) to the number density of electrons $N(0)$

at the bottom of the potential ramp (in the n-type material). Equation (7.43) shows how the equilibrium number density of electrons depends on the voltage.

If we look at the problem in more detail there is a hole current in addition to the electron current in a semiconductor. We denote the electron current density from the n-type region to the p-type region (diffusion current) as J_{de}, the electron current density from the p-type region to the n-type region (field current) as J_{ve}. Similarly the hole current densities can be denoted J_{dh} (diffusion) and J_{vh} (field).

7.8.5 Dependence of current on voltage across a pn junction

How can we explain the current–voltage characteristics of a pn junction?

When a bias voltage V_{app} is applied to the pn junction it causes currents to flow across the junction. A positive voltage by convention reduces the potential barrier for both electrons and holes, while a negative voltage by convention increases the barrier. As the voltage is increased the net electron current density flowing from the n-type region increases. Likewise the net hole current density flowing from the p-type region increases. The current density for the diffusion of electrons can be determined from,

$$J_{de} = \frac{e}{A}\left(\frac{dN(0)}{dt}\right) \tag{7.48}$$

$$= \frac{e}{A}\frac{d}{dt}\left(N(V)\exp\left(\frac{eV_{app}}{k_B T}\right)\right) \tag{7.49}$$

$$= \frac{e}{A}\frac{dN(V)}{dt}\exp\left(\frac{eV_{app}}{k_B T}\right), \tag{7.50}$$

and replacing $\frac{e}{A}\frac{dN(V)}{dt}$ by J_{vc} we arrive at

$$J_{de} = J_{ve}\exp\left(\frac{eV_{app}}{k_B T}\right), \tag{7.51}$$

and similarly for the hole current

$$J_{dh} = J_{vh}\exp\left(\frac{eV_{app}}{k_B T}\right). \tag{7.52}$$

The net electron current density is $J_e = J_{de} - J_{ve}$ and the net hole current density is $J_h = J_{dh} - J_{vh}$. The total current flowing is then,

$$J_{tot} = J_e + J_h \tag{7.53}$$

$$= (J_{vh} + J_{vc})\left\{\exp\frac{eV_{app}}{k_B T} - 1\right\} \tag{7.54}$$

$$= J_s\left\{\exp\left(\frac{eV_{app}}{k_B T}\right) - 1\right\}. \tag{7.55}$$

This is the diode equation, where J_s is the sum of the current densities carried by the minority carriers across the junction. We can now express the current flowing through a pn junction in terms of the voltage across it. When $V_{app} = 0$ there is no current. When $V_{app} > 0$ the current increases exponentially with applied voltage. When $V_{app} < 0$ the current is small and negative. This equation describes the current–voltage characteristics shown in Fig. 7.19.

7.8.6 Temperature dependence of the current–voltage characteristics of a pn junction

How does the pn junction behave if the temperature changes?

The full temperature dependence of J_{tot} is not shown explicitly in the above equation because J_s itself is temperature dependent. J_s is known as the 'reverse saturation current density'. The reason for this is that when a very large reverse bias (negative voltage) is applied to the junction, the current density J_{tot} approaches J_s. Specifically,

$$\lim_{V_{app}\to-\infty} J_{tot} = -J_s, \tag{7.56}$$

as can be seen from Fig. 7.19 and eqn (7.47). As a further example, when $V_{app} < -3k_BT/e$ the current density J_{tot} is 95% of J_s.

This reverse saturation current density is determined by the band gap energy E_g and the temperature T. In general, there is a contribution to J_s from both electrons and holes, so that

$$J_s = eD_e \frac{dN_e}{dx} + eD_h \frac{dN_h}{dx}. \tag{7.57}$$

This is known as Shockley's equation. D_e and D_h are the diffusion coefficients for electrons and holes, respectively, N_e is the number density of electrons in the conduction band of the p-type material, and N_h is the number density of holes in the valence band of the n-type material. From the symmetry of the situation, the number densities of these 'minority carriers' on either side of the junction can reasonably be expected to be equal, and since the movement of holes is simply the movement of electrons in the opposite direction, it is also reasonable to expect the diffusion coefficients to be equal. Therefore,

$$J_s = 2eD_e \frac{dN_e}{dx}. \tag{7.58}$$

In this equation the temperature dependence of J_s is determined principally by the temperature dependence of dN_e/dx. If the distance within the junction over which the voltages of the two bands change is ℓ then a linear approximation gives $dN_e/dx = N_e/\ell$. An expression for N_e can be obtained from eqn (7.24) so that

$$N_e = \frac{N_c N_v}{N_h} \exp\left(-\frac{E_g}{k_B T}\right), \tag{7.59}$$

which gives an expression for the temperature dependence of the reverse saturation current density as

$$J_{\mathrm{s}} = \frac{2eD_{\mathrm{e}}}{\ell}\frac{N_{\mathrm{c}}N_{\mathrm{v}}}{N_{\mathrm{h}}}\exp\left(-\frac{E_{\mathrm{g}}}{k_{\mathrm{B}}T}\right), \tag{7.60}$$

and as shown by Sze [9] the terms in front of the exponential are also weakly temperature dependent; however, the exponential term dominates the temperature dependence of J_{s}.

As the temperature increases, the value of the exponential term approaches unity, giving a high-temperature limit for J_{s}

$$\lim_{T\to\infty} J_{\mathrm{s}} = \frac{2eD_{\mathrm{e}}}{\ell}\frac{N_{\mathrm{c}}N_{\mathrm{v}}}{N_{\mathrm{h}}} = J_{\infty}. \tag{7.61}$$

The expression for the total current density can, therefore, be written as

$$J_{\mathrm{tot}} = J_{\infty}\exp\left(-\frac{E_{\mathrm{g}}}{k_{\mathrm{B}}T}\right)\left(\exp\left(\frac{eV_{\mathrm{app}}}{k_{\mathrm{B}}T}\right)-1\right), \tag{7.62}$$

and since J_{∞} is not dependent on temperature, the other terms in the equation determine the temperature dependence of the current density through the pn junction.

7.8.7 Gallium arsenide and related compounds

What is the reason for the great interest in gallium arsenide and related compounds? Although the present day semiconductor industry is based largely on silicon, there has been widespread interest in gallium arsenide and related compounds based on elements from group III and V of the periodic table. This has been a long-standing interest. Gallium arsenide is of great technical interest for three reasons. It has

(i) a high electron mobility μ_{e},
(ii) a large band gap E_{g},
(iii) a direct band gap.

The large electron mobility results from the small effective mass m^* of the electrons in this material. This is caused by the relatively large upward curvature of the conduction band near to the Γ point in the Brillouin zone. As a result, devices can be fabricated from this material which can operate over a very short time period. The material therefore finds applications where the speed of operation of electronic systems is of paramount importance.

The band gap of 1.5 eV makes gallium arsenide sensitive to wavelengths of light of 827 nm and shorter. This means that electrons dropping from the bottom of the conduction band to the top of the valence band will emit a photon in the red end of the visible spectrum. It can therefore be used in light emitting diodes (LED), which emit red light. By combining it with other materials such as gallium phosphide, other colours can be produced by engineering the size of the band gap. A wide range of optoelectronic semiconductors based on III–V compounds have been produced.

The optical properties of GaAs, specifically its high efficiency in detection and generation of light, derive from its direct band gap. This means that it can be used in fabricating lasers [10] and in optical communication devices for computers. In this application information is transmitted by photons instead of by electrons. These optical computers are much faster than conventional computers. Recent developments in optical computing have again brought this technology to the forefront.

7.8.8 Summary of gallium arsenide properties

What are the advantages of gallium arsenide over silicon and germanium for fabrication of devices?
These include:

(i) GaAs circuits are faster and operate at equal, or lower, power than silicon circuits.
(ii) The separation between the conduction and valence bands is more easily controlled in GaAs and related compounds than in silicon.
(iii) GaAs can radiate and detect near infrared and visible red radiation depending on its band gap.
(iv) GaAs can support optoelectronic functions while silicon cannot.

REFERENCES

1. A. H. Wilson (1931) The theory of electronic semiconductors, *Proc. Roy. Soc.* A133, 458.
2. A. H. Wilson (1931) *Proc. Roy. Soc.* A134, 277.
3. W. Kleinmann and J. C. Phillips (1960) *Phys. Rev.* 118, 1164.
4. D. Brust *et al.* (1964) *Phys. Rev.* 134, 1337.
5. M. Pollak *et al.* (1966) *Physics of Semiconductor Devices*, Kyoto.
6. C. M. Osburn and A. Reisman (1987) Challenges in advanced semiconductor technology in the VLSI era for computer applications, *J. Electron. Mater.* 16, 223.
7. J. R. Lowney and S. Mayo (1992) Analysis of persistent photo conductivity due to potential barriers, *J. Electron. Mater.* 21, 731.
8. K. von Klitzing, G. Dorda and M. Pepper (1980) *Phys. Rev. Lett.* 45, 494.
9. S. M. Sze (1981) *Physics of Semiconductor Devices*, 2nd edn, John Wiley and Sons, New York, p. 88
10. W. D. Goodhue *et. al.* (1990) Monolithic two dimensional GaAsAlGaAs laser arrays fabricated by chlorine ion beam assisted micromachining, *J. Electron. Mater.* 19, 463.

FURTHER READING

J. D. Morant (1970) *Introduction to Semiconductor Devices*, 2nd edn, Harrap, London.
L. Solymar and D. Walsh (1984) *Lectures on the Electrical Properties of Materials*, 3rd edn, Oxford University Press, Oxford.
K. D. Leaver (1989) *Microelectronic Devices*, Longmans, Harlow, England.
S. O. Kasap (1997) *Principles of Electrical Engineering Materials and Devices*, Irwin, Chicago.
S. M. Sze (1998) *Modern Semiconductor Device Physics*, John Wiley and Sons, New York.

EXERCISES

Exercise 7.1 Approximation to the Fermi function in semiconductors
If the band gap in a semiconductor is $E_g = 0.5\,\text{eV}$ and the temperature is 300 K, show that the Fermi function for electrons at the bottom of the conduction band can be approximated by

$$f(E_g) = \exp(-E_g/2k_BT),$$

and that, in general, the probability of an electron being found in the conduction band at an energy ΔE above the Fermi level is

$$f(\Delta E) = \exp(-\Delta E/k_BT).$$

Exercise 7.2 Temperature dependence of conductivity in intrinsic semiconductors
A sample of Ge exhibits intrinsic conductivity at 300 K. If the absorption edge is at 1771 nm, estimate the temperature rise that will result in a 30% increase in conductivity.

Exercise 7.3 Electronic properties of gallium arsenide, silicon and germanium
Compare the known electronic properties (e.g. band gap, electron mobility, conductivity, optical properties, etc.) of silicon, germanium and gallium arsenide. Indicate engineering applications where GaAs has a distinct advantage over the others. If the E–k relationship for the bottom of the conduction band of a specimen of GaAs is of the form $E = Ak^2$, where $A = 7.5 \times 10^{-38}\,\text{J}\,\text{m}^2$ find the effective mass of the conduction electrons in kilograms, and as a ratio of the free electron rest mass.

Exercise 7.4 Electron band gap and conductivity at finite temperature
Calculate the intrinsic conductivity at 300 K of the compound semiconductor gallium antimonide, GaSb, for which the band gap is 0.7 eV, assuming the number density of electron states close to the top of the valence band is $2.5 \times 10^{25}\,\text{m}^{-3}$ and the electron and hole mobilities are $2.3\,\text{m}^2\text{V}^{-1}\text{s}^{-1}$ and $0.01\,\text{m}^2\text{V}^{-1}\text{s}^{-1}$ respectively.

If donor impurities add one conduction electron per atom of impurity how much n-type impurity is required to give GaSb a conductivity of $100\,\text{S}\,\text{m}^{-1}$? Would the same quantity of p-type impurity (instead of n-type impurity) suffice to increase the conductivity to this value?

Exercise 7.5. Impurity levels and resistivity of semiconductors
A semiconductor is doped with 1.0×10^{22} donors and 5.0×10^{21} acceptors per cubic metre. The donor and acceptor levels are 0.01 eV from their respective band edges. If the carrier mobility is $0.2\,\text{m}^2\text{V}^{-1}\text{s}^{-1}$, calculate the resistivity at a temperature of 20 K. (Hint: look up the effect of 'compensation'.)

Exercise 7.6. Effect of temperature on pn junction
What effect has an increase in temperature on the current–voltage characteristics of the pn junction diode shown in the figure below? Discuss these effects both in

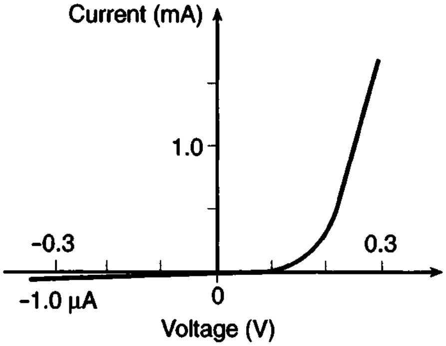


Figure 7.20 Current–voltage characteristics for a pn junction, $i = -1.0 \times 10^{-6}$ A at -0.3 V.

the forward-bias and the reverse-bias regions. Assuming a band gap of 0.7 eV calculate the current through the diode under a reverse bias of 0.3 V at a temperature of 60°C, given that Fig. 7.20 is correct at 20°C with a current of $-1.0\,\mu$A at a reverse bias voltage of -0.3 V.

8 ELECTRICAL AND THERMAL PROPERTIES OF MATERIALS

OBJECTIVE

In previous chapters, we have built up an understanding of the electronic properties of materials on the microscopic scale. These theories have involved first considering the electrons as classical particles of a free electron gas and later as free electron waves contained within the material. We then found that the electrons occupy allowed energy bands and we were able to distinguish between on the one hand metals and on the other semiconductors and insulators on the basis of their electron band structures. We also found that these electron bands were anisotropic and so plots of allowed energy against position in *k*-space were necessary. Now we must bring all of these ideas together to account for the macroscopic electrical and thermal properties of materials. In this chapter, therefore, we look at the relationship between macroscopic measurable electrical properties and the underlying electronic properties such as mobility, effective mass and number density of electrons. Then we look at various thermoelectric effects which span the interface between electrical and thermal effects. Finally, we discuss the phenomenon of thermoluminescence which bridges the gap between thermal and optical properties of materials, and therefore provides a link to the next chapter on optical properties.

8.1 MACROSCOPIC ELECTRICAL PROPERTIES

How can the macroscopic properties of materials be described on the basis of the preceding microscopic theories?
Our objective here is to describe the macroscopic electrical properties of materials and then explain these through models of the microscopic mechanisms inside the material. The familiar macroscopic properties are the conductivity and the Hall effect.

8.1.1 Ohm's law and conductivity

Can the relation between current and voltage be explained on the basis of the classical free electron theory?
Although Ohm's law is often written in the elementary form $V = IR$ where I is the electrical current flowing in a resistance R under an applied voltage V, the law can also be written in the equivalent form,

$$J = \sigma\xi, \tag{8.1}$$

where J is the current density, σ is the conductivity and ξ is the electric field. As shown in Section 3.2.1 the current density J can be written in terms of the number density of electrons N, their charge e and their average drift velocity v:

$$J = Nev, \tag{8.2}$$

and the drift velocity v can be expressed in terms of the mobility μ which is defined as the drift velocity per unit field. (We have already encountered this mobility briefly in Example 3.3.)

$$v = \mu\xi. \tag{8.3}$$

Combining these equations gives the conductivity as

$$\sigma = Ne\mu \tag{8.4}$$

and so,

$$J = Ne\mu\xi. \tag{8.5}$$

In metals, the charge carriers are electrons and so we are concerned in this case only with the number density, charge and mobility of electrons. In semiconductors, both electrons and holes contribute to the conduction, as discussed in Section 7.2.2, so that using a similar equation for the current density in terms of both contributions from holes and electrons leads to

$$J = (N_e\mu_e + N_h\mu_h)e\xi. \tag{8.6}$$

In order for Ohm's law to hold, we see that neither n nor μ can be a function of electric field ξ. In fact, under certain conditions both n and μ can become functions of ξ and Ohm's law is no longer valid. We may, however, regard the above equation as a more generalized form of Ohm's law in which this dependence can be taken into account.

8.1.2 Temperature dependence of conductivity in metals

Can the temperature dependence of conductivity in metals be described by the classical free electron theory?

If we assume one type of charge carrier for simplicity, then the temperature dependence of the conductivity is dependent on the temperature dependence of N and μ, because e is clearly constant according to our present understanding. In a metal N is the density of valence (= conduction) electrons. This has a value of typically $N = 10^{28}\ \text{m}^{-3}$ in a metal, and is largely temperature independent. Therefore the temperature dependence of conductivity should be due to a temperature dependence of mobility.

The mobility of electrons in metals is of the order of $\mu = 10^{-3}$ to 10^{-1} $\text{m}^2\,\text{V}^{-1}\,\text{s}^{-1}$, and so this leads to a conductivity σ of typically 10^6 to 10^8 $(\Omega\,\text{m})^{-1}$. All of the observed temperature dependence of σ in metals arises from the temperature dependence of the electron mobility μ which is affected by phonon scattering and impurity scattering of electrons in the metal. The classical free electron theory does not give any inherent indication of how μ should vary with temperature. Therefore the temperature dependence of electrical conductivity can only be described by an ad hoc variation of mobility with temperature in the classical free electron model.

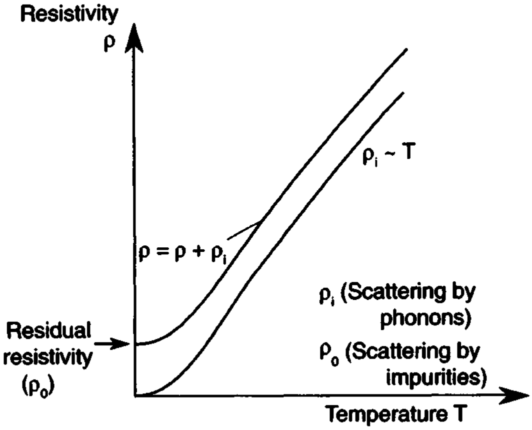


Figure 8.1 Resistivity as a function of temperature in metals. This consists of two components; one due to impurity scattering ρ_0 which persists even at zero temperature, and one due to phonon scattering ρ_i. The relation $\rho = \rho_0 + \rho_i$ is known as Matthieson's rule.

8.1.3 Temperature dependence of conductivity in semiconductors

Can the temperature dependence of conductivity in semiconductors be described by the classical electron theory?

In intrinsic semiconductors, the number density of charge carriers increases with temperature according to the equation

$$N = N_0 \exp\left(\frac{-E_g}{2k_B T}\right), \tag{8.7}$$

where E_g is the band gap. The above equation assumes that the Fermi level is in the middle of the band gap. This equation shows that there is an increase in the number density of conduction electrons with temperature. In addition, there is a change in mobility of the charge carriers with temperature, but this is less significant than the change in charge carrier density. Therefore in semiconductors the temperature variation of N dominates the temperature dependence of conductivity.

8.1.4 Temperature dependence of mobility

How can the temperature dependence of mobility of electrons be explained?

If we return to the classical description of electrons moving in a material, their motion is continually disrupted by scattering. If the mean free time between collisions is τ, the charge e and the mass m, then the mobility μ is given by

$$\mu = \frac{e\tau}{m} = \frac{e}{\gamma}, \tag{8.8}$$

and it can be seen that it is the temperature dependence of τ which determines the mobility, or alternatively we can view this as the temperature dependence of the resistive coefficient γ in the equation of motion of the electrons. Therefore, in a metal γ increases with temperature leading to a reduction in mobility μ and so a decrease in conductivity σ.

The classical model gives no indication of the temperature dependence of γ, although it is reasonable to suppose that, as the temperature is raised, the increased vibrations of the lattice will cause more collisions with the free electrons and contribute to a higher resistive coefficient γ or shorter mean free time τ.

8.1.5 Different types of mobility

How can we define electron mobility in a material?

There are four different kinds of mobility of electrons in a material that must be distinguished.

(i) Microscopic mobility

$$\mu_{\text{mic}} = \frac{v}{\xi}. \tag{8.9}$$

This is defined for a particular electron moving with velocity v in an electric field ξ. It therefore cannot easily be experimentally verified, and so remains only a concept from which a more practical description of collective mobility of electrons can be developed.

(ii) Conductivity mobility

$$\mu_{\text{con}} = \frac{e\tau}{m^*}. \tag{8.10}$$

This is the macroscopic or average mobility which is determined from measurement of electrical conductivity σ

$$\sigma = Ne\mu_{\text{con}}, \tag{8.11}$$

assuming N and e are both known.

(iii) Hall mobility

$$\mu_{\text{H}} = \sigma R_{\text{H}} = \frac{\sigma \xi_{\text{Hall}}}{\mu_0 JH} \tag{8.12}$$

is the mobility of charge carriers as determined from a Hall effect measurement.

(iv) Drift mobility

$$\mu_{\text{d}} = \frac{d}{\xi t}. \tag{8.13}$$

This is determined from measurement of the time t required for carriers to travel a distance d in the material under the action of an electric field ξ.

8.2 QUANTUM-MECHANICAL DESCRIPTION OF CONDUCTION ELECTRON BEHAVIOUR

Do all 'conduction' electrons actually contribute to the electrical conductivity?
As we have shown in Chapter 5 electrons in a material behave not like classical particles but like waves [1]. This leads to properties which can be different from classical expectations. In the absence of an electric field, the valence electrons in a metal have no net or preferential velocity in any direction. If we plot the vectors of these electrons in velocity space, then for a free electron metal we obtain a velocity sphere, the surface of which corresponds to the Fermi velocity. All points inside the Fermi sphere are occupied. Integrating over the entire sphere we obtain zero drift velocity.

When an electric field is applied, the Fermi sphere is displaced as shown in Fig. 8.2. Still the majority of electron velocities cancel, but now some are uncompensated and it is these electrons which cause the electric current. We note therefore the important result that only certain specific electrons which are close to the Fermi surface can contribute to the conduction mechanism. Note that a similar effect was found for heat capacity where only those electrons within $k_B T$ of the Fermi level could contribute to the heat capacity.

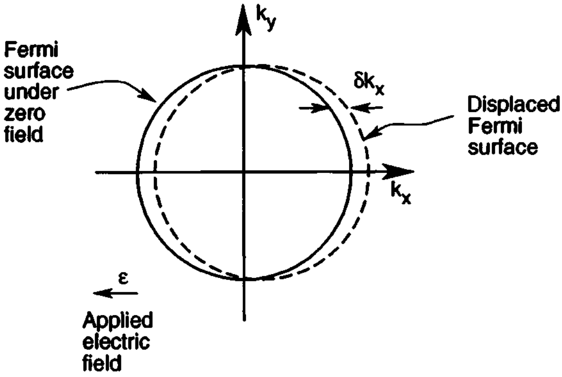


Figure 8.2 Velocity of free electrons within the Fermi sphere under zero electric field and under an applied field ξ along the x direction.

8.2.1 Quantum corrections to the conductivity in Ohm's law

How is Ohm's law modified if only the electrons close to the Fermi surface contribute to the conductivity of a metal?
The highest energy that electrons can take in a metal in its ground state is the Fermi energy E_F. We also know that the density of occupied states is highest around E_F, since for a free electron model, for example, the density of states $D(E)$ has the following form, as shown in Section 4.4.7:

$$D(E) = \frac{V}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2}. \tag{8.14}$$

This means that only a small change of energy ΔE is needed to raise a large number of electrons above the Fermi level. We will consider that the velocity of the uncompensated electrons under the action of the field ξ is close to the Fermi velocity. This will be a reasonable simplifying assumption. With this in mind, we can calculate the electric conductivity σ, based on quantum-mechanical considerations.

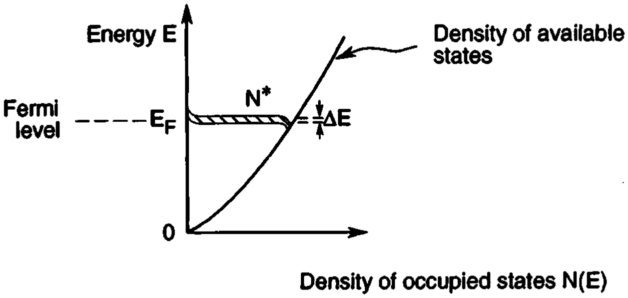


Figure 8.3 Population density of occupied electron states versus energy for free electrons according to the free electron theory.

Our Ohm's law equation of Section 8.1.1 needs to be slightly modified to take into account the fact that not all free electrons contribute to the conductivity. Hence,

$$J = N^* e v_F, \tag{8.15}$$

where v_F is the velocity of electrons at the Fermi level and N^* is the number of displaced electrons, that is those in the region of Fig. 8.2 which contribute to the conductivity.

8.2.2 Number of 'conduction' electrons contributing to conduction

How can we find out how many of the so-called conduction electrons in a metal actually contribute to electrical conduction?

We need to obtain an expression for N^*. This will clearly be dependent on the density of states at the Fermi level $D(E)$ and the displacement energy ΔE

$$N^* = D(E)\Delta E, \tag{8.16}$$

and consequently

$$J = D(E)\Delta E e v_F \tag{8.17}$$

$$= v_F e D(E)\frac{dE}{dk}\Delta k. \tag{8.18}$$

The term dE/dk is determined from the energy versus wave vector diagram for the given case. For free electrons, we have $E = \hbar^2 k^2/2m$ and hence

$$\frac{dE}{dk} = \frac{\hbar^2 k}{m} = \hbar v_F, \tag{8.19}$$

and this yields,

$$J = v_F^2 e D(E)\hbar\Delta k. \tag{8.20}$$

8.2.3 Displacement of a spherical Fermi surface under the action of an electric field

How does the displacement of the electron wave vectors depend on other factors such as the mean free time of electrons between collisions?

We will examine the displacement Δk of the Fermi sphere under the influence of an electric field ξ. Since we know that $m\,\mathrm{d}v/\mathrm{d}t = e\xi$, and since $p = \hbar k$ is the momentum, it follows that the force on the electrons can be expressed as,

$$F = m\frac{\mathrm{d}v}{\mathrm{d}t} = \hbar\frac{\mathrm{d}k}{\mathrm{d}t} = e\xi. \tag{8.21}$$

Hence,

$$\mathrm{d}k = \frac{e\xi}{\hbar}\mathrm{d}t \tag{8.22}$$

or,

$$\Delta k = \frac{e\xi}{\hbar}\Delta t = \frac{e\xi}{\hbar}\tau, \tag{8.23}$$

where τ is the mean free time of the electrons between collisions. With this expression for Δk we arrive at the following expression for the current density,

$$J = v_{\mathrm{F}}^2 e^2 D(E)\xi\tau. \tag{8.24}$$

Only the projections of v_{F} along the direction of the electric field ξ, that is $v_{\mathrm{F}}\cos\theta$, contribute to the current.

$$J = e^2 D(E)\xi\tau\int_{-\pi/2}^{+\pi/2}(v_{\mathrm{F}}\cos\theta)^2\frac{\mathrm{d}\theta}{\pi} \tag{8.25}$$

$$= \tfrac{1}{2}e^2 D(E)\xi\tau v_{\mathrm{F}}^2. \tag{8.26}$$

For a spherical Fermi surface there is a slight correction that gives

$$J = \tfrac{1}{3}e^2 D(E)\xi\tau v_{\mathrm{F}}^2, \tag{8.27}$$

and finally, the conductivity is given by $\sigma = J/\xi$, so that

$$\sigma = \tfrac{1}{3}e^2 v_{\mathrm{F}}^2\tau D(E). \tag{8.28}$$

This quantum-mechanical expression for conductivity shows that not all ‘conduction’ electrons contribute to the conductivity, but only those close to the Fermi surface. In addition, the conductivity is determined by the density of states near the Fermi level. For metals such as copper, which has one conduction electron per atom, this density is high, leading to high conductivity. For bivalent metals such as calcium, which has two conduction electrons per atom, this density is small, leading to a relatively low conductivity. Therefore, it is the density of states at the Fermi surface, and not the classically expected total number of conduction electrons, that determines the conductivity of a material.

8.3 DIELECTRIC PROPERTIES

How can we represent the response of a nonconducting material to an electric field? Most electronic applications involve the use of alternating electric fields or currents. In these cases the atoms in insulators oscillate under the action of the applied electric field, and these oscillations can be expressed in terms of the dielectric constant, ε. This is often expressed in terms of real and imaginary components $\varepsilon_1, \varepsilon_2$

$$\varepsilon = \varepsilon_0 \varepsilon_r = \varepsilon_1 + i\varepsilon_2. \tag{8.29}$$

This dielectric 'constant' is actually dependent on the frequency of the applied electric field. When considering its dependence on the frequency of electromagnetic radiation it is often represented as $\varepsilon(\omega)$.

8.3.1 Polarization

How do we quantify the displacement of charge in a material under the action of an electric field?

The net result of applying an electric field to an insulator is to cause the positive and negative charges within the material to become displaced in opposite directions and the material consequently becomes electrically polarized. The polarization can result from the relative displacement of the electrons and ionic cores, or alternatively from the relative displacement of positive and negative ionic cores [2].

The force F on a charge e under the action of an electric field is,

$$F = e\xi \tag{8.30}$$

and it is this force which causes polarization of a material by displacing the positive and negative charges within an atom in opposite directions, or by displacing the ionic cores within the lattice.

The electric polarization of the material is denoted by the symbol P. This is an electric dipole moment per unit volume, which is measured in coulomb metres per cubic metre (or effectively coulombs per square metre). The equation for P is

$$P = Np, \tag{8.31}$$

where p is the dipole moment of an individual atom and N is the number of atoms per unit volume. P can also be defined as the surface density of charge which appears on the faces of the specimen when placed in a field. The polarization can be expressed in terms of the electric field by the equation,

$$P = \chi_e \varepsilon_0 \xi = (\varepsilon_r - 1)\varepsilon_0 \xi, \tag{8.32}$$

where ε_o is the permittivity or dielectric constant of free space and ε_r is the relative permittivity. We see therefore that the dielectric constant is a measure of the amount of electric polarization induced by unit field strength.

Alternatively, the dielectric permittivity can be expressed as the ratio of polarization to electric field strength P/ξ. A high dielectric constant means that a material is easily polarized in an electric field. Typical values of the relative dielectric permittivity ε_r are in the range 1.0 to 10 (dimensionless), although its value can be much higher in some special materials, for example ε_r is 94 in TiO_2 [3. p. 267].

8.3.2 Dielectric field strength

How high an electric field can a material withstand before it suffers electrical breakdown?

The dielectric field strength is a measure of the largest electric field strength that an insulating material can sustain before the electrostatic forces holding the atoms in place are overcome. Once this happens the material suffers electrical breakdown and suddenly becomes an electrical conductor. This breakdown voltage has been discussed in Chapter 1, for example. Typical values of the dielectric strength are in the range of megavolts per metre. However, it should be noted that the breakdown strength often increases with frequency, and in particular for most materials breakdown is somewhat inhibited above 10^8 Hz [3. p. 272].

8.3.3 Electrical properties of noncrystalline materials

What about materials that do not have a regular crystalline lattice?

We have looked in detail at the electrical properties of crystalline materials. In these cases, the regular periodicity of the atoms on the lattice sites leads to relative simplicity of calculation. However, we need not restrict ourselves entirely to these materials since electrical conduction also occurs in polymers, ceramics, and amorphous materials and there is a need to provide theories and models for these materials also.

8.3.4 Polymers

Why are some polymers found to be good conductors?

Most polymers are insulators, of course, but conducting polymers exist which have electrical properties resembling those of conventional metals or semiconductors [4]. Polyacetylene contains a high degree of crystallinity and a relatively high conductivity compared with other polymers. Trans-polyacetylene has a conductivity that is comparable to silicon. The electron band structure of this polymer has even been calculated and it has been found that when all carbon bond lengths are equal, this material has a band structure which is reminiscent of a metal. When the carbon bonds alternate in length it is found that band gaps appear in the structure.

8.4 OTHER EFFECTS CAUSED BY ELECTRIC FIELDS, MAGNETIC FIELDS AND THERMAL GRADIENTS

What other effects occur when a material is subjected to external influences such as electric, magnetic and thermal fields?

There are a number of other phenomena that occur when a material is subjected to electric, magnetic or thermal fields. We will mention only the most important of these: magnetoresistance, the Seebeck effect, the Peltier effect, the Nernst effect and the Ettingshausen effect.

8.4.1 Magnetoresistance

What happens to the electrical resistance when a material is subjected to a magnetic field?

The magnetoresistance is the change in electrical conductivity associated with an applied magnetic field. It cannot be explained on the classical (Drude) electron model since with one carrier conductivity, constant relaxation time τ and constant effective mass the magnetoresistance is identically zero.

An example of zero magnetoresistance occurs in our explanation of the Hall effect to which the deflection of charge carriers causes the build-up of a transverse electric field which exactly counteracts the effect of the magnetic field. In this case, under equilibrium conditions the motion of the charge carriers is identical in the presence or absence of a magnetic field, because of this transverse electric field, resulting in zero magnetoresistance.

However, if all charge carriers do not have the same properties the current flow is disturbed by the presence of a magnetic field and some of the charge carriers travel a longer distance between electrodes than in the absence of a field. This leads to a larger observed resistivity and the difference between the zero field resistivity and the measured resistivity under the applied magnetic field known as magnetoresistance.

The resistivity is defined as $\rho = \xi/J$ where ξ and J, the electric field and current density, are measured along the same direction. Since the resistivity is, in general, dependent on the magnetic field we find that

$$\rho = \xi/J = \rho_0(1 + \omega_c^2 \tau), \tag{8.33}$$

where $\omega_c = eB_\perp/m$ and $B_\perp$ is the magnetic induction perpendicular to the direction of measurement of current,

$$\rho = \rho_0\left(1 + \frac{e^2 B_\perp^2}{m^2}\tau\right) \tag{8.34}$$

and τ is the mean free time of the electrons between collisions. The magnetoresistive term is then simply

$$\rho_{\text{mag}} = \rho_0 \frac{e^2 B^2}{m^2}\tau. \tag{8.35}$$

8.4.2 Thermoelectric power (Seebeck effect)

What happens to the voltage across a material when it is subjected to a temperature field (temperature gradient)?

If a material is subjected to a temperature gradient, the energy of the carriers at the hot end is greater than at the cold end and this leads to a carrier concentration gradient along the material. Displaced charge resulting from this concentration gradient generates a counteracting electric field ξ until the total current becomes zero. The magnitude of this electric field in terms of the voltage per degree difference is known as the thermoelectric power α. In a metal,

$$\alpha = \frac{\mathrm{d}V}{\mathrm{d}T} \cong \frac{\pi^2 k_B^2 T}{eE_F}, \tag{8.36}$$

where E_F is the Fermi energy. In a metal, α is typically a few microvolts per Kelvin.

In a semiconductor, for an n-type material

$$\alpha = -\frac{k_B}{e}\left(A + \frac{E_c - E_F}{k_B T}\right), \tag{8.37}$$

and for a p-type material

$$\alpha = -\frac{k_B}{e}\left(A + \frac{E_F - E_v}{k_B T}\right). \tag{8.38}$$

In a semiconductor, α is typically a few millivolts per Kelvin. Here A is a constant which depends on the specific scattering mechanism, $A = 2$ for lattice scattering and $A = 4$ for charged impurity scattering, E_c is the energy level at the bottom of the conduction band E_v is the energy at the top of the valence band and E_F is the Fermi energy.

The Seebeck effect is utilized in the thermocouple which is used for measuring temperature. The thermoelectic power α is determined from the open-circuit electric field ξ caused by a temperature gradient $\mathrm{d}T/\mathrm{d}x$

$$\alpha = \frac{\xi}{(\mathrm{d}T/\mathrm{d}x)} = \frac{J}{J_Q}\frac{K}{\sigma}. \tag{8.39}$$

8.4.3 Peltier effect

What happens to the temperature gradient when a current flows in a material?
When a current flows in a material, a temperature gradient is developed. This, of course, is the inverse of the Seebeck effect and is used in some cases for temperature control. The Peltier coefficient π is simply the ratio of the electrical current density J to the thermal current density J_Q

$$\pi = \frac{J_Q}{J} = \frac{-K(\mathrm{d}T/\mathrm{d}x)}{\sigma\xi}. \tag{8.40}$$

8.4.4 Nernst effect

What happens when both a magnetic field and a temperature field (temperature gradient) are applied simultaneously to a material?
When a magnetic field is applied at right angles to a temperature gradient, the diffusing charge carriers are deflected in the same way as when the magnetic field is applied at right angles to a conventional electric current. The result is a Nernst voltage. However, since charge carriers of both signs diffuse in the same direction the polarity of the Nernst voltage is not dependent on the sign of the charge carrier.

8.4.5 Ettingshausen effect

Do we therefore also get a transverse temperature field (temperature gradient) in the Hall effect?
In the Hall effect, the application of a magnetic field normal to the passage of an electric current leads not only to a transverse voltage but also to a transverse temperature gradient. The appearance of this temperature gradient is known as the Ettingshausen effect. This arises because charge carriers with different energies (velocities) are deflected differently by the magnetic field. This is a small effect which adds to the Hall voltage.

8.5 THERMAL PROPERTIES OF MATERIALS

Which factors determine the thermal properties of materials?
The thermal properties of materials can be determined principally by the electrons, as in the case of thermal conductivity of metals, or principally by the lattice, as in the case of thermal conductivity of insulators or of specific heat capacity [5].

8.5.1 Thermal conductivity

How does thermal conduction take place in materials?
Thermal conductivity of materials varies from $6 \times 10^3\,\mathrm{W\,m^{-1}\,K^{-1}}$ in silver and copper to $5 \times 10^{-2}\,\mathrm{W\,m^{-1}\,K^{-1}}$ in sulphur [6]. In the case of metals, the thermal conduction mechanism is similar to the electrical conduction mechanism and proceeds via the free electrons which migrate throughout the material. In semiconductors, conduction can take place by the electrons which are thermally stimulated into the conduction band.

In insulators, another mechanism must be involved and in this case the thermal conduction is due to phonons which are created at the hot part of a solid and destroyed at the cold part. These phonons provide the mechanism by which energy is transfered though the material. In metals the phonon contribution to thermal conductivity is also present, but the electronic contribution is so much greater that in these cases the phonon contribution is neglected.

We have already defined the thermal conductivity K in Section 1.5.1. It is the ratio of the heat flux $J_Q(=Q/A)$ to the thermal gradient $\mathrm{d}T/\mathrm{d}x$

$$K = -\frac{1}{(\mathrm{d}T/\mathrm{d}x)}, \tag{8.41}$$

and its units are $\mathrm{J\,s^{-1}\,m^{-1}\,K^{-1}}$.

8.5.2 Mechanism of thermal conduction

How can we develop a theory of thermal conduction based on our knowledge of the electronic properties?
If we begin from the assumption that thermal conduction can arise from both the motion of free electrons and phonons, we can derive a theory of the thermal

conductivity. Again, as in electrical conductivity, only those electrons close to the Fermi surface can contribute to the thermal conductivity.

8.5.3 Thermal conductivity of metals

How is heat conducted in metals?

From quantum mechanics we have shown that in the electrical conduction process, only those electrons close to the Fermi surface can absorb energy and hence contribute to the conductivity. The same is also true for the thermal conductivity. Therefore, to a very good approximation the velocity of those electrons contributing to the thermal conductivity is the velocity at the Fermi level v_F.

The number of participating electrons N^* is determined by the population density at the Fermi energy $N(E_F)$. To a first approximation, this is about 1% of the number of free electrons per unit volume. As we have already seen in Section 4.5.1 this gives the following contribution to the electronic heat capacity,

$$C_V^e = \frac{\pi^2}{2} N k_B \frac{T}{T_F}, \tag{8.42}$$

$$C_V^e = \frac{\pi^2}{2} \frac{N k_B^2 T}{E_F}. \tag{8.43}$$

We will now show that a relationship exists between the specific heat capacity and the thermal conductivity.

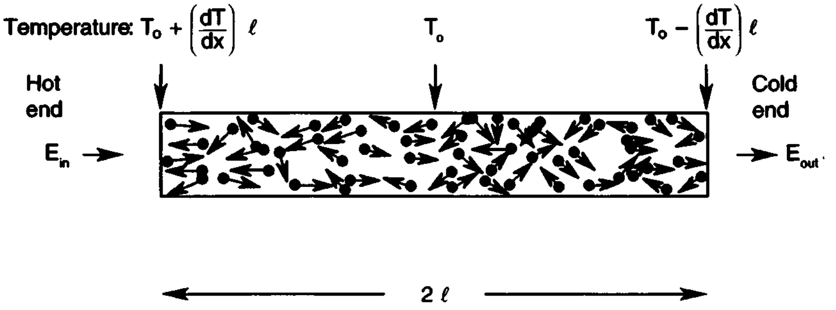


Figure 8.4 Thermal conduction by electrons in a linear section of material under the action of a thermal gradient.

We consider the situation depicted in Fig. 8.4. where ℓ is the mean free path length of the electrons between collisions. Assume that heat flow is linear only along the x direction and is zero in the plane perpendicular to the x direction. In this case if we consider a section of length 2ℓ with unit cross-sectional area, the heat flux is,

$$J_Q = E_{out} - E_{in}, \tag{8.44}$$

where E_{in} is the heat energy flowing in per unit time at the left end, and E_{out} is the heat energy flowing out per unit time at the right end:

$$= 2z\frac{3}{2}k_B\ell\frac{dT}{dx}, \tag{8.45}$$

where z, the number of electrons per unit time impinging on the end face, is

$$z = \tfrac{1}{6}N^*v. \tag{8.46}$$

We might reasonably assume that N^*, the number density of free electrons contributing to thermal conductivity, is similar to the number contributing to the thermal heat capacity $N^* \cong 2C_V^e/3k_B$, since in both cases the electrons must be able to absorb heat energy:

$$J_Q = -\frac{N^*}{2}vk_B\ell\frac{dT}{dx} = \frac{1}{3}C_V^e v\ell\frac{dT}{dx}. \tag{8.47}$$

From the earlier equation for the thermal flux $J_Q = K\,dT/dx$, it is clear that

$$K = \tfrac{1}{3}C_V^e v\ell. \tag{8.48}$$

Notice the important result that the thermal conductivity and electronic heat capacity are related. This is at first a surprising result. Here, ℓ is the mean path length between two electron collisions and v is the average velocity of electrons.

Since only electrons close to the Fermi level can take part in this we will replace N^* with

$$N_F\left(=N(E_F)k_BT = \frac{\pi^2}{6}\frac{N}{E_F}k_BT\right)$$

the number density of electrons at the Fermi surface, replace ℓ with ℓ_F the mean free path of electrons at the Fermi level, and replace v with v_F the velocity of electrons at the Fermi surface. Substituting into the above expression for C_V^e and using the quantum-mechanical expression for heat capacity

$$C_V^e = \frac{\pi^2}{2}\frac{Nk_B^2T}{E_F}$$

from eqn (4.55), the equation for K becomes

$$K = \frac{\pi^2}{6E_F}Nk_B^2Tv_F\ell_F. \tag{8.49}$$

Therefore, the thermal conductivity increases with mean free path ℓ_F, number of electrons per unit volume, and velocity of electrons at the Fermi surface v_F. Remembering that $E_F = m^*v_F^2/2$, and that $\ell_F = \tau v_F$

$$K = \frac{\pi^2}{3m^*}Nk_B^2T\tau. \tag{8.50}$$

The thermal conductivity increases with N_F, T and τ and decreases with m^*. We can reasonably have expected the conductivity to have increased with the density of states at the Fermi level and the mean free time between electron collisions. Similarly we expect conductivity to increase with increasing electron mobility (or decreasing effective mass). However, we might at first be surprised at the increase with temperature. This, however, is taken care of in the temperature dependence of τ, which can be used to compensate for this.

8.5.4 Thermal conductivity of insulators

How are thermal conductivity and heat capacity related in insulators?

Once again the thermal conductivity K is related to the heat capacity by the expression

$$K = \tfrac{1}{3} C_v^l v \ell, \tag{8.51}$$

but now C_v^l is the lattice heat capacity of the phonons, v is the phonon velocity and ℓ is the phonon mean free path.

8.6 THERMOLUMINESCENCE

What other thermal properties are of interest?

We now go on to consider other thermal properties that have important but less wide-ranging application. One of these is thermoluminescence. Thermoluminescence is the emission of electromagnetic radiation, in the visible spectrum, when certain materials are heated [7]. These materials must be either insulators or semiconductors, and they must have a large number of electrons trapped in impurity states in the band gap. The emitted radiation from thermoluminescent materials is different from the well-known black-body radiation (incandescence) which depends on the fourth power of the absolute temperature, Stefan's law. A typical thermoluminescence 'glow curve' is shown in Fig. 8.5.

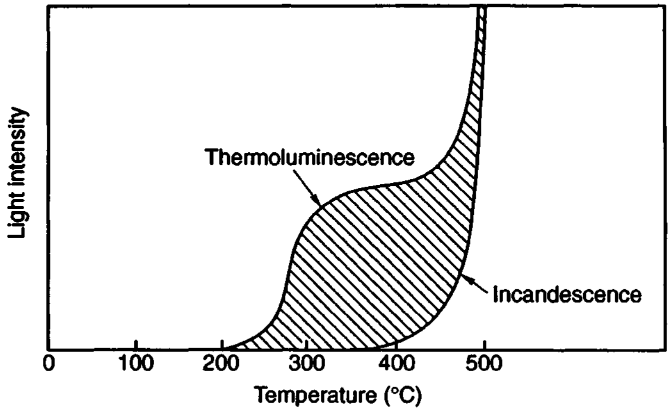


Figure 8.5 Thermoluminescent glow curve of emitted light intensity versus temperature.

8.6.1 Mechanism of thermoluminescence

What distinguishes thermoluminescence from incandescence in terms of the electronic properties?

If we have an electron band structure in which there is a band gap, with a number of isolated defects or impurity states in the band gap, and a Fermi level between the bands as shown in Fig. 8.6, then electrons can be trapped in these impurity states.

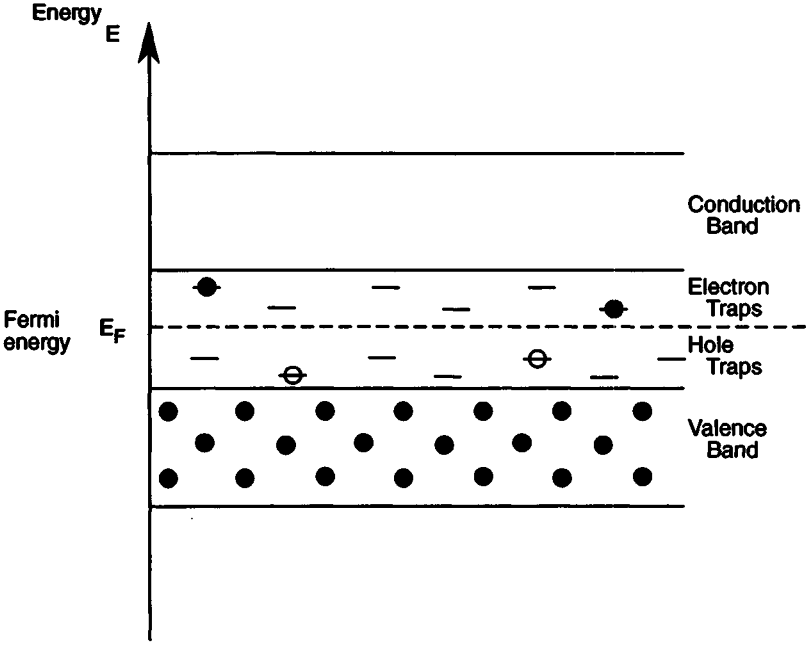


Figure 8.6 Schematic band structure diagram showing band gap with electron and hole traps. Charged particles are held in the traps for long periods. When they escape they recombine and emit light.

Electrons become trapped in these localized energy levels by being stimulated into the conduction band and then dropping down into the localized energy states in the band gap instead of back into the valence band.

The lifetime of electrons in the traps depends on a number of factors, including the prevailing temperature T and the depth of the trap below the conduction band. The lifetime can actually be many years, and this is made use of in thermoluminescent dating of pottery and other ceramics, for example, and in radiation dose monitoring using thermoluminescent sensors. The requirements for a material to be able to exhibit thermoluminescence are:

(a) presence of a band gap,
(b) presence of impurity energy states in the band gap,

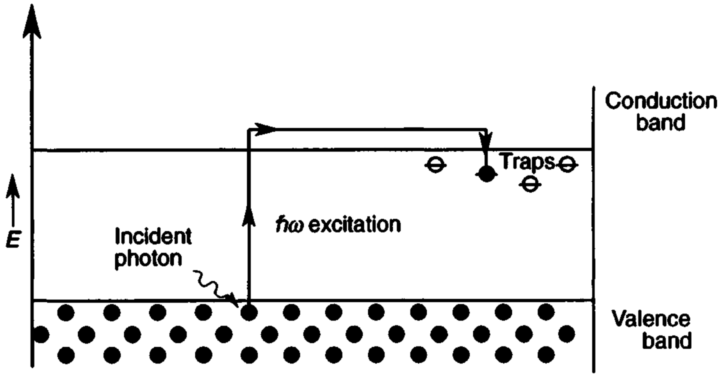


Figure 8.7 Mechanism of electron trapping in thermoluminescence. (1) Electron is excited into the conduction band; (2) electron moves freely within the conduction band; (3) electron falls into localized energy state (electron trap).

(c) long lifetime of electrons in traps,
(d) material must have been subjected to radiation to excite electrons from valence band before becoming trapped,
(e) material must not have been inadvertently heated, which could empty electrons from traps.

8.6.2 Theory of thermoluminescence

How is the light emitted in thermoluminescence?

Once we have electrons located in traps in the band gap we need to explain how this leads to the emission of light. Essentially, electrons are thermally stimulated from the traps into the conduction band and later they fall back into the valence band, emitting a photon as they do so.

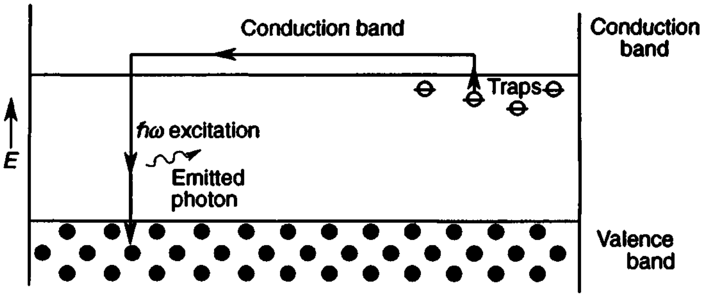


Figure 8.8 Mechanism of electron excitation and thermoluminescent emission of light. (1) Thermal stimulation from trap to conduction band, (2) Movement within conduction band, (3) Transition to valence band with photon emission.

If we have an electron located in a trapped state at an energy ΔE below the conduction band, then the probability of the electron being thermally stimulated into the conduction band in unit time is given by the Arrhenius equation,

$$p = s\exp\left(\frac{-\Delta E}{k_B T}\right), \tag{8.52}$$

where s is a constant, with dimensions time^{-1} and typically of magnitude 10^{11} to $10^{17}\,\mathrm{s}^{-1}$. This means that there is a time frame associated with the occupancy of the electron trap once the electron is there.

8.6.3 Occupation and vacation of trapped states by electrons

How does the occupancy of electron traps vary with time?

The probability of filling any state in the band gap will also be dependent on time. If $\mathrm{d}N/\mathrm{d}t$ is the rate of stimulation of electrons from traps into the conduction band, then

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -Np, \tag{8.53}$$

where N is the number of electrons in traps and p the probability of escape in unit time. This simply states that the number of events leading to stimulation of electrons into the conduction band is proportional to the number of electrons sitting in traps. Integrating this equation gives,

$$-\log_e\left(\frac{N}{N_0}\right) = \int_0^t p\,\mathrm{d}t, \tag{8.54}$$

$$N = N_0\exp(-pt), \tag{8.55}$$

and,

$$N = N_0\exp\left(-s\exp\left(\frac{-\Delta E}{k_B T}\right)t\right). \tag{8.56}$$

This is the Randall–Wilkins equation [8] which describes the number of electrons remaining in traps as a function of both time t and temperature T. We know that eventually an electron must escape from a trap, and so the integral of p over the time interval $t = 0$ to $t = \infty$ must be unity

$$\int_{t=0}^{\infty} p\,\mathrm{d}t = 1. \tag{8.57}$$

8.6.4 Lifetime of electrons in traps

How can we determine the time an electron will, on average, stay in the electron trap?

Clearly the lifetime of occupancy of an electron state is inversely proportional to the probability of a transition in unit time p. We may define this lifetime τ as a function of temperature T by

$$\tau(T) = \frac{1}{p(T)} \tag{8.58}$$

$$= \frac{1}{s} \exp\left(\frac{\Delta E}{k_B T}\right). \tag{8.59}$$

Therefore raising the temperature T decreases the expected lifetime of the electrons in the traps. This is what we should expect, since more thermal energy increases the probability of the electron escaping by thermal stimulation.

From the exponential decay equation $N = N_0 \exp(-pt)$ it is possible to define a half-life for the occupancy of the electron traps. Simply, when the number of traps remaining occupied has declined to half, $N = N_0/2$ we have the half-life of the occupancy $\tau_{1/2}$

$$\tfrac{1}{2} = \exp(-p\tau_{1/2}), \tag{8.60}$$

$$\tau_{1/2} = \frac{1}{p} \log_e 2. \tag{8.61}$$

8.6.5 Intensity of light emitted during thermoluminescence

What factors determine the intensity and frequency of light emitted during thermoluminescence?

The intensity of light emitted during thermoluminescence is dependent on the rate of emptying of the electron traps dN/dt. If we assume that every electron removed from a trap enters the bottom of the conduction band and then instantaneously falls back to the top of the valence band with emission of a photon of energy equal to the band gap energy, then the following will occur:

(i) Emission of light of a single frequency, $\nu = E_g/h$,
(ii) Intensity of light will be equal to the rate of emptying of electron traps.

$$I = -\frac{dN}{dt} \tag{8.62}$$

$$= Np \tag{8.63}$$

$$= Ns \exp\left(-\frac{\Delta E}{k_B T}\right), \tag{8.64}$$

where ΔE is the energy difference between the traps and the conduction band.

8.6.6 Emission of light on heating

How does the emitted light intensity depend on time and temperature for a single type of electron trap?

Suppose then the temperature of the specimen is raised at a constant rate,

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \beta, \tag{8.65}$$

then the fractional change in occupancy $\mathrm{d}N/N$ is,

$$\frac{\mathrm{d}N}{N} = -p\,\mathrm{d}t \tag{8.66}$$

$$\frac{\mathrm{d}N}{N} = -s\exp\left(\frac{-\Delta E}{k_\mathrm{B}T}\right)\mathrm{d}t. \tag{8.67}$$

Replacing $\mathrm{d}t$ with $\mathrm{d}T/\beta$ gives

$$\frac{\mathrm{d}N}{N} = -\frac{s}{\beta}\exp\left(\frac{-\Delta E}{k_\mathrm{B}T}\right)\mathrm{d}T. \tag{8.68}$$

Integrating this expression to give the number of occupied states leads to the following expressions,

$$\log_\mathrm{e}\left(\frac{N}{N_0}\right) = -\int_{T_1}^{T_2}\frac{s}{\beta}\exp\left(\frac{-\Delta E}{k_\mathrm{B}T}\right)\mathrm{d}T, \tag{8.69}$$

$$N = N_0\exp\left\{-\int_{T_1}^{T_2}\frac{s}{\beta}\exp\left(\frac{-\Delta E}{k_\mathrm{B}T}\right)\mathrm{d}T\right\}, \tag{8.70}$$

and since we have stated that the intensity of radiation is given by

$$I = Ns\exp\left(\frac{-\Delta E}{k_\mathrm{B}T}\right)$$

we are led to the conclusion

$$I = N_0 s\exp\left\{-\int_{T_1}^{T_2}\frac{s}{\beta}\exp\left(\frac{-\Delta E}{k_\mathrm{B}T}\right)\mathrm{d}T\right\}\exp\left(\frac{-\Delta E}{k_\mathrm{B}T}\right) \tag{8.71}$$

This emission assumes a single type of trap at an energy ΔE below the conduction band, a constant rate of change of temperature and a constant value of s for all traps of the given type.

8.6.7 Location of the peaks in thermoluminescent intensity

How can the depths of the electron traps below the conduction band be studied from the thermoluminescent glow curve?

When intensity of emission I is measured as a function of temperature T as the temperature is swept at a fixed rate, peaks in the intensity will correspond to the depth of electron traps below the conduction band.

An empirical relationship has been given between the depth ΔE in electron volts (eV) and the peak temperature T^* by Urbach [9]

$$\Delta E = \frac{T^*}{500}. \tag{8.72}$$

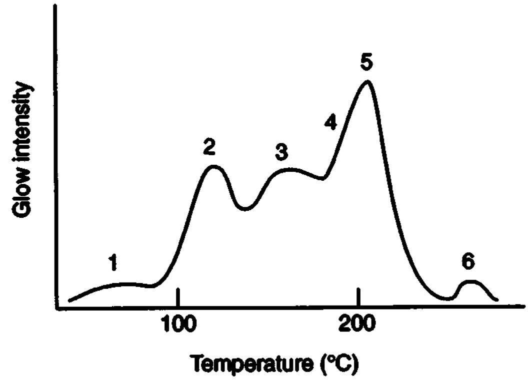


Figure 8.9 Thermoluminescent glow curve for TLD 100 dosimeter with several intensity peaks corresponding to several depths of electron traps. Reprinted from R. K. Bull, *Nucl. Tracks and Radiation Meas.*, **11**, p. 108, copyright 1986, with kind permission from Elsevier Science.

8.6.8 Applications of thermoluminescence

How is thermoluminescence used in its primary applications of radiation dosimetry and archaelogical dating?

Thermoluminescence finds applications in radiation dosimetry [10], geological, and cosmological dating and in the investigation of radiation damage in solids. The thermoluminescent glow curve gives information about the total radiation dose absorbed by the material. For dosimetry this is all that is required. The elapsed time since formation of a solid can be calculated for dating purposes assuming a certain average background radiation intensity over a period of time, assuming the material has not been heated in the interim. If the material has been heated this would empty some or all of the electron traps, effectively resetting the thermoluminescent clock. A useful review of thermoluminescence and its applications has been given by Bull [11].

REFERENCES

1. L. Brillouin (1953) *Wave Propagation in Periodic Structures*, Dover Press, New York.
2. J. C. Anderson (1964) *Dielectrics*, Chapman and Hall, London.
3. L. Solymar and D. Walsh (1984) *Lectures on the Electrical Properties of Materials*, Oxford University Press.
4. J. Mort and G. Pfister (1982) *Electronic Properties of Polymers*, J. Wiley & Sons, New York.
5. C. Kittel and H. Kroemer (1980) *Thermal Physics*, 2nd edn, W. H. Freeman, San Francisco.
6. R. P. Tye (1969) *Thermal Conductivity*, Academic Press.
7. G. F. J. Garlick and A. F. Gibson (1948) *Proc. Phys. Soc.* A60, 574.
8. J. T. Randall and M. H. F. Wilkins (1945) *Proc. Roy. Soc. Lond.* A84, 366.
9. F. Urbach (1930) *Wiener Berichte.* IIA, 39, 363.

10. A. F. McKinlay (1981) *Thermoluminescence Dosimetry*. Adam Hilger, Bristol.
11. R. K. Bull (1980) *Nucl. Tracks Radiation Meas.*, 11, 105.

FURTHER READING

J. C. Anderson, K. D. Leaver, R. D. Rawlings and J. M. Alexander (1990) *Materials Science*, 4th edn, Chapman and Hall, London, Chapters 7 and 13.

R. Hummel (1993) *Electronic Properties of Materials*, 2nd edn, Springer-Verlag, Berlin. Parts II and V.

C. Kittel and H. Kroemer (1980) *Thermal Physics*, 2nd edn, W. H. Freeman, San Francisco.

N. F. Mott and E. A. Davis (1971) *Electronic Processes in Non-crystalline Materials*, Clarendon Press, Oxford.

J. W. Blatchford and A. J. Epstein (1996) *Electronic Polymers and their Applications, Amer. J. Phys.* 64 (2), 120.

EXERCISES

Exercise 8.1 Drift velocity of conduction electrons
The Fermi energy of aluminium is 12 eV and its electrical resistance at 300 K is $3 \times 10^{-8}\ \Omega\,\text{m}$. Calculate the mean free path of the conduction electrons and their mean drift velocity in a field of $10\ \text{V m}^{-1}$. (Atomic weight of aluminium = 27, density = $2700\ \text{kg m}^{-3}$).

Exercise 8.2 Conductivity in intrinsic and extrinsic semiconductors
A sample of n-type germanium contains 10^{23} ionized donors per cubic metre. Estimate the ratio at room temperature of the conductivity of this material to that of high-purity intrinsic germanium. Assume the band gap in germanium is 0.7 eV.

Exercise 8.3 Thermoluminescence and lifetime of electrons in traps
If the lifetime of trapped electrons in a particular ceramic at 273 K is 10^{10} s (320 y) calculate the depth of the electron traps below the conduction band. If the frequency parameter $s = 4.64 \times 10^{17}\ \text{s}^{-1}$, then calculate the temperature at which the peak occurs in the thermoluminscent glow curve, and calculate the lifetime of the same electrons in the same traps at a temperature of 373 K.

Use classical electron theory in the following three exercises
Exercise 8.4 Electron and lattice contributions to the thermal conductivity
The thermal conductivity of germanium at a temperature of 300 K is 80 $\text{J s}^{-1}\,\text{K}^{-1}\,\text{m}^{-1}$, and its electrical resistivity is $1 \times 10^{-5}\ \Omega\,\text{m}$. Calculate the ratio of electronic heat conductivity to lattice heat conductivity of the material.

Exercise 8.5 Classical explanation of thermal conductivity
Determine the thermal conductivity of a metal assuming the mean time between collisions of the electrons is 3×10^{-14} s at a temperature of 300 K, and the number of free electrons per cubic metre is 2.5×10^{28}.

Exercise 8.6 Mean free path of 'free' electrons

Find the mean free path of conduction electrons in silver at room temperature from the known values of heat capacity and thermal conductivity. You can assume a Fermi energy of $E_F = 5.5\,\mathrm{eV}$, a thermal conductivity of $K = 410\ \mathrm{J\,s^{-1}\,K^{-1}\,m^{-1}}$ and that the electronic heat capacity is 1% of the lattice heat capacity. The density of silver is $10\,500\,\mathrm{kg\,m^{-3}}$. [Hint: Remember that the heat capacity C_V^e in the equation $K = C_V^e v\ell/3$ is given per unit volume.]

9 OPTICAL PROPERTIES OF MATERIALS

OBJECTIVE

We have touched briefly on the optical properties of materials in the early chapters, but here we must bring together the concepts of electron structure and the known optical properties of materials. This is done by identifying the allowed energy transitions which determine the main features of the optical spectrum. This means that we need to connect measured optical properties with the allowed electron energy levels. The major classification of electron transitions is here between transitions within the same band (intraband) and transitions between different bands (interband). The former are lower-energy transitions which lead to the high reflectivity of metals in the visible spectrum. The latter are higher-energy transitions which can lead to specific colours in materials. Various methods for measuring the optical properties are discussed including both conventional static optical measurements and differential techniques under external modulation of field, temperature or stress. Finally, the specialized topics of photoluminescence, and electroluminescence are discussed.

9.1 OPTICAL PROPERTIES

What quantities need to be measured to completely determine the optical properties of materials?

In previous chapters we have shown that the optical properties of materials can be described in terms of two constants. These are the refractive index n and the extinction coefficient k. Alternatively we can choose the real and imaginary components ε_1 and ε_2 of the dielectric 'constant' or complex permittivity. The reflectance R can be expressed in terms of either of these two pairs of parameters [1].

The so-called optical constants change with frequency of incident light (electromagnetic radiation), or equivalently, we can say that they change with the energy of the incident photons. When discussing the 'optical' properties of materials, the term is often interpreted widely to include other parts of the electromagnetic spectrum and not just the visible range of frequencies. Table 9.1 shows the frequencies, wavelengths and photon energies of different types of electromagnetic radiation.

9.1.1 Penetration depth δ, and absorption coefficient α

How can we describe empirically the reduction in intensity of light when it passes through a material?

When discussing the electronic transitions in materials that arise from the absorption of photons we should remember that these do not necessarily take place throughout the bulk of the specimen. The depth of penetration of incident

Table 9.1 Frequencies, wavelengths and photon energies of various types of radiation within the electromagnetic spectrum.

Type	**Frequency (10^{14} Hz)**	**Wavelength (nm)**	**Energy**	
			(10^{-19} J)	**(eV)**
Gamma-rays and X-rays	1000 and above	3 and below	660 and above	412 and above
Ultraviolet (typical value)	10	300	6.6	4.12
Violet	7.50–7.05	400–424	4.97–4.69	3.10–2.92
Blue	7.05–6.10	424–491	4.69–4.05	2.92–2.53
Green	6.10–5.21	491–575	4.05–3.46	2.53–2.16
Yellow	5.21–5.12	575–585	3.46–3.40	2.16–2.12
Orange	5.12–4.63	575–647	3.40–3.07	2.13–1.92
Red	4.63–4.28	647–700	3.07–2.84	1.92–1.77
Infrared (typical value)	3	1000	2	1.25
Microwaves and radio waves	0.003 and below	10^6 and above	0.002 and below	0.00125 and below

light depends on the frequency of the light and the optical constants of the material. The depth at which the intensity of the incident electromagnetic wave is attenuated to $1/e$ of its value is called the penetration depth δ. This is expressed by the following equation,

$$I = I_0 \exp(-z/\delta), \tag{9.1}$$

where z is the distance into the material. Replacing δ by the attenuation coefficient α, which is also widely used to characterize materials, gives the relation,

$$I = I_0 \exp(-\alpha z), \tag{9.2}$$

which is known as Lambert's law or Beer's law. In transparent materials, such as various different types of glass, δ is large being of the order of 0.1–0.3 m, while in metals δ is very small being of the order of 10^{-8} m.

9.1.2 Physical significance of the optical constants *n* and *k*

How do the observed optical constants relate to the absorption of a wave in a material medium?

The solution of the wave equation in a material with optical constants n and k leads to the following equation for the electric vector ξ [2],

$$\xi_x(z) = \underbrace{\xi_0}_{\text{incident amplitude}} \underbrace{\exp\left(\frac{-\omega k z}{c}\right)}_{\text{damping term}} \underbrace{\exp\left(\mathrm{i}\omega\left\{t - \frac{nz}{c}\right\}\right)}_{\text{oscillatory term}}. \tag{9.3}$$

ξ_x is the electric field component parallel to the surface, ω the frequency of the incident radiation, z is the distance normal to the surface of the material, x is a direction parallel to the surface of the material and c is the velocity of the incident light wave. The optical constants n and k have been defined in Sections 1.4.1 and 1.4.2. Since $\omega/c = 2\pi/\lambda$ this equation can be expressed alternatively in terms of the wavelength λ:

$$\xi_x(z) = \xi_0 \exp\left(\frac{-2\pi k z}{\lambda}\right) \exp\left(\mathrm{i}\left\{\omega t - \frac{2\pi n z}{\lambda}\right\}\right). \tag{9.4}$$

9.1.3 Dielectric constants of materials

How are the optical constants of a material related to the dielectric constants? The above equation for the parallel component of ξ as a function of depth z contains two terms, an exponentially decaying term which is dependent on k and an undamped wave term which is dependent on n. Therefore n affects the phase of the light wave in the material and k affects its amplitude. The optical properties can equally be expressed in terms of the real and imaginary parts of the dielectric constant ε_r as follows,

$$\varepsilon_r = \varepsilon_1 + \mathrm{i}\varepsilon_2 = (n + \mathrm{i}k)^2, \tag{9.5}$$

where the total dielectric constant is $\varepsilon = \varepsilon_0(\varepsilon_1 + \mathrm{i}\varepsilon_2)$.

The intensity of light, which is proportional to ξ^2, is given by,

$$I = \xi^2 \tag{9.6}$$

$$= I_0 \exp\left(\frac{-2\omega k z}{c}\right), \tag{9.7}$$

and from this equation we can use the definition of the penetration depth δ as the distance required to decrease the intensity by a factor of 1/e:

$$\frac{I}{I_0} = \exp\left(\frac{-2\omega k \delta}{c}\right) = \frac{1}{\mathrm{e}}. \tag{9.8}$$

Under these conditions the penetration depth δ is

$$\delta = \frac{c}{2\omega k} = \frac{\lambda}{4\pi k}, \tag{9.9}$$

and the attenuation coefficient is

$$\alpha = \frac{2\omega k}{c} = \frac{4\pi k}{\lambda}. \tag{9.10}$$

Notice that δ depends on k but not on n. Some typical values of extinction coefficient k and penetration depth δ in the visible range of the spectrum are given in Table 9.2.

Table 9.2 Values of extinction coefficient and penetration depth for various materials in the visible range of the spectrum.

Material	k	δ (m)
Water	1.4×10^{-7}	0.32
Glass	1.5×10^{-7}	0.29
Graphite	0.8	6×10^{-8}
Gold	3.2	1.5×10^{-8}

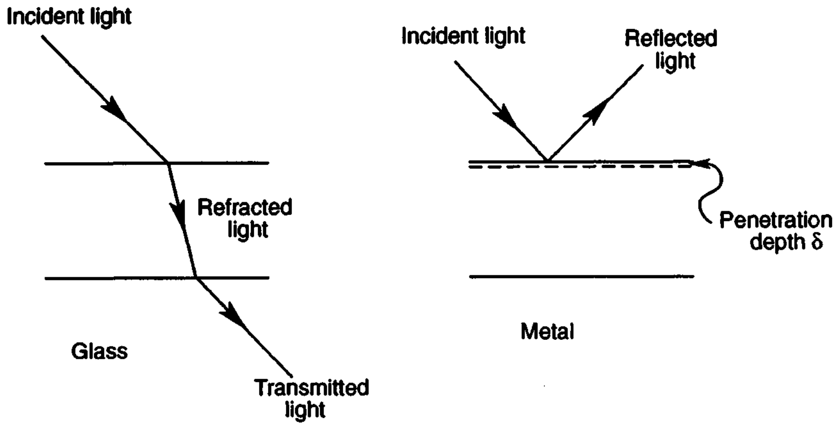


Figure 9.1 Refraction and reflection of light by a material medium.

We see therefore that while optical properties of materials such as water and glass are the result of a bulk measurement, in graphite and gold they are restricted to measurements made over a few nanometres at the surface as indicated in Fig. 9.1. Once again the fact that light only penetrates a few nanometres in some materials implies that those materials must have high reflectance. Reflectance measurements on metals are highly sensitive to the surface condition (e.g. presence of oxide coating) and a question also remains whether a surface measurement under these conditions is representative of bulk material.

9.2 INTERPRETATION OF OPTICAL PROPERTIES IN TERMS OF SIMPLIFIED ELECTRON BAND STRUCTURE

How can the features of the optical reflectance spectrum be related to the electron band structure?

We have also shown in Section 6.2.1 that the optical properties can be explained in terms of the electronic properties of the materials. The high reflectance of metals is a result of the partially filled conduction band, as shown in Fig. 9.2, which allows photons to be absorbed and reflected over a wide range of energies, forming a continuum of energies from the infrared up to the visible range. In the visible or ultraviolet, however, an energy is reached beyond which the absorption and reflection usually decrease markedly in metals, as shown in Fig. 3.7, due to an inability to excite electrons to just above the top of the conduction band. The energy at which this occurs represents the energy between the Fermi level and the top of the conduction band.

In semiconductors, reflectance and absorption are low in the infrared, but absorption becomes possible as soon as the photon energy becomes larger than the band gap. Consequently, semiconductors have higher absorption and reflection in the ultraviolet. This has been shown in Fig. 7.12.

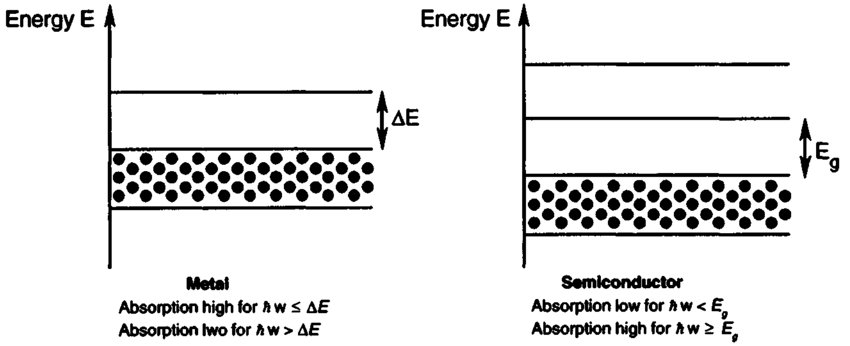


Figure 9.2 Schematic band structure diagrams for a metal and semiconductor.

9.2.1 Summary of optical absorption processes

How can the various electronic transitions be classified?

(a) Interband transitions

(i) The highest-energy transitions are those from the bottom of the valence band to the top of the conduction band. Changes in the density of states across these bands affect the absorption of light at differenct energies and so give rise to a frequency dependence of the absorption coefficient.

(ii) Other lower-energy interband transitions from the top of the valence band to the bottom of the conduction band also occur. The 'absorption edge' occurs at $\hbar\omega = E_g$ the gap energy. The extinction coefficient k in semiconductors is usually in the range (10^7–10^8 m^{-1}) for energies above the band gap energy E_g. However, k decreases by several orders of magnitude once the energy drops below the band gap energy E_g because there are no longer energy states for the excited electrons to occupy so they cannot absorb the energy of the incident photons.

(iii) Another electronic process is known as exciton generation. It is an excitation which produces a bound electron-hole pair. The electron is trapped in a localized energy level in the band gap while the hole remains mobile in the valence band. The exciton can dissociate into independent free carriers or can recombine with the emission of a photon or phonon.

(iv) Excitation of electrons from localized trap sites in the band gap into the conduction band can occur at energies lower than E_g. This usually occurs from optical absorption, although it can also arise from thermal excitation. The optical absorption arising from this process is much lower than for interband transitions because there are relatively few trapped electrons compared with electrons in the valence band.

(b) Intraband transitions

In metals, the absorption of photons by electrons occurs over a continuous wide range of energies beginning effectively from zero energy. This usually involves the absorption or emission of phonons to conserve momentum. In this process

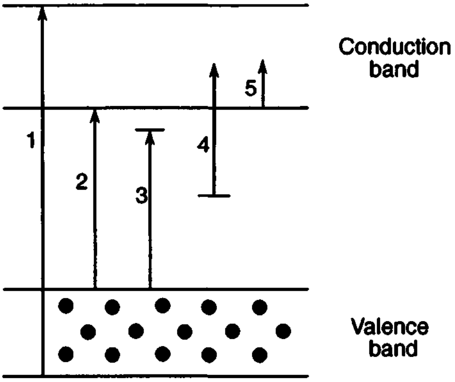


Figure 9.3 Classification of the principal types of electron transitions: (1) high energy interband transitions, (2) transitions across the band gap–absorption edge, (3) exciton generation (bound electron-hole pair), (4) impurity level excitation, (5) intraband transition.

the electrons move between energy states in the same band. The intraband transitions occur in metals, and they are responsible for the high reflectivity of metals at low energies.

The various types of electronic transitions are represented on the 'flat band' diagram of Fig. 9.3.

9.2.2 Colour of semiconductors

How is the colour of a material determined by its electronic structure?

The band gaps of certain semiconductors, such as the III-V semiconductors can be changed by alloying. If the band gap of a range of semiconductors varies from 3.5 eV (ultraviolet) to 1.5 eV (infrared) then when these are illuminated with white light the colour of the materials by transmission changes as progressively more of the visible spectrum is absorbed, beginning from the high-energy blue end of the spectrum and ending with all optical energies being absorbed. The colours change from colourless to yellow, orange, red and finally black, depending on whether all of the visible spectrum or only a portion of the longer wavelength region is transmitted. Here, of course, certain colours by transmission are not possible because of the nature of the absorption process.

We should also note that colour by transmission and colour by reflection will be different in these cases because of the interband absorption process. So a material with a band gap in the yellow region of the visible spectrum, 2 eV might, when illuminated with white light, appear orange-red by transmission but blue-green by reflection, since only the blue-green portion of the spectrum can be absorbed and hence reflected.

Colour itself is a subjective phenomenon. For example, a suitable combination of yellow and blue light may appear green to the eye, even though a spectral analysis would reveal that each of the original frequency components is still present. This is simply due to the physiology of the human eye which interprets the presence of certain frequencies of light as colour in a nonunique way. The eye detects colours only in terms of the combinations of 'primary' colours: red (565 nm, 2.18 eV), green (535 nm, 2.3 eV) and blue (445 nm, 2.8 eV). It is possible to persuade the eye that certain colours (i.e. frequencies) are present even when the actual spectrum is merely a suitable combination of these 'primary' colours.

9.2.3 Direct and indirect transitions across the band gap

How does the probability of an interband transition depend on the difference between the energy of the photon and the band gap?

If the probability of a direct, that is *k*-conserving, transition such as shown in Fig. 9.4(a) is calculated, it is found to be dependent on the square root of the difference between the photon energy and the band gap energy. The absorption is proportional to this probability and therefore,

$$\alpha = \alpha_d(\hbar\omega - E_g)^{1/2}. \tag{9.11}$$

Naturally the probability of a transition is zero when $\hbar\omega < E_g$. Consequently, a plot of α^2 against $\hbar\omega_{\text{photon}}$ gives a straight line with intercept equal to the band gap energy E_g.

The probability of indirect, that is phonon-assisted, transitions such as shown in Fig. 9.4(b) is much lower than for direct transitions. This leads to a lower value of absorption coefficient, by typically two or three orders of magnitude. The effects

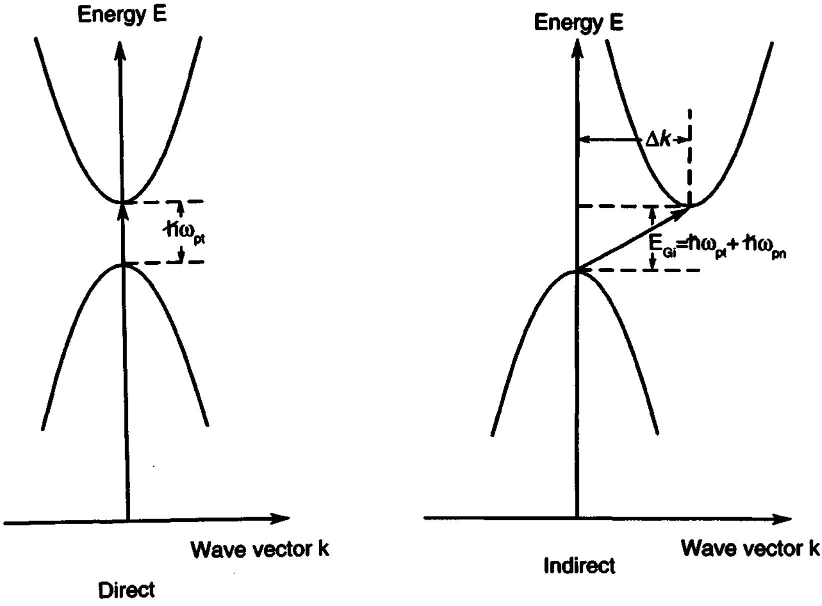


Figure 9.4 Direct and indirect interband transitions shown on an *E* versus *k* diagram.

of indirect interband transition on the optical properties of solids are therefore only noticeable in the absence of direct transitions. However, the transition depends on $(\hbar\omega - E_g)^2$, that is

$$\alpha = \alpha_i(\hbar\omega - E_g)^2. \tag{9.12}$$

Here, α_i for indirect transitions is much smaller than α_d for direct transitions.

9.2.4 Impurity level excitation

How can electrons escape from local traps in the band gap?

The elevation of a trapped electron from an impurity level (or electron trap) in the band gap to the conduction band can occur either by thermal excitation, or alternatively by absorption of a photon. Usually, in order that thermal stimulation can occur the energy of the trap has to be close enough to the conduction band so that $\Delta E \approx k_B T$. Excitation by a photon is simply dependent on the energy of the photon being greater than the difference in energy between the trapped state and the bottom of the conduction band, $\hbar\omega \geq \Delta E$.

Imperfections, defects, or impurity levels are localized and so do not extend throughout the solid. Therefore they are represented as a short line on the energy band diagram. These imperfections usually are one of the following types: (i) point defects, (ii) point impurities, (iii) dislocations and grain boundaries.

9.2.5 Purity of semiconductor materials

How closely is the impurity content controlled in 'electronic quality' semiconductor material?

In semiconductor materials, in which the engineering of band structures to meet stringent requirements is essential, it is clear that the presence of unanticipated defects must be kept to an absolute minimum. Otherwise the materials will have unexpected and undesirable electronic properties. Therefore, production of semiconductor materials takes place in extremely clean environments. The purity of semiconductor materials is frequently better than one part in 10^6 (excluding the doping materials which are also on the level of parts in 10^6). In metals, 'high purity' usually means something like 99.9% or one part in 10^3. We see therefore that much greater care is needed in the fabrication of semiconductors than is needed for metals.

9.2.6 Identification of the occurence of interband transitions from band structure diagrams

How can we locate the electronic transitions from an E versus k diagram?

It is possible to interpret the optical spectra in terms of the electron band structure of a material [3]. If we take a very simple example of a metal with a band structure as shown in Fig. 9.5, we can identify the optical transitions that are possible.

Here, the minimum separation between the bands occurs at the zone boundary. Since both bands are empty at this point, no transitions can occur here. A direct transition can occur at the Fermi surface to the next-highest band with energy $\hbar\omega_0$.

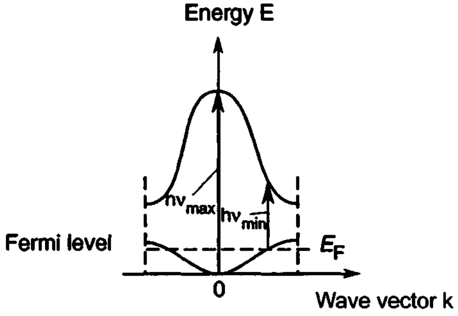


Figure 9.5 Direct electronic transitions at different locations within the Brillouin zone.

Transitions are then possible from all other occupied lower-energy states to the corresponding points in the upper band. These are k-conserving. The highest-energy transition is $\hbar\omega_m$ from the bottom of the occupied band to the top of the next unoccupied band.

Electronic interband transitions are only possible from the region of k-space from $-k_F$ to $+k_F$ in Fig. 9.5, which in this case represents the occupied states below the Fermi level. The transitions with the lowest energy occur at $k = -k_F$ from the Fermi level to the unoccupied level in the conduction band at $k = k_F$. That is to say these are direct (k-conserving) transitions:

$$\Delta E_{\min} = \hbar\omega_0. \tag{9.13}$$

Other transitions are possible. Direct transitions over the range of energies from $\hbar\omega_0$ to $\hbar\omega_m$ are possible. The highest-energy direct interband transition $\hbar\omega_m$ occurs in this case at the centre of the zone, the Γ point.

Indirect, that is phonon-assisted, non-k-conserving, interband transitions are also possible, but these occur with much lower probability. An example of an indirect interband transition is given in Fig. 9.6. The change in momentum of the electron is $+\hbar\Delta k$, and therefore a phonon of momentum $-\hbar\Delta k$ must be emitted to conserve momentum.

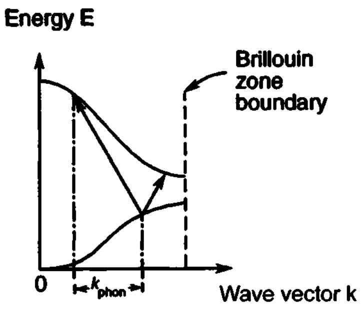


Figure 9.6 Indirect interband transitions described within the first Brillouin zone.

9.2.7 Intraband transitions

What low energy transitions are possible in metals?

Transitions between electron states in the same band are called intraband transitions. These always need the assistance of a phonon, and so are indirect transitions. They can only occur from occupied states to unoccupied states at the Fermi level and above.

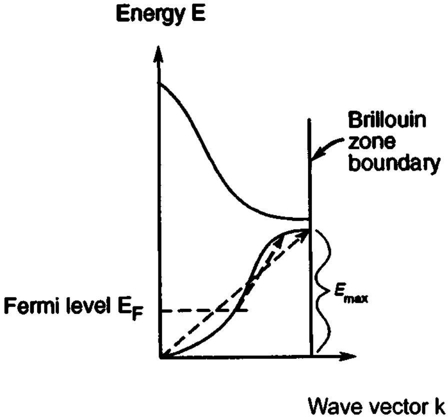


Figure 9.7 Indirect, phonon assisted intraband transition within the conduction band of a metal.

These transitions are the archetypal 'free electron' transitions which are used to describe the optical absorption and reflection process on the classical model. These can occur in metals only, and are responsible for infrared absorption and high reflectance in the optical frequency range in metals.

9.3 BAND STRUCTURE DETERMINATION FROM OPTICAL SPECTRA

How can the details of the electron band structure be investigated through measurements?

The principal methods of determining the electron band structures of materials are optical methods, although a range of other techniques are used to give supplementary information, including photoemission studies, de Haas–van Alphen effect and theoretical band structure calculations [4].

Experimentally, reflectance is the easiest optical property to measure, but in most cases R is a rather slowly varying function of wavelength and this makes it very difficult to locate the exact energies of interband transitions. Also, R itself does not contain all the available optical information. The absorption ε_2, or the extinction coefficient k, are much more useful since they have rather sharper features, but even these need supplementing with ε_1 or n data, respectively, to completely specify the optical properties. As an example consider the reflectance of aluminium which is shown in Fig. 9.8.

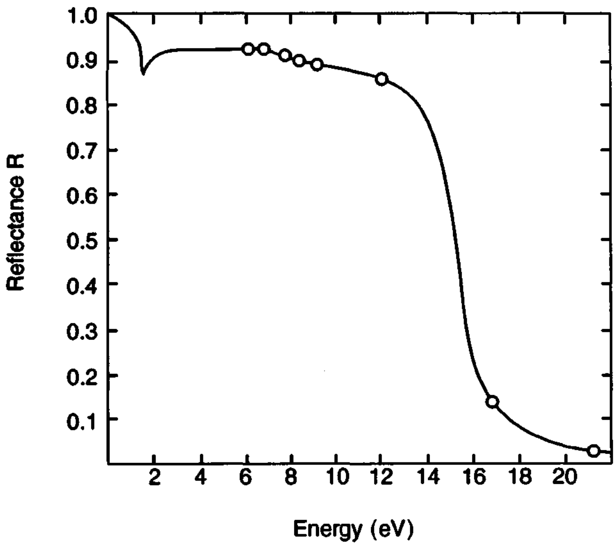


Figure 9.8 Optical reflectance spectra of aluminium after Ehrenreich [5]. © IEEE 1965.

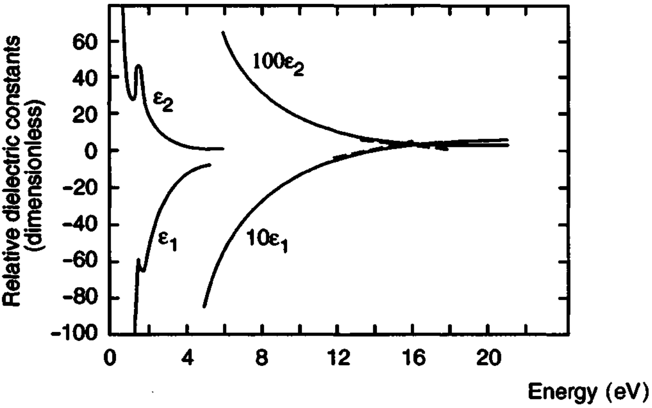


Figure 9.9 Optical constants of aluminium after Ehrenreich [5]. © IEEE 1965.

Here we see a fairly typical metallic reflectance spectrum with high reflectance at low energy and a sharp decay at about 15 eV, much as we might expect on the basis of the Drude model. Notice some structure at about 1.5 eV, but otherwise the spectrum is featureless. The absorption ε_2 as shown in Fig. 9.9 has sharper features, so that the existence of a transition at 1.5 eV is clearly indicated.

9.3.1 Case studies: optical reflectance and band structure

What do the electron band structures of real materials look like?

The optical absorption spectrum of aluminium is shown in Fig. 9.9. The main spectral features occur at an energy of 1.5–2.0 eV. These can be attributed to

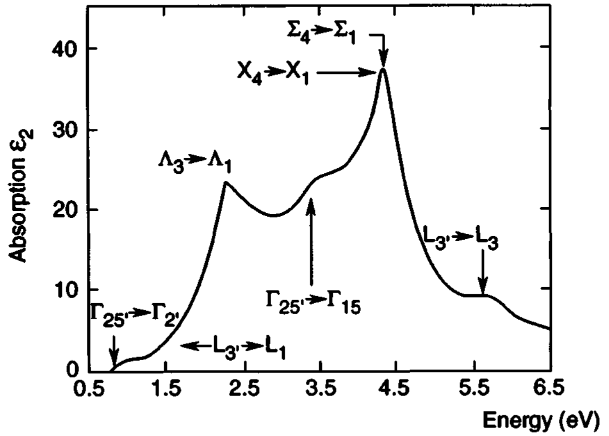


Figure 9.10 Absorption coefficient ε_2 spectrum of germanium. Reproduced with permission of F. Abeles, *Optical Properties and Electronic Structure of Metals and Alloys*, published by Elsevier, 1966.

parallel band absorption along the Z direction between the X and W points in the Brillouin zone, where the band separation is about 2 eV over a wide region between the W point and the Fermi surface [6] as shown in Fig. 5.17. Another region of parallel band absorption is along the Σ direction between the zone centre Γ and the K point. Here, the parallel bands are typically 1.5 eV apart.

The optical absorption spectrum for germanium is shown in Fig. 9.10. The main features occur at energies of 0.8, 2.3 and 4.5 eV. These correspond to the transitions $\Gamma_{25'} \to \Gamma_{2'}$, $\Lambda_3 \to \Lambda_1$ and $\Sigma_4 \to \Sigma_1$ as shown in the band structure diagram of

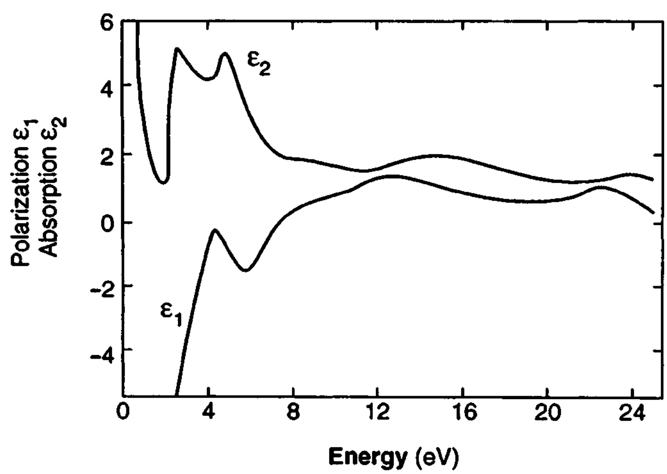


Figure 9.11 Polarization ε_1 and absorption coefficient ε_2 spectra of copper. Reproduced with permission from H. Ehrenreich and H. R. Philips, *Phys. Rev.*, **128**, 1962, p. 1622.

Fig. 7.3. The first of these is a direct transition representing the energy difference across the Γ point which is not quite the minimum 'band gap' energy in germanium.

Figure 9.11 shows the absorption spectrum for copper which is fairly typical of a free electron metal with high absorption at low energies. Spectral features occur at 2, 5, 6 and 7 eV; these features correspond, respectively, to interband transitions $Q_1 \rightarrow Q_2$ at the Fermi level, $X_5 \rightarrow X_4$, $L_2 \rightarrow L_1$ and $\Sigma_2 \rightarrow \Sigma_3$ at the K point as shown in the band structure diagram of Fig. 5.16.

9.3.2 Modulation spectroscopy

Is it possible to accentuate the interband spectral features relative to the broad backgound intraband absorption ?

Further enhancement of the spectra can be obtained by differentiation. This is achieved experimentally by a collection of techniques known as modulation spectroscopy. In these methods the optical spectrum is modulated by the superposition of alternating strain (piezoreflectance) [7], temperature (thermoreflectance), electric field (electroreflectance), magnetic field (magnetoreflectance) and wavelength [8, 9].

All of these, with the exception of wavelength modulation, cause cyclic changes, or perturbations in the electron band structure. These emerge as changes in the optical properties. They cause an enhancement of the interband transitions over the intraband transitions in the optical spectrum because the intraband contribution to the optical properties is almost independent of the modulation, even though the energy levels do change; whereas the sensitivity of the band gap to modulation causes significant changes in the available interband transition energies. An example of the enhancement due to strain of the reflectance spectrum of aluminium is shown in Fig. 9.12.

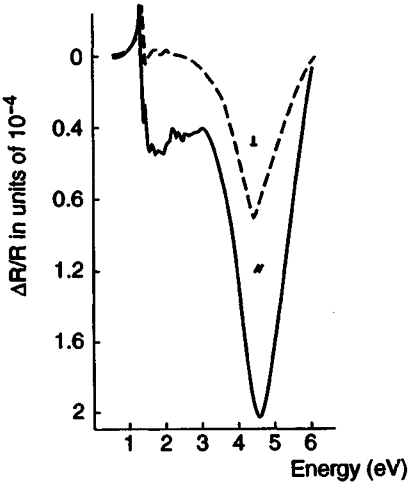


Figure 9.12 Piezoreflectance spectrum of aluminium [7]. Reprinted from D. C. Jiles, *Solid State Communications* **47**, p. 38, 1983. With permission from Elsevier Science.

The modulation spectroscopy techniques therefore lead to an enhancement of features in the optical spectra by measuring the derivative of the optical parameters R, n and k (or R, ε_1 and ε_2) with respect to strain, or electric field, or temperature, or magnetic field.

9.4 PHOTOLUMINESCENCE AND ELECTROLUMINESCENCE

What other methods are there for causing emission of light from semiconductors and insulators?

In the previous chapter we discussed thermoluminescence which is the phenomenon of light emission due to interband transitions when a semiconductor or insulator is heated. This effect was first discovered in quartz. It is different from the familiar black-body glow radiation known as incandescence. Now we will investigate other mechanisms for luminescence in materials. These include photoluminescence, optically stimulated emission of light, and electroluminescence, electrically stimulated emission of light.

9.4.1 Photoluminescence: phosphorescence and fluorescence

How do phosphorescent and fluorescent materials work?

Phosphorescent materials are used widely for dials on clocks and watches because they glow in the dark. By comparison, fluorescent materials glow in the light and are used for 'day glow' colours. Whenever an electron is excited into a higher-energy state it must eventually revert to a lower unoccupied state, and this occurs with the emission of a photon. If the initial excitation is by incident light this process is called photoluminescence.

The lifetime of the electrons in the higher-energy states determines the duration of the emission process. If the lifetime is short then the emission of photons occurs almost immediately and the luminescence stops when the light source is switched off. This process is called fluorescence. If the lifetime extends over a period of

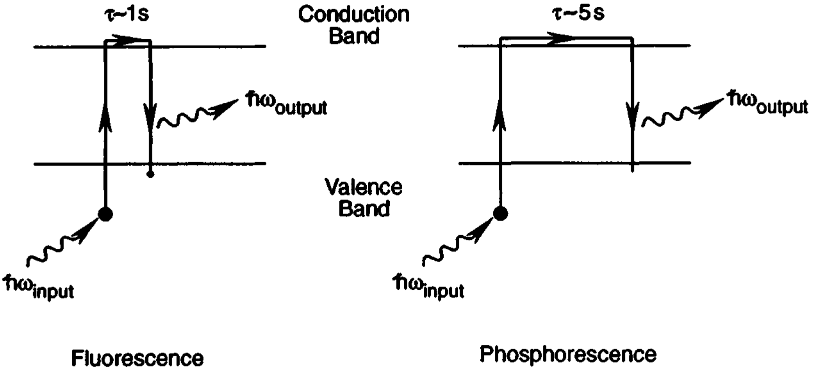


Figure 9.13 Fluorescence and phosphorescence: electron transition diagrams. Large τ corresponds to phosphorescence, small τ corresponds to fluorescence.

several seconds or even a few minutes then the luminescence continues even after the light source is removed. This process is called phosphorescence.

Both of these processes are caused by spontaneous emission of light. That is to say there is no underlying mechanism to stimulate the reversion of the electrons to lower energy states other than spontaneous transition. In general, the wavelengths of the light emitted in fluorescence and phosphorescence are different from the wavelengths of the incident light, and are usually of a well-defined wavelength that is determined by the band gap energy.

9.4.2 Electroluminescence

How can light emission be stimulated by an electric field?

Optical emission, or luminescence, in solids can be caused by a variety of mechanisms. These are classified by the method of excitation. We have studied the two best-known mechanisms, thermal stimulation, (thermoluminescence), and light stimulation (photoluminescence). A third method is electroluminescence which is the excitation of electrons by an electric field [10]. This is used in the creation of semiconductor light sources.

We will consider injection electroluminescence in a single pn junction of a semiconductor. The electronic properties of such a junction have been discussed in Section 7.8. When the n-type and p-type materials are placed in contact, electrons flow into the p-type material leaving it with a negative charge and the n-type with a positive charge.

If the p-type side of such a junction is then connected to the positive terminal of a voltage supply, current is carried by the flow of electrons into the p-type material where there are already free holes in equilibrium. Recombination of electrons and holes can take place and this results in emission of photons.

In a material such as GaAs, which is a direct gap semiconductor, the absorption edge rises very rapidly with photon energy so that the probability of radiative

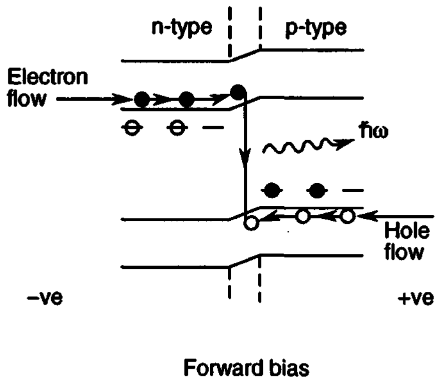


Figure 9.14 Electroluminescence: electron transition diagram in a forward-biased pn junction. The electrons are injected from the n-type material.

recombination is very high. In a positive p-sided pn junction with low impurity concentrations strong emission of light occurs at low temperatures. However, at higher impurity concentrations and higher temperatures the conductivity of the material is too high and emission occurs principally by conduction band to impurity site transitions. These result in the emission of photons of much lower energy, usually in the infrared region of the spectrum.

The peak of the emission increases to higher photon energies as the current through the junction increases until at a critical current the emission peak sharpens considerably and laser action begins. We will discuss the operation of these solid-state lasers in Chapter 12.

REFERENCES

1. F. Wooten (1972) *Optical Properties of Solids*, Academic Press.
2. F. Abeles (1972) *Optical Properties of Solids*, North Holland, Amsterdam.
3. R. E. Hummel (1971) *Optische Eigenschaften von Metallen und Legierungen*, Springer, Berlin.
4. F. Abeles (1966) Optical Properties and Electronic Structure of Metals and Alloys, *Proc. Int. Conf.*, North Holland, Amsterdam.
5. H. Ehrenreich *et al.* (1965) *IEEE Spectrum* **2**, 162.
6. D. Brust *et al.* (1964) *Phys. Rev.* **A134**, 1337.
7. D. C. Jiles and M. P. Staines (1983) *Solid State Commun.* **47**, 37.
8. J. G. Swanson and V. Montgomery (1990) Optoelectronic modulation spectroscopy, *J. Electron. Mater.* **19**, 13.
9. M. DiMarco and J. G. Swanson (1992) Temperature effects in GaAs OEMS spectra, *J. Electron. Mater.* **21**, 619.
10. M. C. Wu and C. W. Chen (1992) Luminescence from AlGaAsP layers grown on GaAsP substrates by liquid phase epitaxy, *J. Electron. Mater.* **21**, 977.

FURTHER READING

O. Svelto (1982) *Principles of Lasers*, 2nd edn, Plenum, London & New York.

H. S. Nalwa (2000) *Handbook of Advanced Electronic and Photonic Materials and Devices*, Academic Press, San Diego.

EXERCISES

Exercise 9.1 Optical properties of metals and insulators

The optical constants n and k of four different materials are given in Table 9.3.

Table 9.3 Optical constants at $\lambda = 1240\,\mathrm{nm}$.

Material	***n***	***k*** **(at $\lambda = 1240\,\mathrm{nm}$)**
1	1.21	12.46
2	0.13	8.03
3	1.51	1.12×10^{-6}
4	1.92	1.5×10^{-6}

From these values determine for each material the attenuation coefficient α, the penetration depth δ, the normal reflectance R, the dielectric constant ε_1 and the absorption ε_2. Determine whether each of these materials is a metal or an insulator based on these optical properties.

Exercise 9.2 Classification of principal electronic transitions

Discuss the principal electronic transitions that can occur in solids and relate them to the band structure diagram. Explain the characteristic colours of materials both in reflectance and transmission in terms of the electronic structure.

Exercise 9.3 Identification of material from optical absorption spectrum

The optical spectrum of an unknown material is given in Fig. 9.15. State whether on the basis of this data the material is a metal or an insulator. Determine the absorption threshold for the material and deduce for which optical wavelengths the material is transparent and for which wavelengths it is opaque.

Using the data in Table 9.4 determine the material assuming it is one of the three shown.

Table 9.4 Electronic properties of three unknown materials.

	A	B	C
Electron mobility ($m^2 s^{-1} V^{-1}$)	0.15	0.39	0.85
Electrical conductivity (at 300 K) $\Omega^{-1} m^{-1}$	9×10^{-4}	2.2	1×10^{-6}
Effective mass of electrons at 4.2 K (m^*/m)	0.98	1.64	0.07
Band gap (eV)	1.1	0.7	1.4
Absorption edge (nm)	1104	1873	871

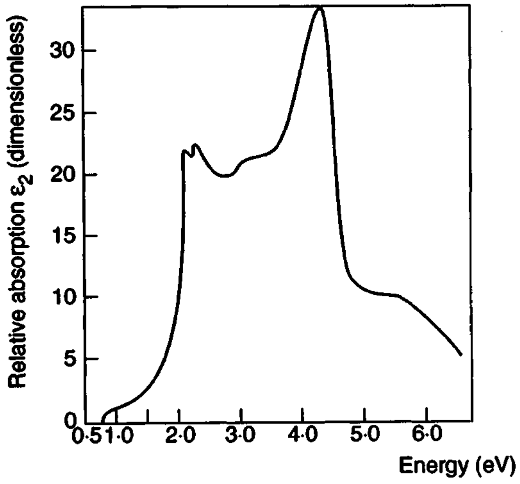


Figure 9.15 Optical spectrum of unknown material.

Exercise 9.4 Equation of motion of 'free' electrons and the absorption of light
If the behaviour of electrons inside a material can be described by the classical free electron model, including both the 'free' and 'bound' characteristics of their motion, write an equation of motion for the electrons inside the material under the action of incident light of amplitude ξ_0 and frequency ω.

Explain the physical significance of the coefficients in this equation. Obtain a solution for the position x of the free electron, and from this derive expressions for the amplitude and phase of the electron's motion. Is there a resonance? If so what would this correspond to in the optical spectrum of the material? Calculate the characteristic penetration depth of light of wavelength 589 nm in aluminium if it has an extinction coefficient of $k = 6$ (dimensionless).

Exercise 9.5 Effects of differences in band gap on optical properties of semiconductors
The semiconductor cadmium sulphide, CdS, has a band gap of 2.4 eV.

(a) What colour will CdS appear by transmitted light?
(b) A dopant (an electron donor) added to CdS has an energy level lying 1.0 eV above the valence band, but below the Fermi level. What colour changes would you expect to see for light transmitted through the material as the amount of dopant is increased from 0 to 1000 parts per million?
(c) If the band gap of CdS decreases linearly with temperature according to the relation $E_g = 2.56 - aT$, where $a = 5.2 \times 10^{-4}$ eV K^{-1}, describe the changes in colour of light transmitted through a pure CdS crystal as the temperature is changed from 0 to 1000 K in 200-degree steps.

Exercise 9.6 Optical properties of direct and indirect band gap materials
Light transmission measurements made on a 1-micrometre thick film of lead sulphide, PbS, at energies near its band gap energy are shown below. These data have been corrected for reflection losses. On the basis of these data is the material a direct or indirect band gap material? What is the band gap energy?

Wavelength (nm)	**% Transmission**
2065	22.8
2155	25.1
2255	28.1
2360	32.0
2480	36.8
2610	44.5
2755	53.3
2915	72.8

10 MAGNETIC PROPERTIES OF MATERIALS

OBJECTIVE

In this chapter we look at the magnetic properties of materials. The magnetic properties are a special subgroup of the electronic properties of materials which really form a separate subject. Nevertheless they can also be considered as an integral part of the electronic properties of materials. The most important and interesting magnetic state of a material is known as ferromagnetism. In this case the relative permeability can be very high. This makes these materials useful in transformers and inductors. Another property of ferromagnets is their retention of magnetization. This is utilized in permanent magnets for both motors and generators. In addition, particulate and thin film magnetic materials are used for magnetic recording purposes. This application represents a very large market, both for magnetic materials and the associated electronic support systems for magnetic recording.

10.1 MAGNETISM IN MATERIALS

What causes magnetism in some materials?

The magnetic properties of materials arise almost exclusively from the motion of the electrons. This motion, in the form of electron spin and electron orbital motion, generates a magnetic moment associated with the electron. Much weaker magnetic moments arise from the nucleus, but these are three orders of magnitude smaller. Compare, for example, the size of the nuclear magneton $\mu_n = 5.051 \times 10^{-27}\,\mathrm{A\,m^2}$ with the Bohr (electron) magneton $\mu_B = 9.274 \times 10^{-24}\,\mathrm{A\,m^2}$.

There are two theories of the origin of magnetization or bulk magnetic moment in solids which represent limiting or extreme cases. These are the localized or atomic theory, and the itinerant or band theory. In the localized model, the electronic magnetic moments are considered to be bound to the ionic cores in the solid. Such a model applies to the lanthanide series of elements in which the 'magnetic' electrons are inner 4f electrons which are closely bound to the nuclei.

In the itinerant model, the magnetic moments are considered to be due to conduction band electrons which originate as the outer electrons on the isolated atoms. When the atoms are brought together, as in a solid, these electrons are shared among the atoms and move freely throughout the material. This model is considered by some authors to be more appropriate for the 3d transition elements iron, cobalt, and nickel. In reality, even in the 3d series metals the itinerant electrons spend more time close to the nuclei, and so the actual situation is somewhere between these extreme or limiting models.

Before proceeding further with these ideas, however, we will need a few definitions.

10.1.1 Magnetic field and magnetic induction

How is a magnetic field generated?

A magnetic field is generated whenever there is electric charge in motion. We denote this field with the symbol H. The magnetic field generated by an elemental length of conductor $\mathrm{d}\ell$ carrying a current i is given by the Biot–Savart law,

$$\mathrm{d}H = \frac{1}{4\pi r^2}\,\mathrm{d}\boldsymbol{\ell} \times \boldsymbol{u}, \tag{10.1}$$

where r is the radial distance from the conductor at which $\mathrm{d}H$ is measured, u is a unit vector along the radial direction, $\mathrm{d}\boldsymbol{\ell}$ is a vector along the direction of the length of the conductor and $\mathrm{d}H$ is the elemental contribution to the total field at r.

The magnetic induction, denoted B, is the response of a medium to the presence of a magnetic field. Therefore, for a given field strength H, the magnetic induction can be different in different media. The relationship between the magnetic induction and the magnetic field is called the permeability μ of the medium.

$$B = \mu H. \tag{10.2}$$

We should note immediately that μ is not necessarily constant for a material, although in most cases it is either constant or nearly so. The important exception is the class of ferromagnetic materials for which μ varies over an extremely wide range.

The permeability of free space is determined, on the basis of our choice of the metre, newton and ampere as units, to be $4\pi \times 10^{-7}$ Henry per metre ($\mathrm{V\,s\,A^{-1}\,m^{-1}}$) and is denoted by the symbol μ_0. Therefore in free space,

$$B = \mu_0 H. \tag{10.3}$$

10.1.2 Magnetization

How do we measure the magnetic response of a material?

When the individual magnetic moments associated with the electrons in a solid are collectively aligned, perhaps by the action of an external magnetic field H, we speak of magnetization. We define the magnetization as the magnetic moment per unit volume and denote it by the symbol M. The magnetization increases as more electronic magnetic moments are aligned in the same direction. When all magnetic moments within a solid are aligned in the same direction, the magnetization cannot get any higher. We therefore call this the saturation magnetization.

The magnetization M contributes, together with the magnetic field H, to the magnetic induction B. Therefore we can write the totally general equation relating M, H and B,

$$B = \mu_0(H + M), \tag{10.4}$$

where $\mu_0 H$ is the induction which would be generated by the field H in free space, and $\mu_0 M$ is the additional induction contributed by the presence of the magnetic material. The magnetization is measured in units of amps per metre. Some authors discuss the magnetization in terms of a bound surface current. We should state clearly that this 'bound current' is a fictitious current which is merely invoked as a convenience so that M and H can be treated equivalently.

We defined the permeability μ in Section 1.6.3 as the ratio of magnetic induction to field

$$\mu = \frac{B}{H}. \tag{10.5}$$

Similarly, we defined the magnetic susceptibility χ in Section 1.6.4 as the ratio of magnetization to field

$$\chi = \frac{M}{H} \tag{10.6}$$

and in the SI system of units $\mu = \mu_0(1 + \chi)$.

10.1.3 Typical values of permeability and susceptibility

What values of permeability and susceptibility do various materials have?
The following table gives the susceptibilities and relative permeabilities of some diamagnets, paramagnets and ferromagnets. In weak magnetic materials such as diamagnets and paramagnets the susceptibility is usually quoted. In ferromagnets permeability is usually quoted. In ferromagnets, because both values are large ($\gg 1$) and the values of relative permeability and susceptibility differ only by $\chi = \mu_r - 1$, permeability and susceptibility are relatively close in value and, in practice, the two terms are often used interchangeably in these materials, although such usage is not strictly correct.

10.2 TYPES OF MAGNETIC MATERIALS

How are magnetic materials classified?
There are several different types of magnetic materials but we shall break them down into three traditional categories,

Table 10.1 Susceptibilities and relative permeabilities.

Material	χ	$\frac{\mu}{\mu_0}$
Bi	-1.31×10^{-6}	0.9999987
Be	-1.85×10^{-6}	0.9999982
Ag	-2.02×10^{-6}	0.9999980
Au	-2.74×10^{-6}	0.9999973
Ge	-0.56×10^{-6}	0.9999994
Cu	-0.77×10^{-6}	0.9999992
β-Sn	0.19×10^{-6}	1.0000002
W	6.18×10^{-6}	1.0000062
Al	1.65×10^{-6}	1.0000016
Pt	21.04×10^{-6}	1.0000210
Mn	66.10×10^{-6}	1.0000661
Fe	$\sim 1 \times 10^{3}$	$\sim 1 \times 10^{3}$

(i) Diamagnets $\chi < 0$; $\mu_r < 1$,
(ii) Paramagnets $\chi > 0$; $\mu_r \cong 1$,
(iii) Ordered magnetic materials (e.g. ferromagnets) for which normally $\chi \gg 0$; $\mu_r \gg 1$.

This categorization is, of course, rather an oversimplification of the different types of magnetic ordering but is still used in traditional magnetism texts. The ordered magnetic materials consist of several subcategories which include: ferromagnets, ferrimagnets, superparamagnets and even two subcategories with low permeabilities, the antiferromagnets and helimagnets.

10.2.1 Diamagnets

How do diamagnets respond to a magnetic field?

These are materials which have no net magnetic moment on their atoms. In other words, the electrons are all paired with spins antiparallel. When a magnetic field H is applied, the orbits of the electrons change in accordance with Lenz's law, and they set up an orbital magnetic moment which opposes the field. Since this moment is in the opposite direction to the field in diamagnets the susceptibility is negative

$$\chi < 0. \tag{10.7}$$

The classical theory of diamagnetism was worked out by Langevin and has been discussed by Cullity [1] and Chen [2].

10.2.2 Paramagnets

How do paramagnets respond to a magnetic field?

Paramagnets are materials which have a net magnetic moment per atom due to an unpaired electron spin. In zero field these individual magnetic moments are randomly aligned, but under the action of an external field H they can be aligned in the field direction. As a result of this alignment of moments in the field direction the magnetization M is parallel to the field and hence the susceptibility is positive,

$$\chi > 0. \tag{10.8}$$

In general, however, very large fields are needed to align all the moments and so the susceptibility, although positive, is usually very small, having a typical value of $\chi \approx 10^{-5}$.

10.2.3 Ordered magnetic materials

What other types of magnetic materials are there?

The third class of magnetic materials are the most interesting. These are the ordered magnetic materials, the most important of which are the ferromagnets. These include iron, cobalt, and nickel and their alloys and compounds, and several of the rare earth elements, notably gadolinium, and their alloys and compounds. Other ordered magnetic materials include the antiferromagnets, chromium and manganese, the ferrimagnets such as iron oxide, and the helimagnets such as dysprosium, terbium, holmium and erbium.

10.2.4 Curie and Néel temperatures

What happens when the temperature of a magnetic material is raised?

The ordered state of any ferromagnet or ferrimagnet breaks down at a temperature known as the Curie point T_c. Above this temperature the material is disordered, that is the electronic magnetic moments point in random directions even on a local scale. Values of T_c for some materials are shown in Table 10.2.

In antiferromagnets and helimagnets the ordering temperature is known as the Néel point T_N. Values of the Néel temperature for some materials are given in Table 10.3.

Table 10.2 Curie temperatures of various ferromagnets.

Material	**T_c (°C)**
Iron	770
Cobalt	1131
Nickel	358
Gadolinium	20

Table 10.3 Néel temperatures of various helimagnets and antiferromagnets.

Material	**T_N (°C)**
Erbium	−253
Holmium	−253
Dysprosium	−93
Terbium	−43
Chromium	35
Manganese	−173

10.3 MICROSCOPIC CLASSIFICATION OF MAGNETIC MATERIALS

How are the electronic magnetic moments arranged in the various different magnetic materials?

The macroscopic classification into the three traditional groups, based on the permeability values, needs significant modification when we consider the magnetic ordering on the sub-microscopic scale of a few atoms. The microscopic classification needs to include the types of order shown in Fig. 10.1.

10.3.1 Electron magnetic moments

Where are the magnetic moments located in a material?

If we consider the classical picture of an electron orbiting a nucleus and also spinning on its axis, as shown in Fig. 10.2, we have charge in motion. Consequently, we must have a contribution to the magnetic field and magnetic induction arising through the Biot–Savart law.

In general terms this is correct; however, the classical picture has many flaws in it. We will soon find that the numerical values of spin magnetic moment, which is the most significant contribution, differ from the expected value.

In reality, there is no electric current here in the classical sense. Therefore the classical model of the electronic magnetic moment merely serves to remind us that

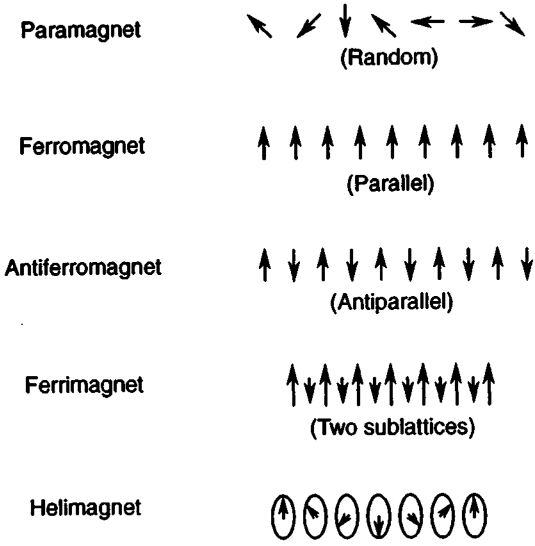


Figure 10.1 Arrangement of magnetic moments on neighbouring atoms of a one-dimensional lattice for various types of magnetic order.

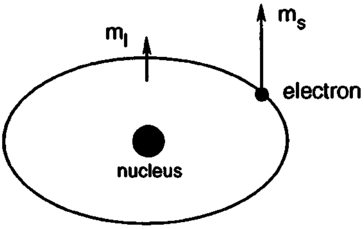


Figure 10.2 Classical model of electron orbiting a nucleus with orbital magnetic moment $\boldsymbol{m}_l$ and spin magnetic moment $\boldsymbol{m}_s$.

there is some link between the angular momentum of an electron and its magnetic moment. However, since the angular momentum of the electron is a quantum phenomenon, it is hardly surprising that the classical prediction breaks down.

We are left with an empirical relation between angular momentum p and magnetic moment m

$$m = \gamma p, \tag{10.9}$$

where the coefficient of proportionality γ is the gyromagnetic ratio. An alternative form of this relation, in terms of the Bohr magneton μ_B and the Landé splitting factor g is,

$$m = \frac{-g\mu_B}{\hbar} p, \tag{10.10}$$

where $g = 2$ for electron spin alone, and $g = 1$ for electron orbital motion alone. Normally, the value of g lies between 1 and 2 indicating some orbital and some spin contributions. In the 3d series elements iron, cobalt and nickel the orbital contribution is negligible.

10.3.2 Order–disorder transitions and the Curie point

What factors determine the Curie temperature of a material?

The ordered magnetic structure which exists in a ferromagnet can be destroyed by raising the temperature. The thermal energy added to the material has a tendency to randomize the orientation of the magnetic moments, while the internal exchange interaction tries to keep them aligned. Eventually, a sufficiently high temperature is reached when the thermal energy overcomes the exchange and the material undergoes an order–disorder transition.

All ordered magnetic materials (including ferromagnets, ferrimagnets, antiferromagnets, and helimagnets) can be made paramagnetic at a sufficiently high temperature. However, not all paramagnets can be converted to ordered magnetic materials by cooling.

The temperature at which the order–disorder transition occurs in ferromagnets and ferrimagnets is known by convention as the Curie temperature T_c. In antiferromagnets and helimagnets, the order–disorder transition temperature is known as the Néel temperature. The behaviour of the susceptibility changes at these ordering temperatures and we say that it exhibits 'critical behaviour' close to T_c or T_N.

Some materials have both a Curie and a Néel temperature because they exhibit more than one ordered magnetic phase. Examples are terbium ($T_N = 230\,\mathrm{K}$, $T_c = 220\,\mathrm{K}$) and dysprosium ($T_N = 180\,\mathrm{K}$, $T_c = 85\,\mathrm{K}$), which undergo transitions paramagnetic → helimagnetic → ferromagnetic as the temperature is reduced.

10.3.3 Temperature dependence of susceptibility

How does the susceptibility of a magnetic material change with temperature, and how can this be quantitatively described?

The Curie and Curie–Weiss laws were empirical discoveries of the temperature dependence of the paramagnetic susceptibility of certain magnetic materials. It is worth noting that these laws are not as widely applicable as is often generally supposed. However, their simple form and their subsequent explanation using classical statistical thermodynamics means that they hold an important place in the historical development of our understanding of magnetism.

The Curie law states that the susceptibility χ of a paramagnet is proportional to the reciprocal of the temperature T in Kelvin

$$\chi = \frac{C}{T}, \tag{10.11}$$

where C is a constant.

The Curie–Weiss law is a generalization of the Curie law to include those materials which undergo an order–disorder transition to ferromagnetism or ferrimagnetism at T_c. In these cases the susceptibility in the paramagnetic phase is also inversely proportional to the temperature according to the relation

$$\chi = \frac{C}{T - T_c}, \tag{10.12}$$

where C is a constant.

It is important to note that the Curie–Weiss law only applies to the susceptibility of a magnetic material in its paramagnetic phase.

10.3.4 The Curie law

Can the Curie law be explained in terms of the statistical behaviour of an array of individual magnetic moments?

The Curie law can be explained on the local moment model using classical Maxwell–Boltzmann statistics. In materials with unpaired electrons, there is a net (or resultant) magnetic moment per atom m which is the vector sum of the spin and orbital magnetic moments. The energy of this moment in a magnetic field H is,

$$E = -\mu_0 \boldsymbol{m} \cdot \boldsymbol{H}. \tag{10.13}$$

If we suppose that the magnetic moments are noninteracting and use classical statistics, then the probability of an electron occupying an energy state E is

$$P(E) = P_0 \exp(-E/k_B T). \tag{10.14}$$

If there are N magnetic moments per unit volume, then the magnetization M, which is the magnetic moment per unit volume, will be found by integrating,

$$M = \int_0^N m \cos\theta \, dn, \tag{10.15}$$

$$M = Nm \frac{\int_0^\pi \cos\theta \sin\theta \exp(\mu_0 m H \cos\theta / k_B T) \, d\theta}{\int_0^\pi \sin\theta \exp(\mu_0 m H \cos\theta / k_B T) \, d\theta}, \tag{10.16}$$

$$M = Nm \left\{ \coth\left(\frac{\mu_0 m H}{k_b T}\right) - \frac{k_B T}{\mu_0 m H} \right\}, \tag{10.17}$$

and saturation magnetization occurs when all moments are aligned parallel. We use here the symbol M_0 to denote saturation magnetization, as distinct from spontaneous magnetization or technical saturation in a ferro- or ferrimagnet, which we shall denote M_s (see Section 10.5.4)

$$M_0 = Nm. \tag{10.18}$$

Therefore,

$$M = M_0 \left\{ \coth\left(\frac{\mu_0 m H}{k_B T}\right) - \frac{k_B T}{\mu_0 m H} \right\}. \tag{10.19}$$

This expression is the Langevin equation for classical paramagnetism based on the local moment model. We can derive the Curie law directly from this. The Langevin function can be expressed as an infinite series in $\mu_0 mH/k_B T$. For high temperatures $\mu_0 mH/k_B T \ll 1$, so that the first term in the series dominates,

$$M = Nm\left\{\frac{\mu_0 mH}{3k_B T}\right\}, \tag{10.20}$$

and substituting $C = \dfrac{N\mu_0 m^2}{3k_B}$, this gives

$$M = \frac{CH}{T}, \tag{10.21}$$

and, since $\chi = \dfrac{M}{H}$, we obtain

$$\chi = \frac{C}{T}, \tag{10.22}$$

which is the Curie law.

10.4 BAND ELECTRON THEORY OF MAGNETISM

Are the 'magnetic' electrons localized on the ionic sites or are they free to move throughout the material?

The itinerant theory of magnetism attributes the magnetic properties of materials to unpaired electrons in the conduction band. These electrons by definition can migrate throughout the whole material and are therefore 'itinerant'. This interpretation is valid for some materials but not for others. It does seem to be broadly applicable to the magnetism of the 3d series of elements such as iron, cobalt and nickel.

10.4.1 Pauli paramagnetism

How can we derive an expression for the paramagnetic susceptibility in terms of the behaviour of conduction electrons?

The itinerant, conduction band theory of paramagnetism was developed by Pauli [3]. This theory leads to a temperature-independent paramagnetic susceptibility. Beginning from our earlier discussion of electron bands, consider, for example, the parabolic free electron band shown in Fig. 10.3. For ease of visualization, we represent the spin-up states on the left and the spin-down states on the right. These we term the spin-up and spin-down half bands. Electrons will occupy the lowest available energy states. In the absence of a field, the energy levels of the spin-up and spin-down states are degenerate. Therefore, the two half-bands are symmetric, occupancies of the two half-bands are equal, and the net magnetic moment per atom is zero.

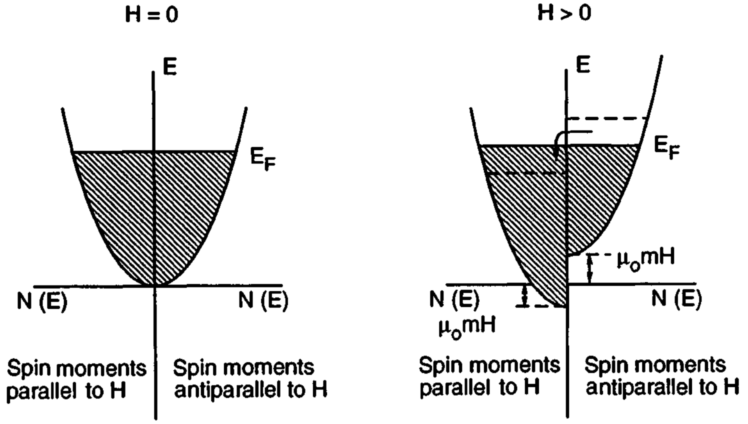


Figure 10.3 Occupation of electron energy levels in a free electron material in zero magnetic field and under an applied magnetic field *H*.

When a field is applied the individual electronic moments acquire an additional energy which depends on the scalar product of the magnetic moment ***m*** with the field ***H***

$$\Delta E = -\mu_0 \boldsymbol{m} \cdot \boldsymbol{H} \qquad (10.23)$$

This energy lifts the degeneracy of the two half-bands, because the spin-up bands move to lower energy while the spin-down bands move to higher energy. As a result, the Fermi energy equalizes in the two half-bands and some electrons switch from spin-down to spin-up. This leads to a net magnetic moment in the spin-up direction. It is also clear from this diagram that only those electrons that are close to the Fermi level will be able to switch direction.

10.4.2 Dependence of magnetization on field in Pauli paramagnetism

How does the susceptibility depend on the Fermi level, number of conduction electrons and magnetic moment per electron?

From our earlier consideration of electron states at the Fermi level, only a fraction T/T_F of the conduction electrons can contribute to the magnetization. This again is similar to the fraction of electrons that contribute to the heat capacity and electrical conductivity. Therefore, according to the classical Curie law we should expect,

$$M = \frac{N^* \mu_0 m^2 H}{3 k_B T}, \qquad (10.24)$$

where M is the magnetization and N^* is the number of electrons per unit volume that can change the orientation of their spins. Since $N^* = NT/T_F$ (as described in

Section 4.5.1) where N is the total number of conduction electrons per unit volume this leads to,

$$M=\frac{N\mu_0 m^2 H}{3k_B T}\left(\frac{T}{T_F}\right), \tag{10.25}$$

where m is the electronic magnetic moment and k_BT is the Boltzmann energy.

If we approach the problem in a more exact way, using quantum mechanics and the band theory of electrons, the above result is modified slightly, but ultimately we arrive at the same form of expression for M. The number density of electrons parallel to the field N_+ is,

$$N_+ = \int f(E)D(E+\mu_0 mH)\,\mathrm{d}E, \tag{10.26}$$

$$N_+ \approx \int f(E)D(E)\,\mathrm{d}E + \mu_0 mHD(E_F), \tag{10.27}$$

and the number density antiparallel to the field N_- is, by a similar argument

$$N_- \approx \int f(E)D(E)\,\mathrm{d}E - \mu_0 mHD(E_F). \tag{10.28}$$

The magnetization is therefore given by

$$M = m(N_+ - N_-) \tag{10.29}$$

$$= 2m^2\mu_0 HD(E_F), \tag{10.30}$$

where $D(E_F)$ is the density of states at the Fermi level. We can see that the resulting magnetization M is dependent not only on the applied field H, but also on the density of states at the Fermi level $D(E_F)$. Using $2D(E_F) = 3N/2k_BT_F$ [4 p. 415],

$$M=\frac{3N}{2k_B T_F}m^2\mu_0 H, \tag{10.31}$$

and therefore the Pauli free electron band theory of paramagnetism leads to the following equation for the susceptibility:

$$\chi=\frac{M}{H}=\frac{3Nm^2\mu_0}{2k_B T_F}. \tag{10.32}$$

This predicts a temperature-independent paramagnetic susceptibility which is observed in a number of metals such as sodium, potassium and rubidium.

10.4.3 Electron band model of ferromagnetism

How can we generalize the above theory of magnetic susceptibility to describe ferromagnetic materials?

The band theory of ferromagnetism is a simple extension of Pauli's band theory of paramagnetism to ferromagnets with the inclusion of an exchange interaction (internal effective magnetic field) to align the electrons in a cooperative manner in the absence of an external applied field. This causes a relative displacement of the spin-up and spin-down half-bands known as the exchange splitting. It is qualitatively similar to that encountered under the action of an applied magnetic field in Pauli paramagnetism, except that here the shift in energy is much larger and occurs in the absence of an applied magnetic field. The net spontaneous magnetization of a material is again determined by the difference in occupancy between the spin-up and spin-down states.

In the 3d series elements, the outer electron bands of interest, which contribute to the magnetic properties, are the 3d (total capacity 10 electrons) and the 4s (total capacity 2 electrons). In iron, cobalt, and nickel, the 4s band is completely filled with its complement of two electrons. Therefore, since this level must be occupied with a spin-up and a spin-down it can have no contribution to the magnetic moment per unit volume. The magnetic properties are therefore determined by the partially filled 3d band.

Suppose that we have a material such as nickel consisting of atoms with 8 electrons each in a partially filled d band which has a capacity of 10 electrons. Then if we consider 10^{29} atoms per m^3 of this material, each atom could have up to 10 of these 3d electrons in the band, and so there will be one electron band consisting of 5×10^{29} energy levels with spin-up and an equal number of energy levels with spin-down states (giving a total of 10^{30} possible electron states). There will be 8×10^{29} electrons in this 3d band and these will occupy the lowest energy states available. With no interactions between the electrons, they will be equally distributed between the two subbands ($3d\uparrow$ and $3d\downarrow$) with no net magnetic moment. So in this case we will again have a Pauli paramagnet.

In order to get the net imbalance of spins necessary for ferromagnetism it is essential to invoke an exchange energy which displaces the energies of the spin-up and spin-down half-bands even in the absence of an external field.

10.4.4 Exchange coupling

Why should there be a net magnetic moment in a partially filled electron band of a ferromagnet?

In a partially filled energy band, it is possible to have an imbalance of spins. This is caused by the presence of an exchange interaction between the electrons in the conduction band which has the effect of aligning the spins parallel. This exchange interaction is quantum mechanical in origin and has no classical analogue. However, following Weiss, it is sometimes modelled as a classical mean field. When this effective field is calculated it is found to be extremely large, being equivalent to a magnetic field of about $10^9\,\mathrm{A\,m^{-1}}$ (10^7 Oe).

The exchange interaction has the effect of reducing the energy of parallel alignment of spins even in the absence of an external field. Therefore the occupancy of the spin-up state becomes energetically favoured over the spin-down state in zero field. This results in a net magnetic moment.

10.4.5 Spin-up and spin-down half-bands

Under what conditions can the exchange interaction lead to a net magnetic moment?

The exchange energy can only alter the alignment of the moments if the reduction in energy due to exchange is greater than the energy difference between the lowest available spin-up state and the highest occupied spin-down state. In other words, the system of spins will always configure itself to the lowest possible energy state, taking the exchange energy into account. If the exchange energy is present but is not large enough to alter the ground state in this way then no net magnetic moment will arise.

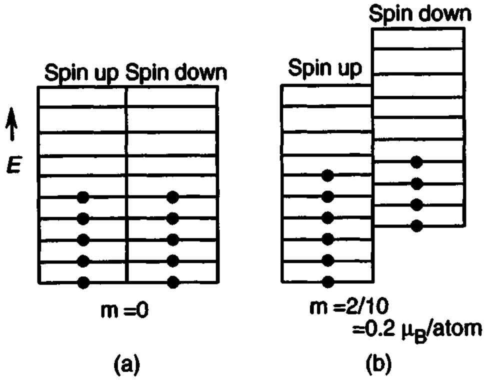


Figure 10.4 Occupation of discrete energy states in the conduction band of an itinerant electron ferromagnet. On the left no exchange splitting results in equal occupancy of the spin-up and spin-down half-bands. On the right the exchange splitting leads to an imbalance of spins and a net magnetic moment per atom.

The energy resulting from the exchange interaction, E_{ex}, is usually represented by the Heisenberg model [5] as,

$$E_{ex} = -J_{ex} s_1 \cdot s_2, \tag{10.33}$$

where J_{ex} is the exchange operator, and s_1 and s_2 are the spins on electrons. When $J_{ex} > 0$, then we have a tendency to parallel alignment which minimizes the exchange energy and leads to ferromagnetism, but this can only occur if

$$|E_{ex}| \geq \Delta E, \tag{10.34}$$

where ΔE is the energy difference between the lowest available spin-up state and the highest occupied spin-down state. This condition simply ensures that any change in orientation of electronic magnetic moment causes a reduction in the total energy of the system.

10.4.6 Magnetic moment per atom

How is it possible to obtain magnetic moments per atom which are nonintegral multiples of the Bohr magneton?

In the example of nickel given above, suppose that in unit volume 3×10^{28} out of the 4×10^{29} (i.e. 3 out of 40) electrons from the spin-down half-band transfer to the spin-up half-band because it is energetically favourable. This results in a net excess of 6×10^{28} spins, or a net magnetic moment of 0.6 Bohr magnetons per atom. We see from this how it becomes possible to explain fractional numbers of Bohr magnetons per atom by the band theory of ferromagnetism.

The electrons fill the spin-up band first as shown in Fig. 10.5. When the half-bands overlap electrons can be added to the spin-down half-band before the spin-up half-band is filled. A complete separation of the energies of the spin-up and spin-down bands would leave the entire spin-up half-band at a lower energy than the spin-down half-band, and therefore all the electrons would fill the spin-up half-band first.

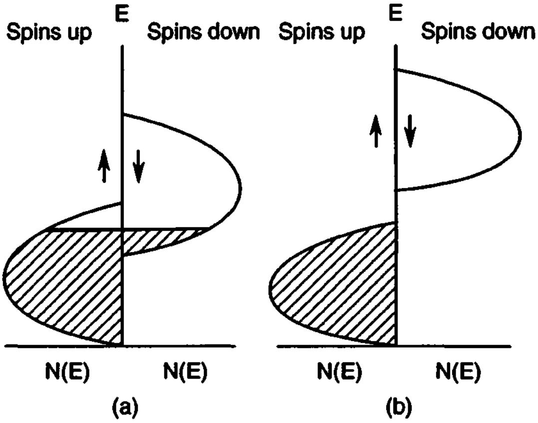


Figure 10.5 Occupation of electron energy levels in a ferromagnet. The two half-bands are split by the exchange coupling. In (a) the exchange splitting does not completely separate the half-bands, while in (b) the exchange splitting is large enough to cause complete separation.

If there were more states in the spin-up band than there were electrons available to fill them, there would be no electrons in the spin-down band. Since in this example each atom contributes one electron to the spin-up band, the net result would be a magnetic moment of one Bohr magneton per atom. The actual electron band structure for nickel is shown in Fig. 10.6. in which the exchange split states are shown as pairs of 'parallel' half-bands. In this case the two half-bands overlap.

10.4.7 Magnetic moments in iron, cobalt and nickel

What are the typical distributions of the outer electrons in iron, nickel and cobalt?

Since it is the outer, unpaired, electrons that contribute to the magnetic properties of the 3d series elements we will consider here the 3d and 4s outer electrons only.

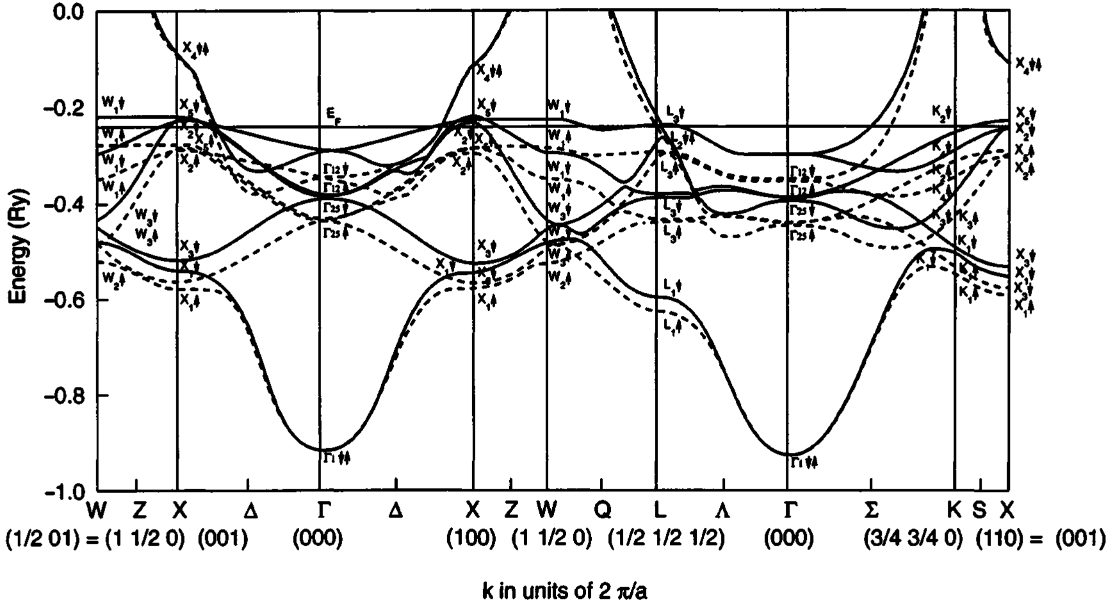


Figure 10.6 Electron band structure of ferromagnetic nickel after Wang and Callaway [6]. Reproduced with permission from C. S. Wang and J. H. Callaway, *Phys. Rev.* **B9**, 1974, p. 4897.

The magnetic moments per atom of these materials when in an aggregate (i.e. solid form) are different from the isolated atoms. This is shown in Table 10.4. The isolated atoms of course do not have 'continuous' electron bands; instead they have discrete energy levels.

Table 10.4 Distribution of 3d and 4s electrons in iron, cobalt and nickel.

Material	Number of 3d electrons	Number of 4s electrons	Net atomic moment (Bohr magnetons)
Fe	6	2	2.2
Co	7	2	1.7
Ni	8	2	0.6

10.4.8 Rigid band model for transition metal alloys

What effect does alloying have on the magnetic moment per atom of the transition metal alloys?

When considering the magnetic properties of the 3d transition metals and alloys, the electronic structure of the conduction band can be approximated by assuming that the electron bands are rigid even on addition of other 3d alloying elements. Therefore by alloying with 3d elements which have more, or less, electrons than the principal component of the alloy, the effect is simply to add, or subtract, electrons from the existing band. Addition or subtraction of electrons simply depends on whether the added element lies to the left or right of the principal component element in the periodic table.

In the 3d transition metals and alloys this occurs without substantially altering the energy levels of the bands. The net result is the variation of magnetic moment per atom with alloy composition known as the Slater–Pauling curve [7, 8]. This

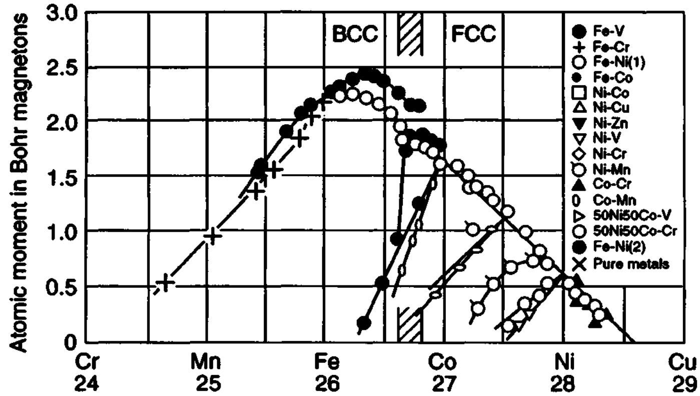


Figure 10.7 The Slater–Pauling curve for 3d transition metals and alloys, showing the variation of net atomic magnetic moment as a function of alloy composition.

describes the magnetic moment in terms of the number of unpaired conduction electrons per atom.

On increasing the number of electrons per atom from zero the magnetic moment at first increases, as the spin-up band is filled, and then decreases, once the spin-down band starts to fill faster than the spin-up band. The assumption of the rigid band approximation which underlies the Slater–Pauling curve should not be taken too literally, but it does seem to work as a first approximation.

10.5 THE LOCALIZED ELECTRON MODEL OF FERROMAGNETISM

Is there an alternative to the itinerant electron theory of magnetism?
In many cases the electronic magnetic moments can be considered to be localized at the atomic or ionic sites. This is particularly appropriate for the lanthanide series of elements in which the 4f 'magnetic' electrons are closely bound to the atomic cores.

10.5.1 The Curie–Weiss law

How can we explain the difference between the Curie–Weiss law and the Curie law?
By introducing a mean field interaction, which is used merely to represent the quantum mechanical exchange interaction in a simplified form, the Curie–Weiss law can be derived from the Curie law. Suppose the interaction can be expressed as a field H_e, which is proportional to M:

$$H_e = \alpha M. \tag{10.35}$$

The total magnetic field H_{Tot} experienced by an individual magnetic moment H is then,

$$H_{Tot} = H + H_e = H + \alpha M. \tag{10.36}$$

The variation of magnetization M with total field H_{Tot} should still obey Curie's law,

$$\frac{M}{H_{Tot}} = \frac{M}{H + \alpha M}, \tag{10.37}$$

but now the measured susceptibility is

$$\chi = \frac{M}{H} \tag{10.38}$$

$$= \frac{C}{T - \alpha C} \tag{10.39}$$

$$= \frac{C}{T - T_c}, \tag{10.40}$$

where $T_c = \alpha N \mu_0 m^2 / 3k_B$ is the Curie temperature. This means that the Curie temperature is directly related to the strength of the exchange interaction as measured by the mean field parameter α.

10.5.2. Classical theory of ferromagnetism

How does the exchange interaction field explain the existence of ferromagnetism? The mean field interaction, as introduced by Weiss, accounts for the observed paramagnetic susceptibility of magnetic materials which undergo a transition to ferromagnetism at the Curie point. This idea can be carried over into the ferromagnetic regime where it explains the existence of magnetic order. A brief outline of the classical Weiss theory follows.

Suppose that any magnetic moment m_i experiences an interaction with any other moment m_j in the material. This can then be expressed as an effective magnetic field experienced by the ith moment,

$$H_{e_{ij}} = \alpha_{ij} m_j / V_j \tag{10.41}$$

where α_{ij} is the interaction between the ith and jth moments and V_j is the volume occupied by m_j.

The total internal interaction field is then given by the sum of the individual interaction fields over the whole material,

$$H_e = \frac{1}{V} \sum_j \alpha_{ij} m_j, \tag{10.42}$$

where V is the volume occupied by the material. This makes it energetically favourable for the electron magnetic moments to align parallel when $\alpha_{ij} > 0$ leading to ferromagnetism.

10.5.3 The mean field approximation

What form does the exchange field take if all interactions are assumed equal? If we assume that the interactions between all moments are identical (i.e. if we make the mean field approximation) with $\alpha_{ij} = \alpha$ for all pairs i, j, then the above summation is simplified to,

$$H_e = \frac{1}{V} \sum_j \alpha m_j \tag{10.43}$$

$$= \frac{\alpha}{V} \sum_j m_j, \tag{10.44}$$

and if M_s is the spontaneous magnetization throughout the volume under consideration (usually a single domain), then

$$H_e = \alpha(M_s - m_i/V) \tag{10.45}$$

and since $m_i/V \ll M_s$, we can replace $M_s - m_i/V$ with M_s

$$H_e \simeq \alpha M_s, \tag{10.46}$$

which is the form of interaction envisaged in a ferromagnetic material in the original theory by Weiss.

The mean field model of ferromagnetism, which provided the first systematic explanation of the phenomenon, has been updated to explain the wider range of phenomena observed since its introduction nearly a century ago. The mean field approach has been applied on the nanoscale [9] to describe such effects as local magnetic fluctuations. The result of this extension of the model now gives results for magnetic susceptibility that agree well with observations close to the Curie temperature. The essential idea in this new development is to overcome the limitations of the traditional 'homogeneous' treatment of the mean field model by considering finite clusters of magnetic moments without restrictions on the cluster sizes. This has enabled a unified description of the paramagnetic behaviour of materials to be developed based on a familiar concept, the mean field model [10].

10.5.4 Magnetic order, spontaneous magnetization and domains

If the magnetic moments are all aligned parallel in ferromagnetism, then how can a ferromagnet ever be demagnetized?

We have discussed the fundamental difference between ferromagnets and paramagnets which is the existence of long-range order in ferromagnets. This means that large numbers of magnetic moments are aligned parallel. This is true of iron, cobalt, and nickel at room temperature.

This leads to an apparent contradiction. Most specimens of iron, cobalt, and nickel do not have a bulk magnetization of the size expected on this basis unless they have been deliberately 'magnetized', that is exposed to a strong external magnetic field. The contradiction is resolved through the existence of magnetic domains which are microscopic volumes in which all magnetic moments are aligned parallel. The direction of this magnetization is different from domain to domain, leading to a low value of bulk magnetization.

We therefore need to distinguish between the macroscopic (or bulk) magnetization which occurs after a ferromagnet has been subjected to a magnetic field, and the microscopic (or spontaneous) magnetization which is the magnetization within a domain. We shall denote the bulk magnetization M and the saturation magnetization M_0. The spontaneous magnetization M_s is close in value to M_0 and as the temperature T is reduced to absolute zero, $M_s(T)$ approaches M_0.

Whereas M can be affected by the presence of an external field, M_s is largely unaffected and M_0 is completely unaffected. The change in magnetization M caused by an external magnetic field arises from reorientation of domain magnetizations, causing the individual domain magnetic moments to line up throughout the material. The description of the dependence of magnetization M on magnetic field H is an important technical problem with no completely general solution.

10.6 APPLICATIONS OF MAGNETIC MATERIALS

What uses do magnetic materials find?

Magnetic materials find their main applications in a variety of technological areas. Soft magnetic materials [11, 12], which have high permeability and low coercivity, are used in electromagnets and inductor cores in which the primary objective is to generate as much magnetic induction B as possible under the action of a magnetic field H. Hard magnetic materials [13], which have low permeability but high coercivity and remanence, are used in permanent magnet applications in which the materials are used to generate a magnetic induction without a conventional electrical power supply.

Magnetic particles and magnetic thin films are used in magnetic recording of information in the form of analogue signals or digital data. The newly emerging field of magnetoelectronics [14], in which magnetic and electronic functions are combined in a single device, is one in which there has been a rapid surge of interest. This interest arises because of the possibilities of dramatic improvements in magnetic data storage densities using these types of materials, both for long-term memory such as disk drives where these materials are used in read heads [15], and for on-line random access memory (MRAM) [16, 17].

REFERENCES

1. B. D. Cullity (1972) *Introduction to Magnetic Materials*, Addison Wesley, Reading, Mass.
2. C. W. Chen (1977) *Magnetism and Metallurgy of Soft Magnetic Materials*, North Holland, Amsterdam.
3. W. Pauli (1926) *Z. Physik* 41, 81.
4. C. Kittel (1986) *Introduction to Solid State Physics*, 6th edn.
5. W. Heisenberg (1928) *Z. Physik* 49, 619.
6. C. S. Wang and J. Callaway (1974) *Phys. Rev. B.*, 9, 4897.
7. J. C. Slater (1937) *J. Appl. Phys.* 8, 385.
8. L. Pauling (1938) *Phys. Rev.* 54, 899.
9. R. V. Chamberlin (2000) *Nature* 408, 337.
10. T. Giebultowicz (2000) *Nature* 408, 299.
11. R. Boll (1979) *Soft Magnetic Materials*, Heyden & Son, London.
12. E. C. Snelling (1988) *Soft Ferrites*, 2nd edn, Butterworths, London.
13. R. J. Parker (1990) *Advances in Permanent Magnetism*, John Wiley & Sons, New York.
14. G. A. Prinz (1999) *J. Magn. Magn. Mater.* 200, 57.
15. J. J. M. Ruigrok, R. Coehoorn, S. R. Cumpson and H. W. Kesteren (2000) *J. Appl. Phys.* 87, 5398.
16. S. S. P. Parkin, K. P. Roche, M. G. Samant, P. M. Rice *et al.* (1999) *J. Appl. Phys.* 85, 5828.
17. J. G Zhu (2000) *J. Appl. Phys.* 87, 6668.

FURTHER READING

A. J. Freeman and S. D. Bader (1999) *Magnetism beyond 2000*, special issue of the *Journal of Magnetism and Magnetic Materials*, 200.

S. Chikazumi (1997) *Physics of Ferromagnetism*, Oxford University Press, Oxford.

D. C. Jiles (1998) *Introduction to Magnetism and Magnetic Materials*, 2nd edn, Chapman & Hall, London.

P. Gruenberg (2001) Layered Magnetic Structures: History, Highlights, Applications. Invited paper number 4Y-01, International Conference on Magnetism, Recife, Brazil, August 2000. To appear in *J. Magn. Magn. Mater.*

EXERCISES

Exercise 10.1 Strength of the exchange field in iron
Iron has a Curie temperature of 1043 K and a magnetic moment of 2.2 Bohr magnetons per ion. Find the strength of its internal exchange field.

Exercise 10.2 Comparison of the magnetic moments on atoms in bulk form and in isolation
Compare the known saturation magnetizations of iron, cobalt and nickel in bulk with the known magnetic moment of the free atoms or ions and comment on the result. Calculate the occupancy of the 3d and 4s bands in iron from these results.

Exercise 10.3 Spontaneous magnetization and the exchange field
Assuming that the Heisenberg exchange interaction can be introduced simply as an effective field that is proportional to the spontaneous magnetization within a domain, derive an expression for the magnetization of a domain as a function of magnetic field H, magnetic moment per atom m and temperature T starting from the classical Langevin equation for paramagnetism.

If $m = 2 \times 10^{-23}\,\mathrm{A\,m^2}$, $N = 9 \times 10^{28}\,\mathrm{m}$ and $T = 300\,\mathrm{K}$, find the value of the mean field parameter which is needed in order to cause spontaneous magnetization.

Exercise 10.4 Saturation magnetization and atomic magnetic moments
Calculate the expected saturation magnetization M_s in iron in $\mathrm{A\,m^{-1}}$ given that as described by the band (or collective electron) theory of magnetism the magnetic moment per atom in bulk iron is 2.2 Bohr magneton. The density of iron is $7.9 \times 10^3\,\mathrm{kg\,m^{-3}}$, and its relative atomic mass is 56.

A toroid of iron has an internal diameter of 0.01 m and external diameter of 0.02 m. The material of the toroid has a square cross-section of 0.005 m. Calculate the total magnetic flux in the toroid when it is magnetized circumferentially by a field of $1000\,\mathrm{A\,m^{-1}}$ if the magnetization is then at 50% of saturation.

Exercise 10.5 Saturation magnetization and electron band structure
A ferromagnetic material has a density of states at the Fermi level of 1.9×10^{49} per Joule per cubic metre. If the exchange coupling between the electrons is $5 \times 10^{-21}\,\mathrm{J}$ calculate the expected saturation magnetization (magnetic moment per unit volume) based on the itinerant or band electron theory of ferromagnetism. If the number of atoms per unit volume is 8.5×10^{28} per cubic metre then calculate the net magnetic moment per atom in Bohr magnetons. (NB Each electron has a net magnetic moment of one Bohr magneton or $9.27 \times 10^{-24}\,\mathrm{A\,m^2}$.)

Exercise 10.6 Hysteresis and energy dissipation
The hysteresis curve in Fig. 10.8 is of a material which has a coercivity of $50\,\mathrm{A\,m^{-1}}$ and a remanence of 0.5 T. Estimate the hysteresis loss per cubic metre of material that is dissipated when the material is driven around one complete hysteresis cycle. If this material is used in a toroidal inductor core of mean circumference 0.05 m,

with a cross-sectional area of $0.25 \times 10^{-4}\,\mathrm{m}^2$, calculate the hysteresis power loss at a frequency of 60 Hz. Is this the total power loss in the inductor at this frequency? What materials properties would you take into account in selecting a material as an inductor core?

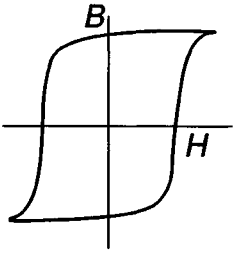


Figure 10.8 Hysteresis loop of magnetic material for Exercise 10.6.

PART THREE

APPLICATIONS OF ELECTRONIC MATERIALS

11 MICROELECTRONICS - SEMICONDUCTOR TECHNOLOGY

OBJECTIVE

In this chapter we look at materials for microelectronics applications. Microelectronics is a vitally important and wide-ranging group of related fields that includes semiconductor materials technology. In a book such as this it is only possible to skim the surface of such a diverse subject. Therefore, the limited objective here is to discuss some of the considerations which go into selecting materials for microelectronics applications. This is based on the specific electronic properties of the materials which have been discussed in previous chapters. We show here how the relative advantages of silicon, germanium and gallium arsenide need to be considered for specific applications.

11.1 USE OF MATERIALS FOR SPECIFIC ELECTRONIC FUNCTIONS

What is 'microelectronics'?

The subject of microelectronics is concerned mainly with semiconductors [1] and semiconductor devices [2]. This is an extremely diverse area of technology which extends from solid-state physics (understanding the electronic properties), through materials science (fabrication of microelectronic circuits), to electrical engineering (performance of components and devices) and computer engineering (integration of large numbers of devices and their application in digital systems). The entire field of microelectronics can be traced back to the development of the pn junction by Ohl [3] and the bipolar junction transistor by Bardeen, Brattain and Shockley [4]. By the mid 1980s the electronics industry had become the largest single manufacturing industry in the world.

Semiconductors are used to produce the key components in the majority of electronic systems including communications, data processing, control and consumer electronics equipment. Of the two main classes of semiconducting materials, intrinsic and extrinsic, it is the extrinsic semiconductors which are of greater technological importance here. The reason for this is that the electrical conductivity can be carefully controlled in extrinsic semiconductors using dopants, whereas in intrinsic semiconductors such fine control is difficult if not impossible to achieve.

11.1.1 Control of electronic properties of semiconductors

What is a dopant?

A dopant is an impurity element which is deliberately added to the semiconductor to change its conductivity. Dopants can provide extra electrons (e.g. phosphorus or arsenic in silicon) to form n-type semiconductors, or they can provide extra holes (e.g. gallium or aluminium in silicon) which are positive charge carriers to

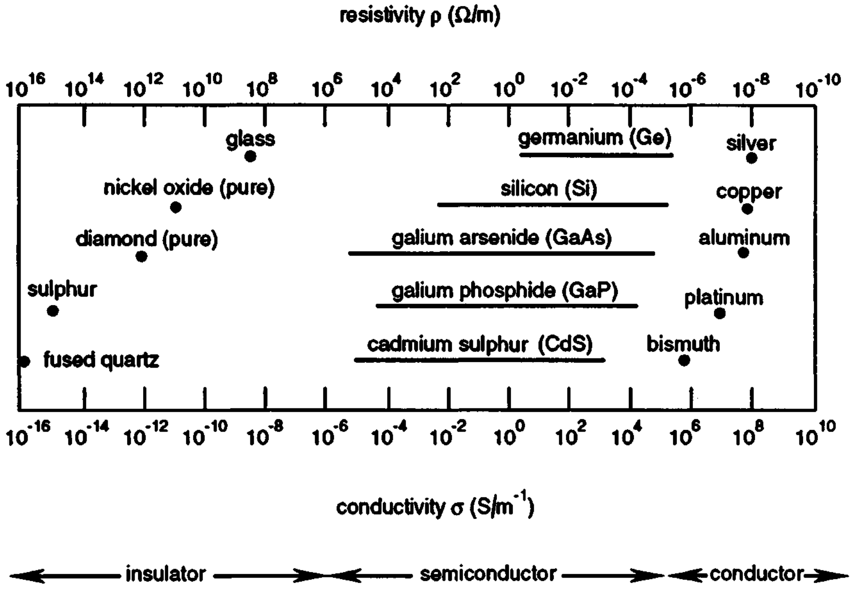


Figure 11.1 Conductivity ranges of semiconductors compared with insulators and conductors.

form p-type semiconductors. Doping concentrations are typically a few parts in 10^6 and this makes it possible to fabricate devices with desired electrical properties. The range of possible conductivities of semiconductors is shown in Fig. 11.1.

11.2 SEMICONDUCTOR MATERIALS

Which are the main semiconductor materials currently available?

The first semiconductor material in widespread use was germanium but this exhibits relatively high leakage currents at moderately elevated temperatures, and it was superseded by silicon in the 1960s. Silicon replaced germanium for almost all applications because of its lower leakage current, resulting from its wider band gap, and the fact that high-quality silicon dioxide can be easily produced to form a good insulating layer on the material where necessary. The silicon dioxide layer strongly adheres to the surface of the silicon and can be used as a mask. In addition, silicon is more abundant than germanium and therefore cheaper. The electrical properties of silicon are also fairly easy to control. The variation of resistivity of silicon with impurity concentration is shown in Fig. 11.2.

The primary material of interest in this chapter is therefore silicon, since most of the semiconductor industry is built around this material and its oxides. It constitutes a large fraction of all semiconductor hardware sold worldwide at present. For use in electronic devices 'electronics grade' silicon is produced. These are single crystals of silicon obtained by slowly withdrawing seed crystals from molten silicon.

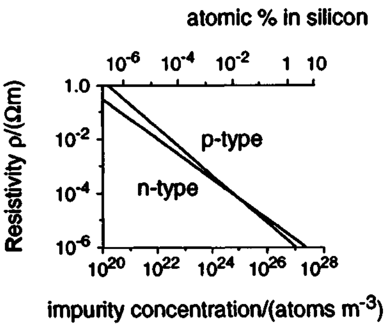


Figure 11.2 Variation of the resistivity of silicon with impurity concentration. Reproduced with permission from N. Braithwaite and G. Weaver, *Electronic Materials*, published by Butterworths, 1990.

11.2.1 Alloy semiconductors

What makes gallium arsenide and related materials so special?

Gallium arsenide and other direct band gap semiconductors are also of interest [5]. Together with other III-V compound semiconductors it is has had enormous impact on the optoelectronics and computer industry because of the possibility of fabricating fast electronic devices from these materials, and because of the ability to support optical functions. The III-V compounds are used in laser diodes, solar cells and light emitting diodes which are discussed in the next chapter. Gallium arsenide and the other III-V semiconductors are the subject of intensive research [6] because of this. Continuing interest stems from the following unusual combination of properties:

(i) High band gap energy which can be engineered (i.e. altered) by combining it with other materials (e.g. InP) to allow optical transitions over a range of energies.
(ii) Direct band gap, making it suitable for optoelectronic applications.
(iii) High electron mobility leading to very fast operation of GaAs devices.

11.2.2 The III-V semiconducting compounds

How can the band gap of these semiconductors be controlled?

One interesting property of these compounds is that they can be mixed together to form solid solutions. For example, aluminium or indium can be substituted for some or all of the gallium; or phosphorus or antimony can be substituted for some or all of the arsenic. This can be used to make subtle changes in the electrical and optical properties of the material. All have the same crystal structure and similar values of lattice constant which is advantageous for fine control of the electronic properties. Composition of the III-V compounds can be carefully controlled by modern materials processing techniques (e.g. growth by molecular beam epitaxy) that allows the band gap to be selected for particular applications. This can result, for example, in materials which emit light over a range of wavelengths close to the optical range which can be used for light emitting diodes of various colours, or for semiconductor lasers.

These materials provide an important avenue of investigation for the development of new optoelectronic devices. The main problem is that the III-V compounds are much more difficult to process than silicon and the raw materials are more expensive too. This leads to a higher cost of devices. However, the possibility of using them for optical applications and the higher electron mobilities, which lead to fast electronic devices, can outweigh the extra cost.

11.3 TYPICAL SEMICONDUCTOR DEVICES

What kinds of devices are fabricated on wafers of semiconductor and how do these work?

We begin by developing an understanding of the electronic processes in some very simple cases. Then we will go on to discuss applications which can be presented without further detailed reference to band structures. Most microprocessor chips contain devices based on only 3 or 4 simple structures. The simple devices that we shall consider in order to develop this understanding are the pn junction and the transistor. Our first objective is merely to describe how these work, using as a basis our understanding of the behaviour of electrons in materials which has been developed in earlier chapters. Only then will we be concerned with their possible applications.

11.3.1 The pn junction

What does the electron band structure look like in the vicinity of a semiconductor junction between p- and n-type material?

The pn junction, which was first demonstrated by Ohl [3], is one of a small number of simple electronic devices that are fabricated on semiconductor chips and which form the basis for the modern microelectronics industry. The other simple structures include the metal/semiconductor junction, the transistor and the heterojunction. These pn junctions are found literally everywhere in microelectronic devices.

Junctions between semiconducting materials are crucial for applications of these materials because in this way very diverse electronic properties can be produced. The electronic behaviour of the pn junction, which consists of two semiconductors with different conduction mechanisms and different Fermi levels joined together, can lead to some quite interesting and useful results. These provide a simple basis for understanding more complex devices. The schematic electron band structure is shown in Fig. 11.3.

As a result of differences in the Fermi levels on each side of the junction, electrons flow from the n-type material to the p-type material. Consequently the n-type material becomes positively charged and the p-type becomes negatively charged. Eventually, the Fermi levels become equalized in the two materials. The region at the interface, which becomes depleted of mobile charge carriers as the mobile

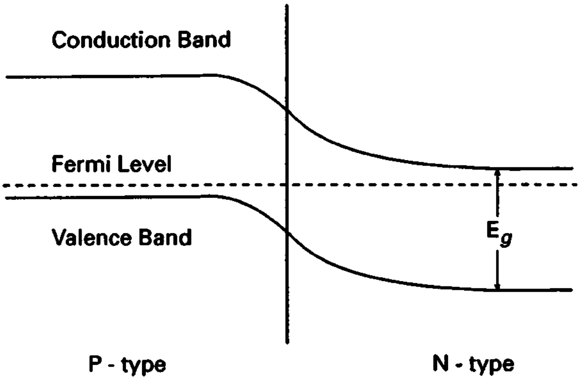
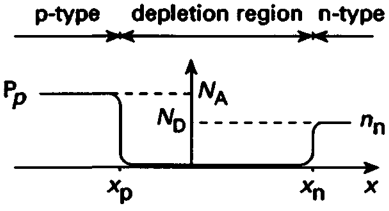


Figure 11.3 Band structure diagram for a pn junction.

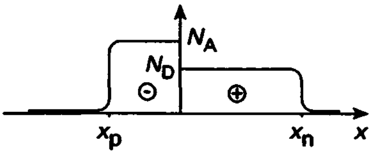
electrons flow into the p-type material and the mobile holes flow into the n-type material, is known as the depletion layer, also sometimes called the space charge region. This is shown in Fig. 11.4. The pn junction can be forward or reverse biased by connecting a voltage source to the two pieces of semiconductor as shown in Fig. 11.5.

11.3.2 Performance characteristics of a pn junction

When subjected to a voltage how does the current through the pn junction change? The current–voltage characteristics of a pn junction are shown in Fig. 11.6. It can be seen that the junction acts like a diode, in which the current rises steeply with applied voltage in one direction (forward bias) but the device remains nonconducting in the other direction (reverse bias). Any electrons trying to move from n-type

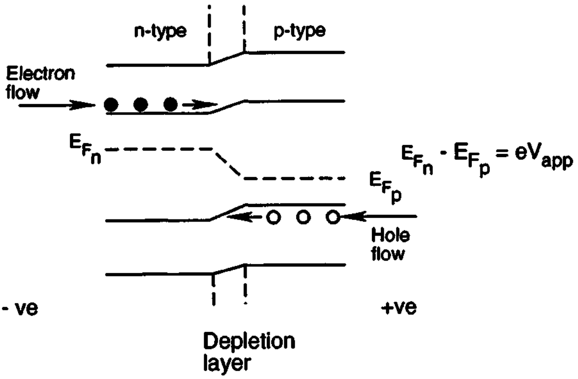


(a) majority carrier density

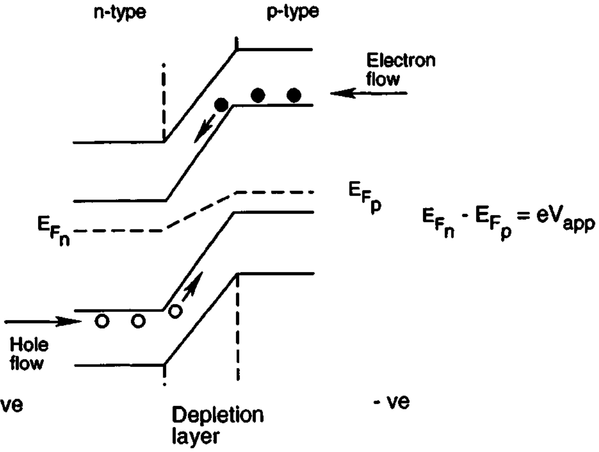


(b) net space charge density

Figure 11.4 Electron and hole densities on either side of a pn junction.



FORWARD BIAS



REVERSE BIAS

Figure 11.5 Connection of a pn junction to a dc power supply which results in biasing of the junction.

to p-type encounter a potential barrier as shown in Fig. 11.5, while those electrons trying to move from p-type to n-type pass easily down the potential ramp.

It should be noted, however, that the current–voltage characteristics of a pn junction are strongly temperature dependent. By raising the temperature from 20°C to 50°C the reverse bias current can be raised by more than an order of

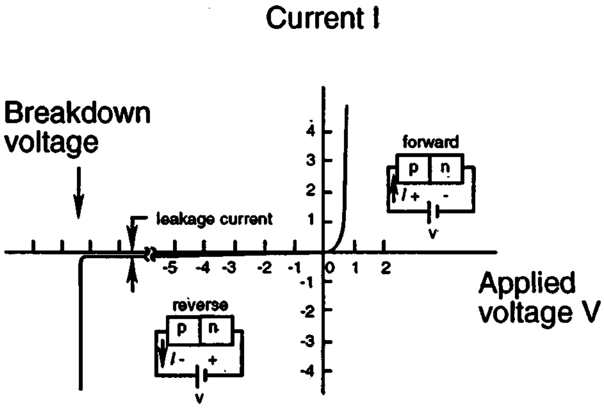


Figure 11.6 Current–voltage characteristics of a pn junction.

magnitude, as a result of thermal stimulation of more electrons across the band gap, resulting in a higher charge carrier density in the conduction band.

11.3.3 Bipolar junction transisitors (BJTs)

How does a transistor work?

Transistors are solid-state devices for amplifying and controlling electrical signals (voltages and currents). The bipolar junction transistor (BJT), which was invented at Bell Laboratories in 1947, is the prototype of many important electronic devices and we shall briefly discuss its operation. This consists of two semiconductor junctions, with one semiconductor region common to both, Fig. 11.7. There are two possible configurations, 'n-p-n' and 'p-n-p'. The first transistors were made of germanium because at the time it could be produced in pure form more easily than other semiconducting materials, such as silicon. The region in the middle of this semiconducting 'sandwich' is the base. The other two are known as the emitter and collector. All three have direct electrical connections in a circuit.

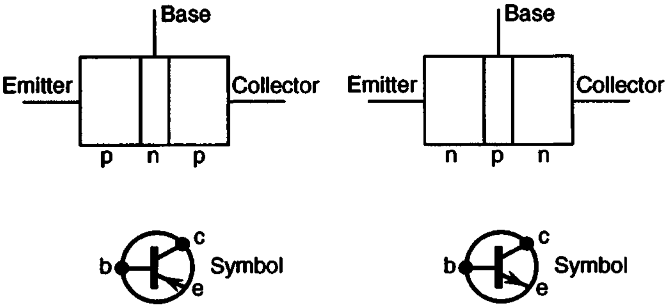


Figure 11.7 Configurations for a transistor. Both involve sandwiching one type of semiconductor between pieces of another type of semiconductor.

11.3.4 Band structures of bipolar junction transistors

What are the special characteristics of the electron band energy levels in the vicinity of the transistor junctions?

In a pnp transistor the valence and conduction bands of the centre or n-type material are lower than for the p-type material on the outside. This is shown in Fig. 11.8.

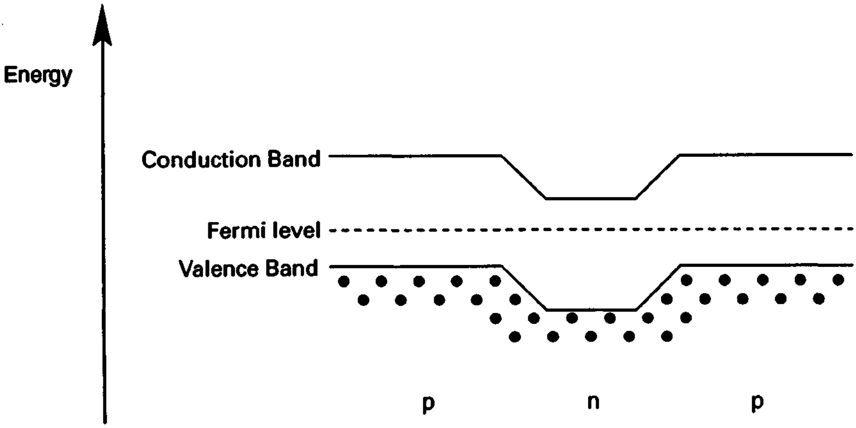


Figure 11.8 Relative energies of conduction and valence bands of a pnp transistor.

In an npn transistor the situation is just the opposite. In this case the energy levels of the centre p-type region are higher than those of the outer n-type region as shown in Fig. 11.9.

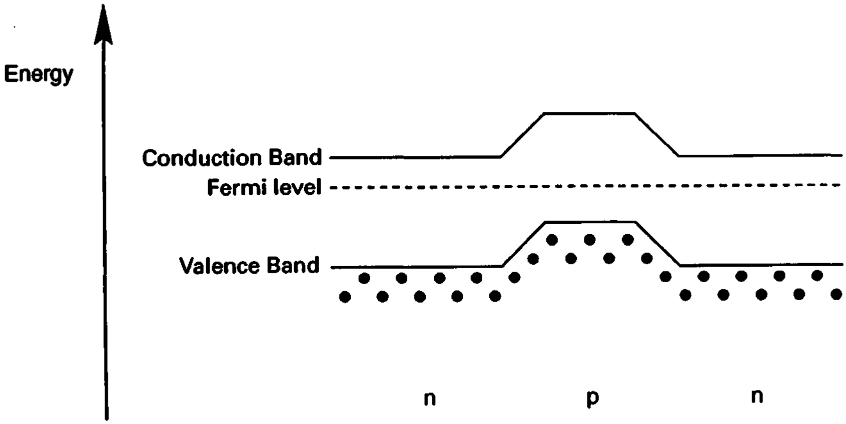


Figure 11.9 Relative energies of conduction and valence bands of an npn transistor.

11.3.5 Effects of biasing the npn transistor

What happens to the electron band structure of an npn transistor when subjected to a bias voltage?

As can be seen in the case of the pn junction the effects of biasing the junction can radically affect the current–voltage characteristics. The same is true for a bipolar junction transistor. The electron band structure of a biased transistor is shown in Fig. 11.10.

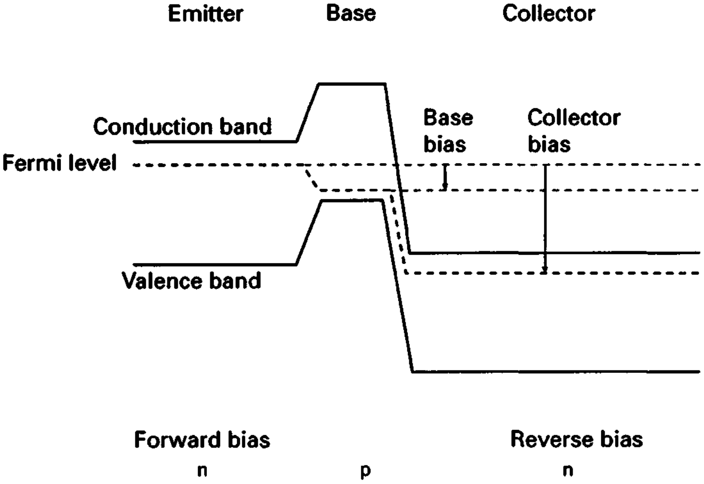


Figure 11.10 Electron band structure of biased npn transistor.

11.3.6 Typical bipolar junction transistor characteristics

How can a transistor be controlled by the bias voltage so that it acts as an amplifier?

A transistor can be used to amplify signal voltages. We will consider an npn transistor as an example. The principles of operation remain the same for a pnp transistor. The emitter–base is forward biased while the base–collector is reverse biased. Electrons climb the barrier between emitter and base being pushed up the potential by the applied voltage, as shown in Fig. 11.11. They then rush down the potential ramp from base to collector, as shown in Fig. 11.12 producing a high current.

Of the three principal configurations of a BJT transistor we will consider only the 'common base' and 'common emitter' arrangements. In the common base configuration the input voltage is V_{be} between the base and the emitter, and the output voltage is V_{cb} between collector and base. The current gain is then simply the change in collector current Δi_c for a given change in emitter current Δi_e

$$\alpha = \frac{\Delta i_c}{\Delta i_e}. \qquad (11.1)$$

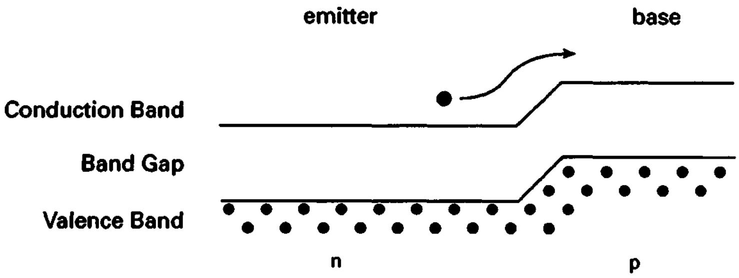


Figure 11.11 Electrons being pushed up the potential ramp from the emitter to the base of an npn transistor.

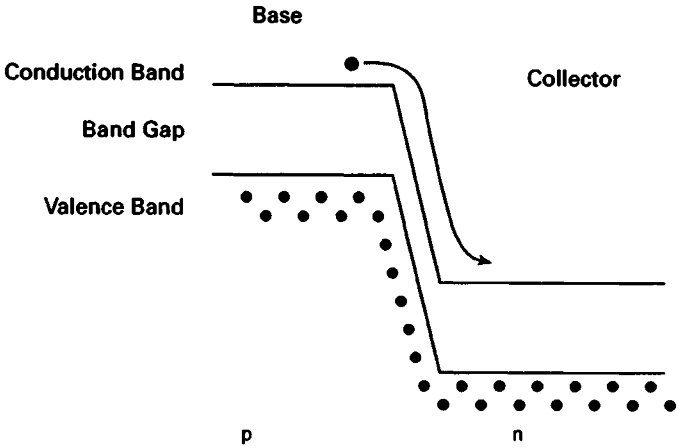


Figure 11.12 Electrons descending the potential ramp from base to collector in an npn transistor.

Since the collector and emitter currents are virtually the same in these devices as explained below, this means that the gain α is close to unity.

In the common emitter configuration the input voltage V_{be} is between the base and emitter and the output voltage V_{ce} is between the collector and emitter. In this case the current gain is the change in collector current Δi_c for a given change in base current Δi_b:

$$\beta = \frac{\Delta i_c}{\Delta i_b}. \tag{11.2}$$

This gain is usually much larger than α. By conservation of charge it follows that

$$i_b + i_c + i_e = 0, \tag{11.3}$$

and therefore

$$\Delta i_b = -(\Delta i_c + \Delta i_e), \tag{11.4}$$

hence

$$\Delta i_b = -\Delta i_c \frac{(\alpha - 1)}{\alpha}, \tag{11.5}$$

leading to

$$\beta = \frac{\Delta i_c}{\Delta i_b} = \frac{\alpha}{(1 - \alpha)}, \tag{11.6}$$

which gives the relationship between the common base and common emitter gains. Since α is close to unity this means that β is very large by comparison, being typically in the range 10–1000. Therefore a small change in the base current can lead to a large change in the collector current.

The explanation of the behaviour of the device is as follows. An increase in the base voltage reduces the potential barrier between emitter and base regions allowing more electrons to pass into the base. The base region itself is very thin, so that a high proportion of the electrons entering from the emitter pass into the collector without recombining with holes in the base. This gives a collector current which is comparable to the emitter current. A large increase in collector current can therefore result from a small increase in base current.

The base current is typically of the order of microamps, while the emitter and collector currents are typically of the order of milliamps. The voltage–current characteristics of a common emitter configuration of an npn transistor are shown in Fig. 11.13.

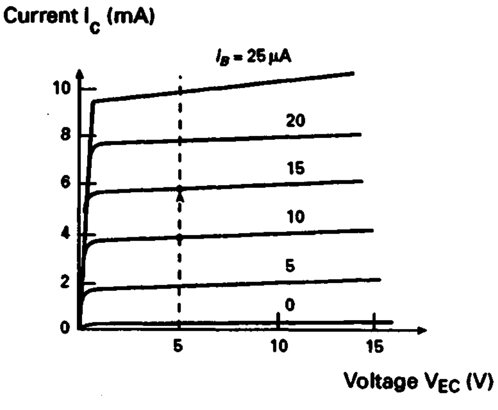


Figure 11.13 Current–voltage characteristics for a BJT.

11.3.7 Charge carrier lifetimes and operational efficiency

How are the performances of junction devices such as pn junctions and transistors affected by the lifetimes of charge carriers?

In devices such as pn junctions and transistors the charge carrier lifetime, and hence the diffusion length, is an important factor in determining operational efficiency. For example, when minority charge carriers are injected into these devices they can contribute to the minority charge carrier concentration only until they recombine in an electron/hole pair. This means that longer charge carrier lifetime results in more charge carriers reaching the depletion (space charge) region and consequently in higher operational efficiencies of such devices.

11.3.8 Field effect transistors (FETs)

What other types of transistor can be produced?

A few years after the invention of the bipolar junction transistor the junction field effect transistor was developed. This is essentially a voltage-controlled resistor. The common feature of field effect transistors is that they employ predominantly one type of carrier. The junction field effect transistor (JFET), which we will consider as an example, is such a unipolar device. This distinguishes it from the BJT, which is a bipolar device wherein both types of carriers are employed.

The JFET consists of either an n-channel or a p-channel as shown in Fig. 11.14. The 'source' and 'drain' are on either side of the 'channel' through which the current flows. The 'gate' is a piece of p-type (or n-type) depending on whether the channel is n-type (or p-type). The voltage applied to the gate controls the flow of charge carriers and so controls the resistance of the device. The variation of the drain current i_d with the voltage across drain and source V_{ds} is nonlinear. When a positive voltage is applied to the drain relative to the source electrons flow from

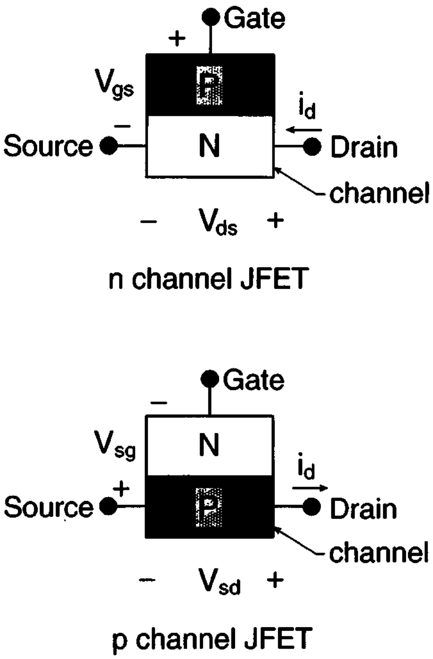


Figure 11.14 Schematic for JFETs.

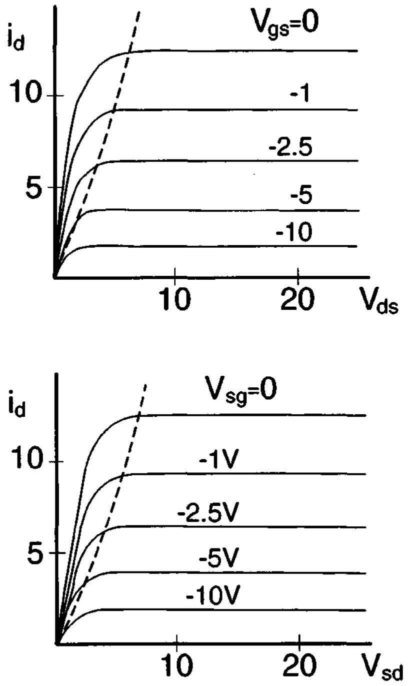


Figure 11.15 Current–voltage characteristics for a JFET.

source to drain. At low values of V_{ds} the drain current i_d increases rapidly with voltage but then saturates as V_{ds} is increased further. The value of i_d at saturation is controlled by voltage across the gate and source V_{gs} as shown in Fig. 11.15. The reason for this is that the voltage on the gate can be used to vary the width of the depletion layer in the channel region. At zero gate voltage the drain current is a maximum. As the gate voltage is made more negative it increases the width of the depletion layer which results in the conduction channel becoming narrower. On the other hand if the gate voltage is made more positive this provides an alternative path for the charge carriers which reduces the flow of electrons from source to drain.

The field effect transistors comprise a whole family of related devices which include, in addition to the JFET, the metal semiconductor field effect transistor (MESFET) and the metal oxide semiconductor field effect transistor (MOSFET).

11.3.9 Metallic spin transistors

Is it possible to make a transistor out of a metal?

Over the entire lifetime of the microelectronics industry, semiconductors have been used for making transistors. However, it has been suggested [7] that magnetic metals can be used to construct a transistor. The advantage of such a device is that the carrier densities in metals are typically 10^{28}–10^{29} m^{-3}, whereas in doped semiconductors, they are 10^{24}–10^{25} m^{-3}. So the devices made from metals could be about ten thousand times smaller than semiconductor devices for the same number of conduction electrons. The magnetic metals can sustain two separate populations of charge carriers distinguished by the orientations of their spins (up and down). These metallic transistors could be used for memory (nonvolatile RAM), amplifiers and logic circuits.

11.4 MICROELECTRONIC SEMICONDUCTOR DEVICES

How many such simple devices can be fabricated on a single silicon chip?
We have seen that it is possible to produce devices with interesting and diverse electrical properties by altering the electronic structure of materials. Three simple examples were the pn junction, the BJT and the FET, but these are only the simplest of devices that can be fabricated from semiconductor materials such as silicon. Large numbers of these devices can be produced on a single wafer of silicon. This process is known as very large scale integration (VLSI) [8]. Fabrication of even larger units, for example complete fabrication of all devices on a single silicon wafer, is known as wafer scale integration (WSI) [9]. The numbers of transistors fabricated on a single silicon chip have grown exponentially since the first integrated circuits were produced. Semiconductor fabrication technology has reached 100 million (10^8) transistors per chip as shown in Fig. 11.16. Furthermore, the number density of components on dynamic RAM chips has always been higher than on microprocessors and this has reached one billion (10^9) transistors per DRAM chip.

11.4.1 Integrated circuits

What is an 'integrated' circuit?
An integrated circuit is an assembly of electrical components that is fabricated as a single unit on a substrate of semiconducting material which is usually silicon. The IC consists of an assembly of electrically isolated circuit elements, both passive (capacitors and resistors) and active (transistors and diodes). These are fabricated together with the necessary electrical connections on a common substrate, which is usually silicon. These circuit elements are arranged in such a way that the whole integrated circuit performs an electrical circuit function. The integrated circuit was invented by Kilby at Texas Instruments in 1958 and independently by Hoerni and Noyce of Fairchild Semiconductor Corporation at about the same time. The first integrated circuits built in 1958 contained 10 components fabricated on a single chip of semiconductor. Modern ICs contain about 10^8 transistors, are typically 5 mm square and are fabricated simultaneously on silicon wafers 100 mm in diameter which are subsequently separated into individual chips.

11.4.2 Microprocessors

What is a microprocessor?
Large-scale integration was developed during the 1970s and this enabled thousands of transistors to be packed on a silicon chip that is 3–4 mm square. This gave

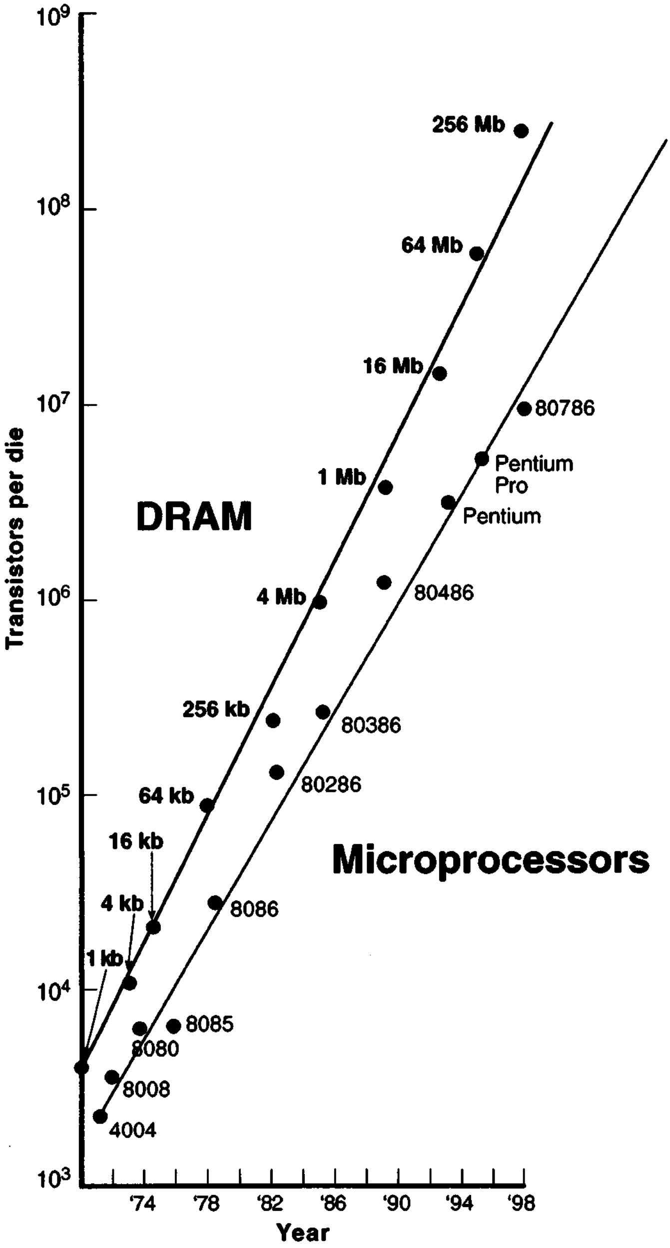


Figure 11.16 Device densities on microprocessors and DRAMs, have followed a revised Moore's Law, which stipulates the doubling in circuit complexity every 18 months. After R. R. Schaller [29]. © 1997 IEEE.

rise to the microprocessor, which is an advanced integrated circuit that, in addition to the conventional circuit functions, contains the arithmetic, logic and control circuitry necessary to carry out the functions of a central processing unit (CPU) of a digital computer. The microprocessor was invented by Intel Corporation in 1971. The first microprocessor had about 1000 transistors. Very-large-scale integration (VLSI) was developed in the 1980s. This further increased the density of devices on the chip. In the early 1990s microprocessors were fabricated with more than 3 million transistors on a single microprocessor. By the late 1990s the number had grown to 10 million transistors on the 80786 microprocessor, and by the year 2001 industry was fabricating 100 million transistors on a chip.

11.4.3 Fabrication procedure

How are the devices actually made out of a single block of silicon?

The fabrication process for producing devices on silicon wafers is quite complicated and involves a large number of steps. The complete process takes several days to complete. However if we simplify the process into its principal stages it proceeds as follows:

(i) Production of single crystal silicon wafer.
(ii) Doping of material, either by diffusion of impurities or by ion implantation, followed by diffusion.
(iii) Growth of epitaxial layers on surface of wafer, such as lightly doped silicon on a heavily doped substrate. This is the stage at which the material is supplied by the wafer manufacturer.
(iv) Ion implantation into the surface to create regions with different electrical properties such as n-type and p-type regions.
(v) Subsequent heat treatments to allow implanted dopants to diffuse farther into the material to establish the desired well structure.
(vi) A thin 0.03-μm layer (gate oxide) of silicon dioxide is produced over the surface, followed by a thin 0.17-μm nitride layer.
(vii) Lithography is the technique that is used to define which areas on the surface are to be etched and which are to be left with the nitride layer. The nitride areas are for sources, drains and channels (electrically active areas).
(viii) A thicker oxide layer 0.6 μm (field oxide) is then deposited on areas without the nitride coating, since these areas are to be covered by insulator.
(ix) The nitride layer is removed by the use of orthophosphoric acid leaving only the thin 'gate oxide' layer in the chosen regions.
(x) The silicon/silicon dioxide interface contains fixed positive charges in the oxide. These are compensated by negative charges just inside the silicon. Since these positive charges are desirable, ion implantation of acceptor atoms (for example BF) in the silicon is used to ensure that it is easier to get mobile positive charges into the oxide layer [10].
(xi) Polycrystalline silicon gates are aligned at the active sites. The silicon is heavily doped to increase its conductivity. These gates are deposited over the thin gate oxide regions. This stage is known as gate deposition.

(xii) Photoresist layers are then placed over the surface in two steps. First, the surface, except for the p wells with their heavily doped polysilicon gates, is covered. Then donor atoms such as arsenic are ion implanted deep enough to penetrate the areas covered only by the thin oxide layers in the p wells. Next, the surface, except for the n wells with their heavily doped polysilicon gates, is covered again with photoresist. Then acceptor atoms such as boron are ion implanted into the areas covered only by the thin oxide layer in the n wells.

(xiii) The photoresist layer is removed using sulphuric acid and hydrogen peroxide. This leaves an uneven surface with a top layer of about 1 μm of silicon dioxide with some contact holes in the surface for electrical connections.

(xiv) The planarization stage involves the use of an organic silicate in alcohol solution which is cured at 800°C to leave a fairly smooth glassy surface ready for attachment of metal electrical contacts.

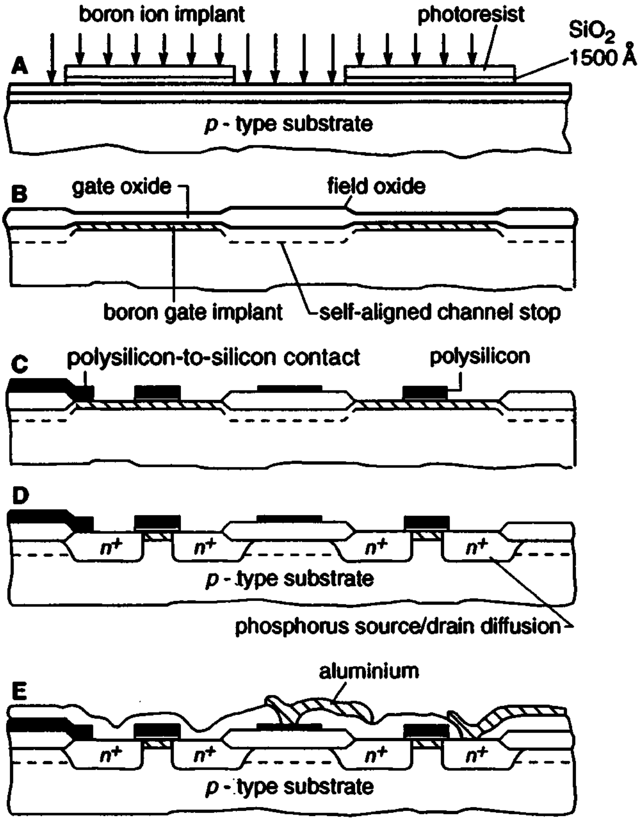


Figure 11.17 Section through a MOSFET transistor at various stages of fabrication.

(xv) Once the surface has been smoothed metal layers can adhere to it and these layers are laid down to form the electrical connections. This process is called metallization. The deposition involves sputtering of a metallic surface layer using conventional lithography to obtain the required pattern. The rate of deposition is typically 0.01 μm per second. The metal layer usually consists of a 0.15-μm titanium–tungsten layer to prevent aluminium migrating into the silicon, and on top of this 0.5-μm layer of Al–4%Cu.

Some of the principal stages in the fabrication process are shown in Fig. 11.17.

11.4.4 Lithography, critical dimensions and feature sizes

How large are the devices that are fabricated on semiconductor 'chips'?
Lithography is the process by which the patterns for chip designs are produced on semiconductor wafers. Optical lithography is the standard technique at present but may soon be reaching its limit as the sizes of devices that need to be patterned continues to be reduced. Lithography requires an exposure tool, a mask technology and a resist technology. The wafer is coated with a layer of photoresist. An image of the mask, usually reduced to one-quarter or one-fifth of its original size, is projected onto the wafer. Exposure to light changes the solubility of the photoresist layer so that the desired pattern emerges when it is developed. The remaining photoresist layer enables selective etching or selective ion implantation to be carried out.

The wavelength of the light used in photolithography must be typically no greater than the critical dimensions. There are some techniques that overcome this limitation such as resolution enhancement technologies ('RET'). These can produce features at roughly half the wavelength of the light by using phase shifting masks. However, even in these cases the wavelength can be no greater than twice the critical dimension. As the technology develops over time the feature size continues to be reduced and therefore the wavelength of the light used in the patterning must also be reduced. The 'critical dimension' is the size of the smallest linewidth that can be patterned onto a semiconductor wafer. This is currently about 0.15 μm but will reach 0.07 μm in the next decade as shown in Fig. 11.18 [11]. To continue to reduce this critical dimension other advances in technology are needed.

With features on the order of 0.15 μm today 'light' of wavelength 150 nm is used for the patterning. Looking further ahead, when the features have critical dimensions of less than 0.07 μm we will need to use extreme ultraviolet (EUV) lithography with electromagnetic radiation of wavelength 70 nm (photon energies of 17 eV) for the patterning. At critical dimensions of 0.07 μm it should be possible to fabricate ICs with about 500 million transistors per chip. According to Geppert [11] there are no fundamental barriers lying between the current technology and such device densities. At smaller critical dimensions, or higher device densities, there are problems in continuing to use 'optical' lithography even with EUV. This dictates that other lithographic methods will need to be used. A candidate for this is electron beam lithography.

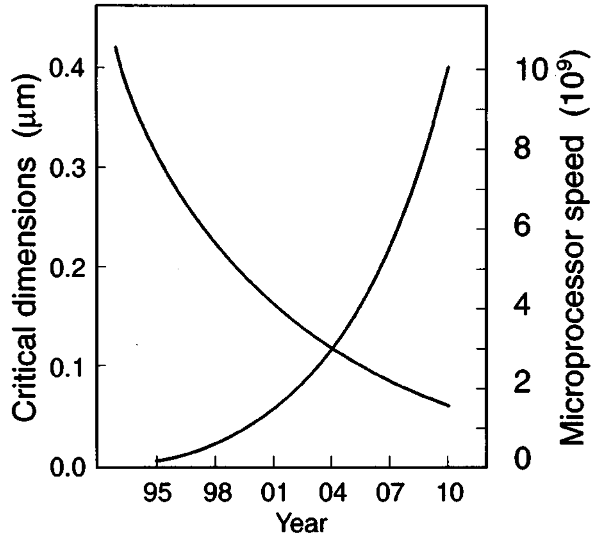


Figure 11.18 The critical dimensions of transistors will decrease from 0.15 μm today to 0.07 μm in 2010. Over the same time span, the performance of general-purpose microprocessors will rise from 100 million operations per second to more than 10 billion [11].

The reduction in feature size also results in a reduction of the power requirements and threshold voltage for operation, as shown in Fig. 11.19 [12]. At the 100 million transistors per chip level, the threshold voltage for operation of the devices is 0.2 V and the operating temperature range is up to 100°C. Of course, it is known that the chips operate better at lower temperatures. At lower temperatures, the electron and hole mobilities are increased and the resistances of the metallic interconnects are reduced, while the threshold voltage for operation decreases. Nevertheless, the incorporation of cooling systems into the microelectronics seems to be impractical at present.

A typical CMOS device of the 0.1 μm lithography generation should look like the diagram in Fig. 11.20. The channel length (which is the distance between the source and drain of the transistor) should have a minimum size of 0.05 μm and an unloaded circuit delay of 10 ps. The gate oxide thickness is typically in the range of $\frac{1}{25}$ to $\frac{1}{50}$ of the 'channel length', which means that the gate oxides will be 1–2 nm in thickness in these devices. This will be a layer that is only a few atoms thick and quantum-tunnelling effects will be significant [13]. This means that the traditional scaling of device sizes, which assumes classical continuum physics, will break down at these dimensions and this will lead to unacceptable device performance [14]. Furthermore, dynamic random access memory chips (DRAM) have even more stringent requirements on leakage currents than logic chips and therefore the insulating gate oxide layers need to be thicker in DRAM.

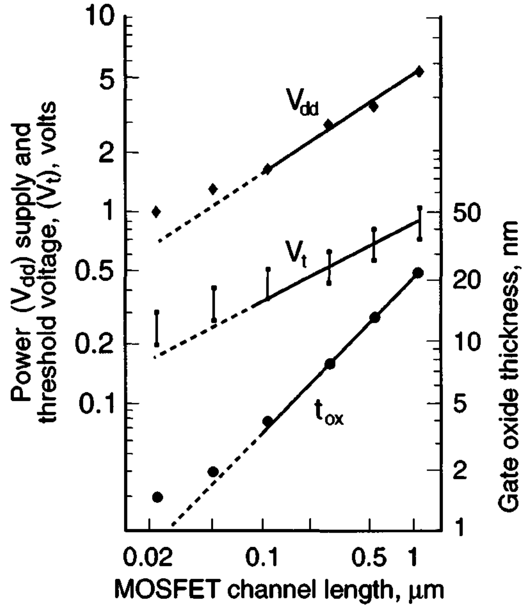


Figure 11.19 As IC manufacturing techniques improve, CMOS devices with shorter channel lengths will switch faster and use less power. Lower power-supply and threshold voltages (V_{dd} and V_t) and thinner gate oxides (t_{ox}) will accompany the shrinking channel lengths. Transistors built between the years 2003 and 2006 will have a minimum channel length of 0.05 μm, a power-supply voltage of 1.2 V, and a threshold voltage near 0.25 V [12].

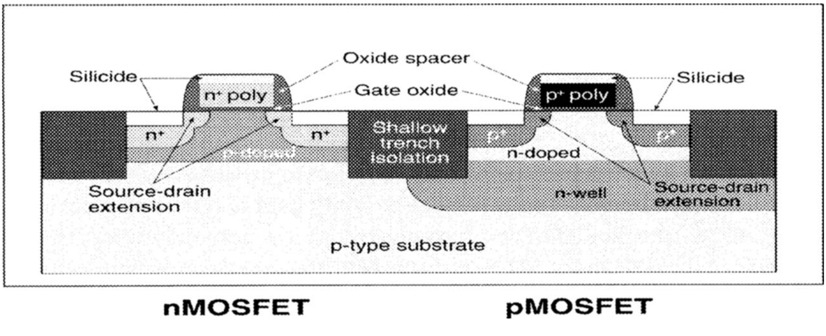


Figure 11.20 A CMOS device in the 0.10–0.13-mm lithography generation will feature: shallow trench isolation, 1.5–2.0-nm insulating gate oxide, n^+-doped and p^+-doped polysilicon gates for nMOSFET and pMOSFET, respectively, 30–50-nm-deep source-drain extension, and self-aligned silicide for the contacts [13].

11.5 FUTURE IMPROVEMENTS IN SEMICONDUCTORS

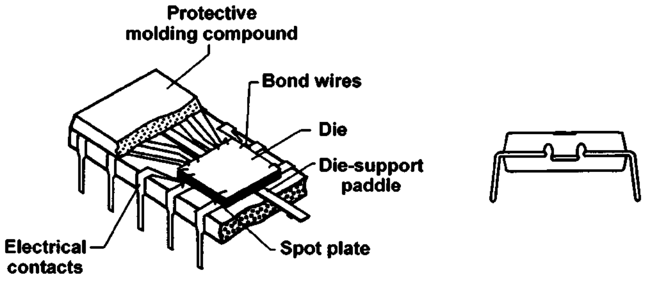
What are the directions in which semiconductors are likely to develop in the near future?

The present-day microelectronics industry is built on the fabrication of microscopic electrical circuits engraved on silicon. These circuits are then sealed, packaged to protect them, and electrical connections to the outside world are added. A typical microelectronic circuit, the proverbial 'computer chip', appears as shown in Fig. 11.21. One direction for further development is in the refinement of techniques for fabrication, particularly in view of the sensitivity of the electronic properties of semiconductors to impurities.

The number of components on a logic chip is a simple function of the feature size. Figure 11.16 shows the progress in reducing feature sizes, and increasing the number of components, from 1970 to the present with extrapolations to the future. It also shows the dates at which various ICs became available and the numbers of transistors incorporated. The plot shown in this figure is an example of Moore's law, which, as originally stated in 1962, predicted that the number of transistors on an IC chip doubled every three years.

With present device sizes, the individual transistors on an IC are switched by the action of about 1000 electrons being added or removed. By extrapolation of the current rate of progress in reducing feature sizes the switching will be achieved by the addition or removal of only 8 electrons by 2010, and by only one electron by 2020 [15].

The operational speed of logic chips is measured in the number of binary operations per second (bops). This is the product of the number of devices (gates) and the clock frequency. In the case of the Teramac computer developed recently by Hewlett Packard Laboratories, there are 10^6 gates operating at 1 MHz giving 10^{12} bops. It is interesting to ask how much electrical energy is involved in this



Cut away section

External end view

Figure 11.21 Cut-away section through an integrated circuit showing the silicon chip together with the necessary circuit connections. The whole is encapsulated in a protective plastic housing.

very large number of operations. Landauer has shown [16] that it is possible in principle to perform 3.5×10^{20} bit operations per second with a power of 1 W. This indicates that there is still plenty of spare capacity left with CMOS technology even at the operational speed of the Teramac. Another direction is the use of amorphous silicon [17] instead of the present single crystal silicon. This has distinct advantages in terms of reduced cost. The main application area for amorphous silicon is in low-cost, high-efficiency photovoltaic cells.

The need for newer, faster semiconductors seems to be continuing presently without relent. Single-element semiconductors based on the group IV elements germanium and silicon are now developed to full maturity, which means there is little scope for further radical improvement using these materials alone. Compound semiconductors composed of elements from groups III and V of the periodic table have been under investigation for some years [18]. Gallium arsenide comes into this category. This material can be used to fabricate high-speed devices because of its high electron mobility (low effective mass of electrons). The detection and generation of light by the semiconductor is another great advantage because of its use in optoelectronic applications for which there is an increasing demand as discussed in Chapter 12.

11.5.1 Fabrication of devices with ever smaller feature sizes

What limitations, if any, are going to be encountered as the sizes of devices on semiconductor chips continue to be reduced?

The width of the smallest feature of a transistor fabricated on a semiconductor is known as the 'linewidth'. As shown in Table 11.1 [19], the linewidth is currently 0.15 μm. In the past the linewidth of transistors on chips has been halved every 6 years; while clock speed, which indicates how fast devices can operate, has doubled every two years. Furthermore, as a result of these advances, the storage density of RAM has doubled every 18 months. The table shows an ambitious set of goals for the industry. However, in order to achieve these, some radical changes in materials and designs are needed.

Electron beam lithography was developed by Bell Laboratories. This uses 100-keV electrons together with a 4 : 1 reduction image projection technique [20]. Electron beam lithography circumvents the problem of diffraction limitations encountered with optical lithography. It also uses an incoherent beam of electrons so that there are no interference effects. The feature sizes that can be produced are much greater than the 3.7-pm electron wavelength, with critical dimensions of

Table 11.1 Shrinking semiconductor chip feature sizes [21].

Year of production	**1999**	**2002**	**2005**	**2008**	**2011**	**2014**
Line width, nm	180	130	100	70	50	35
Microprocessor gate length, nm	140	85–90	65	45	30–32	20–22
Oxide thickness, nm	1.9–2.5	1.5–1.9	1.0–1.5	0.8–1.2	0.6–0.8	0.5–0.6
Junction depth, nm	47–70	25–43	20–33	16–26	11–19	8–13
Interconnect dielectric constant, *k*	3.5–4.0	2.7–3.5	1.6–22	1.5	<1.5	<1.5

35 nm being achievable using this method. The drawbacks of electron beam lithography are its low throughput capability and its high cost. This technique is one of two being considered for further development for sub-70-nm lithography. The other is the continuation of deep ultraviolet lithography [21].

The progressive reduction in the physical sizes of components on chips only provides solutions over the next few years. Beyond that the electrical insulation requirements become a problem because of the small feature sizes. To attain 100-nm linewidths, the oxide layer thickness will need to be reduced to about 1 nm (or about 3 atoms), and this means that the electrical insulation provided by the current generation of materials will be inadequate because of quantum-tunnelling currents.

Alternative materials with higher dielectric coefficients such as silicon nitride are coming under consideration. In fact, SiN is already being mixed into the gate oxide of chips to improve insulation. Another way to solve the insulation problem as features get smaller, is to use the 'silicon on insulator' (SOI) technology. In this case, the insulating silicon dioxide layer is buried under the transistors [22]. The silicon on insulator chips can operate at lower voltages, thereby reducing the problem of leakage currents. These chips can also operate at higher temperatures than the conventional chips, and the capacitance of devices fabricated on the chips are lower, leading to potential for operation at higher frequencies.

The overall speed of operation of devices on chips depends on the resistance and capacitance of the components. SOI offers the prospect of reduced capacitance, and resistance could be reduced if copper interconnects were used instead of aluminium. However, as is well known, copper 'poisons' silicon. Nevertheless, the advantages of reduced resistance are so great that chip producers are changing to copper connections and placing the necessary chemical barriers between Cu and Si. However, this has the disadvantage of taking up valuable space on the chip.

Semiconductor fabrication and lithography systems currently routinely produce Si wafers of diameter 200 mm complete with electronic components fabricated on them. In the near future wafer-handling capabilities will be extended to 300-mm diameter [23].

11.5.2 Gallium arsenide

What are the relative advantages and disadvantages of gallium arsenide over silicon?
The properties of GaAs and Si are compared in Table 11.2. From this table, it is clear that the electron mobilities in GaAs are much higher than in silicon, and that the electron band gap is direct in GaAs while indirect in Si. Figures 7.4 and 7.5 showed the relevant positions of the electron band structures of these two semiconductors, from which it is clear that the valence bands of the two semiconductors are quite similar, but the conduction bands are very different [24].

The lower effective mass of the electrons in GaAs arises from the curvature of the conduction band, particularly close to the centre of the Brillouin zone. Charge carriers move faster in n-type semiconductors (where they are electrons) than in p-type semiconductors (where they are holes). The ratio between the mobilities of the charge carriers in GaAs is greater (10:1) than in silicon (3:1).

Table 11.2 Comparison of the properties of gallium arsenide and silicon [24].

Property	**GaAs**	**Silicon**
MOS grade thermal oxide	No	Yes
Thermal conductivity (W/cm K at 300 K)	0.46	1.45
Linear thermal expansion coefficient (°C)$^{-1}$	5×10^{-6}	2.6×10^{-6}
Electrical breakdown field (V/cm)	4	3
Relative dielectric constant	12	11.8
High-quality n-type diffusion	No	Yes
Electron mobility in n-type material (Surface/bulk cm^2/V s[a])	N.A./4000	400/900
Bandgap	Direct	Indirect
Total dose radiation tolerance (rads)	$<10^8$	10^6–10^7

[a] Where $N_D = 10^{17}/\text{cm}^3$ and $T = 300$ K

As a direct result of the lower effective mass of electrons in GaAs, the charge carriers can cross the channel (the region of the transistor between the emitter and collector) in less time. This means that GaAs devices have higher switching speeds and hence have a higher maximum frequency of operation, than silicon-based devices. In the present 0.5 μm technology, the switching time for a GaAs device is about 70×10^{-12} s, whereas for Si, it is about 150×10^{-12} s. The principal areas of application of GaAs have been described by Chang and Kai [25]. The main application for GaAs microprocessors is in the manipulation of higher-speed serial digital data, while other applications include optoelectronic devices which take advantage of the direct band gap.

Gallium arsenide circuits, in addition to being faster than similar silicon circuits, consume less power, have lower noise, can radiate and detect light more efficiently in the visible wavelength range and have an easily engineered band gap when alloyed with other similar semiconductor materials. For example, when alloyed with gallium phosphide the band gap can be altered from 1.42 to 2.26 eV corresponding to wavelengths in the optical range from red to green. In view of all these advantages it might well be asked why there has not been a wholesale switch from silicon technology to gallium arsenide technology in the semiconductor industry. The reasons are, however, economic: gallium arsenide is both expensive to produce and difficult to fabricate into integrated, single chip, circuits and devices, while silicon is very plentiful and is relatively easily fabricated.

On the other hand, the most rapidly developing applications of semiconductors are in photonic transmission of information ('optical computing') and a high level of interest has been aroused in gallium arsenide and related materials. Modern computers can produce data streams which are too fast for conventional copper wire conductors to transmit over distances greater than about 200 metres. The result of conventional transmission under these conditions is that consecutive signals begin to blur. Optical communication overcomes this limitation and allows computers to communicate at high speed over very large distances using optical fibres and light emitting diodes or diode lasers. It is in this particular application that the optical semiconductors are finding important uses, since despite the difficulty of fabrication they provide functions that other semiconductors cannot.

11.5.3 Band gap engineering with various III-V semiconductor

Since the band gap is the dominant characteristic of a semiconductor how can materials with predetermined band gaps be fabricated?

A large band gap in a semiconductor eliminates much of the intrinsic contribution to the conductivity by making thermal stimulation of electrons across the band gap less probable. The larger the band gap the lower the number density of electrons in the conduction band. This means that the conducting properties of the semiconductor can be more precisely controlled by the addition of donor or acceptor elements if the band gap is larger.

The III-V semiconductor compounds include Ga/Al/In-As/P/Sb which can be alloyed together to form solid solutions. This allows the band gap in these materials to be varied simply by changing the chemical composition [26]. The various band gaps of the alloys are shown in Fig. 11.22.

One of the first alloy semiconductors of this group to be used was the ternary compound $GaAs_{1-x}P_x$ which was used for red light emitting diodes. The maximum brightness of this alloy as an LED was found to occur with $x = 0.4$ which gave an energy gap of 1.9 eV. Other compound III-V semiconductors used for optical applications include $Al_xGa_{1-x}As$ and $Ga_xIn_{1-x}As_yP_{1-y}$. The ability to change the band gap, as in the GaAs-GaP and GaAs-AlAs compounds, is important for optical communication and emission and detection of light at different wavelengths [27].

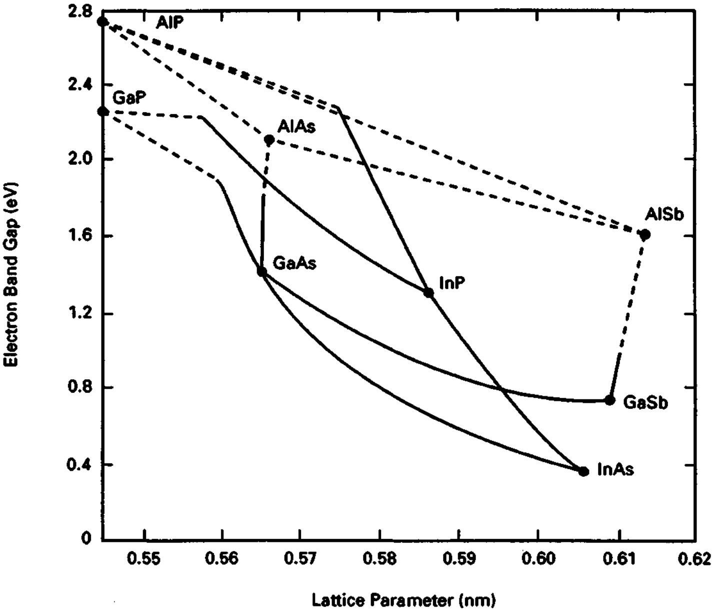


Figure 11.22 Variation of the band gap energy with lattice parameter for III-V semiconductors.

11.5.4 Silicon carbide

What other semiconductor material are being considered in order to produce microprocessors and integrated circuits with improved performances?
For high-temperature, high-power, or high-frequency operation of semiconductor devices silicon carbide has come under consideration in those operational ranges that are not suitable for either silicon or gallium arsenide. The material can operate at temperatures of 300 °C and also can be used in high radiation level environments. The material has a relatively large indirect band gap ranging from 2.4 eV to 3.35 eV depending on the particular structure of silicon carbide produced. In this respect, it competes with other large band gap materials including diamond and the nitrides of the group III elements. Among other interesting properties, it has a very high breakdown field strength of $3.8 \times 10^8\,\mathrm{V\,m^{-1}}$, a high electron drift velocity of $2 \times 10^5\,\mathrm{m\,s^{-1}}$ and a high thermal conductivity of $490\,\mathrm{J\,s^{-1}\,m^{-1}\,K^{-1}}$. Although preparation of adequate quality device structures from silicon carbide was initially a problem, because of the difficulty in eliminating defects in the material, this has now been overcome [28].

REFERENCES

1. S. M. Sze (1998) *Modern Semiconductor Device Physics*, John Wiley, New York.
2. J. Allison (1990) *Electronic Engineering Semiconductors and Devices*, McGraw Hill, London.
3. M. Riordan and L. Hoddeson (1997) *IEEE Spectrum* 34 (6), 46.
4. J. Bardeen and W. H. Brattain (1949) *Phys. Rev* 74, 230, and 75, 1208.
5. A. R. Peaker (Ed.) (1990) *Properties of Gallium Arsenide*, IEE, Inspec, London.
6. M. H. Brodsky (1990) Progress in gallium arsenide semiconductors, *Scientific American* 262 (2), 68.
7. M. Johnson (1994) *IEEE Spectrum* 31 (5), 47.
8. S. K. Ghandi (1977) VLSI fabrication principles, (1983). *Scientific American* 237 (3).
9. J. F. McDonald (1984) The trials of wafer scale integration, *IEEE Spectrum* 21 (10), 32.
10. J. Irby *et al.* (1992) In situ B doped Si epitaxial films grown at 450°C by remote plasma enhanced CVD: physical and electrical characterization, *J. Electronic Mater.* 21, 543.
11. L. Geppert (1996) *IEEE Spectrum* 33 (4), 34.
12. Y. Taur (1999) *IEEE Spectrum* 36 (7), 25.
13. Y. Taur and T. H. Ning (1998) *Fundamentals of Modern VLSI Devices*, Cambridge University Press, New York.
14. P. S. Peercy (2000) *Nature* 406, 1023.
15. J. Birnbaum and R. S. Williams (2000) *Physics Today* 53(1), 39.
16. R. Landauer (2000) *Physics Today* 53 (1), 39.
17. A. Madan (1986) Amorphous silicon: from promise to practice, *IEEE Spectrum* 23 (9), 38.
18. T. E. Bell (1985) Japan reaches beyond silicon, *IEEE Spectrum* 22 (10), 46.
19. E. J. Lerner (2000) *The Industrial Physicist* 6 (3), 18.
20. L. R. Harriot (1999) *IEEE Spectrum* 36 (7), 41.
21. T. Ito and S. Okazaki (2000) *Nature* 406, 1027.
22. J. P. Collinge and R. W. Bower (1998) *MRS Bull.* 23 (12), 13.
23. Y. Taur (1999) *IEEE Spectrum* 36 (7), 25.

24. I. Deyhimy (1995) *IEEE Spectrum* 32(2), 33.
25. C. Y. Chang and F. Kai (1994) *GaAs High Speed Devices*, John Wiley, New York.
26. T. J. Drummond (1988) Quantum tailored solid state devices, *IEEE Spectrum* 25 (6), 33.
27. W. D. Goodhue *et al.* (1990) Monolithic two dimensional GaAsAlGaAs laser arrays fabricated by chlorine ion beam assisted micromachining, *J. Electron. Mater.* 19, 463.
28. M. A. Capano and R. J. Trew (1997) *MRS Bull.* 22 (3), 19.
29. R. R. Schaller (1977) *IEEE Spectrum* 34 (6), 53.

FURTHER READING

P. S. Peercy, T. Ito, S. Okazaki, A. I. Kingon, J. P. Maria, S. K. Streiffer, M. H. Devoret, R. J. Schoelkopf and S. Lloyd (2000) The future of microelectronics, *Nature* 406, 1023.

B. J. Skromme, S. S. Bose and G. E. Stillman (1986) New shallow acceptor levels in GaAs, *J. Electron. Mater.* 15, 345.

C. K. Williams (1992) Kinetics of trapping, detrapping and trap generation, *J. Electron. Mater.* 21, 711.

W. C. Holton, and R. K. Cavin (1986) *Proc. IEEE* 74, 1646.

A. Madan and M. Shaw (1987) *Physics of Amorphous Semiconducting Devices*, Academic Press, New York.

S. O. Kasap (1997) *Principles of Electrical Engineering Materials and Devices*, Irwin, Chicago.

H. S. Nalwa (2000) *Handbook of Advanced Electronic and Photonic Materials and Devices*, Academic Press, San Diego.

S. Mahajan and K. S. Sree Harsha (1999) *Principles of Growth and Processing of Semiconductors*, McGraw-Hill.

12 OPTOELECTRONICS - SOLID STATE OPTICAL DEVICES

OBJECTIVE

The purpose of this chapter is to describe one of the fastest developing areas of electronic materials. The need for optical communications between computers, or for telecommunications, is driven by the need for faster and more precise transmission of data. Conventional electrical communications suffer from problems which are related to the speed of communications and the distance over which such communication needs to be made. Optical communications through fibres can meet these requirements. There are four main components to optoelectronics for communications: light generation, transmission, detection and user interfacing. The devices for these functions are: light emitting diodes, optical fibres, photodetectors and displays. This chapter presents a selected overview of some of the materials used for optoelectronics in each of these main areas.

12.1 ELECTRONIC MATERIALS WITH OPTICAL FUNCTIONS

What do we mean by 'optoelectronics'?

Optoelectronics is the combination of optical and electronic processes in materials. For example, an electronic transition across the band gap of a semiconductor from conduction band to valence band with the emission of a photon is an optoelectronic process. In optoelectronics, we are principally concerned with the generation of light as a result of electronic processes in materials, efficient transmission of light for communications purposes and detection of incident light. The main topics of interest here are therefore lasers and light emitting diodes, photodetectors, fibre optics and optical displays.

The whole field of optoelectronics is currently receiving great attention because of the possibility of developing high-speed computers capable of communicating at a rate of more than 10 gigabits per second using optical methods. By comparison, standard electronic computers operate at frequencies of typically 50 megabits per second. The improvement in speed is, therefore, likely to be up to a factor of 200. Other developments include the ability to produce low-power, low-cost optical displays (LEDs and LCDs) using semiconductors, and the ability to fabricate inexpensive, low-power semiconductor lasers.

12.1.1 Photodetectors

How can we detect light and convert it into an electrical voltage?

A photodetector is a semiconductor device that converts light into an electrical voltage [1]. There are two types of photodetector, the photoconductor and the photodiode. The photoconductor is a material in which electrons are stimulated from the valence band to the conduction band by incident light. This leads to an increase in conductivity which can be related to the incident light intensity and wavelength.

A photodiode is a reverse-biased pn junction in which the stimulation of electrons from the valence band to the conduction band leads to a current under the action of an electric field (potential gradient). This current is proportional to the intensity of incident light provided the energy of the incident photons is sufficient to excite electrons across the band gap. The solar cell is an example of such a device. It consists of a silicon pn junction which, when placed in sunlight, generates an electric current that can flow in an external circuit connecting the two semiconductors.

Photodetectors therefore depend on the photoelectric effect which we have discussed in a previous chapter, in which absorption of photons causes electrons to be stimulated to higher energy levels, leading to either an increase in conductivity, or if a potential gradient is present, as in a pn junction, to the flow of current. The photoelectric effect in semiconductors causes a marked change in conductivity which is quite different from that observed in metals because, whereas the metals already have large numbers of electrons in the conduction band, the semiconductors do not. The relative effect on the conductivity is therefore much greater in semiconductors.

In discussing the photoelectric effect here we draw a distinction between it and photoemission in which the electrons are actually liberated from the material by the incident photons. Photoemission is characteristic of metals which have higher-energy conduction electrons available to be emitted from the material as a result of optical stimulation.

12.1.2 The pn junction as a detector

How can the simple pn junction be used as a detector?

In order to act as a detector the pn junction is reverse biased, since for sensitive detection a large fractional change in the current is required. It is fairly easy to detect a change of 1 μA if the current doubles to 2 μA, but somewhat more difficult to detect a change of 1 μA in 1 mA. Under reverse-biased conditions relatively little conventional current flows in the pn junction.

As a result of the reverse bias the energy differential across the junction is enhanced, which makes it very difficult for electrons to pass from the n-type to the p-type material because of the large energy barrier. This ensures that there is a very low dark current, and so any photocurrent appears to be relatively large by comparison. However, electrons in the conduction band of the p-type material can easily pass into the n-type material.

When photons with sufficient energy to excite electrons across the band gap ($\hbar\omega > E_g$) are incident on a pn junction, electrons are stimulated into the conduction band. These are then driven into the n-type material by the potential gradient across the reverse-biased junction. The result is a change in conductivity

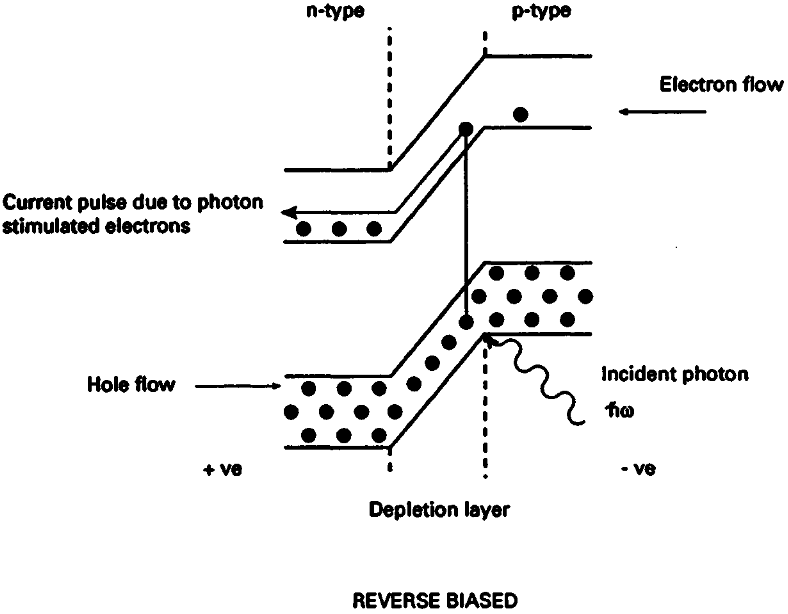


Figure 12.1 Principle of operation of a photodetector.

(increased current for the same applied voltage) and this is used to measure the intensity of the incident light.

12.1.3 Semiconductor light sources

How can a semiconductor be used as a light source?

Conventional light sources emit light as a result of their high temperature in a process known as incandescence. In this case, most of the energy supplied results in heat rather than light, so as a light source the process is relatively inefficient. Luminescence is a different process which is the result of electronic excitation in which electrons at a high energy fall to an available state at a lower energy. This results in the emission of a photon of energy equal to the difference in electron energy before and after the event. Light emitting diodes (LEDs) are used for this purpose.

Semiconductor light sources operate in the inverse way to the semiconductor detectors. Again the device can be a simple pn junction which operates as a diode. It can act as a light source when forward biased. An LED is a form of luminescent lamp. When current flows through the semiconductor, electrons recombine with holes by decaying to a lower energy level and emitting a photon.

Attractive features of light emitting diodes are that they are energy efficient, have a long lifetime, do not overheat and in most cases are relatively inexpensive. Currently, LEDs do not provide enough light for illumination but are often used for information displays or data presentation. In the future it is quite likely that

LEDs will be developed for illumination and in this case, because of their high efficiency and low heating loss, they are likely to become the preferred mode of lighting.

The wavelength of the emitted light is governed by the semiconductor band gap because the initial high-energy electron state is most often at the bottom of the conduction band and the final low-energy electron state is at the top of the valence band. Optical transitions such as this occur with high probability in direct band gap semiconductors but not in indirect band gap semiconductors. Therefore silicon is not a suitable material for LEDs, but gallium arsenide and the family of III-V compound semiconductors are well suited for this application.

12.2 MATERIALS FOR OPTOELECTRONIC DEVICES

Which materials are widely used for optoelectronic devices?

Present-day optoelectronic devices incorporate exotic materials processed with the utmost care, and they function by the manipulation of electrons and holes by electrical and optical means. Both LEDs and laser diodes are used as visual indicators. The market for light emitting diodes in the visible range of the electromagnetic spectrum is currently about 2×10^9 \$ per year. The choice of materials to meet these needs is still quite limited, but is expanding. Absorption and emission of light are related, so the materials with high absorption coefficients in the visible spectrum are the materials of choice for both types of applications.

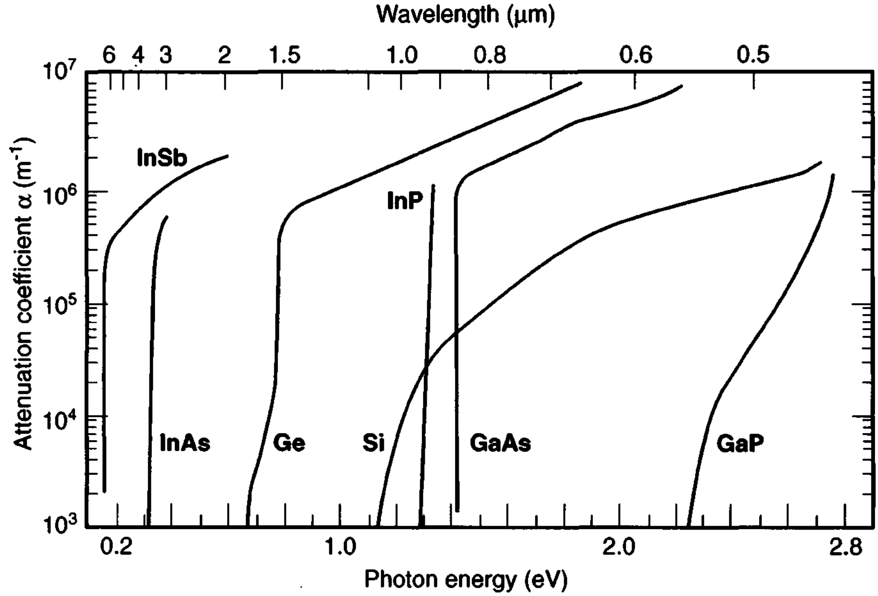


Figure 12.2 Absorption coefficient versus photon energy for Ge, Si, GaAs, and selected other III-V binary semiconductors at $T = 300\,\text{K}$ [2].

The absorption coefficients of several semiconductor materials are shown in Fig. 12.2. These show a very rapid onset of absorption in the direct band gap materials such as GaAs, InP and InAs. The indirect band gap materials such as Si and GaP show a much lower rate of increase of absorption with photon energy. However, Ge, which is known to be an indirect band gap material, shows a quite rapid increase of absorption with energy above 0.8 eV. This is because Ge has a direct transition at 0.86 eV which is at an energy just slightly larger than its band gap. In this respect, Ge behaves more like a direct band gap material above this energy and is a good photodetector for $\lambda \sim 1.3$–$1.5\,\mu$m.

12.2.1 Advances in materials for optoelectronic applications

How are optoelectronic materials being improved and the range of available properties expanded?

Optoelectronic devices are fabricated at present mostly from GaAs or variants such as InGaAs. However, the range of materials options available has expanded as other direct band gap materials have been explored. In particular, materials with larger direct band gaps have been developed and improved, and these now allow green and blue LEDs to be fabricated. Materials such as InGaAsP are already well developed and so these are likely to remain the principal materials of choice for longer wavelength devices in the red-orange-yellow range of the optical spectrum.

Improved energy conversion efficiencies ('quantum efficiencies') have been achieved in recent years with AlGaAs and AlInGaP materials, which now have such improved light emission intensities that they can be used for outdoor displays [3].

Problems with producing green LEDs arose because GaP, which has a suitable band gap, has a quantum efficiency of only 0.1% due to its indirect gap. The alternative material, AlInGaP also has a quantum efficiency of only 0.1%. Gallium nitride GaN and related materials such as AlGaInN are better suited for optical applications because they have direct band gaps and higher quantum efficiencies. AlGaInN in particular forms solid solutions so that a continuous range of band gaps can be obtained by varying the composition.

Improvements in the optoelectronic efficiency of indirect band gap materials, such as silicon, have brought these materials into consideration for optoelectronic devices. Furthermore, there have been a number of advances in polymeric materials in recent years, which now make these organic compounds viable for optoelectronic devices.

12.2.2 Band gaps of semiconductors

How can the colour of light emitting diodes be selected?

Most LEDs have emissions in the infrared (2000–700 nm) or the optical range (700–400 nm). Visible LEDs are used as numeric displays or as small indicator or warning lamps. Infrared LEDs are used as optical isolators or as sources in optical communications systems such as computer communications.

The usual LED materials are aluminum-gallium arsenide $Al_xGa_{1-x}As$, gallium-indium arsenide-phosphide, $Ga_xIn_{1-x}As_yP_{1-y}$, and gallium arsenide-phosphide $GaAs_yP_{1-y}$. The band gap of these materials is dependent on the alloy composition,

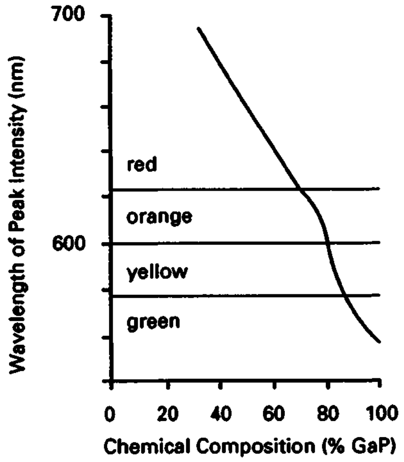


Figure 12.3 Dependence of the wavelength of optical emissions from GaAs/GaP, the first LED material in widespread use.

and can be engineered to a high precision by alloying. Fortunately, the III-V elements form solid solutions which makes alloying simple in these cases and so the band gap is relatively easy to control in this way. This means that the wavelength of the emitted light can be selected by choosing the appropriate alloy composition as shown in Fig. 12.3. Wavelengths throughout the optical range and into the infrared are easily available with these alloys. The III-V alloys still have direct band gaps so they remain optically useful.

In gallium arsenide, up to 30% of the input energy can be converted to light, but this occurs at a wavelength of 900 nm which is in the infrared and so is of no use as an optical display, for which the wavelengths of emissions must be detectable by the human eye. Gallium phosphide gives visible light at about 550 nm, but its efficiency is low because of its indirect band gap. The alloys of these materials have intermediate wavelengths and efficiencies.

12.2.3 Wide band gap semiconductors

What materials can be used for optoelectronic devices at shorter wavelengths?

Wide band gap semiconductors ($E_g > 3$ eV) have been the subject of intense interest because of the possibility of providing low-power solid-state lasers and LEDs at the blue end of the visible spectrum. If emissions at short wavelengths in the blue end of the optical spectrum are required there are other options. Expensive SiC semiconductors can be used. However the latest development is the II-VI family of semiconducting compounds such as ZnS and ZnSe which have band gaps in the range 3.6 eV to 1.5 eV as shown in Fig. 12.4 spanning the entire optical spectrum. These materials and others, such as CdS and SrS, are now being investigated for use as full-colour, thin film, electroluminescent displays [4]. Progress has been made using GaN doped with magnesium to form electroluminesent LEDs ranging from blue to green wavelengths.

12.2.4 Gallium nitride and other group III nitrides

What other III-V semiconductors may be useful?

Materials such as gallium nitride, aluminium nitride and indium nitride, can now be produced with device quality. These materials have direct band gaps in the energy range 1.9–6.2 eV (650–200 nm), and consequently developments in these materials have opened up a number of possibilities for optoelectronic devices in the short wavelength range of the visible spectrum and into the ultraviolet. Some electronic properties of these materials are shown in Table 12.1 [3].

Table 12.1 Properties of GaN, AlN, and InN.

Property	**GaN**	**AlN**	**InN**
Energy gap (eV) (direct)	3.4	6.2	1.89
Lattice constant, *a* (Å)	3.189	3.11	3.54
Lattice constant, *c* (Å)	5.185	4.98	5.70
Density (g/cm^3)	6.1	3.26	6.88
Static dielectric constant	9.5 (8.9)	8.5	19.6
Dynamic dielectric constant	5.3	4.84	9.3
Electron mobility (cm^2/V s)	1000		
Hole mobility (cm^2/V s)	30	14	
Breakdown field (V/cm)	$>5 \times 10^6$		
Saturation velocity (m/s)	2.5×10^5		
Electron effective mass	0.2	0.314	0.11
Light hole mass	0.259	0.471	
Sound velocity (m/s)	5×10^3	10^4	
Optical polar phonon energy (eV)	0.092	0.11	
Deformation potential (eV)	8	19	
K_{th}. (W/cm °C)	1.5	2	
Melting temperature (C)	>1700	3000	≈1100

Gallium nitride emits light in the visible range with a blue-green colour, but again with rather low efficiency. The efficiency of $GaAs_{1-x}P_x$ decreases as the composition x increases. Beyond $x = 0.45$ the band gap becomes indirect and the probability of transition, and therefore the optical efficiency, decreases rapidly. However, the addition of nitrogen in place of phosphorus leads to a more efficient radiative process.

The prospects for using low-power solid-state lasers at the blue end of the visible spectrum ($\lambda \sim 380$–$450\,\text{nm}$) instead of the red end of the spectrum ($\lambda \sim 380$–$780\,\text{nm}$) has opened up the possibility of improved data storage densities using either optical or magneto-optic means because the data storage density increases with the square of the inverse of the wavelength of the light that is used to write data on the disks.

Another area of possible applications of GaN and related materials is in high-temperature, high-power electronic devices. The other material that comes into consideration here is SiC, and more progress has been made to date with SiC than GaN because SiC has been under investigation longer. Nevertheless, the group III-nitrides have excellent properties for such applications and may be expected to be serious candidates for these applications in the near future.

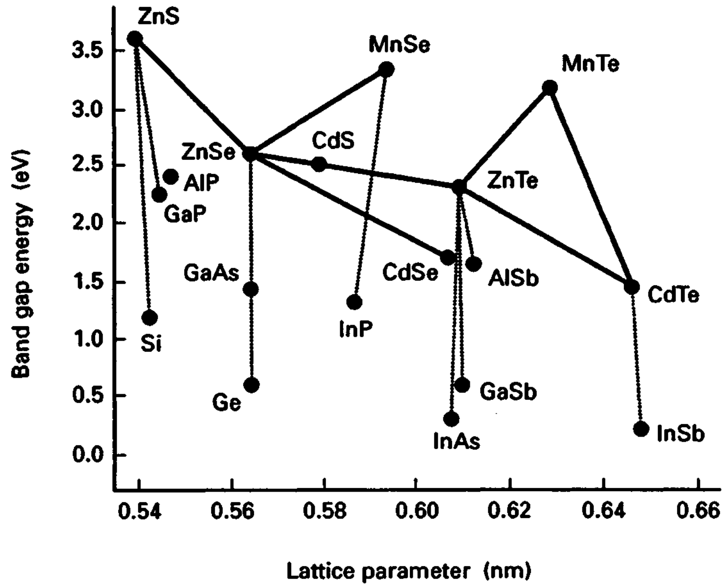


Figure 12.4 Band gap energies of II-VI compound and alloy semiconductors compared to other semiconductors.

12.2.5 Silicon

Is it possible to use silicon in optoelectronic devices?

Silicon is the material of choice for most semiconductor/electronic applications. It has a well-established processing technology that is the result of many years of industrial experience with the material. However, it does have an indirect band gap, which makes its quantum efficiency low for optoelectronic functions.

The demand for high-frequency, long-distance communications indicates the need for integration of optical and electronic functions in which standard microelectronic devices produce electrical signals, which are then converted to optical signals for long-distance transmission, and then reconverted to electrical signals. Consequently, there could be significant advantages derived from improving the optical functions of silicon to allow efficient, silicon-based optoelectronic devices to be constructed. In this respect progress has been made on a variety of silicon or silicon-based materials. These include nanocrystalline silicon, porous silicon, silicon-germanium and rare earth-doped silicon, particularly erbium-doped silicon. The benchmark for acceptable quantum efficiency in weakly optoelectronic materials is 1%. This was achieved in nanocrystalline silicon by Canham in 1990, although this is still an indirect band gap material.

Nanocrystalline silicon is still receiving much attention as a candidate material for optoelectronic applications [5]. The search for improved quantum efficiencies continues, and as a result a whole range of different forms of silicon have been produced with improved light-emitting and absorbing properties [6]. Porous silicon, in the form of a network of nanowires, has a different band structure from normal bulk silicon. This difference is caused by the fact that a high volume fraction of the material is close to the surface, and the electrons in these near-surface regions behave differently [7, 8]. Consequently, a large volume fraction of electrons in the porous silicon behave differently from electrons in the bulk material. As a result, it has been found that optoelectronic properties of silicon are significantly modified [9] when in porous form. In particular, the quantum efficiency of the material for optical functions is greatly enhanced.

In addition, because of the advantages that would arise from being able to integrate optical components with silicon-based devices, efforts have been made to identify direct band gap materials that can be grown directly on silicon. Results that have identified two direct band gap semiconductors, with gaps in the range 0.7–1.0 eV that can be lattice-matched to silicon, have been reported by Zhang *et al.* [10].

12.2.6 Rare-earth-doped silicon

Is it possible to modify silicon to make it more suitable for optoelectronic devices?
Rare earth elements can be used to dope materials to improve their optoelectronic functions for a variety of applications. This includes ultraviolet, visible and infrared lasers and light emitting diodes. In particular, the doping of silicon with erbium has resulted in a material that can be used in electroluminescent devices at $\lambda = 154\,\mu\text{m}$ and this has opened up the prospect of silicon based optoelectronic devices with high quantum efficiencies [11].

The most technologically important of the rare-earth-doped semiconductors is at present erbium-doped silicon which has revolutionized optical fibre communications by allowing 'all optical' amplification of signals being transmitted down the fibre without the need to convert to electronic signals, amplify and then reconvert to optical signals. However, the addition of other rare earth ions can be used to produce a variety of colours in the visible range. Colours that are available now include red (Pr^{3+} and Eu^{3+}), green (Tb^{3+} and Er^{3+}) and blue (Ce^{3+} and Tm^{3+}) [12]. These are being used in flat panel displays, such as plasma displays, electroluminescent displays and field emission displays, where emission in the visible spectrum is essential.

The development of rare-earth-doped solid-state lasers is proceeding very quickly with numerous applications at the shorter wavelength end of the visible spectrum. The best known rare-earth-doped laser material is 'Nd : YAG' (neodymium yttrium aluminium garnet) which has a chemical composition of $Y_3Al_5O_{12}$ together with Nd^{3+} dopants. This produces light with a wavelength of 1064 nm.

12.2.7 Polymeric materials

Can polymers be used as the basis for optoelectronic devices?
The development of polymeric materials with similar optoelectronic properties and functions to those of the conventional semiconductors is a relatively recent

advance. Reports of light emitting devices fabricated from organic materials began to appear in the 1980s. Polymeric optoelectronic materials have attracted attention because of their ease of fabrication in large areas of electroluminescent material with an almost unlimited choice of band gaps, and hence a complete range of colours [13]. These polymeric electroluminescent materials behave very similarly to conventional semiconductor LEDs. Their band gaps can be fairly easily controlled, leading to easy selection of the colour of displays. The optical efficiency of these materials can also be surprisingly high. As an example, polymers have been reported with quantum efficiency greater than 70% [14].

The main interest in these materials at present is in relatively small band gap materials. A summary of much of the important research in the field has been given by Blatchford and Epstein [15]. Among the many possible optoelectronic applications of polymeric materials are as backlighting for flat panel displays, alphanumeric displays (as a replacement for conventional semiconductor LEDs) or high-density information displays. In this last application a layer of electroluminescent polymer is sandwiched between a pair of thin film electrodes. Electrons are injected using a high electric field at the cathode, while holes are injected similarly at the anode. The process then continues in much the same way as in a conventional semiconductor optoelectronic devices. Electrons and holes recombine with emission of photons of a characteristic energy. The polymers usually have low levels of dopants and so behave like extrinsic semiconductors.

12.2.8 Minority carrier injection: injection diodes

How is electron-hole recombination brought about in the light emitting diode?
The process of minority carrier injection is used in light emitting diodes. These are forward-biased pn junctions. It leads to recombination of electrons and holes with the emission of light. The process occurs at the boundary of a pn junction, the depletion layer, if a forward-bias voltage is applied. Electrons are injected from the n-type towards the p-type material by the voltage and holes are driven in the opposite direction. Once in the depletion layer electrons recombine with holes and emit light. Injection diodes are often used as small indicator lamps. Their great advantage is that they have high efficiency, a long lifetime and in most cases are relatively inexpensive.

12.2.9 Recombination process and light emission in LEDs

Where does the electron/hole recombination occur?
The forward biasing of the pn junction, as shown in Fig. 12.5, forces electrons up the potential ramp of the conduction band so that they reach locations in *k*-space where lower-energy states are available in the valence band. An optical transition can then occur from the conduction band into an available state in the valence band. This occurs in the depletion region (space charge region) of the junction. The result is the emission of a photon of energy equal to the band gap. There is also a rapid increase in current with forward-bias voltage.

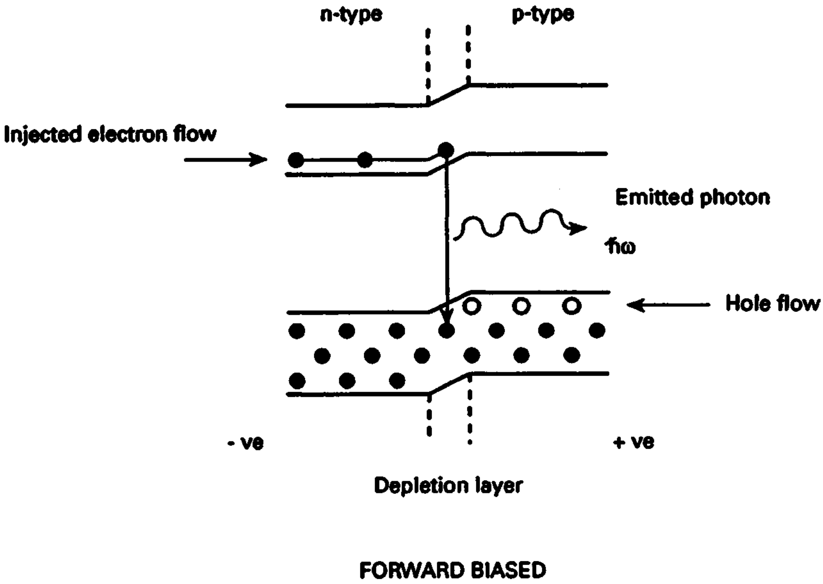


Figure 12.5 Principle of operation of a photoemitter.

The wavelength of the emitted light is determined by the band gap of the material. By engineering the band gap of the material light emitters of different but specific frequencies can be fabricated.

12.2.10 Light detection and generation

What other applications for optoelectronic semiconductors are anticipated?

We have discussed above the operation of phase-coherent light using laser sources. The pn junction described can also be used to detect and generate noncoherent light. It is suggested that these luminescent semiconductor light sources will be the preferred method of illumination in the future, replacing conventional incandescent filament lamps and fluorescent tubes. They will be more energy efficient, will not generate heat and will have virtually limitless lifetimes. These are significant advantages which are now leading to the development of such light sources for the consumer market.

12.3 LASERS

How does a laser semiconductor light source differ from a normal luminescent semiconductor light source?

A laser is a device which emits an intense beam of light composed of photons all of the same wavelength, and all in phase. This is known as a coherent light source. The laser gives an intense energy concentration because the beam divergence is small, as a result of the method of generation of the light in the laser cavity.

There are several different types of lasers:

(i) Semiconductor lasers
(ii) Optically pumped solid-state lasers
(iii) Dye lasers
(iv) Gas discharge lasers
(v) Gas dynamic lasers
(vi) Chemical lasers
(vii) Liquid laser
(viii) Free electron lasers

Of these we will only be concerned here with the semiconductor laser. Semiconductor (junction) lasers are often fabricated from aluminium gallium arsenide or gallium arsenide [16]. When an electrical current is passed through such a device laser light emerges from the junction.

When an electron has been excited, for example into the conduction band of a semiconductor, it must return finally to a lower energy state, either in the valence band or a localized impurity state in the band gap. This occurs with the emission of a light photon. In many cases this emission is spontaneous as in the light emitting diode. Consequently, the radiation is phase incoherent, and therefore this does not constitute laser action. However, in a laser this emission of light occurs by controlled stimulation rather than spontaneously, and this gives phase-coherent photon emission which is characteristic of laser light.

12.3.1 Emission of laser light as a result of electron de-excitation

How are the conditions for stimulated emission created?

In the very simplest case, consider two energy levels, one at the top of the valence band and the other at the bottom of the conduction band, as shown in Fig. 12.6.

In order to get light emission we must have energy stored somewhere in the system waiting to be released. This is achieved in the form of electrons in excited states. In a laser the term 'population inversion' is used to describe the presence of a large number of excited electrons. The lifetime of the electrons in the excited state must also be long enough to ensure that the transition to a lower-energy state does not occur spontaneously, but only under controlled conditions. The presence

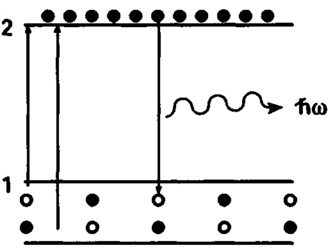


Figure 12.6 Electron energy levels of a two-level laser.

of such a population inversion is a common feature of all lasers since large numbers of electrons need to be maintained in a high-energy state over an extended period of time. This requires rather special conditions which we will discuss.

12.3.2 Production and maintenance of the population inversion

How are large numbers of electrons maintained in a metastable higher-energy state?
Optical pumping is used to produce a large number of electrons in a metastable high-energy state. This involves a high-intensity light source to stimulate electrons into the higher-energy level. The electrons later decay to their original lower level with the emission of light. When the electron reverts to the lower-energy state a photon is emitted with energy equal to the difference in energy of the two states and frequency ω given by

$$\hbar\omega = E_2 - E_1. \tag{12.1}$$

Population inversion is the redistribution of electrons among available energy states by stimulation into metastable higher-energy states. This results in more electrons occupying high-energy states than can be indefinitely sustained. It represents energy stored in the material. Population inversion can be achieved by a number of mechanisms of which 'optical pumping' is only one. Other methods for producing the population inversion include: electrical discharge, chemical reactions, nuclear reactions, electron beam injection and conventional current injection.

The pumping of electrons into excited states is most effective if there is a broad range of energies in the upper energy band. The pumping can then be achieved by a broad range of wavelengths from the light source. This means that the simple two-level laser is not the most efficient configuration for producing laser light from a semiconductor.

If the semiconductor has a range of excited energy levels then the light source can have a continuous range of wavelengths (i.e. can be polychromatic). Electrons are then excited into a range of energy states in the conduction band where they remain until they are caused to de-excite, with the emission of a photon. Typical optical sources are xenon and tungsten iodide filament lamps.

12.3.3 Stimulated emission

How is the stored energy released?
The release of stored energy from the laser occurs when excited electrons drop to a lower-energy state as a result of being stimulated by a photon of the correct energy passing through the material. This leads to the emission of a photon with energy $\hbar\omega$ corresponding to the difference in energy of the electron before and after decay.

When electrons occupy an elevated energy state they can be forced to return to a lower unoccupied state by a kind of resonance involving a photon of the same energy as the transition. The emitted light is phase coherent with the stimulating light. In laser action, the large number of electrons occupying high-energy metastable states can be caused to return to available lower-energy states by an avalanche effect of photons in a chain reaction. This leads to the production of a high-intensity, monochromatic, phase-coherent beam of light which is known as a laser beam.

12.3.4 The emission process

Why is the emitted photon phase coherent with the stimulating photon?

Experimental evidence shows that the passage of a photon through the material can stimulate the transition of other electrons to lower-energy states with the same transition energy through a resonance. The emitted photon is found experimentally to be phase coherent with the stimulating photon. If this process continues, further photons are produced with the same energy and in phase coherence with the first photon. The result is light amplification by stimulated emission of radiation. Laser light is monochromatic because the difference in energy levels, and hence the difference in energy of the electrons before and after emission, are the same throughout the material. Therefore, the photons emitted at any location in the material will have the same energy.

Laser light is also strongly collimated as a result of the method of generation in a long narrow cavity. The semiconductor has mirrored end faces. The laser light is reflected many times between the end mirrors of the cavity leading to a beam with very low divergence. Any photon which is not closely parallel to the axis of the cavity is soon absorbed, leaving only those with low divergence. This is shown in Fig. 12.7.

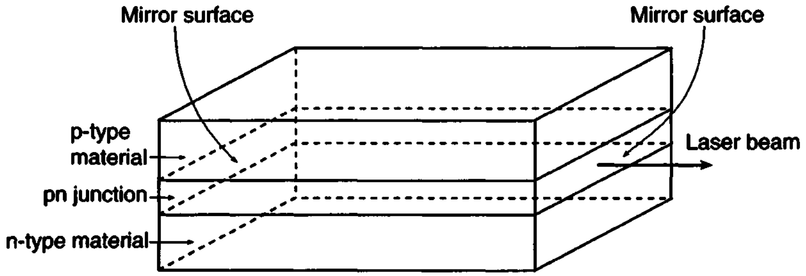


Figure 12.7 Amplification of light photons by stimulated emission of radiation in the cavity of a laser.

12.3.5 Three-level laser

How can the performance of a laser be improved by the use of semiconductor material with different electronic properties?

The need for high-efficiency pumping and for a well-defined transition energy are mutually exclusive in a two-level system. This is because a well-defined optical transition energy requires the upper electron level to be narrow. The simplest way around this problem is to have three electron levels as shown in Fig. 12.8, with a broad upper energy band, with short lifetime for high pumping efficiency, from

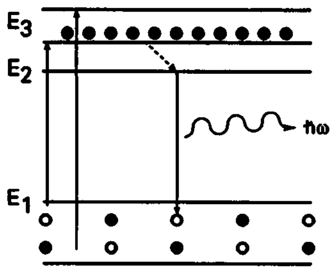


Figure 12.8 Electron energy band diagram for a three-level laser.

which the electrons de-excite into a narrow-energy, long-lifetime level. The long lifetime of electrons in this narrow band allows high population inversion to be created and maintained.

The three-level laser has a number of advantages, one of which is that a range of exciting frequencies can be used to stimulate the electrons initially. This makes the pumping process more efficient, leading to a higher population inversion. In this case, the high-intensity light source has a higher frequency than that which the laser is to emit. Electrons are stimulated into the higher-energy, short-lifetime states of the 'pump band' and then de-excite to the narrow intermediate level. The lifetime of the electrons in this narrow intermediate energy level is much longer, being typically milliseconds, and this allows the population inversion to be maintained. The electrons are then stimulated to de-exite to the lowest level with the emission of a photon.

12.3.6 Four-level laser

Are there further refinements of the electronic structure which can lead to improved performance of the laser?

A further refinement is the four-level laser which has two lower energy levels as shown in Fig. 12.9. This results in emptying of the energy level E_2 into energy level E_1, and hence an even larger population inversion between levels E_2 and E_3 between which the stimulated emission of light occurs.

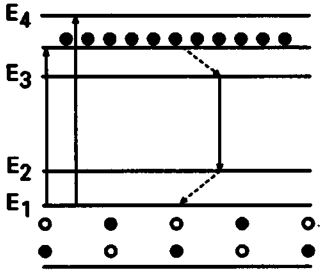


Figure 12.9 Electron energy level diagram for a four-level laser.

12.3.7 Laser materials

What materials are used in solid-state lasers?

A wide range of materials are currently used in the construction of solid-state devices for the generation of laser light. These include

(i) ruby (Al_2O_3, Cr^{3+} doped)
(ii) glass (Nd^{3+} glass or YAG)
(iii) gases (helium, neon)
(iv) vapours (mercury, cadmium)
(v) CO_2
(vi) liquids (dye lasers)
(vii) semiconductors (gallium arsenide and related III-V semiconductors)

12.3.8 Semiconductor lasers (laser diodes)

How does the semiconductor laser work?

The semiconductor laser can be described fairly easily in terms of the electronic properties of materials discussed earlier. The cavity consists of pieces of doped n- and p-type semiconductors in the form of a single pn junction, which we have described above. In such a junction, the electron band structure is distorted under forward biasing, as shown in Fig. 12.5. This gives a light emitting diode, but by arranging that the light emission only occurs when stimulated by photons of the same energy, laser action can be set up.

The stimulated emission only occurs where there is a high population of excited electrons created by 'pumping'. Pumping occurs by direct injection of electrons into the depletion layer where the electrons remain in a metastable state until they are stimulated to return to the valence band by the passage of a photon of the correct energy. The transition only occurs in the depletion layer.

12.3.9 Types of semiconductor junction lasers

What different types of semiconductor laser are there?

When the two parts of the junction are made from the same material then the device is known as a homojunction laser. When the p and n parts of the junction are made from different materials the device is known as a heterojunction laser.

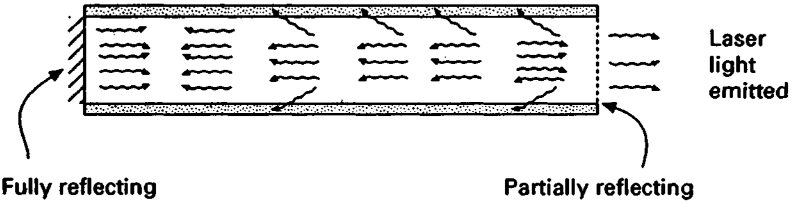


Figure 12.10 Stimulated emission of radiation in the semiconducting material of a laser. The two ends contain parallel mirrored surfaces, and this, combined with the elongated junction region, ensures low beam divergence.

The heterojunction lasers have some advantages if constructed in such a way that the refractive index of the active region is larger than the surrounding regions. This produces an optical waveguide which means that the laser action is confined to the junction layer and no heat is absorbed by the remaining material. The heterojunction laser has higher efficiency because the laser light confined to the junction layer also helps to stimulate more electrons to de-excite.

12.3.10 Applications of semiconductor lasers

Where do semiconductor lasers find specialized applications that differ from the larger high-power lasers?
Helium-neon lasers which have ultra-low beam divergence are used for alignment purposes. Highly focused pulsed laser beams can be used for drilling narrow holes. Optical disks for audio and digital recording applications need to be encoded and read using a laser beam. The laser diode is also used in digital audio and video disk players to read the information encoded on tracks molded onto the disks. Lasers are also used as a cutting tool in laser surgery.

In telecommunications the high frequency of the laser light (10^{14} Hz) enables intensity to be rapidly altered to encode complex signals. The emitted light signal can be modulated at high frequencies by the voltage applied to the semiconductor [17, 18]. This allows the possibility of optical communications at high frequencies and over long distances.

12.4 FIBRE OPTICS AND TELECOMMUNICATIONS

What are the advantages of optical communications over conventional electrical communications?
In many cases optical fibres are used as waveguides to ensure that emission of light from a source reaches the appropriate detector. These optical fibres are transparent materials fabricated with a high length-to-diameter ratio to ensure that light travelling down them is totally internally reflected from the side walls and therefore that very little energy is lost.

Since the emission of light as a result of laser action is dependent on the applied electric field, or on the injected current, it becomes possible with a semiconductor laser to modulate the optical signal through control of the electric field or the injected current. This means that optical communication can be achieved in this way [19], and such communication has distinct advantages over conventional electrical communication in terms of speed and distance over which communication can be maintained.

12.4.1 Fibre optical networks

How do optical fibres compare with conventional metal wires for communications?
In order to satisfy the increasing domain for digital communications over the internet, there has been a rapid increase in the number and size of fibre optic networks over the last ten years. This trend continues to increase resulting in even greater demands for high-frequency, long-distance digital communications. The

best copper wire communications systems have a bandwidth of 10 Mb/s and are limited to typically 200 m at the higher end of the data transfer rate. The attenuation in copper is 2.5 dB/km at 1 MHz and 50 dB/km at 1 GHz. Silicon dioxide optical fibres have a bandwidth of 100 Gb/s and can transmit data over distances of typically 80 km. The wavelength of minimum attenuation in silicon dioxide is $\lambda = 1.55\ \mu m$ when attenuation is only 0.2 dB/km [20].

Optical fibres are made from glass or plastic ('plexiglass' or 'perspex'). These are solid tubes which act as waveguides for light. High quality optical fibres are made from glass as the attenuation is lower than for plastic fibres. The high attenuation of plastic fibres means that they are only used in a limited number of applications, usually for shorter-distance communications. The glass optical fibres are thinner than a human hair and consist of a core region with larger refractive index and an outer cladding region with lower refractive index. The refractive index of the material is changed using dopants such as germanium dioxide, phosphoric oxide or boric oxide. This arrangement confines the light to the core region. Core diameters are typically 1–100 μm in diameter, while the cladding is typically 100–300 μm in diameter.

Optical fibres can be made in single- and multimode form. The single-mode fibre is thinner in diameter with a 1–10-μm core. The refractive index of the outer cladding layer is 0.1–0.3% lower than the core region and there is usually a discontinuous change in refractive index from the core to the cladding. The multimode fibres have a core diameter of typically 40–100 μm and a refractive index change of 0.8–3.0% from the core to the cladding. These multimode fibres can have either discontinuous or graded refractive index change from core to cladding. In these multimode fibres multiple electromagnetic field configurations can propagate down the fibre simultaneously [21].

Despite the performance characteristics of SiO_2, attenuation of the optical signal eventually limits the range of communications frequencies over larger distances so that unless some form of signal amplification is used, the highest data rate is limited by the distance over which the data is being transferred. For example, over the TransAtlantic and TransPacific optical fibre cables, the data transmission rate was limited to 250 Mb/s. However, the discovery of optical signal amplification using erbium-doped silicon dioxide has radically changed the situation, as described in Section 12.2.6. These optical amplifiers can simultaneously amplify signals over a 3-THz frequency range.

The erbium-doped fibre optic amplifier comprises several metres of silica doped with Er^{3+}. The Er^{3+} is excited to a metastable higher-energy state so that a population inversion is created with more ions in the high-energy state than the low-energy state. This region then can provide an active amplification of signals at the appropriate frequency. The principles of the operation of the erbium-doped fibre optic amplifier are very similar to those of the semiconductor laser, except, of course, that the photons only pass through the amplifying region once causing de-excitation of the Er^{3+} ions.

In order to get the erbium ions into a higher state, energy is supplied through the use of an infrared optical pump which is coupled to the erbium-doped region of the optical cable. When they de-excite, the erbium ions emit photons in the 1530–1560 nm wavelength range which corresponds to the range used in silicon dioxide fibres for communications. Amplifier gains on commercial units are typically 20–30 dB, but high-performance research systems have even reached 40 dB [20].

The erbium-doped silica optical amplifier is in high demand at present for digital signal communications over longer distances (>80 km). However, this material is limited to its effective wavelength range of 1530–1560 nm. The need for increased bandwidth, which will allow the number of communication channels to be expanded within a single optical fibre, means that there is a need to identify other rare earth materials that can be used to dope silica and provide efficient optical amplification at other wavelengths [22].

12.4.2 Large scale optical fibre communications

How long is the largest optical fibre network today?

The longest man-made structure today is the 27 300 km fibre optic link around the globe ('FLAG') which provides an optical communications link from the UK to Japan via Europe, the Middle East, India, and China [23]. This comprises what is called the third generation of optical fibre technology. The first two generations were able to operate at data transfer rates of 280 Mb/s and 560 Mb/s, respectively. The current third generation operates at 5.3 Gb/s, almost an order of magnitude improvement over the previous generation.

The cable construction is shown in Fig. 12.11. It consists of two 5.3 Gb/s optical fibre pairs each carrying 60 000 × 64 kb/s. This amounts to a data rate of 3.84 Gb/s, while the remaining 1.46 Gb/s is used for 'overhead' purposes. The optical fibres are SiO_2 glass drawn down to 0.125 mm and protected with a polymer coating. The light signals at the optimum wavelength of 1558 nm are generated by semiconductor lasers.

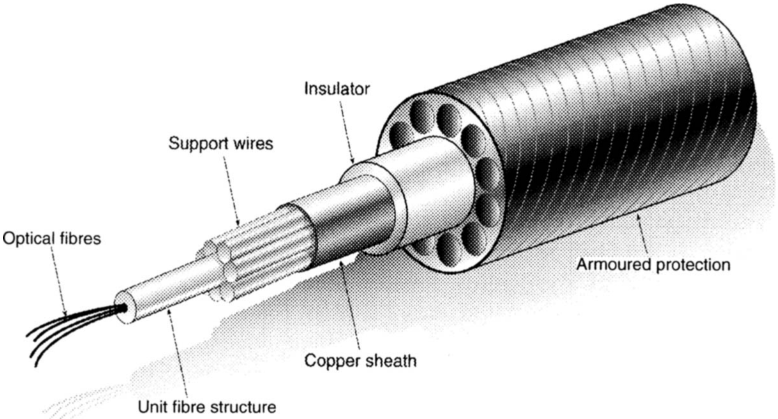


Figure 12.11 The twin pairs of optical fibres in the FLAG system cable can carry 10.6 Gb/s, the equivalent of 120 000 simultaneous telephone conversations. By contrast, only 36 conversations could be handled by the first copper transatlantic telephone cable. After Deniston and Runge [23].

The FLAG system uses 'all-optical' amplification of signals using the Er^{3+}-doped SiO_2 materials described above. These are situated at regular distances of 45–85 km along the length of the cable. The optical amplification is one of the principal reasons that the FLAG system can operate at such high data-transfer rates compared with conventional undersea optical cables. While systems such as FLAG operate at 10 Gb/s, systems with data transfer rates as high as 100 Gb/s are now under consideration as the bandwidth of communications is expanded.

12.5 FLAT PANEL DISPLAYS

How are flat panel displays constructed?

The market for flat panel displays has grown from typically tens of millions of dollars per year in the 1980s to about twenty billion dollars a year today. These flat panel displays come in a variety of different forms, the most important of which are active matrix liquid crystal displays (AMLCDs), electroluminescent displays (ELDs), field emission displays (FEDs) and colour plasma displays (CPDs) [24]. Flat panel displays provide one of the many examples of how a diverse variety of materials and electronic technologies are integrated to produce a device with widespread applications in the display of digital information. The construction and features of some of the principal types of flat panel displays are shown in Fig. 12.12.

The rapid development of flat panel displays has been driven by three main factors: the need to produce more compact, lighter and more energy-efficient means to display information. The primary application has been in portable computers, but there is an increasing demand for large flat panel displays for televisions that can be hung on a wall. The flat panel displays are competing with, and in many cases exceeding, the performance of traditional displays based on the cathode ray tube. Improvements to the flat panel displays continue to be demanded by the needs of various applications. These improvements include increased brightness of displays, increased contrast, improved colour purity, higher resolution, longer lifetime, greater viewing angle, reduced power consumption and reduced cost.

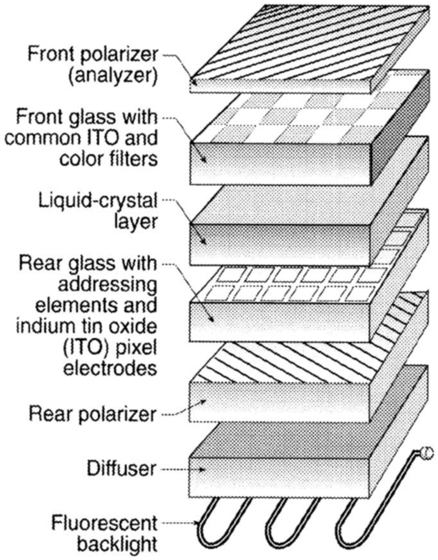
In selecting materials for flat panel displays in the future, the fact that rare earth ions can be used as dopants that emit light in the visible spectrum, including red, green, and blue light, means that they are particularly well suited for applications in displays. In all of the main flat panel display technologies rare-earth-doped materials can provide significant enhancements to the optical functions and performance. The luminescent dopant that is added to the host material can have a narrow emission spectrum (by adding a single type of ion) or a broader spectral range (by adding several different ions). There appears to be no restriction on the number of different types of dopant that can be added to each pixel of host material, allowing several colours or even white light to be emitted.

12.5.1 Liquid crystal displays

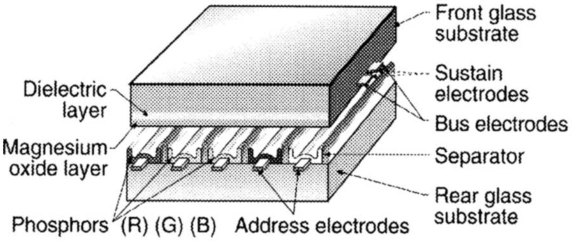
How do liquid crystal displays work?

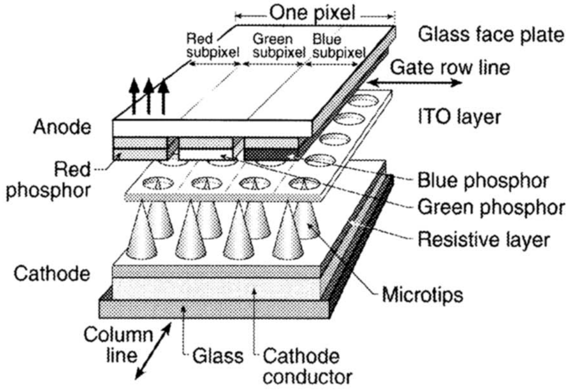
Liquid crystals combine the fluidity of liquids with the orientational (anisotropic) properties of solid crystals. The viscosity of the liquid is simliar to machine oil.

a) Liquid-crystal display



b) Plasma display





c) Field emission display

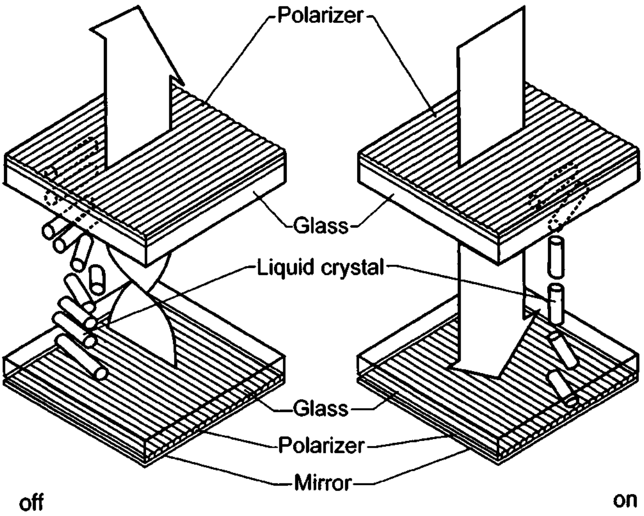


Figure 12.13 Arrangement of a liquid crystal display showing the liquid crystal elongated organic molecules, polarizing filters, mirror and glass substrates.

Nematic liquid crystals, which are often used for displays, are composed of elongated organic molecules which align in a preferred direction [25]. In displays, the molecules of the liquid must be alignable in some way, and this is usually achieved by an electric field. Polarized light is employed through the use of polarized filters in the LCD display screen. The back surface of the LCD display screen has a mirror to reflect incident light. The orientation of the molecules in the liquid crystals changes the direction of polarization of light. This can then be used to produce bright or dark regions on the surface of the display screen.

Liquid crystal displays (LCDs) are widely used as displays and have replaced LED displays in many applications. They are used as flat panel displays for computer screens where they are rapidly replacing the rather bulky cathode ray tubes (CRTs). In addition to their small size one of the other advantages of these displays is their low power consumption. The LCD modifies the ambient light instead of generating its own light as the LED does. Some LCD displays can operate for more than a year on small batteries and so these types of display are favoured for pocket calculators. Another type of display involving the use of ferroelectric materials instead of liquid crystals is also currently under investigation.

Figure 12.12 The three main flat panel display technologies are shown: (a) The liquid crystal display contains polarizers, thin-film transistor plate, and colour matrix filter. (b) The plasma-panel display, in which the phosphor is deposited in the cells as a thick-film paste, (c) The field-emission display in which a high electric field strips electrons from the tips of the cathodes which excite phosphors producing a CRT-like image. After [27] © IEEE 1997.

12.5.2 Active matrix liquid crystal displays

How does an 'active matrix' liquid crystal display work?

The active matrix liquid crystal display is currently the most widely used type of flat panel display. Each pixel has its own switching element for controlling the state of the liquid crystal cell. The switching elements are mostly thin film transistors, and there are three of these per pixel. So for a typical display of 1920 × 1035 pixels there are about 6×10^6 thin film transistors needed to control the display [26].

Since LCDs produce no light of their own, they are used to form an image either by controlling the transmission of light through each pixel (using backlighting) or by controlling the light reflected from each pixel. The best-quality images are produced by backlighting and this is used in laptop computers. For small displays, such as on hand-held calculators, the reflective LCD is still widely used. Ferroelectric liquid crystals have very fast switching speeds for LCDs and these are now finding applications in small high-resolution displays [27].

The main advantages of AMLCDs are that they are light and compact and, therefore, easily portable. They provide high resolution at a high-speed response and have low energy consumption. The main application is in portable computers such as 'laptop' and 'notebook' style computers.

12.5.3 Electroluminescent displays

How does an electroluminescent flat panel display work?

Electroluminescent displays operate using a sandwich structure consisting of two insulators separated by an intermediate layer of phosphor such as ZnS. Electrons are injected from one insulator into the phosphor where they cause luminescence by decaying from their excited state to a lower energy state with the emission of a photon. The construction of electroluminescent displays (ELDs) is simpler than other forms of flat panel display as can be seen in Fig. 12.12. This means that electroluminescent displays are generally more rugged and reliable than the other forms of display. They can provide better performance than other display technologies in certain applications, for example, where a wide viewing angle, high contrast or wide range of operating temperatures are needed. For example, conventional flat panel displays using liquid crystals operate well over the temperature range −25°C to +65°C, whereas electroluminescent displays can operate over the range −60°C to +100°C [28].

12.5.4 Field emission displays

How does a 'field emission display' work?

Field emission displays (FEDs) operate using cathode luminescence which is the same principle that is used in the cathode ray tubes. However, in FEDs each pixel has its own pointed electrode which emits the electrons, which then accelerate across to the anode that is equipped with a phosphor so that light is emitted once the electrons strike it. Essentially, therefore, the FED is an assembly of very small cathode ray tubes, as shown in Fig. 12.12.

The FEDs can be constructed in a thin flat panel which has performance comparable to that of the more widely used liquid crystal displays. The fabrication requires no lithography beyond that needed to define the individual pixels. These features can be produced easily in many cases by using printed circuit board methods. FEDs are now being produced with improved appearance, longer lifetimes, improved stability and lower power consumption compared with a few years ago.

12.5.5 Colour plasma displays

How does a colour plasma display work?

Colour plasma displays (CPDs) now include a full range of colours in the visible spectrum. They provide high levels of brightness and contrast, wide viewing angles and a rapid refresh rate. The principles of operation of the display are the same as those of fluorescent tube lighting that is used for ambient light sources. Recent improvements in depth of colour, efficiency and luminescence have enabled this type of display to be used in televisions. Colour plasma displays can be produced with lifetimes in excess of 10 000 h, which meets the requirements of commercial television screens. Colour plasma displays are therefore ideal for the large wall-hung television screens, and they are currently being used in 42" (1.07 m) display screens. In the future, this type of display is likely to be the first choice technology for large-area displays. CPDs with 60" (1.5 m) diagonals have been produced [29].

12.5.6 Ultrafast electrochromic cells

What are the most recent developments in flat panel displays?

The speed at which flat panel displays can change colour has been a limitation on their use until now [30]. Recently, a new type of display, the electrochromic cell, has been reported that overcomes many of the limitations [31]. In this device two metal films form the positive and negative electrodes. The negative electrode is coated with titanium dioxide and then with methyl viologen molecules, which form a blue colour when electrons are added. The positive electrode is coated with tin oxide doped with antimony and then with phosphonated phenothiazine molecules, which turn red when electrons are removed. These electrochromic cells can be switched in 250 ms using a potential difference of 1.2 volts.

12.6 OPTICAL DISKS FOR DATA STORAGE

How does a DVD work?

The digital versatile disk, now known universally as the 'DVD', became widely available in 1996. A complete range of DVDs are now available including video, audio, ROM (for software) and RAM (for general data) [32]. The original application for DVDs was in the home video market. This application, therefore, dictated the capacity of 4.7 Gbyte and the playback time of 135 min, at a data rate of 5 Mb/s (0.58 MB/s) [33]. The DVD-ROM functions in much the same way as the CD-ROM whereby pits are embossed on the surface of the disk. These pits are typically five times more dense on the DVD-ROM, which can actually store about seven times as much data as a CD-ROM. However, DVD media are more expensive than CD media and this can be a barrier to use. Therefore reduction in cost is desirable to make them more commercially competitive.

REFERENCES

1. S. R. Forrest (1986) Optical detectors: three contenders, *IEEE Spectrum* 23 (5), 76.
2. G. E. Stillman, V. M. Robbins and N. Tabatabaie (1984) III-V compound semiconductor devices: optical detectors, *IEEE Trans. on Electron Devices* 31, 1643.
3. S. J. Pearton and C. Kuo (1997) *MRS Bull.* 22 (2), 17.
4. J. W. Li, Y. K. Su and M. Yokoyama (1992) The construction and characteristics of multi-colour or full colour ACTFEL display devices, *J. Electron. Mater.* 21, 659.
5. L. Tsybeskov (1998) *MRS Bull.* 23 (4), 33.
6. S. S. Iyer, R. T Collins and L. T. Canham (1992) Light emission from silicon, *MRS Symp. Proc.* 256, Pittsburgh, PA.
7. L. Canham (1990) *Appl. Phys. Lett.* 57, 1046.
8. K. D. Hirshman, L. Tsybekov, S. P. Dattagupta and P. M. Fauchet (1996) *Nature* 384, 338.
9. P. Bell (2001) *Nature* 409, 974.
10. P. Zhang, V. Crespi, E. Chang, S. G. Louie and M. L. Cohen (2001) *Nature* 409, 69.
11. S. Coffa, A. Polman, and R. N. Schwartz (1996) Rare earth doped semiconductors, *MRS Symp. Proc.* 422, Pittsburgh, PA.
12. A. J. Steckl and J. M. Zavada (1999) *MRS Bull.* 24 (9), 16.
13. Y. Yang (1997) *MRS Bull.* 22 (6), 31.
14. Q. Pei and Y. Yang (1996) *J. Am. Chem. Soc.* 118, 7416.
15. J. W. Blatchford and A. J. Epstein (1996) *Am. J. Phys.* 64 (2), 120.
16. W. D. Goodhue *et al.* (1990) Monolithic two-dimensional GaAsAlGaAs laser arrays fabricated by chlorine ion beam assisted micro machining, *J. Electron. Mater.* 19, 463.
17. T. E. Bell (1986) Optical computing: a field in flux, *IEEE Spectrum* 23 (8), 38.
18. S. Kobayashi (1984) Semiconductor optical amplifiers, *IEEE Spectrum* 21 (5), 26.
19. F. Guterl and G. Zorpette (1985) Fiber optics: poised to displace satellites, *IEEE Spectrum* 22 (8), 30.
20. A. E. Wilner (1997) *IEEE Spectrum* 34 (4), 32.
21. T. Kimura (1988) Factors affecting fibre-optic transmission quality, *J. Lightwave Technol.* 6, 611.
22. M. Dejneka and B. Samson (1999) *MRS Bull.* 24 (9), 39.
23. F. J. Denniston and P. K. Runge (1995) *IEEE Spectrum* 32 (10), 24.
24. J. S. Im and A. Chiang (1996) *MRS Bull.* 21 (3), 27.
25. F. J. Kahn (1982) The molecular dynamics of liquid crystal devices, *Physics Today* 35, 66.
26. J. Hanna and I. Shimizu (1996) *MRS Bull.* 21 (3), 35.
27. K. Werner (1997) *IEEE Spectrum* 34 (5), 40.
28. P. D. Rack, A. Naman, P. H. Holloway, S. S. Sun and R. T. Tuenge (1996) *MRS Bull.* 21 (3), 49.
29. L. F. Weber and J. D. Birk (1996) *MRS Bull.* 21 (3), 65.
30. M. Gratzel (2001) *Nature* 409, 575.
31. D. Cummins (2000) *J. Phys. Chem.* B104, 11 449.
32. A. E. Bell (1996) *Scientific American* 275, 42.
33. A. E. Bell (1999) *IEEE Spectrum* 36 (10), 28.

FURTHER READING

T. Inagaki (1989) Hologram lenses lead to compact scanners, *IEEE Spectrum* 26 (3), 39.

R. Gunshor, A. Nurmikko and M. Kobayashi (1992) *Physics World* 5 (3), 46.

E. Desurvire (1994) *Erbium Doped Fibre Amplifiers: Principles and Applications*, John Wiley, New York.

I. Kaminow and T. Koch (1999) *Optical Fiber Communications III*, Academic Press, New York.

H. S. Nalwa (2000) *Handbook of Advanced Electronic and Photonic Materials and Devices*, Academic Press, San Diego.

P. Zhang, V. Crespi, E. Chang, S. G. Louie and M. L. Cohen (2001) *Nature* 409, 69.

13 SUPERCONDUCTIVITY AND SUPERCONDUCTING MATERIALS

OBJECTIVE

The main objective of this chapter is to give an overview of superconductivity which includes a description of the basic observations of the phenomenon and an indication of the principal applications. We discuss the emergence of superconductivity in certain materials and how these materials are used in four main groups of applications: superconducting solenoids, superconducting magnetometers (SQUIDs), superconducting logic devices and superconducting power electronics devices. Both flux pinning by a superconductor and the Meissner effect are explained, together with the differences between Type I and Type II superconductors. The onset of superconductivity is discussed as a discontinuous reduction in conductivity to a state with zero dc resistance. It is shown that the resistanceless state is insufficient to explain the Meissner effect in which magnetic flux is completely excluded from the bulk of a superconducting material. Conditions for establishing the presence of superconductivity are given.

13.1 QUANTUM EFFECTS IN ELECTRICAL CONDUCTIVITY

Which are the areas in which quantum effects radically alter the electronic properties of materials on the bulk scale?

By quantum electronics we mean any electronic behaviour where quantum effects make their presence felt. Of course, we can argue that all electronics is quantum electronics. So here we are interested only in those effects where the quantum description is very different from the classical description. Superconductivity, the apparently complete loss of resistivity in some materials at low temperatures, is a quantum phenomenon with no adequate explanation on the classical scale. It results in the passage of an electric current in a material without a potential difference to drive it. We shall consider how this can be used for particular functions. In fact, there are several areas of use: superconducting magnets, superconducting detectors (SQUIDs), and superconducting devices for electronics. Other applications include superconducting light detectors and superconducting devices for generation storage and transmission of electrical energy.

13.1.1 Reduction in resistance on cooling

What happens to the resistance of a material when it goes superconducting?

Superconductivity was first discovered in mercury by Kamerlingh Onnes. The metal exhibited a very rapid reduction in resistivity at 4.2 K, as shown in Fig. 13.1. The fact that the resistivity is very low (typically $< 10^{-25}\,\Omega\,\mathrm{m}$) is a necessary but insufficient condition for the existence of superconductivity. We, of course, expect a

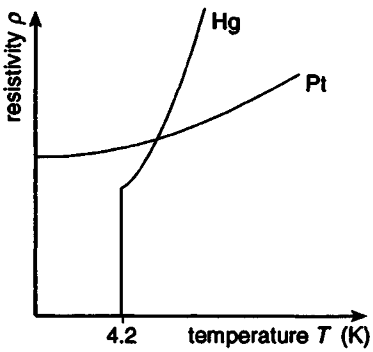


Figure 13.1 Variation of resistance with temperature in mercury and platinum. The superconducting transition in mercury begins at about 4.26 K as the material is cooled. In platinum, which exhibits conventional behaviour, the contributions to resistivity include impurity or defect scattering which is represented by the resistivity at $T = 0$, and phonon scattering which is temperature dependent and represented by $\rho(T) - \rho(0)$.

reduction of resistivity as a result of the reduction of lattice vibrations which in turn reduces the amount of phonon scattering of electrons. However, impurity scattering should continue even at the lowest temperatures and so a residual resistance to the motion of electrons is expected on the basis of classical theory. The reduction of resistivity observed in normal conductors is a continuous process, whereas in superconducting materials there is a sharp phase transition between the normal and superconducting states. Superconductivity therefore involves something else.

13.1.2 Superconductivity

How do we recognize the existence of superconducitivity?

The two essential criteria for determining the existence of superconductivity are,

(i) Complete disappearance of resistivity below a critical temperature T_c.
(ii) Exclusion of magnetic flux from the body of a superconductor – the Meissner effect.

The first condition means that an electric current continues to flow even in the absence of an applied voltage. A further observation is that there exists a critical applied field H_c above which the material is driven 'normal.' This critical field is temperature dependent and becomes zero at the critical temperature T_c.

13.1.3 Nature of the superconducting transition

What are the principal changes in material properties that are observed as a result of superconductivity?

The onset of superconductivity leads to some quite dramatic changes in properties. The following list indicates some of the major observations:

- There is a discontinuous reduction in resistivity.
- In the superconducting state there is a different conduction mechanism.

- AC (eddy current) losses remain, even though there are no dc losses
- Resistivity is more than 13 orders of magnitude lower than the best high-purity annealed copper, $10^{-25}\,\Omega\,\mathrm{m}$ in a superconductor compared with $10^{-12}\,\Omega\,\mathrm{m}$ in high-purity copper and $10^{-8}\,\Omega\,\mathrm{m}$ in normal copper used for conducting wires.
- There is a critical temperature T_c above which the material reverts to normal conducting behaviour.
- There is a critical field H_c above which the material reverts to normal conducting behaviour.

The critical field strength and the critical temperature are related by an equation of the form

$$H_c = H_0(1 - T^2/T_c^2), \tag{13.1}$$

where H_0 is the critical field strength at 0 K.

Below the transition temperature the material has zero resistance unless the current passing through it becomes too large (i.e. reaches the critical current density J_c) in which case the coupling between the electrons is destroyed and the material returns to its normal resistive state.

Table 13.1 Critical temperatures and critical fields for superconducting elements.

Material	Critical temperature (K)	Critical field (kA/m)
Aluminium	1.2	7.9
Cadmium	0.5	2.4
Indium	3.4	22
Lead	7.2	64
Mercury	4.2	33
Niobium	9.3	Type II
Tantalum	4.5	66
Tin	3.7	24
Zinc	0.9	4.2
Zirconium	0.8	3.7

13.2 THEORIES OF SUPERCONDUCTIVITY

How can superconductivity be explained?

Over the years there have been three main theories of superconductivity, the two-fluid model of Gorter and Casimir, the electrodynamic theory of London and London and the paired electron theory of Bardeen *et al.* [1]. Here we will discuss only explanations arising from the last of these.

In normal metals individual electrons are scattered by impurities and by phonons. This causes resistance in the material to the passage of electrons which results in macroscopic resistivity, as described in Chapter 2. Under certain conditions, however, the electrons at energies near the Fermi level, which are the electrons which contribute to the conductivity, can couple together. These electrons then move throughout the solid as a coherent group without scattering.

Under these conditions the phonons and impurities are too weak to scatter the electrons and so the electric current once started moves through the material without experiencing resistance. In the Bardeen–Cooper–Schrieffer (BCS) theory the electrons are paired with opposite spins and opposite wave vectors which results in no net spin and no net momentum. These 'Cooper pairs' are not scattered by the normal mechanisms because their combined wave vector k is zero, corresponding to an infinite wavelength.

13.2.1 Conditions for superconductivity to occur

Why does superconductivity not occur in all materials?

In order for superconductivity to arise there should not be too many normal conduction electrons around to conduct the current, since unpaired 'normal' electrons will get in the way of the paired superconducting electrons. In addition, the thermal energy must be so low that superconducting electrons do not get decoupled. Therefore superconductivity occurs at low temperatures in materials with a low density of states at the Fermi level.

If the temperature is raised, lattice vibrations begin to disrupt the interactions between the coupled electrons until, at a critical temperature T_c, the lattice vibrations completely destroy the coupling. The superconducting electrons then revert to their normal state. The thermal energy needed to decouple the electrons is determined by the superconducting energy gap. For an energy gap of 10^{-3} eV this gives a critical temperature of a few Kelvin, which is typical of many superconducting materials as shown in Table 13.1.

Therefore, it can be seen that the superconducting state will be disrupted and cease to exist at higher temperatures, with the value of the transition temperature being different for different materials depending on the strength of the coupling between the superconducitng electrons. Furthermore, since strong interaction between conduction electrons and the lattice is a prerequisite for superconductivity, it follows that metals such as copper, silver and gold, which are good conventional electrical conductors (and therefore have weak interactions between the conduction electrons and the lattice), do not exhibit superconductivity.

13.2.2 The Meissner effect

How does a superconductor respond to an external field?

When a magnetic field is applied to a superconductor it has the effect of destroying the superconductivity above the critical field H_c. In this case the material will again revert to its normal resistive state. At lower field strengths the superconducting material may completely exclude the magnetic field, a phenomenon known as the Meissner effect. This is shown schematically in Fig. 13.2.

In this case the superconducting electrons set up circulating currents at the surface of the material which counteract the applied field and so cancel the field exactly to zero within the material. In some superconductors, known as Type II superconductors, there also exists a state at intermediate field strengths in which the field penetrates local regions of the superconductor. This field penetration is

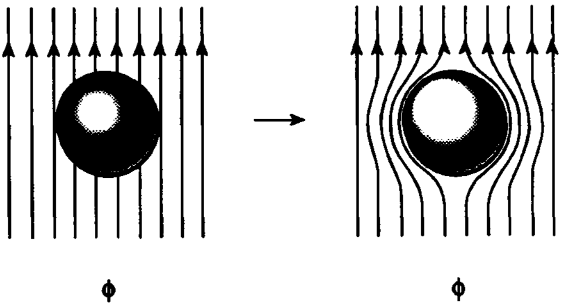


Figure 13.2 Diagram showing flux exclusion from the superconducting material once it has undergone a superconducting transition.

through flux tubes of normal material embedded in a matrix of superconducting material. This state is known as the 'mixed' or 'vortex' state.

13.2.3 Type I and type II superconductors

Is the transition to the superconducting state continuous or discontinuous when a magnetic field is applied ?

Superconductors fall into two categories based on the mechanism of the transition from superconducting to normal state in the presence of a magnetic field. These are known as type I and type II superconductors.

Type I superconductors exhibit only two phases: normal and superconducting. The transition between these phases is very sharp and occurs at a particular critical field strength H_c. Most pure metals are type I superconductors.

Type II superconductors exhibit an intermediate phase known as the 'mixed' or 'vortex' state. This state exists at field strengths between H_{c1}, the critical field for transition to the superconducting state, and H_{c2}, the critical field for transition to the normal state. In this intermediate state the material consists of a honeycomb structure of normal material within a superconducting matrix. Lines of magnetic flux that cannot penetrate the superconducting material because of the Meissner

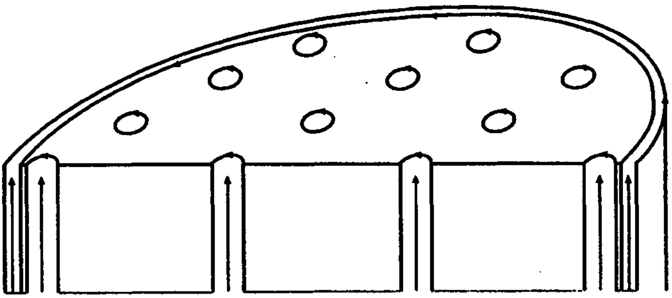


Figure 13.3 Flux lines in a superconductor in its mixed state. The tubular regions are normal material with magnetic flux penetration. The remaining matrix is superconducting material.

effect, can enter through the normal material. Each 'vortex' of normal material carries one flux quantum of 2.07×10^{-15} Wb.

Most alloys are type II superconductors. The matrix of flux vortices in the material of a type II superconductor in its intermediate mixed state is shown in Fig. 13.3. The tubular regions are where the magnetic flux quanta emerge perpendicular to the surface through the normal material. The other regions are the superconducting material.

13.2.4 Flux pinning and flux exclusion

Is a superconductor merely a perfect conductor or is there a greater significance to the Meissner effect?

For some years after the discovery of superconductors, it was assumed that the behaviour of such a material in the presence of a magnetic field would be as shown in Fig. 13.4. That is to say, the supercurrents would prevent any change in the flux ϕ passing through the material because by the Faraday–Lenz law of electromagnetic induction they would set up an induced current. In a perfect conductor

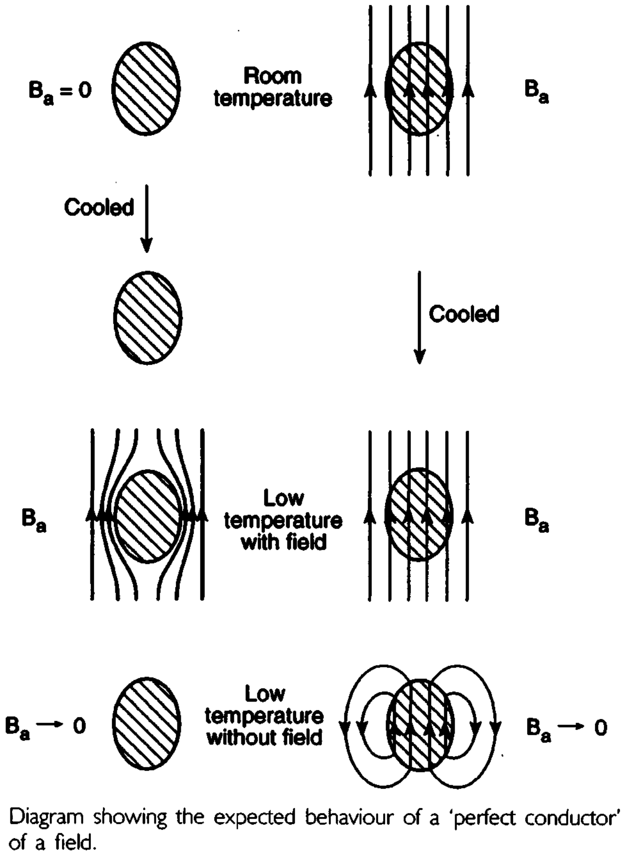


Figure 13.4 Diagram showing the expected behaviour of a 'perfect conductor' in the presence of a field.

this induced current should produce an opposing flux change which exactly counteracts the flux change producing it. In fact, the situation is quite different inside a superconductor in its superconducting state. The magnetic flux ϕ is zero, except for a thin boundary layer at the surface. (We must note immediately that this condition is not exactly fulfilled in the mixed state because of the prescence of normal material.) The Meissner effect does, however, demonstrate that the superconducting state is something more than just a state with perfect conductivity, since the exclusion of flux is an additional property that a merely resistanceless material would not possess.

13.2.5 Surface currents and the Meissner effect

How can we explain the Meissner effect?

When a superconductor is cooled in a magnetic field persistent currents arise on the surface of the material at the critical temperature and these circulate so as to exactly cancel the flux density inside. The surface supercurrents are determined only by the strength of the external prevailing magnetic field.

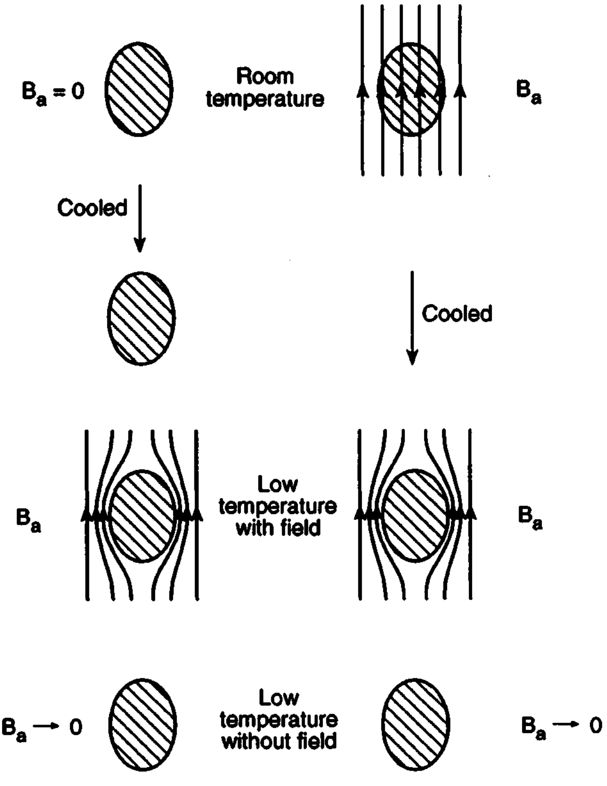


Figure 13.5 Equivalent behaviour of a superconductor (compare with Fig. 13.4) which exhibits the Meissner effect or flux exclusion.

The emergence of the surface currents when a material is cooled through its superconducting transition lies beyond the concept of 'perfect' conductivity. In order not to get an infinite current density, these surface currents must exist over a finite depth. In fact, the surface currents decay exponentially with depth, and this means that the magnetic field does penetrate at the surface of the superconductor to some extent. This is expressed by the penetration depth λ. Typical penetration depths in superconductors are 10^{-8} m. Some values are shown in Table 13.2.

Table 13.2 Penetration depths in superconductors.

Material	**Penetration depth λ (10^{-8} m)**
Aluminium	1.6
Cadmium	11
Lead	3.7
Niobium	3.9
Tin	3.4

13.2.6 Flux pinning in a superconducting circuit

What happens when a magnetic flux passes through the middle of a circuit of superconducting material?

We know that the Meissner effect leads to the exclusion of magnetic flux passing through the body of a superconducting material. A related phenomenon is flux trapping through a superconducting circuit, as shown in Fig. 13.6.

If we have a toroid of superconductor with a flux ϕ passing through the circuit as shown, then the flux passing through the circuit cannot change. The reason lies again in the Faraday–Lenz law of electromagnetic induction which states that any change in flux linking a circuit $d\phi/dt$ sets up a counteracting current in the circuit. In the case of a superconductor, this current produces a flux which exactly

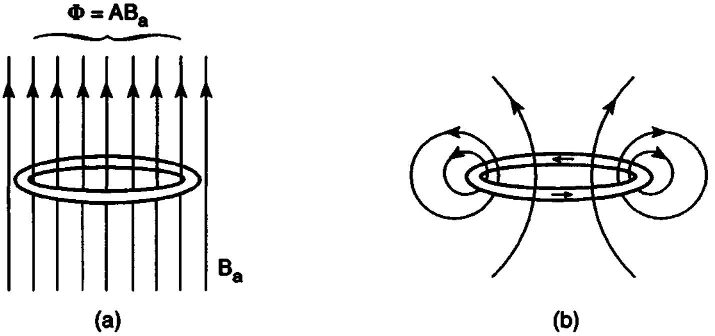


Figure 13.6 Flux trapping by a closed circuit of superconductor: (a) flux driven through a ring of normal material by an applied field; (b) flux trapped by the ring when the ring goes superconducting and the applied magnetic field is subsequently removed.

counteracts the flux change producing it. This is the principle behind the SQUID (superconducting quantum interference device), which is used for measuring changes in magnetic flux to extremely high resolution.

13.3 HIGH-TEMPERATURE SUPERCONDUCTORS

How have the ceramic superconductors altered the perspective for applications of superconducting materials?

General awareness of superconductors was raised by the discovery of materials which are superconducting at temperatures above 77 K. Until early 1986, the highest known critical temperature for superconductors was 23.2 K in NbGe, and, in fact, over the previous 75 years the critical temperatures of superconductors had been raised only very gradually.

Rapid developments in superconductivity began with Bednorz and Muller [2] who discovered that La-Ba-Cu-O is superconducting at 30 K. Within a year an alloy of Y-Ba-Cu-O had been found with a T_c of 95 K and within two years an alloy of Th-Ba-Ca-Cu-O with T_c of 125 K had been discovered. These discoveries were of great scientific interest because they have raised the possibility of finding a room-temperature superconductor. Also, there has been renewed interest in research in two areas: (i) development of new superconductors and (ii) theoretical explanation of the conduction mechanism in these superconductors.

The ceramic superconductors have naturally received much attention in recent years because their high critical temperatures offer a number of possibilities for applications in practical devices. Much of the attention has been directed towards two of these superconductor materials, specifically Y-Ba-Cu-O ('YBCO') and Bi-Sr-Ca-Cu-O ('BSCCO'). The former is usually studied at the chemical composition $YBa_2Cu_3O_7$ and is of interest because of its high critical current density. The latter has been studied in a variety of different chemical compositions and is of interest because it can be fabricated in the form of long wires and tapes.

13.3.1 Critical current density and critical field strength

What range of critical current densities and critical field strengths have been found?

In ceramic superconductors the critical field strength is generally extremely high and has been estimated in one case at $H_c = 3 \times 10^6$ Oe ($0.24 \times 10^9\,\mathrm{A\,m^{-1}}$). The critical current densities J_c of several 'high-temperature' ceramic superconductors are shown as a function of temperature in Fig. 13.7.

The critical current density J_c for YBCO is 40 kA/cm^2 at 77 K compared with 5 kA/cm^2 for $Bi_2Sr_2Ca_2Cu_3O_8$ ('BSCCO-2223') [3]. YBCO can have even higher critical current densities when deposited in thin films on a substrate. For example, when deposited on sapphire (Al_2O_3) using a CeO_2 buffer layer a J_c of 3×10^6 A/cm^2 has been observed at 77 K in a film of thickness 200 nm. This compares with a J_c of 0.2×10^6 A/cm^2 when deposited directly on sapphire [4].

For the production of wires and tapes of superconductor BSCCO-2223 is currently the material of choice. It is formed into wires using the 'powder in tube'

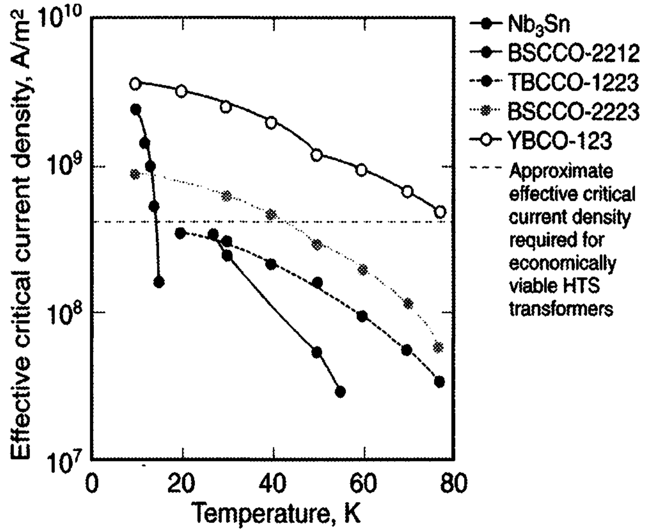


Figure 13.7 Dependence of critical current density on temperature in various ceramic superconducting materials [3].

method whereby it is packed inside tubes of silver, or silver alloy matrix, after which the tube is deformed by drawing and rolling to produce long lengths of wire. Lengths of BSCCO-2223 in excess of 1 km can be produced by this method. Consequently, BSCCO continues to be used for wires; however, the material is more susceptible to the presence of external magnetic fields than other materials such as YBCO, and in this respect, and also because of its lower critical current density, it is presently considered to be less desirable than YBCO for most applications.

13.4 APPLICATIONS OF SUPERCONDUCTORS

What are the major technological applications of these materials?

The major applications of superconductors can be categorized into the following main areas: (i) generation of high magnetic fields using superconducting solenoids, (ii) high-resolution detection of magnetic flux using superconducting quantum interference device magnetometers (SQUIDs), (iii) small low-power electronic devices based mostly on the Josephson effect and (iv) power generation, storage and transmission. The basic science and engineering of superconductivity has been dealt with by Orlando and Delin [5], in which superconducting junctions and devices are discussed. Another useful guide to engineering applications of superconductivity is the work by Doss [6], which although ultimately directed towards ceramic superconductors, nevertheless also contains a detailed general discussion of applications of superconductivity. The discovery of ceramic superconductors with critical temperatures above the boiling point of liquid nitrogen has affected each of the main areas of application.

13.4.1 Superconducting solenoids and magnets

How are superconducting materials used to produce high magnetic fields?

One of the principal applications of superconductors is in the generation of high-intensity, high-stability magnetic fields for both scientific investigations and medical applications such as magnetic resonance imaging (MRI). In this case, superconducting wires need to be fabricated to form the coils of the solenoid. These wires usually consist of niobium-titanium or niobium-tin in a matrix of copper which is extruded into a wire about 0.5 mm in diameter. These multifilament wires are shown in cross section in Fig. 13.8.

In magnetic resonance imaging for medical diagnostics the magnetic moment of the nucleus can be detected through nuclear magnetic resonance. The resonant frequency of a particular nucleus is dependent on its mass and the field strength it is subjected to. The resonance can be caused by using an rf coil with an adjustable excitation frequency and detected by using a pick-up coil.

A field gradient is normally used to determine the spatial locations of the nuclei. The measured resonance frequency indicates the field strength which a particular nucleus experiences. From a three-dimensional map of the field strength over a given volume, the location of the nucleus within that volume can be found. In this application it is essential to have a strong magnetic field which is stable and precisely controllable over the working volume. A high field strength is advantageous because it gives a higher resonant frequency and therefore a stronger signal-to-noise ratio. In this type of instrument the superconducting magnet is therefore a critical component.

Superconducting magnet systems which can generate magnetic flux densities of up to $B = 15\,\mathrm{T}$ ($H = 12 \times 10^6\,\mathrm{A\,m}$, or 150 kOe in free space) have been available for many years. The wires which are used to make the coils of the superconducting solenoid can carry much higher current densities than conventional conductors. For example, in niobium-tin the critical current density, which provides an upper limit to the current density the material can sustain before making a transition to the normal or resistive state, is typically $10\,\mathrm{MA\,cm^{-2}}$ ($10^{11}\,\mathrm{A\,m^{-2}}$).

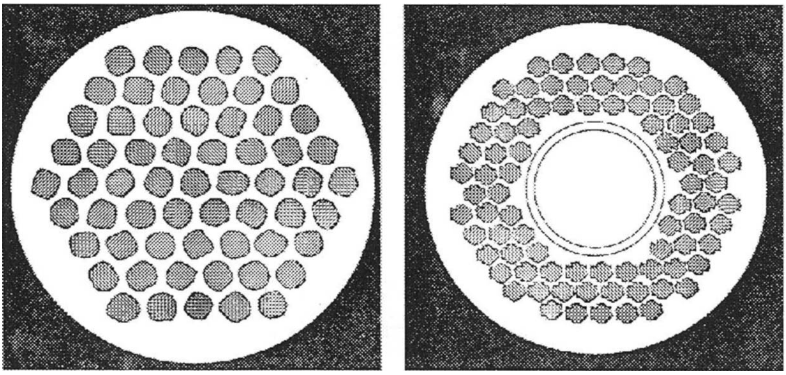


Figure 13.8 Cross section of a superconducting wire consisting of a superconductor in a metal matrix.

13.4.2 Superconducting magnetometers

How can superconducting flux pinning be used to develop high-sensitivity magnetometers?

In a superconductor-insulator-normal metal (SIN) junction, quantum-mechanical tunnelling of the electrons across the energy barrier presented by the insulator can occur provided the insulating layer is sufficiently thin. This can be compared with tunnelling of electrons into a finite potential barrier as discussed in Section 4.3.2. Single-electron tunnelling was first demonstrated by Giaever [7]. Later, Josephson [8] predicted that in a superconductor-insulator-superconductor (SIS) junction, quantum-mechanical tunnelling of Cooper pairs could occur with a higher probability than had been generally recognised before. This effect was demonstrated experimentally by Anderson and Rowell [9].

When two such SIS or Josephson junctions are connected in parallel there is a quantum interference between the electron wavefunctions. This leads to an oscillatory variation in the voltage across the device with magnetic flux linking the circuit. This can be used as a high resolution flux-counting device known as a superconducting quantum interference device (SQUID). SQUID magnetometers can be used to measure changes in flux down to 2.07×10^{-15} Wb. They are now used in medical diagnostics, mineral surveying, submarine detection, motion detection, materials evaluation and scientific measurements [10].

If we consider first a superconducting circuit enclosing a flux ϕ, as shown in Fig. 13.9 then there will be an outer supercurrent i_0 which exactly compensates the external field in order to prevent flux passing through the superconducting material. This outer current does not depend on the field exposure history, it depends only on the external field strength, that is it depends entirely on the prevailing conditions. In order to maintain the flux through the hole in the centre

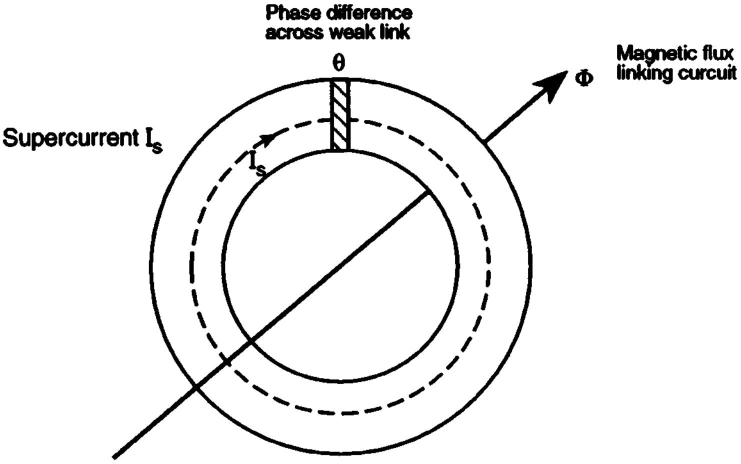


Figure 13.9 Arrangement of a circuit of superconductor with a weak link to form a Josephson junction.

of the toroid there must be an inner current i_i which depends on the field exposure history of the toroid. This current is determined by the amount of flux that was passing through the circuit at the time the material became superconducting.

The relationship between the flux density threading the superconducting circuit and the flux density of the applied field is

$$B = B_a + \frac{LI_s}{A}, \tag{13.2}$$

where B_a is the flux density of the applied field, I_s is the total superconducting current, L is the inductance of the ring, and A is its cross-sectional area. The supercurrent I_s is related to the critical current I_c which is determined by the properties of the weak link:

$$I_s = I_c \sin\theta, \tag{13.3}$$

where θ is the phase difference of the electron wavefunction across the weak link. Therefore, we can write the relationship between the flux densities as

$$B = B_a + \frac{L}{A} I_c \sin\theta. \tag{13.4}$$

If we have a completely superconducting ring, then the flux through the ring must be an integral number of flux quanta

$$\phi = N\phi_0, \tag{13.5}$$

where N is an integer and ϕ_0 is a flux quantum. With the weak link or Josephson junction in the circuit the phase angle θ depends on the flux in the following way,

$$\theta = 2\pi N - 2\pi \frac{\phi}{\phi_0}, \tag{13.6}$$

and since N is an integer we must have

$$\sin\theta = \sin 2\pi\left(N - \frac{\phi}{\phi_0}\right) \tag{13.7}$$

$$= -\sin\left(2\pi \frac{\phi}{\phi_0}\right) \tag{13.8}$$

so that

$$B = B_a - \frac{L}{A} I_c \sin\left(\frac{2\pi\phi}{\phi_0}\right). \tag{13.9}$$

This means that each time ϕ equals an integral multiple of 2.07×10^{-15} Wb, the flux density in the ring is equal to the flux density of the applied field. However, at intermediate values of flux a supercurrent flows in the superconductor. This supercurrent is determined by the flux entering the ring and it can be measured. If a loop of wire is wound on to the superconducting ring, then the voltage induced in the loop is a periodic function of the flux linking the circuit. This can therefore be used to count the changes in flux quanta.

13.4.3 Principles of operation of a SQUID

How does a SQUID count the flux changes?

If we now consider the situation depicted in Fig. 13.10, a current I flows through the two paths of the interferometer device. We assume that the device is symmetric so that $I/2$ flows through each arm of the device.

Assuming $I < 2I_c$ then we will have a phase angle across each weak link as described above. Now suppose a flux ϕ is introduced into the loop. This will cause a superconducting current I_s, which will add to the existing current in one arm but subtract from the current in the other arm. The phase angles across the the two weak links will be $\theta + \delta$ and $\theta - \delta$ so that

$$\tfrac{1}{2} I + I_s = I_c \sin(\theta + \delta), \qquad (13.10)$$

$$\tfrac{1}{2} I - I_s = I_c \sin(\theta - \delta). \qquad (13.11)$$

Summing these currents gives,

$$I = I_c[\sin(\theta + \delta) + \sin(\theta - \delta)] \qquad (13.12)$$

$$= 2I_c \sin\theta \cos\delta, \qquad (13.13)$$

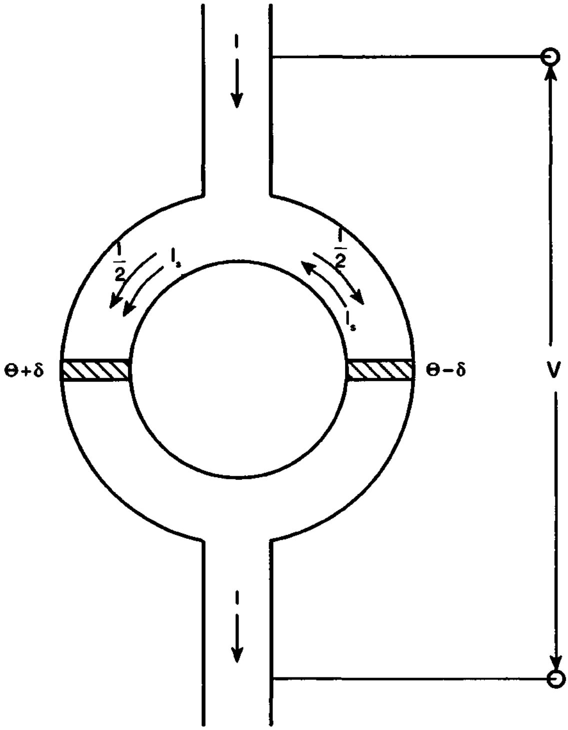


Figure 13.10 Electrical circuit showing connection of two Josephson junctions to form a SQUID.

and the total change of phase across the two weak links is 2δ which must equal $2\pi N - 2\pi\phi/\phi_0$, so that

$$\delta = \pi N - \frac{\pi\phi}{\phi_0} \tag{13.14}$$

where N is an integer as before. Therefore

$$I = 2I_c \cos(\pi N)\cos\left(\frac{\phi\pi}{\phi_0}\right)\sin\theta. \tag{13.15}$$

Therefore as the flux is increased the current varies periodically with flux, and hence so does the voltage V across the device, as shown in Fig. 13.11. The period is one flux quantum ϕ_0, so the device can be used to count flux quanta. Note, however, that in practice the current never changes direction, as the above equation seems to imply. This is because the electron pairs adjust their phase to ensure that the current I is always flowing in the same direction. Note that the device measures changes in magnetic field rather than the absolute value of magnetic field. The device can therefore be used to count the number of flux quanta which enter the ring.

A SQUID magnetometer without Josephson junctions was proposed by Fink *et al.* [11]. In this device the supercurrents in the two parallel branches of the device result in a critical current which varies in an oscillatory way with magnetic flux linking the circuit. Moshchalkov *et al.* [12] have shown that this quantum interference does occur in a superconducting loop of length no greater than the Cooper pair coherence length, which is typically a few micrometres and is much smaller than the length of loop used in a conventional SQUID.

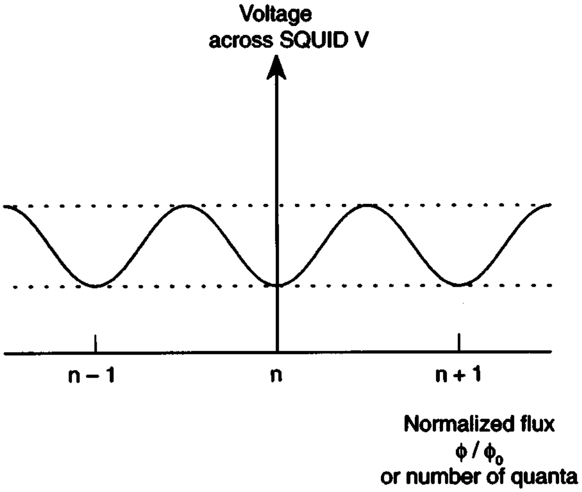


Figure 13.11 Variation of the voltage across a SQUID with the number of flux quanta in the external field attempting to pass through the ring.

This leads to cyclic variation in the voltage across the loop as a function of magnetic flux linking the loop. The mechanism of the interference in these mesoscopic loops is quite distinct from the effect observed in Josephson junction devices. However, there are some similarities in performance. Regions of low density of Cooper pairs, which are analogous to the weak links in the Josephson junction devices, occur periodically in these mesoscopic loops at half-integer flux quanta as a result of the effects of shielding and transport currents in the loop. The phase coherence of electrons in the loop ensures that when the 'pseudo weak link' appears in one branch of the loop it must simultaneously occur in the other branch.

There is a long-term prospect for the development of electronic circuit elements based on the critical current oscillations in these mesoscopic loops. However, there are significant practical problems to be overcome. Specifically, in order to observe the effect temperatures below 1 K are required with voltages of less than 0.1 mV. The response of the device is also very sensitive to changes in sample configuration. For example, the interference pattern can be changed even by moving a single impurity in the device.

13.4.4 Developments and applications in SQUID magnetometers

Where are SQUIDs used and how are they being improved?

SQUIDs provide the most sensitive field detectors in the range 0.1–100 Hz. They are widely used for measuring low-level, low-frequency magnetic fields. Examples of their use can be found in biomagnetism and medical diagnostics, where they are used to detect magnetic field fluctuations produced by the brain and the heart. Other examples of their use are found in nondestructive evaluation where small changes in magnetic fields can be caused by structural changes and degradation in materials [13].

A low-noise SQUID gradiometer for measurement of biomagnetic fields has been developed by Lee *et al.* [14]. This device incorporates a planar thin film pick-up coil integrated onto the same substrate as the SQUID. As a result, it has a flux noise level of 4×10^{-6} flux quanta per $\mathrm{Hz}^{-1/2}$ corresponding to a field gradient noise of $10^{-13}\,\mathrm{T\,m^{-1}\,Hz^{-1/2}}$, which is suitable for the measurement of the magnetic fields generated by neurons in the body.

13.4.5 Superconducting electronic devices

Can superconducting switching effects be used to construct logic devices for computers?

We consider now 'electronic' or low-current applications as distinct from 'electrical' or high-current applications of superconductors. In these applications the electronic characteristics of the superconducting junctions perform traditional electronic functions, although in many cases with improved performance over conventional (i.e. nonsuperconducting) materials. For example, superconducting devices are often faster, or can operate at higher frequencies, or have lower losses than conventional materials.

Along with the change in resistivity of superconductors other equally dramatic changes can occur in the electronic properties. This includes single-electron tunnelling in which electrons tunnel through a thin insulating layer between a superconducting material and a normal material known as an SIN junction. The Josephson effect [8] is a related phenomenon in which a coupled pair of electrons (a Cooper pair) passes from one superconducting region to another through a thin insulating layer known as an SIS junction. This effect can be used to construct devices which can change from one electrical state to another in very short times. This offers the opportunity for constructing logic devices and small computer circuits which operate very rapidly, with switching speeds that can be up to 10- to 100-times faster than conventional silicon devices.

The main electronic device and circuit applications of superconductors are broadly in the following areas: radio frequency and microwave devices (such as filters, resonators and phase shifters); high-speed digital logic devices and circuits; low-noise, high-frequency analogue devices (for example components for high-speed oscilloscopes); thin film devices; hybrid superconductor/semiconductor devices and circuits; and optical detectors. These various applications of superconductors have been discussed in detail by Van Duzer and Turner [15], and by Ruggiero and Rudman [16].

Ralston *et al.* [17] have given a review of developments in low-current device applications of superconductors. In work performed under an industry-university consortium for superconducting electronics, four main areas of interest have been explored: (i) materials and processing, (ii) superconducting junctions, (iii) networks and (iv) circuits.

In materials and processing it was necessary to develop large-area substrates in order to prepare the thin films. This meant that both the lattice parameters and the thermal expansion coefficient of the substrate should match those of the thin film. The main candidate materials are neodymium gallate ($NdGaO_3$) and lanthanum aluminate ($LaAlO_3$). The deposition of the thin films of superconductors onto the substrates can proceed by one of several methods including sputtering, coevaporation in vacuum, laser ablation, molecular beam epitaxy (MBE), and organometallic deposition.

Materials which are receiving much attention for thin film devices are the yttrium-barium-copper oxide material ($YBa_2Cu_3O_x$ or 'YBCO'), thallium-barium-calcium-copper oxide (TlBaCaCuO or 'TBCCO') and bismuth-strontium-calcium-copper oxide (BiSrCaCuO or 'BSCCO'). Another material which is being investigated is barium-potassium bismuthate $(BaK)BiO_3$ which has a very long coherence length, that is the range over which electrons states are correlated. In addition its electronic properties are isotropic. These characteristics make the material a good candidate for devices based on tunnelling effects and junctions, despite its comparatively low critical temperature of 30 K.

A prototype junction device using thin films of ceramic superconductor was developed and this is shown in Fig. 13.12. This is a Josephson junction device made from YBCO thin films which are separated by a layer of praesodymium-barium-copper oxide $PrBa_2Cu_3O_x$. The films were grown with the copper oxide

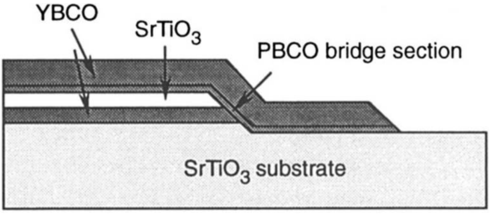


Figure 13.12 Schematic diagram of a thin film Josephson junction using ceramic superconducting YBCO.

planes parallel to the substrate surface, giving the highest critical current density in this direction. The step structure of the films is arranged to ensure that the weak link is also in the direction of maximum coherence length. This geometry therefore optimizes the performance of the device.

In the area of networks, the superconductors can be used to interconnect between semiconductors. The superconducting thin films have low resistivity at microwave frequencies and so low-loss compact microwave filters can be made from them. For example, at 77 K and 4 GHz an YBCO filter has losses which amount to only 25% of an equivalent filter fabricated from silver. The high-performance superconducting microwave filters can have Q values of better than 1000, compared with Q values of 250 in filters fabricated from normal metals operating under similar conditions.

Low-noise SQUIDs can be made using a three-layer process employing niobium–aluminium oxide–niobium films. These are produced by sputtering followed by a planarization process similar to that used in the production of silicon devices (see Section 11.4.3). Devices such as SQUIDs, gradiometers and oscillators can be fabricated with junction areas as low as $0.77\,\mu m \times 0.77\,\mu m$.

Light interacts with materials via the electrons. We have already considered this interaction in semiconductors and normal metals. The electrons in superconductors are sensitive to light with photon energies as low as 0.01 eV and so superconducting materials can be used as photodetectors. These applications have been discussed by Richards and Hu [18].

13.4.6 Applications of superconductors in power electronics

How can superconductors be used in electrical energy conversion and transfer?

One of the main applications for which superconductors seem to hold great promise is in the efficient generation, storage transmission and conversion of electrical energy. However, this prospect seemed far from realization until the discovery of the ceramic 'high-temperature' superconductors. The main benefit that these materials bring is that their critical temperatures are above that of liquid nitrogen, which opens up the possibility of a relatively cheap means of cooling the materials into their superconducting state.

Now that these superconductors can be fabricated into long lengths of wire or tape the prospects for practical power electronics applications seem much nearer.

In principle, these materials can be used to construct more compact and energy-efficient devices for (i) fault current emitters, (ii) electric power generators, (iii) motors, (iv) transformers, (v) power cables, and (vi) energy storage devices. However, there are still significant improvements that need to be made in the materials and a number of technical problems to be overcome before the materials are likely to see widespread implementation in power electronics devices.

Superconducting motors and generators should have typically only half of the power losses of conventional copper wire equivalents [19]. Energy-storage devices include superconducting magnetic energy storage systems ('SMES') and rotating flywheel type devices that can be used to store electrical energy during off-peak periods and then return the energy to the power grid during times of peak demand. These will enable the power grid to become more robust in handling large fluctuations in power demand [20, 21], which has caused some major problems in regions with high energy consumption. Energy-conversion and energy-transmission devices including superconducting transformers, with reduced size and weight and increased capability for over-capacity operation, and superconducting cables with higher current-carrying capabilities and lower power losses have been developed to the prototype stage.

However, despite all of the attractive potential applications, there are still a number of unanswered questions. The superconducting materials suffer from ac losses and the materials themselves are still relatively expensive. This means that their usage involves significant initial capital investment. Also, while reducing the power losses in some devices by half may sound impressive, the economic benefits may not be sufficient to justify implementation if the conventional devices are already above 99% efficient, as is the case with transformers.

13.4.7 High-capacity superconducting wires and cables

How can superconductors be used to produce low loss electrical wires and cables? Transmission of electric power with high efficiency through superconducting wires or cables without resistive losses has obvious attractions if the cost of keeping the materials cool can be made low enough to realize a net improvement in overall energy efficiency. With increased interest in the transmission of power through superconducting wires and cables the nature and magnitude of the ac losses becomes a significant concern for the optimization of device designs.

Progress in producing superconducting wires and tapes has been reasonably good in recent years. BSCCO tapes have been produced with a critical current of 80 A in a cross section of 4 mm × 0.4 mm [22]. Modelling of the ac losses in BSCCO has been demonstrated [23], in which it was shown that the conventional critical state model of Bean [24] was not applicable to this material. The model of Tami *et al.* allowed the field and current profiles inside the tapes to be calculated, and from this the power loss was determined by integrating throughout the material. This showed that for current densities well below the critical current density, the ac losses were mainly hysteretic, whereas when the current density approaches the critical value J_c flux creep and flow dominate the loss mechanism as the material approaches a resistive state.

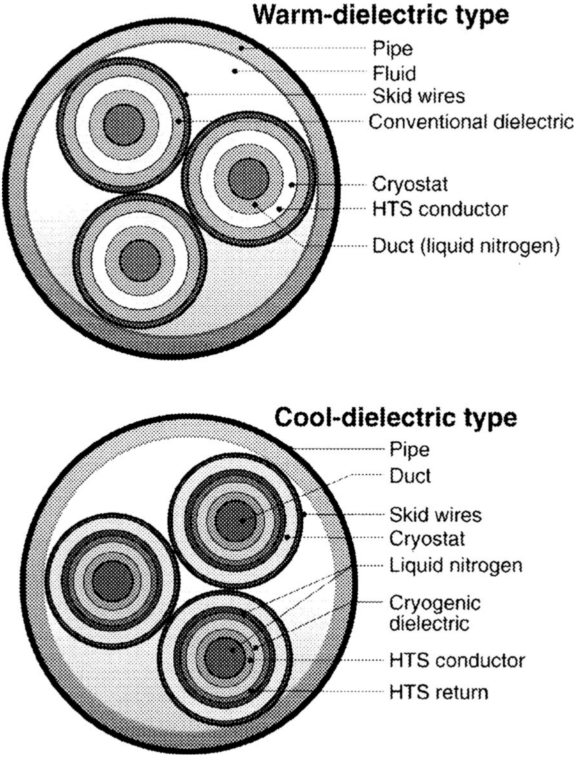


Figure 13.13 Two possible configurations for superconducting cables: (a) the warm dielectric and (b) the cold dielectric [22].

Superconducting cable designs seem to be converging on two possible configurations. These are the so-called 'warm dielectric' and 'cold dielectric' configurations [22] as shown in Fig. 13.13. The warm dielectric configuration can carry more than twice the power of conventional copper cable with the same losses. The cold dielectric configuration can carry four times the power at only two thirds of the losses. The warm dielectric configuration is designed to allow retrofitting into existing tunnels and pipes, thereby minimizing the initial investment costs, whereas the cold dielectric configuration, while more energy efficient, cannot be so easily retrofitted. In either case, the resulting superconducting cable network will need to be completely compatible with existing components and must be at least as reliable as existing cable networks.

13.4.8 Superconducting transformers

How can superconductors be used in energy conversion devices such as transformers?

Current densities in superconductors that are 10–100-times greater than in copper wires, together with zero dc resistance, mean that there are possibilities for dramatic improvements in transformer size, weight and performance. Again, the ability to produce long lengths of superconducting wire was crucial to the application, but now that this has been overcome, prototype superconducting transformers have been built and tested.

Conventional transformers in the 30 MVA range are already typically 99.3 to 99.7% efficient. Even so, the transformer losses amount to about 2×10^9 \$/yr in the US and account for about one quarter of the transmission and distribution power losses. Most of this loss is caused by resistive heating in the windings, as opposed to magnetic losses in the core. So replacement of these with superconductors would appear to be an appropriate way to make further improvements in energy efficiency. Compared with resistive and eddy current losses in conventional transformers the losses in superconducting transformers are small. However, this is not the only factor to be taken into account since power is also needed to keep the windings cool and this reduces the overall efficiency of the superconducting transformer.

Consequently, the prototype superconducting transformers that have been built to date are no more efficient overall than conventional transformers. The 500-kVA unit made by Sumitomo using BSCCO 2223 wires had an efficiency of 99.1% at 77 K and 99.3% at 66 K [3], which is comparable to conventional transformers.

Open-cycle cooling with liquid nitrogen is cheapest but can be inconvenient due to the necessary continuous maintenance. Closed-cycle cooling using cryocoolers may be more acceptable in the long run. In either case the refrigeration technology will need to be developed to ensure rugged, reliable, low maintenance, cost-effective superconducting transformers.

13.4.9 Superconducting motors and generators

Can superconductors be used to produce more efficient motors and generators?

Superconducting motors and generators offer some energy-efficiency improvements over conventional devices. For example, a 3.7-MW motor with conventional windings has an efficiency of about 96%. A motor of the same power rating using superconducting windings has an efficiency of 98%. Of course, the power needed to maintain the windings in their superconducting state also needs to be considered. This would amount to a reduction in efficiency of about 0.1% at 77 K and 0.3% at 30 K resulting in net improvements in energy conversion efficiency of 1.9% and 1.7%, respectively, over a conventional motor [25].

13.4.10 Superconducting devices for energy storage

Can superconductors be used to store electrical energy?

The temporary storage of electric power, over periods of less than 24 h, would allow energy to be accumulated during times when the electric power grid is under-utilized (for example during the night) while allowing utilities to draw on these reserves during times of peak demand. Two classes of superconducting devices have been designed to provide such storage: superconducting flywheels and superconducting magnetic energy storage systems (SMES).

Superconducting bearings, using a combination of permanent magnet and superconductor and based on the Meissner effect, can reduce frictional losses by up to two orders of magnitude compared with the best mechanical bearings. With such superconducting bearings it is possible to reduce idling losses to 0.1% per hour, allowing daily storage efficiencies of about 90% [20]. Designs for superconducting flywheels based on YBCO and storing up to 10 MWh have been proposed. At present the experimental versions of these devices are in the range of 1 kWh.

Energy storage can also be achieved using large superconducting coils. Smaller versions of this, the so-called lSMES (microSMES) have recently become available commercially for ensuring quality control of power variability to sensitive electric loads. Design studies for optimization of the lSMES have recently been reported for a 0.7-MJ version of the device [21].

REFERENCES

1. J. Bardeen, L. N. Cooper and J. R.Schrieffer (1957) *Phys. Rev.* **108**, 1175.
2. J. G. Bednorz and A. Muller (1986) *Z. Physik* **B64**, 189.
3. S. P .Mehta, N. Aversa and M. S. Walker (1997) *IEEE Spectrum* 34 (7), 43.
4. I. S. Kim, H. R. Lim, D. H. Kim and Y. K. Park (1991) *IEEE Trans. Magn.* 35, 4073.
5. T. P. Orlando and K. A. Delin (1991) *Foundations of Applied Superconductivity*, Addison Wesley, Reading.
6. J. D. Doss (1989) *Engineer's Guide to High Temperature Superconductivity*, John Wiley & Son, New York.
7. I. Giaever (1960) *Phys. Rev. Lett.* 5, 147.
8. B. D. Josephson (1962) *Phys. Lett.* 1, 251.
9. P. W. Anderson and J. M. Rowell (1963) *Phys. Rev. Lett.* **10**, 230.
10. J. Clark (1990) in *Superconducting Devices* (S. T. Ruggiero and D. A. Rudman eds), Academic Press, New York.
11. H. J. Fink, A. Lopez and R. Maynard (1982) *Phys. Rev.* **B26**, 5237.
12. V. V. Moschalkov, L. Gielen, M. Dhalle, C. Van Haesendonck and Y. Bruynseraede (1993) *Nature* **361**, 617.
13. J. P. Wikswo (1995) *IEEE Trans. Appl. Supercond.* 5, 74.
14. Y. H. Lee, H. C. Kwon, J. M. Kim, Y. K. Park and J. C. Park (1999) *IEEE Trans. Magn.* 35, 4097.
15. T. Van Duzer and C. W. Turner (1981) *Principles of Superconductive Devices and Circuits*, Elsevier/North Holland, New York.
16. S. T. Ruggiero and D. A. Rudman (eds) (1990) *Superconducting Devices*, Academic Press, New York.
17. R. W. Ralston, M. A. Kastner, W. J. Gallagher and B. Batlogg (1992) *IEEE Spectrum* 29 (8), 50.
18. P. L. Richards and Q. Hu (1989) *Proc. IEEE* 77, 1233.
19. U. Balachandran (1997) *IEEE Spectrum* 34 (7), 18.
20. J. R. Hull (1997) *IEEE Spectrum* 34 (7), 20.
21. K. Ryu, H. J. Kim, K. C. Seong, J. W. Cho, E. Y. Lee, H. B. Jin and K. S. Ryu (1999) *IEEE Trans. Magn.* 35, 4103.
22. M. M. Rahman and M. Nassi (1997) *IEEE Spectrum* 34 (7), 31.

23. A. Tami, A. E. Mahdi and D. J. Mapps (1999) *IEEE Trans. Magn.* 35, 4088.
24. C. P. Bean (1964) *Rev. Mod. Phys.* 36, 31.
25. R. Blaugher (1997) *IEEE Spectrum* 34 (7) 36.

FURTHER READING

T. P. Orlando and K. A. Delin (1991) *Foundations of Applied Superconductivity*, Addison Wesley, Reading.

S. T. Ruggiero and D. A. Rudman (eds) (1990) *Supeconducting Devices*, Academic Press, New York.

T. Van Duzer and C. E. Taylor (eds) (1989) *Proc. IEEE* 77, 1107.

S. Washburn (1993) *Nature* **361**, 587.

T. P. Sheahen (1994) *Introduction to High Temperature Superconductivity*, Plenum, New York.

14 MAGNETIC RECORDING

OBJECTIVE

In this chapter we look at the various magnetic methods available for recording of information, images and sound. The most important of the recording media today are magnetic disks which provide the main method for storing information on computers, and magnetic tapes which are widely used for both audio and video recording. Magnetic recording can be conveniently separated into two groups of related technologies: media and recording heads. 'Media' consists of all technologies concerned with the production and use of magnetic disks and tapes for storing information, while 'heads' covers all technologies concerned with the processes of writing information on media, or reading information from media.

14.1 MAGNETIC RECORDING MEDIA

How is hysteresis in magnetic materials used to record information?

The hysteresis of magnetization versus magnetic field in ferromagnets and ferrimagnets can be used to good effect in magnetic recording. Without hysteresis the magnetic state of the material in zero field would be independent of the field that it had last experienced. However, in hysteretic systems the remanent magnetization acts as a memory of the last field maximum, both in magnitude and direction, experienced by the magnetic material. Therefore data, either in digital form for computers and related devices, or analogue signals as in sound recording, can be stored as magnetic 'imprints' on magnetic media.

Of course, to make this of any practical use it must be possible to store large amounts of data in as small a space as possible. So in the recording industry there is a continual need to increase the recording density of storage media. Currently, magnetic storage densities are increasing at a rate of more than 100% per year or equivalently by a factor of 10 every 3 years. The storage densities are conventionally measured in 'bits per square inch'. The present generation of commercial hard disk drives have storage densities of typically 2×10^{10} bits in^{-2} (30×10^6 bits mm^{-2}).

The information must also be able to be retrieved with a minimum of distortion, that is it must not be easily erased or changed by the exposure to extraneous magnetic fields since the information should be capable of being stored permanently. Nor should it be altered by the reading process since it is usually necessary to reread the data many times without loss of information. Furthermore it should be written and read with minimal power requirements.

The magnetic recording medium must have high saturation magnetization to give as large a signal as possible during the reading process. The coercivity must be sufficient to prevent erasure, but small enough to allow the material to be reused for recording. Coercivities in the range of 20–100 kA m^{-1} are common for magnetic recording tapes and disks, although in the future, in order to continue to improve storage densities, the coercivity of the media will need to increase. Therefore coercivities in the range 200–240 kA m^{-1} (2.5–3 kOe) are anticipated [1].

14.1.1 History and background of magnetic recording

How has magnetic recording developed and what is the state of the art?

Analogue magnetic recording of the human voice was first demonstrated by Poulsen, a Danish engineer. In his device, called the 'telegraphone', acoustic signals were recorded on a ferromagnetic wire using an electromagnet connected to a microphone. However, the reproduction was very weak due to the absence of an amplifier. With the development of amplifiers the signals from the magnetic medium could be recreated more strongly and the sound reproduction was easily audible. However there was also a low signal-to-noise ratio, due to the nonlinear nature of the recording process, which still meant that the quality of the sound was not good. The ac biasing method of recording resulted in much better signal-to-noise ratios because the recorded magnetization could be made linearly dependent on the signal level. Magnetic tape was invented simultaneously in both the USA, using a paper tape coated with dried ferrimagnetic liquid, and in Germany using a tape containing iron powder. Oxide tapes were developed for the commercial market by 3M Corporation and as a result audio recorders became available in 1948 and video recorders in 1956. Digital recording for storage of computer information was developed by IBM and the first magnetic disk drive became available in 1957.

This original hard disk drive, developed by IBM in 1957 and known as the 'RAMAC', had a data storage density of 2×10^3 bits in^{-2} (3 bits mm^{-2}). The increase in data storage densities since then has been continual, so that present storage densities of 20 Gbit in^{-2} represent an increase of a factor of 10^7. This progress in storage densities is shown in Fig. 14.1. The rate of increase in storage densities has accelerated in recent years due to the shift to smaller disks, the use of thin film recording media, and the development of advanced read/write heads with improved signal to noise ratios [2]. Projections for the future are shown in Fig. 14.2 in which it is envisaged that the development of smaller magnetoresistive read heads and improvements in actuation and micromechanics will provide the next stages of evolution, followed eventually by STM ('scanning tunnelling microscope')-like storage.

Since the 1950s there has been a growth in digital magnetic recording for the storage of computer data which, together with the consumer demand for audio analogue magnetic recording, particularly recording of music, form commercially the most important areas of the magnetic recording industry. The industry is currently worth about \$100 billion a year.

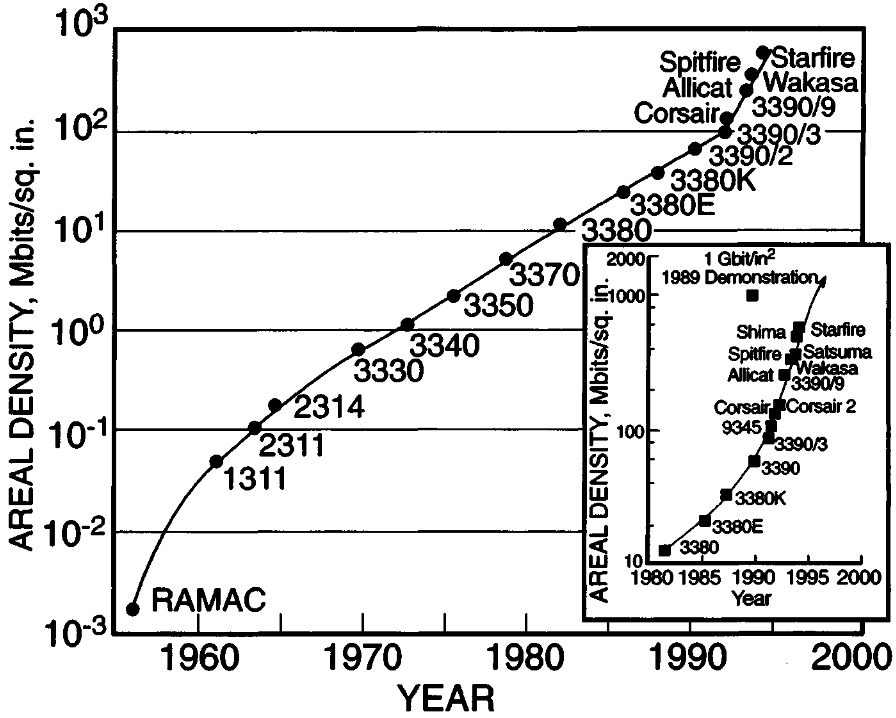


Figure 14.1 Progress in magnetic recording storage densities (after Grochowski and Thompson [2] © IEEE 1994).

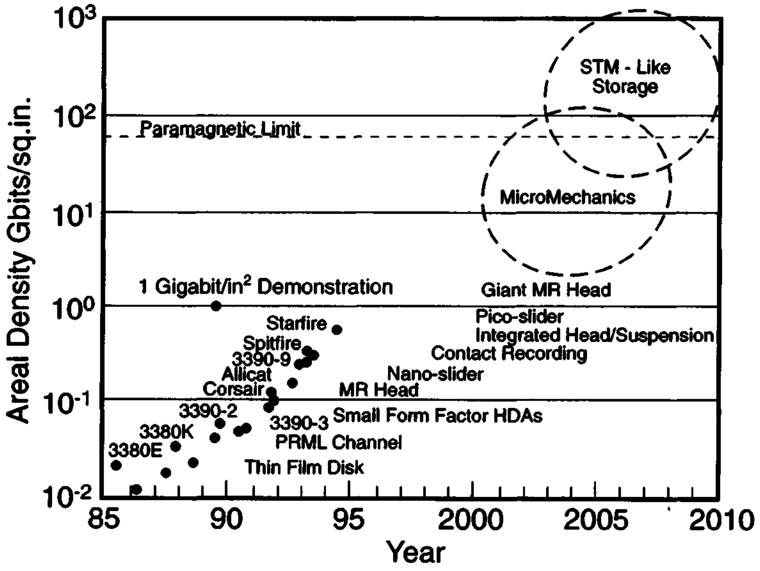


Figure 14.2 Improvements in storage densities with different magnetic recording technologies with projections to the year 2010 (after Grochowski and Thompson [2] © IEEE 1994).

Hard disk drives alone account for over \$50 billion per year of this market. By 2005 the hard disk drive market is expected to reach \$70 billion per year [1].

14.1.2 Magnetic tapes

How are magnetic recording tapes produced?

Magnetic tapes are the most widely used recording medium for audio and video signals. They are produced in two main forms known as particulate and metal evaporated (ME) tapes. The advantage of particulate magnetic tapes is that they can be produced in a wide variety of widths at high coating speeds and at low cost. The disadvantage is that the magnetic particles only occupy 40% of the tape volume. The current trend is toward the use of smaller particles with higher packing densities. The thin film ME tapes have generally better performance characteristics, but are more costly to produce.

Particulate magnetic recording tapes consist of a coating of magnetic material, usually gamma ferric oxide or a cobalt-modified variant of gamma ferric oxide. The coating thickness is 3–6 μm on a flexible, nonmagnetic substrate which is usually PET (polyethylene tetraphthalate). The thickness of the substrate is 10–25 μm, but thinner 5 μm tapes of aramid are used for long-playing tapes, although this is more expensive and has lower demand. The coating consists of the magnetic particles, a binder to contain and disperse the particles, lubricants to ease the motion of the tape and abrasives such as particles of Al_2O_3 to reduce wear of the tape. The lubricants in particulate tape are almost always 'internal' which means that they are included as an integral part of the coating. The magnetic particle sizes vary depending on the choice of the magnetic material, but are acicular (elongated), and are typically 0.25 μm in length and 0.05 μm in width. These are single-domain particles which can be easily magnetized parallel to their long axes. The saturation magnetizations and Curie temperatures of materials used in these tapes are discussed in Section 14.1.4.

Magnetic tapes are magnetically anisotropic. At present the particles are aligned in the plane of the tape as shown in Fig. 14.3. In order to align the single-domain particles the tapes are placed in a magnetic field oriented in the plane of the tape. The field is applied before evaporation of the solvent which carries the magnetic particles leaving the dry binder which carries the magnetic particles. The tape is then heated to completely dry the coating and is rolled or squeezed to densify the coating. Cobalt-modified gamma iron oxide is widely used in magnetic tapes because of the need for increased coercivity. About 2–3 wt% cobalt is impregnated at the surface of the γ-Fe_2O_3 particles, and this results principally in an increase of coercivity from about $30\,\mathrm{kA\,m^{-1}}$ to about $60\,\mathrm{kA\,m^{-1}}$. There is also a slight increase in saturation magnetization and an improvement in the temperature sensitivity of the coercivity. However, it is important that the cobalt is not absorbed into the bulk of the material as this leads to a deterioration in magnetic properties, in particular the coercivity becomes highly sensitive to temperature, which is disadvantageous. The cobalt-surface-modified gamma iron oxide is now widely used in VHS videotapes, some audio tapes and floppy disks.

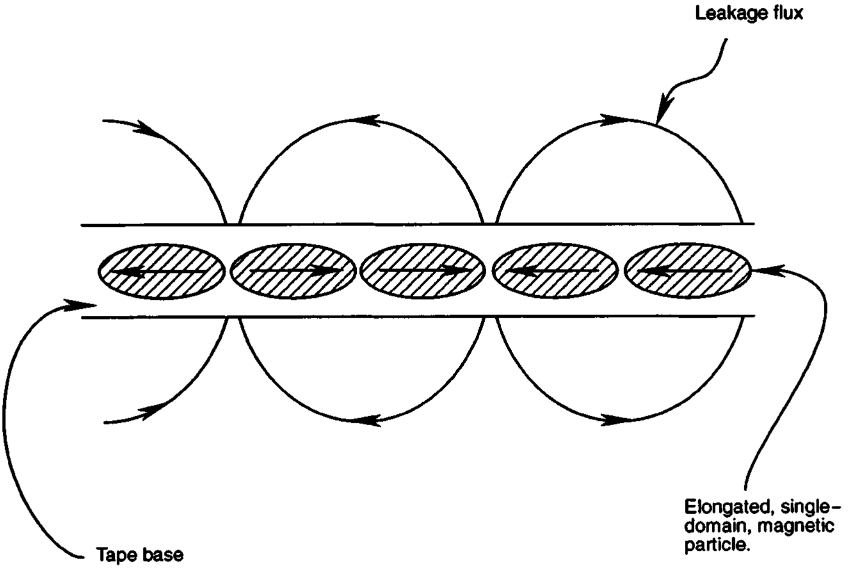


Figure 14.3 Arrangement of acicular magnetic particles in conventional 'longitudinal' magnetic recording tapes.

Metallic particle tapes were first made available commercially for audio tapes by 3M Corporation in 1979. Later Sony used this form of tape in its video cassettes in 1985. The advantages of this form of tape include the high saturation magnetizations that can be achieved, up to $1.7 \times 10^6\,\mathrm{A\,m^{-1}}$ in iron tapes, particle sizes of below 0.4 μm, and coercivities above $100\,\mathrm{kA\,m^{-1}}$. In fact coercivities above $200\,\mathrm{kA\,m^{-1}}$ and particle sizes below 0.05 μm have been achieved [3]. Metal evaporated tapes, or simply 'ME' tapes, were also introduced in 1979 for audio cassettes by Matsushita in Japan, and later for video cassettes. In 1995, ME tape for digital video cassettes became available. Recording densities on these tapes are

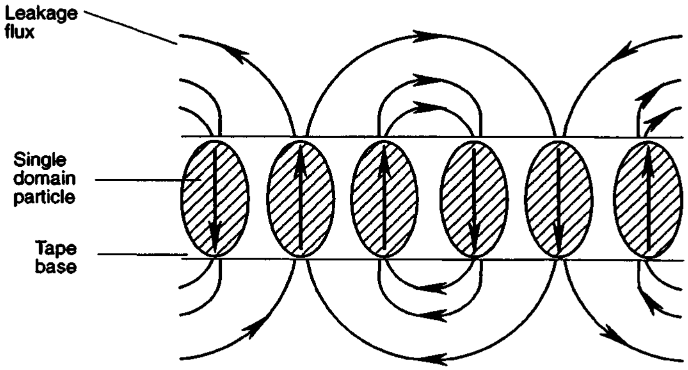


Figure 14.4 Arrangement of acicular magnetic particles in 'perpendicular' magnetic recording tapes.

currently 500 Mbits in^{-2}, which is the highest among tape-based systems. There seems no reason to doubt that 1 Gbit in^{-2} tapes will be produced in the near future.

Efforts have been made to develop tapes and disks for so-called perpendicular recording media in which the long axes of the particles are aligned at right angles to the surface as shown in Fig. 14.4. This should, in principle, allow higher recording densities to be achieved although several difficulties have been encountered in the development of perpendicular recording including the height of the head above the medium [4], noise during the reading process [5] and mechanical stability of the recording medium. The materials under consideration for perpendicular recording are cobalt/chromium and barium ferrite.

14.1.3 Magnetic disks

How are magnetic recording disks produced?

The principles of recording on magnetic disks are almost identical to those of recording on magnetic tapes and these will be discussed shortly. Magnetic recording disks come in two categories: floppy disks and hard disks. The materials used as a magnetic recording medium on floppy disk are also broadly similar to those used on tapes. Floppy disks are made in the same way as tapes, and are usually $3\frac{1}{2}$ inch in diameter.

Hard disks were developed to provide data storage for large 'main frame' computers, and originally were used exclusively for this application. Today hard disks are supplied as standard items on small personal computers because they provide greater storage capacity than floppy disks. In 2000 a typical personal computer came supplied with a hard disk of capacity between 50 and 80 Gbytes of memory. However, with disk storage densities now doubling in less than a year it is reasonable to expect these capacities to reach close to 1000 Gbytes by the year 2005.

Magnetic hard disks are produced by forming several layers of material, including a nonmagnetic underlayer, a magnetic layer, an overcoat and a layer of lubricants on a nonmagnetic disk substrate. The substrate is usually made of aluminium or an aluminium-4% magnesium alloy in the form of a disk 1.3–1.9-mm thick and 95–356 mm in diameter. An underlayer of nickel phosphide is then evaporated to a thickness of 10 μm onto the substrate to improve adhesion, followed by a 5–10-nm thick layer of chromium or chromium/vanadium which helps to control the crystallographic orientation and grain size of the magnetic layer by epitaxy. The magnetic layer, which consists of a cobalt-based alloy, is then evaporated onto the surface to form a layer 50 nm thick. Addition of platinum and tantalum has been found to enhance the anisotropy of the cobalt layer.

An overcoat is deposited to provide a protective layer 15–20 nm thick. This is usually a carbon-based material, but sometimes zirconia or tin are used. Finally, disk lubricants are added to limit friction and increase the lifetime of the disk by reducing wear. These lubricants assist in-contact, slow-speed sliding of the read/write assembly, which usually rests on the disk when stationary. As the disk begins to rotate several tens of revolutions are needed before the air bearing begins to operate fully and lift the read/write head above the surface of the disk. The lubricants remain on the surface (unlike in floppy disks and particulate tape where they are embedded in the material) and are known as 'topical' lubricants. These usually consist of perfluoropolyether, a long-chain fluorocarbon compound which is applied in monolayer thicknesses.

One of the advantages of disks over tapes is that access time is much shorter on disks. This is mainly because the read heads can be moved quickly to the right sector of the disk, whereas in tape recording it is necessary to rewind the tape to find the data. In disk recording the access time can also be improved by rotating the disk at a higher angular velocity. This by itself can bring problems of additional wear on floppy disks but not on hard disks. The reason for this is that on floppy disks the read/write head is in contact with the disk during the reading and writing process, but in hard disks there is no direct contact during reading or writing.

In magnetic tape recording the contact between tape and the read and write heads is a crucial factor in determining performance, but in that case actual contact is acceptable because of the relatively low number of replay head/magnetic tape passes expected. The heads are even contoured to improve contact with tape. In magnetic floppy disks the head also rides in contact with the disk. On hard disks the read/write head is not in contact with the disk except when the disk is stationary and briefly as the disk comes up to full speed. In order to optimize the performance of hard disks while ensuring that there is no direct contact between head and disk during the read/write process, an air bearing is used. In this way the head can be maintained close to, but not actually in direct contact with, the disk. This is the so-called 'flying head'. The air flow is caused by the relative velocity

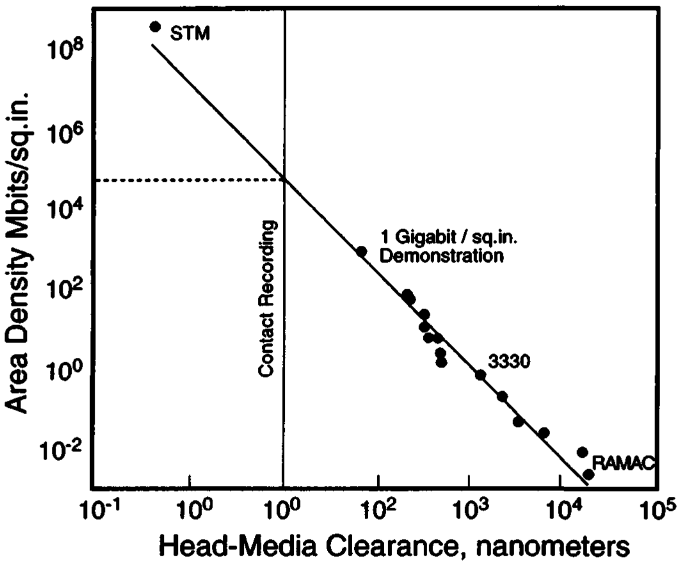


Figure 14.5 Relationship between head-media clearance (flying height) and storage densities (after Grochowski and Thompson [2] © IEEE 1994).

between the disk and head and this maintains a small gap. The typical head-to-disk separation in today's hard disk drives is 50 nm. When this arrangement fails, as it does occasionally, we encounter the so-called 'disk head crash' which usually results in some damage in the form of lost data.

The correlation between the head-disk separation, or flying heights, and storage density is clearly shown in Fig. 14.5. The head-disk separation can be reduced to about 10 nm before the situation becomes effectively 'contact' rather than 'non-contact' recording, because at this separation the thickness of the wear-resistant coating and warping of the disk have to be taken into consideration. The track density to linear bit density ratio is currently around 1:16. Track densities on hard disks are 25 000 TPI ('tracks per inch', or 1000 tracks per mm) while linear bit densities along these tracks are 400 000 BPI ('bits per inch' or 16 000 bits per mm) [6].

14.1.4 Materials for magnetic recording media

What different types of magnetic materials are currently used for recording media?

Gamma ferric oxide

The most widely used magnetic recording material is gamma ferric oxide (γ-Fe_2O_3) which has been used in magnetic tapes since 1937. Gamma ferric oxide is not a commonly occurring form of Fe_2O_3 but is produced by oxidation of specially prepared Fe_3O_4. The coercivity of these tapes is in the range 20–30 kA m^{-1} (250–375 Oe) [7]. The particle size used is typically a few tenths of a micrometre with a length-to-diameter ratio of anything from 10 : 1 to 3 : 1. The shape anisotropy of the particles, of course, also determines their magnetization characteristics such as coercivity. Saturation magnetization of the gamma ferric oxide is 350 kA m^{-1} (350 emu cc^{-1}) while the Curie temperature is about 600°C which is sufficiently high to avoid undue temperature dependence of the properties of the medium while operating under normal conditions in the vicinity of room temperature. Other properties are shown in Table 14.1.

Table 14.1 Magnetic properties of materials used in particulate magnetic recording media.

	γFe_20_3	$Co/\gamma Fe_20_3$	CrO_2	Fe	$BaO.6Fe_20_3$
M_s (kA m^{-1})	350	370	480	1700	370
T_c (°C)	600	600	128	770	320
H_c (kA m^{-1})	25–30	30–70	35–75	100–200	50–200
σ_s (A m^2 kg^{-1})	75	78	95	220	70
ρ (kg m^{-3})	4900	4900	4900	7800	5300
Particle size (μm)	0.5 × 0.1	0.5 × 0.1	0.4 × 0.05	0.15 × 0.05	0.15 × 0.05

Cobalt surface-modified gamma ferric oxide

Cobalt surface-modified gamma ferric oxide is now used as a magnetic recording medium because it has a higher coercivity than gamma ferric oxide [8]. The cobalt accumulates preferentially in the surface of the tape to a depth of about 3 nm. The addition of cobalt increases the anisotropy of the material leading to higher coercivity, for amounts of cobalt adsorbed on the surface up to 2%. Above this amount of cobalt, coercivity remains stable, but the saturation magnetization begins to decrease [9]. The cobalt is added to the ferric oxide at the last stage of processing before it is coated onto the substrate. Most video tape now contains cobalt surface-modified ferric oxide that has a coercivity of 48 kA m^{-1} (600 Oe). Note that absorption of cobalt into the bulk of the material increases the temperature sensitivity of magnetic properties such as coercivity, and is therefore disadvantageous.

Chromium dioxide

Chromium dioxide was also popular as a high-performance material for audio recording, before the cobalt-doped surface modification process was invented, in order to produce a magnetic recording material with higher coercivity than gamma ferric oxide. Chromium dioxide has a coercivity of 40–80 kA m^{-1}. Its saturation magnetization is slightly higher than gamma ferric oxide at 480 kA m^{-1} (480 emu cc^{-1}), but it has a rather low Curie temperature of 128°C which makes its performance more temperature sensitive, a factor which is a distinct disadvantage. It is also more expensive than iron oxide which reduces its commercial attraction. It has been replaced by the cobalt-doped gamma ferric oxide as a high-performance recording material.

Ferromagnetic powders

Iron powder is also used as a recording medium. This has higher saturation magnetization than the oxide particulate media described above and so can be used in thinner coatings. The coercivity of these fine particles is typically 120 kA m^{-1}. The production of the iron particle tapes is a modification of the production process for iron oxide tapes in which the oxide is finally reduced to metallic iron under a hydrogen atmosphere at 300°C. However, these tapes also need a surface coating of tin to prevent sintering whereby the particles coalesce and are no longer single domains. Typical magnetic properties are saturation magnetization 1700 kA m^{-1} (1700 emu cc^{-1}), and coercivity 120 kA m^{-1} (1450 Oe).

Thin metallic films

Thin metallic films are now the principal magnetic medium for high-performance, high-storage density, hard disk drives. The main requirement in hard disk drive media is to obtain an adequate coercivity because ultimately this determines the bit sizes which can practically be used, and hence the recording density. The magnetic layer is 50 nm thick and consists of an assembly of almost noninteracting particles of diameter 50 nm. It is usually cobalt with other chemical additives such as phosphorus, chromium, nickel, tantalum or platinum. These additives are used to increase the anisotropy and hence the coercivity of the cobalt film and to reduce the coupling between grains. The increased coercivity of the film reduces the noise level caused by fluctuations in leakage field. Typically coercivities of these thin film media are up to $H_c = 80$ kA m^{-1} (1 kOe), with remanences of $M_R = 800$ kA m^{-1} ($B_R = 1$ T).

Cobalt-chromium thin film media have improved corrosion resistance compared with pure cobalt. Originally this was the reason for the chromium additions. However, it was found that the chromium accumulates in paramagnetic intergranular regions thereby reducing the exchange coupling between the grains, increasing coercivity and reducing noise fluctuations in the film. Tantalum has also been found to reduce noise fluctuations by reducing grain sizes in the magnetic film [10].

Metallic films are also used in some recording tapes because of their high saturation magnetization and remanence. They can be used in the form of very thin coatings since the leakage magnetic fields, which are used in the reading process, are proportional to the remanent magnetization on the tape. The higher saturation magnetization therefore ensures that these leakage fields are rather larger than for similar thin films of other materials. The pickup voltage in the read head is proportional to the magnetic field from the tape. Thinner recording media allow higher recording densities since the rate of change of field with distance dH/dx along the tape can be made larger.

In the last 10–20 years serious efforts have been made to develop metallic thin film recording tapes as an alternative to the widely used particulate tapes. The magnetic layer is usually an evaporated film of cobalt-nickel alloy which is formed as slanting columns in a porous matrix of cobalt and nickel oxides. These have remanences of $M_r = 0.3 \times 10^6\,\mathrm{A\,m^{-1}}$ ($B_r = 0.4\,\mathrm{T}$) and coercivities of $H_c = 80\,\mathrm{kA\,m^{-1}}$ (1 kOe). The disadvantage of these tapes is that they do not wear well and so their lifetimes are relatively short. They need surface coating of lubricants (topical lubricants) which are usually fluorocarbons similar to those used on metallic disks, and corrosion inhibitor. However, whereas hard disk drive units are hermetically sealed to retain lubricant and keep out contaminants, the metallic tapes are necessarily exposed and therefore retention of the lubricant is a problem.

Hexagonal ferrites

Hexagonal ferrites have much higher coercivities than any of the above materials and are used for more specialized applications such as credit cards where there is less likelihood of a need for rerecording, but where it is imperative that there is little chance of demagnetization by unanticipated exposure to low and moderate external magnetic fields.

The main materials of interest are barium ferrite, and to a lesser extent strontium ferrite. Barium ferrite $BaO.6Fe_2O_3$ has a coercivity of up to $480\,\mathrm{kA\,m^{-1}}$ in its pure form, with saturation magnetization of $370\,\mathrm{kA\,m^{-1}}$ (4.6 kOe). Disk-shaped barium ferrite particles of diameters 50 nm can be produced in films 5 nm thick on substrates by controlling the particle growth conditions. These small particle sizes allow the possibility of very high storage densities (250×10^9 bits $\mathrm{in^{-2}}$) and high signal-to-noise ratios. The particles grow with their easy axis normal to the plane of the film and therefore barium ferrite is one of two leading candidate materials for perpendicular recording media, as discussed below. The addition of cobalt and/or titanium to barium ferrite reduces its anisotropy and so leads to a reduction in coercivity to the range $40\text{–}160\,\mathrm{kA\,m^{-1}}$ (0.5–2 kOe).

Perpendicular recording media

Perpendicular recording media in which the magnetic domains are oriented with magnetizations normal to the plane of the medium have been pioneered by Iwasaki in Japan [11]. These media offer potential for higher recording density than conventional or 'longitudinal' media but so far seem to suffer from other problems which have prevented them becoming viable, such as the need for a very small head-to-medium distance and noise problems in the reading process. The material that has been used for this is a sputtered cobalt-chromium film 100-nm thick containing greater than 14% Cr, which forms columns about 50–100 nm in diameter normal to the surface of the substrate. In addition, due to the nature of the growth of these films, the magnetic moments remain perpendicular to the plane of the film, unlike the previous materials in which the magnetic moments lie in the plane of the material. When the moments are perpendicular to the plane it is believed that the transitions between neighbouring 'bits' become much sharper leading to an increased recording density. Another material that has been tried is oriented barium ferrite, which is produced on a plastic substrate by a method similar to that used to produce particulate tapes.

In principle, perpendicular recording has been shown to be possible; however, it has yet to become more than merely of scientific interest. In practice there are problems with mechanical failure of the medium on floppy disks and also the small head-to-medium separation needed is less than present technology is capable of handling [12].

14.1.5 Magneto-optic recording

Can magneto-optic reading methods be used to develop recording disks with higher storage densities or faster access times?

Another area of interest in magnetic recording is that of magneto-optic devices. These make use of the Faraday and Kerr effects in which the direction of polarization of light is rotated in the presence of a magnetic field. In this way, two oppositely magnetized regions on a magnetic medium can be distinguished. The advantage of magneto-optical disks is that the storage density can be 1000-times greater than for floppy disks [13, 14], while access time for magneto-optic disks are 40–100 ms which are about 10-times faster than for floppy disks but are not yet competitive with access times for hard disks, which are typically 20–60 ms [14].

In 1988 the ISO standard $5\frac{1}{4}$ inch magneto-optic drive was introduced which was a double-sided drive with a total storage capacity of 750 Mbytes. The second-generation ISO standard was introduced in 1991 with a capacity of 1300 Mbytes at a density of 400 Mbit in^{-2} (0.64 Mbits mm^{-2}), and in 1994 the third-generation ISO standard magneto-optic drive became available with 2 Gbytes of storage capacity at a density of 600 Mbits in^{-2} (0.96 Mbit mm^{-2}). This third-generation drive is still commercially available. All three systems used 780 nm wavelength laser light. Access times for these disks have remained at about 100 ms and the data transfer rate is 5–10 Mbit s^{-1} [15].

The recording of information depends on thermomagnetic magnetization in which an intense light source such as a focused laser beam is used to heat a small region of a thin film of ferrimagnetic material above its Curie temperature and then it is allowed to cool again. If the material is exposed to a reverse magnetic field throughout the process (i.e. is operating in the second quadrant of its magnetic hysteresis loop) then we know from earlier discussion of the anhysteretic magnetization that its optimum magnetic energy state corresponds to magnetization in the opposite direction. As it cools through the Curie point the magnetization obtained in the region exposed to the laser beam will be the anhysteretic magnetization under the prevailing field, which will be in the third quadrant. This means that the regions which have been exposed to the laser beam will be magnetized in the opposite direction, as shown in Fig. 14.6.

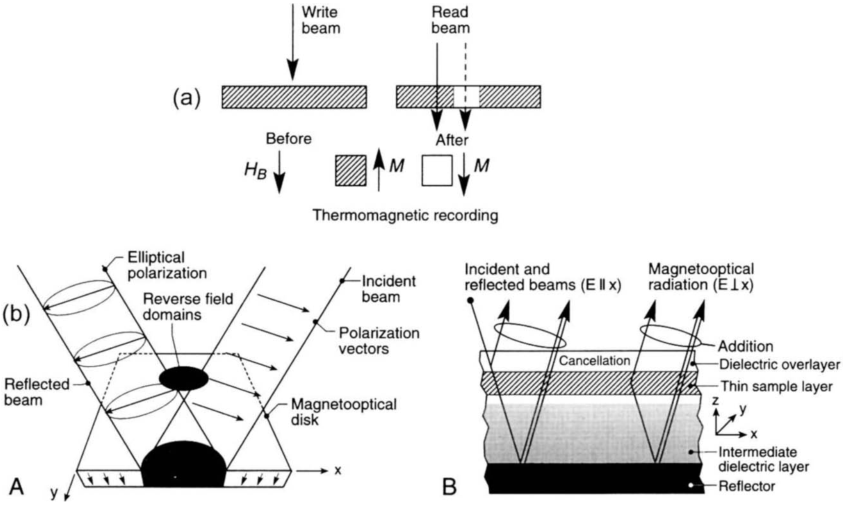


Figure 14.6 Magneto-optic reading and writing processes.

The subsequent reading of magnetic information on the medium depends on the magneto-optic Kerr effect. A polarized laser beam of weaker intensity than that used for writing is reflected from the surface of the magnetic recording medium, as in Fig. 14.7, and is then passed through a polarized analyser before being detected. The presence or absence of the reverse domains can then represent either ‘0’ or ‘1’. The film can later be wiped clean by saturating the magnetization in the original direction.

It should be noted that for purposes of detection this technique works best in perpendicularly magnetized media. Signal-to-noise ratios are comparable with conventional magnetic disk recording. Magneto-optic disks made by Philips have a 50-nm thick magnetic coating on a transparent 3-mm thick plastic substrate. Their method uses a 3-mW laser with a spot size of 2 × 5 μm reading the disk by the Faraday effect.

Two types of magnetic materials are generally used in the production of magneto-optic disks. The most widely used material is a rare earth-iron-cobalt

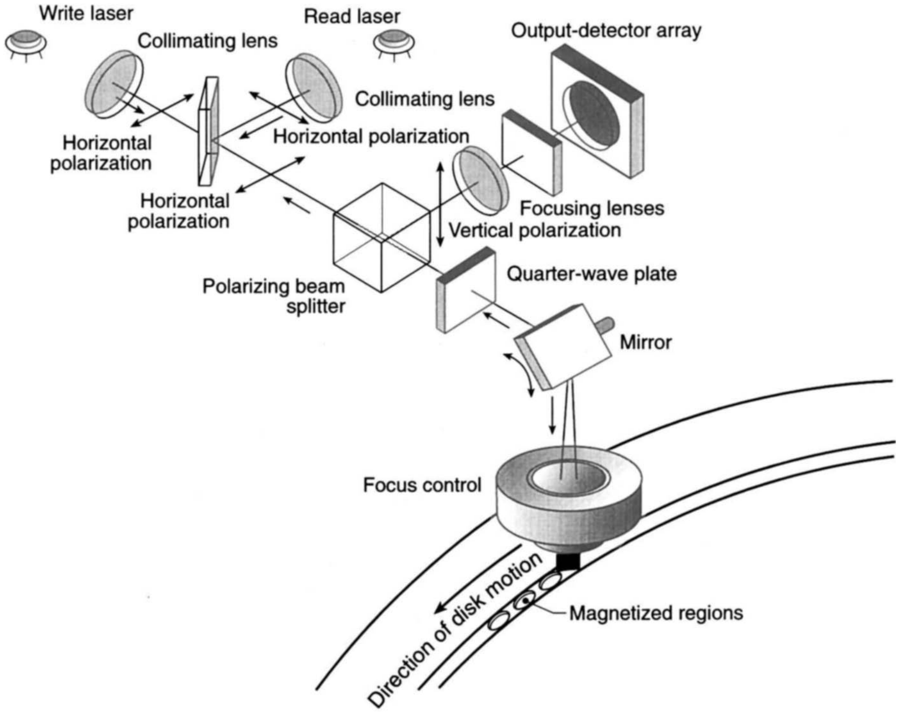


Figure 14.7 Components of a magneto-optic disk recording system. Reproduced with permission of R. M. White, *Introduction to Magnetic Recording*, published by IEEE Press, 1985.

film of general composition $R_x(FeCo)_{1-x}$, R is usually terbium plus gadolinium, although sometimes dysprosium is used. The film is produced as an amorphous layer 50-nm thick on a rigid plastic substrate, which is 3-mm thick. The rare earth-iron-cobalt films have high perpendicular anisotropy ($K_u > 105\,J\,m^{-3}$), a coercivity of typically $240\,kA\,m^{-1}$ (3 kOe), a high squareness ratio M_r/M_s and a Curie temperature of 250–300°C depending on the composition. The advantages of using an amorphous film include (i) high uniformity of properties due to lack of crystallinity, (ii) low-cost deposition by sputtering, and (iii) fine control of the magnetic properties, such as saturation magnetization and coercivity, by changes of the chemical composition of the film.

The main alternative material for magneto-optic disks is cobalt/platinum, which is deposited in a multilayer. These multilayers consist of 10–30 pairs of layers, each pair consisting of 0.3–0.5 nm of cobalt and 0.8–1.2 nm of platinum. The high uniaxial anisotropy of the multilayers ($K_u > 10^5\,J\,m^{-3}$ in these films) ensures that the domain magnetizations align perpendicular to the plane of the film. The coercivities of these films are above $80\,kA\,m^{-1}$ (1 kOe), but perhaps more significantly they have high Kerr rotations at short wavelengths. One disadvantage of these Co/Pt multilayers is the relatively high Curie temperature of about 400°C, which means that the disks have to be heated to a high temperature in order to change the direction of magnetization.

14.2 MAGNETIC RECORDING HEADS AND THE RECORDING PROCESS

How is information stored on, and retrieved from, the recording medium?
The recording process involves the mechanism by which a magnetic imprint is left on the magnetic medium and the mechanism by which this imprint is read from the medium and the original information, whether an audio signal or some digital data, is recreated.

The writing process is the means of transferring electrical impulses in a coil wound on an electromagnet (the writing head) into magnetic patterns on the storage medium. The reading process is the inverse of this mechanism. The reading process is quite well understood since it requires no knowledge of the magnetization characteristics of the medium. Only the remanent magnetization of the medium determines the response. However, the writing process, which involves the effect of an applied field on the magnetization of a magnetic medium, is by comparison more difficult to model. This is because it is difficult to model the dependence of magnetization of the medium on the magnetic field even when the field is completely uniform, and in these cases the field is not even uniform.

Magnetic recording heads are either inductive heads for writing, as shown in Fig. 14.8, or magnetoresistive heads for reading, as shown in Fig. 14.9. Magnetoresistive heads confer significant advantages over the inductive heads for reading because of their low power requirements and high sensitivity. The magnetoresistive read heads that are used in disk drives were originally developed employing conventional anisotropic magnetoresistance in permalloy. This exhibits changes in

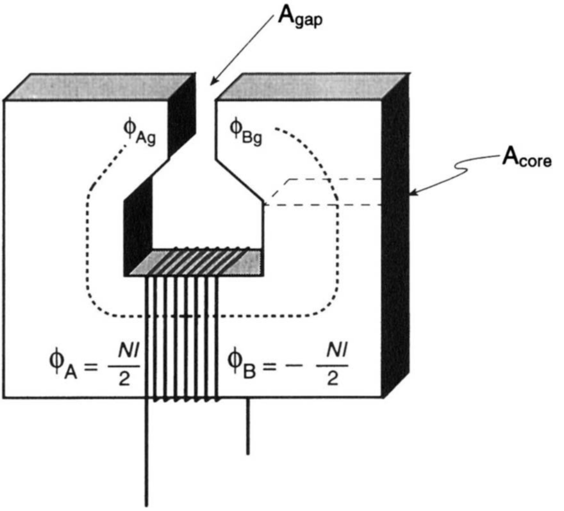


Figure 14.8 An inductive write head.

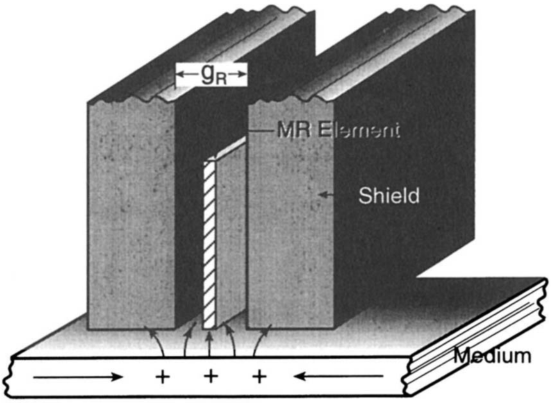


Figure 14.9 A magnetoresistive read head.

resistance $\Delta R/R$ of typically 2–3% in fields of 300–400 A m^{-1} (4–5 Oe). The giant magnetoresistive read heads that are used in commercial disk drives have $\Delta R/R$ of typically ten times greater than this. The writing or imprinting of magnetic signals or data on the hard disks is still performed by inductive write heads. Therefore, the current generation of disk drives employ separate technologies for the reading and writing functions, unlike earlier disk drives which used inductive heads to perform both operations.

14.2.1 Inductive recording heads

How does the current generation of 'write heads' work?

Inductive recording heads that are used for writing data on disks and tapes consist of a 'C'-shaped high permeability core with a gap known as the 'head gap', of

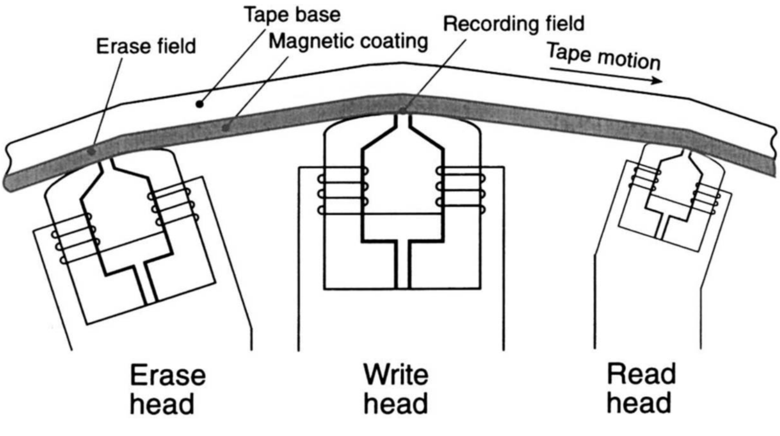


Figure 14.10 Magnetic erase, read and write heads in a magnetic tape recording system.

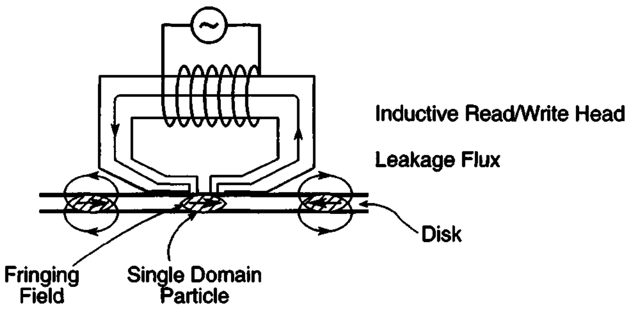


Figure 14.11 Schematic diagram of an inductive read head above a magnetic disk.

width 0.1 μm, as shown in Fig. 14.8. The core is wound with a flux coil in order to generate the necessary magnetic flux density in the gap when a current is passed through the coil. The field in the gap is called the 'fringing field'. In magnetic tape recording the heads are in direct contact with the tape, as shown in Fig. 14.10, and usually there are separate heads for reading, writing and erasing. In disk recording there is only one inductive head which rides on an air bearing above the surface of the disk, as shown schematically in Fig. 14.11. The write head material must have high saturation magnetization in order to leave a large imprint (high magnetization) on the tape but it must also have low remanence to ensure that there is no

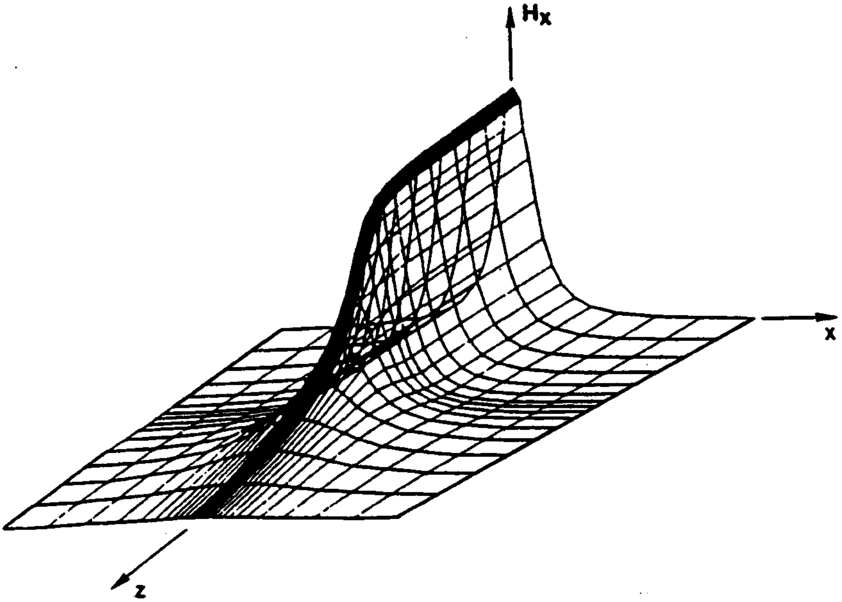


Figure 14.12 The magnetic fringing field in the gap of a recording head as a function of position: the x direction is given in Fig. 14.13 and the y direction is normal to the plane of Fig. 14.13.

writing when the current in the coil is zero. Further, it is also clear that a low coercivity is desirable. Write heads are constructed of magnetically soft material, these include soft ferrites ($MO.Fe_2O_3$), sendust Al-Fe-Si and Al-Fe, permalloy (Ni-Fe), and amorphous cobalt-zirconium.

In writing mode the magnetic tape or disk passes the head where the fringing field causes a realignment of the magnetization within the single-domain particles. The magnetization is then a record of the strength of the field in the gap of the recording head at the time that the disk or tape passed it. In reading mode the passage of the tape causes a variation in flux density in the read head which is converted into a voltage in the coil wound on the read head. The signal is then amplified and, in the case of audio recording, used to activate a loudspeaker.

The magnetic field in the gap of the write head, which is the main region of interest to the magnetic recording engineer in the writing process, can be determined by the finite-element techniques described in Chapter 1. This fringing field is shown in Fig. 14.12. With present technology that uses conventional, 'in-plane' recording media, the tape responds to the component of the fringing field which is parallel to the tape surface. However for perpendicular recording the fringing field needs to be normal to the plane of the tape or disk and this leads to differences in the design of the read and write heads.

14.2.2 Magnetoresistive read heads

How does the current generation of 'read heads' work?

Magnetic recording technology has recently been making rapid progress in storage densities due principally to the development of the magnetoresistive read head, which has led to the production of a new generation of high-sensitivity, high spatial resolution read heads [16]. Commercially available magnetoresistive heads today employ giant magnetoresitive multilayers. These multilayers of iron/chromium or cobalt/copper can exhibit changes in resistance $\Delta R/R$ of 80%, although the fields needed for this are still higher than in the anisotropic magnetoresistive material.

The magnetoresistive head detects the leakage flux from the locally magnetized regions of the recording medium by means of a two-point resistance measurement, as shown in Fig. 14.9. Spatial resolution is determined by the distance between the electrical contacts in the perpendicular or 'cross-track' direction, which is typically around $2\,\mu m$, and by the shield separation along the track direction, which is typically $0.1\,\mu m$. To obtain an antisymmetric response that is approximately linear with strength of the leakage field, the magnetoresistive material also needs to be biased with another constant applied field. This is often supplied by using an adjacent magnetized film. The voltage generated by the magnetoresistive element also depends on the current, and so it is desirable to operate at high current to maximize sensitivity. Current densities of $2 \times 10^{-3}\,A\,m^{-2}$ are typically used in these sensors at present. The operating current density is, of course, limited by the power dissipation in the head.

So-called 'colossal magnetoresistance' (CMR) has been discovered in manganate perovskite thin films in which the fractional change in resistance $\Delta R/R$ can be as high as 99.9%, but only at the rather low temperature of 77 K and at a high field of

$4.6 \times 10^6\,\mathrm{A\,m^{-1}}$ (6 T). This may indicate the future direction for magnetoresistive read heads, but at present it is not clear whether these materials will ultimately prove useful in magnetic recording applications. Certainly the effect will need to be observed at higher temperatures and at lower field strengths before it can be considered as a potential practical alternative to the current technology.

14.2.3 Writing head efficiency

How can the performance of a recording head be determined?

Head efficiency is the ratio of magnetomotive force obtained across the head gap to the magnetomotive force supplied by the energizing coil. This is determined from consideration of the magnetic circuit formed by the magnetic core and air gap of the head.

In the air gap

$$B_g = \mu_0 H_g. \tag{14.1}$$

In the core

$$B_c = \mu_r \mu_0 H_c, \tag{14.2}$$

where μ_r is the relative permeability of the core and in this case H_c is the field in the core. The magnetomotive force of the driving coil is also the magnetomotive force across the whole magnetic circuit. If ℓ_g is the length of the gap and ℓ_c is the length of the ferromagnetic core

$$N_i = H_c \ell_c + H_g \ell_g. \tag{14.3}$$

The efficiency of the core, being simply the ratio of magnetomotive force across the gap $H_g \ell_g$ to the magnetomotive force supplied from the coil N_i, is then

$$\eta = \frac{H_g \ell_g}{(H_c \ell_c + H_g \ell_g)} = \frac{H_g \ell_g}{N_i} \tag{14.4}$$

The efficiency can also be expressed in terms of the reluctances of the magnetic paths R_c in the core and R_g in the gap:

$$\eta = \frac{R_g}{(R_c + R_g)} = \frac{(\ell_g/A_g)}{(\ell_g/A_g) + (\ell_c/\mu_r A_c)}, \tag{14.5}$$

where A_c is the cross-sectional area of the core and A_g is the cross-sectional area of the gap. This means that a large gap field, and hence a large fringing field, requires a large permeability μ_r in the core and a large ratio A_c/A_g.

The head gap together with the saturation magnetization of the head material determines the field in the gap. This determines the fringing field in the vicinity of the gap and consequently the maximum coercivity of the recording medium that can usefully be used with the head.

Karlqvist heads are an idealization often used for calculating the field in the vicinity of the gap. In Karlqvist heads the pole gap is small compared with the pole

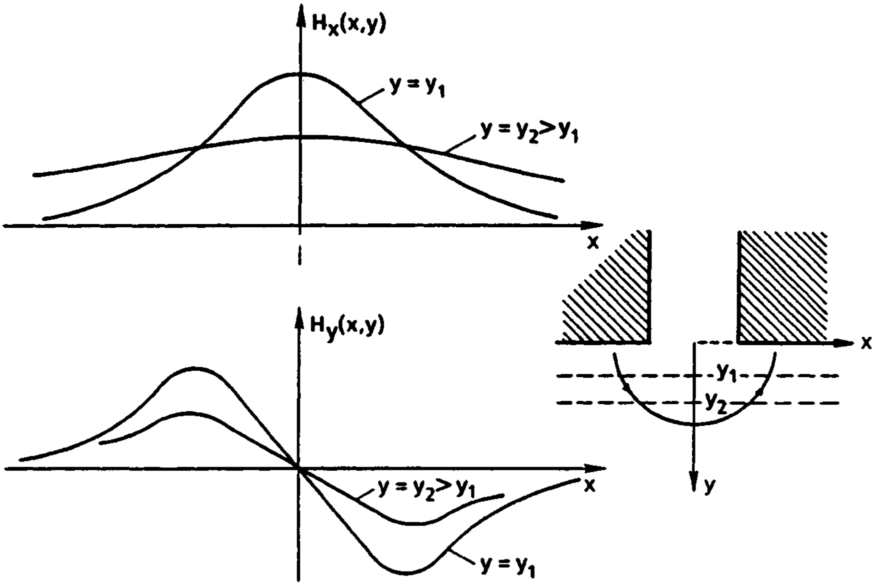


Figure 14.13 Vertical and horizontal components of the magnetic field in the vicinity of a Karlqvist head.

tip lengths, such as in the head depicted in Fig. 14.13. In the past these provided a relatively simple geometrical situation for determining the magnetic field in the gap. With three-dimensional finite-element codes available now for calculating the fields in the gap the Karlqvist approximation is less important.

14.2.4 The writing process

What happens when information is written onto a recording medium?

We now consider the process of magnetizing the recording medium. Specifically we wish to know how the medium responds to an applied field in the head gap. In general, this is a rather difficult problem which is not very well understood, although some empirically based models can be used to good effect. One problem that arises as the medium passes close to the head is that the magnetic fields at different depths in the magnetic medium are different, as shown in Fig. 14.14. Secondly, as the medium passes the gap the field it experiences changes with time.

To give an example of the writing process we will suppose that a given region of the recording medium passing the head begins at the positive remanence point on its hysteresis curve, as shown in Fig. 14.15. Then, as it passes the gap with a negative field the material passes down the second quadrant of the hysteresis loop to its coercive point $-H_c$, for example. As it passes the head the magnetic field it experiences from the head gap decreases to zero and the material magnetization passes along a recoil minor hysteresis loop to $H = 0$, ending with a small positive remanence.

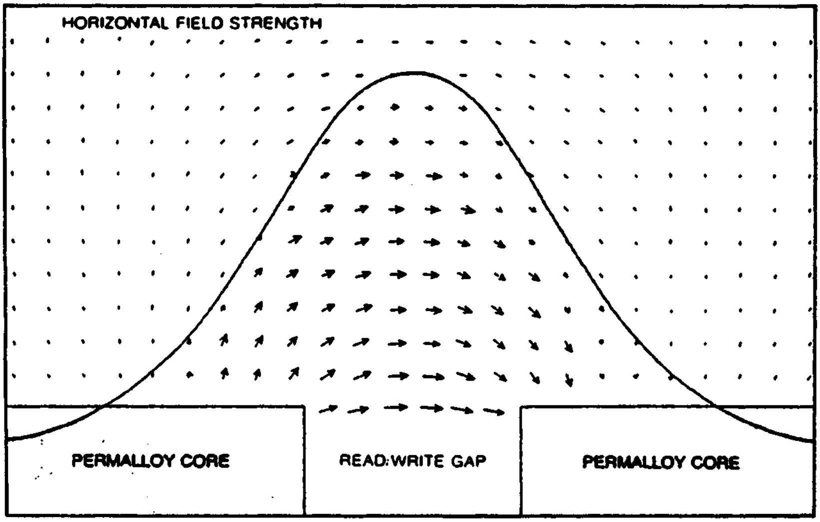


Figure 14.14 Variation of the magnetic field strength in both direction and magnitude close to a permalloy inductive recording head used for writing. © 1985 IEEE.

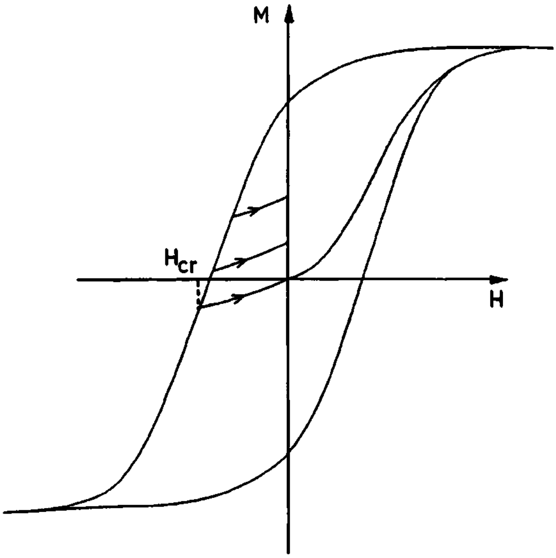


Figure 14.15 Recoil minor loops during the writing process for different regions of a magnetic recording medium as it passes a recording head.

This means that even where the magnetization had been reduced to zero at the coercive point the magnetization will actually finally increase again to give a positive remanent magnetization. Therefore, to result in a completely demagnetized state $M = 0$ the field experienced by the medium must be greater than $-H_c$, a point referred to as the remanent coercivity H_{cr}. A very square hysteresis loop in which the recoil minor loops are very flat $(\mathrm{d}M/\mathrm{d}H) \cong 0$ is therefore desirable.

During the writing process the time-varying current in the writing head coil changes thereby altering the field in the gap. This causes localized changes in the magnetization in the recording medium which passes the write head at a constant speed. It has proved difficult to determine the magnetization of the disk or tape in two dimensions and therefore theoretical models have only limited usefulness in the predictions of tape magnetization.

Models for the magnetization of the recording medium usually make use of some very simplistic approximations to the magnetization characteristics of the medium such as assuming M is a single-valued function of field. One such model is the Williams–Comstock [17] model. This is a one-dimensional model which employs the arctangent function. The magnetization $M(x)$ in the tape can be expressed as a function of distance in response to a step function change in field in the gap by the equation

$$M(x) = \left(\frac{2M_{\mathrm{R}}}{\pi}\right) \arctan\left(\frac{x}{a}\right), \tag{14.6}$$

where x is the distance along the tape or disk, a is an adjustable parameter which is determined by the rate of change of magnetization with distance and M_{R} is the remanent magnetization.

14.2.5 Recording density

How much information can be stored on a magnetic disk?

The recording density in a medium depends on the magnetic properties of the medium and the characteristics of the writing head. The recording density is determined by the product of bits per inch (BPI) and the number of tracks per inch (TPI). Currently, hard disks have TPI values of 25 000 and BPI values of 400 000. The maximum attainable BPI can be measured by the parameter a, known as the transition length, which is the minimum distance along the tape that is needed to reverse completely the magnetization from saturation remanence in one direction to saturation remanence in the other direction.

The transition length in which a signal can be made to change is dependent on $\mathrm{d}M/\mathrm{d}x$ in the recording medium. This can be expressed as the product $(\mathrm{d}M/\mathrm{d}H)(\mathrm{d}H/\mathrm{d}x)$, where $(\mathrm{d}M/\mathrm{d}H)$ is a property of the medium, specifically the slope of the hysteresis curve, while $(\mathrm{d}H/\mathrm{d}x)$ is a property of the writing head. If we make the approximation that the slope of the hysteresis loop is constant then for a fixed field gradient in the head gap we have the following expression for the transition length a

$$a = \frac{(2M_{\mathrm{R}}/\pi)}{(\mathrm{d}M/\mathrm{d}x)_{\max}} = \frac{(2M_{\mathrm{R}}/\pi)}{(\mathrm{d}M/\mathrm{d}H)(\mathrm{d}H/\mathrm{d}x)_{\max}} \tag{14.7}$$

This can be verified from the Williams–Comstock equation above. The transition length a is therefore made smaller for large field gradient $\mathrm{d}H/\mathrm{d}x$ and large

dM/dH on the major hysteresis loop, that is for a square hysteresis loop. Of course, there are other factors which have not been taken into account in this simple analysis, such as the demagnetizing field in the tape and the spatial variation in the transition region. However, these do not alter the basic conclusion about the desirability of high field gradient in the gap and square hysteresis loop materials. Notice that while a large value of dM/dH on the major hysteresis loop is desirable this should coincide with small dM/dH on the recoil minor loops.

14.2.6 AC bias recording

Are there any ways to improve the quality of the imprinted magnetic signal on the recording medium?

When signals are recorded in analogue form, such as in audio recording, it is advantageous to have the recorded signal proportional to the amplitude of the input signal. The recorded signal is the remanent magnetization on a region of the magnetic tape, while the input signal voltage is converted to an applied field in the recording head gap. Because the initial part of the magnetization curve is nonlinear the remanent magnetization on the tape would be a nonlinear function of the applied field if the dc initial magnetization curve of the tape were used. However by ac biasing, in which a sinusoidal field of decaying amplitude is superimposed on the dc field, the nonlinearity can be overcome. This produces the anhysteretic remanent magnetization curve. This variation of remanence with applied field is linear at low fields, and produces a more desirable linear recording characteristic which also improves the signal-to-noise ratio.

14.2.7 The reading process

How is the information read from a hard disk?

The reading process in magnetic recording is relatively well understood. The tape or disk passes below the read head and causes a fluctuation in the flux density in the magnetic core of the read head. The fringing fields from the tapes can be dealt with using simple models. For example, consider the situation shown in Fig. 14.11. As the disk or tape passes near to the reading head the stray field associated with the magnetic imprint on the medium enters the reading head. At the gap this field passes through the coil giving a voltage that, as we have shown earlier, is proportional to $-dB/dt$, the rate of change of magnetic induction linking the coil. Therefore the voltage in the reading head will be dependent on the stray magnetic induction emanating from the tape which is collected by the head and passed through the coil. The reading head efficiency is defined as the ratio of the tape flux entering the reading head that actually passes through the sensing coil.

14.2.8 Various types of recording devices

What other types of magnetic recording devices are available?

In the past, the most common form of magnetic recorders were audio recorders. These have traditionally been analogue recording devices which use ac bias recording, that is they make use of the linearity of the anhysteretic remanent magnetization curve to avoid distortion of the reproduced signal. By this method it is possible to make the magnetization imprinted on the recording medium proportional to the amplitude of the signal. In audio recording, particularly music, any distortion of the signal is undesirable. Therefore the reading and recording processes take place relatively slowly. The typical tape velocity in professional audio recording is $0.4\,\mathrm{m\,s^{-1}}$ [18], while on audio cassettes it is less than $50\,\mathrm{mm\,s^{-1}}$ [19].

Video recorders use frequency modulation in which the signal $S(t)$ imprinted on the tape is related to the original input signal $f(t)$ by

$$S(t) = \cos\left(\omega t + 2\pi\beta \int f(t)\,\mathrm{d}t\right), \tag{14.8}$$

where ω is the carrier frequency and β is the modulation index. The video signals range from 30 Hz to 7 MHz so that tape velocities are relatively fast and may be up to $5\,\mathrm{m\,s^{-1}}$ [18].

The most common form of magnetic recording devices, digital recorders, are in most cases peripheral devices for computers, whether disks or tapes. In digital recording it is only necessary to distinguish between '0' and' l' so these devices can function with much lower signal to noise ratios than are acceptable in analogue recording. Furthermore since in digital recording the actual level of the signal is not really crucial, providing that a '0' and' l' can be distinguished, the reading and writing process is very fast. However, even though the signal-to-noise ratio can be relatively small in digital recording the tolerable error rate is also very low.

14.3 MODELLING THE MAGNETIC RECORDING PROCESS

What computer models of the magnetic recording processes have been developed? For the purposes of designing new and improved magnetic read and write heads and magnetic recording media it is important to have a good quantitative understanding of the magnetization processes in the materials which can be expressed through validated computer models or simulations. Several different models are currently used. Here, two of these models are discusssed.

14.3.1 The Preisach model

What model is available to describe magnetic switching processes in their most general form?

In the magnetic recording industry the magnetic properties of the medium are usually represented using a model for magnetization as a function of field that was devised in the 1930s by Preisach. This model really does not give much physical insight into the magnetic properties of materials, being in essence a complicated mathematical data fitting procedure, but it can be used to give reasonable mathematical representations of hysteresis curves once the curves are already known. It has been found useful for modeling the magnetic properties of the recording media and is quite widely used in the recording industry [20].

The essential idea of the Preisach model is that the observed bulk magnetic hysteresis loop of a material is due to a summation of more elementary hysteresis loops of domains with differing switching fields (coercivities). These 'pseudo-domains' can only have two states within the confines of the model, with magnetization parallel or antiparallel to a given direction. The model relies on a density function called the Preisach function which is defined on a plane described by the positive and negative switching fields $H+$ and $H-$. This function is used to determine how many domains switch their orientation from + to −, or vice versa, as the field is swept between limiting values of magnetic field H.

The model works fairly well for weak interactions between domains such as occur in these recording media, which are usually aggregates of single-domain particles, and because the magnetic moments within the elongated single-domain particles can only have magnetic moments along one axis leading to a magnetization either parallel or antiparallel to the long axis of the particles. A comprehensive treatment of the Preisach model can be found in Mayergoyz's book on modeling of hysteresis [21].

14.3.2 Stoner–Wohlfarth theory

What model is available to describe switching processes in particulate magnetic recording materials?

The Stoner–Wohlfarth theory [22] has more relevance to particulate recording media than to the permanent magnets for which it was originally developed. In particulate recording media the isolated single-domain particles are deliberately created on the tapes or disks and these are clearly well-suited for the application of the Stoner–Wohlfarth theory. The model has therefore found appropriate applications in determining the magnetization characteristics of fine-particle recording media [23].

14.4 MAGNETIC RANDOM ACCESS MEMORIES

Is it possible to replace semiconductor RAM with magnetic RAM?

The random access memories that are currently in use in computers (dynamic RAM or 'DRAM', and static RAM or 'SRAM') are both based solely on semiconductor technology. As a result of this, data that is stored in these memories is lost when the power supply is interrupted. This can arise when the voltage supply is deliberately switched off, or when there are unplanned interruptions or significant fluctuations in the voltage supply. On the other hand magnetic memories offer the prospect of developing nonvolatile RAM whereby the magnetic states of the memory bits are not altered by the voltage interruptions and therefore data is maintained. Such memory has obvious beneficial attractions including the fact that computers would not need to be 'rebooted' when switched on. Such memories are also not susceptible to radiation exposure.

The new field of magnetoelectronics, in which magnetic and electronic functions are integrated into a single device, offers the prospect of significant improvements in the speed, power consumption and reliability of data memory devices

[24]. Prototype MRAM devices have already been shown to have improved on the speed and power consumption of conventional nonvolatile semiconductor memory by several orders of magnitude, and have reached data storage densities and speeds that are comparable with volatile dynamic RAM.

There are three main technologies that are coming under serious consideration: hybrid semiconductor/magnetic devices, magnetic tunnel junctions and all metal spin transistors and spin valves [25]. Of these, the hybrid devices have the advantage, being compatible with existing CMOS fabrication and processing procedures. These devices employ magnetic elements which can be magnetized in one of two directions. The leakage field from the magnetic elements can be detected by a Hall effect sensor and from the sign of the Hall voltage the 'orientation' of the memory bit can be determined.

Magnetic tunnel junctions consist of a simple three layer sandwich of two magnetic layers separated by a thin, 0.5–1.0 nm, nonmagnetic, nonconducting layer [26]. The tunnelling current that passes between the two magnetic layers through the dielectric differs depending on the relative orientations of the magnetization vectors in the two magnetic layers. The fractional change in resistance $\Delta R/R$ of such devices is now typically 12%. The tunnel junctions have several attractive features, not least of which is the possibility of fabricating MRAM with extremely high data-storage densities [27]. However, there are also some problems with this technology, including the fact that magnetic tunnel junctions are not compatible with normal CMOS processing methods, specifically the annealing that is normally performed at temperatures in the range 400–800°C to repair damage to the CMOS structures. These temperatures damage the magnetic tunnel junction which can only sustain temperatures in the range 200–300°C.

All metallic magnetoresistive devices such as the ferromagnet-metal-ferromagnet sandwiches and multilayers include the magnetic spin valve. Such devices have similarities to the magnetic tunnel junction. They employ either current in the plane 'cip' or current perpendicular to the plane 'cpp' configurations. The latter is also the configuration used in magnetic tunnel junctions. In these all-metallic structures the resistances are lower than in the tunnel junctions and the fractional change in resistance is only about 5%. Most device designs are analogous to those using magnetic tunnel junctions. The all-metallic devices have the possibility of high storage densities like magnetic tunnel junctions. However, because of the lower resistance they have low read-out voltages. The slow response time means longer read/write times of about 100 ns. Together, these two factors have prevented them reaching performance levels that are acceptable for MRAM applications.

Generation and control of spin-polarized currents in magnetoelectronic devices such as spin valves has been demonstrated in a metallic mesoscopic spin valve device by Jedema *et al.* [28] in which ferromagnetic electrodes of permalloy (Ni-Fe) are used to drive a spin-polarized current into copper. This allows control of spin-polarized currents in solid-state devices at room temperature through the use of ferromagnetic contacts and marks a significant development towards 'all-metallic' electronic devices, such as diodes and transistors, as an alternative to conventional semiconductor devices.

REFERENCES

1. M. H. Kryder (1996) *MRS Bull.* 21 (9), 17.
2. E. Grochowski and D. A. Thompson (1994) *IEEE Trans. Magn.* 30, 3797.
3. S. Onodera, H. Kondo and T. Kawana (1996) *MRS Bull.* 21 (9), 35.
4. D. Jeanniot and S. A. Bull (1988) *IEEE Trans. Magn.* 24, 2476.
5. R. W. deBie, S. B. Luitjens, V. Zieren, C. P. G. Schrauwen and J. P. C. Bernards (1987) *IEEE Trans. Magn.* 23, 2091.
6. M. H. Kryder, W. Messner and L. R. Carley (1996) *J. Appl. Phys.* 79, 4485.
7. J. C. Mallinson (1987) *The Foundations of Magnetic Recording,* Academic Press, San Diego.
8. M. Camras, *Magnetic Recording Handbook,* Van Nostrand, New York.
9. C. D. Mee and E. D. Daniel (1996) *Magnetic Recording Technology,* McGraw-Hill, New York, p. 341.
10. M. F. Doerner and R. L. White (1996) *MRS Bull.* 21 (9), 28.
11. S. Iwasaki and K. Ouchi (1978) *IEEE Trans. Magn.* 14, 849.
12. C. Bonnebat, C. (1987) *IEEE Trans. Magn.* 23, 9.
13. R. M. White (1983) *IEEE Spectrum* 20 (8), 32.
14. R. P. Freese (1988) *IEEE Spectrum* 25 (2), 41.
15. D. S. Bloomberg *et al.* (1996) in *Magnetic Recording Technology* (eds C. D. Mee and E. D. Daniel), McGraw-Hill, New York.
16. J. A. Brug, T. C. Anthony and J. H. Nickel (1996) *MRS Bull.* 21 (9), 23.
17. M. L. Williams and R. L. Comstock (1971) *AIP Conf Proc.* No. 5, 758.
18. H. N. Bertram (1986) *Proc. IEEE* 74, 1512.
19. J. Crangle (1977) *The Magnetic Properties of Solids,* Edward Arnold, London.
20. E. Della Torre and G. Kadar (1988) *J. Appl. Phys.* 63, 3004.
21. I. D. Mayergoyz (1991) *Mathematical Models of Hysteresis,* Springer-Verlag, New York.
22. E. C. Stoner and E. P. Wohlfarth (1948) *Phil. Trans. Roy. Soc. Lond.* **A240**, 599.
23. R. W. Chantrell, K. O'Grady, A. Bradbury, S. W. Charles and N. Hopkins (1987) *IEEE Trans. Magn.* 23, 204.
24. M. Johnson (2000) *IEEE Spectrum* 37 (2), 33.
25. P. S. Anil Kumar and J. C. Lodder (2000) *J. Phys. D (Appl. Phys.)* 33, 2911.
26. J. S. Moodera and G. Mathon (1999) *J. Magn. Magn. Mater.* 200, 248.
27. J. G. Zhu (2000) *J. Appl. Phys.* 87, 6668.
28. F. J. Jedema, A. T. Filip and B. J. van Wees (2001) *Nature* 410, 345.

FURTHER READING

H. N. Bertram (1994) *Theory of Magnetic Recording,* Cambridge University Press, Cambridge.

M. Camras (1988) *Magnetic Recording Handbook,* Van Nostrand, New York.

J. C. Mallinson (1993) *The Foundations of Magnetic Recording,* 2nd edn, Academic Press, San Diego.

C. D. Mee and E. D. Daniel (1996) *Magnetic Recording Technology,* McGraw-Hill, New York.

T. Suzuki (1996) *MRS Bull.* 21 (9), 42.

15 ELECTRONIC MATERIALS FOR TRANSDUCERS: SENSORS AND ACTUATORS

OBJECTIVE

In this chapter we will be concerned mainly with materials that are used in energy conversion devices. The most common are those which convert electrical into mechanical energy and vice versa. These are the electrostrictive and piezoelectric transducers. Most of the discussion of these transducers can also be applied to magnetostrictive transducers. We will look at the different types of materials for these transducers and define their performance parameters. Among these materials the most important class is the ferroelectrics which have high relative permittivities and high strain coefficients. Finally, we look at applications of these materials in microelectronics, such as microelectromechanical machines (MEMs), and data storage.

15.1 TRANSDUCERS

What do we mean by the term transducer?

A transducer is any device that converts one form of input energy into a different form of output energy. In other words it is an energy-conversion device. A common example is a device which converts mechanical energy into electrical energy, such as a piezoelectric transducer. In fact, the original use of the term transducer was specifically for a device which sensed mechanical input energy and converted it into electrical output energy. However, the term is now used to include any device which converts one form of energy to another. The efficiency of a transducer is a useful parameter which measures the ratio of output energy to input energy. Examples of transducers are loudspeakers, ultrasonic vibration generators, thermocouples, microphones and various forms of magnetometer.

Except for some rather minor differences of meaning the terms transducer, sensor and detector are regarded as synonymous. In these transducers the output bears a known relation to the input. Therefore the output can be controlled through the input, or alternatively the input determined by measuring the output. Another way of viewing transducers is simply as a means of interacting between electronic instrumentation and the outside world. In most cases, therefore, we are concerned with devices that act as detectors (sensors) which convert the external energy into an electrical voltage, or emitters (actuators) which convert the electrical voltage into external energy.

15.1.1 Classification of transducers

What terminology is used for the various types of transducers?

The general categories of transducers are shown in Table 15.1.

Type of transducer	Input energy	Output energy
Electrostrictive	Electric	Mechanical
Piezoelectric	Mechanical	Electrical
Electroacoustic	Electrical	Sound
Photoelectric	Light	Electrical
Magnetoelectric	Magnetic	Electrical
Thermoelectric	Thermal	Electrical
Magnetostrictive	Magnetic	Mechanical
Piezomagnetic	Mechanical	Magnetic
Pyroelectric	Infrared radiation	Electrical

Table 15.1 Nomenclature of transducers in terms of their forms of input and output energy.

The nonlinearity of a transducer can be an important factor. Nonlinearity refers to the deviation from a directly proportional dependence of output on input. For example, when the mechanical force produced by an electrostrictive transducer varies with the square of the applied electric field. Ferroelectric and ferromagnetic materials also exhibit a hysteretic relationship between input and output which can be a hindrance in certain transducer applications, particularly for positioning devices.

The most common form of transducers are those which convert to and from acoustic energy. These are used in loudspeakers, microphones and acoustic systems. The three types of acoustic transducer that are widely used are piezoelectric, electrostrictive and magnetostrictive. These acoustic transducers are usually operated at, or close to, mechanical resonance in order to obtain maximum energy conversion efficiency. For wide bandwidth operation the transducer must necessarily be operated at frequencies well away from the resonant frequency. This is achieved in some cases by sandwiching layers of transducer material with plates of metal to reduce the sharpness of the resonance.

There is a wide range of physical sizes for transducers from a few square millimetres and weights of a few grams up to square metres and weights of tens of kilograms. Transducer phased arrays are sometimes used instead of a single transducer in order to steer an acoustic beam in particular directions. In some cases special transducer arrays several metres across and weighing about a ton have been constructed for submarine surveillance applications.

15.1.2 Energy conversion in transducers

What properties must a material have to make it suitable for use as a transducer? In order for a material to be useful as a transducer it must have a high efficiency for converting one form of energy into another. The input energy is usually supplied by an external influence such as an applied stress, an electric field or a magnetic field. The response of the material, in the form of a voltage or a strain, can be measured and used to detect the external field, as in a sensor, or to cause movement, as in an actuator.

The most common forms of transducer materials are piezoelectric, electrostrictive and magnetostrictive materials. Piezoelectric/electrostrictive transducer materials include quartz, ammonium dihydrogen phosphate, tourmaline, lithium sulphate, barium titanate, lead zirconate and lead magnesium niobate, and Rochelle salt (potassium sodium tartrate tetrahydrate, $KNaC_4H_4O_6.4H_2O$). Magnetostrictive transducer materials include nickel alloys and some of the more recent rare earth-iron alloys such as Terfenol (terbium-dysprosium-iron).

15.2 TRANSDUCER PERFORMANCE PARAMETERS

What materials properties are important for transducer applications?
Ferroelectrics are widely used as transducer materials, since they give a high polarization to field strength ratio. These materials need to be polarized or 'poled' in order to ensure that the piezoelectric effect is observed. On the other hand the dependence of strain on electric field is usually relatively high in these materials, although it is also hysteretic, as shown in Fig. 15.1. This is analogous to the variation of bulk magnetostriction with magnetic field, and produces the well-known 'butterfly curve' of strain versus field.

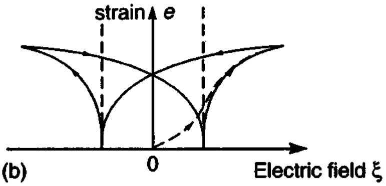


Figure 15.1 Dependence of strain on electric field in a ferroelectric material showing hysteresis. Reproduced with permission of N. Braithwaite and G. Weaver, *Electronic Materials*, published by Butterworths, 1990.

15.2.1 Strain

How do we define electrostrictive strain in these materials?
The strain is given simply as the fractional change in length,

$$e = \left(\frac{\Delta \ell}{\ell_0}\right), \tag{15.1}$$

where ℓ_0 is the original length and $\Delta \ell$ is the change in length. The length ℓ_0 is usually taken as the length in the depolarized state (or in magnetostrictive materials the demagnetized state) for convenience and the changes are therefore measured by convention relative to the depolarized (or demagnetized) state. The above equation is itself merely a definition and tells us nothing about the performance of the materials or how the strain varies with field or polarization. The strain derivatives, $de/d\xi$ for electrostrictive/piezoelectric materials and de/dH for magnetostrictive materials, are rather more meaningful representations of materials properties, since they tell us how rapidly the strain changes with the relevant applied field.

15.2.2 Motor coefficient (strain derivative)

What does the derivative of strain with respect to field tell us about the suitability of a material for transducer applications?

The derivative of the strain e with respect to the applied electric field ξ is one of the most useful parametres for quantifying the performance of a transducer material. This is known as the motor coefficient. For piezoelectric/electrostrictive materials, this is

$$d = \left(\frac{\mathrm{d}e}{\mathrm{d}\xi}\right)_{\sigma}, \tag{15.2}$$

and is measured in $\mathrm{V^{-1}\,m}$ (or equivalently $\mathrm{C\,N^{-1}}$). For magnetostrictive materials it is the derivative of strain with respect to the magnetic field H,

$$d = \left(\frac{\mathrm{d}e}{\mathrm{d}H}\right)_{\sigma}, \tag{15.3}$$

and is measured in $\mathrm{A^{-1}\,m}$. When changes expressed by differentials are small and reversible these d coefficients are also equal to $(\mathrm{d}D/\mathrm{d}\sigma)_{\xi}$ and $(\mathrm{d}B/\mathrm{d}\sigma)_{H}$, respectively, where D is the electric flux density, B is the magnetic flux density and σ is the stress. This is also sometimes known as the strain coefficient, strain derivative, or simply as the '*d* coefficient'. It is a useful measure of the performance of a material since as we shall see it is closely related to the energy transfer efficiency of the material. It is often quoted in the range of picometres per volt for piezoelectric/electrostrictive materials or in nanometres per amp in magnetostrictive materials. Typical values of the d coefficient for various materials are: lead zirconate titanate $5 \times 10^{-10}\,\mathrm{m\,V^{-1}}$, ammonium dihydrogen phosphate $5 \times 10^{-11}\,\mathrm{m\,V^{-1}}$, lead magnesium niobate $1.5 \times 10^{-9}\,\mathrm{m\,V^{-1}}$, iron $\sim 3 \times 10^{-9}\,\mathrm{m\,A^{-1}}$, and terbium dysprosium iron $30 \times 10^{-9}\,\mathrm{m\,A^{-1}}$.

At the origin of the hysteresis curve, that is in the unpolarized state, the strain derivative is very small, as can be seen by reference to Fig. 15.1. This condition is not very useful as a transducer. However, the material can be used in its 'poled' (magnetized) condition whereupon the strain derivative will be considerably larger and the strain will remain almost linear with change in field for a range of field strengths. In active devices it is also possible to use the material under a biased electric field which can be adjusted to find the maximum of $\mathrm{d}e/\mathrm{d}\xi$ or $\mathrm{d}e/\mathrm{d}H$.

Most piezoelectric transducers are ferroelectrics which are 'poled' to produce a remanent polarization and then are operated over a range of electric fields which is small enough to ensure that depolarization does not occur. The exceptions are the single crystal piezoelectrics which are not ferroelectric and therefore cannot be poled.

15.2.3 Generator coefficient

How is the dependence of polarization on strain quantified?

When using a piezoelectric material to generate a voltage by application of a stress the parameter of interest is the rate of change of electric field with stress. This is called the generator coefficient g since it generates a voltage from applied stress at constant electric flux density D

$$g = -\left(\frac{\mathrm{d}\xi}{\mathrm{d}\sigma}\right)_D = \frac{1}{\varepsilon_0 \chi_e}\left(\frac{\mathrm{d}P}{\mathrm{d}\sigma}\right)_D . \tag{15.4}$$

This is measured in $\mathrm{V\,N^{-1}\,m}$, (or equivalently $\mathrm{m^2\,C^{-1}}$). It is also equal to the derivative of the polarization P with respect to stress at constant electric flux density D as shown. The analogous coeeficient for a magnetostrictive transducer is the derivative of the magnetic field H with respect to stress,

$$g = -\left(\frac{\mathrm{d}H}{\mathrm{d}\sigma}\right)_B = \left(\frac{\mathrm{d}M}{\mathrm{d}\sigma}\right)_B . \tag{15.5}$$

This is measured in $\mathrm{A\,N^{-1}\,m}$. It is also equal to the derivative of the magnetization M with respect to stress at constant magnetic flux density B, as shown.

When the changes represented by the differentials are small and reversible these g coefficients are also equal to $-(\mathrm{d}e/\mathrm{d}D)_\sigma$ and $-(\mathrm{d}e/\mathrm{d}B)_\sigma$ respectively, where D is the electric flux density and B is the magnetic flux density. This is also sometimes known as the piezoelectric coefficient or polarization coefficient. It is often quoted in the range of millivolt metres per newton. Typical values of g for different materials are; lead zirconate titanate $30 \times 10^{-3}\,\mathrm{V\,m\,N^{-1}}$, 'PXE' (a commercial form of lead zirconate titanate) $10 \times 10^{-3}\,\mathrm{V\,m\,N^{-1}}$ and potassium sodium niobate $20 \times 10^{-3}\,\mathrm{V\,m\,N^{-1}}$.

15.2.4 Energy coupling coefficient

How can the energy conversion efficiency be represented?

One of the most useful parameters for a transducer, whether it is electrostrictive or magnetostrictive, is its energy conversion efficiency. This can be expressed via the square of the coupling coefficient k^2 which is defined as

$$k^2 = \frac{\text{energy output}}{\text{energy input}} . \tag{15.6}$$

In the case of an electrostrictive transducer the value of k^2 is given by

$$k^2 = \frac{E_Y}{\varepsilon_0 \varepsilon_r} d^2 , \tag{15.7}$$

where d, the strain derivative, has been defined in Section 15.2.2. In the case of a magnetostrictive transducer k^2 is given by

$$k^2 = \frac{E_Y}{\mu_0 \mu_r} d^2 . \tag{15.8}$$

However, it must be remembered that in both ferroelectric and ferromagnetic materials, ε_r and μ_r are not constants. This means that the energy conversion efficiency changes with applied field strength. Therefore it is not a material constant.

15.3 TRANSDUCER MATERIALS CONSIDERATIONS

What considerations determine the choice of a particular material for a transducer?
The first problem in fabricating a transducer for a particular application is to find a material which exhibits the right effect for converting the input energy into the required output energy. Once this has been achieved it is usually necessary to modify the material itself to optimize its performance. This means enhancing the desired properties to increase the energy conversion efficiency. This can be achieved in some cases by adjusting the chemical composition. For example, in recent magnetostrictive transducer materials the anisotropy has been reduced by a suitable choice of alloying components leading to an increase in the rate of change of strain with applied field.

15.3.1 Piezoelectricity

Can a voltage be induced in a material by the application of stress?
The piezoelectric effect is the production of an electric charge on the surface of a material, and hence a voltage across the material, as a result of the application of stress. The piezoelectric effect, therefore, allows a conversion of mechanical energy into electrical energy through a material transducer.

This occurs in materials where the application of stress causes a change in electric polarization by separating the centres of positive and negative charge in the crystal. In crystalline materials the piezoelectric effect only occurs in a limited class of materials of low crystal symmetry in which the application of stress deforms the crystal structure and leads to the generation of an electric dipole moment as shown in Fig. 15.2. These materials necessarily have a crystal structure which lacks a centre of symmetry [1, p. 273].

The converse effect also occurs: when an electric field is applied to a piezoelectric material a strain is produced. This is somewhat similar to electrostriction (see Section 15.3.2), but in this case the strain is antisymmetric with respect to the electric field, which means that it is proportional to odd powers of the field strength. Therefore, the strain in a piezoelectric material changes sign when the

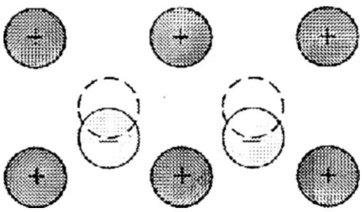


Figure 15.2 Noncentrosymmetric crystal classes can exhibit piezoelectricity because the application of a stress changes the separation between positive and negative ions leading to a net polarization.

electric field is reversed in direction. In the majority of cases, piezoelectric materials exhibit greater strain under the action of an electric field than conventional electrostrictive materials, although some of the ferroelectric electrostrictive materials exhibit comparable strains to piezoelectric materials.

An important application of piezoelectricity is in the quartz crystal resonator in which the strain amplitude can become very large when the applied ac voltage signal coincides with the mechanical resonance of the quartz crystal. Similarly, when the frequency of mechanical excitation coincides with the resonant frequency a large electrical signal is produced.

15.3.2 Electrostriction

What happens to the shape of a dielectric when it is subjected to an electric field?
All dielectric materials undergo a strain when subjected to an applied electric field. It results in a slight change in shape. This change in length is termed electrostriction.

The exact definition of electrostriction is somewhat more restricted than simply the strain produced by an applied electric field. In order to distinguish it from being the converse of the piezoelectric effect, the term 'electrostriction' is sometimes used to refer only to those strains which do not change sign if the direction of the electric field is reversed.

Some materials, such as the ferroelectrics, can exhibit either piezoelectricity or electrostriction depending on the conditions under which they are operated. In fact, electrostriction always occurs in dielectric materials under the action of an electric field, but when both types of strain occur together the electrostrictive strains are much smaller than the piezoelectric strains. In these cases we simply refer to the material as piezoelectric.

Compared with the antisymmetric strains induced in the inverse piezoelectric effect the symmetric electrostrictive strains are, in most cases, typically $\lambda \approx 10^{-4}$ [2]. In some ferroelectrics, however, the electrostrictive strains can be much larger, for example $\lambda \approx 10^{-2}$ to 10^{-3} in lead zirconate titanate [3,4] or lead magnesium niobate [5]. This latter material, which at lower temperatures is a ferroelectric, is usually operated above its Curie temperature in transducers.

15.3.3 Magnetostriction

What happens to the shape of a magnetic material when it is subjected to a magnetic field?
Magnetostriction is a property of materials which change length when they are magnetized either spontaneously by virtue of a magnetic phase transition (spontaneous magnetostriction) or under the action of a magnetic field (field-induced magnetostriction). This effect is only significant in magnetically ordered materials such as ferromagnets, ferrimagnets and antiferromagnets. Typical magnetostrictive strains of magnetic materials are of the order of $\lambda \approx 10^{-5}$ to 10^{-6} [6], although more recent advanced materials, such as terbium dysprosium iron, have

magnetostrictions as high as 10^{-3} [7, 8]. If the ferromagnet is demagnetized the magnetostrictive strains are symmetric with the applied field, but if the ferromagnet is in a remanent magnetized state the magnetostrictive strains are antisymmetric with the applied field.

Much progress has been made in recent years in the development of magnetostrictive transducers based on giant magnetostrictive materials such as Terfenol-D ($Tb_xDy_{1-x}Fe_2$; $x < 1$). Reliable model descriptions of these materials have been developed at the same time, and these can now provide valuable predictive simulations of transducer performance. Calkins *et al.* [9] have developed a magnetostrictive transducer model based on energy considerations which gave very good agreement with performance data of Terfenol-D transducers. The model described the B, H hysteresis loop of the material, and then incorporated a quadratic relationship between magnetostriction and magnetization. In this way, a hysteretic relationship between magnetostrictive strain and applied magnetic field was obtained. The model was able to describe strain under both major and minor loop excursions of magnetization versus magnetic field. The simulator was then used as part of a control system for the transducer in which the current to the coil was the input control parametre. An important factor in the use of this model for control applications is that the speed of computation is fast when compared with other hysteretic methods, allowing rapid response and thereby operation at higher frequency. The model is also compact because it has only a small number of model parametres and can be operated under a wide variety of conditions including nonlinear regimes.

Dapino *et al.* [10] have shown how to include nonlinear and hysteretic effects directly into the transducer equation for strain. In this case, both the elastic compliance and the piezomagnetic coefficient d were allowed to be variable and path dependent. This generalization of the magnetostrictive model is suitable for describing the behaviour of magnetostrictive transducers at high drive levels by providing a more complete description of the relationship between input current to the coil and output strain of the transducer.

15.3.4 Piezomagnetism

Is there an inverse effect to magnetostriction?

Under certain conditions the application of stress to a magnetic material can cause a change in magnetization. In unmagnetized materials this is the magnetic analogue of the piezoelectric effect in unpolarized dielectric materials. Only a few examples are known of magnetic materials which can change from a demagnetized to a magnetized state under the action of a stress. One of these is the antiferromagnet CoF_2 in which a small magnetization of the order of 10^3 A/m can be produced by large shear stresses [11].

A related and much more common phenomenon is the tendency of magnetic materials to change their magnetization under the action of an applied stress when already magnetized [12]. This effect is only of significant size in ferromagnets and is rarely employed in applications. There are no known cases of ferromagnets which become magnetized under the action of a stress on the unmagnetized state.

15.3.5 Mechanism of piezoelectricity

Why does a voltage change arise in a piezoelectric as a result of stress?

The phenomena of piezoelectricity and electrostriction are of great practical importance in electromechanical transducers. Consider a piezoelectric material which elongates along the direction of polarization. If this material is subjected to a compressive stress along the axis of polarization, and if we consider only reversible processes, the material can best respond to the compressive stress by polarizing itself at right angles. This will minimize the strain energy of the system and leads to ionic displacements at right angles to the original direction of polarization and compressive stress. The result is the generation of a voltage along the direction perpendicular to the stress where there was previously no voltage. Therefore the compressive stress has generated a change in voltage. This is the piezoelectric effect.

The inverse effect is the change in strain along the axis of polarization caused by the application of an electric field. Suppose an electric field is applied perpendicularly to the direction of polarization. This will eventually result in a rotation of polarization into the field direction, and this causes a change in strain along the direction of the field. By the Poisson effect there will also be a change in strain at right angles to this direction.

15.3.6 Comparison of electrostriction and magnetostriction

Are electrostriction and magnetostriction analogous effects under the action of electric and magnetic fields, respectively?

According to the conventional definition, electrostrictive strain is dependent on even powers of the electric field strength (i.e. symmetric with field strength). This definition is suitable for most dielectrics, but can cause a problem in ferroelectrics, (see Section 15.4) because of hysteresis in the relationship between polarization and electric field. If we wish to maintain the analogy with magnetostriction, the electrostriction should be defined as a strain which is invariant under the reversal of the direction of polarization, because the magnetostrictive strain can be either symmetric or antisymmetric with field strength, depending on whether the magnetic material was already magnetized.

The analogy with magnetostriction is clear. In many cases, ferromagnetic magnetostrictive transducers and actuators are operated in a 'biased' condition which is equivalent to the 'poled' condition of a ferroelectric electrostrictive transducer or actuator. The objective of biasing (poling) is merely to find a point on the magnetostriction (electrostriction) curve at which the strain derivative is greater than at the unmagnetized (unpolarized) state. The dependence of strain on magnetization (polarization) remains symmetric despite the biasing. However, the dependence of strain on magnetic field (electric field) becomes antisymmetric once the material has been biased (poled). In magnetic materials the phenomenon is referred to as magnetostriction under both conditions as there has not been any fundamental change in the nature of the material or its response to the applied field.

15.3.7 Piezoelectric and piezomagnetic response times

What is the relative speed of response in piezoelectric electrostrictive and magnetostrictive materials?

Generally, the speed of the piezoelectric response is much faster than the analogous response of a magnetic material to field or stress. Piezoelectric transducers can operate at frequencies of up to tens of megahertz whereas magnetostrictive transducers are restricted to a few kilohertz. This is because most highly electrostrictive materials are insulators, while most highly magnetostrictive materials are conductors. In this latter case, the generation of eddy currents at higher frequencies limits the penetration of the alternating magnetic field into the material.

15.4 FERROELECTRIC MATERIALS

Can materials be spontaneously polarized?

Ferroelectricity is a property of certain dielectrics which exhibit electrical polarization in the absence of an applied electric field. On the scale of a few atoms ferroelectrics consist of structural units with tiny localized electric dipoles. These dipoles are grouped locally within domains which have spontaneous polarization, as shown in Fig. 15.3, in the same way that a ferromagnet is spontaneously magnetized within domains. The bulk polarization can nevertheless be zero, since the polarizations of all the domains can sum vectorially to zero.

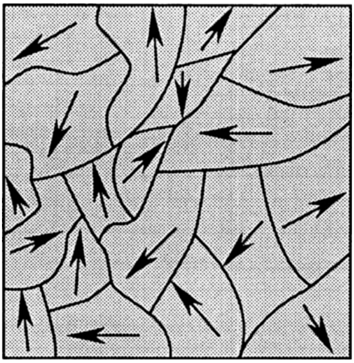


Figure 15.3 Schematic diagram of electric domains with spontaneous polarization within a ferroelectric material. The arrows show the directions of the electric polarization vectors.

Ferroelectric materials have a macroscopic electric polarization P which is both large and nonlinear as a function of electric field ξ. In particular, the variation of P with ξ exhibits hysteresis as shown in Fig. 15.4. The name ferroelectric was derived from the name ferromagnet because both show hysteresis, not because of any association between ferroelectrics and the element iron. The early theory of

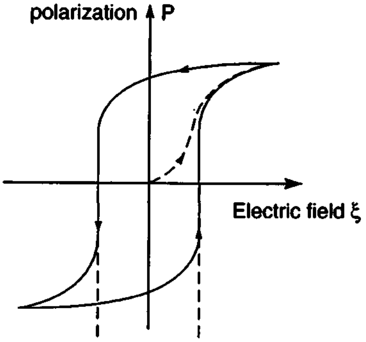


Figure 15.4 Hysteresis in the dependence of electric polarization on electric field in ferroelectrics.

ferroelectricity was analogous to the classical Weiss mean field theory of ferromagnetism. Examples of ferroelectric materials include barium titanate, lead zirconate titanate, Rochelle salt (potassium sodium tartrate tetrahydrate) and zinc oxide. Typically, ferroelectrics have a high relative dielectric constant (relative permittivity) ε_r. A good example is barium titanate $BaTiO_3$. The variation of its relative permittivity with temperature is shown in Fig. 15.5. This can reach a value of several thousand just below the Curie point.

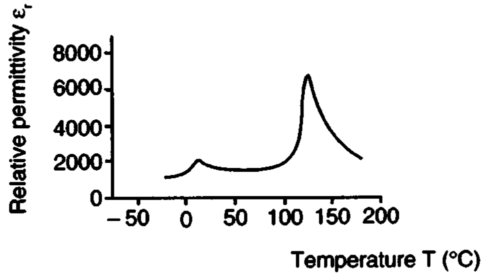


Figure 15.5 Variation of relative permittivity of barium titanate with temperature.

15.4.1 Electric polarization in dielectric materials

How can an electric polarization be produced in a material ?

An electric polarization can be produced in certain types of materials in two ways. The polarization can arise spontaneously within domains, so that the material forms a low symmetry structure at temperatures below a transition temperature known as the Curie point. Alternatively, the polarization can be caused by the application of stress, in which the electric charges in a noncentrosymmetric crystal can be displaced by the stress. In the first case we have a ferroelectric material, in the second case a piezoelectric material.

15.4.2 Depolarized ferroelectrics

Under what conditions does a ferroelectric have the largest differential permittivity in zero electric field?

When a ferroelectric with high relative permittivity ε_r is used in a capacitor it is used in the depolarized condition. From this state the differential permittivity ($= dP/d\xi$) is larger than in the remanent polarized condition and therefore for a given amplitude of electric field ξ the change in polarization P is greater since $P = \varepsilon_0 \chi_e \xi$.

15.4.3 Domains and domain walls

How is the polarization process explained in terms of changes in the domain configuration?

Ferroelectric materials consist of a number of localized regions or volumes in which the electric polarization vectors are aligned parallel, but from one region to the next the directions of the vectors change. These volumes, within which the polarization vectors are parallel, are known as domains. The regions at the boundaries of the domains are known as domain walls. In barium titanate, for example, the difference between directions of polarization on either side of the domain walls can be 180° or 90° because the crystallographic directions of polarization are restricted, by crystalline anisotropy, to be along the [100] family of directions as shown in Fig. 15.6.

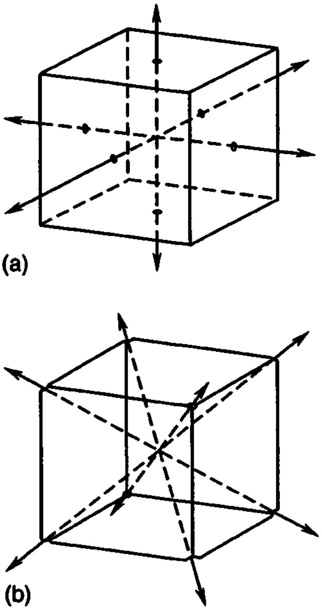


Figure 15.6 Tetragonal and rhombohedral structures both of which occur in barium ferroelectrics. In the tetragonal phase the ferroelectric domain walls must be either 90° or 180°. In the rhombohedral structure the domain walls can be either 180°, 71° or 109°.

The high relative differential permittivity of depolarized ferroelectrics arises from the motion of 180° domain walls, which are easy to move. At higher polarizations it becomes necessary to move more 90° walls which require higher energy because of the associated change in strain. This leads to a lower differential permittivity.

Since the directions of polarization have associated with them a spontaneous strain, the strain energy is a very significant factor in determining the energy of the domain wall. Consequently, it is relatively easy to move a 180° domain wall under the action of a field because it does not result in any change in strain. However, 90° walls are difficult to move because successive regions of the material need to change shape to accommodate the movement. This requires higher energy to overcome obstructions to the necessary deformation from neighbouring grains which must be strained to accommodate the local change in shape. These strains can even be large enough in some cases to cause plastic deformation or even cause the material to fracture.

15.4.4 Paraelectric phase

What happens to ferroelectrics at higher temperatures?

At the Curie temperature and above, the ferroelectric domains are destroyed and the material becomes 'paraelectric'. The Curie point itself depends on chemical composition of the material. In the paraelectric phase polarization is induced by application of an electric field, but when the field is removed it reverts to the unpolarized condition. In this phase, the electric dipole vectors are disordered in the absence of an applied field. This is analogous to the behaviour of the magnetization of a paramagnet when subjected to a magnetic field.

15.4.5 Ferroelectric phase

What happens when the disordered paraelectric is cooled below its Curie temperature?

Above the Curie temperature, since the electric dipoles are randomly oriented, no electric domains can exist. Once the material has been cooled below its Curie temperature it becomes ferroelectric with localized spontaneous polarization within domains as shown in Fig. 15.3. The ferroelectric phase is characterized by a very high relative permittivity ε_r. All ferroelectric materials exhibit a paraelectric phase at higher temperatures.

15.4.6 Antiferroelectric phase

Is there an ordered polarized phase that is analogous to antiferromagnetism with antiparallel alignment of polarizations?

Compositions within 10% of pure $PbZrO_3$ are orthorhombic and below the Curie temperature are antiferroelectric, which means that the neighbouring unit cells of the crystal have polarizations in opposite directions. This is analogous to the antiferromagnetic phase in magnetic materials.

15.5 FERROELECTRICS AS TRANSDUCERS

Which materials are most widely used as transducers?

Ferroelectric materials are particularly useful in transducers. All insulators, whether ferroelectric or not, exhibit electrostriction to some degree. This results directly from the polarization of the material. Since ferroelectrics give much higher polarizations for a given field strength we might expect that these materials will, on the whole, exhibit higher electrostrictive strains. This, of course, is not a totally general result, but the materials with the highest electrostrictions are also ferroelectrics. These materials, such as barium titanate, are therefore widely used as transducers.

Other ferroelectric materials such as lead zirconate titanate and related materials (PZT), lithium niobate, lead germanate, potassium dihydrogen phosphate, strontium barium niobate have ranges of Curie temperature up to 1500 K and spontaneous polarizations of up to 4 coulombs per square metre.

15.5.1 Piezoelectricity in ferroelectrics

Does a ferroelectric material exhibit electrostriction or piezoelectric strain under the action of an electric field?

In the 'poled' state the action of a stress on a ferroelectric produces a change in electric voltage. These materials can therefore act as generators of electric field. The 'poled' state of the material is one in which it has been subjected to an electric field and left with a remanent polarization. This is equivalent to remanent magnetization in a ferromagnet. The applied stress causes a rearrangement of the electric domains. This changes the bulk polarization leading to a change in voltage across the material.

Note that while all ferroelectrics can exhibit piezoelectricity (according to the conventional definition (see Section 15.3.6)), not all piezoelectric materials are ferroelectrics. For example, quartz is piezoelectric but not ferroelectric.

15.5.2 Polycrystalline transducer materials

How does the energy conversion efficiency vary from single crystal to polycrystalline materials?

The above discussion of electrostriction and piezoelectricity needs to be modified in the case of polycrystalline materials because the change in shape of the grains is opposed by their neighbouring grains. These neighbouring grains will not, in general, need to reorient, or change shape, at the same field strength. The mechanical work done internally as a consequence of these strains results in lower energy conversion efficiencies.

The use of ferroelectric materials has allowed polycrystalline materials to be employed as transducers. In the early days of piezoelectricity the connection between piezoelectricity and crystal symmetry was so strongly established that the analogy between electrostrictive strain of a ferroelectric and the magnetostrictive strain of a ferromagnet was difficult to recognize. Therefore the piezoelectrics which received attention were crystalline materials without centres of symmetry.

Once it was established that the polarity needed to impart piezoelectric properties could be achieved by the temporary application of a strong electric field to ferroelectric materials, a whole range of new polycrystalline piezoelectric electrostrictive materials became available [13].

15.5.3 Ageing

Does the bulk polarization of a ferroelectric change with time?

A very slow change in remanent polarization occurs with time in ferroelectrics. This results in a decay of remanence through relaxation of the material to its global energy minimum which in zero field occurs at zero bulk polarization *P*. As in ferromagnets the magnetic domain walls are pinned by defects and this increases both the coercivity and the ageing rate of the material by making changes in polarization arising from domain wall motion more difficult.

15.5.4 Lead zirconate titanate (PZT)

What is the most widely used ferroelectric material?

An important category of ferroelectrics is based on lead, zirconium and titanium oxides and is known as 'PZT'. These may be considered the archetypal ferroelectric materials. Suitable adjustment of the chemical composition and microstructure of these materials can result in a wide range of possible properties suitable for many technological needs. In this respect PZT plays a role in ferroelectrics that is similar to the role of iron in ferromagnets.

The perovskite crystal structure which occurs in barium titanate also occurs in the alloys of the system $PbZrO_3$–$PbTiO_3$, lead zirconate titanate. By varying the

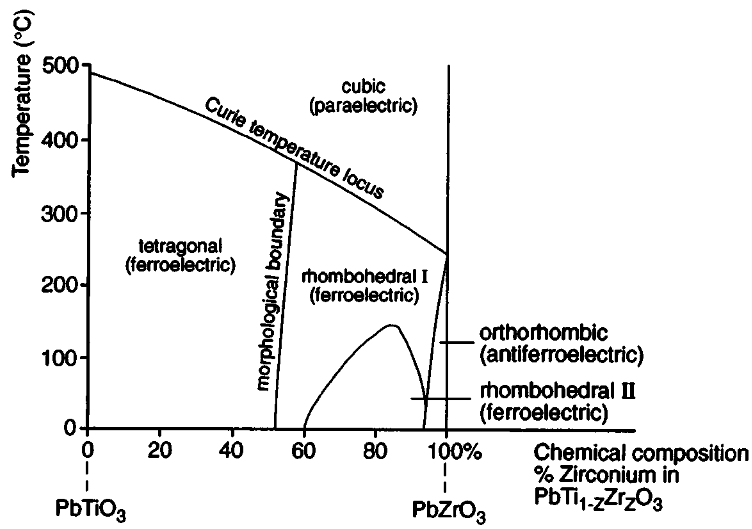


Figure 15.7 Structural phase diagram of PZT–lead zirconate titanate. Reproduced with permission of N. Braithwaite and G. Weaver, *Electronic Materials*, published by Butterworths, 1990.

ratio of Zr to Ti the properties can be engineered to meet many transducer specifications. The phase diagram of PZT is shown in Fig. 15.7.

15.5.5 Chemical additions to PZT

How can the piezoelectric properties of PZT be controlled by selection of chemical composition?

The PZT class of ferroelectrics are very important because of the diversity of properties that they can display. The materials properties can be engineered by chemical additions and control of microstructure. This diversity of ferroelectric properties makes the material very versatile and of crucial importance to the subject of ferroelectricity. The properties of PZT can be strongly influenced by the addition of other metal oxides. These are used to manipulate properties such as conductivity, electric coercivity and elastic modulus. In the PZT compounds the energy conversion efficiency also depends on the chemical composition. If the composition is close to the rhombohedral-tetragonal transition which occurs at about 50%Zr-50%Ti then the energy conversion efficiency rises to about 25% ($k = 0.5$), from values in the range of 2% ($k = 0.15$) on either side of this composition, as shown in Fig. 15.8.

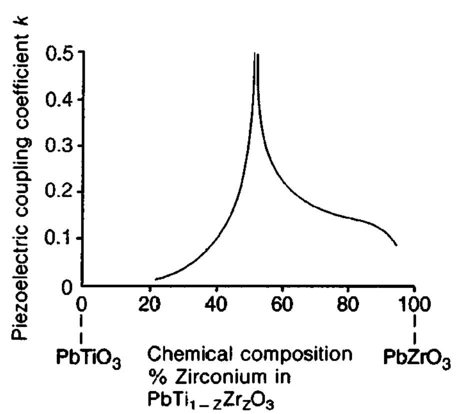


Figure 15.8 Variation of the coupling coefficient *k* of PZT with chemical composition. Reproduced with permission of N. Braithwaite and G. Weaver, *Electronic Materials*, published by Butterworths, 1990.

The reason for this high conversion efficiency at 50%Zr-50%Ti composition is that the grains of the material can find shear transformations between both rhombohedral and tetragonal crystal classes because the energy is finely balanced between these two crystal structures at this composition. These shear transformations therefore occur as well as the conventional electric polarization rotations within either the rhombohedral or tetragonal crystal grains. This allows the rotations to proceed more easily via two mechanisms and results in smaller internal energy losses.

15.5.6 Ferroelectric thin films for microelectronic applications

Can ferroelectric materials be combined with semiconductors to produce microelectronic devices?

There is an increasing need to integrate ferroelectric materials with semiconductors in devices in order to take advantage of the wide range of dielectric, piezoelectric and electrostrictive properties of ferroelectrics. For example, ferroelectrics can be used to fabricate microcapacitors for dynamic random access memories (DRAM) and nonvolatile ferroelectric random access memories (NVFRAM). The high dielectric permitivities of perovskite materials such as lead zirconate titanate (PZT) can be used to make DRAMs with higher storage density and faster access times, while the large values of remanent polarization of these materials are suitable for NVFRAMs. The high piezoelectric coefficients of some ferroelectrics are suitable for microelectromechanical machines (MEMs) in which both electronic and microscopic mechanical functions are integrated onto the same substrate. MEMs embrace a wide range of functions including accelerometers, displacement transducers, micropumps, pressure sensors and microactuators.

These materials can be produced by a variety of thin film fabrication techniques such as sputter deposition, pulsed laser deposition, chemical vapour deposition, plasma-enhanced metal-organic chemical vapour deposition and solution deposition [14]. In many cases lead zirconate titanate (PZT) has been the material of first choice for prototype NVFRAMs and barium strontium titanate has been used for DRAMs. A variety of ferroelectric materials can be used for MEMs, although PZT and related materials are the leading candidates.

15.5.7 Ferroelectrics for data storage

What are the prospects for using ferroelectric media instead of magnetic media for recording applications?

In much the same way that ferro- and ferrimagnetic materials have been used for data storage, ferroelectrics could also be used in principle, and in fact they have come under consideration for this from time to time. In the past, considerable effort has been devoted to this but the power input needed to obtain fast switching speeds has been a limitation. More recently there has been renewed interest by Chikarmane *et al.* [15] and Scott *et al.* [16]. It seems from this work that thin films of lead zirconate titanate may have applications in dynamic random access memories because of the large areal charge storage density ($19.6\ \mu C/cm^2$) and low leakage current density ($1.32 \times 10^{-7}\ A/cm^2$). The remanent polarization of lead zirconate titanate is typically $10\ \mu C/cm^2$.

For nonvolatile random access memories Ramesh *et al.* [17] have studied heterostructures composed of layers of lead zirconate titanate (PZT) and yttrium barium copper oxide (YBCO). The principal advantage of ferroelectrics over ferromagnetic materials for data storage lies in the high switching speeds which can be achieved. Ferroelectrics can respond to applied fields at frequencies up to the megahertz range, while magnetic materials, particularly electrically conducting magnetic materials, can only respond at kilohertz frequencies. On the other hand, the access time for data storage and retrieval is not limited by the response time of the storage medium alone, and this means that other factors limiting the response time need to be improved before the speed of the magnetic material itself becomes the critical parameter.

15.5.8 Dynamic random access memory (DRAM)

What advantages do ferroelectric materials have over conventional semiconductors for dynamic random access memory?

DRAM forms a large market segment of the computer industry that is viewed as a technology leader for semiconductor devices because of its crucial impact on computer performance. There are significant potential benefits that can be achieved through the incorporation of thin film ferromagnetic materials into DRAM. In the last ten years, there has been substantial progress in the integration of ferroelectrics such as barium strontium titanate into high-density prototype DRAM devices. The main objective is to increase storage densities and reduce access times through the use of ferroelectric materials.

Ferroelectric memory devices can be produced in perpendicular arrays of rows and columns with a ferroelectric capacitor located at each crossing point. Each capacitor will, therefore, represent one digital bit of data in a structure that is analogous to the magnetic core memories used in early computers. The data storage density increases as the area of an individual capacitor in DRAM decreases, being 256 Mbit for devices with individual capacitor areas of $0.4\,\mu m^2$ and 1 Gbit for devices with capacitor areas of $0.2\,\mu m^2$. Using standard low dielectric constant materials, it is difficult to increase the capacitance per unit area, as is necessary for increased data storage densities. There are difficulties in getting beyond $9\,fF/\mu m^2$ using the standard materials such as silicon dioxide or silicon dioxide/silicon nitride. However, high permitivity materials can allow the capacitance per unit area to be enhanced. Two materials currently under consideration for this application are tantalum oxide and barium strontium titanate (BST). Of these, BST offers the higher potential. The highest capacitance per unit area reported to date for a BST-based DRAM is $145\,fF/\mu m^2$ in a 2-nm thick film with a relative permittivity of 325 [18].

15.5.9 Non-volatile ferroelectric random access memory (FRAM)

What advantages do ferroelectric materials have over magnetic materials for 'nonvolatile' random access memory?

The main classes of materials currently used for nonvolatile memory are the ferromagnets and ferrimagnets. Ferroelectric materials provide an alternative means for producing nonvolatile memories for computers. In ferroelectric memories, the read/write times of 1–35 ns are much faster than existing electronically erasable programmable ROMs (EEPROMs) which take 120–150 ns for read/write operations [19]. Nonvolatile FRAMs have been produced in fully functional 256-kbit devices using standard CMOS methods. These incorporated one transistor and one capacitor per data cell with 32K × 8 cells operating at 3 V. The feature sizes were $1.2\,\mu m$.

The main materials used in NVFRAM are PZT and strontium-bismuth-tantalate (SBT). However, thin film ferroelectrics can encounter some performance problems as a result of leakage currents. These can limit the endurance time of the memory meaning that it is not truly nonvolatile. PZT suffers from this problem, but SBT appears to have much less difficulty. On the other hand, SBT has a lower remanent polarization than PZT and this is disadvantageous. Polarization 'fatigue' is another factor that has to be taken into account for this application. This is the reduction in polarization that results from repeated read/erase/rewrite operations. It causes a change in the ferroelectric hysteresis loop of the material rendering it progressively less suitable for the application. SBT is superior to PZT in this respect, even when layer thicknesses are less than 100 nm, because it maintains better ferroelectric properties including more stable remanence.

Ferroelectric models of these materials are needed in order to understand their properties better and to use in computational materials simulators. In particular, there is a need to be able to describe the P, ξ (polarization, electric field) hysteresis loop characteristics of these materials and how they evolve with time or ageing of the material. Such models would allow property predictions and simulations of device performance which could be used to improve device designs. Three areas are currently in need of investigation: (i) phenomenology of the dependence of remanent polarization on coercivity, dielectric coefficients and capacitance; (ii) dependence of electrostatic breakdown field on thermal conductivity and specific heat; and (iii) dependence of switching time on coercivity. The above require the development of detailed microscopic models in order to better understand the behaviour of the materials. Such modeling will need to take into account microstructure (grain size and texture), depletion effects, ferroelectric domain wall motion and will even need to consider the nature of the electrodes and interfaces in the final analysis.

15.5.10 Microelectrical mechanical machines (MEMs)

How can electrostrictive/ferroelectric materials be used in microscopic devices with moving parts?

MEMs, which combine traditional electronic functions with mechanical functions on a single semiconductor chip, can be produced with device dimensions in the range of 10 μm–10 mm. These devices are aimed towards full integration of sensing and actuating functions on a common substrate. The materials that are used in MEMs include the usual semiconductor materials Si, SiO_2, Si_3N_4 together now with other classes of materials such as piezoelectrics or ferroelectrics. The performance of ferroelectric materials in electromechanical energy conversion make them suitable for many MEMs device applications. Among the desirable properties of ferroelectrics for MEMs devices, the high piezoelectric coefficient allows large displacements and high forces to be generated for relatively low voltages. Furthermore, the large changes in polarization that are generated by strain in these materials lead to voltage changes that can be relatively easily sensed using a capacitive method.

Poling, or polarizing of ferroelectric materials, enables polycrystalline materials with piezoelectric properties to be produced in thin films instead of the traditional piezoelectric crystalline form. Materials such as lead zirconate titanate are particularly well adapted for thin film MEMs uses.

In the design of MEMs devices incorporating ferroelectric materials it is important to know in advance the ferroelectric properties of the material, particularly the piezoelectric coefficient d which can be obtained from the basic actuator differential equation. In the case of ferroelectric materials it should be remembered that d is a hysteretic variable and, therefore, it is necessary to know how it varies with electrical field strength and field exposure history.

MEM devices are usually fabricated on either silicon or silicon nitride substrates. Silicon suffers from a disadvantage in this respect because in order to remove the high fabrication stresses an anneal at about 1000°C is necessary. This temperature is often incompatible with other materials fabricated on the substrate for MEMs applications. On the other hand, silicon nitride can be produced with much lower levels of stress which do not require annealing [20]. Finally, when considering the economics of fabricating MEMs devices it is worth noting that this is rather a small fraction of the total manufacturing cost. With these devices packaging and testing accounts for about 90% of the total manufacturing costs.

REFERENCES

1. M. E. Lines and A. M. Glass (1977) *Principles and Applications of Ferroelectrics and Related Materials*, Clarendon Press, Oxford.
2. U. Kumar, A. Halliyal and L. E. Cross (1986) *IEEE Sixth Symposium on Applications of Ferroelectrics, Lehigh University, Pennsylvania*, IEEE Press, New York, p. 633.
3. T. Mitsui, I. Tatsuzaki and E. Nakamura (1976) *An Introduction to the Physics of Ferroelectrics*, Gordon & Breach Publishers, London, p. 77.
4. K. Uchino (1986) *IEEE Sixth Symposium on Applications of Ferroelectrics, Lehigh University, Pennsylvania*, IEEE Press.
5. R. E. Newnham, Q. C. Xu, S. Kumar and L. E. Cross (1990) Smart ceramics, *Ferroelectrics* **102**, 259.
6. B. D. Cullity (1971) Fundamentals of magnetostriction, *J. Metals* 23, 35.
7. A. E. Clark (1980) Magnetostrictive rare earth iron compounds, in *Ferromagnetic Materials*, Vol. 1 ed. E. P. Wohlforth, North Holland Publishing, Amsterdam.
8. D. C. Jiles (1990) Development and characterization of the highly magnetostrictive alloy Tb-Dy-Fe for use in sensors and actuators, in *New Materials and Their Applications*, Institute of Physics Publishing, Bristol.
9. F. T. Calkins, R. C. Smith and A. B. Flatau (2000) *IEEE Trans. Magn.* 36, 429.
10. M. J. Dapino, R. C. Smith and A. B. Flatau (2000) *IEEE Trans. Magn.* 36, 545.
11. M. M. Schieber (1967) *Experimental Magnetochemistry*, John Wiley & Sons, New York, pp. 55, 401.
12. D. C. Jiles and D. L. Atherton (1984) *J. Phys. D. (Appl. Phys.)* **17**, 1265.
13. B. Jaffe, W. R. Cook and H. Jaffe (1971) *Piezoelectric Ceramics*, Academic Press.
14. O. Auciello and R. Ramesh, (1996) *MRS Bull.* **21** (6), 21.
15. V. Chikarmane, J. Kim, C. Sudhama, J. Lee, A. Tasch and S. Novak (1992) *J. Electron. Mater.* 21, 503.

16. J. F. Scott, C. A. Araujo, H. Brett Meadows, L. D. McMillan and A. Shawabkeh (1989) *J. Appl. Phys.* 66, 1444.
17. R. Ramesh *et al.* (1992) *J. Electron. Mater.* 21, 513.
18. A. I. Kingon, S. K. Streiffer, C. Basceri and S. R. Summerfelt (1996) *MRS Bull.* 21 (7), 46.
19. J. F. Scott, F. M. Ross, C. A. Paz de Araujo, M. C. Scott and M. Huffman (1996) *MRS Bull.* 21 (7), 33.
20. D. L. Polla and L. F. Francis (1996) *MRS Bull.* 21 (7), 59.

FURTHER READING

C. Z. Rosen, B. V. Hiremath, and R. E. Newnham (1992) *Piezoelectricity*, Key papers in Physics Series, American Physical Society, New York.

R. E. Newnham (1989) Electroceramics, *Rep. Prog. Phys.* 52, 123.

M. E. Lines and A. M. Glass (1977) *Principles and Applications of Ferroelectrics and Related Materials*, Clarendon Press, Oxford.

V. M. Fridkin (1980) *Ferroelectric Semiconductors*, Consultants Bureau, New York.

Proceedings of the Sixth IEEE International Symposium on Applications of Ferroelectrics (1986) Lehigh University, Bethlehem, Pennsylvania, IEEE Press.

N. Braithwaite and G. Weaver (1990) *Electronic Materials*, Butterworths, London, Chapter 4.

S. M. Sze (1994) *Semiconductor Sensors*, John Wiley, New York.

J. B. Restorff (1994) Magnetostrictive materials and devices, *Encyclopedia of Applied Physics* 9, 603.

J. Fraden (1993) *Handbook of Modern Sensors: Physics, Designs and Applications*, American Institute of Physics, New York.

SOLUTIONS

Example 1.1 The Wiedemann–Franz law
This states that the thermal and electrical conductivities of metals are related by the equation

$$\frac{K}{\sigma T} = L,$$

where K is the thermal conductivity in $\mathrm{J\,m^{-1}\,K^{-1}\,s^{-1}}$, σ is the electrical conductivity in $\Omega^{-1}\,\mathrm{m}^{-1}$ and T is the temperature in Kelvin. L is the Lorentz number which has the value $2.4 \times 10^{-8}\,\mathrm{J\,\Omega\,K^{-2}\,s^{-1}}$.

Example 1.2 The Hagen–Rubens relation
This relates the optical reflectivity and electrical conductivity of a material. The relation is only valid for longer wavelengths, that is in the infrared region of the spectrum ($\nu < 10^{14}$ Hz, $\lambda > 3\,\mu\mathrm{m}$)

$$R = 1 - 2\sqrt{\frac{4\pi\varepsilon_0 \nu}{\sigma}},$$

where R is the reflectance, which is dimensionless, ν is the frequency in Hz and σ is the electrical conductivity in $\Omega^{-1}\,\mathrm{m}^{-1}$.

Example 1.3 The Dulong–Petit law
This law relates the specific heat capacity of a solid to the number of atoms. It was originally observed that the molar heat capacity of solids was similar for a large number of materials, being typically $25\,\mathrm{J\,mol^{-1}\,K^{-1}}$. Later, this was explained in terms of the vibrational modes of the atoms within the solid. If each atom has three degrees of freedom and behaves as a harmonic oscillator with energy $k_B T$ along each of these degrees of freedom, then the internal energy of the solid will be

$$U = 3Nk_B T,$$

where N is the number of atoms per unit volume, k_B is Boltzmann's constant and T is the absolute temperature. If we consider a mole of the material, then $N = N_0 = 6.025 \times 10^{23}$ = Avogadro's number and the molar heat capacity $C = \mathrm{d}U/\mathrm{d}T$, is

$$\begin{aligned} C &= 3N_0 k_B \\ &= 24.96\,\mathrm{J\,mol^{-1}\,K^{-1}} \end{aligned}$$

Exercise 1.4 Macroscopic properties

These properties are simply the response of material to an external stimulus. Permeability = magnetic induction/magnetic field, conductivity = heat flux/ temperature gradient, elastic modulus = stress/strain, extinction coefficient = decay in light intensity/number of wavelengths, heat capacity = change in temperature/heat transfer.

Exercise 1.5 Empirical laws

Wiedemann–Franz: relates electrical conductivity to thermal conductivity
Hagen–Rubens: relates optical reflectance to electrical conductivity
Dulong–Petit: relates heat capacity to the number of atoms
Curie–Weiss: relates magnetic susceptibility to temperature
The continuum model provides no explanation for these observations

Exercise 1.6 Electronic and lattice contributions to properties

Heat capacity – dominated by lattice vibrations with a small component due to electrons.

Thermal conductivity – in metals dominated by electrons, in insulators dominated by lattice vibrations.

Permeability – determined by electrons.

Example 2.1 Elastic modulus of linear atomic lattice

In this situation, there is an equation for the potential energy with two unknowns and two pieces of information which allow the unknowns to be calculated

$$E_p = \alpha_1 a^{-9} - \alpha_2 a^{-1}.$$

At equilibrium separation a_0 we must have the derivative of energy with respect to displacement equal to zero, $(dE_p/da)_{a_0} = 0$. Therefore

$$0 = -9\alpha_1 a_0^{-10} + \alpha_2 a_0^{-2},$$

with $a_0 = 0.3 \times 10^{-9}$ m. The dissociation energy is the difference in energy between this equilibrium position, and complete separation of atoms at $a = \infty$. Therefore setting $E(\infty) = 0$ and

$$\begin{aligned} E_p(a_0) &= -4\,\text{eV} \\ &= -6.4 \times 10^{-19}\,\text{J} \end{aligned}$$

we obtain,

$$-6.4 \times 10^{-19} = \alpha_1 a_0^{-9} - \alpha_2 a_0^{-1},$$

with $a_0 = 0.3 \times 10^{-9}$ m. This gives two simultaneous equations in the two unknowns. Solving these for α_1 and α_2 gives

$$\begin{aligned} \alpha_1 &= 1.57 \times 10^{-105}\,\text{J m}^9 \\ \alpha_2 &= 2.16 \times 10^{-28}\,\text{J m}. \end{aligned}$$

If we consider the elastic modulus E_Y as the derivative of the applied force F_{app} per unit area with respect to the strain e, then in the case of a linear lattice the concept of 'pressure' is not very meaningful. Suppose we consider the effective cross-sectional area per linear lattice chain to be a_0^2 (i.e. equivalent to a simple cubic lattice), then the stress becomes equal to F/a_0^2, and therefore the elastic modulus is

$$E_Y = \frac{1}{a_0^2}\left(\frac{dF_{app}}{de}\right),$$

and since the strain e can be represented as $(a - a_0)/a_0$ where a_0 is the equilibrium separation of the atoms, and a is the separation at strain e,

$$\frac{de}{da} = \frac{1}{a_0}.$$

Therefore

$$E_Y = \frac{1}{a_0}\left(\frac{dF_{app}}{da}\right)_{a_0}$$

and $F_{app} = -F_{internal} = dE_p/da$, so

$$E_Y = \frac{1}{a_0}\left(\frac{d^2E_p}{da^2}\right)_{a_0}$$

$$= \frac{1}{(0.3 \times 10^{-9})}(90\alpha_1 a_0^{-11} - 2\alpha_2 a_0^{-3})$$

$$E_Y = 0.213 \times 10^{12}\,\text{Pa}.$$

The force on the lattice is related to the energy $E_p(a)$ by the equation,

$$F = -\left(\frac{dE_p(a)}{da}\right)_{a_1},$$

where $a_1 = 0.99$, $a_0 = 0.297\,\text{nm}$ is the lattice spacing at 1% strain,

$$F = 9\alpha_1 a_1^{-10} - \alpha_2 a_1^{-2}$$

$$= 1.92 \times 10^{-10}\,\text{N}.$$

Example 2.2. Lattice stabilized by electrostatic repulsion

In this case consider the force on a given atom in the linear lattice due to its interactions with its two nearest-neighbours. If x_o is the equilibrium separation, Δx is the displacement of one atom from equilibrium and q is the charge per ion, then the force on that atom will be

$$F(\Delta x) = \frac{q^2}{4\pi\varepsilon_0}\left(\frac{1}{(x_0 + \Delta x)^2} - \frac{1}{(x_0 - \Delta x)^2}\right),$$

which from the previous considerations is known to be equal to the negative derivative of the energy E_p with respect to position x, that is $-dE_p/dx$. The elastic modulus

$$E_Y = \frac{x_0}{A}\left(\frac{dF}{dx}\right)_{x_0},$$

where A is the cross-sectional area. If we assume a simple cubic lattice then the cross-sectional area per linear lattice chain will be x_0^2. Therefore, the elastic modulus is

$$E_Y = \frac{1}{x_0}\left(\frac{dF}{dx}\right)_{x_0}.$$

We have defined the force constant k as

$$k = \left(\frac{dF}{dx}\right)_{x_0},$$

so that in this case $E_Y = k/x_0$, then from the wave equation, Section 2.2.2, the velocity of longitudinal waves v is given by,

$$v = \sqrt{\frac{kx_0^2}{m}} = \sqrt{\frac{E_Y x_0^3}{m}}.$$

Differentiating the expression for F, and allowing for the nearest neighbours, gives

$$k = \left(\frac{dF}{dx}\right)_{x_0} = 4\frac{q^2}{4\pi\varepsilon_0}\frac{1}{x_0^3},$$

giving

$$k = 7.36\,\mathrm{N\,m^{-1}}.$$

Now, substituting in the values of k, x_0 and m into the equation for velocity gives

$$v = 4.15 \times 10^3\,\mathrm{m\,s^{-1}}$$

and the elastic constant is,

$$E_Y = 1.47 \times 10^{10}\,\mathrm{Pa}.$$

Example 2.3 Classical and Debye theories of specific heat

The Dulong–Petit law states that the specific heat capacity of one mole of material is

$$C = 3N_0k_B,$$

where N_0 is Avogadro's number and k_B is Boltzmann's constant. Since the heat capacity is simply the derivative of the internal energy U with respect to temperature,

$$C = \frac{dU}{dT},$$

therefore the internal energy U is given by

$$U = \int C\,\mathrm{d}T$$
$$= 3N_0k_\mathrm{B}T$$

with $T = 300\,\mathrm{K}$,

$$U = 7477\,\mathrm{J}.$$

In order to demonstrate that the Debye theory gives the classically expected Dulong–Petit result at high temperature, we can start from the Debye expression for the internal energy

$$U = \frac{9Nk_\mathrm{B}T^4}{\theta_\mathrm{D}^3}\int_0^{x_\mathrm{max}} \frac{x^3}{\mathrm{e}^x - 1}\,\mathrm{d}x,$$

where $x = \hbar\omega/k_\mathrm{B}T$ and $x_\mathrm{max} = \theta_\mathrm{D}/T$. At high temperatures x becomes small, so that

$$\mathrm{e}^x = 1 \approx x.$$

Hence

$$U = \frac{9Nk_\mathrm{B}T^4}{\theta_\mathrm{D}^3}\int_0^{x_\mathrm{max}} x^2\,\mathrm{d}x$$
$$= \frac{9Nk_\mathrm{B}T^4}{\theta_\mathrm{D}^3}\left(\frac{x^3}{3}\right)_0^{\theta_\mathrm{D}/T},$$

leading to the following result for the internal energy of one mole ($N = N_0$) at high temperature,

$$U = 3N_0k_\mathrm{B}T,$$

which is the classically expected result.

Using Fig. 2.9, which shows the variation in heat capacity with temperature, the energy is then simply the integral $\int_0^T C_\mathrm{v}\,\mathrm{d}T = U(T)$. At $T = 300\,\mathrm{K}$, $T/\theta_\mathrm{D} = 0.70$, where the integral is clearly about half of the classical value of $U(T)$. Therefore

$$U(T) \approx 3750\,\mathrm{J}.$$

Exercise 2.4 Lattice vibrations

The equation for harmonic motion is

$$m\frac{\mathrm{d}^2y}{\mathrm{d}t^2} + b\frac{\mathrm{d}y}{\mathrm{d}t} + ky = 0,$$

where y is the position vector. In the special case of undamped simple harmonic motion the value of $b = 0$. The difference between damped and undamped vibration is therefore merely the value of the coefficient of the first derivative term. In the case of a discrete lattice the equation applies to each individual atom:

$$m\frac{\mathrm{d}^2y_i}{\mathrm{d}t^2} + b\frac{\mathrm{d}y_i}{\mathrm{d}t} + ky_i = 0,$$

where y_i represents the position vector of the ith atom in the lattice. Solutions of the above equation are exponentially decaying vibrations in which the rate of decay of vibration is determined by the value of b/m

$$\frac{d^2y}{dt^2}+\frac{b}{m}\frac{dy}{dt}+\frac{k}{m}y=0.$$

Rewriting the coefficients as γ, the damping, and ω, the natural frequency of vibration,

$$\frac{d^2y}{dt^2}+2\gamma\frac{dy}{dt}+\omega^2y=0.$$

The following expression for y solves the differential equation

$$y=Ae^{-2\gamma t}e^{i\omega t}.$$

This can be demonstrated by substitution

$$\begin{aligned}\frac{dy}{dt}&=(-2\gamma+i\omega)(A\,e^{-2\gamma t}\,e^{i\omega t})\\ \frac{d^2y}{dt^2}&=(i\omega-2\gamma)^2A\,e^{-2\gamma t}\,e^{i\omega t}\\ &=(i\omega-2\gamma)^2y\\ &=2\gamma(-i\omega+2\gamma)y-\omega^2y,\end{aligned}$$

which confirms the original differential equation. These equations can then be applied to describe the displacement of each atom in a discrete lattice.

A boundary condition is a known constraint that can be applied to the equation to limit the allowed solutions of the equation. The boundary conditions are usually restrictions on displacement y in the above equations, but there can also be restrictions on the values of dy/dt or d^2y/dt^2. For example, in the finite discrete lattice, imposition of boundary conditions on the displacement of atoms at the ends of the lattice shows that the number of identifiably unique solutions is equal to the number of atoms.

Exercise 2.5 Interatomic potential

If the potential seen by each atom is given by

$$E_p=\tfrac{1}{2}ku^2,$$

where u is the displacement from equilibrium,

$$F=\frac{dE}{du}=ku$$

and,

$$\sigma=\frac{F}{A}=\frac{ku}{a^2},$$

while strain $e = u/a$ and therefore in this case the elastic modulus is strain independent,

$$E_y = \frac{k}{a}.$$

But, if $E_p = \frac{1}{2}ku^2 - \frac{1}{3}fu^3$ then

$$F = \frac{dE_p}{du} = ku - fu^2$$

$$\sigma = \frac{ku - fu^2}{a^2} = \frac{ke}{a} - fe^2,$$

and in this case the elastic modulus is strain dependent

$$E_y = \frac{k}{a} - fe.$$

Exercise 2.6 Heat capacity

According to Einstein's theory of heat capacity

$$C_v = 3N_0k_B\left(\frac{\hbar\omega}{k_BT}\right)^2 \frac{\exp(\hbar\omega/k_BT)}{(\exp(\hbar\omega/k_BT) - 1)^2}.$$

If $\hbar\omega = k_BT$, then

$$C_v = 3N_0k_B\frac{e}{(e-1)^2} = 3N_0k_B\frac{2.7183}{(1.7183)^2}$$

$$C_v = 3N_0k_B \times 0.92,$$

so at $\theta_D = \hbar\omega_0/k_B$ the heat capacity $C_v = 92\%$ of the classical value.

Since $\omega_0 = k_B\theta_D/\hbar$ then for the various solids

Pb: $\omega_0 = 12.4 \times 10^{12}$

Au: $\omega_0 = 22 \times 10^{12}$

NaCl: $\omega_0 = 36 \times 10^{12}$

Fe: $\omega_0 = 47 \times 10^{12}$

Se: $\omega_0 = 85 \times 10^{12}$

C: $\omega_0 = 242 \times 10^{12}$.

On the basis of the harmonic model of the lattice

$$E_P = ku^2$$

$$F = -\frac{dE}{du} = -2ku = m\frac{d^2u}{dt^2}$$

$$\frac{d^2u}{dt^2} = -\frac{2k}{m}u,$$

the natural frequency of oscillation is therefore

$$\omega_0^2 = \frac{2k}{m},$$

and the Debye temperature is

$$\theta_D = \frac{\hbar\omega_0}{k_B} = \frac{\hbar}{k_B}\sqrt{\frac{2k}{m}}.$$

Therefore, high values of the Debye temperature θ_D occur for high values of the Debye frequency which correspond to large k (i.e. stiff materials) and low m (i.e. light atoms).

Example 3.1 Drude free electron theory of metals

The Drude theory attempts to explain only those properties of a metal which arise from the electronic properties. These include the relationship between electrical and thermal conductivity known as the Wiedemann–Franz law. In this respect the 'theory' is, in fact, only a rather limited model of certain restricted properties of metals. The main assumptions of the theory: (i) collisions between electrons are instantaneous and elastic and this is the mechanism by which thermal conductivity takes place; (ii) other interactions between the electrons or the electrons and ions can be neglected; and (iii) the mean free time of electrons between collisions is independent of the electron position and velocity. The Drude model gives the equation of motion of the electrons in the metal as

$$m\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} = e\xi$$

where m is the electronic mass, e the electronic charge, ξ the applied electric field and γ is a damping factor, which prevents the electrons from accelerating indefinitely under the action of a field ξ. The Lorentz theory, by including the bound electrons, arrived at an equation of motion of the form

$$m\frac{d^2x}{dt^2} + \gamma\frac{dx}{dt} = kx = e\xi$$

for the bound electrons where k is the measure of the binding strength of the electrons to their atoms. The optical properties of a metal could then be described in terms of both types of electron, that is both free and bound. The instantaneous velocity of the conduction electrons at 300 K on the basis of this model is

$$\nu = \sqrt{\frac{3k_BT}{m}} = 1 \times 10^5\,\mathrm{m\,s^{-1}}.$$

Example 3.2 Reflectivity based on Drude theory

As shown in the chapter, the dielectric constants ε_1 and ε_2 can be derived on the basis of the Drude model in terms of the two frequencies ν_1 and ν_2

$$\varepsilon_1(\nu) = n^2 - k^2 = 1 - \left(\frac{\nu_1^2}{\nu^2 + \nu_2^2}\right),$$

$$\varepsilon_2(\nu) = 2nk = \frac{\nu_2}{\nu}\left(\frac{\nu_1^2}{\nu^2 + \nu_2^2}\right),$$

where ν_1 is the 'plasma' frequency and ν_2 is the 'damping' frequency. From classical optics the reflectance R is related to the dielectric constants by the equations

$$R = \frac{\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + 1 - \sqrt{2\left(\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + \varepsilon_1\right)}}{\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + 1 + \sqrt{2\left(\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + \varepsilon_1\right)}}.$$

The predictions of the Drude theory can be seen easily from this equations at very high and very low frequencies. However, at intermediate frequencies the evaluation of R is somewhat cumbersome. Nevertheless, starting with the limit as $\nu \to 0$,

$$\lim_{\nu \to 0} \varepsilon_1 = 1 - \nu_1^2/\nu_2^2,$$

$$\lim_{\nu \to 0} \varepsilon_2 = \infty.$$

Consequently, at low frequencies ε_2 dominates in the above equation for reflectance R and therefore,

$$\lim_{\nu \to 0} R = 1.$$

At the high-frequency limit as $\nu \to \infty$ it is easily seen that,

$$\lim_{\nu \to \infty} \varepsilon_1 = 1$$

$$\lim_{\nu \to \infty} \varepsilon_2 = 0.$$

Consequently at high frequencies the reflectance goes to zero:

$$\lim_{\nu \to \infty} R = 0.$$

A sketch of the value of R as a function of frequency ν for the Drude model is given in Fig. 3.6.

The Drude model for reflectance of metals works reasonably well for low frequencies, specifically in the infrared range of the spectrum. It fails to account for the optical properties of metals at higher frequencies, in the optical range and beyond.

Example 3.3 Electrical and optical properties of a classical free electron metal
The resistive coefficient γ is related to the mobility of the electrons by the relation

$$\gamma = \frac{e}{\mu}.$$

Therefore, inserting the value of μ for copper gives

$$= \frac{1.602 \times 10^{-19}\,\mathrm{C}}{3.5 \times 10^{-3}\,\mathrm{m^2\,V^{-1}\,s^{-1}}}$$

$$\gamma = 0.46 \times 10^{-16}\,\mathrm{N\,s\,m^{-1}}.$$

The electrical conductivity σ is itself related to the resistive coefficient γ by the equation

$$\sigma = \frac{n_f e^2}{\gamma} = n_f e \mu,$$

and from the density, atomic weight and number of electrons per atom in copper the number of conduction electrons in a cubic metre of copper is $N_f \approx 8.5 \times 10^{28}\,\mathrm{m^{-3}}$. Therefore the electrical conductivity is

$$\sigma = \frac{(8.5 \times 10^{28})(1.602 \times 10^{-19})^2}{0.46 \times 10^{-16}}$$

$$\sigma = 0.476 \times 10^{8}\,\Omega^{-1}\,\mathrm{m^{-1}}.$$

The time constant between collisions of the electrons τ is

$$\tau = \frac{m}{\gamma}$$

$$= \frac{9.109 \times 10^{-31}}{0.46 \times 10^{-16}}$$

$$\tau = 19.8 \times 10^{-15}\,\mathrm{s}.$$

Exercise 3.4 Classical free electron description of resistivity

$$\text{Number atoms per unit volume } N = \frac{\text{density}}{\text{atomic weight}} \times N_0$$

$$= \frac{8.95 \times 10^3}{63.5} \times 6.02 \times 10^{26}\,\mathrm{m^{-3}}$$

$$= 8.5 \times 10^{28}\,\text{atoms}\,\mathrm{m^{-3}}$$

In copper each atom has one 'free' conduction electron:

Number of free electrons $N_f = 8.5 \times 10^{28}$ electrons $\mathrm{m^{-3}}$.

The current density J is given by

$$J = N_f e v_{\text{drift}}$$

$$v_{\text{drift}} = \frac{J}{N_f e}$$

$$v_{\text{drift}} = 0.74 \times 10^{-6}\,\text{m}\,\text{s}^{-1}$$

and if the relaxation time is $\tau = 5 \times 10^{-14}$ s then

$$\rho = \frac{m}{N_f e^2 \tau}$$

$$\rho_{\text{imp}} = 8.28 \times 10^{-9}\,\Omega\,\text{m}.$$

The resistivity of the pure metal is

$$\rho_{\text{pm}} = 18 \times 10^{-9}\,\Omega\,\text{m}$$

$$\begin{aligned} \rho_{\text{tot}} &= \rho_{\text{pm}} + \rho_{\text{imp}} \\ &= 26.28 \times 10^{-9}\,\Omega\,\text{m}. \end{aligned}$$

Exercise 3.5 Mobility of classical free electrons

Calculate the mobility μ of the electrons in each material using

$$\mu = \frac{\sigma}{N_f e},$$

where σ is given, N_f needs to be calculated, and $e = 1.6 \times 10^{-19}$ C

$$N_f = N_A \times \frac{\text{density}}{\text{atomic mass}} \times \text{valence}$$

$$\text{Cu:}\quad N_f = \frac{6.02 \times 10^{26} \times 8.96 \times 10^3}{63.5} = 8.5 \times 10^{28}\,\text{m}^{-3}.$$

$$\text{Ag:}\quad N_f = \frac{6.02 \times 10^{26} \times 10.5 \times 10^3}{108} = 5.8 \times 10^{28}\,\text{m}^{-3}.$$

$$\text{Au:}\quad N_f = \frac{6.02 \times 10^{26} \times 19.32 \times 10^3}{197} = 5.9 \times 10^{28}\,\text{m}^{-3}.$$

$$\text{Cd:}\quad N_f = \frac{6.02 \times 10^{26} \times 8.65 \times 10^3}{112} \times 2 = 9.3 \times 10^{28}\,\text{m}^{-3}.$$

$$\text{Zn:}\quad N_f = \frac{6.02 \times 10^{26} \times 7.13 \times 10^3}{65.4} \times 2 = 13.1 \times 10^{28}\,\text{m}^{-3}.$$

$$\text{Al:}\quad N_f = \frac{6.02 \times 10^{26} \times 2.7 \times 10^3}{27} \times 3 = 1.8 \times 10^{29}\,\text{m}^{-3}.$$

Therefore the mobilities μ are

	N_f (m^{-3})	$\mu = \sigma/N_f e$ (m^2 V^{-1} s^{-1})
Cu	8.5×10^{28}	4.1×10^{-3}
Ag	5.8×10^{28}	6.7×10^{-3}
Au	5.9×10^{28}	4.4×10^{-3}
Cd	9.2×10^{28}	0.9×10^{-3}
Zn	13.2×10^{28}	0.79×10^{-3}
Al	18.0×10^{28}	1.28×10^{-3}

Cold working will increase the dislocation density and so increase the number of scattering sites for electrons. This will result in increased resistivity. The material can be restored to its former properties by using a strain relief anneal (~500°C) to eliminate the dislocations.

Exercise 3.6 Absorption of light by a metal

The attenuation coefficient α is given by

$$\alpha = -\frac{1}{x}\log_e\left(\frac{I}{I_0}\right).$$

In the case of the metal,

$$x = 50\,\mathrm{nm}$$

$$I/I_0 = 0.33$$

$$\alpha = \frac{1}{50 \times 10^{-9}}\log_e(0.33) = 22 \times 10^6\,\mathrm{m}^{-1}.$$

In the case of the glass

$$x = 0.2\,\mathrm{nm}$$

$$I/I_0 = 0.1$$

$$\alpha = -\frac{1}{0.2}\log_e(0.1) = 11.5\,\mathrm{m}^{-1}.$$

To calculate the extinction coefficient k the wavelength of light is needed

$$E = 1.5\,\mathrm{eV}$$

$$E = 0.24 \times 10^{-18}\,\mathrm{J}$$

$$\nu = E/h = 0.362 \times 10^{15}\,\mathrm{Hz}$$

$$\lambda = \frac{c}{\nu} = 826\,\mathrm{nm},$$

and

$$k = \frac{\lambda\alpha}{4\pi}$$

$$k = 1.45 \text{ for metal}$$

$$k = 0.756 \times 10^{-6} \text{ for glass,}$$

$n = 0.5$ is given and therefore, for the metal

$$\varepsilon_1 = n^2 - k^2 = (0.5)^2 - (1.45)^2 = -1.85$$

$$\varepsilon_2 = 2nk = 2(0.5)k = 1.45.$$

Example 4.1 Fermi energy for a free electron metal
The number of states per unit volume of k-space is $1/4\pi^3$ which allows for one spin-up and one spin-down state at each k-state. Assuming a spherically symmetric distribution in k-space because the electrons are, by definition, completely free, then the region of k-space between k and $k + \mathrm{d}k$ has a volume

$$\mathrm{d}V = 4\pi k^2 \mathrm{d}k,$$

and therefore the number of electron states between k and $k + \mathrm{d}k$ is simply this volume multiplied by $1/4\pi^3$

$$N(k)\,\mathrm{d}k = \frac{1}{4\pi^3}\,4\pi k^2\,\mathrm{d}k$$

$$= \frac{k^2}{\pi^2}\,\mathrm{d}k.$$

This can now easily be transformed into an energy expression $N(E)\,\mathrm{d}E$, since for completely free electrons the energy is

$$E = \frac{\hbar^2 k^2}{2m},$$

so $k^2 = 2mE/\hbar^2$, and $\mathrm{d}k = \frac{1}{2}(2m/\hbar^2)^{1/2}E^{1/2}\,\mathrm{d}E$.

Therefore, the number of electron states between energies E and $E + \mathrm{d}E$ is obtained by substitution into the expression for $N(k)\,\mathrm{d}k$, giving

$$N(E)\,\mathrm{d}E = \frac{1}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2}\,\mathrm{d}E.$$

The total number of electrons per unit volume of space will then be the integral of this expression, and is therefore

$$N = 2N_0(E) = \frac{1}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{3/2} = \frac{k^3}{3\pi^2}.$$

This expression includes all electrons from energy zero up to an energy E. Therefore, if the free electron Fermi energy is E_F it can be related to the total number of electrons per unit volume N_{Tot} $(=2N_0(E_F))$ by the equation

$$E_F = (3\pi^2 N_{Tot})^{2/3}\frac{\hbar^2}{2m}.$$

It must be remembered that this expression for the Fermi energy only applies to free electrons.

Example 4.2 Solution of wave equation in a finite square well

We consider a one-dimensional finite square-well potential of height V_0 and spatial extent $2a$ as shown in Fig. 4.4. and applying the time-independent Schrödinger wave equation,

$$-\frac{\hbar^2}{2m}\frac{d^2\Psi(x)}{dx^2} + V(x)\Psi(x) = E\Psi(x),$$

which is simply a statement that the kinetic energy plus the potential energy equals the total energy. We must now find the set of wave functions $\Psi_i(x)$ which will satisfy this equation with the given boundary conditions. Note that without boundary conditions all wavefunctions $\psi(x)$ will satisfy the equation.

Within the box the potential is zero, so the wavefunction must satisfy the equation

$$-\frac{\hbar^2}{2m}\frac{d^2\Psi}{dx^2} = E\Psi(x),$$

which has solutions inside the box of the form

$$\Psi(x) = A\,e^{ik_1x} + B\,e^{-ik_1x},$$

where

$$k_1^2 = \frac{2m}{\hbar^2}E.$$

Outside the box the wavefunction must satisfy the equation

$$-\frac{\hbar^2}{2m}\frac{d^2\Psi(x)}{dx^2} + V_0\Psi(x) = E\Psi(x)$$

and therefore has solutions of the form,

$$\Psi(x) = C\,e^{k_2x} + D\,e^{-k_2x},$$

where

$$k_2^2 = -\frac{2m}{\hbar^2}(E - V_0).$$

Now we use the boundary conditions to determine the coefficients A, B, C and D. Outside the box, if we consider what happens when $x \rightarrow \infty$ we can look at the wavefunction at infinity. We know that the probability of observing the electron must remain finite so we must have

$$\lim_{x\rightarrow\infty} \Psi(x) = 0.$$

So, outside the box we must have

$$\Psi(x) = D\,\mathrm{e}^{-k_2 x} \quad \text{for } x > a$$

$$\Psi(x) = C\,\mathrm{e}^{k_2 x} \quad \text{for } x < -a.$$

These are functions that decay exponentially with x in both directions moving away from $x = 0$.

The potential well is centred at $x = 0$, so we must have a symmetic distribution of the electrons with respect to x. The observed properties of electrons are represented by the square of the wavefunction $|\Psi(x)|^2$. Therefore, the wavefunction squared should be invariant with inversion of x. This constraint merely requires the wavefunction $\Psi(x)$ itself to be symmetric or antisymmetric with respect to inversion of x.

We will continue by considering only the symmetric (or even parity) wavefunction solutions (a similar approach can be used for the antisymmetric wavefunction). This imposes the following condition,

$$\Psi(x) = \Psi(-x)$$

$$A\,\mathrm{e}^{ik_1 x} + B\,\mathrm{e}^{-ik_1 x} = A\,\mathrm{e}^{-ik_1 x} + B\,\mathrm{e}^{ik_1 x},$$

which implies that $A = B$. Therefore

$$\Psi(x) = A(\mathrm{e}^{ik_1 x} + \mathrm{e}^{-ik_1 x})$$

$$\Psi(x) = 2A\cos(k_1 x).$$

We then need to apply the boundary conditions at $x = \pm a$. The boundary conditions merely require that Ψ and $\mathrm{d}\Psi/\mathrm{d}x$ remain continuous throughout. Therefore at $x = a$ we must have

$$2A\cos(k_1 a) = D\,\mathrm{e}^{-k_2 a}$$

and

$$-2k_1 A\sin(k_1 a) = -k_2 D\,\mathrm{e}^{-k_2 a}.$$

These two conditions can then be used to obtain the coefficients A and D. A similar approach can be used at $x = -a$ to obtain a relation between the coefficients A and C. Dividing these equations gives

$$k_1\tan(k_1 a) = k_2,$$

and so substituting for k_1 and k_2 from the earlier results gives

$$\frac{2mE}{\hbar^2}\tan^2\left(\frac{(2mE)^{1/2}}{\hbar}a\right)=\frac{2mE}{\hbar^2}(E-V_0)$$

$$E-\tan^2\left(\frac{(2mE)^{1/2}}{\hbar}a\right)=V_0.$$

The result is that for $E > V_0$ the electrons are free with all possible values of energy E although the wavefunction is perturbed in the vicinity of the potential well. For $E < V_0$ the electrons are contained within the box (with some quantum tunnelling at the boundaries) and have a set of discrete allowed energies determined by the box parameters a and V_0. The situation is perhaps best depicted as shown in Fig. 4.5.

If there is a periodic potential within the square well, then the situation can be described as a perturbation of the above solution. In fact, this case quickly begins to resemble a real solid, with a large square-well barrier at the boundary of the solid and localized potential wells representing the 'atomic cores'. In this case, which is depicted in Figs. 5.1 and 5.2, for energies less than V_1 the electrons are highly localized at the 'atomic cores.' For energies between V_1 and V_2 the electrons are quasi-free conduction electrons because they can migrate throughout the 'solid' but are constrained by the solid boundaries. For energies greater than V_2 the electrons are free and can have a continuous range of energies, but these allowed energy states correspond to electrons that have completely escaped from the solid.

Only certain energies are allowed for the 'quasi-free' electrons in the Sommerfeld model because the boundary conditions imposed by the potential well can only be satisfied by certain values of the wave vector k. Since the energy is determined by the wave vector (to a first approximation) by $E = \hbar^2k^2/2m$, it follows that only restricted energies of the wavefunction can meet the boundary conditions.

Example 4.3 Electronic specific heat of copper at 300 K

This can be calculated by finding the kinetic energy of the electrons according to the quantum-mechanical free electron model and then differentiating with respect to temperature.

The kinetic energy of the electrons is given classically by

$$E_k = \tfrac{3}{2}Nk_BT,$$

where N is the number of electrons. According to quantum mechanics only those electrons close to the Fermi level can contribute to the specific heat, and therefore the value of N that should be used for specific heat calculations is smaller than the total number of electrons. One way of determining the effective number of electrons N^* is to use the density of states close to the Fermi level and make an estimation

$$N^* = N(E_F)k_BT,$$

so that if we suppose that $N(E)$ does not vary too rapidly with energy close to E_F and we include all electrons within an energy range k_BT of the Fermi level, we can find the effective number N^* of electrons contributing to the specific heat capacity. So the collective kinetic energy of these electrons is

$$E_k = N^* \tfrac{3}{2} k_B T$$

$$E_k = \tfrac{3}{2} k_B^2 T^2 N(E_F),$$

and if kinetic energy is the only relevant energy of the electrons in this model, then the internal energy $U = E_k$, and the electronic heat capacity is therefore

$$C_v^e = \frac{dE_k}{dT} = 3k_B^2 T N(E_F).$$

Here, the density of states at the Fermi level is unaffected by temperature but we still need an expression for $N(E_F)$. The approximate expression for $N(E_F)$ that can be used is

$$N(E_F) = \frac{3N}{2E_F},$$

where N is the total number of conduction electrons. Therefore

$$C_v^e = \frac{9Nk_B^2 T}{2E_F}.$$

You can also use the alternative expression $(\pi^2/2)Nk_B^2(T/E_F)$ for the heat capacity with only minor differences to the final numerical values. If we consider the molar specific heat then $N = N_0$, the number of conduction electrons per mole. In a monovalent metal such as copper this number is equal to Avogadro's number. We know from Example 4.1 that

$$E_F = (3\pi^2 N_{tot})^{2/3} \frac{\hbar^2}{2m},$$

where N_{tot} is the total number of conduction electrons per unit volume in copper. This is $8.5 \times 10^{28}\ \mathrm{m^{-3}}$, and hence $E_F = 1.13 \times 10^{-18}$ J ($=7.05$ eV). Inserting the values $k_B = 1.38 \times 10^{-23}\ \mathrm{J\,K^{-1}}$, $T = 300$ K, $N_0 = 6.02 \times 10^{23}$ electrons/mole and $E_F = 1.13 \times 10^{-18}$ J into the previous equation we obtain the following molar heat capacity for the electrons

$$C_v^e = 0.151\ \mathrm{J\,mol^{-1}\,K^{-1}}.$$

If we consider the lattice specific heat C_v^l at low temperatures to be given by the approximate expression obtained from the Debye model,

$$C_v^l = \frac{12\pi^4}{5} k_B N_0 \left(\frac{T}{\theta_D}\right)^3,$$

then the temperature at which the heat capacities of the electrons and lattice are identical is,

$$C_{\mathrm{v}}^{\mathrm{l}} = C_{\mathrm{v}}^{\mathrm{e}}$$

$$\frac{12\pi^4}{5} k_{\mathrm{B}} N_0 \left(\frac{T}{\theta_{\mathrm{D}}}\right)^3 = \frac{9 k_{\mathrm{B}} N_0 T}{2 T_{\mathrm{F}}},$$

and substituting the given value of the Debye temperature $\theta_0 = 348\,\mathrm{K}$

$$T^2 = 9.942$$
$$T \cong 3.15\,\mathrm{K}.$$

A second 'mathematical solution' exists at high temperatures when

$$\frac{9}{2} N_0 k_{\mathrm{B}} \frac{T}{T_{\mathrm{F}}} = 3 N_0 k_{\mathrm{B}},$$

where T_{F} is the Fermi temperature ($T_{\mathrm{F}} = E_{\mathrm{F}}/k_{\mathrm{B}} = 81.9 \times 10^3\,\mathrm{K}$) and hence

$$T = \tfrac{6}{9} T_{\mathrm{F}}$$
$$= 0.667 T_{\mathrm{F}}$$
$$T = 54.6 \times 10^3\,\mathrm{K},$$

but the material will be gaseous at this temperature, so it is not a physical solution.

Exercise 4.4 Fermi energy of electrons

$$v_{\mathrm{ms}} = \sqrt{\frac{2E}{m}}.$$

For aluminium $E = 11.7 \times 1.6 \times 10^{-19}\,\mathrm{J}$ and the electron mass is $m = 9.1 \times 10^{-31}\,\mathrm{kg}$, therefore

$$\text{Al:} \quad v_{\mathrm{rms}} = 2.03 \times 10^6\,\mathrm{m\,s^{-1}}.$$

Similarly for copper and gold, $E = 7.0 \times 1.6 \times 10^{-19}\,\mathrm{J}$ and $5.5 \times 1.6 \times 10^{-19}\,\mathrm{J}$, respectively, therefore

$$\text{Cu:} \quad v_{\mathrm{rms}} = 1.57 \times 10^6\,\mathrm{m\,s^{-1}}$$
$$\text{Au:} \quad v_{\mathrm{rms}} = 1.39 \times 10^6\,\mathrm{m\,s^{-1}}.$$

The mean free time τ between collisions is related to the conductivity

$$\tau = \frac{m\sigma}{N_{\mathrm{f}} e^2},$$

and mean free path length ℓ can be found from the velocity and mean free time between collisions

$$\ell = v_{\mathrm{rms}} \tau = \frac{v_{\mathrm{rms}} m \sigma}{N_{\mathrm{f}} e^2}.$$

The number of free electrons per unit volume N_f is given by

$$N_\mathrm{f} = \frac{\text{density}}{\text{atomic weight}} \times \text{valence} \times 6.02 \times 10^{26}$$

Aluminium

$$N_\mathrm{f} = 1.8 \times 10^{29}\ \text{electrons/m}^3$$
$$\tau = 7.3 \times 10^{-15}\,\text{s}$$
$$\ell = v\tau = 15 \times 10^{-9}\,\text{m}$$

Copper

$$N_\mathrm{f} = 8.5 \times 10^{28}\ \text{m}^{-3}$$
$$\tau = 2.32 \times 10^{-14}\,\text{s}$$
$$\ell = v\tau = 36 \times 10^{-9}\,\text{m}$$

Gold

$$N_\mathrm{f} = 5.90 \times 10^{28}\ \text{m}^{-3}$$
$$\tau = 2.5 \times 10^{-14}\,\text{s}$$
$$\ell = v\tau = 35 \times 10^{-9}\,\text{m}$$

Exercise 4.5 Diffraction of electrons at the Fermi surface by the crystal lattice
The wave vector k can be expressed in terms of the energy E by

$$\frac{2\pi}{\lambda} = k = \sqrt{\frac{2mE}{\hbar^2}}$$

$$\lambda = \frac{2\pi\hbar}{\sqrt{2mE}}.$$

The wavelengths at the Fermi surface are therefore

Na: $\lambda_\mathrm{f} = 0.69\,\text{nm}$

Cu: $\lambda_\mathrm{f} = 0.46\,\text{nm}$

Ag: $\lambda_\mathrm{f} = 0.52\,\text{nm}$.

For diffraction, the following condition is needed: $n\lambda \leq 2d\sin\theta$, and therefore if $\lambda > 2d$ no diffraction can occur. Maximum wavelengths for the possibility of diffraction in the three solids are

Na: $\lambda_\mathrm{max} = 0.60\,\text{nm}$

Cu: $\lambda_\mathrm{max} = 0.42\,\text{nm}$

Ag: $\lambda_\mathrm{max} = 0.48\,\text{nm}$.

Therefore no diffraction by the lattice occurs for the electrons at the Fermi energy in these metals.

Exercise 4.6 Diffraction of electrons and validity of the quantum free electron model

The energies and velocities of electrons that can be just diffracted by the lattice can be calculated from the maximum wavelength values for diffraction determined above. The wave vector k of an electron with energy E is

$$k = \frac{2\pi}{\lambda} = \sqrt{\frac{2mE}{\hbar^2}},$$

and therefore,

$$E = \frac{4\pi^2\hbar^2}{2m}\frac{1}{\lambda^2} = \frac{(241 \times 10^{-39})}{\lambda^2}\,\mathrm{J}$$

and the velocity is given by

$$v = \sqrt{\frac{2E}{m}}.$$

The energies and velocities needed for diffraction by the lattice are therefore

$$\text{Cu:}\quad E = \frac{241 \times 10^{-39}}{(0.42 \times 10^{-9})^2} = 1.37 \times 10^{-18}\,\mathrm{J} = 8.54\,\mathrm{eV}$$

$$v = 1.73 \times 10^6\,\mathrm{m\,s^{-1}}.$$

$$\text{Ag:}\quad E = \frac{241 \times 10^{-39}}{(0.48 \times 10^{-9})^2} = 1.046 \times 10^{-18}\,\mathrm{J} = 6.54\,\mathrm{eV}$$

$$v = 1.52 \times 10^6\,\mathrm{m\,s^{-1}}.$$

$$\text{Na:}\quad E = \frac{241 \times 10^{-39}}{(0.60 \times 10^{-9})^2} = 0.669 \times 10^{-18}\,\mathrm{J} = 4.18\,\mathrm{eV}$$

$$v = 1.21 \times 10^6\,\mathrm{m\,s^{-1}}.$$

Since the energies at the Fermi level are below the energy needed for diffraction by the lattice the free electron model with no scattering is a valid approximation for the metals.

Example 5.1. Effective mass of electrons in bands

A free electron with wave vector k, energy E and mass m will, since it is free and subject to no potential energy, obey the equation

$$-\frac{\hbar^2}{2m}\nabla^2\psi(x) = E\psi(x)$$

and for a plane wave, free electron, with a wave function

$$\psi(x) = A\,\mathrm{e}^{ikx} + B\,\mathrm{e}^{-ikx}$$

this yields the solution

$$E = \frac{\hbar^2 k^2}{2m},$$

which is a useful simple relationship between the energy E, wave vector k and mass m.

When an electron is in a solid, however, its movement is affected not only by any external electric fields but also by an internal potential $V(x)$ caused by the periodic atomic array and the other electrons. Therefore, the energy equation becomes

$$\frac{-\hbar^2}{2m}\nabla^2\psi(x) + V(x)\psi(x) = E\psi(x).$$

Under these circumstances the simple relationship between E, k, and m given above no longer holds.

However, there will still be some relation between E and k, and if we compare this with the relationship which holds for free electrons, then we can introduce a correction to this equation by describing the motion in terms of a free electron with a modified effective mass,

$$E = \frac{\hbar^2 k^2}{2m^*},$$

where m^* is the effective mass. This happens to be a useful practical result, but we should not lose sight of the fact that it is an artificial procedure in many respects, because it is not really the mass of the electron which is changing but the relationship between E and k, and we are simply choosing to represent this mathematically by incorporating the changes into the expression through the mass of the electron.

If we now consider how the effective masses of the electrons in an energy band are related to the form of the energy band itself, we can obtain another simple relationship. We know that force equals mass times acceleration

$$F = m^* \frac{\mathrm{d}v}{\mathrm{d}t},$$

and the velocity can be represented as energy differentiated with respect to momentum,

$$v = \frac{1}{\hbar}\frac{\mathrm{d}E}{\mathrm{d}k},$$

therefore

$$F = m^* \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\hbar}\frac{\mathrm{d}E}{\mathrm{d}k}\right)$$

$$= \frac{m^*}{\hbar}\frac{\mathrm{d}k}{\mathrm{d}t}\frac{d^2E}{\mathrm{d}k^2}.$$

But, we also know that force is the derivative of momentum with respect to time,

$$F = \frac{\mathrm{d}p}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(\hbar k)$$

$$F = \hbar\frac{\mathrm{d}k}{\mathrm{d}t}.$$

Equating the expressions for F gives

$$m^* = \frac{\hbar^2}{(\mathrm{d}^2E/\mathrm{d}k^2)}.$$

Here, $(\mathrm{d}^2E/\mathrm{d}k^2)$ is the curvature of the energy bands in k-space. Therefore, the effective mass of the electrons is determined solely by the curvature of the electron bands, an interesting and rather surprisingly simple result.

Example 5.2 Origin of electron bands in materials
The description of the emergence of electron bands in solids can be approached either as a perturbation of the free electron model or as a perturbation of the isolated energy levels of bound electrons within a single atom.

As we have shown in Chapter 5 a free electron can have any energy value and still satisfy the Schrödinger (total energy) wave equation. However, once constraints are put on the electrons, in the form of arbitrary potentials, the allowed energy states become discretized as a requirement for meeting the boundary conditions. In the infinite square-well potential the energy levels closely resemble those in an isolated atom because the electron is constrained by the potential to a limited region of space. When the potential is finite the separation of the energy levels depends entirely on how far below the top of the energy well the electron energy is. Those deep-lying, low-energy states are widely separated. Those that are nearer to the top of the well are closer together in energy. The free electron perturbation works very well for the 'quasi-free' conduction band electrons in a solid, which are those higher-energy electrons that migrate throughout the solid.

Alternatively, if we begin from the energy levels in a single isolated atom, then these levels are completely separated, as in the infinite potential well. However, as we bring together more and more atoms to form a solid, the potential wells of these atoms start to overlap producing a perturbation of the potential well of the individual atom. This perturbation is more significant for the outer electrons, which are at higher energies, but is less so for the inner core electrons.

The result is that for the outer electrons there begins to be some spatial overlap of the electron wavefunctions, and there is a resulting coalescence of available states, which then form an allowed energy band. If there are N identical atoms in the solid, then for each atomic energy level there will arise N energy levels in the solid. By Pauli's exclusion principle no two electrons can have an identical set of quantum numbers and therefore the energies of these N levels will be different. The result is an energy band that is a quasicontinuous range of allowable energy states. This is shown in Fig. 5.5.

The key to the calculation of the band gaps is that they are equal to the Fourier coefficients of the crystal potential. Therefore for the first band gap we need to find the first coefficient of the Fourier series expansion. For a square wave of periodicity a and with width of $0.8a$ and height V, the Fourier expansion is

$$f(x) = A_0 + \sum_{n=1}^{\infty} A_n \cos\left(\frac{n\pi x}{a}\right).$$

The Fourier expansion for a periodic square well potential can be found in 'CRC Standard Mathematical Tables' by W. H. Beyer, 27th edn, 1984 on page 403. The expansion is

$$f(x) = -V\left\{0.8 + \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\sin(0.8n\pi)\cos\left(\frac{n\pi x}{a}\right)\right\}.$$

So the first Fourier coefficient, with $n = 1$, is

$$\begin{aligned} E_g &= V\frac{2}{\pi}\sin(0.8\pi) \\ &= 0.374\,\mathrm{V}, \end{aligned}$$

and since $V = 2$ electron volts the energy gap is,

$$E_g = 0.748\,\mathrm{eV}.$$

Example 5.3 Number of conduction electrons in a Fermi sphere of known radius
Since the problem states that the metal is a quasi-free electron metal, and has a spherical Fermi surface in k-space, then the free electron approximation can be used to give the density of states in a volume V of a material. As given in Section 4.4.7, this is

$$\frac{\mathrm{d}N_0(E)}{\mathrm{d}E} = D(E) = \frac{V}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2},$$

where V is the volume of the sample. In this case we will use only a unit cell of the simple cubic lattice; therefore, if the lattice parameter is a the volume will be

$$V = a^3.$$

Integrating the above equation and setting E equal to the Fermi energy E_F, this gives the total number of conduction electron states:

$$N_0(E_F) = \frac{V}{6\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E_F^{3/2}.$$

Allowing two electrons per state, one with spin-up the other with spin-down, the total number of conduction electrons will be $N = 2N_0(E_F)$

$$N = \frac{V}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E_F^{3/2}.$$

At this stage we note that for free electrons $E = \hbar^2k^2/2m$, so that the Fermi energy is,

$$E_F = \frac{\hbar^2 k_F^2}{2m}.$$

Consequently the total number of conduction electrons is,

$$N = \frac{V}{3\pi^2}k_F^3.$$

At the Brillouin zone boundary of this simple cubic lattice we must have

$$k_F = \frac{\pi}{a}.$$

Therefore, substituting in the values for V and k_F we arrive at the following number of conduction electrons per atom needed to just cause the Fermi surface to touch the Brillouin zone boundary:

$$N = \frac{\pi}{3} = 1.047.$$

Exercise 5.4 Boundary conditions and solutions of the wave equation

$$\Psi = A\exp(ikx) + B\exp(-ikx).$$

Putting this into the Schrödinger equation

$$-\frac{\hbar^2}{2m}\frac{d^2\Psi}{dx^2} + V\Psi = E\Psi,$$

gives

$$\frac{\hbar^2}{2m}k^2\Psi + (V - E)\Psi = 0,$$

which is a solution of the wave equation provided

$$k^2 = \frac{2m(E - V)}{\hbar^2}.$$

When boundary conditions are applied, certain contraints on the allowed solutions of the wavefunction are imposed with the result that there are restrictions on the allowed values of k which meet the boundary conditions, and have restrictions on the allowed energies.

In general, the wave function Ψ is a complex number. The probability of observation must always be a real number. The amplitude of a wavefunction $|\Psi|^2$ is always a real number and, in fact, this magnitude represents the energy density of the wavefunction at a given location. Therefore, it is postulated that the probability of observation is $|\Psi|^2$ on $\Psi^*\Psi$. This is the observable quantity while Ψ, being complex, is not observable. Note that this means two wavefunctions $\Psi = a + ib$ and $\Psi = a - ib$, although physically distinct, will apparently seem to be identical when observed, since both have $\Psi^*\Psi = a^2 + b^2$.

Exercise 5.5 Electrons in a periodic potential
If the periodicity of the crystal potential is expressed as

$$V(x + a_n) = V(x),$$

for all $a_n = na$ where n is an integer and a is the lattice parameter, then we can write Schrödinger's equation

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Psi(x) + V(x)\Psi(x) = E\Psi(x),$$

and for an identical wave function at $x + a_\mathrm{n}$

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Psi(x + a_n) + V(x + a_n)\Psi(x + a_n) = E\Psi(x + a_n).$$

Since we require that $V(x + a_n) = V(x)$ this leads to two simultaneous equations

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Psi(x + a_n) = (E - V(x))\Psi(x + a_n)$$

and

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\Psi(x) = (E - V(x))\Psi(x).$$

Solutions of the wave equation have the form

$$\Psi(x) = A\,\mathrm{e}^{\mathrm{i}kx} + B\,\mathrm{e}^{-\mathrm{i}kx}$$

$$\Psi(x \pm a_n) = A\,\mathrm{e}^{\mathrm{i}k(x\pm a_n)} + B\,\mathrm{e}^{-\mathrm{i}k(x\pm a_n)},$$

and therefore

$$\Psi(x \pm a_n) = \Psi(x)\,\mathrm{e}^{\pm \mathrm{i}ka_n},$$

and $a_n = na$, so that

$$\Psi(x + na) = \Psi(x)\,\mathrm{e}^{\mathrm{i}nka}.$$

Now, if we impose boundary conditions so that the wavefunction $\Psi(x)$ must be the same at both ends, $x = 0$ and $x = L$ of the solid $\Psi(0) = \Psi(L)$, and hence

$$\Psi(L) = \Psi(0)\,\mathrm{e}^{\mathrm{i}kL} = \Psi(0),$$

and, therefore,

$$k = \frac{2n\pi}{L} = \frac{2n\pi}{Na},$$

where N is the number of atoms in the lattice ($Na = L$) and n is simply any integer.

If we impose a more restrictive boundary condition such that the wavefunction will be the same at the boundaries of each unit cell, then $\Psi(0) = \Psi(na)$ for any integer n, including $n = 1$ and hence

$$\Psi(na) = \Psi(0)\,\mathrm{e}^{\mathrm{i}nka} = \Psi(0)$$

$$\mathrm{e}^{\mathrm{i}nka} = 1$$

$$k = \frac{2n\pi}{a},$$

and since

$$\Psi(x + na) = \Psi(x)\,\mathrm{e}^{\mathrm{i}nka}$$

it follows that for any integer n

$$\Psi(x + na) = \Psi(x).$$

According to Bloch's theorem the wavefunction of an electron in a lattice should obey

$$\Psi(x) = u(x)\exp(\mathrm{i}kx),$$

where $u(x)$ has the periodictiy of the lattice, i.e. $u(x) = u(x + na)$ and n is any integer. The potential itself obeys

$$V(x) = V(x + na).$$

Now, on translating once around the entire array of N atoms

$$\Psi(x) = \Psi(x + Na).$$

If this is true, then

$$\begin{aligned}\Psi(x + Na) &= u(x + Na)\exp(\mathrm{i}k(x + Na))\\ \Psi(x + Na) &= u(x)\exp(\mathrm{i}kx)\exp(\mathrm{i}kNa)\\ &= \Psi(x)\exp(\mathrm{i}kNa)\\ &= \Psi(x),\end{aligned}$$

and therefore $\exp(\mathrm{i}kNa) = 1$ and hence $k = 2n\pi/Na$ where n is any integer.

If we have a wavefunction

$$\Psi_k(x) = A\exp(\mathrm{i}kx) + B\exp(-\mathrm{i}kx)$$

then when $k = 0$,

$$\Psi_0(x) = A + B,$$

but we must also have a finite probability of observing the electron when integrating over all space.

$$\int_{\text{all } x} \Psi^*(x)\Psi(x)\,\mathrm{d}x = 1.$$

The only way this can be satisfied is if $A = B = 0$, otherwise the integral will be infinite.

Exercise 5.6 Electron energy bands

The first one is a metal because energy states exist at the Fermi level. The actual diagram is of copper.

The second one is a semiconductor/insulator because the Fermi level lies in the band gap. The band gap is 1 eV, so it is a semiconductor. The actual band structure diagram is for silicon. $E_g = 1$ eV so that $\lambda_{\text{thresh}} = hc/E = 1240$ nm. For $\lambda < 1240$ nm the material is absorbent. For $\lambda > 1240$ nm the material is transparent.

Example 6.1 Brillouin zones in a two-dimensional lattice

If the lattice parameters in real space are $a = 0.2 \times 10^{-9}$ m and $b = 0.4 \times 10^{-9}$ m, then the reciprocal lattice vectors are

$$k_x = \frac{\pi}{a},$$

$$k_y = \frac{\pi}{b}.$$

The reciprocal lattice of a rectangular lattice is also rectangular (see Fig. S1).

Dimensions of the first Brillouin zone are twice the lengths of the reciprocal lattice vectors

$$k_x = \frac{2\pi}{a} = 31.4\,\text{nm}^{-1}$$

$$k_y = \frac{2\pi}{b} = 15.7\,\text{nm}^{-1}.$$

If the atom has a valence of 1 then it will have one electron per atom in the conduction band. This conduction band is according to the statement of the problem of a free electron sphere. In two dimensions the free electrons will occupy a circle, the radius of this 'Fermi circle' is k_F. Therefore

$$\text{'Area' of the Fermi circle} = \pi k_F^2.$$

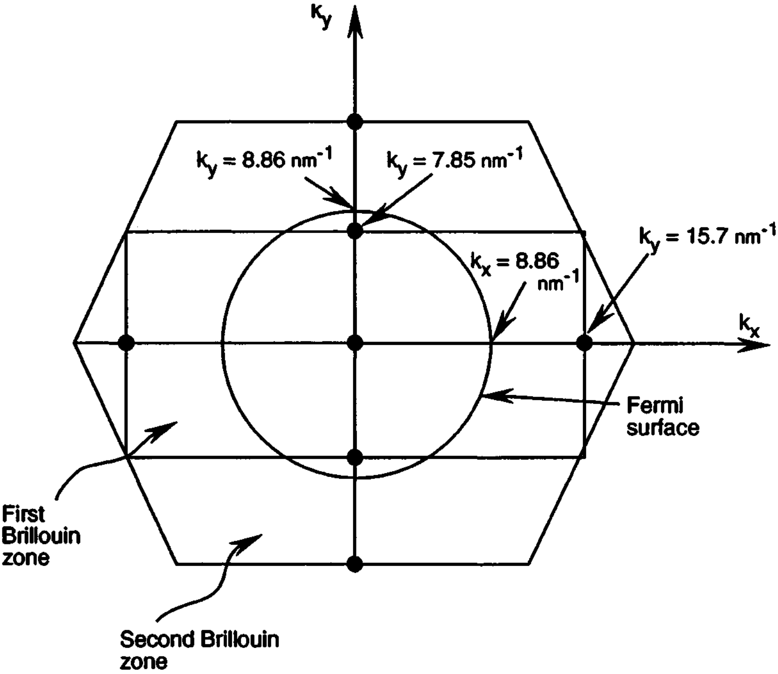


Figure S1 Brillouin zones of the two-dimensional rectangular lattice in reciprocal space, and the Fermi 'sphere' superimposed on the zones.

The total number of available k-states contained in a free electron circle radius k_F is simply the area of the Fermi circle πk_F^2 divided by the area of a k-state. The area of each k-state is

$$A_k = \left(\frac{2\pi}{a}\right)\left(\frac{2\pi}{b}\right) = \frac{4\pi^2}{ab}.$$

Therefore, the number of k-states between zero energy and the Fermi energy E_F in this two-dimensional case is

$$N_0(E_F) = \frac{ab}{4\pi}k_F^2.$$

The total number of electrons will be double this number because of the possibility of accommodating a 'spin-up' and 'spin-down' electron at each k-state

$$N_0(E_F) = \frac{ab}{2\pi}k_F^2.$$

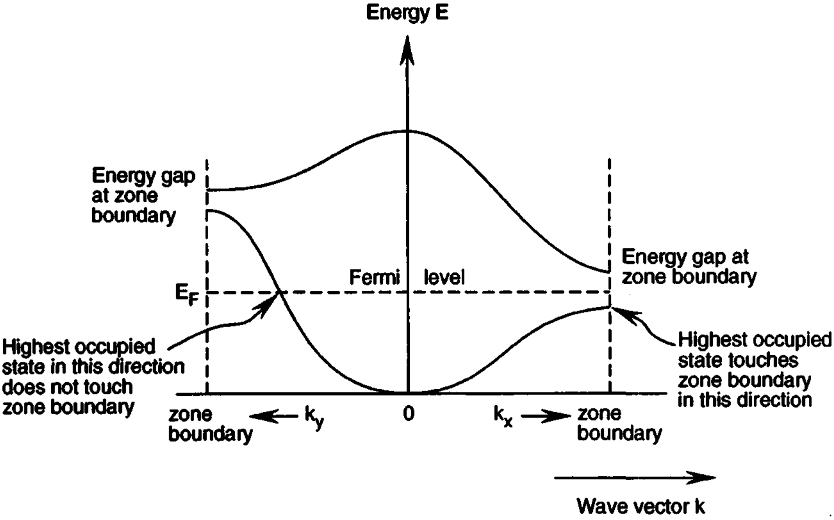


Figure S2 Electron band structure of the two-dimensional rectangular lattice.

And since the number of electrons per atom must be 1, $N_0(E_F) = 1$ and therefore,

$$\begin{aligned} k_F &= \sqrt{\frac{2\pi}{ab}} \\ &= \sqrt{\frac{2\pi}{0.08 \times 10^{-18}}} \\ &= 8.86\,\mathrm{nm}^{-1}. \end{aligned}$$

Since the dimensions of the first Brillouin zone are $k_x = 15.7\,\mathrm{nm}^{-1}$ and $k_y = 7.85\,\mathrm{nm}^{-1}$, the free electron sphere crosses the zone boundary along the k_y direction, but does not cross the zone boundary along the k_x direction. The electron band structure is therefore as shown in Fig. S2.

Example 6.2 Number of k-states in reciprocal space

Consider a simple cubic lattice with lattice parameter a and with N^3 primitive cells. This will give a cube of side L such that

$$L = Na.$$

If we apply periodic boundary conditions the allowed values of k are as follows:

$$k_x, k_y, k_z = 0, \pm\frac{2\pi}{L}, \pm\frac{4\pi}{L}, \pm \cdots \pm \frac{2N\pi}{L}.$$

Ignoring the case $k = 0$, which is physically identical to $k = 2N\pi/L$, this gives N allowed k-states for each of k_x, k_y, k_z. That is, there are N^3 allowed k-states. Since all of the allowed states can be represented in the first Brillouin zone in the reduced zone scheme this means that the number of k-states in the first Brillouin zone of the reduced zone scheme is N^3. Alternatively, we may say that the volume of k-space up to the Nth Brillouin zone is

$$V_N = \left(\frac{2\pi N}{a}\right)^3.$$

Furthermore, it is known that the volume occupied by a k-state in reciprocal space is

$$V_k = \left(\frac{2\pi}{a}\right)^3.$$

Therefore, the number of allowed k-states is simply the volume of the k-space representation of the crystal up to the Nth Brillouin zone divided by the volume occupied by a single k-state:

$$\text{No. of } k\text{-states} = N^3.$$

Example 6.3 Fermi energy of sodium and aluminium

If we assume that both metals are free electron-like, then in both cases the Fermi energy is related to the wave vector k_F at the Fermi surface by the equation

$$E_F = \frac{\hbar^2}{2m}k_F^2.$$

The total number of k-states contained in a volume of k-space of radius k_F is equal to the volume of the Fermi sphere in k-space ($\frac{4}{3}\pi k_F^3$) divided by the volume of a Brillouin zone $(8\pi^3/a^3) = (8\pi^3/V)$

$$N_0(k_F) = \frac{(4\pi k_F^3/3)}{(8\pi^3/a^3)}$$

$$N_0(k_F) = \frac{a^3}{6\pi^2}k_F^3,$$

where V is the volume of the specimen and a is the linear dimension of a unit cell. Since each electron can have 'spin-up' or 'spin-down' the number of electrons that can be contained in such a volume is twice the number of k-states:

$$N_0(k_F) = \frac{V}{3\pi^2}k_F^3.$$

If we consider a unit cell of each metal, then the volume of this crystallographic unit cell is $V = (0.43 \times 10^{-9}\,\text{m})^3 = 79.5 \times 10^{-30}\,\text{m}^{-3}$ for sodium, and $V = (0.4 \times 10^{-9}\,\text{m})^3 = 64 \times 10^{-30}\,\text{m}^3$ for aluminium. Within these unit cells there are effectively two atoms in a bcc lattice and four atoms in an fcc lattice. Therefore, the above unit cells contain 2 conduction electrons in sodium but 12 conduction electrons in aluminium because aluminium is trivalent.

For the case of sodium therefore we must have $N_0(k_F) = 2$ and as

$$k_F^3 = \frac{6\pi^2}{79.5 \times 10^{-30}}\ \mathrm{m}^{-3},$$

therefore

$$k_F = 9.06 \times 10^9\ \mathrm{m}^{-1},$$

and consequently the Fermi energy is

$$\begin{aligned} E_F &= \hbar^2 k_F^2/2m \\ E_F &= \frac{(1.054 \times 10^{-34})^2}{2(9.109 \times 10^{-31})}(9.07 \times 10^9)^2 \\ &= 5.016 \times 10^{-19}\ \mathrm{J} \\ &= 3.13\ \mathrm{eV}. \end{aligned}$$

For the case of aluminium we must have $N_0(k_F) = 12$ and so

$$k_F^3 = \frac{36\pi^2}{64 \times 10^{-30}},$$

therefore

$$k_F = 17.71 \times 10^9\ \mathrm{m}^{-1}$$

and the Fermi energy is,

$$\begin{aligned} E_F &= \frac{(1.054 \times 10^{-34})^2}{2(9.109 \times 10^{-31})}(17.71 \times 10^9)^2 \\ &= 1.913 \times 10^{-18}\ \mathrm{J} \\ E_F &= 11.9\ \mathrm{eV}. \end{aligned}$$

Exercise 6.4 Brillouin zones

For free electrons the energy depends on the components of the wave vector according to the following

$$E = \frac{\hbar^2}{2m}(k_x^2 + k_y^2 + k_z^2).$$

Consider, therefore, a simple cubic lattice, with lattice parameter a. Then, at the zone boundary

$$k_x = \frac{\pi}{a}, \quad k_y = \frac{\pi}{a}, \quad k_z = \frac{\pi}{a}.$$

At a zone face,

$$k_x = \frac{\pi}{a}, \quad k_y = 0, \quad k_z = 0$$

$$E_{100} = \frac{\hbar^2}{2m}\left(\frac{\pi^2}{a^2}\right).$$

In a corner of the zone, $k_x = k_y = k_z = \pi/a$

$$E_{111} = \frac{\hbar^2}{2m}\left(\frac{3\pi^2}{a^2}\right).$$

Therefore

$$\frac{E_{111}}{E_{100}} = 3.$$

If the Fermi surface just touches the Brillouin zone boundary, then $k = \pi/a$, and so, assuming free electron behaviour with $a = 0.3 \times 10^{-9}$ m

$$\begin{aligned} E &= \frac{\hbar^2}{2m}k^2 \\ &= \frac{\hbar^2}{2m}\left(\frac{\pi}{a}\right)^2 \\ E &= \frac{h^2}{8ma^2} = 6.7 \times 10^{-19}\,\text{J} \\ E &= 4.2\,\text{eV}. \end{aligned}$$

The number of free electrons per unit volume is

$$N_0(E) = \frac{V}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E_F^{3/2}$$

Using the relation $E_F = \hbar^2 k_F^2/2m$

$$\begin{aligned} N_0(E) &= \frac{V}{3\pi^2}k_F^3 \\ &= \frac{V}{3}\frac{\pi}{a^3}. \end{aligned}$$

So the total number of conduction electrons needed to just reach the Brillouin zone boundary is

$$N_0(E) = 3.88 \times 10^{28} \times V.$$

And if we consider that one atom occupies a volume a^3, then the number of conduction electrons per atom needed to just touch the Brillouin zone boundary is

$$N_0(E) = \frac{\pi}{3} \text{ electrons per atom.}$$

The density of states $D(E)$ is given by

$$D(E) = \frac{V}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2} = \frac{V}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)k$$

$$D(E_F) = \frac{Vm}{2\pi^2\hbar^2}\frac{\pi}{a} = \frac{Vm}{2\pi a\hbar^2}.$$

So at the Fermi level, the density of states is

$$D(E_F) = V\frac{(9.1 \times 10^{-31})}{2(3.142)(0.3 \times 10^{-9})(1.05 \times 10^{-34})^2}\ \mathrm{J}^{-1}$$

$$= 4.35 \times 10^{46} \times V\ \mathrm{J}^{-1},$$

and the density of states per unit volume is, therefore

$$\frac{D(E_F)}{V} = 4.35 \times 10^{46}\ \mathrm{J}^{-1}\ \mathrm{m}^{-3}$$

or

$$\frac{D(E_F)}{V} = 0.70 \times 10^{28}\ \mathrm{eV}^{-1}\ \mathrm{m}^{-3}.$$

Exercise 6.5 Electron density of states at the Fermi level

If the electrons behave as idealized free electrons, then the density of states is

$$D(E) = \frac{V}{4\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2},$$

and so if $E_F = 7\,\mathrm{eV}$ ($= 11.2 \times 10^{-19}$ J), then the density of states at the Fermi level is simply,

$$D(E_F) = \left(\frac{1}{39.5}\right)(1.64 \times 10^{36})^{3/2}(1.12 \times 10^{-18})^{1/2} = 56.27 \times 10^{45}\ \mathrm{J}^{-1}\ \mathrm{m}^{-3}$$

$$= 9 \times 10^{27} \text{ states per eV per cubic metre,}$$

$$N_0(E) = \frac{V}{3\pi^3}\left(\frac{2m}{\hbar^2}E\right)^{3/2}$$

So that

$$E_F = \left(\frac{3\pi^2 N}{V}\right)^{2/3}\frac{\hbar^2}{2m}.$$

Now, assuming $N/V = 6.0 \times 10^{28}\ \mathrm{m}^{-3}$

$$E_F = 9.09 \times 10^{-21}\ \mathrm{J} = 5.67\ \mathrm{eV}.$$

Exercise 6.6 The de Haas–van Alphen effect and the Fermi surface

If the susceptibility of gold and aluminium exhibit periodicity in field $\Delta(1/B)$ of $2 \times 10^{-3}\,\mathrm{T}^{-1}$ and $1 \times 10^{-5}\,\mathrm{T}^{-1}$ respectively, calculate the external area of the Fermi surface normal to the field, the wave vector k_F at the Fermi surface and the Fermi energy E_F in electron volts, assuming that both can be treated as free-electron-like

$$\Delta(1/B) = \frac{2\pi e}{\hbar}\frac{1}{A_{\text{ext}}}$$

$$A_{\text{ext}} = \frac{2\pi e}{\hbar}\frac{1}{\Delta(1/B)}$$

$$= 9.54 \times 10^{15}\frac{1}{\Delta(1/B)}\,\mathrm{m}^{-2}.$$

For gold, $\Delta(1/B) = 2 \times 10^{-5}\,\mathrm{T}^{-1}$ and therefore

$$A_{\text{ext}} = 4.77 \times 10^{20}\,\mathrm{m}^{-2}$$

$$k_F = \sqrt{\frac{A_{\text{ext}}}{\pi}} = 1.23 \times 10^{10}\,\mathrm{m}^{-1}$$

$$E_F = \frac{\hbar^2 k_F^2}{2m} = 9.26 \times 10^{-19}\,\mathrm{J} = 5.8\,\mathrm{eV}.$$

For aluminium, $\Delta(1/B)$ of $1 \times 10^{-5}\,\mathrm{T}^{-1}$, and therefore

$$A_{\text{ext}} = 9.54 \times 10^{20}\,\mathrm{m}^{-2}$$

$$k_F = \sqrt{\frac{A_{\text{ext}}}{\pi}} = 1.74 \times 10^{10}\,\mathrm{m}^{-1}$$

$$E_F = \frac{\hbar^2 k_F^2}{2m} = 1.85 \times 10^{-18}\,\mathrm{J} = 11.6\,\mathrm{eV}.$$

Example 7.1 Approximation to the Fermi function in semiconductors

The Fermi function $f(E)$, which describes the probability that an electron occupies an energy level E at temperature T, is

$$f(E) = \frac{1}{1 + \exp((E - E_F)/k_B T)},$$

where E_F is the Fermi level. If we consider this simply in terms of a probability as a function of the difference in energy ΔE then

$$f(\Delta E) = \frac{1}{1 + \exp(\Delta E/k_B T)}.$$

In the example under consideration the band gap is 0.5 eV. Considering the lowest-available energy state in the conduction band and locating the Fermi level at the midpoint of the band gap, it is clear that the lowest possible value of ΔE for states in the conduction band is 0.25 eV above the Fermi level,

$$\Delta E \geq 0.25\,\text{eV}.$$

Consequently for all energy levels in the conduction band $\Delta E/k_{\mathrm{B}}T \geq 9.67$ and so

$$\exp\left(\frac{\Delta E}{k_{\mathrm{B}}T}\right) \geq 15.8 \times 10^3.$$

Therefore the exponential term dominates the constant term in the denominator and,

$$f(\Delta E) = \frac{1}{1 + \exp(\Delta E/k_{\mathrm{B}}T)} \approx \exp\left(\frac{-\Delta E}{k_{\mathrm{B}}T}\right).$$

Since this is true for the energy levels just above the band gap it will also be true for higher-energy levels, and for temperatures of 300 K and below. Note, however, that for electrons at the bottom of the conduction band $\Delta E = E_{\mathrm{g}}/2$ and not $\Delta E = E_{\mathrm{g}}$ so that the probability of occupancy of energy levels at the bottom of the conduction band can also be written

$$f(E_g) = \exp\left(\frac{-E_{\mathrm{g}}}{2k_{\mathrm{B}}T}\right).$$

Example 7.2 Temperature dependence of conductivity in intrinsic semiconductors
If the wavelength of the absorption edge is $\lambda = 1771$ nm, then the band gap energy will be

$$\begin{aligned} E_{\mathrm{g}} &= \frac{hc}{\lambda} \\ &= \frac{(2.99 \times 10^8)(6.62 \times 10^{-34})}{1.771 \times 10^{-6}} \\ &= 1.12 \times 10^{-19}\,\text{J} \\ E_{\mathrm{g}} &= 0.698\,\text{eV}. \end{aligned}$$

Since the material is described as an intrinsic conductor, its conductivity will obey the relation

$$\sigma = \sigma_0 \exp\left(-\frac{E_{\mathrm{g}}}{2k_{\mathrm{B}}T}\right).$$

If we let the conductivity increase by 30% at a temperature $T + \Delta T$, then,

$$1.3 = \frac{\exp(-E_g/2k_B(T+\Delta T))}{\exp(-E_g/2k_BT)}$$

$$1.3 = \exp\left(\frac{-E_g}{2k_B}\left(\frac{1}{T+\Delta T} - \frac{1}{T}\right)\right),$$

therefore

$$T + \Delta T = \left(\frac{1}{T} - \frac{2k_B}{E_g}\log_e(1.3)\right)^{-1}$$

$$T + \Delta T = 305.9\,\mathrm{K}.$$

Consequently, a temperature rise of $\Delta T = 5.9\,\mathrm{K}$ from 300 K to 305.9 K will lead to a 30% rise in conductivity.

Exercise 7.3 Electronic properties of gallium arsenide, silicon and germanium
The following is a comparision of the electronic properties of the three materials.

	Si	**Ge**	**GaAs**
Band gap (eV)	1.1	0.7	1.4
Electron mobility ($\mathrm{m^2/V\,s}$)	0.15	0.39	0.85
Effective mass of electrons m^*/m_e	0.97	1.64	0.07
Effective mass of holes m^*/m_h	0.6	0.3	0.5
k at valence band maximum	0	0	0
k at conduction band minimum	0	at zone edge in (1, 1, 1)	near zone edge in (1, 0, 0)
Electrical conductivity at 300 K ($\Omega^{-1}\,\mathrm{m}^{-1}$)	9×10^{-4}	2.2	1×10^{-6}
Absorption edge (nm)	1104 (IR)	1873 (IR)	871 (visible)

Applications in which gallium arsenide has an advantage over the others include those involving emission and absorption of light (e.g. lasers and optical communications), in which the direct band gap gives a higher probability of transition, and those requiring fast response times, in which the high electron mobilities in GaAs are advantageous (e.g. high-speed computer applications).

Since the relation between the electron energy E and the wave vector k is given as

$$E = Ak^2,$$

we can calculate the effective mass from the equation,

$$m^* = \frac{\hbar^2}{(\mathrm{d}^2E/\mathrm{d}k^2)},$$

where

$$\frac{\mathrm{d}^2E}{\mathrm{d}k^2} = 2A,$$

therefore

$$m^* = \frac{\hbar^2}{2A},$$

and with $A = 7.5 \times 10^{-38}\,\mathrm{J\,m^2}$,

$$m^* = 74 \times 10^{-33}\,\mathrm{kg},$$

and since the rest mass of an electron is

$$m_0 = 9.1 \times 10^{-31}\,\mathrm{kg}$$

giving

$$m^*/m_0 = 0.08.$$

Exercise 7.4 Electron band gap and conductivity at finite temperature
The number of electrons per unit volume in the conduction band is given by,

$$\begin{aligned} N &= N_0 \exp\left(-\frac{E_g}{2k_B T}\right) \\ N &= (2.5 \times 10^{25}) \exp(-13.52) \\ N &= 33.4 \times 10^{18}\,\mathrm{m^{-3}}, \end{aligned}$$

and conductivity is given by

$$\begin{aligned} \sigma &= Ne\mu \\ &= Ne(\mu_e + \mu_h) \\ \sigma &= 12.33\,\Omega^{-1}\,\mathrm{m^{-1}} \end{aligned}$$

and

$$\rho = 0.081\,\Omega\,\mathrm{m}.$$

In order to get $\sigma = 100\,\Omega^{-1}\,\mathrm{m^{-1}}$ we need to add donors or acceptors to make up the difference between this and $12.33\,\Omega^{-1}\,\mathrm{m^{-1}}$. If N_d is the number density of donors,

$$\begin{aligned} N_d &= \frac{100 - 12.33}{e\mu} \\ N_d &= 2.38 \times 10^{20}\,\mathrm{m^{-3}}. \end{aligned}$$

For the number of acceptors N_a the calculation is similar, except μ is different

$$\begin{aligned} N_a &= \frac{100 - 12.33}{e\mu} \\ N_a &= 5.48 \times 10^{22}\,\mathrm{m^{-3}}. \end{aligned}$$

Therefore, different number densities of acceptor or donor impurities are needed to reach the same conductivity.

Exercise 7.5 Impurity levels and resistivity of semiconductors

Due to compensation, the acceptors will negate the effect of an equivalent number of donors. Therefore, the effective number density of donors $N_{de} = N_d - N_a$, $N_d = 10^{22}$, $N_a = 5 \times 10^{21}$,

$$N_{de} = 5 \times 10^{21}\,\mathrm{m}^{-3}$$

$$N_c = N_{de} \exp\left(-\frac{\Delta E}{k_B T}\right)$$

$$= 5 \times 10^{21} \exp(-5.797)\,\mathrm{m}^{-3}$$

$$N_c = 15.18 \times 10^{18}\,\mathrm{m}^{-3}$$

$$\sigma = Ne\mu$$

$$\sigma = 0.485\,\Omega^{-1}\,\mathrm{m}^{-1}$$

and therefore

$$\rho = 2.06\,\Omega\,\mathrm{m}.$$

Exercise 7.6 Effect of temperature on a pn junction

As shown in the text, the ideal diode equation is

$$J_{tot} = J_s\left(\exp\left(\frac{eV_{app}}{k_B T}\right) - 1\right).$$

When temperature changes, however, J_s is also a variable. So from eqn (7.47)

$$J_{tot}(T, V_{app}) = J_s(E_g, T)\left(\exp\left(\frac{eV_{app}}{k_B T}\right) - 1\right).$$

In the forward-bias region, the term $(\exp(eV_{app}/k_B T) - 1)$ dominates the temperature dependence. In the forward-bias region the current decreases with temperature for a given voltage V_{app}.

In the reverse-bias region, however, $\exp(eV_{app}/k_B T) \ll 1$ when $eV_{app}/k_B T$ is less than -3. This is clearly true for $V_{app} = -0.3\,\mathrm{V}$. Therefore, in the reverse-bias region it is the temperature dependence of $J_s = J_\infty \exp(-E_g/k_B T)$ which dominates, so that $J_{tot}(333) \cong J_{tot}(293) \times [J_s(333)/J_s(293)]$. Therefore using eqn (7.54)

$$J_{tot} = J_\infty \exp\left(-\frac{E_g}{k_B T}\right)\left(\exp\left(\frac{eV_{app}}{k_B T}\right) - 1\right),$$

where J_∞ is temperature independent gives

$$J_s(E_g T) = J_\infty \exp\left(-\frac{E_g}{k_B T}\right),$$

and taking the ratio of J_s at 333 K and 293 K gives

$$J_s(333) = J_s(293)\frac{\exp(-E_g/333k_B)}{\exp(-E_g/293k_B)}$$
$$= J_s(293)\exp\left(-\frac{E_g}{k_B}\left(\frac{1}{333}-\frac{1}{293}\right)\right),$$

and $E_g/k_B = 13.6 \times 10^3$ and

$$\left(\frac{1}{333}-\frac{1}{293}\right) = -4.1 \times 10^{-4},$$

therefore,

$$J_{tot}(333) = J_{tot}(293)\exp(5.56)$$
$$= J_{tot}(293) \times 260$$
$$= 0.26\,\mathrm{mA}.$$

So the current in the reverse-bias region increases with increasing temperature.

Example 8.1 Drift velocity of conduction electrons

If we use the free electron approximation, then the following relation can be used,

$$\frac{1}{\rho} = \sigma = \frac{Ne^2}{m}\tau_F,$$

where ρ is resistivity, σ is conductivity, N is the number of conduction electrons per unit volume, e is the electronic charge, m the electronic mass and τ_F is the mean time between collisions at the Fermi level.

If we let ℓ be the mean free path of the conduction electrons then,

$$v_F = \frac{\ell}{\tau_F},$$

where v_F is the Fermi velocity. Therefore

$$\rho = \frac{mv_F}{Ne^2\ell}.$$

The energy E_F and velocity v_F of electrons at the Fermi level are related by the equation,

$$E_F = \tfrac{1}{2}mv_F^2,$$

so

$$v_F = \sqrt{2E_F/m},$$

and substituting $E_F = 1.922 \times 10^{-18}\,\mathrm{J}$ (= 12 eV) and $m = 9.1 \times 10^{-31}\,\mathrm{kg}$

$$v_F = 2.05 \times 10^6\,\mathrm{m\,s^{-1}},$$

and rearranging the equation for the resistivity

$$\ell = \frac{mv_F}{Ne^2\rho}.$$

This means that we need to calculate N, the number of conduction electrons per unit volume. Since each atom gives three conduction electrons to the conduction band

$$N = 3N_a,$$

where N_a is the number density of atoms. The density of Al = 2700 kg m^{-3} and the atomic weight is 27. Therefore 6.02×10^{26} atoms weigh 27 kg. Hence, the number density of aluminium atoms is

$$N_a = 6.02 \times 10^{28} \text{ atoms per m}^3,$$

and the number of conduction electrons per unit volume N is three times greater,

$$N = 3N_a$$
$$N = 1.806 \times 10^{29}\,\text{m}^{-3}$$

The mean free path length is then,

$$\ell = \frac{mv_F}{Ne^2\rho}$$
$$\ell = \frac{(9.1 \times 10^{-31})(2.05 \times 10^6)}{(1.806 \times 10^{29})(1.602 \times 10^{-19})^2(3 \times 10^{-8})}$$
$$\ell = 13.4 \times 10^{-9}\,\text{m}.$$

The mean free path of the electrons is related to the mean drift velocity v_d under the action of an electric field ξ. The current density J is given by

$$J = \sigma\xi,$$

where σ is the conductivity. Furthermore,

$$J = Nev_d,$$

where N = number density of conduction electrons, e = electronic charge and v_d = drift velocity:

$$v_d = \frac{\sigma\xi}{Ne} = \frac{\xi}{\rho Ne}$$
$$= \frac{(1 \times 10^3)}{(3 \times 10^{-8})(18 \times 10^{28})(1.6 \times 10^{-19})}$$
$$= 1.15\,\text{m}\,\text{s}^{-1}.$$

Example 8.2 Conductivity in intrinsic and extrinsic semiconductors

The problem states that the doped semiconductor, the n-type germanium, contains 10^{23} ionized donors per cubic metre. This means that it has 10^{23} electrons per cubic metre in the conduction band. The intrinsic semiconductor will have a number n of electrons per unit volume in its conduction band, where n is determined by the temperature and the electronic band gap.

The conductivity of the extrinsic germanium σ_{ex} is determined only by the electrons according to the relation,

$$\sigma_{\mathrm{ex}} = Ne\mu_{\mathrm{e}}$$

where N is the number of electrons per unit volume in the conduction band, e is the electronic charge and μ is the electron mobility. In this case, $N = 1 \times 10^{23}\,\mathrm{m}^{-3}$, $e = 1.602 \times 10^{-19}\,\mathrm{C}$ and $\mu_{\mathrm{c}} \cong 0.39\,\mathrm{m}^2\,\mathrm{s}^{-1}\,\mathrm{V}^{-1}$ for electrons, and $\mu_{\mathrm{h}} = 0.19$ $\mathrm{m}^2\,\mathrm{s}^{-1}\,\mathrm{V}^{-1}$ for holes.

In the intrinsic germanium the conductivity is determined by both the electrons and holes, and is given by

$$\sigma_{\mathrm{in}} = eN_{\mathrm{e}}\mu_{\mathrm{e}} + eN_{\mathrm{h}}\mu_{\mathrm{h}},$$

where N_{e} is the number density of electrons in the conduction band and μ_{e} is their mobility, and N_{h} is the number density of holes in the valence band and μ_{h} is their mobility. It can reasonably be assumed that $N_{\mathrm{e}} = N_{\mathrm{h}} = N$ in an intrinsic semiconductor, and also that e is the same for electrons and holes:

$$\sigma_{\mathrm{in}} = Ne(\mu_{\mathrm{e}} + \mu_{\mathrm{h}}).$$

The number density of electrons in the conduction band N can be obtained from thermodynamic considerations. Using Fermi–Dirac statistics the number density is then given by the equation,

$$\begin{aligned} N(E) &= 2f(E)D(E) \\ &= \frac{2}{1 + \exp((E - E_{\mathrm{F}})/k_{\mathrm{B}}T)} \frac{\pi}{4} \left(\frac{2m}{\pi^2 \hbar^2}\right)^{3/2} E^{1/2} \end{aligned}$$

and assuming that the exponential term in the denominator is much greater than unity,

$$N(E) \cong \frac{1}{2\pi^2} \left(\frac{2m}{\hbar^2}\right)^{3/2} E^{1/2} \exp\left(\frac{-(E - E_{\mathrm{F}})}{k_{\mathrm{B}}T}\right).$$

Integrating this equation, assuming that $E = k_{\mathrm{B}}T$ and $E - E_{\mathrm{F}} \cong E_{\mathrm{g}}/2$ for the conduction electrons at the bottom of the conduction band, leads to,

$$N = \frac{1}{4} \left(\frac{2m}{\hbar^2}\right)^{3/2} (k_{\mathrm{B}}T)^{3/2} \left(\frac{1}{\pi}\right)^{3/2} \exp\left(\frac{-E_{\mathrm{g}}}{2k_{\mathrm{B}}T}\right).$$

Then, rearranging and correcting for the effective mass of the electrons gives

$$N = \frac{1}{4}\left(\frac{2mk_B}{\pi\hbar^2}\right)^{3/2}\left(\frac{m^*}{m}\right)^{3/2} T^{3/2}\exp\left(\frac{-E_g}{2k_BT}\right).$$

The first term on the right-hand side is a constant with value $4.82 \times 10^{21}\,\mathrm{K}^{-3/2}\,\mathrm{m}^{-3}$ so this gives the final expression for N as

$$N = 4.82 \times 10^{21}\left(\frac{m^*}{m_0}\right)^{3/2} T^{3/2}\exp\left(\frac{-E_g}{2k_BT}\right),$$

where m^* is the effective mass of electrons and m_0 is the rest mass of free electrons:

$$\sigma_{in} = 4.82 \times 10^{21}\left(\frac{m^*}{m_0}\right)^{3/2} T^{3/2} e(\mu_e + \mu_n)\exp\left(\frac{-E_g}{2k_BT}\right)$$

and $(m^*/m_0) \approx 0.8, T = 300\,\mathrm{K}, \mu_e = 0.36\,\mathrm{m}^2\,\mathrm{s}^{-1}\,\mathrm{V}^{-1}, \mu_h = 0.18\,\mathrm{m}^2\,\mathrm{s}^{-1}\,\mathrm{V}^{-1}, E_g = 0.7\,\mathrm{eV}\ (1.12 \times 10^{-19}\,\mathrm{J}), e = 1.602 \times 10^{-19}\,\mathrm{C}$. Therefore,

$$\sigma_{in} = (4.82 \times 10^{21})(0.71)(5196)(1.602 \times 10^{-19})(0.58)\exp(-13.53)$$
$$\sigma_{in} = 2.2\,\Omega^{-1}\,\mathrm{m}^{-1}$$

and

$$\sigma_{ex} = (1 \times 10^{23})(1.602 \times 10^{-19})(0.39)$$
$$= 6.248 \times 10^3\,\Omega^{-1}\,\mathrm{m}^{-1}.$$

Therefore, the ratio of extrinsic to intrinsic conductivities is,

$$\sigma_{ex}/\sigma_{in} = 0.3 \times 10^4.$$

Example 8.3 Thermoluminescence and lifetime of electrons in traps
The depth of traps below the conduction band is ΔE, such that,

$$\tau(T) = \frac{1}{s}\exp\left(\frac{\Delta E}{k_BT}\right)$$

therefore,

$$\Delta E = k_BT\log_e(s\tau)$$
$$\Delta E = (1.38 \times 10^{-23})(273)(63.70)$$
$$\Delta E = 2.4 \times 10^{-19}\,\mathrm{J}$$
$$\Delta E = 1.5\,\mathrm{eV}.$$

So the depth of traps below the conduction band is 1.5 eV (or 2.4×10^{-19} J). Using the empirical relationship of Urbach, the temperature T^*, at which the peak of the glow curve occurs is,

$$T^* = 500\Delta E,$$

where T^* is measured in Kelvin and ΔE is measured in electron volts. Therefore

$$T^* = 750\,\mathrm{K},$$

or 477°C.

The lifetime $\tau(T)$ of electrons in traps at a given temperature T is

$$\tau(T) = \frac{1}{s}\exp\left(\frac{\Delta E}{k_B T}\right).$$

Therefore, if the lifetimes at 273 K and 373 K are compared,

$$\begin{aligned}\frac{\tau(373)}{\tau(273)} &= \frac{\exp(\Delta E/373k_B)}{\exp(\Delta E/273k_B)}\\ &= \exp\left(\frac{\Delta E}{k_B}\left(\frac{1}{373}-\frac{1}{273}\right)\right),\\ \tau(373) &= (1.0\times 10^{10})\exp\left(\frac{2.4\times 10^{-19}}{1.38\times 10^{-23}}(0.00268-0.00366)\right)\\ &= (1.0\times 10^{10})\exp(-17.04)\\ \tau(373) &= 3.96\times 10^{2}\\ &= 396\,\mathrm{s}.\end{aligned}$$

Exercise 8.4 Electron and lattice contributions to the thermal conductivity

Thermal conductivity of Ge is given as $K = 80\,\mathrm{W\,m^{-1}\,K^{-1}}$ and electrical resistivity is given as $\rho = 1\times 10^{-5}\,\Omega\,\mathrm{m}$, hence electrical conductivity is $\sigma = 1/\rho = 10^5\,\Omega^{-1}\,\mathrm{m}^{-1}$. From the Wiedemann–Franz law, we can calculate the electronic contribution to thermal conductivity

$$\begin{aligned}\frac{K}{\sigma T} &= L = 2.4\times 10^{-8}\\ K_e &= L\sigma T\\ K_e &= 0.732\,\mathrm{W\,m^{-1}\,K^{-1}}.\end{aligned}$$

The electrical contribution K_e is therefore much smaller than the lattice contribution K_l

$$K_l/K_e = (K_{tot} - K_e)/K_e = 109,$$

or K_e comprises 0.92% of the total thermal conductivity.

Exercise 8.5 Classical explanation of thermal conductivity

Determine the thermal conductivity of a metal, assuming $\tau = 3\times 10^{-14}$ s at 300 K and the number of free electrons is $N_f = 2.5\times 10^{28}\,\mathrm{m}^{-3}$.

The thermal conductivity of a metal can be expressed either in quantum or classical terms:

Quantum: $$K = \frac{\pi^2 k_B^2}{3m} T\tau N_f$$

Classical: $$K = \frac{3k_B^2}{m} T\tau N_f.$$

Inserting the appropriate values,

Quantum: $K = 155\,\mathrm{J\,s^{-1}\,m^{-1}\,K^{-1}}$

Classical: $K = 140.9\,\mathrm{J\,s^{-1}\,m^{-1}\,K^{-1}}$.

Exercise 8.6 Mean free path of 'free' electrons
The mean free path of electrons ℓ in a metal is related to the thermal conductivity K by

$$K = \tfrac{1}{3} c_V^e v \ell,$$

where v is the velocity of the electrons. This velocity can be determined from the Fermi energy, since only electrons at the Fermi level contribute to the conduction process

$$v = \sqrt{\frac{2E_F}{m}}.$$

Therefore,

$$\ell = \frac{3K}{c_V^e v} = \frac{3K}{c_V^e}\sqrt{\frac{m}{2E_F}}.$$

Now, an expression is needed for the electronic heat capacity. We know that the classical expression for the lattice heat capacity above the Debye temperature is

$$c_V^\ell = 3Nk_B,$$

where, in this case, N is the number of atoms per unit volume:

$$N = \frac{\text{density}}{\text{atomic wt.}} \times \text{Avogadro's number}$$

$$= \frac{10\,500}{108} \times 6.02 \times 10^{26}\,\mathrm{m^{-3}}$$

$$N = 5.85 \times 10^{28}\,\mathrm{m^{-3}}.$$

Therefore,

$$c_V^\ell = 2.42 \times 10^6\,\mathrm{J\,m^{-3}\,K^{-1}},$$

and using the given relation $c_v^e = (0.01)c_v^\ell$ gives

$$c_v^e = 24.2 \times 10^3\,\mathrm{J\,m^{-3}\,K^{-1}}.$$

Consequently,

$$\ell = \frac{3K}{c_v^e}\sqrt{\frac{m}{2E_F}},$$

and the Fermi energy $E_F = 5.5\,\mathrm{eV} = 8.8 \times 10^{-19}\,\mathrm{J}$, and thermal conductivity $K = 410\,\mathrm{J\,s^{-1}\,m^{-1}\,K^{-1}}$ gives

$$\ell = 36.5 \times 10^{-9}\,\mathrm{m}.$$

Example 9.1 Optical properties of metals and insulators

The various optical constants α, δ, R, ε_1, and ε_2, can be determined from the refractive index n and the extinction coefficient k by the following equations

$$\alpha = \frac{4\pi k}{\lambda}\,\mathrm{m}$$

$$\delta = \frac{\lambda}{4\pi k}\,\mathrm{m}^{-1}$$

$$R = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2}$$

$$\varepsilon_1 = n^2 - k^2$$

$$\varepsilon_2 = 2nk.$$

The values for the four materials, as calculated from n and k at $\lambda = 1240\,\mathrm{nm}$ ($\hbar\omega = 1\,\mathrm{eV}$), are

Material	n	k	R	α
1	1.21	12.46	0.94	0.126×10^9
2	0.13	8.03	0.99	0.08×10^9
3	1.51	1.12×10^{-6}	0.04	11.35
4	1.92	1.5×10^{-6}	0.10	15.21

Material	d	ε_1	ε_2
1	7.92×10^{-9}	-153.8	30.15
2	12.28×10^{-9}	-64.5	2.09
3	0.088	2.28	3.4×10^{-6}
4	0.066	3.69	5.8×10^{-6}

Materials 1 and 2 have very high reflectance R and high absorption ε_2 at 1 eV and are therefore metals. Both materials 3 and 4 have low reflectance R and absorption ε_2, and a very low extinction coefficient k. These therefore must have a band gap greater than 1 eV which prevents absorption of light at this wavelength. Both are therefore semiconductors or insulators. In fact, material 1 is aluminium, 2 is gold and 3 and 4 are different types of glass.

Example 9.2 Classification of principal electronic transitions
The principal types of electronic transitions that can occur are as follows:

Interband transitions
(i) High-energy transitions (from bottom of valence band to top of conduction band)
(ii) Band gap edge transitions (from top of valence band to bottom of conduction band).

Impurity level transitions
(iii) Exciton generation (from valence band to trap)
(iv) Impurity level excitation (from trap to conduction band).

Intraband transitions (metals only)
(v) Transitions within partially filled band.

Characteristic colours
The characteristic colours of materials are determined mainly by the band gap energy. Absorption, and hence reflectance, can only occur when there is an allowed electronic transition of the appropriate energy. Therefore, in large band gap materials (e.g. diamond) all optical wavelengths are transmitted. As the band gap becomes smaller, however, certain wavelengths at the shorter wavelength, higher-energy end of the spectrum start to get absorbed. This can lead to green, yellow, orange, and red transmission as the band gap decreases. Finally, the transmission in the visible range goes to zero as the band gap falls below the visible red end of the spectrum at about 1.7 eV, leading to a black colour.

Colours by reflection, however, will be different because absorption of a given wavelength allows it to be reflected. Therefore, semiconductors with band gaps in the range 2.5–3.0 eV may appear bluish by reflection but yellow or orange by transmission.

Certain coloured metals such as copper and gold also have their reflectances altered by the presence of characteristic interband transitions at the appropriate energies to cause a yellow or reddish tinge in their reflectance spectrum.

Example 9.3 Identification of material from optical absorption spectrum
The optical spectrum is that of an insulator or semiconductor because of the low absorption at long wavelengths (low energies) and higher absorption at short wavelengths (high energies).

The absorption edge, which corresponds to the energy at which electrons can just begin to cross the band gap occurs in this material at about 0.75 eV, where the absorption increases from zero as the photon energy increases. From the data on the band gaps and the absorption edges of the three materials A, B and C, as given in the table, the spectrum must correspond to material B.

In fact, the remainder of the data in the tables is irrelevant for interpreting which of the materials corresponds to the given spectrum.

The material will be transparent for wavelengths at which the absorption ε_2 is close to zero. This corresponds to all wavelengths longer than 1653 nm (equivalent to 0.75 eV). As a result it will be opaque, and hence reflecting, for all shorter wavelengths, including the optical region of the spectrum (750–450 nm).

The materials are: A – silicon, B – germanium and C – gallium arsenide.

Exercise 9.4 Equation of motion of 'free' electrons and the absorption of light

The equation of motion for electrons, using the classical free electron model, under the action of an external field $\xi_0 \exp(\mathrm{i}\omega t)$ is

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + \gamma\frac{\mathrm{d}x}{\mathrm{d}t} + kx = e\xi_0 \exp(\mathrm{i}\omega t).$$

The physical signficance of the terms is:

'$e\xi_0 \exp(\mathrm{i}\omega t)$' is the instantaneous force on the electrons due to the electromagnetic field

'γ' is the damping coefficient which dissipates the electron energy and hence slows the motion ($\gamma = 0$ is undamped)

'k/m' is the square of the natural frequency of oscillation ($k = m\omega_0^2$) of the electrons.

The solution is the sum of a transient and a steady-state term. The transient term is,

$$x = A\exp\left(-\frac{\gamma}{2m}t\right)\exp\left(\mathrm{i}\left(\frac{k}{m} - \frac{\gamma^2}{4m^2}\right)^{1/2} t\right),$$

which decays away exponentially with time constant $2m/\gamma$.

The steady-state term is

$$x = \frac{e\xi_0}{\omega Z_m}\{\sin(\omega t - \phi) - \mathrm{i}\cos(\omega t - \phi)\},$$

where $Z_m = [\gamma^2 + (\omega m - k/\omega)^2]^{1/2}$.

Amplitude

The amplitude of oscillation of the electrons is

$$A = \frac{e\xi_0}{\sqrt{\omega^2\gamma^2 + (\omega^2 m - k)^2}}.$$

Under these conditions it is easily seen that resonance must occur when the denominator of the expression for A reaches a minimum. This occurs when

$$\omega^2 = \frac{k}{m} - \frac{\gamma^2}{2m^2},$$

and since $k = m\omega_0^2$

$$\omega^2 = \omega_0^2 - \frac{\gamma^2}{2m^2},$$

where ω_0 is the natural frequency of the undamped, unforced electrons.

Amplitude resonance therefore occurs when $\omega = \omega_r$

$$\omega_r = \sqrt{\omega_0^2 - \frac{\gamma}{2m^2}}.$$

Resonance can therefore only occur when the damping coefficient is sufficiently small that

$$\gamma^2 < 2m^2\omega_0^2$$

or

$$\gamma^2 < 2mk.$$

Phase is given by

$$\phi = \arctan\left(\frac{m\omega - (k/\omega)}{\gamma}\right)$$

$$= \arctan\left(\frac{m\omega^2 - k}{\gamma\omega}\right).$$

The penetration depth of light or other forms of electromagnetic radiation is δ, given by the Lambert–Beer law:

$$I = I_0 \exp\left(-\frac{z}{\delta}\right),$$

where $\delta = 1/\alpha$ and, furthermore, we know that

$$\delta = \frac{\lambda}{4\pi k},$$

when $\lambda = 589\,\text{nm}$ and $k = 6$

$$\delta = \frac{589 \times 10^{-9}}{24\pi}\ \text{m}$$

$$\delta = 7.81 \times 10^{-9}\ \text{m}.$$

Exercise 9.5 Effects of differences in band gap on optical properties of semiconductors

(a) If the band gap is 2.4 eV, then all energies above 2.4 eV are absorbed, and all energies below are transmitted. 2.4 eV corresponds to the blue/green region of the visible spectrum. Therefore, yellow, orange and red are transmitted.

(b) As dopant is added, the amount of electron donors increases at a depth of 1.4 eV below the conduction band. Therefore, the number of photons in the range of 1.4–2.4 eV that are transmitted is gradually reduced. The colour will gradually fade from yellow to black because the intensity of light that is transmitted will decrease with dopant concentration.

(c) Considering the effects of temperature on the band gap, $E_g = 2.56 - (5.2 \times 10^{-4})T$ eV:

T (K)	**E_g (eV)**	**Colour**
0	2.56	green/blue
200	2.46	green
400	2.35	yellow/green
600	2.25	yellow
800	2.14	orange/yellow
1000	2.04	orange
2000	1.52	red

The initial wavelengths transmitted are green and all longer wavelengths. As the temperature is raised, the shortest transmitted wavelength is gradually increased, so that the range of energies transmitted is reduced. This means that the colour changes from green/blue to orange as shown in the table.

Exercise 9.6 Optical properties of direct and indirect band gap materials
Transmission of light through 1 μm thin layer of PbS.

λ (μm)	**E (eV)**	**Transmission (I/I_0)**	**α (m^{-1}) (10^6)**	**α^2 (m^{-2}) (10^{12})**
2.065	0.597	0.228	1.478	2.18
2.155	0.572	0.251	1.382	1.91
2.255	0.547	0.281	1.269	1.61
2.360	0.523	0.320	1.139	1.298
2.480	0.497	0.368	0.999	0.998
2.610	0.473	0.445	0.810	0.656
2.755	0.448	0.533	0.629	0.396
2.915	0.423	0.728	0.317	0.100

The attenuation coefficient α is obtained from the Lambert–Beer law,

$$\frac{I}{I_0} = \exp(-\alpha x)$$

$$\alpha = \frac{1}{x}\log\left(\frac{I_0}{I}\right).$$

Now, plotting the absorption in terms of the attenuation coefficient α against the photon energy $\hbar\omega$ gives either the form of direct band gap attenuation

$$\alpha = \alpha_d(\hbar\omega - E_g)^{1/2},$$

or the form of indirect band gap attenuation

$$\alpha = \alpha_i(\hbar\omega - E_g)^2.$$

In this case the plot of α^2 against $\hbar\omega$ gives a straight line with intercept 0.416 eV implying a direct transition (direct band gap) with energy gap 0.416 eV.

Example 10.1 Strength of exchange field in iron

A relationship exists between the Curie temperature of a ferromagnet and the exchange interaction. The exchange field H_{ex} is given by

$$H_{ex} = \alpha M,$$

and the paramagnetic susceptibility is

$$\chi = \frac{C}{T - T_c} = \frac{M}{H + H_{ex}},$$

where C is the Curie constant, T is the absolute temperature and T_c is the Curie temperature,

$$C = \frac{N\mu_0 m^2}{3k_B},$$

where N = number of atoms per unit volume, μ_0 is the permeability of free space, k_B is Boltzmann's constant and m is the moment per atom. Rearranging the Curie–Weiss law equation gives

$$T_c = \alpha C = \frac{\alpha N\mu_0 m^2}{3k_B},$$

and therefore,

$$\alpha = \frac{3k_B T_c}{N\mu_0 m^2}.$$

Since $H_{ex} = \alpha M$, and within a single domain the magnetization is saturated so that $M = M_s = Nm$ this leads to

$$\begin{aligned} H_{ex} &= \alpha M_s \\ &= \frac{3k_B T_c}{\mu_0 m}. \end{aligned}$$

The value of m needs to be in A m^2 instead of Bohr magnetons. This can be calculated from the relation 1 Bohr magneton = 9.27×10^{-24} A m^2. Therefore for iron, $m = 2.04 \times 10^{-23}$ A m^2

$$H_{ex} = 1.68 \times 10^9\ \mathrm{A\,m^{-1}},$$

which is a surprisingly high value of magnetic field.

Example 10.2 Comparison of the magnetic moments on atoms in bulk form and in isolation

The values of saturation magnetization M_s of the three metals in bulk and the magnetic moments of the isolated atoms of each metal are given in the table below.

The magnetic moment per atom in the bulk, which is calculated from the saturation magnetization by dividing by the number of atoms per unit volume in the metal, is also shown.

	Isolated ion magnetic moment (Bohr magnetons)	**Saturation magnetization of bulk material (A/m)**	**Calculated moment per atom in bulk material (Bohr magnetons)**
Fe^{2+}	5.4	1.71×10^6	2.22
Co^{2+}	4.8	1.42×10^6	1.72
Ni^{2+}	3.2	0.48×10^6	0.54

The important result here is that the magnetic moments on the atoms in bulk form (e.g. solids) are substantially different from the magnetic moments on the same atoms in isolation. Therefore the interactions between the electrons on atoms in bulk material cause significant modifications of the observed magnetic moments.

If we consider the electronic structure of the iron, cobalt, and nickel, we find that the isolated atoms have the following numbers of electrons in the outer shells:

	3d	**4s**
Fe	6	2
Co	7	2
Ni	8	2

The magnetic properties of these metals are due to the d electrons. However, in bulk material the 3d and 4s electrons occupy similar energy levels, and in fact the broadening of the 3d and 4s levels in the solid causes these levels to overlap. This leads to s-d electron mixing, also known as 'hybridization.'

In the solid, therefore, the 3d and 4s levels can be treated as a single energy band, with twelve possible electron states. The difference in the number of spin-up electrons n_+ and spin-down electrons n_- will give the magnetic moment per atom:

	$(n_+ - n_-)$	$(n_+ + n_-)$
Fe	2.22	8
Co	1.72	9
Ni	0.54	10

Consequently, the number of spin-up and spin-down electrons in the 3d/4s band is

	Spin-up	**Spin-down**
Fe	5.11	2.89
Co	5.36	3.64
Ni	5.27	4.73

Example 10.3 Spontaneous magnetization and the exchange field
The Langevin expression for the magnetization M of a classical paramagnet with localized magnetic moments m on each atomic site is

$$M = Nm\left\{\coth\left(\frac{\mu_0 mH}{k_B T}\right) - \left(\frac{k_B T}{\mu_0 mH}\right)\right\}.$$

If an exchange field H_{ex} is introduced that is proportional to the spontaneous magnetization M within a domain, then the effective field becomes

$$H_{eff} = H + \alpha M.$$

Substituting this into the Langevin expression gives

$$M = Nm\left\{\coth\left(\frac{\mu_0 m(H + \alpha M)}{k_B T}\right) - \left(\frac{k_B T}{\mu_0 m(H + \alpha M)}\right)\right\}.$$

For spontaneous magnetization ordering will occur in the absence of an external field. Setting the external field H equal to zero in this equation gives,

$$M = Nm\left\{\coth\left(\frac{\mu_0 m\alpha M}{k_B T}\right) = \frac{k_B T}{\mu_0 m\alpha M}\right\}$$

$$M = Nm\left\{\coth\left(\frac{\mu_0 \alpha m^2 N}{k_B T}\right) - \frac{k_B T}{\mu_0 \alpha N m^2}\right\}.$$

In order for ferromagnetism to occur, within a domain there must be a spontaneous magnetization. Consequently dM/dH must be infinite at the origin of the M, H plane:

$$\left(\frac{dM}{dH}\right)_{\substack{H=0\\M=0}} = \frac{Nm}{(3k_B T/\mu_0 m) - \alpha Nm}.$$

Therefore, for ferromagnetism to occur we must have the denominator equal to zero (actually negative values also give ferromagnetism). This means that α must have a value of

$$\alpha \geq \frac{3k_B T}{\mu_0 N m^2} = \frac{(3)(1.38 \times 10^{-23})(300)}{(4\pi \times 10^{-7})(9 \times 10^{28})(2 \times 10^{-23})^2},$$

therefore, for ferromagnetic ordering to occur we must have

$$\alpha \cong 274.5.$$

Exercise 10.4 Saturation magnetization and atomic magnetic moments
For iron: the atomic magnetic moment is $m = 2.2$ Bohr magnetons $= 2.04 \times 10^{-23}\,\mathrm{A\,m^2}$, density $= 7.9 \times 10^3\,\mathrm{kg\,m^{-3}}$ and atomic wt. $= 56$. Therefore the number of atoms per unit volume N is given by

$$N = N_A \times \frac{\text{density}}{\text{atomic weight}} = 6.02 \times 10^{26} \times \frac{7.9 \times 10^3}{56}$$

$$N = 8.5 \times 10^{28}\,\mathrm{m^{-3}}.$$

The saturation magnetization is

$$M_s = Nm = 1.73 \times 10^6\,\mathrm{A\,m^{-1}}.$$

Flux density in toroid

$$B = \mu_0(H + M) = \mu_0(H + 0.5M_s)$$

$$= \mu_0(1000 + (0.85 \times 10^6))$$

$$= 1.069\,\mathrm{T}.$$

Cross-sectional area $A = (0.005)^2\,\mathrm{m^2} = 25 \times 10^{-6}\,\mathrm{m^2}$. So the total flux is given by

$$\phi = BA$$

$$\phi = 2.67 \times 10^{-5}\,\mathrm{Wb}.$$

Exercise 10.5 Saturation magnetization and electron band structure
Given $D(E_F) = 1.9 \times 10^{49}\,\mathrm{J^{-1}\,m^{-3}}$ and $\Delta E_{ex} = 5 \times 10^{-21}\,\mathrm{J}$ the excess number of electrons in the spin-up band over the spin-down band will be

$$\Delta N = N_\uparrow - N_\downarrow$$

$$= 2D(E_F)\Delta E_{ex}$$

$$\Delta N = 190 \times 10^{27}\,\mathrm{m^{-3}}.$$

Since each electron has a magnetic moment of $m = 9.27 \times 10^{-24}$, the expected saturation magnetization will be

$$M_s = \Delta N m\ \mathrm{A\,m^{-1}}$$

$$= 1.76 \times 10^6\,\mathrm{A\,m^{-1}}.$$

The number of atoms per unit volume is $N_A = 8.5 \times 10^{28}\,\mathrm{m}^{-3}$. Therefore, the net moment per atom in Bohr magnetons is

$$\begin{aligned} m &= \Delta N/N_A \\ &= 2.23\mu_B, \end{aligned}$$

or in terms of $\mathrm{A\,m^2}$, the net magnetic moment per atom will be

$$m = \frac{M_s}{N_A}$$

$$m = 2.07 \times 10^{-23}\,\mathrm{A\,m^2}.$$

Exercise 10.6 Hysteresis and energy dissipation

If the coercivity is $50\,\mathrm{A\,m^{-1}}$ and the resonance is 0.5 T, then to a first approximation

$$\text{Area of loop} = 2H_c 2B_r = 100\,\mathrm{T\,A\,m^{-1}}.$$

Now the base units of Tesla are $\mathrm{kg\,A^{-1}\,s^{-2}}$. Consequently $\mathrm{T\,A\,m^{-2}} = \mathrm{kg\,m^{-1}\,s^{-2}} = \mathrm{J\,m^{-3}}$. Therefore a first approximation to the hysteresis loss, that is the energy dissipated per cycle per unit volume, is

$$W_H = 100\,\mathrm{J\,m^{-3}}.$$

The transformer core has the following properties

$$\begin{array}{ll} \text{Area} & A = 0.25 \times 10^{-4}\,\mathrm{m^2} \\ \text{Length} & \ell = 0.05\,\mathrm{m} \\ \text{Volume} & V = 1.25 \times 10^{-6}\,\mathrm{m^3}. \end{array}$$

Hysteresis power loss P_H is equal to the product of frequency, volume and hysteresis loss per unit volume:

$$P_H = \nu V W_H.$$

At $\nu = 60$ Hz this is

$$P_H = 0.0075\,\mathrm{W}.$$

This hysteresis loss is not the only power dissipation mechanism. Two other types of losses occur, both related to eddy currents: classical eddy current losses P_c, which are due to movement of electrons under a time dependent field, and 'anomalous' losses P_a which are due to domain wall motion.

SUBJECT INDEX

Page numbers appearing in bold refer to figures and page numbers appearing in *italic* refer to tables.

AUTHOR INDEX