EIGHTH EDITION

Introduction to Solid State Physics

CHARLES KITTEL

Name	Symbol	Name	Symbol	Name	Symbol
Actinium	Ac	Hafnium	Hf	Praseodymium	Pr
Aluminum	Al	Helium	He	Promethium	Pm
Americium	Am	Holmium	Ho	Protactinium	Pa
Antimony	Sb	Hydrogen	H	Radium	Ra
Argon	Ar	Indium	In	Radon	Rn
Arsenic	As	Iodine	I	Rhenium	Re
Astatine	At	Iridium	Ir	Rhodium	Rh
Barium	Ba	Iron	Fe	Rubidium	Rb
Berkelium	Bk	Krypton	Kr	Ruthenium	Ru
Beryllium	Be	Lanthanum	La	Samarium	Sm
Bismuth	Bi	Lawrencium	Lr	Scandium	Sc
Boron	B	Lead	Pb	Selenium	Se
Bromine	Br	Lithium	Li	Silicon	Si
Cadmium	Cd	Lutetium	Lu	Silver	Ag
Calcium	Ca	Magnesium	Mg	Sodium	Na
Californium	Cf	Manganese	Mn	Strontium	Sr
Carbon	C	Mendelevium	Md	Sulfur	S
Cerium	Ce	Mercury	Hg	Tantalum	Ta
Cesium	Cs	Molybdenum	Mo	Technetium	Tc
Chlorine	Cl	Neodymium	Nd	Tellurium	Te
Chromium	Cr	Neon	Ne	Terbium	Tb
Cobalt	Co	Neptunium	Np	Thallium	Tl
Copper	Cu	Nickel	Ni	Thorium	Th
Curium	Cm	Niobium	Nb	Thulium	Tm
Dysprosium	Dy	Nitrogen	N	Tin	Sn
Einsteinium	Es	Nobelium	No	Titanium	Ti
Erbium	Er	Osmium	Os	Tungsten	W
Europium	Eu	Oxygen	O	Uranium	U
Fermium	Fm	Palladium	Pd	Vanadium	V
Fluorine	F	Phosphorus	P	Xenon	Xe
Francium	Fr	Platinum	Pt	Ytterbium	Yb
Gadolinium	Gd	Plutonium	Pu	Yttrium	Y
Gallium	Ga	Polonium	Po	Zinc	Zn
Germanium	Ge	Potassium	K	Zirconium	Zr
Gold	Au				

Periodic Table, with the Outer Electron Configurations of Neutral Atoms in Their Ground States

The notation used to describe the electronic configuration of atoms and ions is discussed in all textbooks of introductory atomic physics. The letters s, p, d, . . . signify electrons having orbital angular momentum 0, 1, 2, . . . in units $\hbar$; the number to the left of the letter denotes the principal quantum number of one orbit, and the superscript to the right denotes the number of electrons in the orbit.

H^{1} $1s$																	He^{2} $1s^2$
Li^{3} $2s$	Be^{4} $2s^2$											B^{5} $2s^22p$	C^{6} $2s^22p^2$	N^{7} $2s^22p^3$	O^{8} $2s^22p^4$	F^{9} $2s^22p^5$	Ne^{10} $2s^22p^6$
Na^{11} $3s$	Mg^{12} $3s^2$											Al^{13} $3s^23p$	Si^{14} $3s^23p^2$	P^{15} $3s^23p^3$	S^{16} $3s^23p^4$	Cl^{17} $3s^23p^5$	Ar^{18} $3s^23p^6$
K^{19} $4s$	Ca^{20} $4s^2$	Sc^{21} $3d$ $4s^2$	Ti^{22} $3d^2$ $4s^2$	V^{23} $3d^3$ $4s^2$	Cr^{24} $3d^5$ $4s$	Mn^{25} $3d^5$ $4s^2$	Fe^{26} $3d^6$ $4s^2$	Co^{27} $3d^7$ $4s^2$	Ni^{28} $3d^8$ $4s^2$	Cu^{29} $3d^{10}$ $4s$	Zn^{30} $3d^{10}$ $4s^2$	Ga^{31} $4s^24p$	Ge^{32} $4s^24p^2$	As^{33} $4s^24p^3$	Se^{34} $4s^24p^4$	Br^{35} $4s^24p^5$	Kr^{36} $4s^24p^6$
Rb^{37} $5s$	Sr^{38} $5s^2$	Y^{39} $4d$ $5s^2$	Zr^{40} $4d^2$ $5s^2$	Nb^{41} $4d^4$ $5s$	Mo^{42} $4d^5$ $5s$	Tc^{43} $4d^6$ $5s$	Ru^{44} $4d^7$ $5s$	Rh^{45} $4d^8$ $5s$	Pd^{46} $4d^{10}$ –	Ag^{47} $4d^{10}$ $5s$	Cd^{48} $4d^{10}$ $5s^2$	In^{49} $5s^25p$	Sn^{50} $5s^25p^2$	Sb^{51} $5s^25p^3$	Te^{52} $5s^25p^4$	I^{53} $5s^25p^5$	Xe^{54} $5s^25p^6$
Cs^{55} $6s$	Ba^{56} $6s^2$	La^{57} $5d$ $6s^2$	Hf^{72} $4f^{14}$ $5d^2$ $6s^2$	Ta^{73} $5d^3$ $6s^2$	W^{74} $5d^4$ $6s^2$	Re^{75} $5d^5$ $6s^2$	Os^{76} $5d^6$ $6s^2$	Ir^{77} $5d^9$ –	Pt^{78} $5d^9$ $6s$	Au^{79} $5d^{10}$ $6s$	Hg^{80} $5d^{10}$ $6s^2$	Tl^{81} $6s^26p$	Pb^{82} $6s^26p^2$	Bi^{83} $6s^26p^3$	Po^{84} $6s^26p^4$	At^{85} $6s^26p^5$	Rn^{86} $6s^26p^6$
Fr^{87} $7s$	Ra^{88} $7s^2$	Ac^{89} $6d$ $7s^2$															

Ce^{58} $4f^2$ $6s^2$	Pr^{59} $4f^3$ $6s^2$	Nd^{60} $4f^4$ $6s^2$	Pm^{61} $4f^5$ $6s^2$	Sm^{62} $4f^6$ $6s^2$	Eu^{63} $4f^7$ $6s^2$	Gd^{64} $4f^7$ $5d$ $6s^2$	Tb^{65} $4f^8$ $5d$ $6s^2$	Dy^{66} $4f^{10}$ $6s^2$	Ho^{67} $4f^{11}$ $6s^2$	Er^{68} $4f^{12}$ $6s^2$	Tm^{69} $4f^{13}$ $6s^2$	Yb^{70} $4f^{14}$ $6s^2$	Lu^{71} $4f^{14}$ $5d$ $6s^2$
Th^{90} – $6d^2$ $7s^2$	Pa^{91} $5f^2$ $6d$ $7s^2$	U^{92} $5f^3$ $6d$ $7s^2$	Np^{93} $5f^5$ $7s^2$	Pu^{94} $5f^6$ $7s^2$	Am^{95} $5f^7$ $7s^2$	Cm^{96} $5f^7$ $6d$ $7s^2$	Bk^{97}	Cf^{98}	Es^{99}	Fm^{100}	Md^{101}	No^{102}	Lr^{103}

Introduction to Solid State Physics

EIGHTH EDITION

Charles Kittel
Professor Emeritus
University of California, Berkeley

Chapter 18, Nanostructures, was written by Professor Paul McEuen of Cornell University.

John Wiley & Sons, Inc

EXECUTIVE EDITOR Stuart Johnson

SENIOR PRODUCTION EDITOR Patricia McFadden

SENIOR MARKETING MANAGER Robert Smith

DESIGN DIRECTOR Madelyn Lesure

SENIOR MEDIA EDITOR Martin Batey

PRODUCTION MANAGEMENT Suzanne Ingrao/Ingrao Associates

This book was set in 10/12 New Caledonia by GGS Books Services, Atlantic Highlands and printed and bound by Malloy Lithographing. The cover was printed by Phoenix Color.

This book is printed on acid free paper. ∞

Library of Congress Cataloging in Publication Data:
Kittel, Charles.
Introduction to solid state physics / Charles Kittel.—8th ed.
p. cm.
ISBN 0-471-41526-X
1. Solid state physics. I. Title.

QC176.K5 2005
530.4'1—dc22 2004042250
ISBN 0-471-41526-X
WIE ISBN 0-471-68057-5

Printed in the United States of America

10 9 8 7 6 5 4

About the Author

Charles Kittel did his undergraduate work in physics at M.I.T and at the Cavendish Laboratory of Cambridge University. He received his Ph.D. from the University of Wisconsin. He worked in the solid state group at Bell Laboratories, along with Bardeen and Shockley, leaving to start the theoretical solid state physics group at Berkeley in 1951. His research has been largely in magnetism and in semiconductors. In magnetism he developed the theories of ferromagnetic and antiferromagnetic resonance and the theory of single ferromagnetic domains, and extended the Bloch theory of magnons. In semiconductor physics he participated in the first cyclotron and plasma resonance experiments and extended the results to the theory of impurity states and to electron-hole drops.

He has been awarded three Guggenheim fellowships, the Oliver Buckley Prize for Solid State Physics, and, for contributions to teaching, the Oersted Medal of the American Association of Physics Teachers. He is a member of the National Academy of Science and of the American Academy of Arts and Sciences.

Preface

This book is the eighth edition of an elementary text on solid state/condensed matter physics for seniors and beginning graduate students of the physical sciences, chemistry, and engineering. In the years since the first edition was published the field has developed vigorously, and there are notable applications. The challenge to the author has been to treat significant new areas while maintaining the introductory level of the text. It would be a pity to present such a physical, tactile field as an exercise in formalism.

At the first edition in 1953 superconductivity was not understood; Fermi surfaces in metals were beginning to be explored and cyclotron resonance in semiconductors had just been observed; ferrites and permanent magnets were beginning to be understood; only a few physicists then believed in the reality of spin waves. Nanophysics was forty years off. In other fields, the structure of DNA was determined and the drift of continents on the Earth was demonstrated. It was a great time to be in Science, as it is now. I have tried with the successive editions of *ISSP* to introduce new generations to the same excitement.

There are several changes from the seventh edition, as well as much clarification:

- An important chapter has been added on nanophysics, contributed by an active worker in the field, Professor Paul L. McEuen of Cornell University. Nanophysics is the science of materials with one, two, or three small dimensions, where "small" means (nanometer 10^{-9} m). This field is the most exciting and vigorous addition to solid state science in the last ten years.
- The text makes use of the simplifications made possible by the universal availability of computers. Bibliographies and references have been nearly eliminated because simple computer searches using keywords on a search engine such as Google will quickly generate many useful and more recent references. As an example of what can be done on the Web, explore the entry http://www.physicsweb.org/bestof/cond-mat. No lack of honor is intended by the omissions of early or traditional references to the workers who first worked on the problems of the solid state.
- The order of the chapters has been changed: superconductivity and magnetism appear earlier, thereby making it easier to arrange an interesting one-semester course.

The crystallographic notation conforms with current usage in physics. Important equations in the body of the text are repeated in SI and CGS-Gaussian units, where these differ, except where a single indicated substitution will translate from CGS to SI. The dual usage in this book has been found helpful and acceptable. Tables are in conventional units. The symbol e denotes the

charge on the proton and is positive. The notation (18) refers to Equation 18 of the current chapter, but (3.18) refers to Equation 18 of Chapter 3. A caret (^) over a vector denotes a unit vector.

Few of the problems are exactly easy: Most were devised to carry forward the subject of the chapter. With few exceptions, the problems are those of the original sixth and seventh editions. The notation *QTS* refers to my *Quantum Theory of Solids, with solutions by C. Y. Fong; TP* refers to *Thermal Physics, with H. Kroemer.*

This edition owes much to detailed reviews of the entire text by Professor Paul L. McEuen of Cornell University and Professor Roger Lewis of Wollongong University in Australia. They helped make the book much easier to read and understand. However, I must assume responsibility for the close relation of the text to the earlier editions, Many credits for suggestions, reviews, and photographs are given in the prefaces to earlier editions. I have a great debt to Stuart Johnson, my publisher at Wiley; Suzanne Ingrao, my editor; and Barbara Bell, my personal assistant.

Corrections and suggestions will be gratefully received and may be addressed to the author by email to kittel@berkeley.edu.

The Instructor's Manual is available for download at: www.wiley.com/college/kittel.

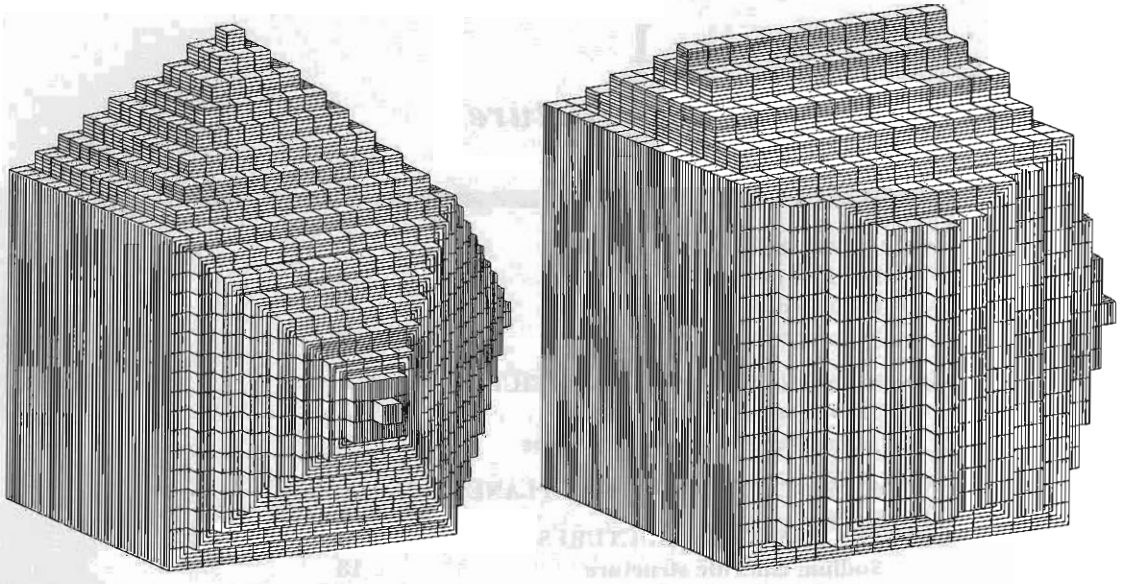
Charles Kittel

Contents

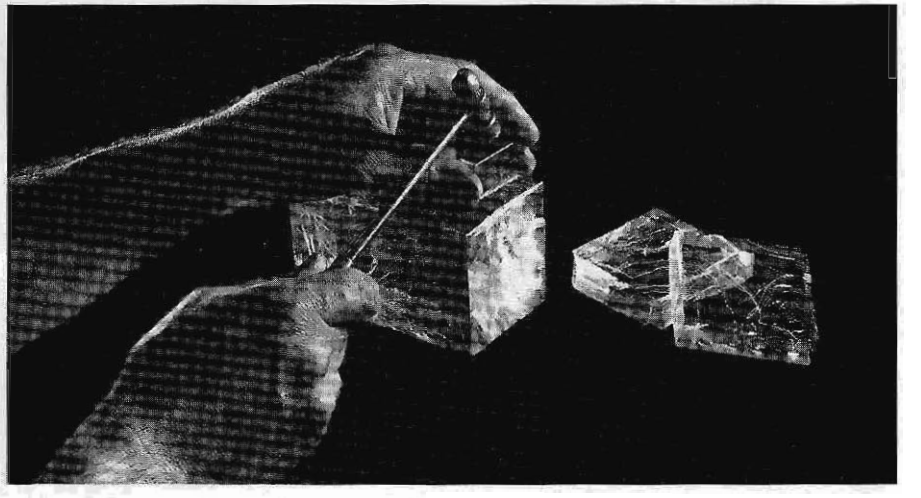
1

Crystal Structure

UNITS: 1 Å = 1 angstrom = 10^{-8} cm = 0.1 nm = 10^{-10} m.



(a) (b)



(c)

Figure 1 Relation of the external form of crystals to the form of the elementary building blocks. The building blocks are identical in (a) and (b), but different crystal faces are developed. (c) Cleaving a crystal of rocksalt.

PERIODIC ARRAYS OF ATOMS

The serious study of solid state physics began with the discovery of x-ray diffraction by crystals and the publication of a series of simple calculations of the properties of crystals and of electrons in crystals. Why crystalline solids rather than noncrystalline solids? The important electronic properties of solids are best expressed in crystals. Thus the properties of the most important semiconductors depend on the crystalline structure of the host, essentially because electrons have short wavelength components that respond dramatically to the regular periodic atomic order of the specimen. Noncrystalline materials, notably glasses, are important for optical propagation because light waves have a longer wavelength than electrons and see an average over the order, and not the less regular local order itself.

We start the book with crystals. A crystal is formed by adding atoms in a constant environment, usually in a solution. Possibly the first crystal you ever saw was a natural quartz crystal grown in a slow geological process from a silicate solution in hot water under pressure. The crystal form develops as identical building blocks are added continuously. Figure 1 shows an idealized picture of the growth process, as imagined two centuries ago. The building blocks here are atoms or groups of atoms. The crystal thus formed is a three-dimensional periodic array of identical building blocks, apart from any imperfections and impurities that may accidentally be included or built into the structure.

The original experimental evidence for the periodicity of the structure rests on the discovery by mineralogists that the index numbers that define the orientations of the faces of a crystal are exact integers. This evidence was supported by the discovery in 1912 of x-ray diffraction by crystals, when Laue developed the theory of x-ray diffraction by a periodic array, and his coworkers reported the first experimental observation of x-ray diffraction by crystals. The importance of x-rays for this task is that they are waves and have a wavelength comparable with the length of a building block of the structure. Such analysis can also be done with neutron diffraction and with electron diffraction, but x-rays are usually the tool of choice.

The diffraction work proved decisively that crystals are built of a periodic array of atoms or groups of atoms. With an established atomic model of a crystal, physicists could think much further, and the development of quantum theory was of great importance to the birth of solid state physics. Related studies have been extended to noncrystalline solids and to quantum fluids. The wider field is known as condensed matter physics and is one of the largest and most vigorous areas of physics.

Lattice Translation Vectors

An ideal crystal is constructed by the infinite repetition of identical groups of atoms (Fig. 2). A group is called the **basis**. The set of mathematical points to which the basis is attached is called the **lattice**. The lattice in three dimensions may be defined by three translation vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$, such that the arrangement of atoms in the crystal looks the same when viewed from the point $\mathbf{r}$ as when viewed from every point $\mathbf{r}'$ translated by an integral multiple of the $\mathbf{a}$'s:

$$\mathbf{r}' = \mathbf{r} + u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + u_3\mathbf{a}_3. \tag{1}$$

Here u_1, u_2, u_3 are arbitrary integers. The set of points $\mathbf{r}'$ defined by (1) for all u_1, u_2, u_3 defines the lattice.

The lattice is said to be **primitive** if any two points from which the atomic arrangement looks the same always satisfy (1) with a suitable choice of the integers u_i. This statement defines the **primitive translation vectors** $\mathbf{a}_i$. There is no cell of smaller volume than $\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3$ that can serve as a building block for the crystal structure. We often use the primitive translation vectors to define the **crystal axes**, which form three adjacent edges of the primitive parallelepiped. Nonprimitive axes are often used as crystal axes when they have a simple relation to the symmetry of the structure.

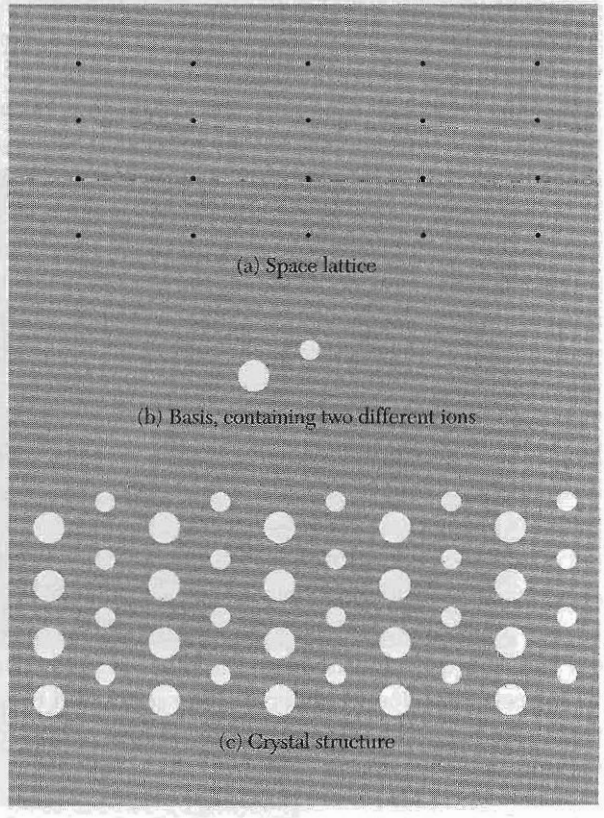


Figure 2 The crystal structure is formed by the addition of the basis (b) to every lattice point of the space lattice (a). By looking at (c), one can recognize the basis and then one can abstract the space lattice. It does not matter where the basis is put in relation to a lattice point.

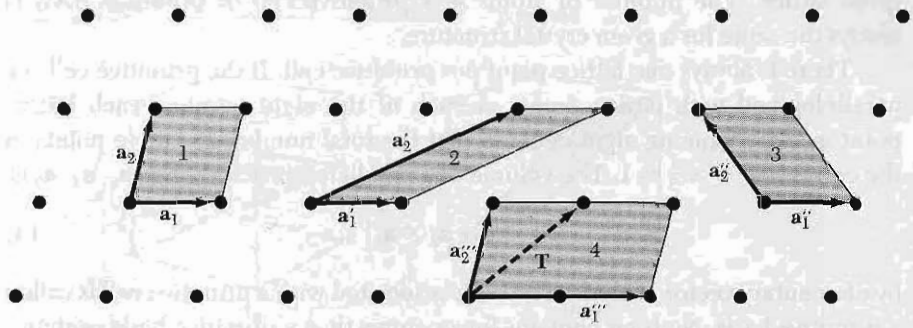
Basis and the Crystal Structure

The **basis** of the crystal structure can be identified once the crystal axes have been chosen. Figure 2 shows how a crystal is made by adding a basis to every lattice point—of course the lattice points are just mathematical constructions. Every basis in a given crystal is identical to every other in composition, arrangement, and orientation.

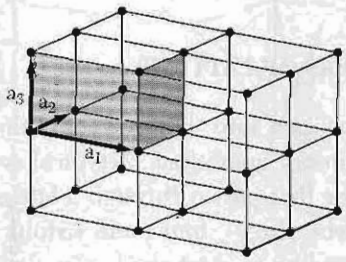
The number of atoms in the basis may be one, or it may be more than one. The position of the center of an atom j of the basis relative to the associated lattice point is

$$\mathbf{r}_j = x_j\mathbf{a}_1 + y_j\mathbf{a}_2 + z_j\mathbf{a}_3 . \tag{2}$$

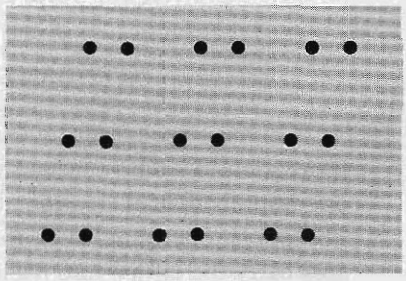
We may arrange the origin, which we have called the associated lattice point, so that $0 \leq x_j, y_j, z_j \leq 1$.



(a)



(b)



(c)

Figure 3a Lattice points of a space lattice in two dimensions. All pairs of vectors $\mathbf{a}_1$, $\mathbf{a}_2$ are translation vectors of the lattice. But $\mathbf{a}_1'''$, $\mathbf{a}_2'''$ are not primitive translation vectors because we cannot form the lattice translation $\mathbf{T}$ from integral combinations of $\mathbf{a}_1'''$ and $\mathbf{a}_2'''$. The other pairs shown of $\mathbf{a}_1$ and $\mathbf{a}_2$ may be taken as the primitive translation vectors of the lattice. The parallelograms 1, 2, 3 are equal in area and any of them could be taken as the primitive cell. The parallelogram 4 has twice the area of a primitive cell.

Figure 3b Primitive cell of a space lattice in three dimensions.

Figure 3c Suppose these points are identical atoms: Sketch in on the figure a set of lattice points, a choice of primitive axes, a primitive cell, and the basis of atoms associated with a lattice point.

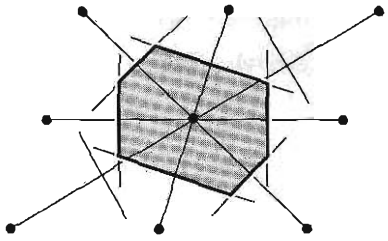


Figure 4 A primitive cell may also be chosen following this procedure: (1) draw lines to connect a given lattice point to all nearby lattice points; (2) at the midpoint and normal to these lines, draw new lines or planes. The smallest volume enclosed in this way is the Wigner-Seitz primitive cell. All space may be filled by these cells, just as by the cells of Fig. 3.

Primitive Lattice Cell

The parallelepiped defined by primitive axes $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ is called a **primitive cell** (Fig. 3b). A primitive cell is a type of cell or unit cell. (The adjective unit is superfluous and not needed.) A cell will fill all space by the repetition of suitable crystal translation operations. A primitive cell is a minimum-volume cell. There are many ways of choosing the primitive axes and primitive cell for a given lattice. The number of atoms in a primitive cell or primitive basis is always the same for a given crystal structure.

There is always one lattice point per primitive cell. If the primitive cell is a parallelepiped with lattice points at each of the eight corners, each lattice point is shared among eight cells, so that the total number of lattice points in the cell is one: $8 \times \frac{1}{8} = 1$. The volume of a parallelepiped with axes $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ is

$$V_c = |\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3| \ , \tag{3}$$

by elementary vector analysis. The basis associated with a primitive cell is called a primitive basis. No basis contains fewer atoms than a primitive basis contains. Another way of choosing a primitive cell is shown in Fig. 4. This is known to physicists as a **Wigner-Seitz cell**.

FUNDAMENTAL TYPES OF LATTICES

Crystal lattices can be carried or mapped into themselves by the lattice translations $\mathbf{T}$ and by various other symmetry operations. A typical symmetry operation is that of rotation about an axis that passes through a lattice point. Lattices can be found such that one-, two-, three-, four-, and sixfold rotation axes carry the lattice into itself, corresponding to rotations by 2π, $2\pi/2$, $2\pi/3$, $2\pi/4$, and $2\pi/6$ radians and by integral multiples of these rotations. The rotation axes are denoted by the symbols 1, 2, 3, 4, and 6.

We cannot find a lattice that goes into itself under other rotations, such as by $2\pi/7$ radians or $2\pi/5$ radians. A single molecule properly designed can have any degree of rotational symmetry, but an infinite periodic lattice cannot. We can make a crystal from molecules that individually have a fivefold rotation axis, but we should not expect the lattice to have a fivefold rotation axis. In Fig. 5 we show what happens if we try to construct a periodic lattice having fivefold

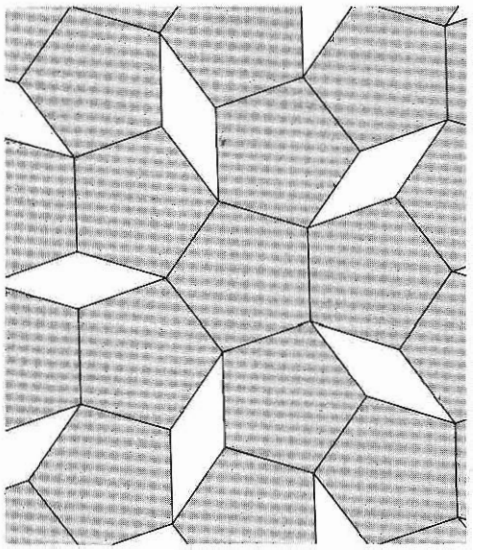


Figure 5 A fivefold axis of symmetry cannot exist in a periodic lattice because it is not possible to fill the area of a plane with a connected array of pentagons. We can, however, fill all the area of a plane with just two distinct designs of "tiles" or elementary polygons.

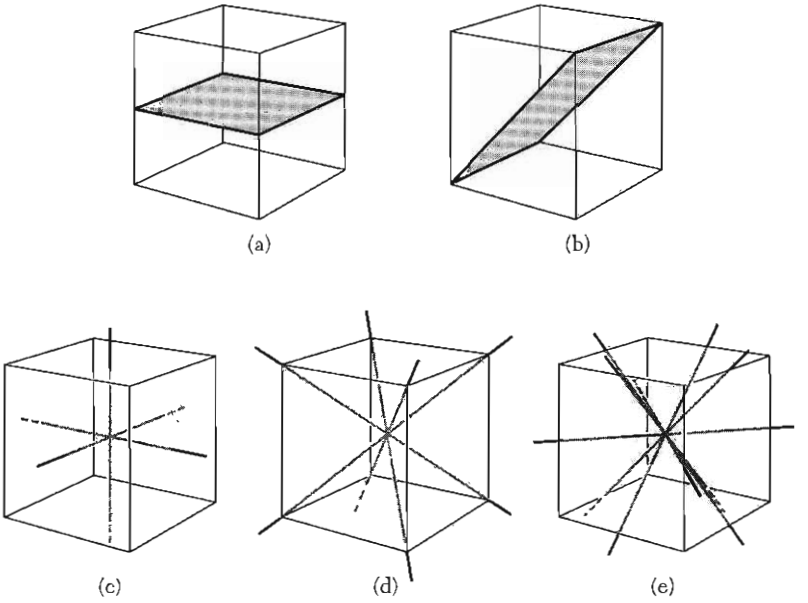


Figure 6 (a) A plane of symmetry parallel to the faces of a cube. (b) A diagonal plane of symmetry in a cube. (c) The three tetrad axes of a cube. (d) The four triad axes of a cube. (e) The six diad axes of a cube.

symmetry: the pentagons do not fit together to fill all space, showing that we cannot combine fivefold point symmetry with the required translational periodicity.

By lattice point group we mean the collection of symmetry operations which, applied about a lattice point, carry the lattice into itself. The possible rotations have been listed. We can have mirror reflections m about a plane through

a lattice point. The inversion operation is composed of a rotation of π followed by reflection in a plane normal to the rotation axis; the total effect is to replace $\mathbf{r}$ by $-\mathbf{r}$. The symmetry axes and symmetry planes of a cube are shown in Fig. 6.

Two-Dimensional Lattice Types

The lattice in Fig. 3a was drawn for arbitrary $\mathbf{a}_1$ and $\mathbf{a}_2$. A general lattice such as this is known as an **oblique lattice** and is invariant only under rotation of π and 2π about any lattice point. But special lattices of the oblique type can be invariant under rotation of $2\pi/3$, $2\pi/4$, or $2\pi/6$, or under mirror reflection. We must impose restrictive conditions on $\mathbf{a}_1$ and $\mathbf{a}_2$ if we want to construct a lattice that will be invariant under one or more of these new operations. There are four distinct types of restriction, and each leads to what we may call a **special lattice type**. Thus there are five distinct lattice types in two dimensions, the oblique lattice and the four special lattices shown in Fig. 7. **Bravais lattice** is the common phrase for a distinct lattice type; we say that there are five Bravais lattices in two dimensions.

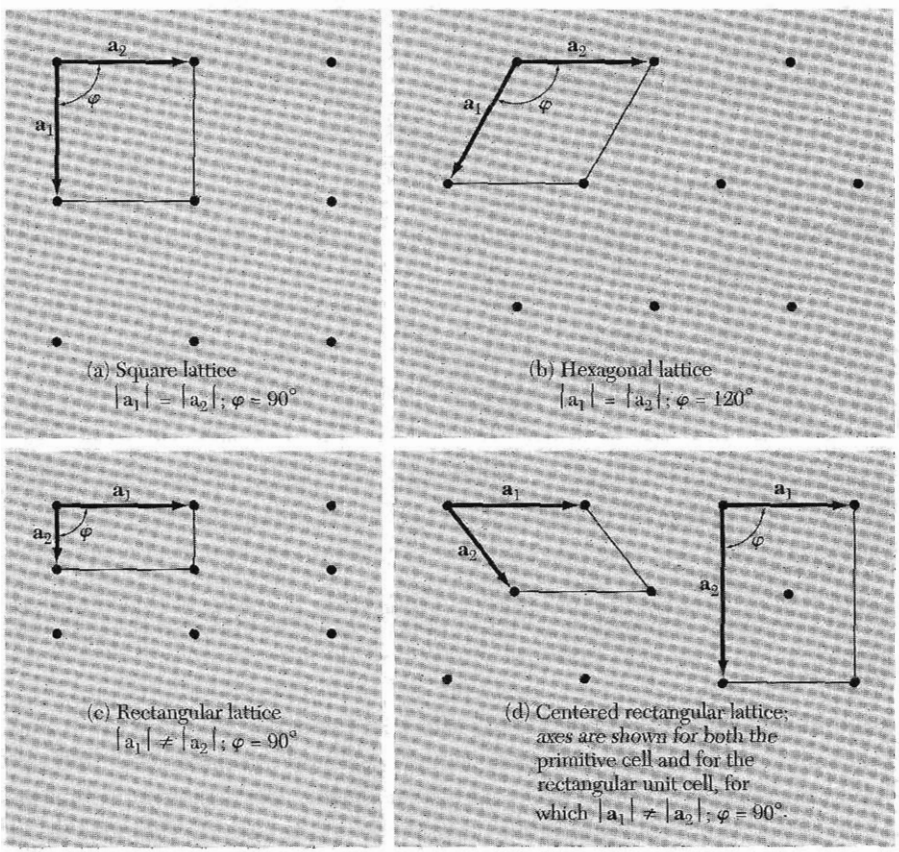


Figure 7 Four special lattices in two dimensions.

Three-Dimensional Lattice Types

The point symmetry groups in three dimensions require the 14 different lattice types listed in Table 1. The general lattice is triclinic, and there are 13 special lattices. These are grouped for convenience into systems classified according to seven types of cells, which are triclinic, monoclinic, orthorhombic, tetragonal, cubic, trigonal, and hexagonal. The division into systems is expressed in the table in terms of the axial relations that describe the cells. The cells in Fig. 8 are conventional cells: of these only the sc is a primitive cell. Often a nonprimitive cell has a more obvious relation with the point symmetry operations than has a primitive cell.

There are three lattices in the cubic system: the simple cubic (sc) lattice, the body-centered cubic (bcc) lattice, and the face-centered cubic (fcc) lattice.

Table 1 The 14 lattice types in three dimensions

System	Number of lattices	Restrictions on conventional cell axes and angles
Triclinic	1	$a_1 \neq a_2 \neq a_3$ $\alpha \neq \beta \neq \gamma$
Monoclinic	2	$a_1 \neq a_2 \neq a_3$ $\alpha = \gamma = 90° \neq \beta$
Orthorhombic	4	$a_1 \neq a_2 \neq a_3$ $\alpha = \beta = \gamma = 90°$
Tetragonal	2	$a_1 = a_2 \neq a_3$ $\alpha = \beta = \gamma = 90°$
Cubic	3	$a_1 = a_2 = a_3$ $\alpha = \beta = \gamma = 90°$
Trigonal	1	$a_1 = a_2 = a_3$ $\alpha = \beta = \gamma < 120°, \neq 90°$
Hexagonal	1	$a_1 = a_2 \neq a_3$ $\alpha = \beta = 90°$ $\gamma = 120°$

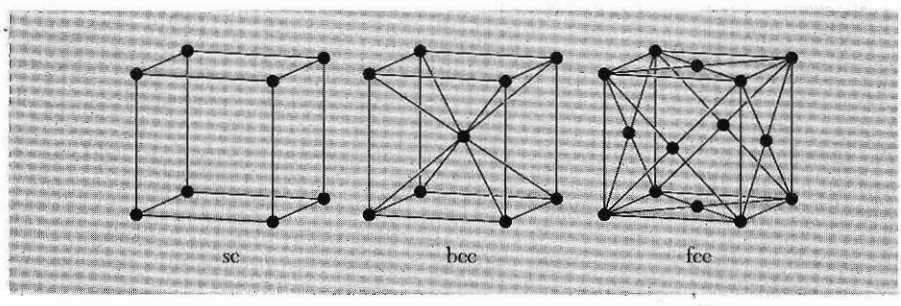


Figure 8 The cubic space lattices. The cells shown are the conventional cells.

Table 2 Characteristics of cubic lattices[a]

	Simple	Body-centered	Face-centered
Volume, conventional cell	a^3	a^3	a^3
Lattice points per cell	1	2	4
Volume, primitive cell	a^3	$\frac{1}{2}a^3$	$\frac{1}{4}a^3$
Lattice points per unit volume	$1/a^3$	$2/a^3$	$4/a^3$
Number of nearest neighbors	6	8	12
Nearest-neighbor distance	a	$3^{1/2}\,a/2 = 0.866a$	$a/2^{1/2} = 0.707a$
Number of second neighbors	12	6	6
Second neighbor distance	$2^{1/2}a$	a	a
Packing fraction[a]	$\frac{1}{6}\pi$	$\frac{1}{8}\pi\sqrt{3}$	$\frac{1}{6}\pi\sqrt{2}$
	$=0.524$	$=0.680$	$=0.740$

[a]The packing fraction is the maximum proportion of the available volume that can be filled with hard spheres.

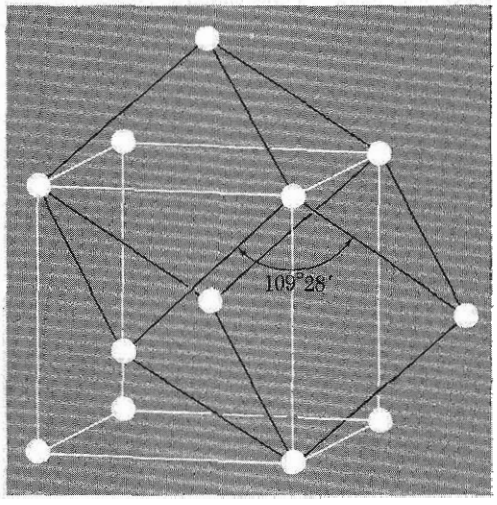


Figure 9 Body-centered cubic lattice, showing a primitive cell. The primitive cell shown is a rhombohedron of edge $\frac{1}{2}\sqrt{3}\,a$, and the angle between adjacent edges is 109°28′.

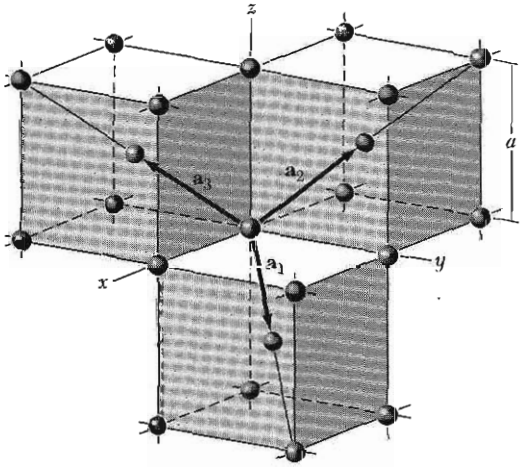


Figure 10 Primitive translation vectors of the body-centered cubic lattice; these vectors connect the lattice point at the origin to lattice points at the body centers. The primitive cell is obtained on completing the rhombohedron. In terms of the cube edge a, the primitive translation vectors are

$$\mathbf{a}_1 = \tfrac{1}{2}a(\hat{\mathbf{x}} + \hat{\mathbf{y}} - \hat{\mathbf{z}}) \; ; \qquad \mathbf{a}_2 = \tfrac{1}{2}a(-\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}}) \; ;$$

$$\mathbf{a}_3 = \tfrac{1}{2}a(\hat{\mathbf{x}} - \hat{\mathbf{y}} + \hat{\mathbf{z}}) \; .$$

Here $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ are the Cartesian unit vectors.

The characteristics of the three cubic lattices are summarized in Table 2. A primitive cell of the bcc lattice is shown in Fig. 9, and the primitive translation vectors are shown in Fig. 10. The primitive translation vectors of the fcc lattice are shown in Fig. 11. Primitive cells by definition contain only one lattice point, but the conventional bcc cell contains two lattice points, and the fcc cell contains four lattice points.

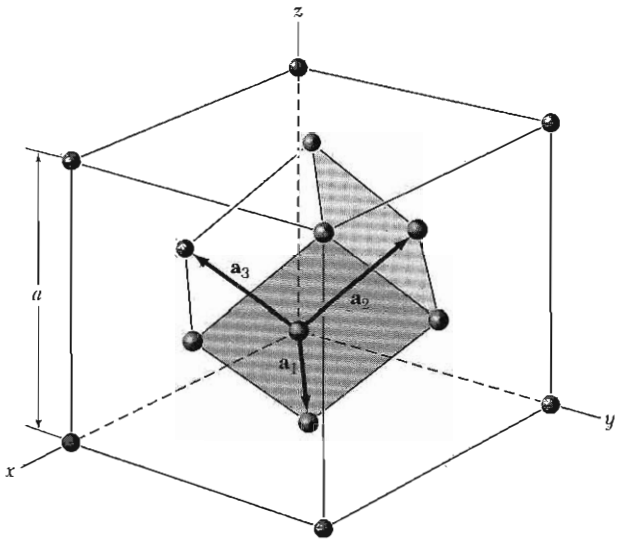


Figure 11 The rhombohedral primitive cell of the face-centered cubic crystal. The primitive translation vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ connect the lattice point at the origin with lattice points at the face centers. As drawn, the primitive vectors are:

$$\mathbf{a}_1 = \tfrac{1}{2}a(\hat{\mathbf{x}} + \hat{\mathbf{y}}) \; ; \quad \mathbf{a}_2 = \tfrac{1}{2}a(\hat{\mathbf{y}} + \hat{\mathbf{z}}) \; ; \quad \mathbf{a}_3 = \tfrac{1}{2}a(\hat{\mathbf{z}} + \hat{\mathbf{x}}) \; .$$

The angles between the axes are 60°.

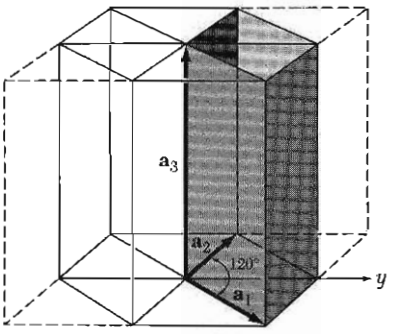


Figure 12 Relation of the primitive cell in the hexagonal system (heavy lines) to a prism of hexagonal symmetry. Here $a_1 = a_2 \neq a_3$.

The position of a point in a cell is specified by (2) in terms of the atomic coordinates x, y, z. Here each coordinate is a fraction of the axial length a_1, a_2, a_3 in the direction of the coordinate axis, with the origin taken at one corner of the cell. Thus the coordinates of the body center of a cell are $\frac{1}{2}\frac{1}{2}\frac{1}{2}$, and the face centers include $\frac{1}{2}\frac{1}{2}0$, $0\frac{1}{2}\frac{1}{2}$; $\frac{1}{2}0\frac{1}{2}$. In the hexagonal system the primitive cell is a right prism based on a rhombus with an included angle of 120°. Figure 12 shows the relationship of the rhombic cell to a hexagonal prism.

INDEX SYSTEM FOR CRYSTAL PLANES

The orientation of a crystal plane is determined by three points in the plane, provided they are not collinear. If each point lay on a different crystal axis, the plane could be specified by giving the coordinates of the points in terms of the lattice constants a_1, a_2, a_3. However, it turns out to be more useful for structure analysis to specify the orientation of a plane by the indices determined by the following rules (Fig. 13).

- Find the intercepts on the axes in terms of the lattice constants a_1, a_2, a_3. The axes may be those of a primitive or nonprimitive cell.

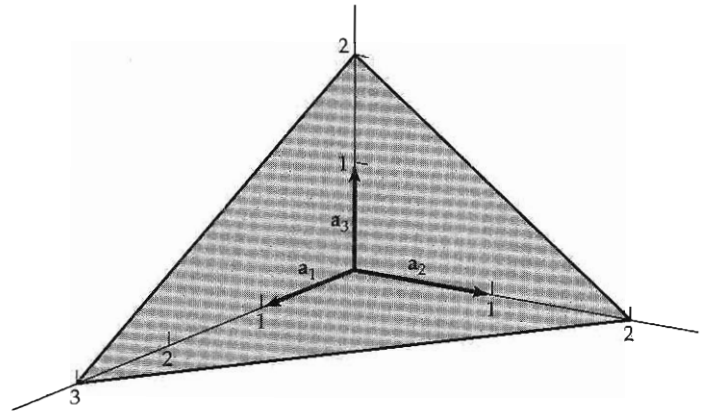


Figure 13 This plane intercepts the $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ axes at $3\mathbf{a}_1$, $2\mathbf{a}_2$, $2\mathbf{a}_3$. The reciprocals of these numbers are $\frac{1}{3}, \frac{1}{2}, \frac{1}{2}$. The smallest three integers having the same ratio are 2, 3, 3, and thus the indices of the plane are (233).

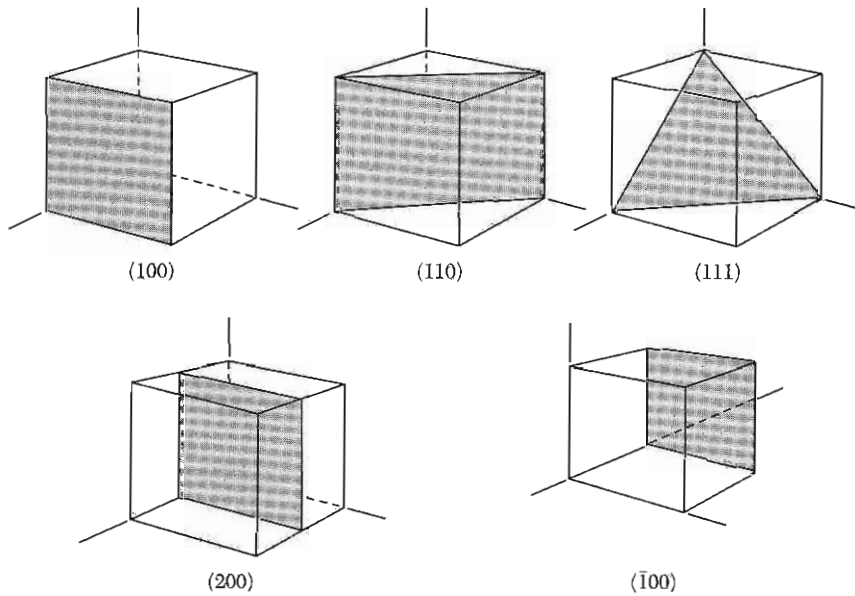


Figure 14 Indices of important planes in a cubic crystal. The plane (200) is parallel to (100) and to (100).

- Take the reciprocals of these numbers and then reduce to three integers having the same ratio, usually the smallest three integers. The result, enclosed in parentheses (*hkl*), is called the index of the plane.

For the plane whose intercepts are 4, 1, 2, the reciprocals are $\frac{1}{4}$, 1, and $\frac{1}{2}$; the smallest three integers having the same ratio are (142). For an intercept at infinity, the corresponding index is zero. The indices of some important planes in a cubic crystal are illustrated by Fig. 14. The indices (*hkl*) may denote a single plane or a set of parallel planes. If a plane cuts an axis on the negative side of the origin, the corresponding index is negative, indicated by placing a minus sign

above the index: $(h\bar{k}l)$. The cube faces of a cubic crystal are (100), (010), (001), $(\bar{1}00)$, $(0\bar{1}0)$, and $(00\bar{1})$. Planes equivalent by symmetry may be denoted by curly brackets (braces) around indices; the set of cube faces is {100}. When we speak of the (200) plane we mean a plane parallel to (100) but cutting the $\mathbf{a}_1$ axis at $\frac{1}{2}a$.

The indices $[uvw]$ of a direction in a crystal are the set of the smallest integers that have the ratio of the components of a vector in the desired direction, referred to the axes. The $\mathbf{a}_1$ axis is the [100] direction; the $-\mathbf{a}_2$ axis is the $[0\bar{1}0]$ direction. In cubic crystals the direction $[hkl]$ is perpendicular to a plane (hkl) having the same indices, but this is not generally true in other crystal systems.

SIMPLE CRYSTAL STRUCTURES

We discuss simple crystal structures of general interest: the sodium chloride, cesium chloride, hexagonal close-packed, diamond, and cubic zinc sulfide structures.

Sodium Chloride Structure

The sodium chloride, NaCl, structure is shown in Figs. 15 and 16. The lattice is face-centered cubic; the basis consists of one Na^+ ion and one Cl^- ion

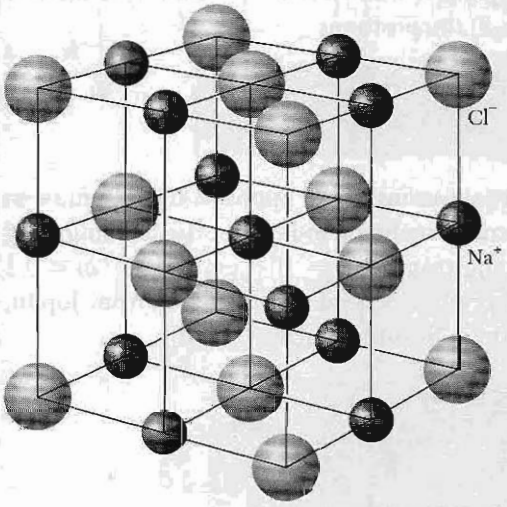


Figure 15 We may construct the sodium chloride crystal structure by arranging Na^+ and Cl^- ions alternately at the lattice points of a simple cubic lattice. In the crystal each ion is surrounded by six nearest neighbors of the opposite charge. The space lattice is fcc, and the basis has one Cl^- ion at 000 and one Na^+ ion at $\frac{1}{2}\frac{1}{2}\frac{1}{2}$. The figure shows one conventional cubic cell. The ionic diameters here are reduced in relation to the cell in order to clarify the spatial arrangement.

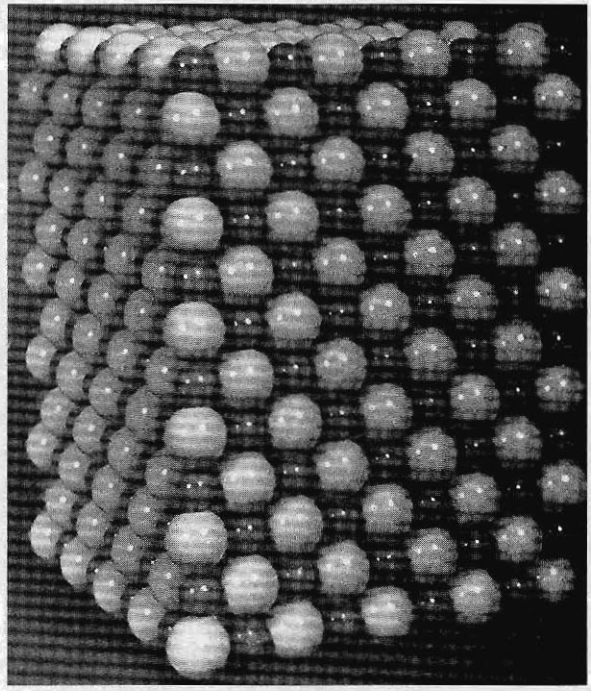


Figure 16 Model of sodium chloride. The sodium ions are smaller than the chlorine ions. (Courtesy of A. N. Holden and P. Singer.)

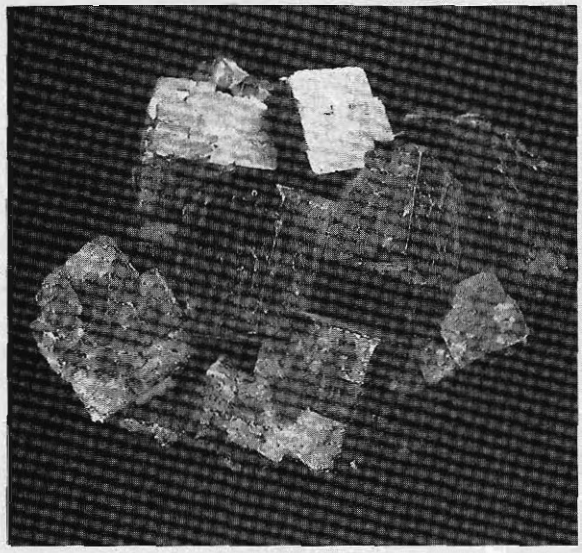


Figure 17 Natural crystals of lead sulfide, PbS, which has the NaCl crystal structure. (Photograph by B. Burleson.)

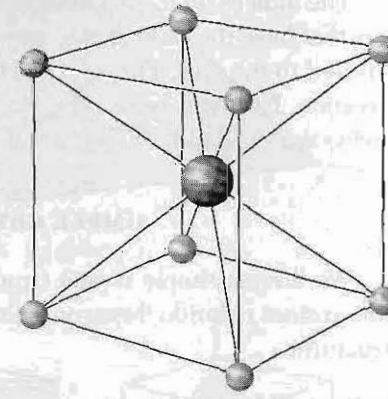


Figure 18 The cesium chloride crystal structure. The space lattice is simple cubic, and the basis has one Cs^+ ion at 000 and one Cl^- ion at $\frac{1}{2}\frac{1}{2}\frac{1}{2}$.

separated by one-half the body diagonal of a unit cube. There are four units of NaCl in each unit cube, with atoms in the positions

Cl:	000 ;	$\frac{1}{2}\frac{1}{2}0$;	$\frac{1}{2}0\frac{1}{2}$;	$0\frac{1}{2}\frac{1}{2}$.
Na:	$\frac{1}{2}\frac{1}{2}\frac{1}{2}$;	$00\frac{1}{2}$;	$0\frac{1}{2}0$;	$\frac{1}{2}00$.

Each atom has as nearest neighbors *six atoms of the opposite kind*. Representative crystals having the NaCl arrangement include those in the following table. The cube edge a is given in angstroms; 1 Å $\equiv 10^{-8}$ cm $\equiv 10^{-10}$ m $\equiv 0.1$ nm. Figure 17 is a photograph of crystals of lead sulfide (PbS) from Joplin, Missouri. The Joplin specimens form in beautiful cubes.

Crystal	a	Crystal	a
LiH	4.08 Å	AgBr	5.77 Å
MgO	4.20	PbS	5.92
MnO	4.43	KCl	6.29
NaCl	5.63	KBr	6.59

Cesium Chloride Structure

The cesium chloride structure is shown in Fig. 18. There is one molecule per primitive cell, with atoms at the corners 000 and body-centered positions $\frac{1}{2}\frac{1}{2}\frac{1}{2}$ of the simple cubic space lattice. Each atom may be viewed as at the center

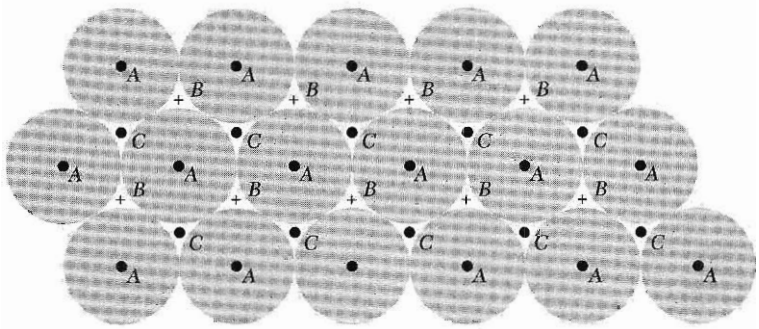


Figure 19 A close-packed layer of spheres is shown, with centers at points marked *A*. A second and identical layer of spheres can be placed on top of this, above and parallel to the plane of the drawing, with centers over the points marked *B*. There are two choices for a third layer. It can go in over *A* or over *C*. If it goes in over *A*, the sequence is *ABABAB* . . . and the structure is hexagonal close-packed. If the third layer goes in over *C*, the sequence is *ABCABCABC* . . . and the structure is face-centered cubic.

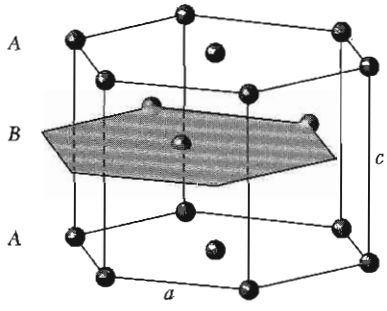


Figure 20 The hexagonal close-packed structure. The atom positions in this structure do not constitute a space lattice. The space lattice is simple hexagonal with a basis of two identical atoms associated with each lattice point. The lattice parameters a and c are indicated, where a is in the basal plane and c is the magnitude of the axis $\mathbf{a}_3$ of Fig. 12.

of a cube of atoms of the opposite kind, so that the number of nearest neighbors or coordination number is eight.

Crystal	a	Crystal	a
BeCu	2.70 Å	LiHg	3.29 Å
AlNi	2.88	NH_4Cl	3.87
CuZn (β-brass)	2.94	TlBr	3.97
CuPd	2.99	CsCl	4.11
AgMg	3.28	TlI	4.20

Hexagonal Close-Packed Structure (hcp)

There are an infinite number of ways of arranging identical spheres in a regular array that maximizes the packing fraction (Fig. 19). One is the face-centered cubic structure; another is the hexagonal close-packed structure (Fig. 20). The fraction of the total volume occupied by the spheres is 0.74 for both structures. No structure, regular or not, has denser packing.

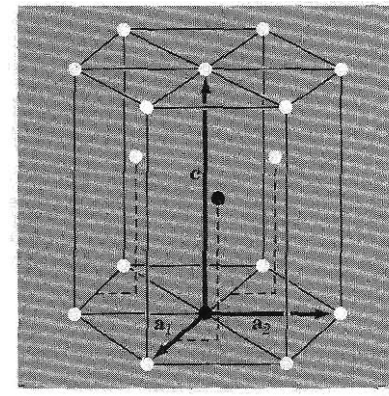


Figure 21 The primitive cell has $a_1 = a_2$, with an included angle of 120°. The c axis (or $\mathbf{a}_3$) is normal to the plane of $\mathbf{a}_1$ and $\mathbf{a}_2$. The ideal hcp structure has $c = 1.633\ a$. The two atoms of one basis are shown as solid circles. One atom of the basis is at the origin; the other atom is at $\frac{2}{3}\frac{1}{3}\frac{1}{2}$, which means at the position $\mathbf{r} = \frac{2}{3}\mathbf{a}_1 + \frac{1}{3}\mathbf{a}_2 + \frac{1}{2}\mathbf{a}_3$.

Spheres are arranged in a single closest-packed layer A by placing each sphere in contact with six others in a plane. This layer may serve as either the basal plane of an hcp structure or the (111) plane of the fcc structure. A second similar layer B may be added by placing each sphere of B in contact with three spheres of the bottom layer, as in Figs. 19–21. A third layer C may be added in two ways. We obtain the fcc structure if the spheres of the third layer are added over the holes in the first layer that are not occupied by B. We obtain the hcp structure when the spheres in the third layer are placed directly over the centers of the spheres in the first layer.

The number of nearest-neighbor atoms is 12 for both hcp and fcc structures. If the binding energy (or free energy) depended only on the number of nearest-neighbor bonds per atom, there would be no difference in energy between the fcc and hcp structures.

Crystal	c/a	Crystal	c/a	Crystal	c/a
He	1.633	Zn	1.861	Zr	1.594
Be	1.581	Cd	1.886	Gd	1.592
Mg	1.623	Co	1.622	Lu	1.586
Ti	1.586	Y	1.570		

Diamond Structure

The diamond structure is the structure of the semiconductors silicon and germanium and is related to the structure of several important semiconductor binary compounds. The space lattice of diamond is face-centered cubic. The primitive basis of the diamond structure has two identical atoms at coordinates 000 and $\frac{1}{4}\frac{1}{4}\frac{1}{4}$ associated with each point of the fcc lattice, as shown in Fig. 22. Because the conventional unit cube of the fcc lattice contains 4 lattice points, it follows that the conventional unit cube of the diamond structure contains $2 \times 4 = 8$ atoms. There is no way to choose a primitive cell such that the basis of diamond contains only one atom.

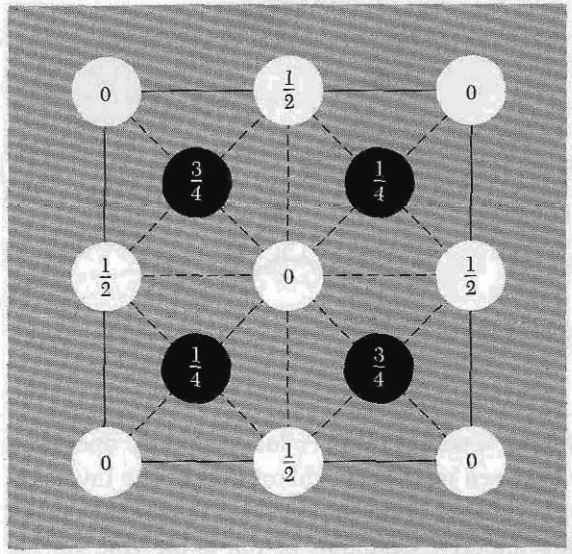


Figure 22 Atomic positions in the cubic cell of the diamond structure projected *on a cube face;* fractions denote height above the base in units of a cube edge. The points at 0 and $\frac{1}{2}$ are on the fcc lattice; those at $\frac{1}{4}$ and $\frac{3}{4}$ are on a similar lattice displaced along the body diagonal by one-fourth of its length. With a fcc space lattice, the basis consists of two identical atoms at 000 and $\frac{1}{4}\frac{1}{4}\frac{1}{4}$.

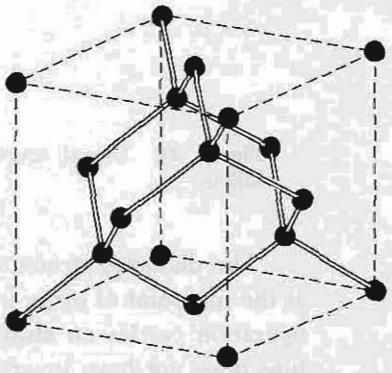


Figure 23 Crystal structure of diamond, showing the tetrahedral bond arrangement.

The tetrahedral bonding characteristic of the diamond structure is shown in Fig. 23. Each atom has 4 nearest neighbors and 12 next nearest neighbors. The diamond structure is relatively empty: the maximum proportion of the available volume which may be filled by hard spheres is only 0.34, which is 46 percent of the filling factor for a closest-packed structure such as fcc or hcp. The diamond structure is an example of the directional covalent bonding found in column IV of the periodic table of elements. Carbon, silicon, germanium, and tin can crystallize in the diamond structure, with lattice constants a = 3.567, 5.430, 5.658, and 6.49 Å, respectively. Here a is the edge of the conventional cubic cell.

Cubic Zinc Sulfide Structure

The diamond structure may be viewed as two fcc structures displaced from each other by one-quarter of a body diagonal. The cubic zinc sulfide (zinc blende) structure results when Zn atoms are placed on one fcc lattice and S atoms on the other fcc lattice, as in Fig. 24. The conventional cell is a cube. The coordinates of the Zn atoms are 000; $0\frac{1}{2}\frac{1}{2}$; $\frac{1}{2}0\frac{1}{2}$; $\frac{1}{2}\frac{1}{2}0$; the coordinates of the S atoms are $\frac{1}{4}\frac{1}{4}\frac{1}{4}$, $\frac{1}{4}\frac{3}{4}\frac{3}{4}$, $\frac{3}{4}\frac{1}{4}\frac{3}{4}$, $\frac{3}{4}\frac{3}{4}\frac{1}{4}$. The lattice is fcc. There are four molecules of ZnS per conventional cell. About each atom there are four equally distant atoms of the opposite kind arranged at the corners of a regular tetrahedron...

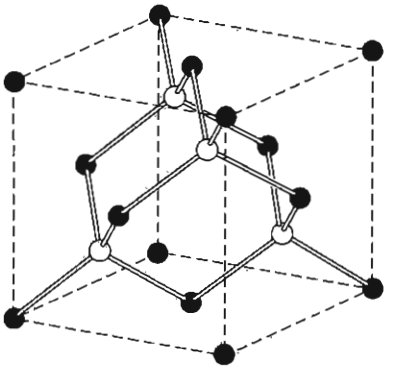


Figure 24 Crystal structure of cubic zinc sulfide.

The diamond structure allows a center-of-inversion symmetry operation at the midpoint of every line between nearest-neighbor atoms. The inversion operation carries an atom at $\mathbf{r}$ into an atom at $-\mathbf{r}$. The cubic ZnS structure does not have inversion symmetry. Examples of the cubic zinc sulfide structure are

Crystal	a	Crystal	a
SiC	4.35 Å	ZnSe	5.65 Å
ZnS	5.41	GaAs	5.65
AlP	5.45	AlAs	5.66
GaP	5.45	InSb	6.46

The close equality of the lattice constants of several pairs, notably (Al, Ga)P and (Al, Ga)As, makes possible the construction of semiconductor heterojunctions (Chapter 19).

DIRECT IMAGING OF ATOMIC STRUCTURE

Direct images of crystal structure have been produced by transmission electron microscopy. Perhaps the most beautiful images are produced by scanning tunneling microscopy; in STM (Chapter 19) one exploits the large variations in quantum tunneling as a function of the height of a fine metal tip above the surface of a crystal. The image of Fig. 25 was produced in this way. An STM method has been developed that will assemble single atoms into an organized layer nanometer structure on a crystal substrate.

NONIDEAL CRYSTAL STRUCTURES

The ideal crystal of classical crystallographers is formed by the periodic repetition of identical units in space. But no general proof has been given that

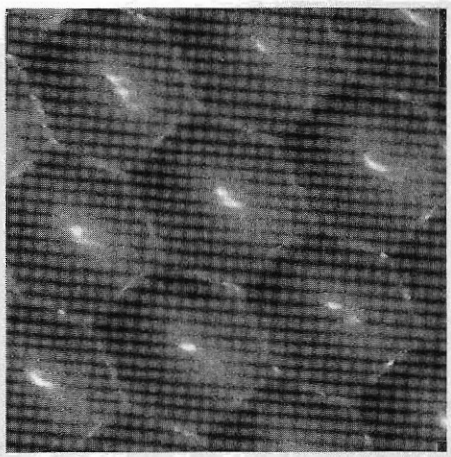


Figure 25 A scanning tunneling microscope image of atoms on a (111) surface of fcc platinum at 4 K. The nearest-neighbor spacing is 2.78 Å. (Photo courtesy of D. M. Eigler, IBM Research Division.)

the ideal crystal is the state of minimum energy of identical atoms at the temperature of absolute zero. At finite temperatures this is likely not to be true. We give a further example here.

Random Stacking and Polytypism

The fcc and hcp structures are made up of close-packed planes of atoms. The structures differ in the stacking sequence of the planes, fcc having the sequence *ABCABC* . . . and hcp having the sequence *ABABAB* . . . . Structures are known in which the stacking sequence of close-packed planes is random. This is known as **random stacking** and may be thought of as crystalline in two dimensions and noncrystalline or glasslike in the third.

Polytypism is characterized by a stacking sequence with a long repeat unit along the stacking axis. The best known example is zinc sulfide, ZnS, in which more than 150 polytypes have been identified, with the longest periodicity being 360 layers. Another example is silicon carbide, SiC, which occurs with more than 45 stacking sequences of the close-packed layers. The polytype of SiC known as 393R has a primitive cell with $a = 3.079$ Å and $c = 989.6$ Å. The longest primitive cell observed for SiC has a repeat distance of 594 layers. A given sequence is repeated many times within a single crystal. The mechanism that induces such long-range crystallographic order is not a long-range force, but arises from spiral steps due to dislocations in the growth nucleus (Chapter 20).

CRYSTAL STRUCTURE DATA

In Table 3 we list the more common crystal structures and lattice structures of the elements. Values of the atomic concentration and the density are given in Table 4. Many elements occur in several crystal structures and transform from

Table 3 Crystal structures of the elements

The data given are at room temperature for the most common form, or at the stated temperature in deg K. (Inorganic Crystal Structure Database (ICSD) online.

H^1 4K hcp 3.75 6.12																	He^4 2K hcp 3.57 5.83
Li 78K bcc 3.491	Be hcp 2.27 3.59											B rhomb.	C diamond 3.567	N 20K cubic 5.66 (N_2)	O complex (O_2)	F	Ne 4K fcc 4.46
Na 5K bcc 4.225	Mg hcp 3.21 5.21	← Crystal structure → ← a lattice parameter, in Å → ← c lattice parameter, in Å →										Al fcc 4.05	Si diamond 5.430	P complex	S complex	Cl complex (Cl_2)	Ar 4K fcc 5.31
K 5K bcc 5.225	Ca fcc 5.58	Sc hcp 3.31 5.27	Ti hcp 2.95 4.68	V bcc 3.03	Cr bcc 2.88	Mn cubic complex	Fe bcc 2.87	Co hcp 2.51 4.07	Ni fcc 3.52	Cu fcc 3.61	Zn hcp 2.66 4.95	Ga complex	Ge diamond 5.658	As rhomb.	Se hex. chains	Br complex (Br_2)	Kr 4K fcc 5.64
Rb 5K bcc 5.585	Sr fcc 6.08	Y hcp 3.65 5.73	Zr hcp 3.23 5.15	Nb bcc 3.30	Mo bcc 3.15	Tc hcp 2.74 4.40	Ru hcp 2.71 4.28	Rh fcc 3.80	Pd fcc 3.89	Ag fcc 4.09	Cd hcp 2.98 5.62	In tetr. 3.25 4.95	Sn (α) diamond 6.49	Sb rhomb.	Te hex. chains	I complex (I_2)	Xe 4K fcc 6.13
Cs 5K bcc 6.045	Ba bcc 5.02	La hex. 3.77 ABAC	Hf hcp 3.19 5.05	Ta bcc 3.30	W bcc 3.16	Re hcp 2.76 4.46	Os hcp 2.74 4.32	Ir fcc 3.84	Pt fcc 3.92	Au fcc 4.08	Hg rhomb.	Tl hcp 3.46 5.52	Pb fcc 4.95	Bi rhomb.	Po sc 3.34	At —	Rn —
Fr —	Ra —	Ac fcc 5.31															

Ce fcc 5.16	Pr hex. 3.67 ABAC	Nd hex. 3.66	Pm —	Sm complex	Eu bcc 4.58	Gd hcp 3.63 5.78	Tb hcp 3.60 5.70	Dy hcp 3.59 5.65	Ho hcp 3.58 5.62	Er hcp 3.56 5.59	Tm hcp 3.54 5.56	Yb fcc 5.48	Lu hcp 3.50 5.55
Th fcc 5.08	Pa tetr. 3.92 3.24	U complex	Np complex	Pu complex	Am hex. 3.64 ABAC	Cm —	Bk —	Cf —	Es —	Fm —	Md —	No —	Lr —

Table 4 Density and atomic concentration

The data are given at atmospheric pressure and room temperature, or at the stated temperature in deg K. (Crystal modifications as for Table 3.)

H 4K																	**He** 2K
0.088																	0.205 (at 37 atm)
Li 78K	**Be**											**B**	**C**	**N** 20K	**O**	**F**	**Ne** 4K
0.542	1.82											2.47	3.516	1.03			1.51
4.700	12.1											13.0	17.6				4.36
3.023	2.22												1.54			1.44	3.16
Na 5K	**Mg**											**Al**	**Si**	**P**	**S**	**Cl** 93K	**Ar** 4K
1.013	1.74	← Density in g cm^{-3} (10^{3}kg m^{-3}) →										2.70	2.33			2.03	1.77
2.652	4.30	← Concentration in 10^{22} cm^{-3} (10^{28} m^{-3}) →										6.02	5.00				2.66
3.659	3.20	← Nearest-neighbor distance, in Å (10^{-10}m) →										2.86	2.35			2.02	3.76
K 5K	**Ca**	**Sc**	**Ti**	**V**	**Cr**	**Mn**	**Fe**	**Co**	**Ni**	**Cu**	**Zn**	**Ga**	**Ge**	**As**	**Se**	**Br** 123K	**Kr** 4K
0.910	1.53	2.99	4.51	6.09	7.19	7.47	7.87	8.9	8.91	8.93	7.13	5.91	5.32	5.77	4.81	4.05	3.09
1.402	2.30	4.27	5.66	7.22	8.33	8.18	8.50	8.97	9.14	8.45	6.55	5.10	4.42	4.65	3.67	2.36	2.17
4.525	3.95	3.25	2.89	2.62	2.50	2.24	2.48	2.50	2.49	2.56	2.66	2.44	2.45	3.16	2.32		4.00
Rb 5K	**Sr**	**Y**	**Zr**	**Nb**	**Mo**	**Tc**	**Ru**	**Rh**	**Pd**	**Ag**	**Cd**	**In**	**Sn**	**Sb**	**Te**	**I**	**Xe** 4K
1.629	2.58	4.48	6.51	8.58	10.22	11.50	12.36	12.42	12.00	10.50	8.65	7.29	5.76	6.69	6.25	4.95	3.78
1.148	1.78	3.02	4.29	5.56	6.42	7.04	7.36	7.26	6.80	5.85	4.64	3.83	2.91	3.31	2.94	2.36	1.64
4.837	4.30	3.55	3.17	2.86	2.72	2.71	2.65	2.69	2.75	2.89	2.98	3.25	2.81	2.91	2.86	3.54	4.34
Cs 5K	**Ba**	**La**	**Hf**	**Ta**	**W**	**Re**	**Os**	**Ir**	**Pt**	**Au**	**Hg** 227	**Tl**	**Pb**	**Bi**	**Po**	**At**	**Rn**
1.997	3.59	6.17	13.20	16.66	19.25	21.03	22.58	22.55	21.47	19.28	14.26	11.87	11.34	9.80	9.31		
0.905	1.60	2.70	4.52	5.55	6.30	6.80	7.14	7.06	6.62	5.90	4.26	3.50	3.30	2.82	2.67	—	—
5.235	4.35	3.73	3.13	2.86	2.74	2.74	2.68	2.71	2.77	2.88	3.01	3.46	3.50	3.07	3.34		
Fr	**Ra**	**Ac**															
		10.07															
—	—	2.66															
		3.76															

Ce	**Pr**	**Nd**	**Pm**	**Sm**	**Eu**	**Gd**	**Tb**	**Dy**	**Ho**	**Er**	**Tm**	**Yb**	**Lu**
6.77	6.78	7.00		7.54	5.25	7.89	8.27	8.53	8.80	9.04	9.32	6.97	9.84
2.91	2.92	2.93	—	3.03	2.04	3.02	3.22	3.17	3.22	3.26	3.32	3.02	3.39
3.65	3.63	3.66		3.59	3.96	3.58	3.52	3.51	3.49	3.47	3.54	3.88	3.43
Th	**Pa**	**U**	**Np**	**Pu**	**Am**	**Cm**	**Bk**	**Cf**	**Es**	**Fm**	**Md**	**No**	**Lr**
11.72	15.37	19.05	20.45	19.81	11.87								
3.04	4.01	4.80	5.20	4.26	2.96	—	—	—	—	—	—	—	—
3.60	3.21	2.75	2.62	3.1	3.61								

one to the other as the temperature or pressure is varied. Sometimes two structures coexist at the same temperature and pressure, although one may be slightly more stable.

SUMMARY

- A lattice is an array of points related by the lattice translation operator $\mathbf{T} = u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + u_3\mathbf{a}_3$, where u_1, u_2, u_3 are integers and $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ are the crystal axes.
- To form a crystal we attach to every lattice point an identical basis composed of s atoms at the positions $\mathbf{r}_j = x_j\mathbf{a}_1 + y_j\mathbf{a}_2 + z_j\mathbf{a}_3$, with $j = 1, 2, \ldots, s$. Here x, y, z may be selected to have values between 0 and 1.
- The axes $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ are primitive for the minimum cell volume $|\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3|$ for which the crystal can be constructed from a lattice translation operator $\mathbf{T}$ and a basis at every lattice point.

Problems

1. ***Tetrahedral angles.*** The angles between the tetrahedral bonds of diamond are the same as the angles between the body diagonals of a cube, as in Fig. 10. Use elementary vector analysis to find the value of the angle.
2. ***Indices of planes.*** Consider the planes with indices (100) and (001); the lattice is fcc, and the indices refer to the conventional cubic cell. What are the indices of these planes when referred to the primitive axes of Fig. 11?
3. ***Hcp structure.*** Show that the c/a ratio for an ideal hexagonal close-packed structure is $(\frac{8}{3})^{1/2} = 1.633$. If c/a is significantly larger than this value, the crystal structure may be thought of as composed of planes of closely packed atoms, the planes being loosely stacked.

2

Wave Diffraction and the Reciprocal Lattice

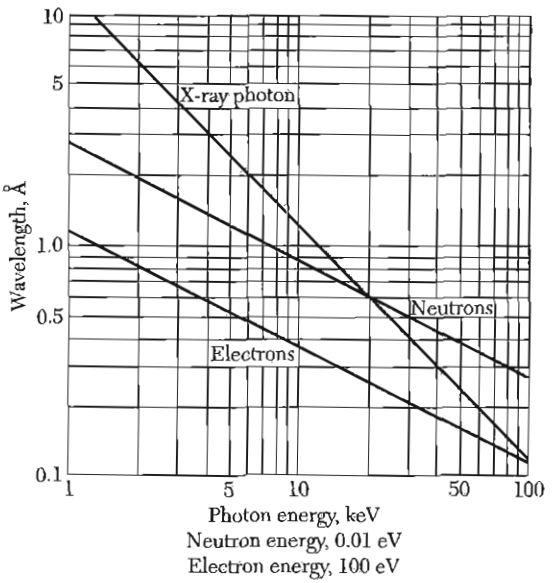


Figure 1 Wavelength versus particle energy, for photons, neutrons, and electrons.

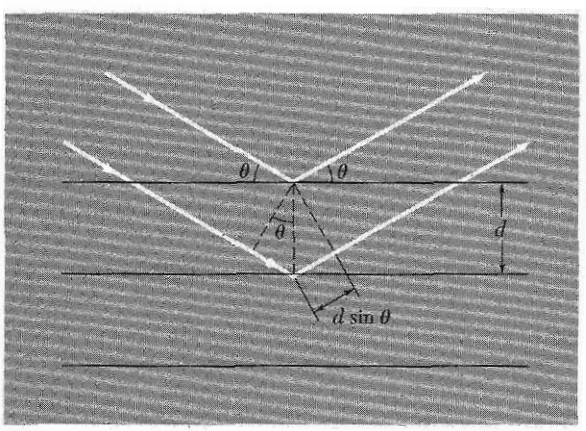


Figure 2 Derivation of the Bragg equation $2d \sin \theta = n\lambda$; here d is the spacing of parallel atomic planes and $2\pi n$ is the difference in phase between reflections from successive planes. The reflecting planes have nothing to do with the surface planes bounding the particular specimen.

CHAPTER 2: WAVE DIFFRACTION AND THE RECIPROCAL LATTICE

DIFFRACTION OF WAVES BY CRYSTALS

The Bragg law

We study crystal structure through the diffraction of photons, neutrons, and electrons (Fig. 1). The diffraction depends on the crystal structure and on the wavelength. At optical wavelengths such as 5000 Å, the superposition of the waves scattered elastically by the individual atoms of a crystal results in ordinary optical refraction. When the wavelength of the radiation is comparable with or smaller than the lattice constant, we may find diffracted beams in directions quite different from the incident direction.

W. L. Bragg presented a simple explanation of the diffracted beams from a crystal. The Bragg derivation is simple but is convincing only because it reproduces the correct result. Suppose that the incident waves are reflected specularly from parallel planes of atoms in the crystal, with each plane reflecting only a very small fraction of the radiation, like a lightly silvered mirror. In specular (mirrorlike) reflection the angle of incidence is equal to the angle of reflection. The diffracted beams are found when the reflections from parallel planes of atoms interfere constructively, as in Fig. 2. We treat elastic scattering, in which the energy of the x-ray is not changed on reflection.

Consider parallel lattice planes spaced d apart. The radiation is incident in the plane of the paper. The path difference for rays reflected from adjacent planes is $2d \sin \theta$, where θ is measured from the plane. Constructive interference of the radiation from successive planes occurs when the path difference is an integral number n of wavelengths λ, so that

$$2d \sin \theta = n\lambda \quad . \tag{1}$$

This is the Bragg law, which can be satisfied only for wavelength $\lambda \leq 2d$.

Although the reflection from each plane is specular, for only certain values of θ will the reflections from all periodic parallel planes add up in phase to give a strong reflected beam. If each plane were perfectly reflecting, only the first plane of a parallel set would see the radiation, and any wavelength would be reflected. But each plane reflects 10^{-3} to 10^{-5} of the incident radiation, so that 10^3 to 10^5 planes may contribute to the formation of the Bragg-reflected beam in a perfect crystal. Reflection by a single plane of atoms is treated in Chapter 17 on surface physics.

The Bragg law is a consequence of the periodicity of the lattice. Notice that the law does not refer to the composition of the basis of atoms associated

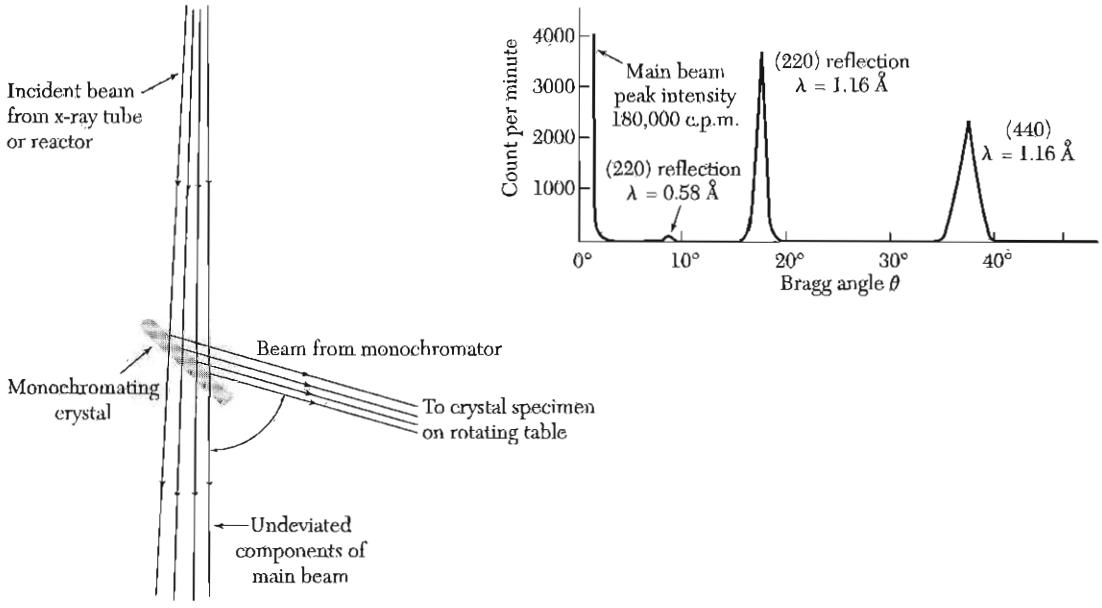


Figure 3 Sketch of a monochromator which by Bragg reflection selects a narrow spectrum of x-ray or neutron wavelengths from a broad spectrum incident beam. The upper part of the figure shows the analysis (obtained by reflection from a second crystal) of the purity of a 1.16 Å beam of neutrons from a calcium fluoride crystal monochromator. (After G. Bacon.)

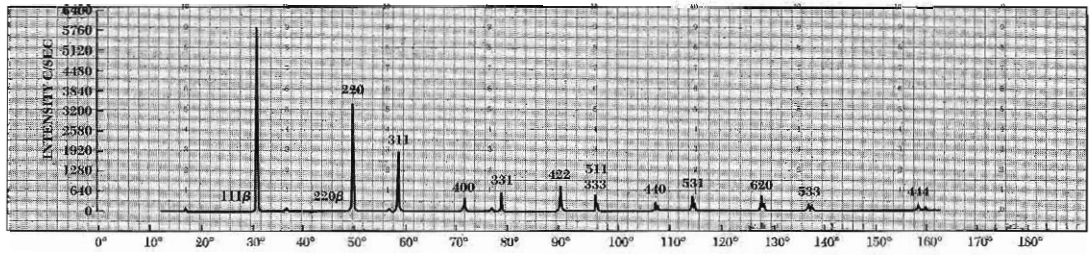


Figure 4 X-ray diffractometer recording of powdered silicon, showing a counter recording of the diffracted beams. (Courtesy of W. Parrish.)

with every lattice point. We shall see, however, that the composition of the basis determines the relative intensity of the various orders of diffraction (denoted by n above) from a given set of parallel planes. Bragg reflection from a single crystal is shown in Fig. 3 and from a powder in Fig. 4.

SCATTERED WAVE AMPLITUDE

The Bragg derivation of the diffraction condition (1) gives a neat statement of the condition for the constructive interference of waves scattered from the lattice points. We need a deeper analysis to determine the scattering

intensity from the basis of atoms, which means from the spatial distribution of electrons within each cell.

Fourier Analysis

We have seen that a crystal is invariant under any translation of the form $\mathbf{T} = u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + u_3\mathbf{a}_3$, where u_1, u_2, u_3 are integers and $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ are the crystal axes. Any local physical property of the crystal, such as the charge concentration, electron number density, or magnetic moment density is invariant under $\mathbf{T}$. What is most important to us here is that the electron number density $n(\mathbf{r})$ is a periodic function of $\mathbf{r}$, with periods $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ in the directions of the three crystal axes, respectively. Thus

$$n(\mathbf{r} + \mathbf{T}) = n(\mathbf{r}) \ . \tag{2}$$

Such periodicity creates an ideal situation for Fourier analysis. The most interesting properties of crystals are directly related to the Fourier components of the electron density.

We consider first a function $n(x)$ in one dimension with period a in the direction x. We expand $n(x)$ in a Fourier series of sines and cosines:

$$n(x) = n_0 + \sum_{p>0} [C_p \cos(2\pi px/a) + S_p \sin(2\pi px/a)] \ , \tag{3}$$

where the p are positive integers and C_p, S_p are real constants, called the Fourier coefficients of the expansion. The factor $2\pi/a$ in the arguments ensures that $n(x)$ has the period a:

$$\begin{aligned} n(x + a) &= n_0 + \sum [C_p \cos(2\pi px/a + 2\pi p) + S_p \sin(2\pi px/a + 2\pi p)] \\ &= n_0 + \sum [C_p \cos(2\pi px/a) + S_p \sin(2\pi px/a)] = n(x) \ . \end{aligned} \tag{4}$$

We say that $2\pi p/a$ is a point in the reciprocal lattice or Fourier space of the crystal. In one dimension these points lie on a line. The **reciprocal lattice points** tell us the allowed terms in the Fourier series (4) or (5). A term is allowed if it is consistent with the periodicity of the crystal, as in Fig. 5; other

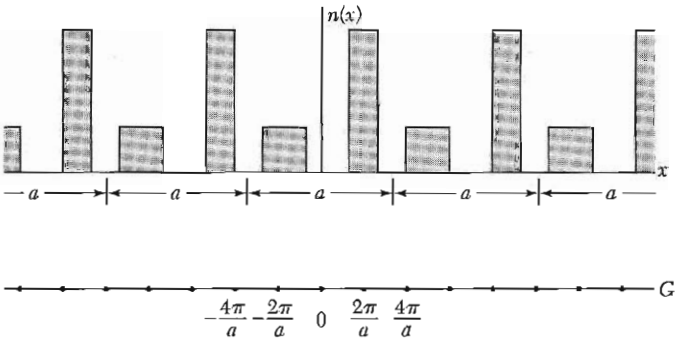


Figure 5 period a. may app $n(x) = \sum$

points in the reciprocal space are not allowed in the Fourier expansion of a periodic function.

It is convenient to write the series (4) in the compact form

$$n(x) = \sum_p n_p \exp(i2\pi px/a) \ , \tag{5}$$

where the sum is over all integers p: positive, negative, and zero. The coefficients n_p now are complex numbers. To ensure that $n(x)$ is a real function, we require

$$n^*_{-p} = n_p \ , \tag{6}$$

for then the sum of the terms in p and $-p$ is real. The asterisk on n^*_{-p} denotes the complex conjugate of n_{-p}.

With $\varphi = 2\pi px/a$, the sum of the terms in p and $-p$ in (5) is real if (6) is satisfied. The sum is

$$\begin{aligned} n_p(\cos\varphi + i\sin\varphi) + n_{-p}(\cos\varphi - i\sin\varphi) \\ = (n_p + n_{-p})\cos\varphi + i(n_p - n_{-p})\sin\varphi \ , \end{aligned} \tag{7}$$

which in turn is equal to the real function

$$2\mathrm{Re}\{n_p\}\cos\varphi - 2\mathrm{Im}\{n_p\}\sin\varphi \tag{8}$$

if (6) is satisfied. Here $\mathrm{Re}\{n_p\}$ and $\mathrm{Im}\{n_p\}$ are real and denote the real and imaginary parts of n_p. Thus the number density $n(x)$ is a real function, as desired.

The extension of the Fourier analysis to periodic functions $n(\mathbf{r})$ in three dimensions is straightforward. We must find a set of vectors $\mathbf{G}$ such that

$$n(\mathbf{r}) = \sum_{\mathbf{G}} n_{\mathbf{G}} \exp(i\mathbf{G}\cdot\mathbf{r}) \tag{9}$$

is invariant under all crystal translations $\mathbf{T}$ that leave the crystal invariant. It will be shown below that the set of Fourier coefficients $n_{\mathbf{G}}$ determines the x-ray scattering amplitude.

Inversion of Fourier Series. We now show that the Fourier coefficient n_p in the series (5) is given by

$$n_p = a^{-1}\int_0^a dx\, n(x)\exp(-i2\pi px/a) \ . \tag{10}$$

Substitute (5) in (10) to obtain

$$n_p = a^{-1}\sum_{p'} n_{p'}\int_0^a dx \exp[i2\pi(p'-p)x/a] \ . \tag{11}$$

If $p' \neq p$ the value of the integral is

$$\frac{a}{i2\pi(p'-p)}(e^{i2\pi(p'-p)}-1)=0 \ ,$$

because $p'-p$ is an integer and $\exp[i2\pi(\text{integer})] = 1$. For the term $p' = p$ the integrand is $\exp(i0) = 1$, and the value of the integral is a, so that $n_p = a^{-1}n_p a = n_p$, which is an identity, so that (10) is an identity.

As in (10), the inversion of (9) gives

$$n_{\mathbf{G}} = V_c^{-1}\int_{\text{cell}} dV\, n(\mathbf{r})\exp(-i\mathbf{G}\cdot\mathbf{r}) \ . \tag{12}$$

Here V_c is the volume of a cell of the crystal.

Reciprocal Lattice Vectors

To proceed further with the Fourier analysis of the electron concentration we must find the vectors $\mathbf{G}$ of the Fourier sum $\Sigma n_{\mathbf{G}} \exp(i\mathbf{G}\cdot\mathbf{r})$ as in (9). There is a powerful, somewhat abstract procedure for doing this. The procedure forms the theoretical basis for much of solid state physics, where Fourier analysis is the order of the day.

We construct the axis vectors $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ of the **reciprocal lattice**:

$$\mathbf{b}_1 = 2\pi\frac{\mathbf{a}_2\times\mathbf{a}_3}{\mathbf{a}_1\cdot\mathbf{a}_2\times\mathbf{a}_3} \ ; \qquad \mathbf{b}_2 = 2\pi\frac{\mathbf{a}_3\times\mathbf{a}_1}{\mathbf{a}_1\cdot\mathbf{a}_2\times\mathbf{a}_3} \ ; \qquad \mathbf{b}_3 = 2\pi\frac{\mathbf{a}_1\times\mathbf{a}_2}{\mathbf{a}_1\cdot\mathbf{a}_2\times\mathbf{a}_3} \ . \tag{13}$$

The factors 2π are not used by crystallographers but are convenient in solid state physics.

If $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ are primitive vectors of the crystal lattice, then $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ are **primitive vectors of the reciprocal lattice**. Each vector defined by (13) is orthogonal to two axis vectors of the crystal lattice. Thus $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ have the property

$$\mathbf{b}_i\cdot\mathbf{a}_j = 2\pi\delta_{ij} \ , \tag{14}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

Points in the reciprocal lattice are mapped by the set of vectors

$$\mathbf{G} = v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + v_3\mathbf{b}_3 \ , \tag{15}$$

where v_1, v_2, v_3 are integers. A vector $\mathbf{G}$ of this form is a **reciprocal lattice** vector.

The vectors $\mathbf{G}$ in the Fourier series (9) are just the reciprocal lattice vectors (15), for then the Fourier series representation of the electron density has the desired invariance under any crystal translation $\mathbf{T} = u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + u_3\mathbf{a}_3$. From (9),

$$n(\mathbf{r}+\mathbf{T}) = \sum_{\mathbf{G}} n_{\mathbf{G}}\exp(i\mathbf{G}\cdot\mathbf{r})\exp(i\mathbf{G}\cdot\mathbf{T}) \ . \tag{16}$$

But $\exp(i\mathbf{G}\cdot\mathbf{T}) = 1$, because

$$\begin{aligned}\exp(i\mathbf{G}\cdot\mathbf{T}) &= \exp[i(v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + v_3\mathbf{b}_3)\cdot(u_1\mathbf{a}_1 + u_2\mathbf{a}_2 + u_3\mathbf{a}_3)] \\ &= \exp[i2\pi(v_1u_1 + v_2u_2 + v_3u_3)]\ .\end{aligned} \tag{17}$$

The argument of the exponential has the form $2\pi i$ times an integer, because $v_1u_1 + v_2u_2 + v_3u_3$ is an integer, being the sum of products of integers. Thus by (9) we have the desired invariance, $n(\mathbf{r} + \mathbf{T}) = n(\mathbf{r}) = \sum n_{\mathbf{G}} \exp(i\mathbf{G}\cdot\mathbf{r})$.

Every crystal structure has two lattices associated with it, the crystal lattice and the reciprocal lattice. A diffraction pattern of a crystal is, as we shall show, a map of the reciprocal lattice of the crystal. A microscope image, if it could be resolved on a fine enough scale, is a map of the crystal structure in real space. The two lattices are related by the definitions (13). Thus when we rotate a crystal in a holder, we rotate both the direct lattice and the reciprocal lattice.

Vectors in the direct lattice have the dimensions of [length]; vectors in the reciprocal lattice have the dimensions of [1/length]. The reciprocal lattice is a lattice in the Fourier space associated with the crystal. The term is motivated below. Wavevectors are always drawn in Fourier space, so that every position in Fourier space may have a meaning as a description of a wave, but there is a special significance to the points defined by the set of $\mathbf{G}$'s associated with a crystal structure.

Diffraction Conditions

Theorem. The set of reciprocal lattice vectors $\mathbf{G}$ determines the possible x-ray reflections.

We see in Fig. 6 that the difference in phase factors is $\exp[i(\mathbf{k} - \mathbf{k}')\cdot\mathbf{r}]$ between beams scattered from volume elements $\mathbf{r}$ apart. The wavevectors of the incoming and outgoing beams are $\mathbf{k}$ and $\mathbf{k}'$. We suppose that the amplitude

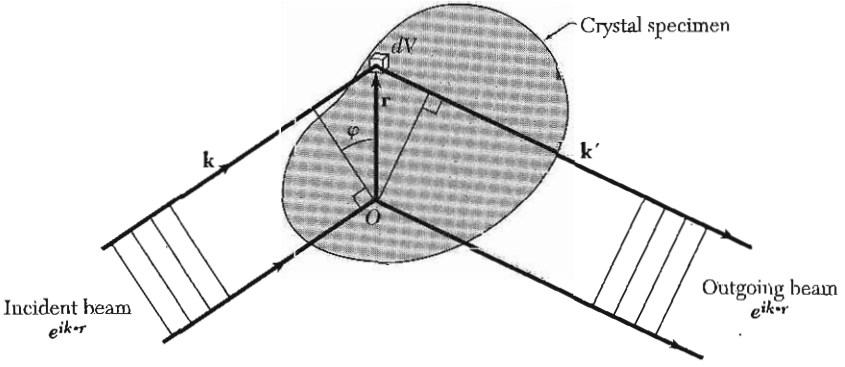


Figure 6 The difference in path length of the incident wave $\mathbf{k}$ at the points O, $\mathbf{r}$ is $r \sin \varphi$, and the difference in phase angle is $(2\pi r \sin \varphi)/\lambda$, which is equal to $\mathbf{k}\cdot\mathbf{r}$. For the diffracted wave the difference in phase angle is $-\mathbf{k}'\cdot\mathbf{r}$. The total difference in phase angle is $(\mathbf{k} - \mathbf{k}')\cdot\mathbf{r}$, and the wave scattered from dV at $\mathbf{r}$ has the phase factor $\exp[i(\mathbf{k} - \mathbf{k}')\cdot\mathbf{r}]$ relative to the wave scattered from a volume element at the origin O.

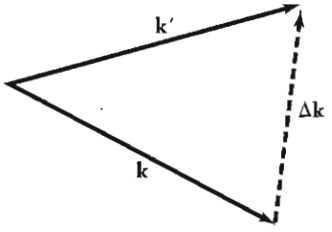


Figure 7 Definition of the scattering vector $\Delta\mathbf{k}$ such that $\mathbf{k} + \Delta\mathbf{k} = \mathbf{k}'$. In elastic scattering the magnitudes satisfy $k' = k$. Further, in Bragg scattering from a periodic lattice, any allowed $\Delta\mathbf{k}$ must equal some reciprocal lattice vector $\mathbf{G}$.

of the wave scattered from a volume element is proportional to the local electron concentration $n(\mathbf{r})$. The total amplitude of the scattered wave in the direction of $\mathbf{k}'$ is proportional to the integral over the crystal of $n(\mathbf{r})\,dV$ times the phase factor $\exp[i(\mathbf{k} - \mathbf{k}') \cdot \mathbf{r}]$.

In other words, the amplitude of the electric or magnetic field vectors in the scattered electromagnetic wave is proportional to the following integral which defines the quantity F that we call the **scattering amplitude**:

$$F = \int dV\, n(\mathbf{r}) \exp[i(\mathbf{k} - \mathbf{k}') \cdot \mathbf{r}] = \int dV\, n(\mathbf{r}) \exp(-i\Delta\mathbf{k} \cdot \mathbf{r}) \ , \tag{18}$$

where $\mathbf{k} - \mathbf{k}' = -\Delta\mathbf{k}$, or

$$\mathbf{k} + \Delta\mathbf{k} = \mathbf{k}' \ . \tag{19}$$

Here $\Delta\mathbf{k}$ measures the change in wavevector and is called the **scattering vector** (Fig. 7). We add $\Delta\mathbf{k}$ to $\mathbf{k}$ to obtain $\mathbf{k}'$, the wavevector of the scattered beam.

We introduce into (18) the Fourier components (9) of $n(\mathbf{r})$ to obtain for the scattering amplitude

$$F = \sum_{\mathbf{G}} \int dV\, n_{\mathbf{G}} \exp[i(\mathbf{G} - \Delta\mathbf{k}) \cdot \mathbf{r}] \ . \tag{20}$$

When the scattering vector $\Delta\mathbf{k}$ is equal to a particular reciprocal lattice vector,

$$\Delta\mathbf{k} = \mathbf{G} \ , \tag{21}$$

the argument of the exponential vanishes and $F = Vn_G$. It is a simple exercise (Problem 4) to show that F is negligibly small when $\Delta\mathbf{k}$ differs significantly from any reciprocal lattice vector.

In elastic scattering of a photon its energy $\hbar\omega$ is conserved, so that the frequency $\omega' = ck'$ of the emergent beam is equal to the frequency of the incident beam. Thus the magnitudes k and k' are equal, and $k^2 = k'^2$, a result that holds also for elastic scattering of electron and neutron beams. From (21) we found $\Delta\mathbf{k} = \mathbf{G}$ or $\mathbf{k} + \mathbf{G} = \mathbf{k}'$, so that the **diffraction condition** is written as $(\mathbf{k} + \mathbf{G})^2 = k^2$, or

$$2\mathbf{k} \cdot \mathbf{G} + G^2 = 0 \ . \tag{22}$$

This is the central result of the theory of elastic scattering of waves in a periodic lattice. If $\mathbf{G}$ is a reciprocal lattice vector, so is $-\mathbf{G}$, and with this substitution we can write (22) as

$$2\mathbf{k} \cdot \mathbf{G} = G^2 \ , \tag{23}$$

This particular expression is often used as the condition for diffraction.

Equation (23) is another statement of the Bragg condition (1). The result of Problem 1 is that the spacing $d(hkl)$ between parallel lattice planes that are normal to the direction $\mathbf{G} = h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3$ is $d(hkl) = 2\pi/|\mathbf{G}|$. Thus the result $2\mathbf{k} \cdot \mathbf{G} = G^2$ may be written as

$$2(2\pi/\lambda) \sin \theta = 2\pi/d(hkl) \ ,$$

or $2d(hkl) \sin \theta = \lambda$. Here θ is the angle between the incident beam and the crystal plane.

The integers hkl that define $\mathbf{G}$ are not necessarily identical with the indices of an actual crystal plane, because the hkl may contain a common factor n, whereas in the definition of the indices in Chapter 1 the common factor has been eliminated. We thus obtain the Bragg result:

$$2d \sin \theta = n\lambda \ , \tag{24}$$

where d is the spacing between adjacent parallel planes with indices h/n, k/n, l/n.

Laue Equations

The original result (21) of diffraction theory, namely that $\Delta\mathbf{k} = \mathbf{G}$, may be expressed in another way to give what are called the Laue equations. These are valuable because of their geometrical representation. Take the scalar product of both $\Delta\mathbf{k}$ and $\mathbf{G}$ successively with $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$. From (14) and (15) we get

$$\mathbf{a}_1 \cdot \Delta\mathbf{k} = 2\pi v_1 \ ; \qquad \mathbf{a}_2 \cdot \Delta\mathbf{k} = 2\pi v_2 \ ; \qquad \mathbf{a}_3 \cdot \Delta\mathbf{k} = 2\pi v_3 \ . \tag{25}$$

These equations have a simple geometrical interpretation. The first equation $\mathbf{a}_1 \cdot \Delta\mathbf{k} = 2\pi v_1$ tells us that $\Delta\mathbf{k}$ lies on a certain cone about the direction of $\mathbf{a}_1$. The second equation tells us that $\Delta\mathbf{k}$ lies on a cone about $\mathbf{a}_2$ as well, and the third equation requires that $\Delta\mathbf{k}$ lies on a cone about $\mathbf{a}_3$. Thus, at a reflection $\Delta\mathbf{k}$ must satisfy all three equations; it must lie at the common line of intersection of *three* cones, which is a severe condition that can be satisfied only by systematic sweeping or searching in wavelength or crystal orientation—or by sheer accident.

A beautiful construction, the Ewald construction, is exhibited in Fig. 8. This helps us visualize the nature of the accident that must occur in order to satisfy the diffraction condition in three dimensions.

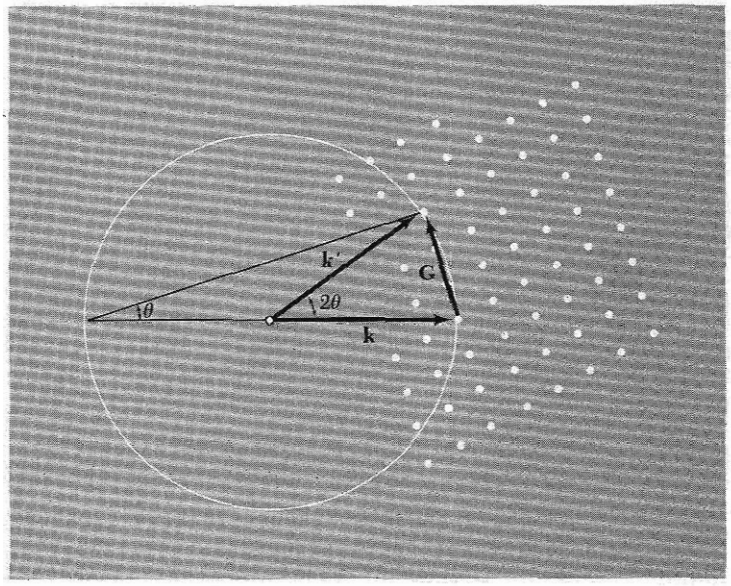


Figure 8 The points on the right-hand side are reciprocal-lattice points of the crystal. The vector **k** is drawn in the direction of the incident x-ray beam, and the origin is chosen such that **k** terminates at any reciprocal lattice point. We draw a sphere of radius $k = 2\pi/\lambda$ about the origin of **k**. A diffracted beam will be formed if this sphere intersects any other point in the reciprocal lattice. The sphere as drawn intercepts a point connected with the end of **k** by a reciprocal lattice vector **G**. The diffracted x-ray beam is in the direction $\mathbf{k}' = \mathbf{k} + \mathbf{G}$. The angle θ is the Bragg angle of Fig. 2. This construction is due to P. P. Ewald.

BRILLOUIN ZONES

Brillouin gave the statement of the diffraction condition that is most widely used in solid state physics, which means in the description of electron energy band theory and of the elementary excitations of other kinds. A *Brillouin zone is defined as a Wigner-Seitz primitive cell in the reciprocal lattice.* (The construction in the direct lattice was shown in Fig. 1.4.) The Brillouin zone gives a vivid geometrical interpretation of the diffraction condition $2\mathbf{k} \cdot \mathbf{G} = G^2$ of Eq. (23). We divide both sides by 4 to obtain

$$\mathbf{k} \cdot (\tfrac{1}{2}\mathbf{G}) = (\tfrac{1}{2}G)^2 \ . \tag{26}$$

We now work in reciprocal space, the space of the **k**'s and **G**'s. Select a vector **G** from the origin to a reciprocal lattice point. Construct a plane normal to this vector **G** at its midpoint. This plane forms a part of a zone boundary (Fig. 9a). An x-ray beam in the crystal will be diffracted if its wavevector **k** has the magnitude and direction required by (26). The diffracted beam will then be in the direction $\mathbf{k} - \mathbf{G}$, as we see from (19) with $\Delta\mathbf{k} = -\mathbf{G}$. Thus the Brillouin construction exhibits all the wavevectors **k** which can be Bragg-reflected by the crystal.

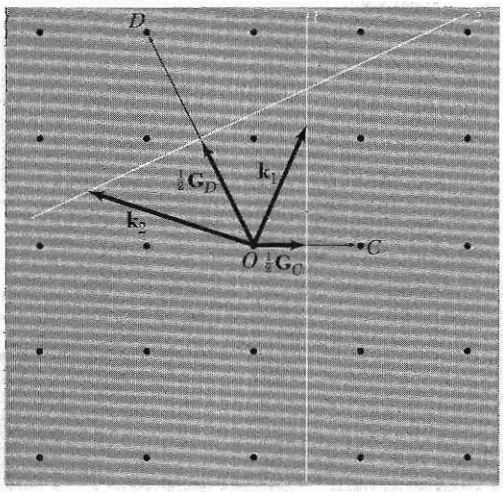


Figure 9a Reciprocal lattice points near the point O at the origin of the reciprocal lattice. The reciprocal lattice vector $\mathbf{G}_C$ connects points OC; and $\mathbf{G}_D$ connects OD. Two planes 1 and 2 are drawn which are the perpendicular bisectors of $\mathbf{G}_C$ and $\mathbf{G}_D$, respectively. Any vector from the origin to the plane 1, such as $\mathbf{k}_1$, will satisfy the diffraction condition $\mathbf{k}_1 \cdot (\frac{1}{2}\mathbf{G}_C) = (\frac{1}{2}\mathbf{G}_C)^2$. Any vector from the origin to the plane 2, such as $\mathbf{k}_2$, will satisfy the diffraction condition $\mathbf{k}_2 \cdot (\frac{1}{2}\mathbf{G}_D) = (\frac{1}{2}\mathbf{G}_D)^2$.

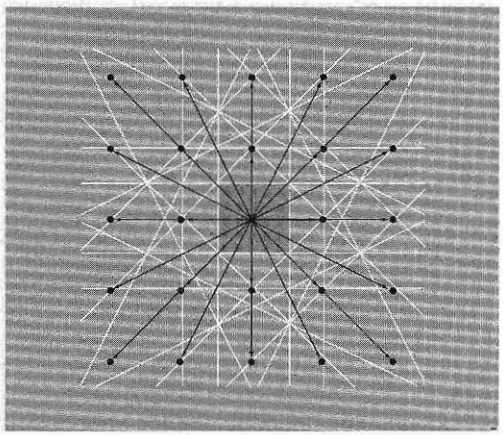


Figure 9b Square reciprocal lattice with reciprocal lattice vectors shown as fine black lines. The lines shown in white are perpendicular bisectors of the reciprocal lattice vectors. The central square is the smallest volume about the origin which is bounded entirely by white lines. The square is the Wigner-Seitz primitive cell of the reciprocal lattice. It is called the first Brillouin zone.

The set of planes that are the perpendicular bisectors of the reciprocal lattice vectors is of general importance in the theory of wave propagation in crystals: A wave whose wavevector drawn from the origin terminates on any of these planes will satisfy the condition for diffraction. These planes divide the Fourier space of the crystal into fragments, as shown in Fig. 9b for a square lattice. The central square is a primitive cell of the reciprocal lattice. It is a Wigner-Seitz cell of the reciprocal lattice.

The central cell in the reciprocal lattice is of special importance in the theory of solids, and we call it the first Brillouin zone. *The first Brillouin zone is the smallest volume entirely enclosed by planes that are the perpendicular bisectors of the reciprocal lattice vectors drawn from the origin.* Examples are shown in Figs. 10 and 11.

Historically, Brillouin zones are not part of the language of x-ray diffraction analysis of crystal structures, but the zones are an essential part of the analysis of the electronic energy-band structure of crystals.

Reciprocal Lattice to sc Lattice

The primitive translation vectors of a simple cubic lattice may be taken as the set

$$\mathbf{a}_1 = a\hat{\mathbf{x}} \; ; \qquad \mathbf{a}_2 = a\hat{\mathbf{y}} \; ; \qquad \mathbf{a}_3 = a\hat{\mathbf{z}} \; . \tag{27a}$$

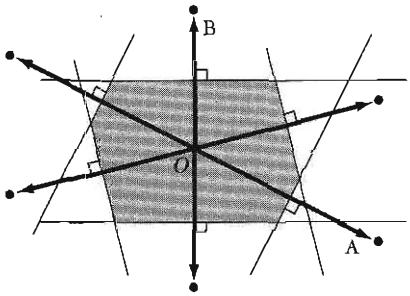


Figure 10 Construction of the first Brillouin zone for an oblique lattice in two dimensions. We first draw a number of vectors from O to nearby points in the reciprocal lattice. Next we construct lines perpendicular to these vectors at their midpoints. The smallest enclosed area is the first Brillouin zone.

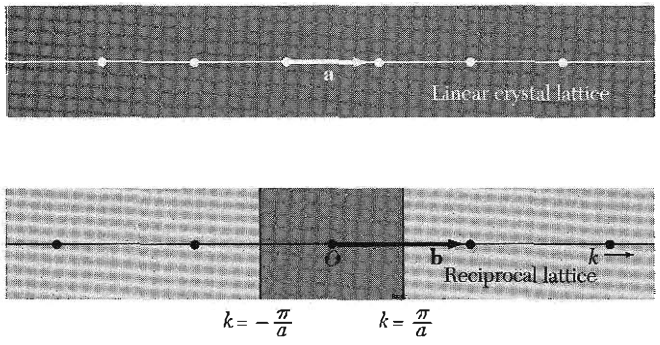


Figure 11 Crystal and reciprocal lattices in one dimension. The basis vector in the reciprocal lattice is **b**, of length equal to $2\pi/a$. The shortest reciprocal lattice vectors from the origin are **b** and $-\mathbf{b}$. The perpendicular bisectors of these vectors form the boundaries of the first Brillouin zone. The boundaries are at $k = \pm\pi/a$.

Here $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$ are orthogonal vectors of unit length. The volume of the cell is $\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3 = a^3$. The primitive translation vectors of the reciprocal lattice are found from the standard prescription (13):

$$\mathbf{b}_1 = (2\pi/a)\hat{\mathbf{x}} \; ; \qquad \mathbf{b}_2 = (2\pi/a)\hat{\mathbf{y}} \; ; \qquad \mathbf{b}_3 = (2\pi/a)\hat{\mathbf{z}} \; . \tag{27b}$$

Here the reciprocal lattice is itself a simple cubic lattice, now of lattice constant $2\pi/a$.

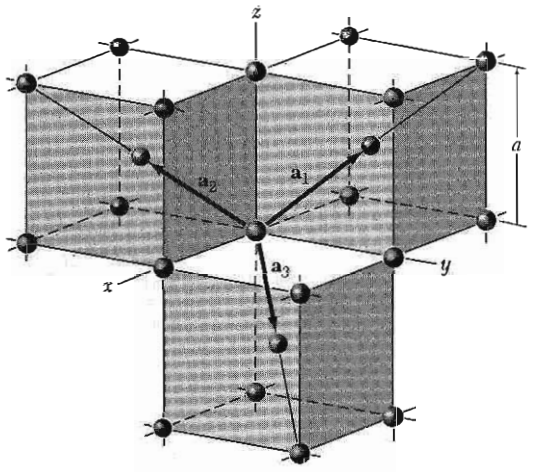


Figure 12 Primitive basis vectors of the body-centered cubic lattice.

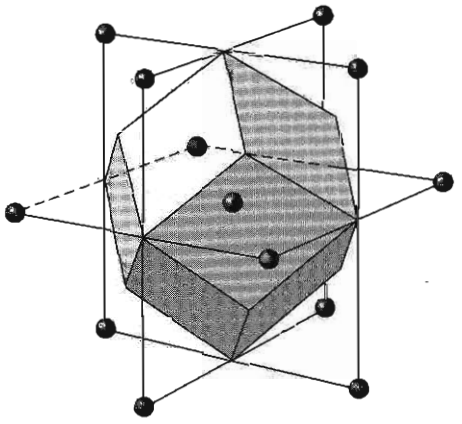


Figure 13 First Brillouin zone of the body-centered cubic lattice. The figure is a regular rhombic dodecahedron.

The boundaries of the first Brillouin zones are the planes normal to the six reciprocal lattice vectors $\pm\mathbf{b}_1$, $\pm\mathbf{b}_2$, $\pm\mathbf{b}_3$ at their midpoints:

$$\pm\tfrac{1}{2}\mathbf{b}_1 = \pm(\pi/a)\hat{\mathbf{x}} \; ; \qquad \pm\tfrac{1}{2}\mathbf{b}_2 = \pm(\pi/a)\hat{\mathbf{y}} \; ; \qquad \pm\tfrac{1}{2}\mathbf{b}_3 = \pm(\pi/a)\hat{\mathbf{z}} \; . \tag{28}$$

The six planes bound a cube of edge $2\pi/a$ and of volume $(2\pi/a)^3$; this cube is the first Brillouin zone of the sc crystal lattice.

Reciprocal Lattice to bcc Lattice

The primitive translation vectors of the bcc lattice (Fig. 12) are

$$\mathbf{a}_1 = \tfrac{1}{2}a(-\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}}) \; ; \quad \mathbf{a}_2 = \tfrac{1}{2}a(\hat{\mathbf{x}} - \hat{\mathbf{y}} + \hat{\mathbf{z}}) \; ; \quad \mathbf{a}_3 = \tfrac{1}{2}a(\hat{\mathbf{x}} + \hat{\mathbf{y}} - \hat{\mathbf{z}}) \; , \tag{29}$$

where a is the side of the conventional cube and $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$ are orthogonal unit vectors parallel to the cube edges. The volume of the primitive cell is

$$V = |\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3| = \tfrac{1}{2}a^3 \; . \tag{30}$$

The primitive translations of the reciprocal lattice are defined by (13). We have, using (28),

$$\mathbf{b}_1 = (2\pi/a)(\hat{\mathbf{y}} + \hat{\mathbf{z}}) \; ; \qquad \mathbf{b}_2 = (2\pi/a)(\hat{\mathbf{x}} + \hat{\mathbf{z}}) \; ; \qquad \mathbf{b}_3 = (2\pi/a)(\hat{\mathbf{x}} + \hat{\mathbf{y}}) \; . \tag{31}$$

Note by comparison with Fig. 14 (p. 37) that these are just the primitive vectors of an fcc lattice, so that an fcc lattice is the reciprocal lattice of the bcc lattice.

The general reciprocal lattice vector is, for integral v_1, v_2, v_3,

$$\mathbf{G} = v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + v_3\mathbf{b}_3 = (2\pi/a)[(v_2 + v_3)\hat{\mathbf{x}} + (v_1 + v_3)\hat{\mathbf{y}} + (v_1 + v_2)\hat{\mathbf{z}}] \; . \tag{32}$$

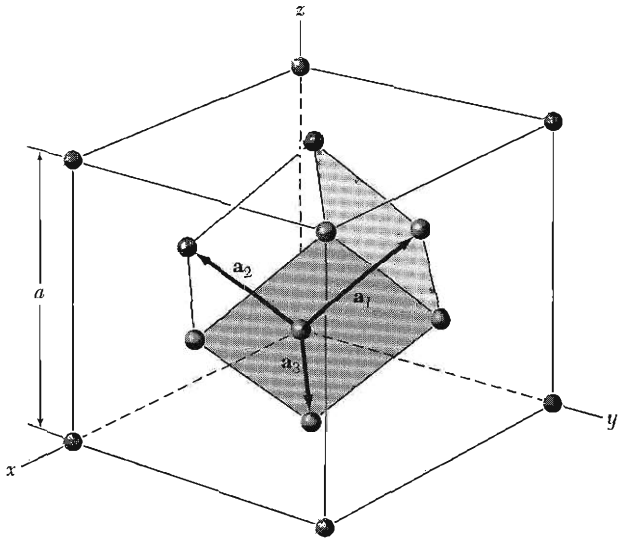


Figure 14 Primitive basis vectors of the face-centered cubic lattice.

The shortest **G**'s are the following 12 vectors, where all choices of sign are independent:

$$(2\pi/a)(\pm\hat{\mathbf{y}} \pm \hat{\mathbf{z}}) \; ; \qquad (2\pi/a)(\pm\hat{\mathbf{x}} \pm \hat{\mathbf{z}}) \; ; \qquad (2\pi/a)(\pm\hat{\mathbf{x}} \pm \hat{\mathbf{y}}) \; . \tag{33}$$

One primitive cell of the reciprocal lattice is the parallelepiped described by the $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ defined by (31). The volume of this cell in reciprocal space is $\mathbf{b}_1 \cdot \mathbf{b}_2 \times \mathbf{b}_3 = 2(2\pi/a)^3$. The cell contains one reciprocal lattice point, because each of the eight corner points is shared among eight parallelepipeds. Each parallelepiped contains one-eighth of each of eight corner points (see Fig. 12).

Another primitive cell is the central (Wigner-Seitz) cell of the reciprocal lattice which is the first Brillouin zone. Each such cell contains one lattice point at the central point of the cell. This zone (for the bcc lattice) is bounded by the planes normal to the 12 vectors of Eq. (33) at their midpoints. The zone is a regular 12-faced solid, a rhombic dodecahedron, as shown in Fig. 13.

Reciprocal Lattice to fcc Lattice

The primitive translation vectors of the fcc lattice of Fig. 14 are

$$\mathbf{a}_1 = \frac{1}{2}a(\hat{\mathbf{y}} + \hat{\mathbf{z}}) \; ; \quad \mathbf{a}_2 = \frac{1}{2}a(\hat{\mathbf{x}} + \hat{\mathbf{z}}) \; ; \quad \mathbf{a}_3 = \frac{1}{2}a(\hat{\mathbf{x}} + \hat{\mathbf{y}}) \; . \tag{34}$$

The volume of the primitive cell is

$$V = |\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3| = \frac{1}{4}a^3 \; . \tag{35}$$

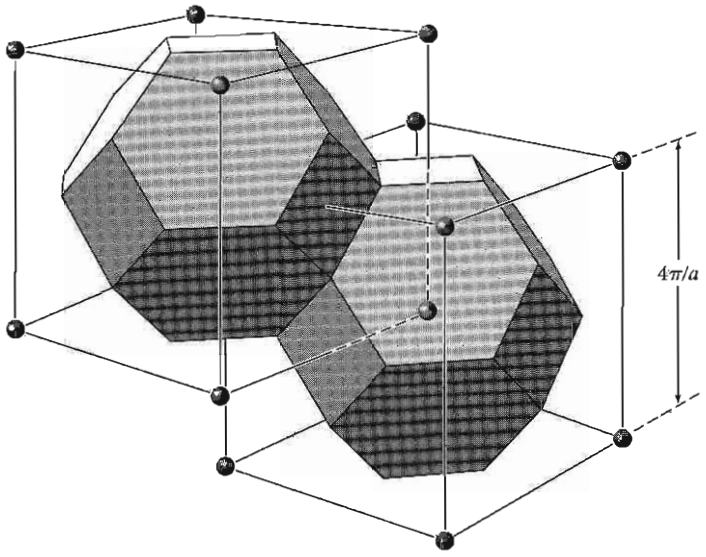


Figure 15 Brillouin zones of the face-centered cubic lattice. The cells are in reciprocal space, and the reciprocal lattice is body centered.

The primitive translation vectors of the lattice reciprocal to the fcc lattice are

$$\mathbf{b}_1 = (2\pi/a)(-\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}}) \ ; \qquad \mathbf{b}_2 = (2\pi/a)(\hat{\mathbf{x}} - \hat{\mathbf{y}} + \hat{\mathbf{z}}) \ ; \qquad \mathbf{b}_3 = (2\pi/a)(\hat{\mathbf{x}} + \hat{\mathbf{y}} - \hat{\mathbf{z}}) \ . \tag{36}$$

These are primitive translation vectors of a bcc lattice, so that the bcc lattice is reciprocal to the fcc lattice. The volume of the primitive cell of the reciprocal lattice is $4(2\pi/a)^3$.

The shortest **G**'s are the eight vectors:

$$(2\pi/a)(\pm\hat{\mathbf{x}} \pm \hat{\mathbf{y}} \pm \hat{\mathbf{z}}) \ . \tag{37}$$

The boundaries of the central cell in the reciprocal lattice are determined for the most part by the eight planes normal to these vectors at their midpoints. But the corners of the octahedron thus formed are cut by the planes that are the perpendicular bisectors of six other reciprocal lattice vectors:

$$(2\pi/a)(\pm 2\hat{\mathbf{x}}) \ ; \qquad (2\pi/a)(\pm 2\hat{\mathbf{y}}) \ ; \qquad (2\pi/a)(\pm 2\hat{\mathbf{z}}) \ . \tag{38}$$

Note that $(2\pi/a)(2\hat{\mathbf{x}})$ is a reciprocal lattice vector because it is equal to $\mathbf{b}_2 + \mathbf{b}_3$. The first Brillouin zone is the smallest bounded volume about the origin, the truncated octahedron shown in Fig. 15. The six planes bound a cube of edge $4\pi/a$ and (before truncation) of volume $(4\pi/a)^3$.

FOURIER ANALYSIS OF THE BASIS

When the diffraction condition $\Delta\mathbf{k} = \mathbf{G}$ of Eq. (21) is satisfied, the scattering amplitude (18) for a crystal of N cells may be written as

$$F_{\mathbf{G}} = N \int_{cell} dV\, n(\mathbf{r}) \exp(-i\mathbf{G} \cdot \mathbf{r}) = NS_{\mathbf{G}} \ . \tag{39}$$

The quantity $S_{\mathbf{G}}$ is called the **structure factor** and is defined as an integral over a single cell, with $\mathbf{r} = 0$ at one corner.

Often it is useful to write the electron concentration $n(\mathbf{r})$ as the superposition of electron concentration functions n_j associated with each atom j of the cell. If $\mathbf{r}_j$ is the vector to the center of atom j, then the function $n_j(\mathbf{r} - \mathbf{r}_j)$ defines the contribution of that atom to the electron concentration at $\mathbf{r}$. The total electron concentration at $\mathbf{r}$ due to all atoms in the single cell is the sum

$$n(\mathbf{r}) = \sum_{j-1}^{s} n_j(\mathbf{r} - \mathbf{r}_j) \tag{40}$$

over the s atoms of the basis. The decomposition of $n(\mathbf{r})$ is not unique, for we cannot always say how much charge density is associated with each atom. This is not an important difficulty.

The structure factor defined by (39) may now be written as integrals over the s atoms of a cell:

$$\begin{aligned} S_{\mathbf{G}} &= \sum_j \int dV\, n_j(\mathbf{r} - \mathbf{r}_j) \exp(-i\mathbf{G} \cdot \mathbf{r}) \\ &= \sum_j \exp(-i\mathbf{G} \cdot \mathbf{r}_j) \int dV\, n_j(\boldsymbol{\rho}) \exp(-i\mathbf{G} \cdot \boldsymbol{\rho}) \ , \end{aligned} \tag{41}$$

where $\boldsymbol{\rho} = \mathbf{r} - \mathbf{r}_j$. We now define the **atomic form factor** as

$$f_j = \int dV\, n_j(\boldsymbol{\rho}) \exp(-i\mathbf{G} \cdot \boldsymbol{\rho}) \ , \tag{42}$$

integrated over all space. If $n_j(\boldsymbol{\rho})$ is an atomic property, f_j is an atomic property.

We combine (41)and (42) to obtain the **structure factor of the basis** in the form

$$S_{\mathbf{G}} = \sum_j f_j \exp(-i\mathbf{G} \cdot \mathbf{r}_j) \ . \tag{43}$$

The usual form of this result follows on writing for atom j:

$$\mathbf{r}_j = x_j\mathbf{a}_1 + y_j\mathbf{a}_2 + z_j\mathbf{a}_3 \ , \tag{44}$$

as in (1.2). Then, for the reflection labelled by v_1, v_2, v_3, we have

$$\begin{aligned} \mathbf{G} \cdot \mathbf{r}_j &= (v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + v_3\mathbf{b}_3) \cdot (x_j\mathbf{a}_1 + y_j\mathbf{a}_2 + z_j\mathbf{a}_3) \\ &= 2\pi(v_1x_j + v_2y_j + v_3z_j) \quad , \end{aligned} \tag{45}$$

so that (43) becomes

$$S_G(v_1v_2v_3) = \sum_j f_j \exp[-i2\pi(v_1x_j + v_2y_j + v_3z_j)] \quad . \tag{46}$$

The structure factor S need not be real because the scattered intensity will involve S^*S, where S^* is the complex conjugate of S so that S^*S is real.

Structure Factor of the bcc Lattice

The bcc basis referred to the cubic cell has identical atoms at $x_1 = y_1 = z_1 = 0$ and at $x_2 = y_2 = z_2 = \frac{1}{2}$. Thus (46) becomes

$$S(v_1v_2v_3) = f\{1 + \exp[-i\pi(v_1 + v_2 + v_3)]\} \quad , \tag{47}$$

where f *is the form factor of an atom.* The value of S is zero whenever the exponential has the value -1, which is whenever the argument is $-i\pi \times$ (odd integer). Thus we have

$$\begin{aligned} S &= 0 \qquad \text{when } v_1 + v_2 + v_3 = \text{odd integer} \ ; \\ S &= 2f \qquad \text{when } v_1 + v_2 + v_3 = \text{even integer} \ . \end{aligned}$$

Metallic sodium has a bcc structure. The diffraction pattern does not contain lines such as (100), (300), (111), or (221), but lines such as (200), (110), and (222) will be present; here the indices $(v_1v_2v_3)$ are referred to a cubic cell. What is the physical interpretation of the result that the (100) reflection vanishes? The (100) reflection normally occurs when reflections from the planes that bound the cubic cell differ in phase by 2π. In the bcc lattice there is an intervening plane (Fig. 16) of atoms, labeled the second plane in the figure, which is equal in scattering power to the other planes. Situated midway between them, it gives a reflection retarded in phase by π with respect to the first plane, thereby canceling the contribution from that plane. The cancellation of the (100) reflection occurs in the bcc lattice because the planes are identical in composition. A similar cancellation can easily be found in the hcp structure.

Structure Factor of the fcc Lattice

The basis of the fcc structure referred to the cubic cell has identical atoms at 000; $0\frac{11}{22}$; $\frac{1}{2}0\frac{1}{2}$; $\frac{11}{22}0$. Thus (46) becomes

$$\begin{aligned} S(v_1v_2v_3) = f\{1 &+ \exp[-i\pi(v_2 + v_3)] + \exp[-i\pi(v_1 + v_3)] \\ &+ \exp[-i\pi(v_1 + v_2)]\} \ . \end{aligned} \tag{48}$$

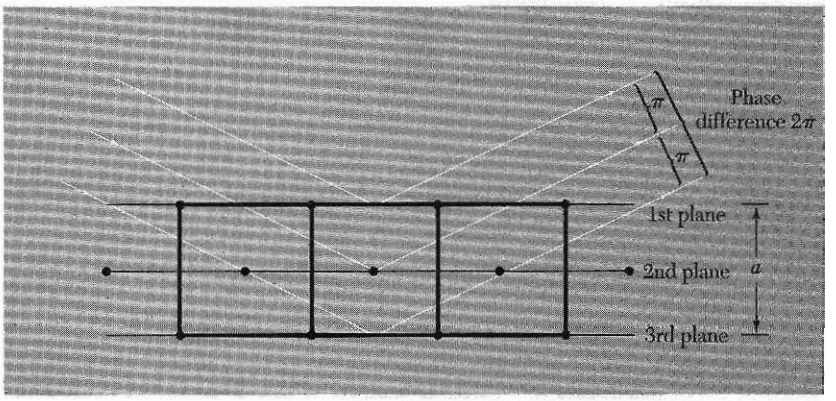


Figure 16 Explanation of the absence of a (100) reflection from a body-centered cubic lattice. The phase difference between successive planes is π, so that the reflected amplitude from two adjacent planes is $1 + e^{-i\pi} = 1 - 1 = 0$.

If all indices are even integers, $S = 4f$; similarly if all indices are odd integers. But if only one of the integers is even, two of the exponents will be odd multiples of $-i\pi$ and S will vanish. If only one of the integers is odd, the same argument applies and S will also vanish. Thus in the fcc lattice no reflections can occur for which the indices are partly even and partly odd.

The point is beautifully illustrated by Fig. 17: both KCl and KBr have an fcc lattice, but $n(\mathbf{r})$ for KCl simulates an sc lattice because the K^+ and Cl^- ions have equal numbers of electrons.

Atomic Form Factor

In the expression (46) for the structure factor, there occurs the quantity f_j, which is a measure of the scattering power of the jth atom in the unit cell. The value of f involves the number and distribution of atomic electrons, and the wavelength and angle of scattering of the radiation. We now give a classical calculation of the scattering factor.

The scattered radiation from a single atom takes account of interference effects within the atom. We defined the form factor in (42):

$$f_j = \int dV\, n_j(\mathbf{r}) \exp(-i\mathbf{G}\cdot\mathbf{r}) \ , \tag{49}$$

with the integral extended over the electron concentration associated with a single atom. Let $\mathbf{r}$ make an angle α with $\mathbf{G}$; then $\mathbf{G}\cdot\mathbf{r} = Gr\cos\alpha$. If the electron distribution is spherically symmetric about the origin, then

$$\begin{aligned} f_j &= 2\pi \int dr\, r^2\, d(\cos\alpha)\, n_j(r) \exp(-iGr\cos\alpha) \\ &= 2\pi \int dr\, r^2 n_j(r) \cdot \frac{e^{iGr} - e^{-iGr}}{iGr} \ , \end{aligned}$$

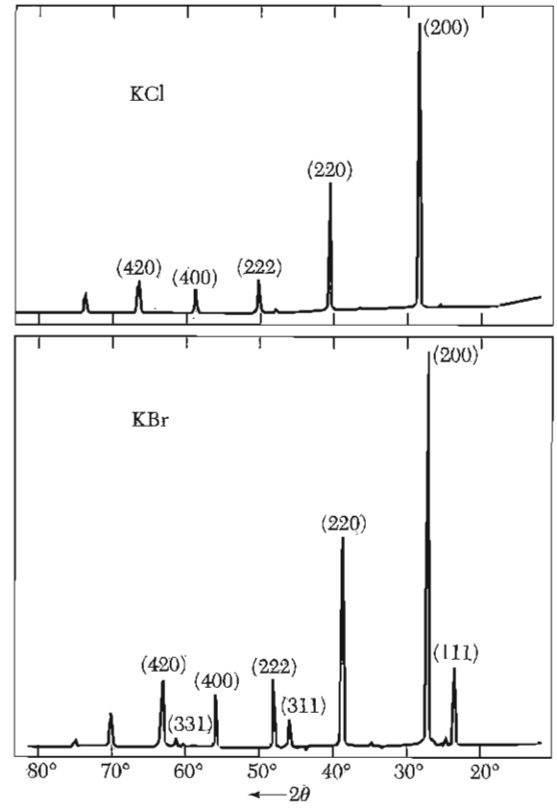


Figure 17 Comparison of x-ray reflections from KCl and KBr powders. In KCl the numbers of electrons of K^+ and Cl^- ions are equal. The scattering amplitudes $f(K^+)$ and $f(Cl^-)$ are almost exactly equal, so that the crystal looks to x-rays as if it were a monatomic simple cubic lattice of lattice constant $a/2$. Only even integers occur in the reflection indices when these are based on a cubic lattice of lattice constant a. In KBr the form factor of Br^- is quite different to that of K^+, and all reflections of the fcc lattice are present. (Courtesy of R. van Nordstrand.)

after integration over $d(\cos\alpha)$ between -1 and 1. Thus the form factor is given by

$$f_j = 4\pi \int dr\, n_j(r) r^2 \frac{\sin Gr}{Gr} \,. \tag{50}$$

If the same total electron density were concentrated at $r = 0$, only $Gr = 0$ would contribute to the integrand. In this limit $(\sin Gr)/Gr = 1$, and

$$f_j = 4\pi \int dr\, n_j(r) r^2 = Z \,, \tag{51}$$

the number of atomic electrons. Therefore f is the ratio of the radiation amplitude scattered by the actual electron distribution in an atom to that scattered by one electron localized at a point. In the forward direction $G = 0$, and f reduces again to the value Z.

The overall electron distribution in a solid as seen in x-ray diffraction is fairly close to that of the appropriate free atoms. This statement does not mean that the outermost or valence electrons are not redistributed somewhat in forming the solid; it means only that the x-ray reflection intensities are represented well by the free atom values of the form factors and are not very sensitive to small redistributions of the electrons.

SUMMARY

- Various statements of the Bragg condition:

$$2d \sin\theta = n\lambda \ ; \qquad \Delta\mathbf{k} = \mathbf{G} \ ; \qquad 2\mathbf{k}\cdot\mathbf{G} = \mathbf{G}^2 \ .$$

- Laue conditions:

$$\mathbf{a}_1\cdot\Delta\mathbf{k} = 2\pi v_1 \ ; \qquad \mathbf{a}_2\cdot\Delta\mathbf{k} = 2\pi v_2 \ ; \qquad \mathbf{a}_3\cdot\Delta\mathbf{k} = 2\pi v_3 \ .$$

- The primitive translation vectors of the reciprocal lattice are

$$\mathbf{b}_1 = 2\pi\frac{\mathbf{a}_2\times\mathbf{a}_3}{\mathbf{a}_1\cdot\mathbf{a}_2\times\mathbf{a}_3} \ ; \qquad \mathbf{b}_2 = 2\pi\frac{\mathbf{a}_3\times\mathbf{a}_1}{\mathbf{a}_1\cdot\mathbf{a}_2\times\mathbf{a}_3} \ ; \qquad \mathbf{b}_3 = 2\pi\frac{\mathbf{a}_1\times\mathbf{a}_2}{\mathbf{a}_1\cdot\mathbf{a}_2\times\mathbf{a}_3} \ .$$

  Here $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ are the primitive translation vectors of the crystal lattice.

- A reciprocal lattice vector has the form

$$\mathbf{G} = v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + v_3\mathbf{b}_3 \ ,$$

  where v_1, v_2, v_3 are integers or zero.

- The scattered amplitude in the direction $\mathbf{k}' = \mathbf{k} + \Delta\mathbf{k} = \mathbf{k} + \mathbf{G}$ is proportional to the geometrical structure factor:

$$S_\mathbf{G} \equiv \sum f_j \exp(-i\mathbf{r}_j\cdot\mathbf{G}) = \sum f_j \exp[-i2\pi(x_j v_1 + y_j v_2 + z_j v_3)] \ ,$$

  where j runs over the s atoms of the basis, and f_j is the atomic form factor (49) of the jth atom of the basis. The expression on the right-hand side is written for a reflection $(v_1 v_2 v_3)$, for which $\mathbf{G} = v_1\mathbf{b}_1 + v_2\mathbf{b}_2 + v_3\mathbf{b}_3$.

- Any function invariant under a lattice translation $\mathbf{T}$ may be expanded in a Fourier series of the form

$$n(\mathbf{r}) = \sum_\mathbf{G} n_\mathbf{G} \exp(i\mathbf{G}\cdot\mathbf{r}) \ .$$

- The first Brillouin zone is the Wigner-Seitz primitive cell of the reciprocal lattice. Only waves whose wavevector $\mathbf{k}$ drawn from the origin terminates on a surface of the Brillouin zone can be diffracted by the crystal.

Crystal lattice	*First Brillouin zone*
Simple cubic	Cube
Body-centered cubic	Rhombic dodecahedron (Fig. 13)
Face-centered cubic	Truncated octahedron (Fig. 15)

Problems

1. ***Interplanar separation.*** Consider a plane hkl in a crystal lattice. (a) Prove that the reciprocal lattice vector $\mathbf{G} = h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3$ is perpendicular to this plane. (b) Prove that the distance between two adjacent parallel planes of the lattice is $d(hkl) = 2\pi/|\mathbf{G}|$. (c) Show for a simple cubic lattice that $d^2 = a^2/(h^2 + k^2 + l^2)$.

2. ***Hexagonal space lattice.*** The primitive translation vectors of the hexagonal space lattice may be taken as

$$\mathbf{a}_1 = (3^{1/2}a/2)\hat{\mathbf{x}} + (a/2)\hat{\mathbf{y}} \ ; \qquad \mathbf{a}_2 = -(3^{1/2}a/2)\hat{\mathbf{x}} + (a/2)\hat{\mathbf{y}} \ ; \qquad \mathbf{a}_3 = c\hat{\mathbf{z}} \ .$$

(a) Show that the volume of the primitive cell is $(3^{1/2}/2)a^2c$.
(b) Show that the primitive translations of the reciprocal lattice are

$$\mathbf{b}_1 = (2\pi/3^{1/2}a)\hat{\mathbf{x}} + (2\pi/a)\hat{\mathbf{y}} \ ; \qquad \mathbf{b}_2 = -(2\pi/3^{1/2}a)\hat{\mathbf{x}} + (2\pi/a)\hat{\mathbf{y}} \ ; \qquad \mathbf{b}_3 = (2\pi/c)\hat{\mathbf{z}} \ ,$$

so that the lattice is its own reciprocal, but with a rotation of axes.
(c) Describe and sketch the first Brillouin zone of the hexagonal space lattice.

3. ***Volume of Brillouin zone.*** Show that the volume of the first Brillouin zone is $(2\pi)^3/V_c$, where V_c is the volume of a crystal primitive cell. Hint: The volume of a Brillouin zone is equal to the volume of the primitive parallelepiped in Fourier space. Recall the vector identity $(\mathbf{c} \times \mathbf{a}) \times (\mathbf{a} \times \mathbf{b}) = (\mathbf{c} \cdot \mathbf{a} \times \mathbf{b})\mathbf{a}$.

4. ***Width of diffraction maximum.*** We suppose that in a linear crystal there are identical point scattering centers at every lattice point $\boldsymbol{\rho}_m = m\mathbf{a}$, where m is an integer. By analogy with (20), the total scattered radiation amplitude will be proportional to $F = \Sigma \exp[-im\mathbf{a} \cdot \Delta\mathbf{k}]$. The sum over M lattice points is

$$F = \frac{1 - \exp[-iM(\mathbf{a} \cdot \Delta\mathbf{k}]}{1 - \exp[-i(\mathbf{a} \cdot \Delta\mathbf{k})]} \ ,$$

by the use of the series

$$\sum_{m=0}^{M-1} x^m = \frac{1 - x^M}{1 - x} \ .$$

(a) The scattered intensity is proportional to $|F|^2$. Show that

$$|F|^2 \equiv F^*F = \frac{\sin^2 \frac{1}{2} M(\mathbf{a} \cdot \Delta\mathbf{k})}{\sin^2 \frac{1}{2} (\mathbf{a} \cdot \Delta\mathbf{k})} \ .$$

(b) We know that a diffraction maximum appears when $\mathbf{a} \cdot \Delta\mathbf{k} = 2\pi h$, where h is an integer. We change $\Delta\mathbf{k}$ slightly and define ϵ in $\mathbf{a} \cdot \Delta\mathbf{k} = 2\pi h + \epsilon$ such that ϵ gives the position of the first zero in $\sin \frac{1}{2} M(\mathbf{a} \cdot \Delta\mathbf{k})$. Show that $\epsilon = 2\pi/M$, so that the width of the diffraction maximum is proportional to $1/M$ and can be extremely narrow for macroscopic values of M. The same result holds true for a three-dimensional crystal.

5. ***Structure factor of diamond.*** The crystal structure of diamond is described in Chapter 1. The basis consists of eight atoms if the cell is taken as the conventional cube. (a) Find the structure factor S of this basis. (b) Find the zeros of S and show that the allowed reflections of the diamond structure satisfy $v_1 + v_2 + v_3 = 4n$, where all indices are even and n is any integer, or else all indices are odd (Fig. 18). (Notice that h, k, l may be written for v_1, v_2, v_3 and this is often done.)

6. ***Form factor of atomic hydrogen.*** For the hydrogen atom in its ground state, the number density is $n(r) = (\pi a_0^3)^{-1} \exp(-2r/a_0)$, where a_0 is the Bohr radius. Show that the form factor is $f_G = 16/(4 + G^2a_0^2)^2$.

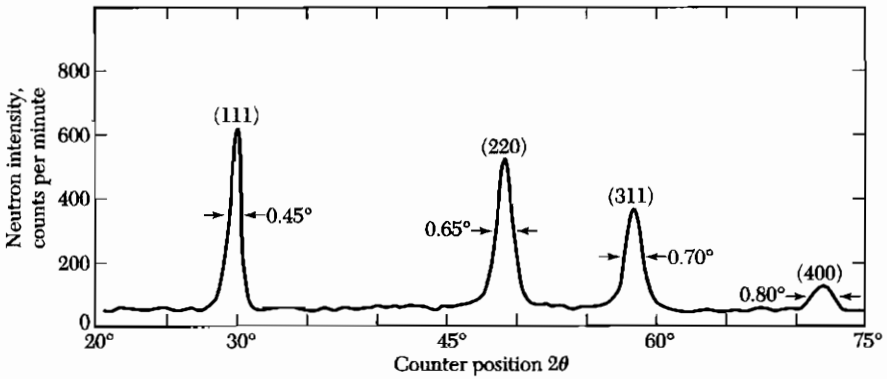


Figure 18 Neutron diffraction pattern for powdered diamond. (After G. Bacon.)

7. ***Diatomic line.*** Consider a line of atoms $ABAB \ldots AB$, with an A—B bond length of $\frac{1}{2}a$. The form factors are f_A, f_B for atoms A, B, respectively. The incident beam of x-rays is perpendicular to the line of atoms. (a) Show that the interference condition is $n\lambda = a \cos\theta$, where θ is the angle between the diffracted beam and the line of atoms. (b) Show that the intensity of the diffracted beam is proportional to $|f_A - f_B|^2$ for n odd, and to $|f_A + f_B|^2$ for n even. (c) Explain what happens if $f_A = f_B$.

3

Crystal Binding and Elastic Constants

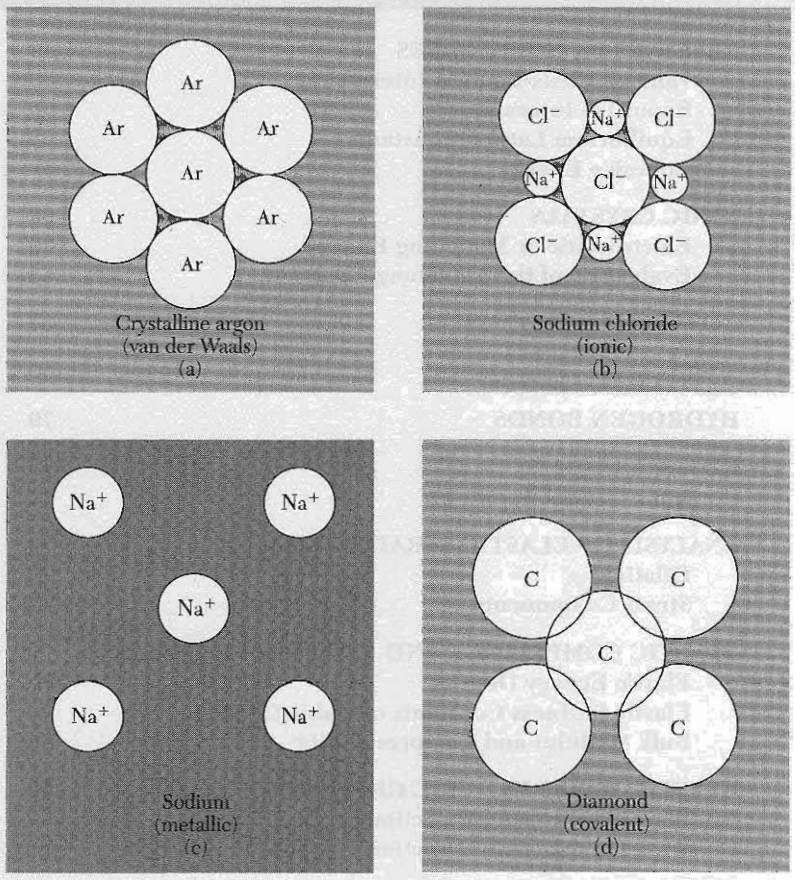


Figure 1 The principal types of crystalline binding. In (a) neutral atoms with closed electron shells are bound together weakly by the van der Waals forces associated with fluctuations in the charge distributions. In (b) electrons are transferred from the alkali atoms to the halogen atoms, and the resulting ions are held together by attractive electrostatic forces between the positive and negative ions. In (c) the valence electrons are taken away from each alkali atom to form a communal electron sea in which the positive ions are dispersed. In (d) the neutral atoms are bound together by the overlapping parts of their electron distributions.

In this chapter we are concerned with the question: What holds a crystal together? The attractive electrostatic interaction between the negative charges of the electrons and the positive charges of the nuclei is entirely responsible for the cohesion of solids. Magnetic forces have only a weak effect on cohesion, and gravitational forces are negligible. Specialized terms categorize distinctive situations: exchange energy, van der Waals forces, and covalent bonds. The observed differences between the forms of condensed matter are caused in the final analysis by differences in the distribution of the outermost electrons and the ion cores (Fig. 1).

The cohesive energy of a crystal is defined as the energy that must be added to the crystal to separate its components into neutral free atoms at rest, at infinite separation, with the same electronic configuration. The term lattice energy is used in the discussion of ionic crystals and is defined as the energy that must be added to the crystal to separate its component ions into free ions *at rest at infinite separation.*

Values of the cohesive energy of the crystalline elements are given in Table 1. Notice the wide variation in cohesive energy between different columns of the periodic table. The inert gas crystals are weakly bound, with cohesive energies less than a few percent of the cohesive energies of the elements in the C, Si, Ge . . . column. The alkali metal crystals have intermediate values of the cohesive energy. The transition element metals (in the middle columns) are quite strongly bound. The melting temperatures (Table 2) and bulk modulii (Table 3) vary roughly as the cohesive energies.

CRYSTALS OF INERT GASES

The inert gases form the simplest crystals. The electron distribution is very close to that of the free atoms. Their properties at absolute zero are summarized in Table 4. The crystals are transparent insulators, weakly bound, with low melting temperatures. The atoms have very high ionization energies (see Table 5). The outermost electron shells of the atoms are completely filled, and the distribution of electron charge in the free atom is spherically symmetric. In the crystal the inert gas atoms pack together as closely as possible[1]: the

[1]Zero-point motion of the atoms (kinetic energy at absolute zero) is a quantum effect that plays a dominant role in He^3 and He^4. They do not solidify at zero pressure even at absolute zero temperature. The average fluctuation at 0 K of a He atom from its equilibrium position is of the order of 30 to 40 percent of the nearest-neighbor distance. The heavier the atom, the less important the zero-point effects. If we omit zero-point motion, we calculate a molar volume of 9 $cm^3\ mol^{-1}$ for solid helium, as compared with the observed values of 27.5 and 36.8 $cm^3\ mol^{-1}$ for liquid He^4 and liquid He^3, respectively.

Table 1 Cohesive energies

Energy required to form separated neutral atoms in their ground electronic state from the solid at 0 K at 1 atm. The data were supplied by Prof. Leo Brewer.

Li	Be											B	C	N	O	F	Ne
158.	320.											561	711.	474.	251.	81.0	1.92
1.63	3.32											5.81	7.37	4.92	2.60	0.84	0.020
37.7	76.5											134	170.	113.4	60.03	19.37	0.46
Na	Mg											Al	Si	P	S	Cl	Ar
107.	145.	← kJ/mol →										327.	446.	331.	275.	135.	7.74
1.113	1.51	← eV/atom →										3.39	4.63	3.43	2.85	1.40	0.080
25.67	34.7	← kcal/mol →										78.1	106.7	79.16	65.75	32.2	1.85
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
90.1	178.	376	468.	512.	395.	282.	413.	424.	428.	336.	130	271.	372.	285.3	237	118.	11.2
0.934	1.84	3.90	4.85	5.31	4.10	2.92	4.28	4.39	4.44	3.49	1.35	2.81	3.85	2.96	2.46	1.22	0.116
21.54	42.5	89.9	111.8	122.4	94.5	67.4	98.7	101.3	102.4	80.4	31.04	64.8	88.8	68.2	56.7	28.18	2.68
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
82.2	166.	422.	603.	730.	658	661.	650.	554.	376.	284.	112.	243.	303.	265.	211	107.	15.9
0.852	1.72	4.37	6.25	7.57	6.82	6.85	6.74	5.75	3.89	2.95	1.16	2.52	3.14	2.75	2.19	1.11	0.16
19.64	39.7	100.8	144.2	174.5	157.2	158.	155.4	132.5	89.8	68.0	26.73	58.1	72.4	63.4	50.34	25.62	3.80
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
77.6	183.	431.	621.	782.	859.	775.	788.	670.	564.	368.	65.	182.	196.	210.	144.		19.5
0.804	1.90	4.47	6.44	8.10	8.90	8.03	8.17	6.94	5.84	3.81	0.67	1.88	2.03	2.18	1.50		0.202
18.54	43.7	103.1	148.4	186.9	205.2	185.2	188.4	160.1	134.7	87.96	15.5	43.4	46.78	50.2	34.5		4.66
Fr	Ra	Ac															
	160.	410.															
	1.66	4.25															
	38.2	98.															

Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
417.	357.	328.		206.	179.	400.	391.	294.	302.	317.	233.	154.	428.
4.32	3.70	3.40		2.14	1.86	4.14	4.05	3.04	3.14	3.29	2.42	1.60	4.43
99.7	85.3	78.5		49.3	42.8	95.5	93.4	70.2	72.3	75.8	55.8	37.1	102.2
Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr
598.		536.	456	347.	264.	385							
6.20		5.55	4.73	3.60	2.73	3.99							
142.9		128.	109.	83.0	63.	92.1							

Table 2 Melting points, in K.

(After R. H. Lamoreaux)

Li 453.7	Be 1562											B 2365	C	N 63.15	O 54.36	F 53.48	Ne 24.56
Na 371.0	Mg 922											Al 933.5	Si 1687	P *w* 317 *r* 863	S 388.4	Cl 172.2	Ar 83.81
K 336.3	Ca 1113	Sc 1814	Ti 1946	V 2202	Cr 2133	Mn 1520	Fe 1811	Co 1770	Ni 1728	Cu 1358	Zn 692.7	Ga 302.9	Ge 1211	As 1089	Se 494	Br 265.9	Kr 115.8
Rb 312.6	Sr 1042	Y 1801	Zr 2128	Nb 2750	Mo 2895	Tc 2477	Ru 2527	Rh 2236	Pd 1827	Ag 1235	Cd 594.3	In 429.8	Sn 505.1	Sb 903.9	Te 722.7	I 386.7	Xe 161.4
Cs 301.6	Ba 1002	La 1194	Hf 2504	Ta 3293	W 3695	Re 3459	Os 3306	Ir 2720	Pt 2045	Au 1338	Hg 234.3	Tl 577	Pb 600.7	Bi 544.6	Po 527	At	Rn
Fr	Ra 973	Ac 1324															

Ce 1072	Pr 1205	Nd 1290	Pm	Sm 1346	Eu 1091	Gd 1587	Tb 1632	Dy 1684	Ho 1745	Er 1797	Tm 1820	Yb 1098	Lu 1938
Th 2031	Pa 1848	U 1406	Np 910	Pu 913	Am 1449	Cm 1613	Bk 1562	Cf	Es	Fm	Md	No	Lw

Table 3 Isothermal bulk modulii and compressibilities at room temperature

After K. Gschneidner, Jr., Solid State Physics 16, 275–426 (1964); several data are from F. Birch, in *Handbook of physical constants*, Geological Society of America Memoir 97, 107–173 (1966). Original references should be consulted when values are needed for research purposes. Values in parentheses are estimates. Letters in parentheses refer to the crystal form. Letters in brackets refer to the temperature:

[a] = 77 K; [b] = 273 K; [c] = 1 K; [d] = 4 K; [e] = 81 K.

Bulk modulus in units 10^{12} dyn/cm² or 10^{11} N/m²
Compressibility in units 10^{-12} cm²/dyn or 10^{-11} m²/N

H [d]																	He [d]
0.002																	0.00
500																	1168
Li	Be											B	C [d]	N [e]	O	F	Ne [d]
0.116	1.003											1.78	4.43	0.012			0.010
8.62	0.997											0.562	0.226	80			100
Na	Mg											Al	Si	P [b]	S [r]	Cl	Ar [a]
0.068	0.354											0.722	0.988	0.304	0.178		0.013
14.7	2.82											1.385	1.012	3.29	5.62		79
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga [b]	Ge	As	Se	Br	Kr [a]
0.032	0.152	0.435	1.051	1.619	1.901	0.596	1.683	1.914	1.86	1.37	0.598	0.569	0.772	0.394	0.091		0.018
31.	6.58	2.30	0.951	0.618	0.526	1.68	0.594	0.522	0.538	0.73	1.67	1.76	1.29	2.54	11.0		56
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn [g]	Sb	Te	I	Xe
0.031	0.116	0.366	0.833	1.702	2.725	(2.97)	3.208	2.704	1.808	1.007	0.467	0.411	1.11	0.383	0.230		
32.	8.62	2.73	1.20	0.587	0.366	(0.34)	0.311	0.369	0.553	0.993	2.14	2.43	0.901	2.61	4.35		
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg [c]	Tl	Pb	Bi	Po	At	Rn
0.020	0.103	0.243	1.09	2.00	3.232	3.72	(4.18)	3.55	2.783	1.732	0.382	0.359	0.430	0.315	(0.26)		
50.	9.97	4.12	0.92	0.50	0.309	0.269	(0.24)	0.282	0.359	0.577	2.60	2.79	2.33	3.17	(3.8)		
Fr	Ra	Ac															
(0.020)	(0.132)	(0.25)															
(50.)	(7.6)	(4.)															

Ce [γ]	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
0.239	0.306	0.327	(0.35)	0.294	0.147	0.383	0.399	0.384	0.397	0.411	0.397	0.133	0.411
4.18	3.27	3.06	(2.85)	3.40	6.80	2.61	2.51	2.60	2.52	2.43	2.52	7.52	2.43
Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr
0.543	(0.76)	0.987	(0.68)	0.54									
1.84	(1.3)	1.01	(1.5)	1.9									

Table 4 Properties of inert gas crystals

(Extrapolated to 0 K and zero pressure)

	Nearest-neighbor distance, in Å	Experimental cohesive energy		Melting point, K	Ionization potential of free atom, eV	Parameters in Lennard-Jones potential, Eq. 10	
		kJ/mol	eV/atom			ϵ, in 10^{-16} erg	σ, in Å
He	(liquid at zero pressure)				24.58	14	2.56
Ne	3.13	1.88	0.02	24.56	21.56	50	2.74
Ar	3.76	7.74	0.080	83.81	15.76	167	3.40
Kr	4.01	11.2	0.116	115.8	14.00	225	3.65
Xe	4.35	16.0	0.17	161.4	12.13	320	3.98

crystal structures (Fig. 2) are all cubic close-packed (fcc), except He^3 and He^4.

What holds an inert gas crystal together? The electron distribution in the crystal is not significantly distorted from the electron distribution around the free atoms because not much energy is available to distort the free atom charge distributions. The cohesive energy of an atom in the crystal is only 1 percent or less of the ionization energy of an atomic electron. Part of this distortion gives the van der Waals interaction.

Van der Waals-London Interaction

Consider two identical inert gas atoms at a separation R large in comparison with the radii of the atoms. What interactions exist between the two neutral atoms? If the charge distributions on the atoms were rigid, the interaction between atoms would be zero, because the electrostatic potential of a spherical distribution of electronic charge is canceled outside a neutral atom by the electrostatic potential of the charge on the nucleus. Then the inert gas atoms could show no cohesion and could not condense. But the atoms induce dipole moments in each other, and the induced moments cause an attractive interaction between the atoms.

As a model, we consider two identical linear harmonic oscillators 1 and 2 separated by R. Each oscillator bears charges $\pm e$ with separations x_1 and x_2, as in Fig. 3. The particles oscillate along the x axis. Let p_1 and p_2 denote the momenta. The force constant is C. Then the hamiltonian of the unperturbed system is

$$\mathcal{H}_0 = \frac{1}{2m}p_1^2 + \tfrac{1}{2}Cx_1^2 + \frac{1}{2m}p_2^2 + \tfrac{1}{2}Cx_2^2 \ . \qquad (1)$$

Each uncoupled oscillator is assumed to have the frequency ω_0 of the strongest optical absorption line of the atom. Thus $C = m\omega_0^2$.

Table 5 Ionization energies

The total energy required to remove the first two electrons is the sum of the first and second ionization potentials. (*Source:* National Bureau of Standards Circular 467.)

H																	He
13.595																	24.58
																	78.98
Li	Be											B	C	N	O	F	Ne
5.39	9.32											8.30	11.26	14.54	13.61	17.42	21.56
81.01	27.53											33.45	35.64	44.14	48.76	52.40	62.63
Na	Mg											Al	Si	P	S	Cl	Ar
5.14	7.64	← Energy to remove one electron, in eV →										5.98	8.15	10.55	10.36	13.01	15.76
52.43	22.67	← Energy to remove two electrons, in eV →										24.80	24.49	30.20	34.0	36.81	43.38
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
4.34	6.11	6.56	6.83	6.74	6.76	7.43	7.90	7.86	7.63	7.72	9.39	6.00	7.88	9.81	9.75	11.84	14.00
36.15	17.98	19.45	20.46	21.39	23.25	23.07	24.08	24.91	25.78	27.93	27.35	26.51	23.81	30.0	31.2	33.4	38.56
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
4.18	5.69	6.5	6.95	6.77	7.18	7.28	7.36	7.46	8.33	7.57	8.99	5.78	7.34	8.64	9.01	10.45	12.13
31.7	16.72	18.9	20.98	21.22	23.25	22.54	24.12	25.53	27.75	29.05	25.89	24.64	21.97	25.1	27.6	29.54	33.3
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
3.89	5.21	5.61	7.	7.88	7.98	7.87	8.7	9.	8.96	9.22	10.43	6.11	7.41	7.29	8.43		10.74
29.0	15.21	17.04	22.	24.1	25.7	24.5	26.		27.52	29.7	29.18	26.53	22.44	23.97			
Fr	Ra	Ac															
	5.28	6.9															
	15.42	19.0															

Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
6.91	5.76	6.31		5.6	5.67	6.16	6.74	6.82				6.2	5.0
Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr
		4.											

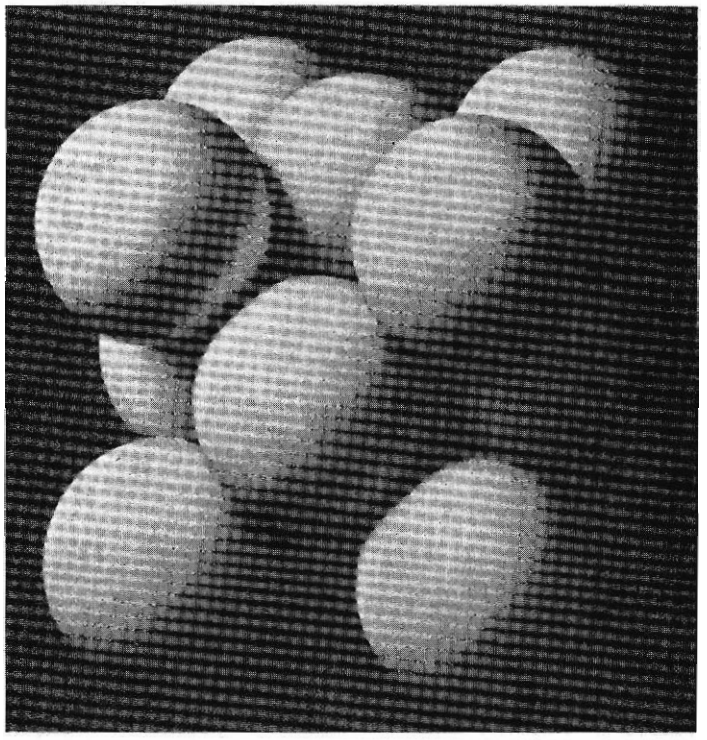


Figure 2 Cubic close-packed (fcc) crystal structure of the inert gases Ne, Ar, Kr, and Xe. The lattice parameters of the cubic cells are 4.46, 5.31, 5.64, and 6.13 Å, respectively, at 4 K.

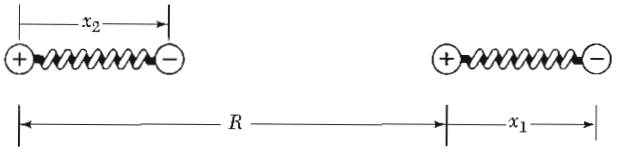


Figure 3 Coordinates of the two oscillators.

Let $\mathcal{H}_1$ be the coulomb interaction energy of the two oscillators. The geometry is shown in the figure. The internuclear coordinate is R. Then

$$\text{(CGS)} \qquad \mathcal{H}_1 = \frac{e^2}{R} + \frac{e^2}{R + x_1 - x_2} - \frac{e^2}{R + x_1} - \frac{e^2}{R - x_2} \; ; \tag{2}$$

in the approximation $|x_1|, |x_2| \ll R$ we expand (2) to obtain in lowest order:

$$\mathcal{H}_1 \cong -\frac{2e^2 x_1 x_2}{R^3} \; . \tag{3}$$

The total hamiltonian with the approximate form (3) for $\mathcal{H}_1$ can be diagonalized by the normal mode transformation

$$x_s \equiv \frac{1}{\sqrt{2}}(x_1 + x_2) \; ; \qquad x_a \equiv \frac{1}{\sqrt{2}}(x_1 - x_2) \; , \tag{4}$$

or, on solving for x_1 and x_2,

$$x_1 = \frac{1}{\sqrt{2}}(x_s + x_a) \; ; \qquad x_2 = \frac{1}{\sqrt{2}}(x_s - x_a) \; . \tag{5}$$

The subscripts s and a denote symmetric and antisymmetric modes of motion. Further, we have the momenta p_s, p_a associated with the two modes:

$$p_1 \equiv \frac{1}{\sqrt{2}}(p_s + p_a) \; ; \qquad p_2 \equiv \frac{1}{\sqrt{2}}(p_s - p_a) \; . \tag{6}$$

The total hamiltonian $\mathcal{H}_0 + \mathcal{H}_1$ after the transformations (5) and (6) is

$$\mathcal{H} = \left[\frac{1}{2m}p_s^2 + \frac{1}{2}\left(C - \frac{2e^2}{R^3}\right)x_s^2\right] + \left[\frac{1}{2m}p_a^2 + \frac{1}{2}\left(C + \frac{2e^2}{R^3}\right)x_a^2\right] . \tag{7}$$

The two frequencies of the coupled oscillators are found by inspection of (7) to be

$$\omega = \left[\left(C \pm \frac{2e^2}{R^3}\right)\Big/ m\right]^{1/2} = \omega_0\left[1 \pm \frac{1}{2}\left(\frac{2e^2}{CR^3}\right) - \frac{1}{8}\left(\frac{2e^2}{CR^3}\right)^2 + \cdots\right] , \tag{8}$$

with ω_0 given by $(C/m)^{1/2}$. In (8) we have expanded the square root.

The zero point energy of the system is $\frac{1}{2}\hbar(\omega_s + \omega_a)$; because of the interaction the sum is lowered from the uncoupled value $2 \cdot \frac{1}{2}\hbar\omega_0$ by

$$\Delta U = \tfrac{1}{2}\hbar(\Delta\omega_s + \Delta\omega_a) = -\hbar\omega_0 \cdot \frac{1}{8}\left(\frac{2e^2}{CR^3}\right)^2 = -\frac{A}{R^6} \; . \tag{9}$$

This attractive interaction varies as the minus sixth power of the separation of the two oscillators.

This is called the van der Waals interaction, known also as the London interaction or the induced dipole-dipole interaction. It is the principal attractive interaction in crystals of inert gases and also in crystals of many organic molecules. The interaction is a quantum effect, in the sense that $\Delta U \to 0$ as $\hbar \to 0$. Thus the zero point energy of the system is lowered by the dipole-dipole coupling of Eq. (3). The van der Waals interaction does not depend for its existence on any overlap of the charge densities of the two atoms.

An approximate value of the constant A in (9) for identical atoms is given by $\hbar\omega_0\alpha^2$, where $\hbar\omega_0$ is the energy of the strongest optical absorption line and α is the electronic polarizability (Chapter 15).

Repulsive Interaction

As the two atoms are brought together, their charge distributions gradually overlap (Fig. 4), thereby changing the electrostatic energy of the system. At sufficiently close separations the overlap energy is repulsive, in large part because of the **Pauli exclusion principle**. The elementary statement of the principle is that two electrons cannot have all their quantum numbers equal. When the charge distributions of two atoms overlap, there is a tendency for electrons from atom B to occupy in part states of atom A already occupied by electrons of atom A, and vice versa.

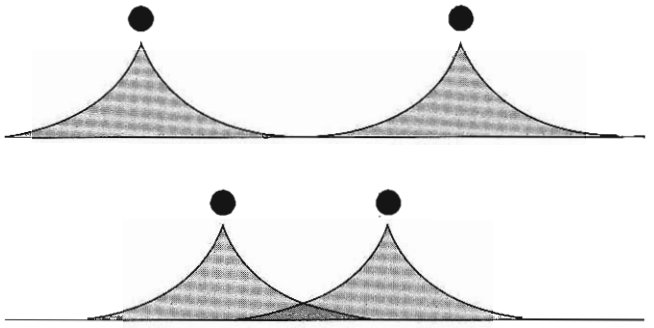


Figure 4 Electronic charge distributions overlap as atoms approach. The solid circles denote the nuclei.

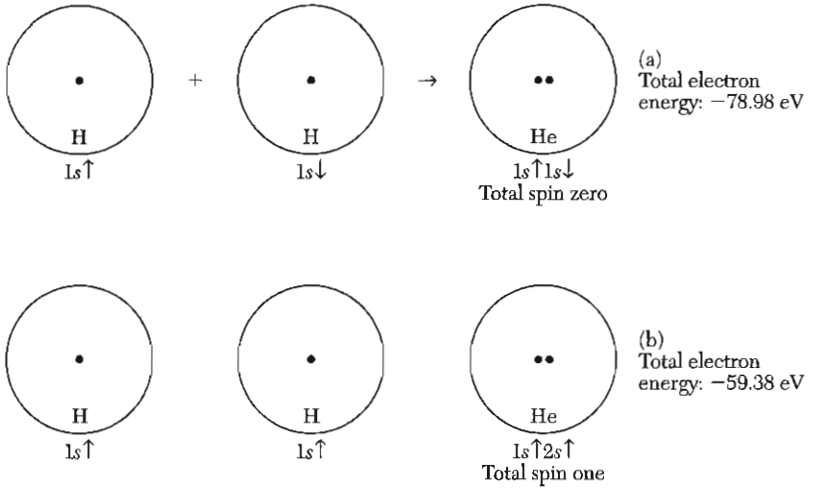


Figure 5 The effect of Pauli principle on the repulsive energy: in an extreme example, two hydrogen atoms are pushed together until the protons are almost in contact. The energy of the electron system alone can be taken from observations on atomic He, which has two electrons. In (a) the electrons have antiparallel spins and the Pauli principle has no effect: the electrons are bound by −78.98 eV. In (b) the spins are parallel: the Pauli principle forces the promotion of an electron from a 1s ↑ orbital of H to a 2s ↑ orbital of He. The electrons now are bound by -59.38 eV, less than (a) by 19.60 eV. This is the amount by which the Pauli principle has increased the repulsion. We have omitted the repulsive coulomb energy of the two protons, which is the same in both (a) and (b).

The Pauli principle prevents multiple occupancy, and electron distributions of atoms with closed shells can overlap only if accompanied by the partial promotion of electrons to unoccupied high energy states of the atoms. Thus the electron overlap increases the total energy of the system and gives a repulsive contribution to the interaction. An extreme example in which the overlap is complete is shown in Fig. 5.

We make no attempt here to evaluate the repulsive interaction[2] from first principles. Experimental data on the inert gases can be fitted well by an empirical repulsive potential of the form B/R^{12}, where B is a positive constant, when used

[2]The overlap energy naturally depends on the radial distribution of charge about each atom. The mathematical calculation is always complicated even if the charge distribution is known.

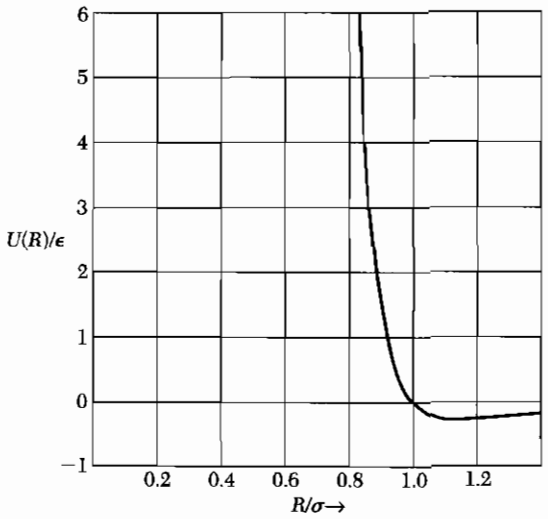


Figure 6 Form of the Lennard-Jones potential (10) which describes the interaction of two inert gas atoms. The minimum occurs at $R/\sigma = 2^{1/6} \cong 1.12$. Notice how steep the curve is inside the minimum, and how flat it is outside the minimum. The value of U at the minimum is $-\epsilon$; and $U = 0$ at $R = \sigma$.

together with a long-range attractive potential of the form of (9). The constants A and B are empirical parameters determined from independent measurements made in the gas phase; the data used include the virial coefficients and the viscosity. It is usual to write the total potential energy of two atoms at separation R as

$$U(R) = 4\epsilon\left[\left(\frac{\sigma}{R}\right)^{12} - \left(\frac{\sigma}{R}\right)^{6}\right], \tag{10}$$

where ϵ and σ are the new parameters, with $4\epsilon\sigma^6 = A$ and $4\epsilon\sigma^{12} = B$. The potential (10) is known as the Lennard-Jones potential, Fig. 6. The force between the two atoms is given by $-dU/dR$. Values of ϵ and σ given in Table 4 can be obtained from gas-phase data, so that calculations on properties of the solid do not involve disposable parameters.

Other empirical forms for the repulsive interaction are widely used, in particular the exponential form $\lambda \exp(-R/\rho)$, where ρ is a measure of the range of the interaction. This is generally as easy to handle analytically as the inverse power law form.

Equilibrium Lattice Constants

If we neglect the kinetic energy of the inert gas atoms, the cohesive energy of an inert gas crystal is given by summing the Lennard-Jones potential (10) over all pairs of atoms in the crystal. If there are N atoms in the crystal, the total potential energy is

$$U_{\text{tot}} = \tfrac{1}{2}N(4\epsilon)\left[\sum_j{}' \left(\frac{\sigma}{p_{ij}R}\right)^{12} - \sum_j{}' \left(\frac{\sigma}{p_{ij}R}\right)^{6}\right], \tag{11}$$

where $p_{ij}R$ is the distance between reference atom i and any other atom j, expressed in terms of the nearest-neighbor distance R. The factor $\frac{1}{2}$ occurs with the N to compensate for counting twice each pair of atoms.

The summations in (11) have been evaluated, and for the fcc structure

$$\sum_j{}' p_{ij}^{-12} = 12.13188 \; ; \qquad \sum_j{}' p_{ij}^{-6} = 14.45392 \; . \tag{12}$$

There are 12 nearest-neighbor sites in the fcc structure; we see that the series are rapidly converging and have values not far from 12. The nearest neighbors contribute most of the interaction energy of inert gas crystals. The corresponding sums for the hcp structure are 12.13229 and 14.45489.

If we take U_{tot} in (11) as the total energy of the crystal, the equilibrium value R_0 is given by requiring that U_{tot} be a minimum with respect to variations in the nearest-neighbor distance R:

$$\frac{dU_{\text{tot}}}{dR} = 0 = -2N\epsilon\left[(12)(12.13)\frac{\sigma^{12}}{R^{13}} - (6)(14.45)\frac{\sigma^6}{R^7}\right] , \tag{13}$$

whence

$$R_0/\sigma = 1.09 \; , \tag{14}$$

the same for all elements with an fcc structure. The observed values of R_0/σ, using the independently determined values of σ given in Table 4, are:

	Ne	Ar	Kr	Xe
R_0/σ	1.14	1.11	1.10	1.09 .

The agreement with (14) is remarkable. The slight departure of R_0/σ for the lighter atoms from the universal value 1.09 predicted for inert gases can be explained by zero-point quantum effects. From measurements on the gas phase we have predicted the lattice constant of the crystal.

Cohesive Energy

The cohesive energy of inert gas crystals at absolute zero and at zero pressure is obtained by substituting (12) and (14) in (11):

$$U_{\text{tot}}(R) = 2N\epsilon\left[(12.13)\left(\frac{\sigma}{R}\right)^{12} - (14.45)\left(\frac{\sigma}{R}\right)^6\right] , \tag{15}$$

and, at $R = R_0$,

$$U_{\text{tot}}(R_0) = -(2.15)(4N\epsilon) \; , \tag{16}$$

the same for all inert gases. This is the calculated cohesive energy when the atoms are at rest. Quantum-mechanical corrections act to reduce the binding by 28, 10, 6, and 4 percent of Eq. (16) for Ne, Ar, Kr, and Xe, respectively.

The heavier the atom, the smaller the quantum correction. We can understand the origin of the quantum correction by consideration of a simple model in which an atom is confined by fixed boundaries. If the particle has the quantum wavelength λ, where λ is determined by the boundaries, then the particle has kinetic energy $p^2/2M = (h/\lambda)^2/2M$ with the de Broglie relation $p = h/\lambda$ for the connection between the momentum and the wavelength of a particle. On this model the quantum zero-point correction to the energy is inversely proportional to the mass. The final calculated cohesive energies agree with the experimental values of Table 4 within 1 to 7 percent.

One consequence of the quantum kinetic energy is that a crystal of the isotope Ne^{20} is observed to have a larger lattice constant than a crystal of Ne^{22}. The higher quantum kinetic energy of the lighter isotope expands the lattice because the kinetic energy is reduced by expansion. The observed lattice constants (extrapolated to absolute zero from 2.5 K) are Ne^{20}, 4.4644 Å; Ne^{22}, 4.4559 Å.

IONIC CRYSTALS

Ionic crystals are made up of positive and negative ions. The ionic bond results from the electrostatic interaction of oppositely charged ions. Two common crystal structures found for ionic crystals, the sodium chloride and the cesium chloride structures, were shown in Chapter 1.

The electronic configurations of all ions of a simple ionic crystal correspond to closed electronic shells, as in the inert gas atoms. In lithium fluoride the configuration of the neutral atoms are, according to the periodic table in the front endpapers of this book, Li: $1s^2 2s$, F: $1s^2 2s^2 2p^5$. The singly charged ions have the configurations Li^+:$1s^2$, F^-: $1s^2 2s^2 2p^6$, as for helium and neon, respectively. Inert gas atoms have closed shells, and the charge distributions are spherically symmetric. We expect that the charge distributions on each ion in an ionic crystal will have approximately spherical symmetry, with some distortion near the region of contact with neighboring atoms. This picture is confirmed by x-ray studies of electron distributions (Fig. 7).

A quick estimate suggests that we are not misguided in looking to electrostatic interactions for a large part of the binding energy of an ionic crystal. The distance between a positive ion and the nearest negative ion in crystalline sodium chloride is 2.81×10^{-8} cm, and the attractive coulomb part of the potential energy of the two ions by themselves is 5.1 eV. This value may be compared (Fig. 8) with the experimental value of 7.9 eV per molecular unit for the lattice energy of crystalline NaCl with respect to separated Na^+ and Cl^- ions. We now calculate the energy more closely.

Electrostatic or Madelung Energy

The long-range interaction between ions with charge $\pm q$ is the electrostatic interaction $\pm q^2/r$, attractive between ions of opposite charge and repulsive

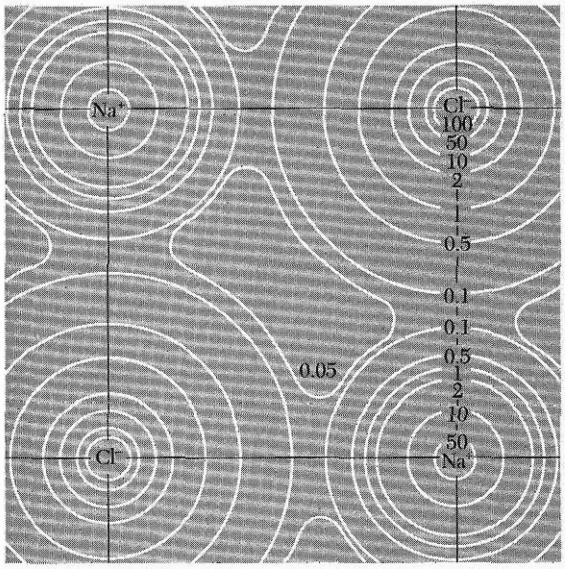


Figure 7 Electron density distribution in the base plane of NaCl, after x-ray studies by G. Schoknecht. The numbers on the contours give the relative electron concentration.

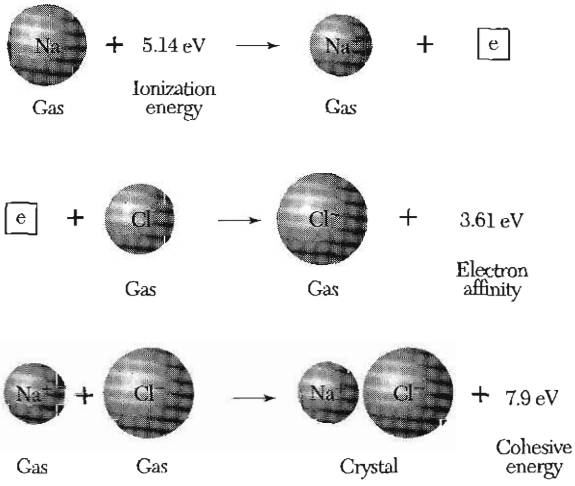


Figure 8 The energy per molecule unit of a crystal of sodium chloride is (7.9 − 5.1 + 3.6) = 6.4 eV lower than the energy of separated neutral atoms. The lattice energy with respect to separated ions is 7.9 eV per molecule unit. All values on the figure are experimental. Values of the ionization energy are given in Table 5, and values of the electron affinity are given in Table 6.

between ions of the same charge. The ions arrange themselves in whatever crystal structure gives the strongest attractive interaction compatible with the repulsive interaction at short distances between ion cores. The repulsive interactions between ions with inert gas configurations are similar to those between inert gas atoms. The van der Waals part of the attractive interaction in ionic crystals makes a relatively small contribution to the cohesive energy in ionic crystals, of the order of 1 or 2 percent. The main contribution to the binding energy of ionic crystals is electrostatic and is called the **Madelung energy**.

Table 6 Electron affinities of negative ions

The electron affinity is positive for a stable negative ion.

Atom	Electron affinity energy eV	Atom	Electron affinity energy eV
H	0.7542	Si	1.39
Li	0.62	P	0.74
C	1.27	S	2.08
O	1.46	Cl	3.61
F	3.40	Br	3.36
Na	0.55	I	3.06
Al	0.46	K	0.50

Source: H. Hotop and W. C. Lineberger, J. Phys. Chem. Ref. Data 4, 539 (1975).

If U_{ij} is the interaction energy between ions i and j, we define a sum U_i which includes all interactions involving the ion i:

$$U_i = \sum_j{}' U_{ij} \ , \tag{17}$$

where the summation includes all ions except $j = i$. We suppose that U_{ij} may be written as the sum of a central field repulsive potential of the form $\lambda \exp(-r/\rho)$, where λ and ρ are empirical parameters, and a coulomb potential $\pm q^2/r$. Thus

$$\text{(CGS)} \qquad U_{ij} = \lambda \exp(-r_{ij}/\rho) \pm q^2/r_{ij} \ , \tag{18}$$

where the + sign is taken for the like charges and the – sign for unlike charges. In SI units the coulomb interaction is $\pm q^2/4\pi\epsilon_0 r$; we write this section in CGS units in which the coulomb interaction is $\pm q^2/r$.

The repulsive term describes the fact that each ion resists overlap with the electron distributions of neighboring ions. We treat the strength λ and range ρ as constants to be determined from observed values of the lattice constant and compressibility; we have used the exponential form of the empirical repulsive potential rather than the R^{-12} form used for the inert gases. The change is made because it may give a better representation of the repulsive interaction. For the ions, we do not have gas-phase data available to permit the independent determination of λ and ρ. We note that ρ is a measure of the range of the repulsive interaction; when $r = \rho$, the repulsive interaction is reduced to e^{-1} of the value at $r = 0$.

In the NaCl structure the value of U_i does not depend on whether the reference ion i is a positive or a negative ion. The sum in (17) can be arranged to converge rapidly, so that its value will not depend on the site of the reference ion in the crystal, as long as it is not near the surface. We neglect surface effects

and write the total lattice energy U_{tot} of a crystal composed of N molecules or $2N$ ions as $U_{tot} = NU_i$. Here N, rather than $2N$, occurs because we must count each *pair* of interactions only once or each bond only once. The total lattice energy is defined as the energy required to separate the crystal into individual ions at an infinite distance apart.

It is convenient again to introduce quantities p_{ij} such that $r_{ij} \equiv p_{ij}R$, where R is the nearest-neighbor separation in the crystal. If we include the repulsive interaction only among nearest neighbors, we have

$$\text{(CGS)} \qquad U_{ij} = \begin{cases} \lambda \exp(-R/\rho) - \dfrac{q^2}{R} & \text{(nearest neighbors)} \\ \pm\dfrac{1}{p_{ij}}\dfrac{q^2}{R} & \text{(otherwise).} \end{cases} \tag{19}$$

Thus

$$\text{(CGS)} \qquad U_{tot} = NU_i = N\left(z\lambda e^{-R/\rho} - \frac{\alpha q^2}{R}\right), \tag{20}$$

where z is the number of nearest neighbors of any ion and

$$\boxed{\alpha \equiv \sum_j{}' \frac{(\pm)}{p_{ij}} \equiv \textbf{Madelung constant}} \ . \tag{21}$$

The sum should include the nearest-neighbor contribution, which is just z. The $(\pm)$ sign is discussed just before (25). The value of the Madelung constant is of central importance in the theory of an ionic crystal. Methods for its calculation are discussed next.

At the equilibrium separation $dU_{tot}/dR = 0$, so that

$$\text{(CGS)} \qquad N\frac{dU_i}{dR} = -\frac{Nz\lambda}{\rho}\exp(-R/\rho) + \frac{N\alpha q^2}{R^2} = 0 \ , \tag{22}$$

or

$$\text{(CGS)} \qquad R_0^2 \exp(-R_0/\rho) = \rho\alpha q^2/z\lambda \ . \tag{23}$$

This determines the equilibrium separation R_0 if the parameters ρ, λ of the repulsive interaction are known. For SI, replace q^2 by $q^2/4\pi\epsilon_0$.

The total lattice energy of the crystal of $2N$ ions at their equilibrium separation R_0 may be written, using (20) and (23), as

$$\text{(CGS)} \qquad U_{tot} = -\frac{N\alpha q^2}{R_0}\left(1 - \frac{\rho}{R_0}\right) \ . \tag{24}$$

The term $-N\alpha q^2/R_0$ is the Madelung energy. We shall find that ρ is of the order of $0.1R_0$, so that the repulsive interaction has a very short range.

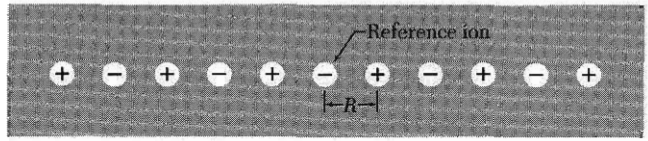


Figure 9 Line of ions of alternating signs, with distance R between ions.

Evaluation of the Madelung Constant

The first calculation of the coulomb energy constant α was made by Madelung. A powerful general method for lattice sum calculations was developed by Ewald and is developed in Appendix B. Computers are now used for the calculations.

The definition of the Madelung constant α is, by (21),

$$\alpha = \sum_j{}' \frac{(\pm)}{p_{ij}} .$$

For (20) to give a stable crystal it is necessary that α be positive. If we take the reference ion as a negative charge, the plus sign will apply to positive ions and the minus sign to negative ions.

An equivalent definition is

$$\frac{\alpha}{R} = \sum_j{}' \frac{(\pm)}{r_j} , \tag{25}$$

where r_j is the distance of the jth ion from the reference ion and R is the nearest-neighbor distance. The value given for α will depend on whether it is defined in terms of the nearest-neighbor distance R or in terms of the lattice parameter a or in terms of some other relevant length.

As an example, we compute the Madelung constant for the infinite line of ions of alternating sign in Fig. 9. Pick a negative ion as reference ion, and let R denote the distance between adjacent ions. Then

$$\frac{\alpha}{R} = 2\left[\frac{1}{R} - \frac{1}{2R} + \frac{1}{3R} - \frac{1}{4R} + \cdots\right] ,$$

or

$$\alpha = 2\left[1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots\right] ;$$

the factor 2 occurs because there are two ions, one to the right and one to the left, at equal distances r_j. We sum this series by the expansion

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots .$$

Thus the Madelung constant for the one-dimensional chain is $\alpha = 2 \ln 2$.

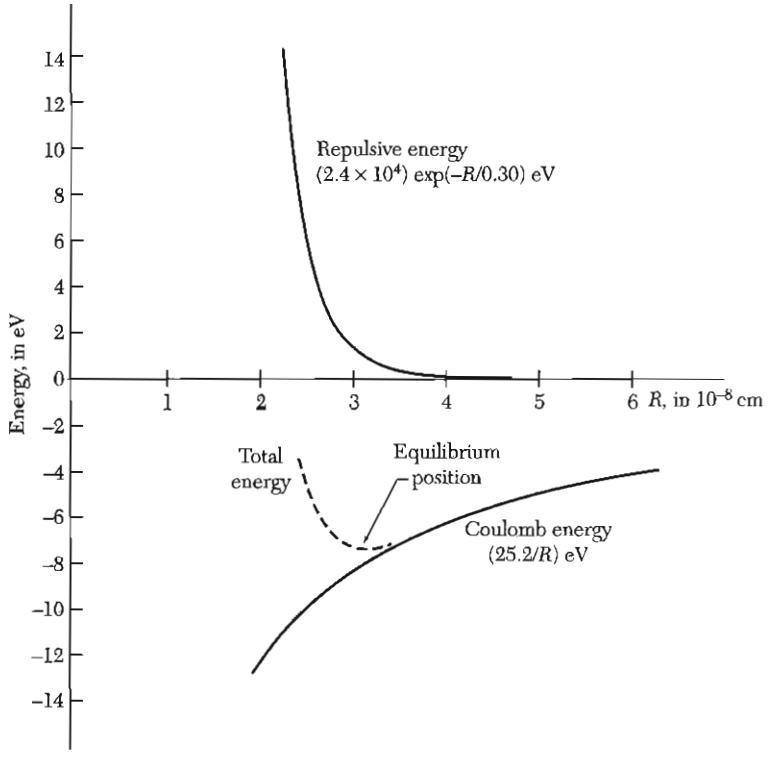


Figure 10 Energy per molecule of KCl crystal, showing Madelung (coulomb) and repulsive contributions.

In three dimensions the series presents greater difficulty. It is not possible to write down the successive terms by a casual inspection. More important, the series will not converge unless the successive terms in the series are arranged so that the contributions from the positive and negative terms nearly cancel.

Typical values of the Madelung constant are listed below, based on unit charges and *referred to the nearest-neighbor distance:*

Structure	α
Sodium chloride, NaCl	1.747565
Cesium chloride, CsCl	1.762675
Zinc blende, cubic ZnS	1.6381

The Madelung and repulsive contributions to the binding of a KCl crystal are shown in Fig. 10. Properties of alkali halide crystals having the sodium chloride structure are given in Table 7. The calculated values of the lattice energy are in exceedingly good agreement with the observed values.

Table 7 Properties of alkali halide crystals with the NaCl structure

All values (except those in square brackets) at room temperature and atmospheric pressure, with no correction for changes in R_0 and U from absolute zero. Values in square brackets at absolute zero temperature and zero pressure, from private communication by L. Brewer.

	Nearest-neighbor separation R_0 in Å	Bulk modulus B, in 10^{11} dyn/cm^2 or 10^{10} N/m^2	Repulsive energy parameter $z\lambda$, in 10^{-8} erg	Repulsive range parameter ρ, in Å	Lattice energy compared to free ions, in kcal/mol	
					Experimental	Calculated
LiF	2.014	6.71	0.296	0.291	242.3[246.8]	242.2
LiCl	2.570	2.98	0.490	0.330	198.9[201.8]	192.9
LiBr	2.751	2.38	0.591	0.340	189.8	181.0
LiI	3.000	(1.71)	0.599	0.366	177.7	166.1
NaF	2.317	4.65	0.641	0.290	214.4[217.9]	215.2
NaCl	2.820	2.40	1.05	0.321	182.6[185.3]	178.6
NaBr	2.989	1.99	1.33	0.328	173.6[174.3]	169.2
NaI	3.237	1.51	1.58	0.345	163.2[162.3]	156.6
KF	2.674	3.05	1.31	0.298	189.8[194.5]	189.1
KCl	3.147	1.74	2.05	0.326	165.8[169.5]	161.6
KBr	3.298	1.48	2.30	0.336	158.5[159.3]	154.5
KI	3.533	1.17	2.85	0.348	149.9[151.1]	144.5
RbF	2.815	2.62	1.78	0.301	181.4	180.4
RbCl	3.291	1.56	3.19	0.323	159.3	155.4
RbBr	3.445	1.30	3.03	0.338	152.6	148.3
RbI	3.671	1.06	3.99	0.348	144.9	139.6

Data from various tables by M. P. Tosi, Solid State Physics **16**, 1 (1964).

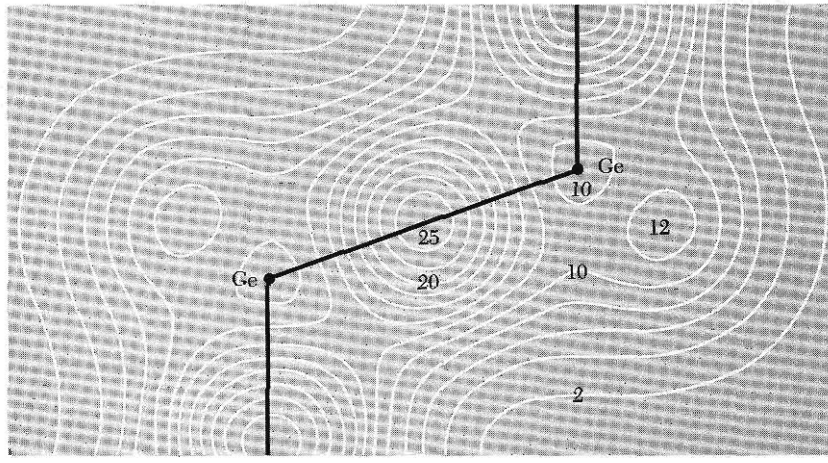


Figure 11 Calculated valence electron concentration in germanium. The numbers on the contours give the electron concentration per primitive cell, with four valence electrons per atom (eight electrons per primitive cell). Note the high concentration midway along the Ge-Ge bond, as we expect for covalent bonding. (After J. R. Chelikowsky and M. L. Cohen.)

COVALENT CRYSTALS

The covalent bond is the classical electron pair or homopolar bond of chemistry, particularly of organic chemistry. It is a strong bond: the bond between two carbon atoms in diamond with respect to separated neutral atoms is comparable with the bond strength in ionic crystals.

The covalent bond is usually formed from two electrons, one from each atom participating in the bond. The electrons forming the bond tend to be partly localized in the region between the two atoms joined by the bond. The spins of the two electrons in the bond are antiparallel.

The covalent bond has strong directional properties (Fig. 11). Thus carbon, silicon, and germanium have the diamond structure, with atoms joined to four nearest neighbors at tetrahedral angles, even though this arrangement gives a low filling of space, 0.34 of the available space, compared with 0.74 for a close-packed structure. The tetrahedral bond allows only four nearest neighbors, whereas a close-packed structure has 12. We should not overemphasize the similarity of the bonding of carbon and silicon. Carbon gives biology, but silicon gives geology and semiconductor technology.

The binding of molecular hydrogen is a simple example of a covalent bond. The strongest binding (Fig. 12) occurs when the spins of the two electrons are antiparallel. The binding depends on the relative spin orientation not because there are strong magnetic dipole forces between the spins, but because the Pauli principle modifies the distribution of charge according to the spin orientation. This spin-dependent coulomb energy is called the **exchange interaction**.

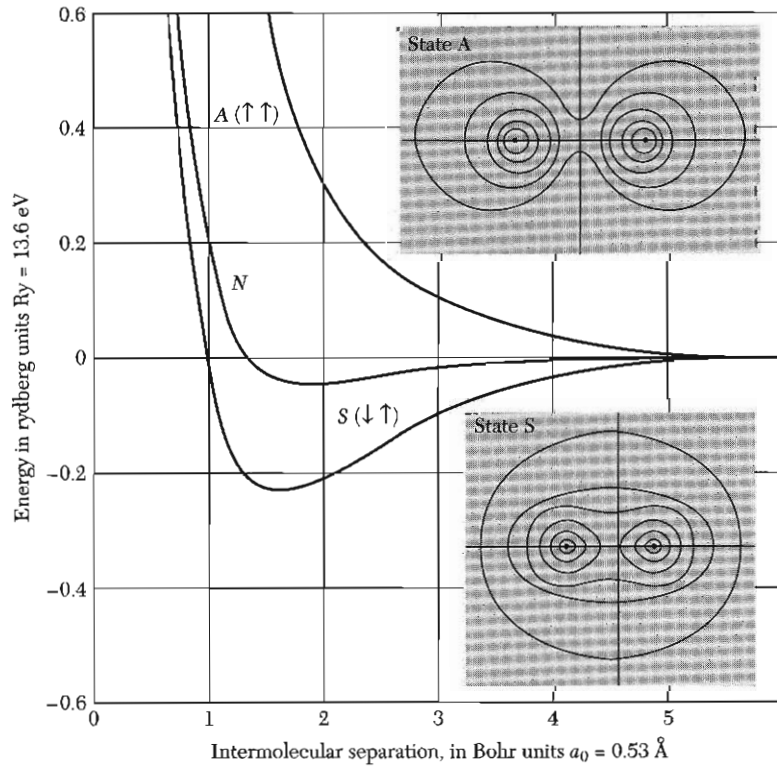


Figure 12 Energy of molecular hydrogen (H_2) referred to separated neutral atoms. A negative energy corresponds to binding. The curve *N* refers to a classical calculation with free atom charge densities; *A* is the result for parallel electron spins, taking the Pauli exclusion principle into account, and *S* (the stable state) for antiparallel spins. The density of charge is represented by contour lines for the states *A* and *S*.

The Pauli principle gives a strong repulsive interaction between atoms with filled shells. If the shells are not filled, electron overlap can be accommodated without excitation of electrons to high energy states and the bond will be shorter. Compare the bond length (2 Å) of Cl_2 with the interatomic distance (3.76 Å) of Ar in solid Ar; also compare the cohesive energies given in Table 1. The difference between Cl_2 and Ar_2 is that the Cl atom has five electrons in the $3p$ shell and the Ar atom has six, filling the shell, so that the repulsive interaction is stronger in Ar than in Cl.

The elements C, Si, and Ge lack four electrons with respect to filled shells, and thus these elements (for example) can have an attractive interaction associated with charge overlap. The electron configuration of carbon is $1s^2 2s^2 2p^2$. To form a tetrahedral system of covalent bonds the carbon atom must first be promoted to the electronic configuration $1s^2 2s 2p^3$. This promotion from the ground state requires 4 eV, an amount more than regained when the bonds are formed.

Table 8 Fractional ionic character of bonds in binary crystals

Crystal	Fractional ionic character	Crystal	Fractional ionic character
Si	0.00		
SiC	0.18	GaAs	0.31
Ge	0.00	GaSb	0.26
ZnO	0.62	AgCl	0.86
ZnS	0.62	AgBr	0.85
ZnSe	0.63	AgI	0.77
ZnTe	0.61	MgO	0.84
CdO	0.79	MgS	0.79
CdS	0.69	MgSe	0.79
CdSe	0.70		
CdTe	0.67	LiF	0.92
		NaCl	0.94
InP	0.42	RbF	0.96
InAs	0.36		
InSb	0.32		

After J. C. Phillips, *Bonds and bands in semiconductors.*

There is a continuous range of crystals between the ionic and the covalent limits. It is often important to estimate the extent a given bond is ionic or covalent. A semiempirical theory of the fractional ionic or covalent character of a bond in a dielectric crystal has been developed with considerable success by J. C. Phillips, Table 8.

METALS

Metals are characterized by high electrical conductivity, and a large number of electrons in a metal are free to move about, usually one or two per atom. The electrons available to move about are called *conduction electrons*. The valence electrons of the atom become the conduction electrons of the metal.

In some metals the interaction of the ion cores with the conduction electrons always makes a large contribution to the binding energy, but the characteristic feature of metallic binding is the lowering of the energy of the valence electrons in the metal as compared with the free atom.

The binding energy of an alkali metal crystal is considerably less than that of an alkali halide crystal: the bond formed by a conduction electron is not very strong. The interatomic distances are relatively large in the alkali metals because the kinetic energy of the conduction electrons is lower at large interatomic distances. This leads to weak binding. Metals tend to crystallize in relatively

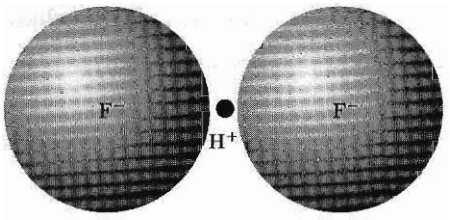


Figure 13 The hydrogen difluoride ion HF_2^- is stabilized by a hydrogen bond. The sketch is of an extreme model of the bond, extreme in the sense that the proton is shown bare of electrons.

close packed structures: hcp, fcc, bcc, and some other closely related structures, and not in loosely-packed structures such as diamond.

In the transition metals there is additional binding from inner electron shells. Transition metals and the metals immediately following them in the periodic table have large *d*-electron shells and are characterized by high binding energy.

HYDROGEN BONDS

Because neutral hydrogen has only one electron, it should form a covalent bond with only one other atom. It is known, however, that under certain conditions an atom of hydrogen is attracted by rather strong forces to two atoms, thus forming a **hydrogen bond** between them, with a bond energy of the order of 0.1 eV. It is believed that the hydrogen bond is largely ionic in character, being formed only between the most electronegative atoms, particularly F, O, and N. In the extreme ionic form of the hydrogen bond, the hydrogen atom loses its electron to another atom in the molecule; the bare proton forms the hydrogen bond. The atoms adjacent to the proton are so close that more than two of them would get in each other's way; thus the hydrogen bond connects only two atoms (Fig. 13).

The hydrogen bond is an important part of the interaction between H_2O molecules and is responsible together with the electrostatic attraction of the electric dipole moments for the striking physical properties of water and ice. It is important in certain ferroelectric crystals and in DNA.

ATOMIC RADII

Distances between atoms in crystals can be measured very accurately by x-ray diffraction, often to 1 part in 10^5. Can we say that the observed distance between atoms may be assigned partly to atom A and partly to atom B? Can a definite meaning be assigned to the radius of an atom or an ion, irrespective of the nature and composition of the crystal?

Strictly, the answer is no. The charge distribution around an atom is not limited by a rigid spherical boundary. Nonetheless, the concept of an atomic

Table 9 Atomic and ionic radii

Values approximate only. Units are 1 Å$=10^{-10}$ m. For original references see W. B. Pearson, *Crystal chemistry and physics of metals and alloys*, Wiley, 1972.

H 2.08																	He
Li 0.68 1.56	Be 0.35 1.06 1.13											B 0.23 0.88 0.98	C 0.15 0.77 0.92	N 1.71 0.70	O 1.40 0.66	F 1.36 0.64	Ne 1.58
Na 0.97 1.91	Mg 0.65 1.40 1.60	← Standard radii for ions in inert gas (filled shell) configuration →										Al 0.50 1.26 1.43	Si 0.41 1.17 1.32	P 2.12 1.10	S 1.84 1.04	Cl 1.81 0.99	Ar 1.88
		← Radii of atoms when in tetrahedral covalent bonds →															
		← Radii of ions in 12-coordinated metals →															
K 1.33 2.38	Ca 0.99 1.98	Sc 0.81 1.64	Ti 0.68 1.46	V 1.35	Cr 1.28	Mn 1.26	Fe 1.27	Co 1.25	Ni 1.25	Cu 1.35 1.28	Zn 0.74 1.31 1.39	Ga 0.62 1.26 1.41	Ge 0.53 1.22 1.37	As 2.22 1.18 1.39	Se 1.98 1.14	Br 1.95 1.11	Kr 2.00
Rb 1.48 2.55	Sr 1.13 2.15	Y 0.93 1.80	Zr 0.80 1.60	Nb 0.67 1.47	Mo 1.40	Tc 1.36	Ru 1.34	Rh 1.35	Pd 1.38	Ag 1.26 1.52 1.45	Cd 0.97 1.48 1.57	In 0.81 1.44 1.66	Sn 0.71 1.40 1.55	Sb 2.45 1.36 1.59	Te 2.21 1.32	I 2.16 1.28	Xe 2.17
Cs 1.67 2.73	Ba 1.35 2.24	La 1.15 1.88	Hf 1.58	Ta 1.47	W 1.41	Re 1.38	Os 1.35	Ir 1.36	Pt 1.39	Au 1.37 1.44	Hg 1.10 1.48 1.57	Tl 0.95 1.72	Pb 0.84 1.75	Bi 1.70	Po 1.76	At	Rn
Fr 1.75	Ra 1.37	Ac 1.11															

Ce 1.01 1.71– 1.82	Pr 1.83	Nd 1.82	Pm 1.81	Sm 1.80	Eu 2.04^{2+} 1.80^{3+}	Gd 1.80	Tb 1.78	Dy 1.77	Ho 1.77	Er 1.76	Tm 1.75	Yb 1.94^{2+}– 1.74^{3+}	Lu
Th 0.99 1.80	Pa 0.90 1.63	U 0.83 1.56	Np 1.56	Pu 1.58– 1.64	Am 1.81	Cm	Bk	Cf	Es	Fm	Md	No	Lr

radius is fruitful in predicting interatomic spacing. The existence and probable lattice constants of phases that have not yet been synthesized can be predicted from the additive properties of the atomic radii. Further, the electronic configuration of the constituent atoms often can be inferred by comparison of measured and predicted values of the lattice constants.

To make predictions of lattice constants it is convenient to assign (Table 9) sets of self-consistent radii to various types of bonds: one set for ionic crystals with the constituent ions 6-coordinated in inert gas closed-shell configurations, another set for the ions in tetrahedrally-coordinated structures, and another set for 12-coordinated (close-packed) metals.

The predicted self-consistent radii of the cation Na^+ and the anion F^- as given in Table 9 would lead to 0.97 Å + 1.36 Å = 2.33 Å for the interatomic separation in the crystal NaF, as compared with the observed 2.32 Å. This agreement is much better than if we assume atomic (neutral) configurations for Na and F, for this would lead to 2.58 Å for the interatomic separation in the crystal. The latter value is $\frac{1}{2}$(n.n. distance in metallic Na+ interatomic distance in gaseous F_2).

The interatomic distance between C atoms in diamond is 1.54 Å; one-half of this is 0.77 Å. In silicon, which has the same crystal structure, one-half the interatomic distance is 1.17 Å. In SiC each atom is surrounded by four atoms of the opposite kind. If we add the C and Si radii just given, we predict 1.94 Å for the length of the C-Si bond, in fair agreement with the 1.89 Å observed for the bond length. This is the kind of agreement (a few percent) that we shall find in using tables of atomic radii.

Ionic Crystal Radii

Table 9 gives the ionic crystal radii in inert gas configurations for 6-fold coordination. The ionic radii can be used in conjunction with Table 10. Let us

Table 10 Use of the standard radii of ions given in Table 9

The interionic distance D is represented by $D_N = R_C + R_A + \Delta_N$, for ionic crystals, where N is the coordination number of the cation (positive ion), R_C and R_A are the standard radii of the cation and anion, and Δ_N is a correction for coordination number. Room temperature. (After Zachariasen.)

N	Δ_N(Å)	N	Δ_N(Å)	N	Δ_N(Å)
1	−0.50	5	−0.05	9	+0.11
2	−0.31	6	0	10	+0.14
3	−0.19	7	+0.04	11	+0.17
4	−0.11	8	+0.08	12	+0.19

consider $BaTiO_3$ with a lattice constant of 4.004 Å at room temperature. Each Ba^{++} ion has 12 nearest O^{--} ions, so that the coordination number is 12 and the correction Δ_{12} of Table 10 applies. If we suppose that the structure is determined by the Ba-O contacts, we have $D_{12} = 1.35 + 1.40 + 0.19 = 2.94$ Å or $a = 4.16$ Å; if the Ti-O contact determines the structure, we have $D_6 = 0.68 + 1.40 = 2.08$ or $a = 4.16$ Å. The actual lattice constant is somewhat smaller than the estimates and may perhaps suggest that the bonding is not purely ionic, but is partly covalent.

ANALYSIS OF ELASTIC STRAINS

We consider the elastic properties of a crystal viewed as a homogeneous continuous medium rather than as a periodic array of atoms. The continuum approximation is usually valid for elastic waves of wavelengths λ longer than 10^{-6} cm, which means for frequencies below 10^{11} or 10^{12} Hz. Some of the material below looks complicated because of the unavoidable multiplicity of subscripts on the symbols. The basic physical ideas are simple: we use Hooke's law and Newton's second law. **Hooke's law** states that in an elastic solid the strain is directly proportional to the stress. The law applies to small strains only. We say that we are in the **nonlinear region** when the strains are so large that Hooke's law is no longer satisfied.

We specify the strain in terms of the components e_{xx}, e_{yy}, e_{zz}, e_{xy}, e_{yz}, e_{zx} which are defined below. We treat infinitesimal strains only. We shall not distinguish in our notation between isothermal (constant temperature) and adiabatic (constant entropy) deformations. The small differences between the isothermal and adiabatic elastic constants are not often of importance at room temperature and below.

We imagine that three orthogonal vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ of unit length are embedded securely in the unstrained solid, as shown in Fig. 14. After a small uniform deformation of the solid has taken place, the axes are distorted in orientation and in length. In a uniform deformation each primitive cell of the crystal is deformed in the same way. The new axes $\mathbf{x}'$, $\mathbf{y}'$, $\mathbf{z}'$ may be written in terms of the old axes:

$$\begin{aligned} \mathbf{x}' &= (1 + \epsilon_{xx})\hat{\mathbf{x}} + \epsilon_{xy}\hat{\mathbf{y}} + \epsilon_{xz}\hat{\mathbf{z}} \ ; \\ \mathbf{y}' &= \epsilon_{yx}\hat{\mathbf{x}} + (1 + \epsilon_{yy})\hat{\mathbf{y}} + \epsilon_{yz}\hat{\mathbf{z}} \ ; \\ \mathbf{z}' &= \epsilon_{zx}\hat{\mathbf{x}} + \epsilon_{zy}\hat{\mathbf{y}} + (1 + \epsilon_{zz})\hat{\mathbf{z}} \ . \end{aligned} \tag{26}$$

The coefficients $\epsilon_{\alpha\beta}$ define the deformation; they are dimensionless and have values $\ll 1$ if the strain is small. The original axes were of unit length, but the new axes will not necessarily be of unit length. For example,

$$\mathbf{x}' \cdot \mathbf{x}' = 1 + 2\epsilon_{xx} + \epsilon_{xx}^2 + \epsilon_{xy}^2 + \epsilon_{xz}^2 \ ,$$

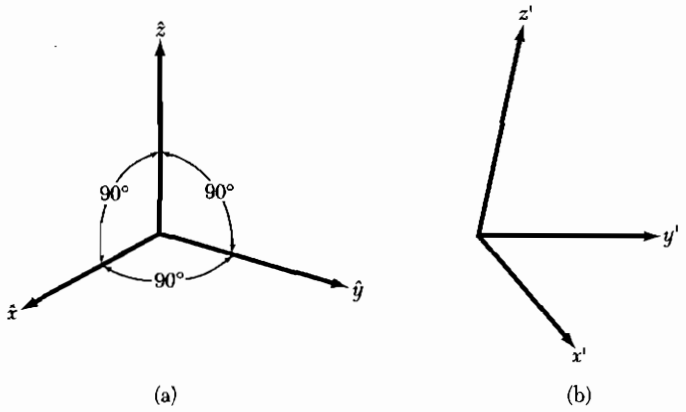


Figure 14 Coordinate axes for the description of the state of strain; the orthogonal unit axes in the unstrained state (a) are deformed in the strained state (b).

whence $x' \cong 1 + \epsilon_{xx} + \cdots$. The fractional changes of length of the $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ axes are ϵ_{xx}, ϵ_{yy}, ϵ_{zz}, respectively, to the first order.

What is the effect of the deformation (26) on an atom originally at $\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$? The origin is taken at some other atom. If the deformation is uniform, then after deformation the point will be at the position $\mathbf{r}' = x\mathbf{x}' + y\mathbf{y}' + z\mathbf{z}'$. This is obviously correct if we choose the $\hat{\mathbf{x}}$ axis such that $\mathbf{r} = x\hat{\mathbf{x}}$; then $\mathbf{r}' = x\mathbf{x}'$ by definition of $\mathbf{x}'$. The displacement $\mathbf{R}$ of the deformation is defined by

$$\mathbf{R} \equiv \mathbf{r}' - \mathbf{r} = x(\mathbf{x}' - \hat{\mathbf{x}}) + y(\mathbf{y}' - \hat{\mathbf{y}}) + z(\mathbf{z}' - \hat{\mathbf{z}}) \ , \tag{27}$$

or, from (26),

$$\mathbf{R}(\mathbf{r}) \equiv (x\epsilon_{xx} + y\epsilon_{yx} + z\epsilon_{zx})\hat{\mathbf{x}} + (x\epsilon_{xy} + y\epsilon_{yy} + z\epsilon_{zy})\hat{\mathbf{y}} + (x\epsilon_{xx} + y\epsilon_{yz} + z\epsilon_{zz})\hat{\mathbf{z}} \ . \tag{28}$$

This may be written in a more general form by introducing u, v, w such that the displacement is given by

$$\boxed{\mathbf{R}(\mathbf{r}) = u(\mathbf{r})\hat{\mathbf{x}} + v(\mathbf{r})\hat{\mathbf{y}} + w(\mathbf{r})\hat{\mathbf{z}} \ .} \tag{29}$$

If the deformation is nonuniform we must relate u, v, w to the local strains. We take the origin of $\mathbf{r}$ close to the region of interest; then comparison of (28) and (29) gives, by Taylor series expansion of $\mathbf{R}$ using $\mathbf{R}(0) = 0$,

$$x\epsilon_{xx} \cong x\frac{\partial u}{\partial x} \ ; \qquad y\epsilon_{yx} = y\frac{\partial u}{\partial y} \ ; \qquad \text{etc.} \tag{30}$$

It is usual to work with coefficients $e_{\alpha\beta}$ rather than $\epsilon_{\alpha\beta}$. We define the strain components e_{xx}, e_{yy}, e_{zz} by the relations

$$e_{xx} \equiv \epsilon_{xx} = \frac{\partial u}{\partial x} \; ; \qquad e_{yy} \equiv \epsilon_{yy} = \frac{\partial v}{\partial y} \; ; \qquad e_{zz} \equiv \epsilon_{zz} = \frac{\partial w}{\partial z} \; , \tag{31}$$

using (30). The other strain components e_{xy}, e_{yz}, e_{zx} are defined in terms of the changes in angle between the axes: using (26) we may define

$$\begin{aligned} e_{xy} &\equiv \mathbf{x}' \cdot \mathbf{y}' \cong \epsilon_{yx} + \epsilon_{xy} = \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \; ; \\ e_{yz} &\equiv \mathbf{y}' \cdot \mathbf{z}' \cong \epsilon_{zy} + \epsilon_{yz} = \frac{\partial v}{\partial z} + \frac{\partial w}{\partial y} \; ; \\ e_{zx} &\equiv \mathbf{z}' \cdot \mathbf{x}' \cong \epsilon_{zx} + \epsilon_{xz} = \frac{\partial u}{\partial z} + \frac{\partial w}{\partial x} \; . \end{aligned} \tag{32}$$

We may replace the $\cong$ signs by $=$ signs if we neglect terms of order ϵ^2. The six dimensionless coefficients $e_{\alpha\beta}(=e_{\beta\alpha})$ completely define the strain.

Dilation

The fractional *increase* of volume associated with a deformation is called the dilation. The dilation is negative for hydrostatic pressure. The unit cube of edges $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ has a volume after deformation of

$$V' = \mathbf{x}' \cdot \mathbf{y}' \times \mathbf{z}' \; , \tag{33}$$

by virtue of a well-known result for the volume of a parallelepiped having edges $\mathbf{x}'$, $\mathbf{y}'$, $\mathbf{z}'$. From (26) we have

$$\mathbf{x}' \cdot \mathbf{y}' \times \mathbf{z}' = \begin{vmatrix} 1 + \epsilon_{xx} & \epsilon_{xy} & \epsilon_{xz} \\ \epsilon_{yx} & 1 + \epsilon_{yy} & \epsilon_{yz} \\ \epsilon_{zx} & \epsilon_{zy} & 1 + \epsilon_{zz} \end{vmatrix} \cong 1 + e_{xx} + e_{yy} + e_{zz} \; . \tag{34}$$

Products of two strain components have been neglected. The dilation δ is then given by

$$\delta \equiv \frac{V' - V}{V} \cong e_{xx} + e_{yy} + e_{zz} \; . \tag{35}$$

Stress Components

The force acting on a unit area in the solid is defined as the stress. There are nine stress components: X_x, X_y, X_z, Y_x, Y_y, Y_z, Z_x, Z_y, Z_z. The capital letter indicates the direction of the force, and the subscript indicates the normal to the plane to which the force is applied. In Fig. 15 the stress component X_x

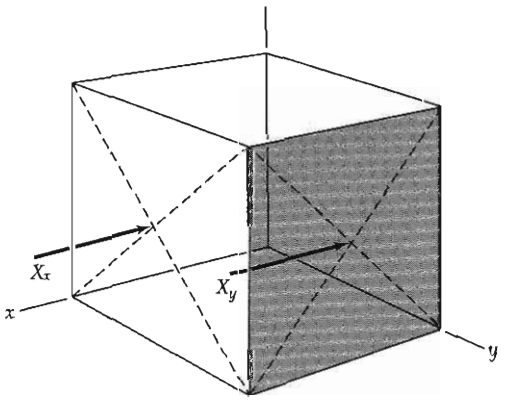


Figure 15 Stress component X_x is a force applied in the x direction to a unit area of a plane whose normal lies in the x direction; X_y is applied in the x direction to a unit area of a plane whose normal lies in the y direction.

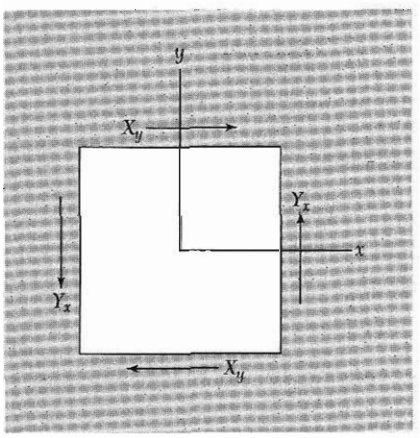


Figure 16 Demonstration that for a body in static equilibrium $Y_x = X_y$. The sum of the forces in the x direction is zero. The sum of the forces in the y direction is also zero. The total force vanishes. The total torque about the origin is also zero if $Y_x = X_y$.

represents a force applied in the x direction to a unit area of a plane whose normal lies in the x direction; the stress component X_y represents a force applied in the x direction to a unit area of a plane whose normal lies in the y direction. The number of independent stress components is reduced from nine to six by applying to an elementary cube (as in Fig. 16) the condition that the angular acceleration vanish, and hence that the total torque must be zero. It follows that

$$Y_z = Z_y \ ; \qquad Z_x = X_z \ ; \qquad X_y = Y_x \ . \tag{36}$$

The six independent stress components may be taken as X_x, Y_y, Z_z, Y_z, Z_x, X_y.

Stress components have the dimensions of force per unit area or energy per unit volume. The strain components are ratios of lengths and are dimensionless.

ELASTIC COMPLIANCE AND STIFFNESS CONSTANTS

Hooke's law states that for sufficiently small deformations the strain is directly proportional to the stress, so that the strain components are linear functions of the stress components:

$$\begin{aligned}
e_{xx} &= S_{11}X_x + S_{12}Y_y + S_{13}Z_z + S_{14}Y_z + S_{15}Z_x + S_{16}X_y \ ; \\
e_{yy} &= S_{21}X_x + S_{22}Y_y + S_{23}Z_z + S_{24}Y_z + S_{25}Z_x + S_{26}X_y \ ; \\
e_{zz} &= S_{31}X_x + S_{32}Y_y + S_{33}Z_z + S_{34}Y_z + S_{35}Z_x + S_{36}X_y \ ; \\
e_{yz} &= S_{41}X_x + S_{42}Y_y + S_{43}Z_z + S_{44}Y_z + S_{45}Z_x + S_{46}X_y \ ; \\
e_{zx} &= S_{51}X_x + S_{52}Y_y + S_{53}Z_z + S_{54}Y_z + S_{55}Z_x + S_{56}X_y \ ; \\
e_{xy} &= S_{61}X_x + S_{62}Y_y + S_{63}Z_z + S_{64}Y_z + S_{65}Z_x + S_{66}X_y \ .
\end{aligned} \tag{37}$$

$$\begin{aligned}
X_x &= C_{11}e_{xx} + C_{12}e_{yy} + C_{13}e_{zz} + C_{14}e_{yz} + C_{15}e_{zx} + C_{16}e_{xy} \ ; \\
Y_y &= C_{21}e_{xx} + C_{22}e_{yy} + C_{23}e_{zz} + C_{24}e_{yz} + C_{25}e_{zx} + C_{26}e_{xy} \ ; \\
Z_z &= C_{31}e_{xx} + C_{32}e_{yy} + C_{33}e_{zz} + C_{34}e_{yz} + C_{35}e_{zx} + C_{36}e_{xy} \ ; \\
Y_z &= C_{41}e_{xx} + C_{42}e_{yy} + C_{43}e_{zz} + C_{44}e_{yz} + C_{45}e_{zx} + C_{46}e_{xy} \ ; \\
Z_x &= C_{51}e_{xx} + C_{52}e_{yy} + C_{53}e_{zz} + C_{54}e_{yz} + C_{55}e_{zx} + C_{56}e_{xy} \ ; \\
X_y &= C_{61}e_{xx} + C_{62}e_{yy} + C_{63}e_{zz} + C_{64}e_{yz} + C_{65}e_{zx} + C_{66}e_{xy} \ .
\end{aligned} \tag{38}$$

The quantities S_{11}, S_{12} ... are called **elastic compliance constants** or elastic constants; the quantities C_{11}, C_{12}, ... are called the **elastic stiffness constants** or moduli of elasticity. The S's have the dimensions of [area]/[force] or [volume]/[energy]. The C's have the dimensions of [force]/[area] or [energy]/[volume].

Elastic Energy Density

The 36 constants in (37) or in (38) may be reduced in number by several considerations. The elastic energy density U is a quadratic function of the strains, in the approximation of Hooke's law (recall the expression for the energy of a stretched spring). Thus we may write

$$U = \frac{1}{2}\sum_{\lambda=1}^{6}\sum_{\mu=1}^{6}\tilde{C}_{\lambda\mu}e_\lambda e_\mu \ , \tag{39}$$

where the indices 1 through 6 are defined as:

$$1 \equiv xx \ ; \quad 2 \equiv yy \ ; \quad 3 \equiv zz \ ; \quad 4 \equiv yz \ ; \quad 5 \equiv zx \ ; \quad 6 \equiv xy \ . \tag{40}$$

The $\tilde{C}$'s are related to the C's of (38), as in (42) below.

The stress components are found from the derivative of U with respect to the associated strain component. This result follows from the definition of potential energy. Consider the stress X_x applied to one face of a unit cube, the opposite face being held at rest:

$$X_x = \frac{\partial U}{\partial e_{xx}} \equiv \frac{\partial U}{\partial e_1} = \tilde{C}_{11}e_1 + \frac{1}{2}\sum_{\beta=2}^{6}(\tilde{C}_{1\beta} + \tilde{C}_{\beta 1})e_\beta \ . \tag{41}$$

Note that only the combination $\frac{1}{2}(\tilde{C}_{\alpha\beta} + \tilde{C}_{\beta\alpha})$ enters the stress-strain relations. It follows that the elastic stiffness constants are symmetrical:

$$C_{\alpha\beta} = \tfrac{1}{2}(\tilde{C}_{\alpha\beta} + \tilde{C}_{\beta\alpha}) = C_{\beta\alpha} \ . \tag{42}$$

Thus the thirty-six elastic stiffness constants are reduced to twenty-one.

Elastic Stiffness Constants of Cubic Crystals

The number of independent elastic stiffness constants is reduced further if the crystal possesses symmetry elements. We now show that in cubic crystals there are only three independent stiffness constants.

We assert that the elastic energy density of a cubic crystal is

$$U = \tfrac{1}{2}C_{11}(e_{xx}^2 + e_{yy}^2 + e_{zz}^2) + \tfrac{1}{2}C_{44}(e_{yz}^2 + e_{zx}^2 + e_{xy}^2) + C_{12}(e_{yy}e_{zz} + e_{zz}e_{xx} + e_{xx}e_{yy}) \ , \tag{43}$$

and that no other quadratic terms occur; that is,

$$(e_{xx}e_{xy} + \cdots) \ ; \qquad (e_{yz}e_{zx} + \cdots) \ ; \qquad (e_{xx}e_{yz} + \cdots) \tag{44}$$

do not occur.

The minimum symmetry requirement for a cubic structure is the existence of four three-fold rotation axes. The axes are in the [111] and equivalent directions (Fig. 17). The effect of a rotation of $2\pi/3$ about these four axes is to interchange the x, y, z axes according to the schemes

$$\begin{aligned} &x \to y \to z \to x \ ; \qquad -x \to z \to y \to -x \ ; \\ &x \to z \to -y \to x \ ; \qquad -x \to y \to z \to -x \ , \end{aligned} \tag{45}$$

according to the axis chosen. Under the first of these schemes, for example,

$$e_{xx}^2 + e_{yy}^2 + e_{zz}^2 \to e_{yy}^2 + e_{zz}^2 + e_{xx}^2 \ ,$$

and similarly for the other terms in parentheses in (43). Thus (43) is invariant under the operations considered. But each of the terms exhibited in (44) is odd in one or more indices. A rotation in the set (45) can be found which will

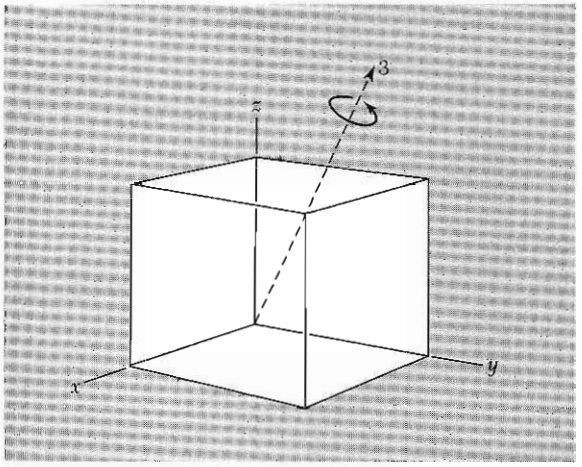


Figure 17 Rotation by $2\pi/3$ about the axis marked 3 changes $x \rightarrow y$; $y \rightarrow z$; and $z \rightarrow x$.

change the sign of the term, because $e_{xy} = -e_{x(-y)}$, for example. Thus the terms (44) are not invariant under the required operations.

It remains to verify that the numerical factors in (43) are correct. By (41),

$$\partial U/\partial e_{xx} = X_x = C_{11}e_{xx} + C_{12}(e_{yy} + e_{zz}) \ . \tag{46}$$

The appearance of $C_{11}e_{xx}$ agrees with (38). On further comparison, we see that

$$C_{12} = C_{13} \ ; \qquad C_{14} = C_{15} = C_{16} = 0 \ . \tag{47}$$

Further, from (43),

$$\partial U/\partial e_{xy} = X_y = C_{44}e_{xy} \ ; \tag{48}$$

on comparison with (38) we have

$$C_{61} = C_{62} = C_{63} = C_{64} = C_{65} = 0 \ ; \qquad C_{66} = C_{44} \ . \tag{49}$$

Thus from (43) we find that the array of values of the elastic stiffness constants is reduced for a cubic crystal to the matrix

	e_{xx}	e_{yy}	e_{zz}	e_{yz}	e_{zx}	e_{xy}
X_x	C_{11}	C_{12}	C_{12}	0	0	0
Y_y	C_{12}	C_{11}	C_{12}	0	0	0
Z_z	C_{12}	C_{12}	C_{11}	0	0	0
Y_z	0	0	0	C_{44}	0	0
Z_x	0	0	0	0	C_{44}	0
X_y	0	0	0	0	0	C_{44}

(50)

For cubic crystals the stiffness and compliance constants are related by

$$C_{14} = 1/S_{44} \; ; \qquad C_{11} - C_{12} = (S_{11} - S_{12})^{-1} \; ;$$
$$C_{11} + 2C_{12} = (S_{11} + 2S_{12})^{-1} \; . \tag{51}$$

These relations follow on evaluating the inverse matrix to (50).

Bulk Modulus and Compressibility

Consider the uniform dilation $e_{xx} = e_{yy} = e_{zz} = \frac{1}{3}\delta$. For this deformation the energy density (43) of a cubic crystal is

$$U = \tfrac{1}{6}(C_{11} + 2C_{12})\delta^2 \; . \tag{52}$$

We may define the **bulk modulus** B by the relation

$$U = \tfrac{1}{2}B\delta^2 \; , \tag{53}$$

which is equivalent to the definition $-V\,dp/dV$. For a cubic crystal,

$$B = \tfrac{1}{3}(C_{11} + 2C_{12}) \; . \tag{54}$$

The **compressibility** K is defined as $K = 1/B$. Values of B and K are given in Table 3.

ELASTIC WAVES IN CUBIC CRYSTALS

By considering as in Figs. 18 and 19 the forces acting on an element of volume in the crystal we obtain the equation of motion in the x direction

$$\rho \frac{\partial^2 u}{\partial t^2} = \frac{\partial X_x}{\partial x} + \frac{\partial X_y}{\partial y} + \frac{\partial X_z}{\partial z} \; ; \tag{55}$$

here ρ is the density and u is the displacement in the x direction. There are similar equations for the y and z directions. From (38) and (50) it follows that for a cubic crystal

$$\rho \frac{\partial^2 u}{\partial t^2} = C_{11} \frac{\partial e_{xx}}{\partial x} + C_{12}\left(\frac{\partial e_{yy}}{\partial x} + \frac{\partial e_{zz}}{\partial x}\right) + C_{44}\left(\frac{\partial e_{xy}}{\partial y} + \frac{\partial e_{zx}}{\partial z}\right) \; ; \tag{56}$$

here the x, y, z directions are parallel to the cube edges. Using the definitions (31) and (32) of the strain components we have

$$\rho \frac{\partial^2 u}{\partial t^2} = C_{11} \frac{\partial^2 u}{\partial x^2} + C_{44}\left(\frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) + (C_{12} + C_{44})\left(\frac{\partial^2 v}{\partial x\,\partial y} + \frac{\partial^2 w}{\partial x\,\partial z}\right) \; , \tag{57a}$$

where u, v, w are the components of the displacement $\mathbf{R}$ as defined by (29).

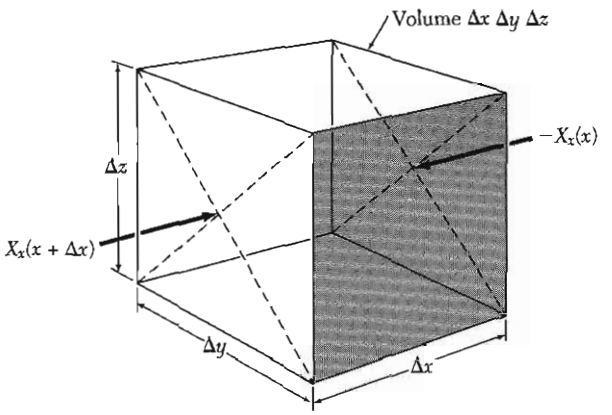


Figure 18 Cube of volume $\Delta x\ \Delta y\ \Delta z$ acted on by a stress $-X_x(x)$ on the face at x, and $X_x(x + \Delta x) = X_x(x) + \dfrac{\partial X_x}{\partial x}\Delta x$ on the parallel face at $x + \Delta x$. The net force is $\left(\dfrac{\partial X_x}{\partial x}\Delta x\right)\Delta y\ \Delta z$. Other forces in the x direction arise from the variation across the cube of the stresses X_y and X_z, which are not shown. The net x component of the force on the cube is

$$F_x = \left(\frac{\partial X_x}{\partial x} + \frac{\partial X_y}{\partial y} + \frac{\partial X_z}{\partial z}\right)\Delta x\ \Delta y\ \Delta z\ .$$

The force equals the mass of the cube times the component of the acceleration in the x direction. The mass is $\rho\ \Delta x\ \Delta y\ \Delta z$, and the acceleration is $\partial^2 u/\partial t^2$.

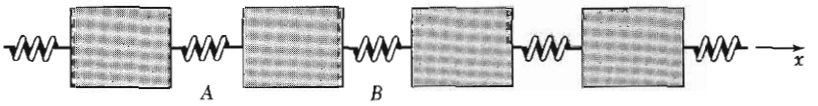


Figure 19 If springs A and B are stretched equally, the block between them experiences no net force. This illustrates the fact that a uniform stress X_x in a solid does not give a net force on a volume element. If the spring at B is stretched more than the spring at A, the block between them will be accelerated by the force $X_x(B) - X_x(A)$.

The corresponding equations of motion for $\partial^2 v/\partial t^2$ and $\partial^2 w/\partial t^2$ are found directly from (57a) by symmetry:

$$\rho\frac{\partial^2 v}{\partial t^2} = C_{11}\frac{\partial^2 v}{\partial y^2} + C_{44}\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial z^2}\right) + (C_{12} + C_{44})\left(\frac{\partial^2 u}{\partial x\,\partial y} + \frac{\partial^2 w}{\partial y\,\partial z}\right)\ ; \tag{57b}$$

$$\rho\frac{\partial^2 w}{\partial t^2} = C_{11}\frac{\partial^2 w}{\partial z^2} + C_{44}\left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2}\right) + (C_{12} + C_{44})\left(\frac{\partial^2 u}{\partial x\,\partial z} + \frac{\partial^2 v}{\partial y\,\partial z}\right)\ . \tag{57c}$$

We now look for simple special solutions of these equations.

Waves in the [100] Direction

One solution of (57a) is given by a longitudinal wave

$$u = u_0 \exp\left[i(Kx - \omega t)\right]\ , \tag{58}$$

where u is the x component of the particle displacement. Both the wavevector and the particle motion are along the x cube edge. Here $K = 2\pi/\lambda$ is the

wavevector and $\omega = 2\pi\nu$ is the angular frequency. If we substitute (58) into (57a) we find

$$\omega^2\rho = C_{11}K^2 \ ; \tag{59}$$

thus the velocity ω/K of a longitudinal wave in the [100] direction is

$$v_s = \nu\lambda = \omega/K = (C_{11}/\rho)^{1/2} \ . \tag{60}$$

Consider a transverse or shear wave with the wavevector along the x cube edge and with the particle displacement v in the y direction:

$$v = v_0 \exp\,[i(Kx - \omega t)] \ . \tag{61}$$

On substitution in (57b) this gives the dispersion relation

$$\omega^2\rho = C_{44}K^2 \ ; \tag{62}$$

thus the velocity ω/K of a transverse wave in the [100] direction is

$$v_s = (C_{44}/\rho)^{1/2} \ . \tag{63}$$

The identical velocity is obtained if the particle displacement is in the z direction. Thus for **K** parallel to [100] the two independent shear waves have equal velocities. This is not true for **K** in a general direction in the crystal.

Waves in the [110] Direction

There is a special interest in waves that propagate in a face diagonal direction of a cubic crystal, because the three elastic constants can be found simply from the three propagation velocities in this direction.

Consider a shear wave that propagates in the xy plane with particle displacement w in the z direction

$$w = w_0 \exp\,[i(K_x x + K_y y - \omega t)] \ , \tag{64}$$

whence (32c) gives

$$\omega^2\rho = C_{44}\,(K_x^2 + K_y^2) = C_{44}K^2 \ , \tag{65}$$

independent of propagation direction in the plane.

Consider other waves that propagate in the xy plane with particle motion in the xy plane: let

$$u = u_0 \exp\,[i(K_x x + K_y y - \omega t)] \ ; \qquad v = v_0 \exp\,[i(K_x x + K_y y - \omega t)] \ . \tag{66}$$

From (57a) and (57b),

$$\begin{aligned} \omega^2\rho u &= (C_{11}K_x^2 + C_{44}K_y^2)u + (C_{12} + C_{44})K_xK_yv \ ; \\ \omega^2\rho v &= (C_{11}K_y^2 + C_{44}K_x^2)v + (C_{12} + C_{44})K_xK_yu \ . \end{aligned} \tag{67}$$

This pair of equations has a particularly simple solution for a wave in the [110] direction, for which $K_x = K_y = K/\sqrt{2}$. The condition for a solution is that the determinant of the coefficients of u and v in (67) should equal zero:

$$\begin{vmatrix} -\omega^2\rho + \tfrac{1}{2}(C_{11} + C_{44})K^2 & \tfrac{1}{2}(C_{12} + C_{44})K^2 \\ \tfrac{1}{2}(C_{12} + C_{44})K^2 & -\omega^2\rho + \tfrac{1}{2}(C_{11} + C_{44})K^2 \end{vmatrix} = 0 \ . \qquad (68)$$

This equation has the roots

$$\omega^2\rho = \tfrac{1}{2}(C_{11} + C_{12} + 2C_{44})K^2 \ ; \qquad \omega^2\rho = \tfrac{1}{2}(C_{11} - C_{12})K^2 \ . \qquad (69)$$

The first root describes a longitudinal wave; the second root describes a shear wave. How do we determine the direction of particle displacement? The first root when substituted into the upper equation of (67) gives

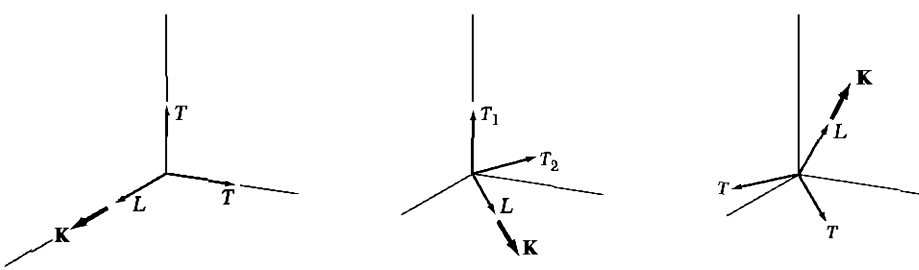
$$\tfrac{1}{2}(C_{11} + C_{12} + 2C_{44})K^2u = \tfrac{1}{2}(C_{11} + C_{44})K^2u + \tfrac{1}{2}(C_{12} + C_{44})K^2v \ , \qquad (70)$$

whence the displacement components satisfy $u = v$. Thus the particle displacement is along [110] and parallel to the **K** vector (Fig. 20). The second root of (44) when substituted into the upper equation of (67) gives

$$\tfrac{1}{2}(C_{11} - C_{12})K^2u = \tfrac{1}{2}(C_{11} + C_{44})K^2u + \tfrac{1}{2}(C_{12} + C_{44})K^2v \ , \qquad (71)$$

whence $u = -v$. The particle displacement is along $[1\bar{1}0]$ and perpendicular to the **K** vector.

Selected values of the adiabatic elastic stiffness constants of cubic crystals at low temperatures and at room temperature are given in Table 11. Notice the general tendency for the elastic constants to decrease as the temperature is increased. Further values at room temperature alone are given in Table 12.



Wave in [100] direction
$L : C_{11}$
$T : C_{44}$

Wave in [110] direction
$L : \frac{1}{2}(C_{11} + C_{12} + 2C_{44})$
$T_1 : C_{44}$
$T_2 : \frac{1}{2}(C_{11} - C_{12})$

Wave in [111] direction
$L : \frac{1}{3}(C_{11} + 2C_{12} + 4C_{44})$
$T : \frac{1}{3}(C_{11} - C_{12} + C_{44})$

Figure 20 Effective elastic constants for the three modes of elastic waves in the principal propagation directions in cubic crystals. The two transverse modes are degenerate for propagation in the [100] and [111] directions.

Table 11 Adiabatic elastic stiffness constants of cubic crystals at low temperature and at room temperature

The values given at 0 K were obtained by extrapolation of measurements carried out down to 4 K. The table was compiled with the assistance of Professor Charles S. Smith.

Crystal	Stiffness constants, in 10^{12} dyne/cm^2 (10^{11}N/m^2)			Temperature, K	Density, g/cm^3
	C_{11}	C_{12}	C_{44}		
W	5.326	2.049	1.631	0	19.317
	5.233	2.045	1.607	300	—
Ta	2.663	1.582	0.874	0	16.696
	2.609	1.574	0.818	300	—
Cu	1.762	1.249	0.818	0	9.018
	1.684	1.214	0.754	300	—
Ag	1.315	0.973	0.511	0	10.635
	1.240	0.937	0.461	300	—
Au	2.016	1.697	0.454	0	19.488
	1.923	1.631	0.420	300	—
Al	1.143	0.619	0.316	0	2.733
	1.068	0.607	0.282	300	—
K	0.0416	0.0341	0.0286	4	
	0.0370	0.0314	0.0188	295	
Pb	0.555	0.454	0.194	0	11.599
	0.495	0.423	0.149	300	—
Ni	2.612	1.508	1.317	0	8.968
	2.508	1.500	1.235	300	—
Pd	2.341	1.761	0.712	0	12.132
	2.271	1.761	0.717	300	—

Table 12 Adiabatic elastic stiffness constants of several cubic crystals at room temperature or 300 K

	Stiffness constants, in 10^{12} dyne/cm^2 or 10^{11} N/m^2		
	C_{11}	C_{12}	C_{44}
Diamond	10.76	1.25	5.76
Na	0.073	0.062	0.042
Li	0.135	0.114	0.088
Ge	1.285	0.483	0.680
Si	1.66	0.639	0.796
GaSb	0.885	0.404	0.433
InSb	0.672	0.367	0.302
MgO	2.86	0.87	1.48
NaCl	0.487	0.124	0.126

There are three normal modes of wave motion in a crystal for a given magnitude and direction of the wavevector **K**. In general, the polarizations (directions of particle displacement) of these modes are not exactly parallel or perpendicular to **K**. In the special propagation directions [100], [111], and [110] of a cubic crystal two of the three modes for a given **K** are such that the particle motion is exactly transverse to **K** and in the third mode the motion is exactly longitudinal (parallel to **K**). The analysis is much simpler in these special directions than in general directions.

SUMMARY

- Crystals of inert gas atoms are bound by the van der Waals interaction (induced dipole-dipole interaction), and this varies with distance as $1/R^6$.
- The repulsive interaction between atoms arises generally from the electrostatic repulsion of overlapping charge distributions and the Pauli principle, which compels overlapping electrons of parallel spin to enter orbitals of higher energy.
- Ionic crystals are bound by the electrostatic attraction of charged ions of opposite sign. The electrostatic energy of a structure of $2N$ ions of charge $\pm q$ is

(CGS) $$U = -N\alpha\frac{q^2}{R} = -N\sum\frac{(\pm)q^2}{r_{ij}} ,$$

where α is the Madelung constant and R is the distance between nearest neighbors.

- Metals are bound by the reduction in the kinetic energy of the valence electrons in the metal as compared with the free atom.
- A covalent bond is characterized by the overlap of charge distributions of antiparallel electron spin. The Pauli contribution to the repulsion is reduced for antiparallel spins, and this makes possible a greater degree of overlap. The overlapping electrons bind their associated ion cores by electrostatic attraction.

Problems

1. ***Quantum solid.*** In a quantum solid the dominant repulsive energy is the zero-point energy of the atoms. Consider a crude one-dimensional model of crystalline He^4 with each He atom confined to a line segment of length L. In the ground state the wave function within each segment is taken as a half wavelength of a free particle. Find the zero-point kinetic energy per particle.

2. ***Cohesive energy of bcc and fcc neon.*** Using the Lennard-Jones potential, calculate the ratio of the cohesive energies of neon in the bcc and fcc structures (Ans. 0.958). The lattice sums for the bcc structures are

$$\sum_j{}' p_{ij}^{-12} = 9.11418 \; ; \qquad \sum_j{}' p_{ij}^{-6} = 12.2533 \; .$$

3. ***Solid molecular hydrogen.*** For H_2 one finds from measurements on the gas that the Lennard-Jones parameters are $\epsilon = 50 \times 10^{-16}$ erg and $\sigma = 2.96$ Å. Find the cohesive energy in kJ per mole of H_2; do the calculation for an fcc structure. Treat each H_2 molecule as a sphere. The observed value of the cohesive energy is 0.751 kJ/mol, much less than we calculated; thus, quantum corrections must be very important.

4. ***Possibility of ionic crystals R^+R^-.*** Imagine a crystal that exploits for binding the coulomb attraction of the positive and negative ions of the same atom or molecule R. This is believed to occur with certain organic molecules, but it is not found when **R** is a single atom. Use the data in Tables 5 and 6 to evaluate the stability of such a form of Na in the NaCl structure relative to normal metallic sodium. Evaluate the energy at the observed interatomic distance in metallic sodium, and use 0.78 eV as the electron affinity of Na.

5. ***Linear ionic crystal.*** Consider a line of $2N$ ions of alternating charge $\pm q$ with a repulsive potential energy A/R^n between nearest neighbors. (a) Show that at the equilibrium separation

(CGS) $$U(R_0) = -\frac{2Nq^2 \ln 2}{R_0}\left(1 - \frac{1}{n}\right) .$$

(b) Let the crystal be compressed so that $R_0 \to R_0(1 - \delta)$. Show that the work done in compressing a unit length of the crystal has the leading term $\frac{1}{2}C\delta^2$, where

(CGS) $$C = \frac{(n-1)q^2 \ln 2}{R_0} .$$

To obtain the results in SI, replace q^2 by $q^2/4\pi\epsilon_0$. Note: We should not expect to obtain this result from the expression for $U(R_0)$, but we must use the complete expression for $U(R)$.

6. ***Cubic ZnS structure.*** Using λ and ρ from Table 7 and the Madelung constants given in the text, calculate the cohesive energy of KCl in the cubic ZnS structure described in Chapter 1. Compare with the value calculated for KCl in the NaCl structure.

7. ***Divalent ionic crystals.*** Barium oxide has the NaCl structure. Estimate the cohesive energies per molecule of the hypothetical crystals Ba^+O^- and $Ba^{++}O^{--}$ referred to separated neutral atoms. The observed nearest-neighbor internuclear distance is $R_0 = 2.76$ Å; the first and second ionization potentials of Ba are 5.19 and 9.96 eV; and the electron affinities of the first and second electrons added to the neutral oxygen atom are 1.5 and −9.0 eV. The first electron affinity of the

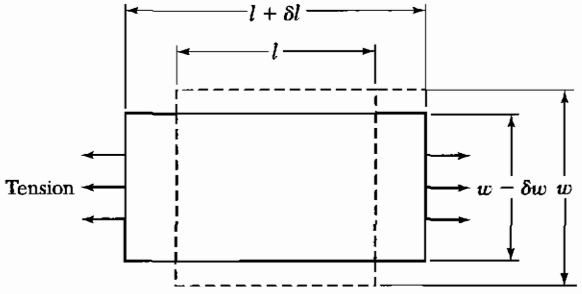


Figure 21 Young's modulus is defined as stress/strain for a tensile stress acting in one direction, with the specimen sides left free. Poisson's ratio is defined as $(\delta w/w)/(\delta l/l)$ for this situation.

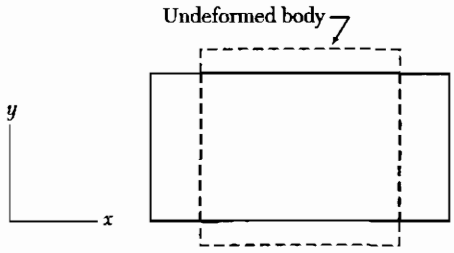


Figure 22 This deformation is compounded from the two shears $e_{xx} = -e_{yy}$.

neutral oxygen atom is the energy released in the reaction $O + e \rightarrow O^-$. The second electron affinity is the energy released in the reaction $O^- + e \rightarrow O^{--}$. Which valence state do you predict will occur? Assume R_0 is the same for both forms, and neglect the repulsive energy.

8. ***Young's modulus and Poisson's ratio.*** A cubic crystal is subject to tension in the [100] direction. Find expressions in terms of the elastic stiffnesses for Young's modulus and Poisson's ratio as defined in Fig. 21.

9. ***Longitudinal wave velocity.*** Show that the velocity of a longitudinal wave in the [111] direction of a cubic crystal is given by $v_x = [\frac{1}{3}(C_{11} + 2C_{12} + 4C_{44})/\rho]^{1/2}$. *Hint:* For such a wave $u = v = w$. Let $u = u_0 e^{iK(x+y+z)/\sqrt{3}} e^{-i\omega t}$, and use Eq. (57a).

10. ***Transverse wave velocity.*** Show that the velocity of transverse waves in the [111] direction of a cubic crystal is given by $v_s = [\frac{1}{3}(C_{11} - C_{12} + C_{44}/\rho]^{1/2}$. *Hint:* See Problem 9.

11. ***Effective shear constant.*** Show that the shear constant $\frac{1}{2}(C_{11} - C_{12})$ in a cubic crystal is defined by setting $e_{xx} = -e_{yy} = \frac{1}{2}e$ and all other strains equal to zero, as in Fig. 22. *Hint:* Consider the energy density (43); look for a C' such that $U = \frac{1}{2}C'e^2$.

12. ***Determinantal approach.*** It is known that an R-dimensional square matrix with all elements equal to unity has roots R and 0, with the R occurring once and the zero occurring $R - 1$ times. If all elements have the value p, then the roots are Rp and 0. (a) Show that if the diagonal elements are q and all other elements are p, then there is one root equal to $(R - 1)p + q$ and $R - 1$ roots equal to $q - p$. (b) Show from the elastic equation (57) for a wave in the [111] direction of a cubic crystal that the determinantal equation which gives ω^2 as a function of K is

$$\begin{vmatrix} q - \omega^2\rho & p & p \\ p & q - \omega^2\rho & p \\ p & p & q - \omega^2\rho \end{vmatrix} = 0 ,$$

where $q = \frac{1}{3}K^2(C_{11} + 2C_{44})$ and $p = \frac{1}{3}K^2(C_{12} + C_{44})$. This expresses the condition that three linear homogeneous algebraic equations for the three displacement

components u, v, w have a solution. Use the result of part (a) to find the three roots of ω^2; check with the results given for Problems 9 and 10.

13. ***General propagation direction.*** (a) By substitution in (57) find the determinantal equation which expresses the condition that the displacement

$$\mathbf{R}(\mathbf{r}) = [u_0\hat{\mathbf{x}} + v_0\hat{\mathbf{y}} + w_0\hat{\mathbf{z}}] \exp [i(\mathbf{K} \cdot \mathbf{r} - \omega t)]$$

be a solution of the elastic wave equations in a cubic crystal. (b) The sum of the roots of a determinantal equation is equal to the sum of the diagonal elements a_{ii}. Show from part (a) that the sum of the squares of the three elastic wave velocities in any direction in a cubic crystal is equal to $(C_{11} + 2C_{44})/\rho$. Recall that $v_s^2 = \omega^2/K^2$.

14. ***Stability criteria.*** The criterion that a cubic crystal with one atom in the primitive cell be stable against small homogeneous deformations is that the energy density (43) be positive for all combinations of strain components. What restrictions are thereby imposed on the elastic stiffness constants? (In mathematical language the problem is to find the conditions that a real symmetric quadratic form should be positive definite. The solution is given in books on algebra; see also Korn and Korn, *Mathematical Handbook*, McGraw-Hill, 1961, Sec. 13.5–6.) *Ans.* $C_{44} > 0$, $C_{11} > 0$, $C_{11}^2 - C_{12}^2 > 0$, and $C_{11} + 2C_{12} > 0$. For an example of the instability which results when $C_{11} \cong C_{12}$, see L. R. Testardi et al., Phys. Rev. Letters **15**, 250 (1965).

4

Phonons I. Crystal Vibrations

Chapter 5 treats the thermal properties of phonons.

	Name	Field
⟶	Electron	—
∿∿∿⟶	Photon	Electromagnetic wave
⟶	Phonon	Elastic wave
⟶	Plasmon	Collective electron wave
⟶	Magnon	Magnetization wave
−	Polaron	Electron + elastic deformation
−	Exciton	Polarization wave

Figure 1 Important elementary excitations in solids.

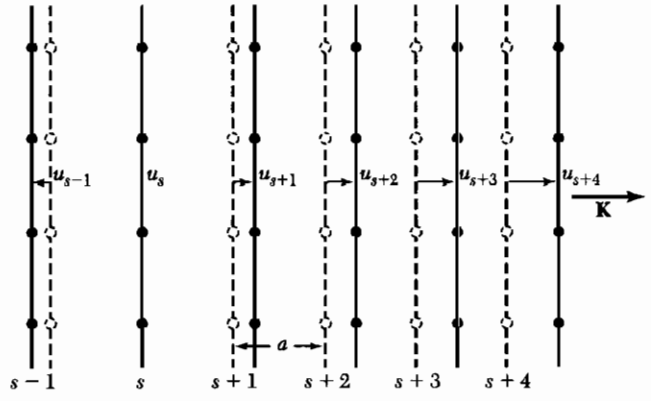


Figure 2 (Dashed lines) Planes of atoms when in equilibrium. (Solid lines) Planes of atoms when displaced as for a longitudinal wave. The coordinate u measures the displacement of the planes.

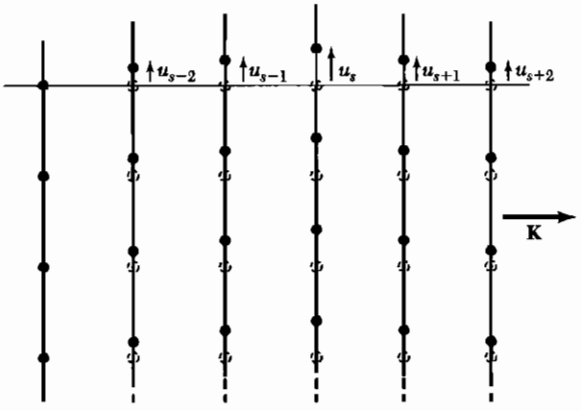


Figure 3 Planes of atoms as displaced during passage of a transverse wave.

VIBRATIONS OF CRYSTALS WITH MONATOMIC BASIS

Consider the elastic vibrations of a crystal with one atom in the primitive cell. We want to find the frequency of an elastic wave in terms of the wavevector that describes the wave and in terms of the elastic constants.

The mathematical solution is simplest in the [100], [110], and [111] propagation directions in cubic crystals. These are the directions of the cube edge, face diagonal, and body diagonal. When a wave propagates along one of these directions, entire planes of atoms move in phase with displacements either parallel or perpendicular to the direction of the wavevector. We can describe with a single coordinate u_s the displacement of the plane s from its equilibrium position. The problem is now one dimensional. For each wavevector there are three modes as solutions for u_s, one of longitudinal polarization (Fig. 2) and two of transverse polarization (Fig. 3).

We assume that the elastic response of the crystal is a linear function of the forces. That is equivalent to the assumption that the elastic energy is a quadratic function of the relative displacement of any two points in the crystal. Terms in the energy that are linear in the displacements will vanish in equilibrium—see the minimum in Fig. 3.6. Cubic and higher-order terms may be neglected for sufficiently small elastic deformations.

We assume that the force on the plane s caused by the displacement of the plane $s + p$ is proportional to the difference $u_{s+p} - u_s$ of their displacements. For brevity we consider only nearest-neighbor interactions, with $p = \pm 1$. The total force on s from planes $s \pm 1$:

$$F_s = C(u_{s+1} - u_s) + C(u_{s-1} - u_s) \ . \tag{1}$$

This expression is linear in the displacements and is of the form of Hooke's law.

The constant C is the **force constant** between nearest-neighbor planes and will differ for longitudinal and transverse waves. It is convenient hereafter to regard C as defined for one atom of the plane, so that F_s is the force on one atom in the plane s.

The equation of motion of an atom in the plane s is

$$M\frac{d^2u_s}{dt^2} = C(u_{s+1} + u_{s-1} - 2u_s) \ , \tag{2}$$

where M is the mass of an atom. We look for solutions with all displacements having the time dependence $\exp(-i\omega t)$. Then $d^2u_s/dt^2 = -\omega^2 u_s$, and (2) becomes

$$-M\omega^2 u_s = C(u_{s+1} + u_{s-1} - 2u_s) \ . \tag{3}$$

This is a difference equation in the displacements u and has traveling wave solutions of the form:

$$u_{s\pm1} = u \exp(isKa) \exp(\pm iKa) \ , \tag{4}$$

where a is the spacing between planes and K is the wavevector. The value to use for a will depend on the direction of K.

With (4), we have from (3):

$$-\omega^2 Mu \exp(isKa) = Cu\{\exp[i(s+1)Ka] + \exp[i(s-1)Ka] - 2\exp(isKa)\} \ . \tag{5}$$

We cancel $u \exp(isKa)$ from both sides, to leave

$$\omega^2 M = -C[\exp(iKa) + \exp(-iKa) - 2] \ . \tag{6}$$

With the identity $2\cos Ka = \exp(iKa) + \exp(-iKa)$, we have the **dispersion relation** $\omega(K)$.

$$\omega^2 = (2C/M)(1 - \cos Ka) \ . \tag{7}$$

The boundary of the first Brillouin zone lies at $K = \pm\pi/a$. We show from (7) that the slope of ω versus K is zero at the zone boundary:

$$d\omega^2/dK = (2Ca/M) \sin Ka = 0 \tag{8}$$

at $K = \pm\pi/a$, for here $\sin Ka = \sin(\pm\pi) = 0$. The special significance of phonon wavevectors that lie on the zone boundary is developed in (12) below.

By a trigonometric identity, (7) may be written as

$$\omega^2 = (4C/M) \sin^2 \tfrac{1}{2}Ka \ ; \qquad \omega = (4C/M)^{1/2} |\sin \tfrac{1}{2}Ka| \ . \tag{9}$$

A plot of ω versus K is given in Fig. 4.

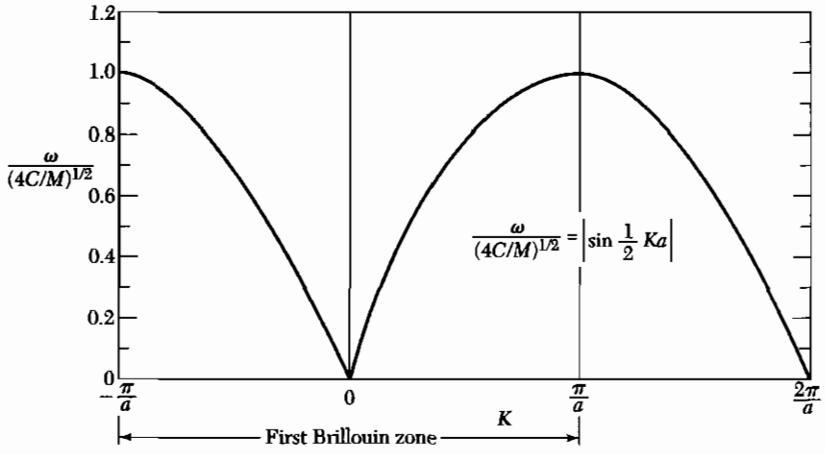


Figure 4 Plot of ω versus K. The region of $K \ll 1/a$ or $\lambda \gg a$ corresponds to the continuum approximation; here ω is directly proportional to K.

First Brillouin Zone

What range of K is physically significant for elastic waves? Only those in the first Brillouin zone. From (4) the ratio of the displacements of two successive planes is given by

$$\frac{u_{s+1}}{u_s} = \frac{u \exp[i(s+1)Ka]}{u \exp(isKa)} = \exp(iKa) \ . \tag{10}$$

The range $-\pi$ to $+\pi$ for the phase Ka covers all independent values of the exponential.

The range of independent values of K is specified by

$$-\pi < Ka \leq \pi \ , \qquad \text{or} \qquad -\frac{\pi}{a} < K \leq \frac{\pi}{a} \ .$$

This range is the first Brillouin zone of the linear lattice, as defined in Chapter 2. The extreme values are $K_{max} = \pm\pi/a$. Values of K outside of the first Brillouin zone (Fig. 5) merely reproduce lattice motions described by values within the limits $\pm\pi/a$.

We may treat a value of K outside these limits by subtracting the integral multiple of $2\pi/a$ that will give a wavevector inside these limits. Suppose K lies outside the first zone, but a related wavevector K' defined $K' = K - 2\pi n/a$ lies within the first zone, where n is an integer. Then the displacement ratio (10) becomes

$$u_{s+1}/u_s = \exp(iKa) \equiv \exp(i2\pi n) \exp[i(Ka - 2\pi n)] \equiv \exp(iK'a) \ , \tag{11}$$

because $\exp(i2\pi n) = 1$. Thus the displacement can always be described by a wavevector within the first zone. We note that $2\pi n/a$ is a reciprocal lattice vector because $2\pi/a$ is a reciprocal lattice vector. Thus by subtraction of an appropriate reciprocal lattice vector from K, we always obtain an equivalent wavevector in the first zone.

At the boundaries $K_{max} = \pm\pi/a$ of the Brillouin zone the solution $u_s = u \exp(isKa)$ does not represent a traveling wave, but a standing wave. At the zone boundaries $sK_{max}a = \pm s\pi$, whence

$$u_s = u \exp(\pm is\pi) = u(-1)^s \ . \tag{12}$$

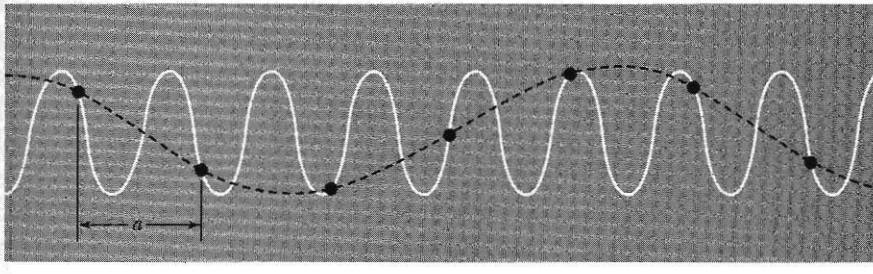


Figure 5 The wave represented by the solid curve conveys no information not given by the dashed curve. Only wavelengths longer than $2a$ are needed to represent the motion.

This is a standing wave: alternate atoms oscillate in opposite phases, because $u_s = \pm 1$ according to whether s is an even or an odd integer. The wave moves neither to the right nor to the left.

This situation is equivalent to Bragg reflection of x-rays: when the Bragg condition is satisfied a traveling wave cannot propagate in a lattice, but through successive reflections back and forth, a standing wave is set up.

The critical value $K_{max} = \pm\pi/a$ found here satisfies the Bragg condition $2d \sin\theta = n\lambda$: we have $\theta = \frac{1}{2}\pi$, $d = a$, $K = 2\pi/\lambda$, $n = 1$, so that $\lambda = 2a$. With x-rays it is possible to have n equal to other integers besides unity because the amplitude of the electromagnetic wave has a meaning in the space between atoms, but the displacement amplitude of an elastic wave usually has a meaning only at the atoms themselves.

Group Velocity

The transmission velocity of a wave packet is the **group velocity**, given as

$$v_g = d\omega/dK \ ,$$

or

$$\mathbf{v}_g = \text{grad}_{\mathbf{K}}\, \omega(\mathbf{K}) \ , \tag{13}$$

the gradient of the frequency with respect to $\mathbf{K}$. This is the velocity of energy propagation in the medium.

With the particular dispersion relation (9), the group velocity (Fig. 6) is

$$v_g = (Ca^2/M)^{1/2} \cos\tfrac{1}{2}Ka \ . \tag{14}$$

This is zero at the edge of the zone where $K = \pi/a$. Here the wave is a standing wave, as in (12), and we expect zero net transmission velocity for a standing wave.

Long Wavelength Limit

When $Ka \ll 1$ we expand $\cos Ka \equiv 1 - \frac{1}{2}(Ka)^2$, so that the dispersion relation (7) becomes

$$\omega^2 = (C/M)K^2a^2 \ . \tag{15}$$

The result that the frequency is directly proportional to the wavevector in the long wavelength limit is equivalent to the statement that the velocity of sound is independent of frequency in this limit. Thus $v = \omega/K$, exactly as in the continuum theory of elastic waves—in the continuum limit $Ka \ll 1$.

Derivation of Force Constants from Experiment

In metals the effective forces may be of quite long range and are carried from ion to ion through the conduction electron sea. Interactions have been found between planes of atoms separated by as many as 20 planes. We can make a statement about the range of the forces from the observed experimental

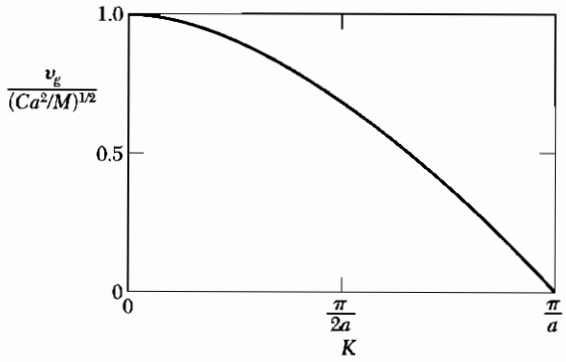


Figure 6 Group velocity v_g versus K, for model of Fig. 4. At the zone boundary $K = \pi/a$ the group velocity is zero.

dispersion relation for ω. The generalization of the dispersion relation (7) to p nearest planes is easily found to be

$$\omega^2 = (2/M) \sum_{p>0} C_p(1 - \cos pKa) \ . \tag{16a}$$

We solve for the interplanar force constants C_p by multiplying both sides by $\cos rKa$, where r is an integer, and integrating over the range of independent values of K:

$$\begin{aligned} M\int_{-\pi/a}^{\pi/a} dK\, \omega_K^2 \cos rKa &= 2\sum_{p>0} C_p \int_{-\pi/a}^{\pi/a} dK\, (1 - \cos pKa) \cos rKa \\ &= -2\pi C_r/a \ . \end{aligned} \tag{16b}$$

The integral vanishes except for $p = r$. Thus

$$C_p = -\frac{Ma}{2\pi}\int_{-\pi/a}^{\pi/a} dK\, \omega_K^2 \cos pKa \tag{17}$$

gives the force constant at range pa, for a structure with a monatomic basis.

TWO ATOMS PER PRIMITIVE BASIS

The phonon dispersion relation shows new features in crystals with two or more atoms per primitive basis. Consider, for example, the NaCl or diamond structures, with two atoms in the primitive cell. For each polarization mode in a given propagation direction the dispersion relation ω versus K develops two branches, known as the acoustical and optical branches, as in Fig. 7. We have longitudinal LA and transverse acoustical TA modes, and longitudinal LO and transverse optical TO modes.

If there are p atoms in the primitive cell, there are $3p$ branches to the dispersion relation: 3 acoustical branches and $3p - 3$ optical branches. Thus germanium (Fig. 8a) and KBr (Fig. 8b), each with two atoms in a primitive cell, have six branches: one LA, one LO, two TA, and two TO.

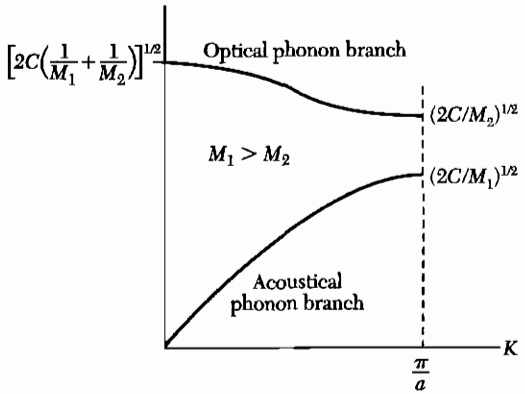


Figure 7 Optical and acoustical branches of the dispersion relation for a diatomic linear lattice, showing the limiting frequencies at $K = 0$ and $K = K_{max} = \pi/a$. The lattice constant is a.

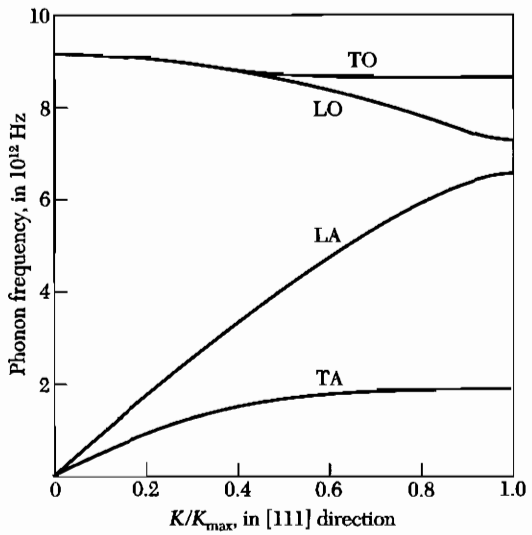


Figure 8a Phonon dispersion relations in the [111] direction in germanium at 80 K. The two TA phonon branches are horizontal at the zone boundary position, $K_{max} = (2\pi/a)(\frac{1}{2}\frac{1}{2}\frac{1}{2})$. The LO and TO branches coincide at $K = 0$; this also is a consequence of the crystal symmetry of Ge. The results were obtained with neutron inelastic scattering by G. Nilsson and G. Nelin.

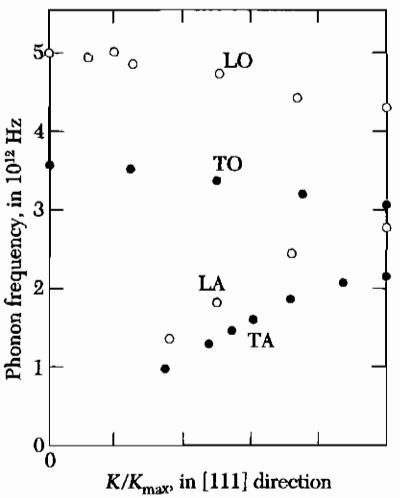


Figure 8b Dispersion curves in the [111] direction in KBr at 90 K, after A. D. B. Woods, B. N. Brockhouse, R. A. Cowley, and W. Cochran. The extrapolation to $K = 0$ of the TO, LO branches are called ω_T, ω_L.

The numerology of the branches follows from the number of degrees of freedom of the atoms. With p atoms in the primitive cell and N primitive cells, there are pN atoms. Each atom has three degrees of freedom, one for each of the x, y, z directions, making a total of $3pN$ degrees of freedom for the crystal. The number of allowed K values in a single branch is just N for one Brillouin zone.[1] Thus the

[1]We show in Chapter 5 by application of periodic boundary conditions to the modes of the crystal of volume V that there is one $\mathbf{K}$ value in the volume $(2\pi)^3/V$ in Fourier space. The volume of a Brillouin zone is $(2\pi)^3/V_c$, where V_c is the volume of a crystal primitive cell. Thus the number of allowed $\mathbf{K}$ values in a Brillouin zone is V/V_c, which is just N, the number of primitive cells in the crystal.

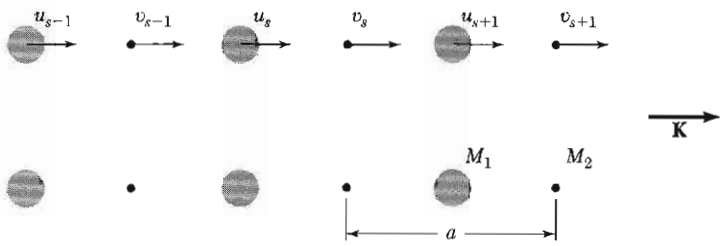


Figure 9 A diatomic crystal structure with masses M_1, M_2 connected by force constant C between adjacent planes. The displacements of atoms M_1 are denoted by u_{s-1}, u_s, u_{s+1}, . . . , and of atoms M_2 by v_{s-1}, v_s, v_{s+1}. The repeat distance is a in the direction of the wavevector K. The atoms are shown in their undisplaced positions.

LA and the two TA branches have a total of $3N$ modes, thereby accounting for $3N$ of the total degrees of freedom. The remaining $(3p - 3)N$ degrees of freedom are accommodated by the optical branches.

We consider a cubic crystal where atoms of mass M_1 lie on one set of planes and atoms of mass M_2 lie on planes interleaved between those of the first set (Fig. 9). It is not essential that the masses be different, but either the force constants or the masses will be different if the two atoms of the basis are in nonequivalent sites. Let a denote the repeat distance of the lattice in the direction normal to the lattice planes considered. We treat waves that propagate in a symmetry direction such that a single plane contains only a single type of ion; such directions are [111] in the NaCl structure and [100] in the CsCl structure.

We write the equations of motion under the assumption that each plane interacts only with its nearest-neighbor planes and that the force constants are identical between all pairs of nearest-neighbor planes. We refer to Fig. 9 to obtain

$$\begin{aligned} M_1 \frac{d^2u_s}{dt^2} &= C(v_s + v_{s-1} - 2u_s) \ ; \\ M_2 \frac{d^2v_s}{dt^2} &= C(u_{s+1} + u_s - 2v_s) \ . \end{aligned} \tag{18}$$

We look for a solution in the form of a traveling wave, now with different amplitudes u, v on alternate planes:

$$u_s = u \exp(isKa) \exp(-i\omega t) \ ; \qquad v_2 = v \exp(isKa) \exp(-i\omega t) \ . \tag{19}$$

We define a in Fig. 9 as the distance between nearest identical planes, not nearest-neighbor planes.

On substitution of (19) in (18) we have

$$\begin{aligned} -\omega^2 M_1 u &= Cv[1 + \exp(-iKa)] - 2Cu \ ; \\ -\omega^2 M_2 v &= Cu[\exp(iKa) + 1] - 2Cv \ . \end{aligned} \tag{20}$$

The homogeneous linear equations have a solution only if the determinant of the coefficients of the unknowns u, v vanishes:

$$\begin{vmatrix} 2C - M_1\omega^2 & -C[1 + \exp(iKa)] \\ -C[1 + \exp(iKa)] & 2C - M_2\omega^2 \end{vmatrix} = 0 \ , \tag{21}$$

or

$$M_1M_2\omega^4 - 2C(M_1 + M_2)\omega^2 + 2C^2(1 - \cos Ka) = 0 \ . \tag{22}$$

We can solve this equation exactly for ω^2, but it is simpler to examine the limiting cases $Ka \ll 1$ and $Ka = \pm\pi$ at the zone boundary. For small Ka we have $\cos Ka \cong 1 - \frac{1}{2}K^2a^2 + \ldots$, and the two roots are

$$\omega^2 \cong 2C\left(\frac{1}{M_1} + \frac{1}{M_2}\right) \qquad \text{(optical branch)} \ ; \tag{23}$$

$$\omega^2 \cong \frac{\frac{1}{2}C}{M_1 + M_2}K^2a^2 \qquad \text{(acoustical branch)} \ . \tag{24}$$

The extent of the first Brillouin zone is $-\pi/a \leq K \leq \pi/a$, where a is the repeat distance of the lattice. At $K_{max} = \pm\pi/a$ the roots are

$$\omega^2 = 2C/M_1 \ ; \qquad \omega^2 = 2C/M_2 \ . \tag{25}$$

The dependence of ω on K is shown in Fig. 7 for $M_1 > M_2$.

The particle displacements in the transverse acoustical (TA) and transverse optical (TO) branches are shown in Fig. 10. For the optical branch at $K = 0$ we find, on substitution of (23) in (20),

$$\frac{u}{v} = -\frac{M_2}{M_1} \ . \tag{26}$$

The atoms vibrate against each other, but their center of mass is fixed. If the two atoms carry opposite charges, as in Fig. 10, we may excite a motion of this

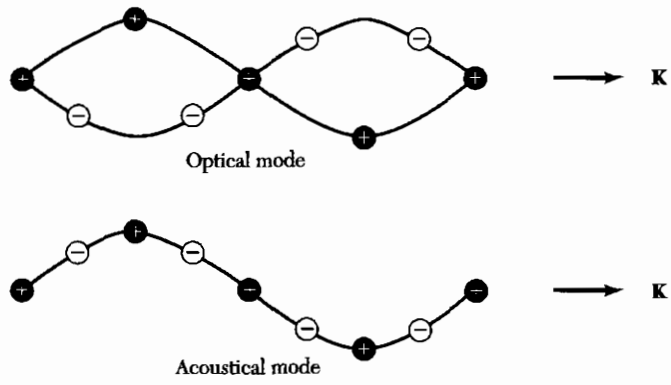


Figure 10 Transverse optical and transverse acoustical waves in a diatomic linear lattice, illustrated by the particle displacements for the two modes at the same wavelength.

type with the electric field of a light wave, so that the branch is called the optical branch. At a general K the ratio u/v will be complex, as follows from either of the equations (20). Another solution for the amplitude ratio at small K is $u = v$, obtained as the $K = 0$ limit of (24). The atoms (and their center of mass) move together, as in long wavelength acoustical vibrations, whence the term acoustical branch.

Wavelike solutions do not exist for certain frequencies, here between $(2C/M_1)^{1/2}$ and $(2C/M_2)^{1/2}$. This is a characteristic feature of elastic waves in polyatomic lattices. There is a frequency gap at the boundary $K_{max} = \pm\pi/a$ of the first Brillouin zone.

QUANTIZATON OF ELASTIC WAVES

The energy of a lattice vibration is quantized. The quantum of energy is called a **phonon** in analogy with the photon of the electromagnetic wave. The energy of an elastic mode of angular frequency ω is

$$\epsilon = (n + \tfrac{1}{2})\hbar\omega \tag{27}$$

when the mode is excited to quantum number n; that is, when the mode is occupied by n phonons. The term $\frac{1}{2}\hbar\omega$ is the zero point energy of the mode. It occurs for both phonons and photons as a consequence of their equivalence to a quantum harmonic oscillator of frequency ω, for which the energy eigenvalues are also $(n + \frac{1}{2})\hbar\omega$. The quantum theory of phonons is developed in Appendix C.

We can quantize the mean square phonon amplitude. Consider the standing wave mode of amplitude

$$u = u_0 \cos Kx \cos \omega t \ .$$

Here u is the displacement of a volume element from its equilibrium position at x in the crystal. The energy in the mode, as in any harmonic oscillator, is half kinetic energy and half potential energy, when averaged over time. The kinetic energy density is $\frac{1}{2}\rho(\partial u/\partial t)^2$, where ρ is the mass density. In a crystal of volume V, the volume integral of the kinetic energy is $\frac{1}{4}\rho V\omega^2 u_0^2 \sin^2 \omega t$. The time average kinetic energy is

$$\tfrac{1}{8}\rho V\omega^2 u_0^2 = \tfrac{1}{2}(n + \tfrac{1}{2})\hbar\omega \ , \tag{28}$$

because $<\sin^2 \omega t> = \frac{1}{2}$. The square of the amplitude of the mode is

$$u_0^2 = 4(n + \tfrac{1}{2})\hbar/\rho V\omega \ . \tag{29}$$

This relates the displacement in a given mode to the phonon occupancy n of the mode.

What is the sign of ω? The equations of motion such as (2) are equations for ω^2, and if this is positive then ω can have either sign, + or −. But the

energy of a phonon must be positive, so it is conventional and suitable to view ω as positive. If the crystal structure is unstable, then ω^2 will be negative and ω will be imaginary.

PHONON MOMENTUM

A phonon of wavevector K will interact with particles such as photons, neutrons, and electrons as if it had a momentum $\hbar K$. However, a phonon does not carry physical momentum.

The reason that phonons on a lattice do not carry momentum is that a phonon coordinate (except for $K = 0$) involves relative coordinates of the atoms. Thus in an H_2 molecule the internuclear vibrational coordinate $\mathbf{r}_1 - \mathbf{r}_2$ is a relative coordinate and does not carry linear momentum; the center of mass coordinate $\frac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2)$ corresponds to the uniform mode $K = 0$ and can carry linear momentum.

In crystals there exist wavevector selection rules for allowed transitions between quantum states. We saw in Chapter 2 that the elastic scattering of an x-ray photon by a crystal is governed by the wavevector selection rule

$$\mathbf{k}' = \mathbf{k} + \mathbf{G} \ , \tag{30}$$

where $\mathbf{G}$ is a vector in the reciprocal lattice, $\mathbf{k}$ is the wavevector of the incident photon, and $\mathbf{k}'$ is the wavevector of the scattered photon. In the reflection process the crystal as a whole will recoil with momentum $-\hbar\mathbf{G}$, but this uniform mode momentum is rarely considered explicitly.

Equation (30) is an example of the rule that the total wavevector of interacting waves is conserved in a periodic lattice, with the possible addition of a reciprocal lattice vector $\mathbf{G}$. The true momentum of the whole system always is rigorously conserved. If the scattering of the photon is inelastic, with the creation of a phonon of wavevector $\mathbf{K}$, then the wavevector selection rule becomes

$$\mathbf{k}' + \mathbf{K} = \mathbf{k} + \mathbf{G} \ . \tag{31}$$

If a phonon $\mathbf{K}$ is absorbed in the process, we have instead the relation

$$\mathbf{k}' = \mathbf{k} + \mathbf{K} + \mathbf{G} \ . \tag{32}$$

Relations (31) and (32) are the natural extensions of (30).

INELASTIC SCATTERING BY PHONONS

Phonon dispersion relations $\omega(\mathbf{K})$ are most often determined experimentally by the inelastic scattering of neutrons with the emission or absorption of a phonon. A neutron sees the crystal lattice chiefly by interaction with the nuclei

of the atoms. The kinematics of the scattering of a neutron beam by a crystal lattice are described by the general wavevector selection rule:

$$\mathbf{k} + \mathbf{G} = \mathbf{k}' \pm \mathbf{K} , \tag{33}$$

and by the requirement of conservation of energy. Here $\mathbf{K}$ is the wavevector of the phonon created (+) or absorbed (−) in the scattering process, and $\mathbf{G}$ is any reciprocal lattice vector. For a phonon we choose $\mathbf{G}$ such that $\mathbf{K}$ lies in the first Brillouin zone.

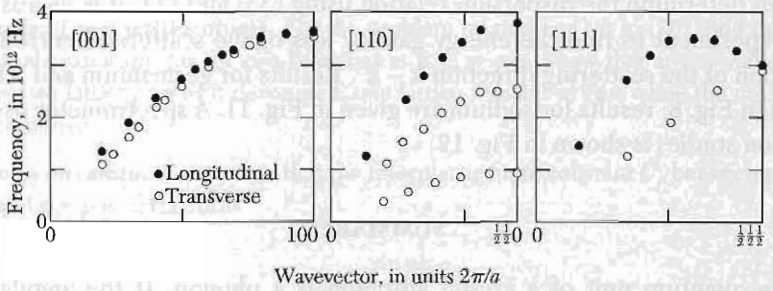


Figure 11 The dispersion curves of sodium for phonons propagating in the [001], [110], and [111] directions at 90 K, as determined by inelastic scattering of neutrons, by Woods, Brockhouse, March and Bowers.

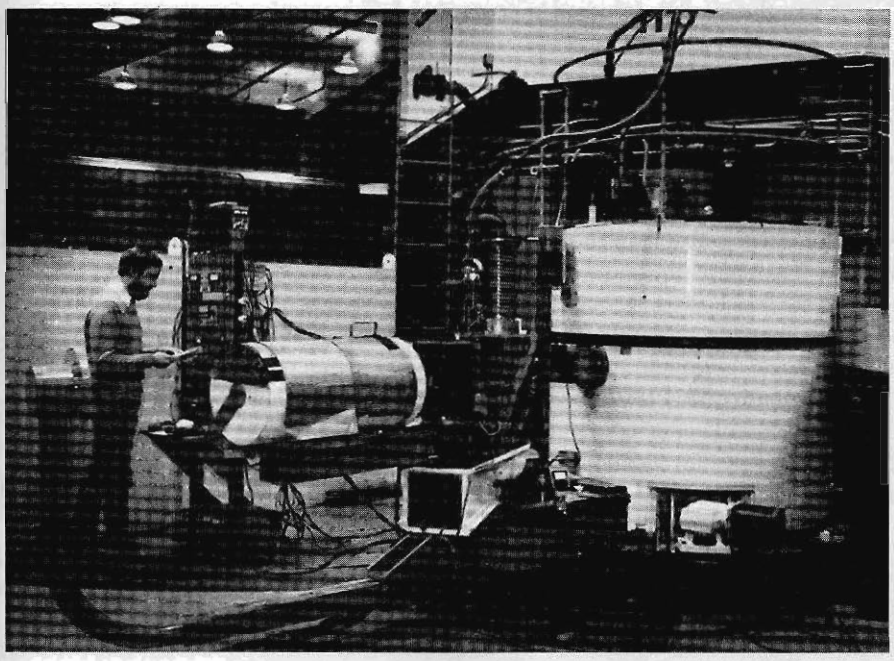


Figure 12 A triple axis neutron spectrometer at Brookhaven. (Courtesy of B. H. Grier.)

The kinetic energy of the incident neutron is $p^2/2M_n$, where M_n is the mass of the neutron. The momentum $\mathbf{p}$ is given by $\hbar\mathbf{k}$, where $\mathbf{k}$ is the wavevector of the neutron. Thus $\hbar^2k^2/2M_n$ is the kinetic energy of the incident neutron. If $\mathbf{k}'$ is the wavevector of the scattered neutron, the energy of the scattered neutron is $\hbar^2k'^2/2M_n$. The statement of conservation of energy is

$$\frac{\hbar^2k^2}{2M_n} = \frac{\hbar^2k'^2}{2M_n} \pm \hbar\omega \ , \tag{34}$$

where $\hbar\omega$ is the energy of the phonon created (+) or absorbed (−) in the process.

To determine the dispersion relation using (33) and (34) it is necessary in the experiment to find the energy gain or loss of the scattered neutrons as a function of the scattering direction $\mathbf{k} - \mathbf{k}'$. Results for germanium and KBr are given in Fig. 8; results for sodium are given in Fig. 11. A spectrometer used for phonon studies is shown in Fig. 12.

SUMMARY

- The quantum unit of a crystal vibration is a phonon. If the angular frequency is ω, the energy of the phonon is $\hbar\omega$.
- When a phonon of wavevector $\mathbf{K}$ is created by the inelastic scattering of a photon or neutron from wavevector $\mathbf{k}$ to $\mathbf{k}'$, the wavevector selection rule that governs the process is

$$\mathbf{k} = \mathbf{k}' + \mathbf{K} + \mathbf{G} \ ,$$

  where $\mathbf{G}$ is a reciprocal lattice vector.
- All elastic waves can be described by wavevectors that lie within the first Brillouin zone in reciprocal space.
- If there are p atoms in the primitive cell, the phonon dispersion relation will have 3 acoustical phonon branches and $3p - 3$ optical phonon branches.

Problems

1. ***Monatomic linear lattice.*** Consider a longitudinal wave

$$u_s = u\cos(\omega t - sKa)$$

which propagates in a monatomic linear lattice of atoms of mass M, spacing a, and nearest-neighbor interaction C.
(a) Show that the total energy of the wave is

$$E = \tfrac{1}{2}M\sum_s (du_s/dt)^2 + \tfrac{1}{2}C\sum_s (u_s - u_{s+1})^2 \ .$$

where s runs over all atoms.

(b) By substitution of u_s in this expression, show that the time-average total energy per atom is

$$\tfrac{1}{4}M\omega^2u^2 + \tfrac{1}{2}C(1 - \cos Ka)u^2 = \tfrac{1}{2}M\omega^2u^2 \ ,$$

where in the last step we have used the dispersion relation (9) for this problem.

2. ***Continuum wave equation.*** Show that for long wavelengths the equation of motion (2) reduces to the continuum elastic wave equation

$$\frac{\partial^2 u}{\partial t^2} = v^2 \frac{\partial^2 u}{\partial x^2} \ ,$$

where v is the velocity of sound.

3. ***Basis of two unlike atoms.*** For the problem treated by (18) to (26), find the amplitude ratios u/v for the two branches at $K_{max} = \pi/a$. Show that at this value of K the two lattices act as if decoupled: one lattice remains at rest while the other lattice moves.

4. ***Kohn anomaly.*** We suppose that the interplanar force constant C_p between planes s and $s + p$ is of the form

$$C_p = A \frac{\sin pk_0a}{pa} \ ,$$

where A and k_0 are constants and p runs over all integers. Such a form is expected in metals. Use this and Eq. (16a) to find an expression for ω^2 and also for $\partial\omega^2/\partial K$. Prove that $\partial\omega^2/\partial K$ is infinite when $K = k_0$. Thus a plot of ω^2 versus K or of ω versus K has a vertical tangent at k_0: there is a kink at k_0 in the phonon dispersion relation $\omega(K)$.

5. ***Diatomic chain.*** Consider the normal modes of a linear chain in which the force constants between nearest-neighbor atoms are alternately C and $10C$. Let the masses be equal, and let the nearest-neighbor separation be $a/2$. Find $\omega(K)$ at $K = 0$ and $K = \pi/a$. Sketch in the dispersion relation by eye. This problem simulates a crystal of diatomic molecules such as H_2.

6. ***Atomic vibrations in a metal.*** Consider point ions of mass M and charge e immersed in a uniform sea of conduction electrons. The ions are imagined to be in stable equilibrium when at regular lattice points. If one ion is displaced a small distance r from its equilibrium position, the restoring force is largely due to the electric charge within the sphere of radius r centered at the equilibrium position. Take the number density of ions (or of conduction electrons) as $3/4\pi R^3$, which defines R. (a) Show that the frequency of a single ion set into oscillation is $\omega = (e^2/MR^3)^{1/2}$. (b) Estimate the value of this frequency for sodium, roughly. (c) From (a), (b), and some common sense, estimate the order of magnitude of the velocity of sound in the metal.

*7. ***Soft phonon modes.*** Consider a line of ions of equal mass but alternating in charge, with $e_p = e(-1)^p$ as the charge on the pth ion. The interatomic potential is

*This problem is rather difficult.

the sum of two contributions: (1) a short-range interaction of force constant $C_{1R} = \gamma$ that acts between nearest neighbors only, and (2) a coulomb interaction between all ions. (a) Show that the contribution of the coulomb interaction to the atomic force constants is $C_{pC} = 2(-1)^p e^2/p^3a^3$, where a is the equilibrium nearest-neighbor distance. (b) From (16a) show that the dispersion relation may be written as

$$\omega^2/\omega_0^2 - \sin^2 \tfrac{1}{2}Ka + \sigma \sum_{p-1}^{\infty} (-1)^p (1 - \cos pKa) p^{-3} ,$$

where $\omega_0^2 \equiv 4\gamma/M$ and $\sigma = e^2/\gamma a^3$. (c) Show that ω^2 is negative (unstable mode) at the zone boundary $Ka = \pi$ if $\sigma > 0.475$ or $4/7\zeta(3)$, where ζ is a Riemann zeta function. Show further that the speed of sound at small Ka is imaginary if $\sigma > (2 \ln 2)^{-1} = 0.721$. Thus ω^2 goes to zero and the lattice is unstable for some value of Ka in the interval $(0, \pi)$ if $0.475 < \sigma < 0.721$. Notice that the phonon spectrum is not that of a diatomic lattice because the interaction of any ion with its neighbors is the same as that of any other ion.

5

Phonons II. Thermal Properties

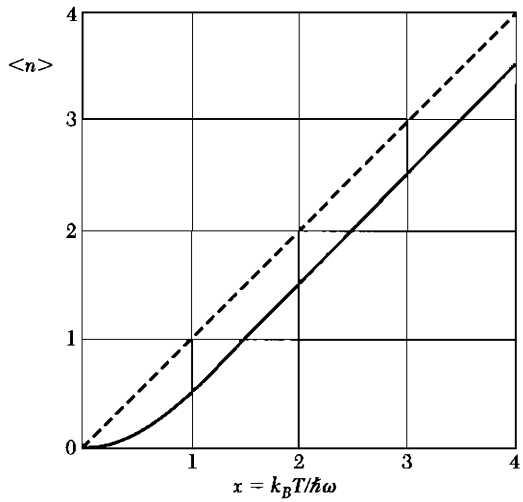


Figure 1 Plot of Planck distribution function. At high temperatures the occupancy of a state is approximately linear in the temperature. The function $\langle n \rangle + \frac{1}{2}$, which is not plotted, approaches the dashed line as asymptote at high temperatures.

We discuss the heat capacity of a phonon gas and then the effects of anharmonic lattice interactions on the phonons and on the crystal.

PHONON HEAT CAPACITY

By heat capacity we shall usually mean the heat capacity at constant volume, which is more fundamental than the heat capacity at constant pressure, which is what the experiments determine.[1] The heat capacity at constant volume is defined as $C_V \equiv (\partial U/\partial T)_V$ where U is the energy and T the temperature.

The contribution of the phonons to the heat capacity of a crystal is called the lattice heat capacity and is denoted by C_{lat}. The total energy of the phonons at a temperature $\tau(\equiv k_B T)$ in a crystal may be written as the sum of the energies over all phonon modes, here indexed by the wavevector K and polarization index p:

$$U_{lat} = \sum_K \sum_p U_{K,p} = \sum_K \sum_p \langle n_{K,p} \rangle \hbar\omega_{K,p} \ , \tag{1}$$

where $\langle n_{K,p} \rangle$ is the thermal equilibrium occupancy of phonons of wavevector K and polarization p. The form of $\langle n_{K,p} \rangle$ is given by the Planck distribution function:

$$\langle n \rangle = \frac{1}{\exp(\hbar\omega/\tau) - 1} \ , \tag{2}$$

where the $\langle \cdots \rangle$ denotes the average in thermal equilibrium. A graph of $\langle n \rangle$ is given in Fig. 1.

Planck Distribution

Consider a set of identical harmonic oscillators in thermal equilibrium. The ratio of the number of oscillators in their $(n + 1)$th quantum state of excitation to the number in the nth quantum state is

$$N_{n+1}/N_n = \exp(-\hbar\omega/\tau) \ , \qquad \tau \equiv k_B T \ , \tag{3}$$

[1] A thermodynamic relation gives $C_p - C_V = 9\alpha^2 BVT$, where α is the temperature coefficient of linear expansion, V the volume, and B the bulk modulus. The fractional difference between C_p and C_V is usually small in solids and often may be neglected. As $T \to 0$ we see that $C_p \to C_V$, provided α and B are constant.

by use of the Boltzmann factor. Thus the fraction of the total number of oscillators in the nth quantum state is

$$\frac{N_n}{\sum_{s=0}^{\infty} N_s} = \frac{\exp(-n\hbar\omega/\tau)}{\sum_{s=0}^{\infty} \exp(-s\hbar\omega/\tau)} . \tag{4}$$

We see that the average excitation quantum number of an oscillator is

$$\langle n \rangle = \frac{\sum_s s \exp(-s\hbar\omega/\tau)}{\sum_s \exp(-s\hbar\omega/\tau)} . \tag{5}$$

The summations in (5) are

$$\sum_s x^s = \frac{1}{1-x} ; \qquad \sum_s s x^s = x \frac{d}{dx} \sum_s x^s = \frac{x}{(1-x)^2} , \tag{6}$$

with $x = \exp(-\hbar\omega/\tau)$. Thus we may rewrite (5) as the Planck distribution:

$$\langle n \rangle = \frac{x}{1-x} = \frac{1}{\exp(\hbar\omega/\tau) - 1} . \tag{7}$$

Normal Mode Enumeration

The energy of a collection of oscillators of frequencies $\omega_{K,p}$ in thermal equilibrium is found from (1) and (2):

$$U = \sum_K \sum_p \frac{\hbar\omega_{K,p}}{\exp(\hbar\omega_{K,p}/\tau) - 1} . \tag{8}$$

It is usually convenient to replace the summation over K by an integral. Suppose that the crystal has $D_p(\omega)d\omega$ modes of a given polarization p in the frequency range ω to $\omega + d\omega$. Then the energy is

$$U = \sum_p \int d\omega \, D_p(\omega) \frac{\hbar\omega}{\exp(\hbar\omega/\tau) - 1} . \tag{9}$$

The lattice heat capacity is found by differentiation with respect to temperature. Let $x = \hbar\omega/\tau = \hbar\omega/k_B T$: then $\partial U/\partial T$ gives

$$C_{\text{lat}} = k_B \sum_p \int d\omega \, D_p(\omega) \frac{x^2 \exp x}{(\exp x - 1)^2} . \tag{10}$$

The central problem is to find $D(\omega)$, the number of modes per unit frequency range. This function is called the density of modes or, more often, density of states.

Density of States in One Dimension

Consider the boundary value problem for vibrations of a one-dimensional line (Fig. 2) of length L carrying $N + 1$ particles at separation a. We suppose

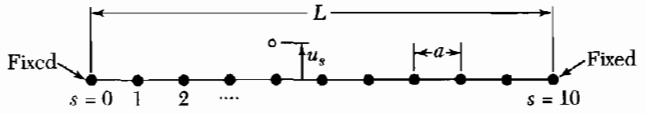


Figure 2 Elastic line of $N + 1$ atoms, with $N = 10$, for boundary conditions that the end atoms $s = 0$ and $s = 10$ are fixed. The particle displacements in the normal modes for either longitudinal or transverse displacements are of the form $u_s \propto \sin sKa$. This form is automatically zero at the atom at the end $s = 0$, and we choose K to make the displacement zero at the end $s = 10$.

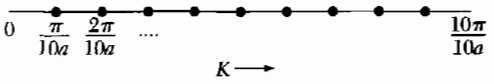


Figure 3 The boundary condition $\sin sKa = 0$ for $s = 10$ can be satisfied by choosing $K = \pi/10a$, $2\pi/10a, \ldots, 9\pi/10a$, where $10a$ is the length L of the line. The present figure is in K space. The dots are not atoms but are the allowed values of K. Of the $N + 1$ particles on the line, only $N - 1$ are allowed to move, and their most general motion can be expressed in terms of the $N - 1$ allowed values of K. *This quantization of K has nothing to do with quantum mechanics but follows* classically from the boundary conditions that the end atoms be fixed.

that the particles $s = 0$ and $s = N$ at the ends of the line are held fixed. Each normal vibrational mode of polarization p has the form of a standing wave, where u_s is the displacement of the particle s:

$$u_s = u(0) \exp(-i\omega_{Kp}t) \sin sKa \ , \tag{11}$$

where ω_{Kp} is related to K by the appropriate dispersion relation.

As in Fig. 3, the wavevector K is restricted by the fixed-end boundary conditions to the values

$$K = \frac{\pi}{L} \ , \quad \frac{2\pi}{L} \ , \quad \frac{3\pi}{L} \ , \ldots, \quad \frac{(N-1)\pi}{L} \ . \tag{12}$$

The solution for $K = \pi/L$ has

$$u_s \propto \sin (s\pi a/L) \tag{13}$$

and vanishes for $s = 0$ and $s = N$ as required.

The solution for $K = N\pi/L = \pi/a = K_{\max}$ has $u_s \propto \sin s\pi$; this permits no motion of any atom, because $\sin s\pi$ vanishes at each atom. Thus there are $N - 1$ allowed independent values of K in (12). This number is equal to the number of particles allowed to move. Each allowed value of K is associated with a standing wave. For the one-dimensional line there is one mode for each interval $\Delta K = \pi/L$, so that the number of modes per unit range of K is L/π for $K \leq \pi/a$, and 0 for $K > \pi/a$.

There are three polarizations p for each value of K: in one dimension two of these are transverse and one longitudinal. In three dimensions the polarizations are this simple only for wavevectors in certain special crystal directions.

Another device for enumerating modes is equally valid. We consider the medium as unbounded, but require that the solutions be periodic over a large

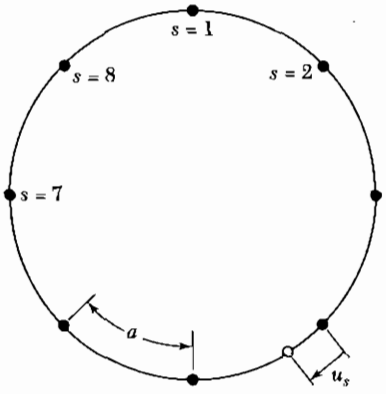


Figure 4 Consider N particles constrained to slide on a circular ring. The particles can oscillate if connected by elastic springs. In a normal mode the displacement u_s of atom s will be of the form sin sKa or cos sKa: these are independent modes. By the geometrical periodicity of the ring the boundary condition is that $u_{N+s} = u_s$ for all s, so that NKa must be an integral multiple of 2π. For $N = 8$ the allowed independent values of K are $0, 2\pi/8a, 4\pi/8a, 6\pi/8a$, and $8\pi/8a$. The value $K = 0$ is meaningless for the sine form, because sin $s0a = 0$. The value $8\pi/8a$ has a meaning only for the cosine form, because sin $(s8\pi a/8a) = \sin s\pi = 0$. The three other values of K are allowed for both the sine and cosine modes, giving a total of eight allowed modes for the eight particles. Thus the periodic boundary condition leads to one allowed mode per particle, exactly as for the fixed-end boundary condition of Fig. 3. If we had taken the modes in the complex form $\exp(isKa)$, the periodic boundary condition would lead to the eight modes with $K = 0, \pm 2\pi/Na, \pm 4\pi/Na, \pm 6\pi/Na$, and $8\pi/Na$. as in Eq. (14).

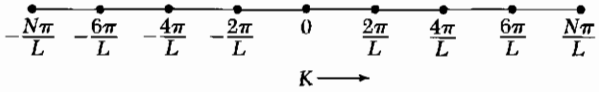


Figure 5 Allowed values of wavevector K for periodic boundary conditions applied to a linear lattice of periodicity $N = 8$ atoms on a line of length L. The $K = 0$ solution is the uniform mode. The special points $\pm N\pi/L$ represent only a single solution because $\exp(i\pi s)$ is identical to $\exp(-i\pi s)$; thus there are eight allowed modes, with displacements of the sth atom proportional to 1, $\exp(\pm i\pi s/4)$, $\exp(\pm i\pi s/2)$, $\exp(\pm i3\pi s/4)$, $\exp(i\pi s)$.

distance L, so that $u(sa) = u(sa + L)$. The method of **periodic boundary conditions** (Figs. 4 and 5) does not change the physics of the problem in any essential respect for a large system. In the running wave solution $u_s = u(0) \exp[i(sKa - \omega_K t)]$ the allowed values of K are

$$K = 0 \ , \quad \pm \frac{2\pi}{L} \ , \quad \pm \frac{4\pi}{L} \ , \quad \pm \frac{6\pi}{L} \ , \ \ldots, \ \frac{N\pi}{L} \ . \tag{14}$$

This method of enumeration gives the same number of modes (one per mobile atom) as given by (12), but we have now both plus and minus values of K, with the interval $\Delta K = 2\pi/L$ between successive values of K. For periodic boundary conditions, the number of modes per unit range of K is $L/2\pi$ for $-\pi/a \leq K \leq \pi/a$, and 0 otherwise. The situation in a two-dimensional lattice is portrayed in Fig. 6.

We need to know $D(\omega)$, the number of modes per unit frequency range for a given polarization. The number of modes $D(\omega)\, d\omega$ in $d\omega$ at ω is given in one dimension by

$$D_1(\omega)\, d\omega = \frac{L}{\pi} \frac{dK}{d\omega} d\omega = \frac{L}{\pi} \cdot \frac{d\omega}{d\omega/dK} \ . \tag{15}$$

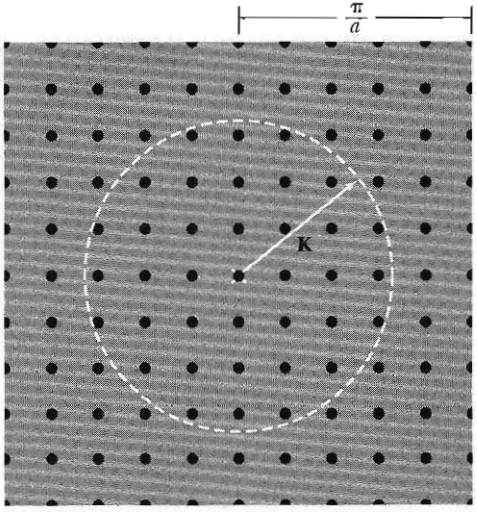


Figure 6 Allowed values in Fourier space of the phonon wavevector K for a square lattice of lattice constant a, with periodic boundary conditions applied over a square of side $L = 10a$. The uniform mode is marked with a cross. There is one allowed value of K per area $(2\pi/10a)^2 = (2\pi/L)^2$, so that within the circle of area πK^2 the smoothed number of allowed points is $\pi K^2(L/2\pi)^2$.

We can obtain the group velocity $d\omega/dK$ from the dispersion relation ω versus K. There is a singularity in $D_1(\omega)$whenever the dispersion relation $\omega(K)$ is horizontal; that is, whenever the group velocity is zero.

Density of States in Three Dimensions

We apply periodic boundary conditions over N^3 primitive cells within a cube of side L, so that $\mathbf{K}$ is determined by the condition

$$\exp[i(K_x x + K_y y + K_z z)] \equiv \exp\{i[K_x(x + L) + K_y(y + L) + K_z(z + L)]\} , \qquad (16)$$

whence

$$K_x, K_y, K_z = 0 \; ; \quad \pm\frac{2\pi}{L} \; ; \quad \pm\frac{4\pi}{L} \; ; \quad \cdots ; \quad \frac{N\pi}{L} . \qquad (17)$$

Therefore, there is one allowed value of $\mathbf{K}$ per volume $(2\pi/L)^3$ in $\mathbf{K}$ space, or

$$\left(\frac{L}{2\pi}\right)^3 = \frac{V}{8\pi^3} \qquad (18)$$

allowed values of $\mathbf{K}$ per unit volume of $\mathbf{K}$ space, for each polarization and for each branch. The volume of the specimen is $V = L^3$.

The total number of modes with wavevector less than K is found from (18) to be $(L/2\pi)^3$ times the volume of a sphere of radius K. Thus

$$N = (L/2\pi)^3(4\pi K^3/3) \qquad (19)$$

for each polarization type. The density of states for each polarization is

$$D(\omega) = dN/d\omega = (VK^2/2\pi^2)(dK/d\omega) \ . \tag{20}$$

Debye Model for Density of States

In the Debye approximation the velocity of sound is taken as constant for each polarization type, as it would be for a classical elastic continuum. The dispersion relation is written as

$$\omega = vK \ , \tag{21}$$

with v the constant velocity of sound.

The density of states (20) becomes

$$D(\omega) = V\omega^2/2\pi^2v^3 \ . \tag{22}$$

If there are N primitive cells in the specimen, the total number of acoustic phonon modes is N. A cutoff frequency ω_D is determined by (19) as

$$\omega_D^3 = 6\pi^2v^3N/V \ . \tag{23}$$

To this frequency there corresponds a cutoff wavevector in **K** space:

$$K_D = \omega_D/v = (6\pi^2N/V)^{1/3} \ . \tag{24}$$

On the Debye model we do not allow modes of wavevector larger than K_D. The number of modes with $K \leq K_D$ exhausts the number of degrees of freedom of a monatomic lattice.

The thermal energy (9) is given by

$$U = \int d\omega \, D(\omega)\langle n(\omega)\rangle\hbar\omega = \int_0^{\omega_D} d\omega \left(\frac{V\omega^2}{2\pi^2v^3}\right)\left(\frac{\hbar\omega}{e^{\hbar\omega/\tau} - 1}\right) \ , \tag{25}$$

for each polarization type. For brevity we assume that the phonon velocity is independent of the polarization, so that we multiply by the factor 3 to obtain

$$U = \frac{3V\hbar}{2\pi^2v^3}\int_0^{\omega_D} d\omega \, \frac{\omega^3}{e^{\hbar\omega/\tau} - 1} = \frac{3Vk_B^4T^4}{2\pi^2v^3\hbar^3}\int_0^{x_D} dx \, \frac{x^3}{e^x - 1} \ , \tag{26}$$

where $x \equiv \hbar\omega/\tau \equiv \hbar\omega/k_BT$ and

$$x_D \equiv \hbar\omega_D/k_BT \equiv \theta/T \ . \tag{27}$$

This defines the **Debye temperature** θ in terms of ω_D defined by (23). We may express θ as

$$\theta = \frac{\hbar v}{k_B} \cdot \left(\frac{6\pi^2N}{V}\right)^{1/3} \ , \tag{28}$$

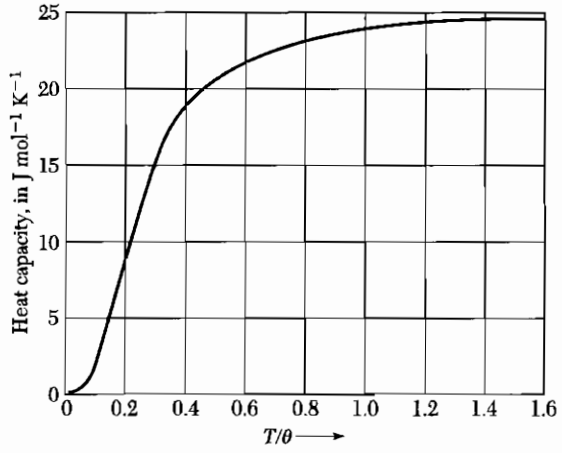


Figure 7 Heat capacity C_V of a solid, according to the Debye approximation. The vertical scale is in J mol^{-1} K^{-1}. The horizontal scale is the temperature normalized to the Debye temperature θ. The region of the T^3 law is below 0.1θ. The asymptotic value at high values of T/θ is 24.943 J mol^{-1} deg^{-1}.

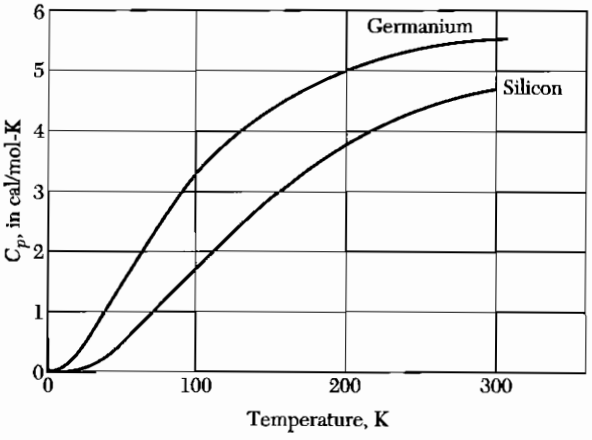


Figure 8 Heat capacity of silicon and germanium. Note the decrease at low temperatures. To convert a value in cal/mol-K to J/mol-K, multiply by 4.186.

so that the total phonon energy is

$$U = 9Nk_BT\left(\frac{T}{\theta}\right)^3 \int_0^{x_D} dx \frac{x^3}{e^x - 1} \quad , \tag{29}$$

where N is the number of atoms in the specimen and $x_D = \theta/T$.

The heat capacity is found most easily by differentiating the middle expression of (26) with respect to temperature. Then

$$C_V = \frac{3V\hbar^2}{2\pi^2 v^3 k_B T^2} \int_0^{\omega_D} d\omega \frac{\omega^4 e^{\hbar\omega/\tau}}{(e^{\hbar\omega/\tau} - 1)^2} = 9Nk_B\left(\frac{T}{\theta}\right)^3 \int_0^{x_D} dx \frac{x^4 e^x}{(e^x - 1)^2} \quad . \tag{30}$$

The Debye heat capacity is plotted in Fig. 7. At $T \gg \theta$ the heat capacity approaches the classical value of $3Nk_B$. Measured values for silicon and germanium are plotted in Fig. 8.

Debye T^3 Law

At very low temperatures we may approximate (29) by letting the upper limit go to infinity. We have

$$\int_0^\infty dx \frac{x^3}{e^x - 1} = \int_0^\infty dx\, x^3 \sum_{s=1}^{\infty} \exp(-sx) = 6 \sum_1^\infty \frac{1}{s^4} = \frac{\pi^4}{15} , \tag{31}$$

where the sum over s^{-4} is found in standard tables. Thus $U \cong 3\pi^4 N k_B T^4/5\theta^3$ for $T \ll \theta$, and

$$C_V \cong \frac{12\pi^4}{5} N k_B \left(\frac{T}{\theta}\right)^3 \cong 234\, N k_B \left(\frac{T}{\theta}\right)^3 , \tag{32}$$

which is the Debye T^3 approximation. Experimental results for argon are plotted in Fig. 9.

At sufficiently low temperature the T^3 approximation is quite good; that is, when only long wavelength acoustic modes are thermally excited. These are just the modes that may be treated as an elastic continuum with macroscopic elastic constants. The energy of the short wavelength modes (for which this approximation fails) is too high for them to be populated significantly at low temperatures.

We understand the T^3 result by a simple argument (Fig. 10). Only those lattice modes having $\hbar\omega < k_B T$ will be excited to any appreciable extent at a low temperature T. The excitation of these modes will be approximately classical, each with an energy close to $k_B T$, according to Fig. 1.

Of the allowed volume in **K** space, the fraction occupied by the excited modes is of the order of $(\omega_T/\omega_D)^3$ or $(K_T/K_D)^3$, where K_T is a "thermal" wavevector defined such that $\hbar v K_T = k_B T$ and K_D is the Debye cutoff wavevector. Thus the fraction occupied is $(T/\theta)^3$ of the total volume in **K** space. There are of the order of $3N(T/\theta)^3$excited modes, each having energy $k_B T$. The energy is $\sim 3Nk_B T(T/\theta)^3$, and the heat capacity is $\sim 12Nk_B(T/\theta)^3$.

For actual crystals the temperatures at which the T^3 approximation holds are quite low. It may be necessary to be below $T = \theta/50$ to get reasonably pure T^3 behavior.

Selected values of θ are given in Table 1. Note, for example, in the alkali metals that the heavier atoms have the lowest θ's, because the velocity of sound decreases as the density increases.

Einstein Model of the Density of States

Consider N oscillators of the same frequency ω_0 and in one dimension. The Einstein density of states is $D(\omega) = N\delta(\omega - \omega_0)$, where the delta function is centered at ω_0. The thermal energy of the system is

$$U = N\langle n\rangle \hbar\omega = \frac{N\hbar\omega}{e^{\hbar\omega/\tau} - 1} , \tag{33}$$

with ω now written in place of ω_0, for convenience.

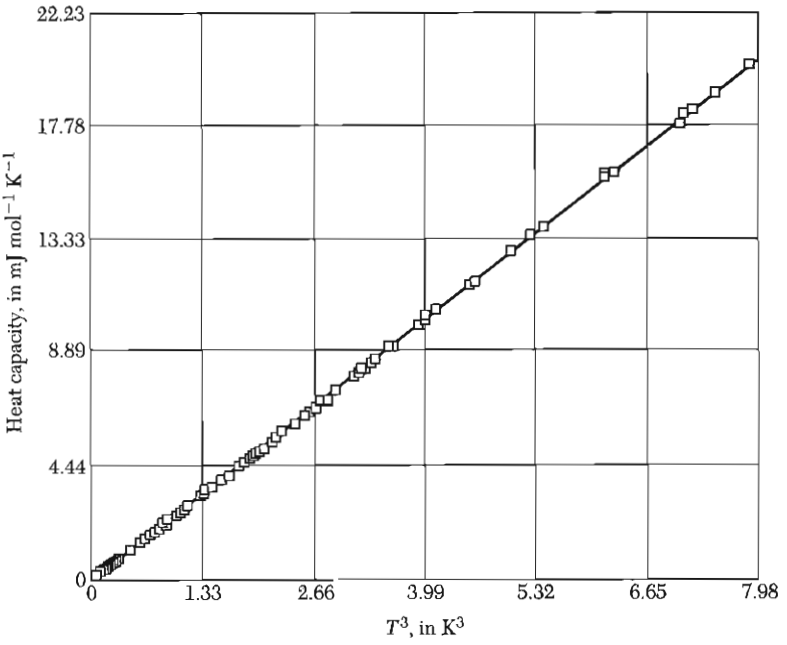


Figure 9 Low temperature heat capacity of solid argon, plotted against T^3. In this temperature region the experimental results are in excellent agreement with the Debye T^3 law with $\theta = 92.0$ K. (Courtesy of L. Finegold and N. E. Phillips.)

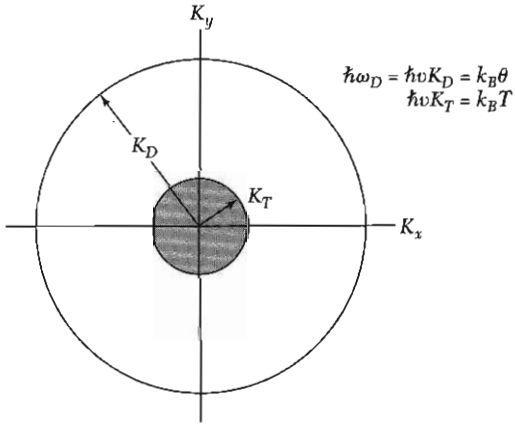


Figure 10 To obtain a qualitative explanation of the Debye T^3 law, we suppose that all phonon modes of wavevector less than K_T have the classical thermal energy k_BT and that modes between K_T and the Debye cutoff K_D are not excited at all. Of the $3N$ possible modes, the fraction excited is $(K_T/K_D)^3 = (T/\theta)^3$, because this is the ratio of the volume of the inner sphere to the outer sphere. The energy is $U \approx k_BT \cdot 3N(T/\theta)^3$, and the heat capacity is $C_V = \partial U/\partial T = 12Nk_B(T/\theta)^3$.

Table 1 Debye temperature and thermal conductivity

Low temperature limit of θ, in Kelvin
Thermal conductivity at 300 K, in W $cm^{-1}K^{-1}$

Li	Be											B	C	N	O	F	Ne
344	1440												2230				75
0.85	2.00											0.27	1.29				
Na	Mg											Al	Si	P	S	Cl	Ar
158	400											428	645				92
1.41	1.56											2.37	1.48				
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
91	230	360	420	380	630	410	470	445	450	343	327	320	374	282	90		72
1.02		0.16	0.22	0.31	0.94	0.08	0.80	1.00	0.91	4.01	1.16	0.41	0.60	0.50	0.02		
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn w	Sb	Te	I	Xe
56	147	280	291	275	450		600	480	274	225	209	108	200	211	153		64
0.58		0.17	0.23	0.54	1.38	0.51	1.17	1.50	0.72	4.29	0.97	0.82	0.67	0.24	0.02		
Cs	Ba	La β	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
38	110	142	252	240	400	430	500	420	240	165	71.9	78.5	105	119			
0.36		0.14	0.23	0.58	1.74	0.48	0.88	1.47	0.72	3.17		0.46	0.35	0.08			
Fr	Ra	Ac															

Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
						200		210				120	210
0.11	0.12	0.16		0.13		0.11	0.11	0.11	0.16	0.14	0.17	0.35	0.16
Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr
163		207											
0.54		0.28	0.06	0.07									

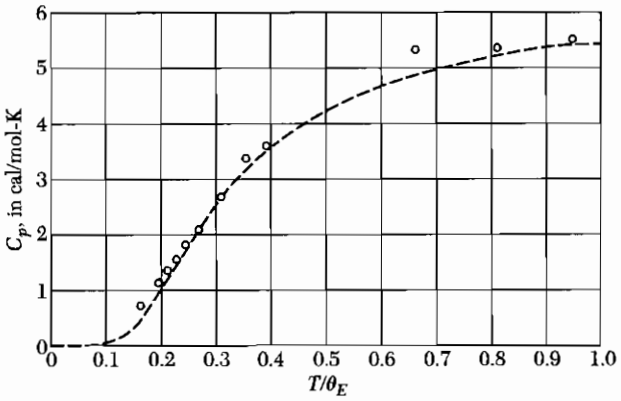


Figure 11 Comparison of experimental values of the heat capacity of diamond with values calculated on the earliest quantum (Einstein) model, using the characteristic temperature $\theta_E = \hbar\omega/k_B = 1320$ K. To convert to J/mol-deg, multiply by 4.186.

The heat capacity of the oscillators is

$$C_V = \left(\frac{\partial U}{\partial T}\right)_V = Nk_B \left(\frac{\hbar\omega}{\tau}\right)^2 \frac{e^{\hbar\omega/\tau}}{(e^{\hbar\omega/\tau} - 1)^2} , \tag{34}$$

as plotted in Fig. 11. This expresses the Einstein (1907) result for the contribution of N identical oscillators to the heat capacity of a solid. In three dimensions N is replaced by $3N$, there being three modes per oscillator. The high temperature limit of C_V becomes $3Nk_B$, which is known as the Dulong and Petit value.

At low temperatures (34) decreases as $\exp(-\hbar\omega/\tau)$, whereas the experimental form of the phonon contribution is known to be T^3as accounted for by the Debye model treated above. The Einstein model, however, is often used to approximate the optical phonon part of the phonon spectrum.

General Result for $D(\omega)$

We want to find a general expression for $D(\omega)$, the number of states per unit frequency range, given the phonon dispersion relation $\omega(\mathbf{K})$. The number of allowed values of $\mathbf{K}$ for which the phonon frequency is between ω and $\omega + d\omega$ is

$$D(\omega)\, d\omega = \left(\frac{L}{2\pi}\right)^3 \int_{\text{shell}} d^3K , \tag{35}$$

where the integral is extended over the volume of the shell in $\mathbf{K}$ space bounded by the two surfaces on which the phonon frequency is constant, one surface on which the frequency is ω and the other on which the frequency is $\omega + d\omega$.

The real problem is to evaluate the volume of this shell. We let dS_ω denote an element of area (Fig. 12) on the surface in $\mathbf{K}$ space of the selected constant

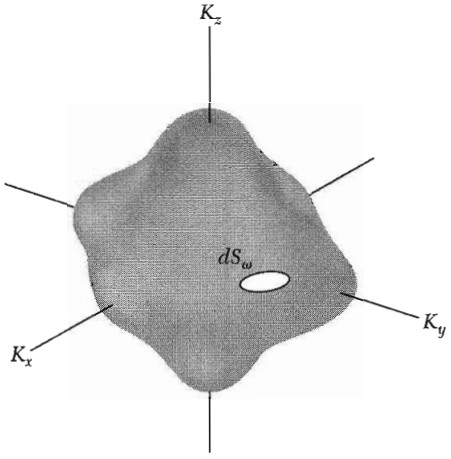


Figure 12 Element of area dS_ω on a constant frequency surface in **K** space. The volume between two surfaces of constant frequency at ω and $\omega + d\omega$ is equal to $\int dS_\omega\, d\omega/|\nabla_{\mathbf{K}}\omega|$.

frequency ω. The element of volume between the constant frequency surfaces ω and $\omega + d\omega$ is a right cylinder of base dS_ω and altitude $dK_\perp$, so that

$$\int_{\text{shell}} d^3K = \int dS_\omega dK_\perp \ , \tag{36}$$

Here $dK_\perp$ is the perpendicular distance (Fig. 13) between the surface ω constant and the surface $\omega + d\omega$ constant. The value of $dK_\perp$ will vary from one point to another on the surface.

The gradient of ω, which is $\nabla_{\mathbf{K}}\omega$, is also normal to the surface ω constant, and the quantity

$$|\nabla_{\mathbf{K}}\omega|\, dK_\perp = d\omega \ ,$$

is the difference in frequency between the two surfaces connected by $dK_\perp$. Thus the element of the volume is

$$dS_\omega\, dK_\perp = dS_\omega \frac{d\omega}{|\nabla_{\mathbf{K}}\omega|} = dS_\omega \frac{d\omega}{v_g} \ ,$$

where $v_g = |\nabla_{\mathbf{K}}\omega|$ is the magnitude of the group velocity of a phonon. For (35) we have

$$D(\omega)\, d\omega = \left(\frac{L}{2\pi}\right)^3 \int \frac{dS_\omega}{v_g}\, d\omega \ .$$

We divide both sides by $d\omega$ and write $V = L^3$ for the volume of the crystal: the result for the density of states is

$$\boxed{D(\omega) = \frac{V}{(2\pi)^3} \int \frac{dS_\omega}{v_g} \ .} \tag{37}$$

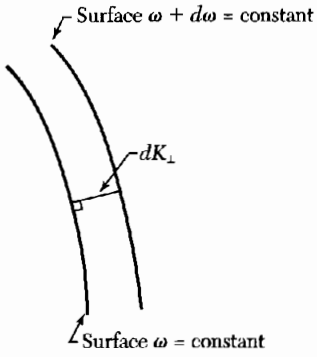


Figure 13 The quantity $dK_\perp$ is the perpendicular distance between two constant frequency surfaces in **K** space, one at frequency ω and the other at frequency $\omega + d\omega$.

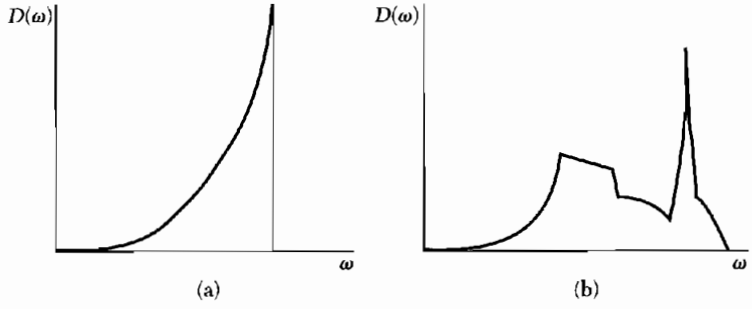


Figure 14 Density of states as a function of frequency for (a) the Debye solid and (b) an actual crystal structure. The spectrum for the crystal starts as ω^2 for small ω, but discontinuities develop at singular points.

The integral is taken over the area of the surface ω constant, in **K** space. The result refers to a single branch of the dispersion relation. We can use this result also in electron band theory.

There is a special interest in the contribution to $D(\omega)$ from points at which the group velocity is zero. Such critical points produce singularities (known as Van Hove singularities) in the distribution function (Fig. 14).

ANHARMONIC CRYSTAL INTERACTIONS

The theory of lattice vibrations discussed thus far has been limited in the potential energy to terms quadratic in the interatomic displacements. This is the harmonic theory; among its consequences are:

- Two lattice waves do not interact; a single wave does not decay or change form with time.
- There is no thermal expansion.
- Adiabatic and isothermal elastic constants are equal.
- The elastic constants are independent of pressure and temperature.
- The heat capacity becomes constant at high temperatures $T > \theta$.

In real crystals none of these consequences is satisfied accurately. The deviations may be attributed to the neglect of anharmonic (higher than quadratic) terms in the interatomic displacements. We discuss some of the simpler aspects of anharmonic effects.

Beautiful demonstrations of anharmonic effects are the experiments on the interaction of two phonons to produce a third phonon at a frequency $\omega_3 = \omega_1 + \omega_2$. Three-phonon processes are caused by third-order terms in the lattice potential energy. The physics of the phonon interaction can be stated simply: the presence of one phonon causes a periodic elastic strain which (through the anharmonic interaction) modulates in space and time the elastic constant of the crystal. A second phonon perceives the modulation of the elastic constant and thereupon is scattered to produce a third phonon, just as from a moving three-dimensional grating.

Thermal Expansion

We may understand thermal expansion by considering for a classical oscillator the effect of anharmonic terms in the potential energy on the mean separation of a pair of atoms at a temperature T. We take the potential energy of the atoms at a displacement x from their equilibrium separation at absolute zero as

$$U(x) = cx^2 - gx^3 - fx^4 \ , \tag{38}$$

with c, g, and f all positive. The term in x^3 represents the asymmetry of the mutual repulsion of the atoms and the term in x^4 represents the softening of the vibration at large amplitudes. The minimum at $x = 0$ is not an absolute minimum, but for small oscillations the form is an adequate representation of an interatomic potential.

We calculate the average displacement by using the Boltzmann distribution function, which weights the possible values of x according to their thermodynamic probability:

$$\langle x \rangle = \frac{\int_{-\infty}^{\infty} dx\, x \exp[-\beta U(x)]}{\int_{-\infty}^{\infty} dx \exp[-\beta U(x)]} \ ,$$

with $\beta \equiv 1/k_B T$. For displacements such that the anharmonic terms in the energy are small in comparison with $k_B T$, we may expand the integrands as

$$\int dx\, x \exp(-\beta U) \cong \int dx\, [\exp(-\beta c x^2)](x + \beta g x^4 + \beta f x^5) = (3\pi^{1/2}/4)(g/c^{5/2})\beta^{-3/2} \ ;$$
$$\int dx \exp(-\beta U) \cong \int dx \exp(-\beta c x^2) = (\pi/\beta c)^{1/2} \ , \tag{39}$$

whence the thermal expansion is

$$\langle x \rangle = \frac{3g}{4c^2} k_B T \tag{40}$$

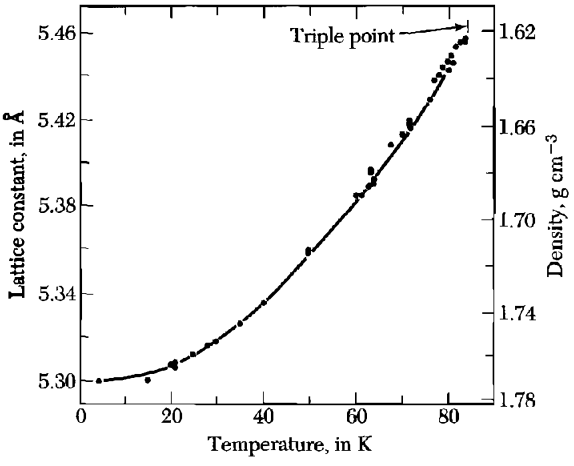


Figure 15 Lattice constant of solid argon as a function of temperature.

in the classical region. Note that in (39) we have left cx^2 in the exponential, but we have expanded $\exp(\beta g x^3 + \beta f x^4) \cong 1 + \beta g x^3 + \beta f x^4 + \cdots$.

Measurements of the lattice constant of solid argon are shown in Fig. 15. The slope of the curve is proportional to the thermal expansion coefficient. The expansion coefficient vanishes as $T \to 0$, as we expect from Problem 5. In lowest order the thermal expansion does not involve the symmetric term fx^4 in $U(x)$, but only the antisymmetric term gx^3.

THERMAL CONDUCTIVITY

The thermal conductivity coefficient K of a solid is defined with respect to the steady-state flow of heat down a long rod with a temperature gradient dT/dx:

$$j_U = -K\frac{dT}{dx}\,, \tag{41}$$

where j_U is the flux of thermal energy, or the energy transmitted across unit area per unit time.

This form implies that the process of thermal energy transfer is a random process. The energy does not simply enter one end of the specimen and proceed directly (ballistically) in a straight path to the other end, but diffuses through the specimen, suffering frequent collisions. If the energy were propagated directly through the specimen without deflection, then the expression for the thermal flux would not depend on the temperature gradient, but only on the difference in temperature ΔT between the ends of the specimen, regardless of the length of the specimen. The random nature of the conductivity process brings the temperature gradient and, as we shall see, a mean free path into the expression for the thermal flux.

Table 2 Phonon mean free paths

[Calculated from (44), taking $v = 5 \times 10^5$ cm/sec as a representative sound velocity. The ℓ's obtained in this way refer to umklapp processes.]

Crystal	T, °C	C, in J cm^{-3}K^{-1}	K, in W cm^{-1}K^{-1}	ℓ, in Å
Quartz[a]	0	2.00	0.13	40
	−190	0.55	0.50	540
NaCl	0	1.88	0.07	23
	−190	1.00	0.27	100

[a] Parallel to optic axis.

From the kinetic theory of gases we find below the following expression for the thermal conductivity:

$$K = \tfrac{1}{3} C v \ell \ , \tag{42}$$

where C is the heat capacity per unit volume, v is the average particle velocity, and ℓ is the mean free path of a particle between collisions. This result was applied first by Debye to describe thermal conductivity in dielectric solids, with C as the heat capacity of the phonons, v the phonon velocity, and ℓ the phonon mean free path. Several representative values of the mean free path are given in Table 2.

We give the elementary kinetic theory which leads to (42). The flux of particles in the x direction is $\frac{1}{2} n \langle | v_x | \rangle$, where n is the concentration of molecules; in equilibrium there is a flux of equal magnitude in the opposite direction. The $\langle \cdots \rangle$ denote average value.

If c is the heat capacity of a particle, then in moving from a region at local temperature $T + \Delta T$ to a region at local temperature T a particle will give up energy $c\,\Delta T$. Now ΔT between the ends of a free path of the particle is given by

$$\Delta T = \frac{dT}{dx} \ell_x = \frac{dT}{dx} v_x \tau \ ,$$

where τ is the average time between collisions.

The net flux of energy (from both senses of the particle flux) is therefore

$$j_U = -n \langle v_x^2 \rangle c \tau \frac{dT}{dx} = -\tfrac{1}{3} n \langle v^2 \rangle c \tau \frac{dT}{dx} \ . \tag{43}$$

If, as for phonons, v is constant, we may write (43) as

$$j_U = -\tfrac{1}{3} C v \ell \frac{dT}{dx} \ , \tag{44}$$

with $\ell \equiv v\tau$ and $C \equiv nc$. Thus $K = \frac{1}{3} C v \ell$.

Thermal Resistivity of Phonon Gas

The phonon mean free path ℓ is determined principally by two processes, geometrical scattering and scattering by other phonons. If the forces between atoms were purely harmonic, there would be no mechanism for collisions between different phonons, and the mean free path would be limited solely by collisions of a phonon with the crystal boundary, and by lattice imperfections. There are situations where these effects are dominant.

With anharmonic lattice interactions, there is a coupling between different phonons which limits the value of the mean free path. The exact states of the anharmonic system are no longer like pure phonons.

The theory of the effect of anharmonic coupling on thermal resistivity predicts that ℓ is proportional to $1/T$ at high temperatures, in agreement with many experiments. We can understand this dependence in terms of the number of phonons with which a given phonon can interact: at high temperature the total number of excited phonons is proportional to T. The collision frequency of a given phonon should be proportional to the number of phonons with which it can collide, whence $\ell \propto 1/T$.

To define a thermal conductivity there must exist mechanisms in the crystal whereby the distribution of phonons may be brought locally into thermal equilibrium. Without such mechanisms we may not speak of the phonons at one end of the crystal as being in thermal equilibrium at a temperature T_2 and those at the other end in equilibrium at T_1.

It is not sufficient to have only a way of limiting the mean free path, but there must also be a way of establishing a local thermal equilibrium distribution of phonons. Phonon collisions with a static imperfection or a crystal boundary will not by themselves establish thermal equilibrium, because such collisions do not change the energy of individual phonons: the frequency ω_2 of the scattered phonon is equal to the frequency ω_1 of the incident phonon.

It is rather remarkable also that a three-phonon collision process

$$\mathbf{K}_1 + \mathbf{K}_2 = \mathbf{K}_3 \tag{45}$$

will not establish equilibrium, but for a subtle reason: the total momentum of the phonon gas is not changed by such a collision. An equilibrium distribution of phonons at a temperature T can move down the crystal with a drift velocity which is not disturbed by three-phonon collisions of the form (45). For such collisions the phonon momentum

$$\mathbf{J} = \sum_{\mathbf{K}} n_{\mathbf{K}} \hbar \mathbf{K} \tag{46}$$

is conserved, because on collision the change in $\mathbf{J}$ is $\mathbf{K}_3 - \mathbf{K}_2 - \mathbf{K}_1 = 0$. Here $n_{\mathbf{K}}$ is the number of phonons having wavevector $\mathbf{K}$.

For a distribution with $\mathbf{J} \neq 0$, collisions such as (45) are incapable of establishing complete thermal equilibrium because they leave $\mathbf{J}$ unchanged. If

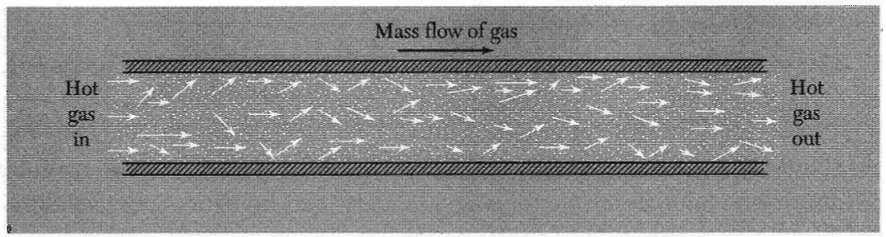


Figure 16a Flow of gas molecules in a state of drifting equilibrium down a long open tube with frictionless walls. Elastic collision processes among the gas molecules do not change the momentum or energy flux of the gas because in each collision the velocity of the center of mass of the colliding particles and their energy remain unchanged. Thus energy is transported from left to right without being driven by a temperature gradient. Therefore the thermal resistivity is zero and the thermal conductivity is infinite.

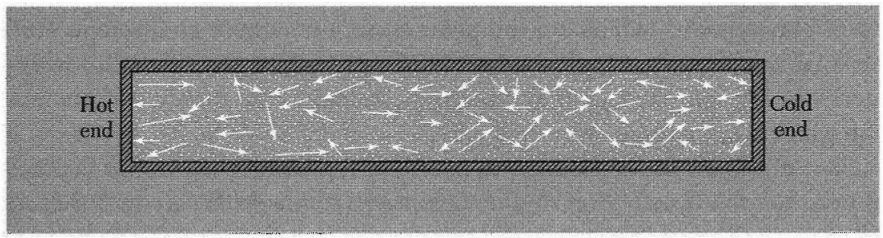


Figure 16b The usual definition of thermal conductivity in a gas refers to a situation where no mass flow is permitted. Here the tube is closed at both ends, preventing the escape or entrance of molecules. With a temperature gradient the colliding pairs with above-average center of mass velocities will tend to be directed to the right, those with below-average velocities will tend to be directed to the left. A slight concentration gradient, high on the right, will be set up to enable the net mass transport to be zero while allowing a net energy transport from the hot to the cold end.

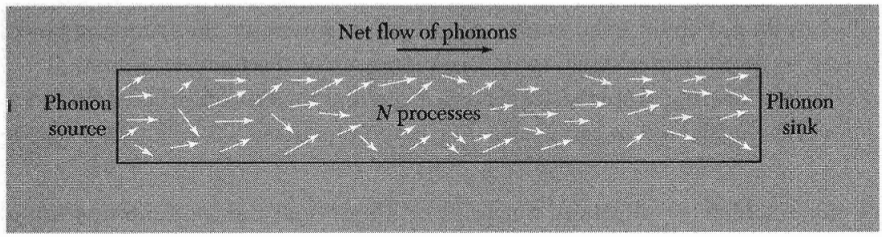


Figure 16c In a crystal we may arrange to create phonons chiefly at one end, as by illuminating the left end with a lamp. From that end there will be a net flux of phonons toward the right end of the crystal. If only N processes ($\mathbf{K}_1 + \mathbf{K}_2 = \mathbf{K}_3$) occur, the phonon flux is unchanged in momentum on collision and some phonon flux will persist down the length of the crystal. On arrival of phonons at the right end we can arrange in principle to convert most of their energy to radiation, thereby creating a sink for the phonons. Just as in (a) the thermal resistivity is zero.

we start a distribution of hot phonons down a rod with $\mathbf{J} \neq 0$, the distribution will propagate down the rod with $\mathbf{J}$ unchanged. Therefore there is no thermal resistance. The problem as illustrated in Fig. 16 is like that of the collisions between molecules of a gas in a straight tube with frictionless walls.

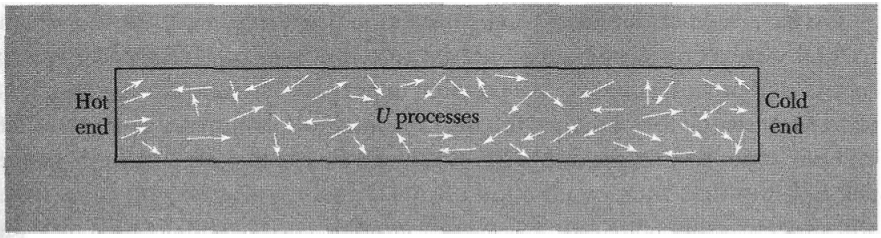


Figure 16d In U processes there is a large net change in phonon momentum in each collision event. An initial net phonon flux will rapidly decay as we move to the right. The ends may act as sources and sinks. Net energy transport under a temperature gradient occurs as in (b).

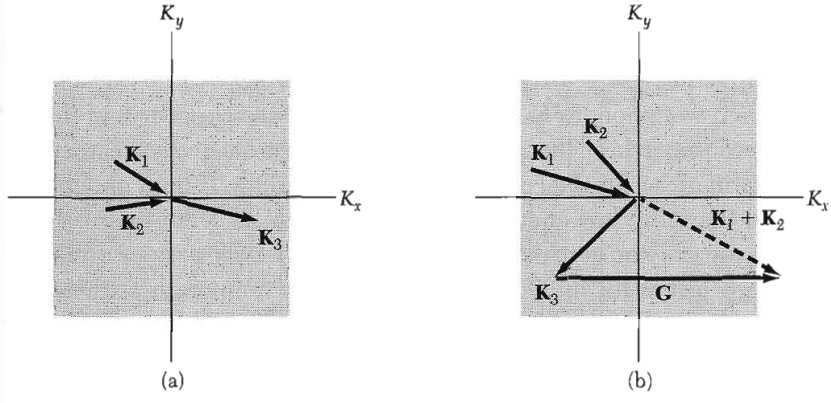


Figure 17 (a) Normal $\mathbf{K}_1 + \mathbf{K}_2 = \mathbf{K}_3$ and (b) umklapp $\mathbf{K}_1 + \mathbf{K}_2 = \mathbf{K}_3 + \mathbf{G}$ phonon collision processes in a two-dimensional square lattice. The square in each figure represents the first Brillouin zone in the phonon $\mathbf{K}$ space; this zone contains all the possible independent values of the phonon wavevector. Vectors $\mathbf{K}$ with arrowheads at the center of the zone represent phonons absorbed in the collision process; those with arrowheads away from the center of the zone represent phonons emitted in the collision. We see in (b) that in the umklapp process the direction of the x-component of the phonon flux has been reversed. The reciprocal lattice vector $\mathbf{G}$ as shown is of length $2\pi/a$, where a is the lattice constant of the crystal lattice, and is parallel to the K_x axis. For all processes, N or U, energy must be conserved, so that $\omega_1 + \omega_2 = \omega_3$.

Umklapp Processes

The important three-phonon processes that cause thermal resistivity are not of the form $\mathbf{K}_1 + \mathbf{K}_2 = \mathbf{K}_3$ in which $\mathbf{K}$ is conserved, but are of the form

$$\mathbf{K}_1 + \mathbf{K}_2 = \mathbf{K}_3 + \mathbf{G} \ , \tag{47}$$

where $\mathbf{G}$ is a reciprocal lattice vector (Fig. 17). These processes, discovered by Peierls, are called **umklapp processes**. We recall that $\mathbf{G}$ may occur in all momentum conservation laws in crystals. In all allowed processes of the form of (46) and (47), energy is conserved.

We have seen examples of wave interaction processes in crystals for which the total wavevector change need not be zero, but may be a reciprocal lattice vector. Such processes are always possible in periodic lattices. The argument is particularly strong for phonons: the only meaningful phonon **K**'s lie in the first Brillouin zone, so that any longer **K** produced in a collision must be brought back into the first zone by addition of a **G**. A collision of two phonons both with a negative value of K_x can by an umklapp process ($G \neq 0$), create a phonon with positive K_x. Umklapp processes are also called U processes.

Collisions in which $\mathbf{G} = 0$ are called **normal processes** or N processes. At high temperatures $T > \theta$ all phonon modes are excited because $k_B T > \hbar\omega_{max}$. A substantial proportion of all phonon collisions will then be U processes, with the attendant high momentum change in the collision. In this regime we can estimate the thermal resistivity without particular distinction between N and U processes; by the earlier argument about nonlinear effects we expect to find a lattice thermal resistivity $\propto T$ at high temperatures.

The energy of phonons $\mathbf{K}_1$, $\mathbf{K}_2$ suitable for umklapp to occur is of the order of $\frac{1}{2}k_B\theta$, because each of the phonons 1 and 2 must have wavevectors of the order of $\frac{1}{2}G$ in order for the collision (47) to be possible. If both phonons have low K, and therefore low energy, there is no way to get from their collision a phonon of wavevector outside the first zone. The umklapp process must conserve energy, just as for the normal process. At low temperatures the number of suitable phonons of the high energy $\frac{1}{2}k_B\theta$ required may be expected to vary roughly as $\exp(-\theta/2T)$, according to the Boltzmann factor. The exponential form is in good agreement with experiment. In summary, the phonon mean free path which enters (42) is the mean free path for umklapp collisions between phonons and not for all collisions between phonons.

Imperfections

Geometrical effects may also be important in limiting the mean free path. We must consider scattering by crystal boundaries, the distribution of isotopic masses in natural chemical elements, chemical impurities, lattice imperfections, and amorphous structures.

When at low temperatures the mean free path ℓ becomes comparable with the width of the test specimen, the value of ℓ is limited by the width, and the thermal conductivity becomes a function of the dimensions of the specimen. This effect was discovered by de Haas and Biermasz. The abrupt decrease in thermal conductivity of pure crystals at low temperatures is caused by the size effect.

At low temperatures the umklapp process becomes ineffective in limiting the thermal conductivity, and the size effect becomes dominant, as shown in Fig. 18. One would expect then that the phonon mean free path would be constant and of the order of the diameter D of the specimen, so that

$$K \approx CvD \ . \tag{48}$$

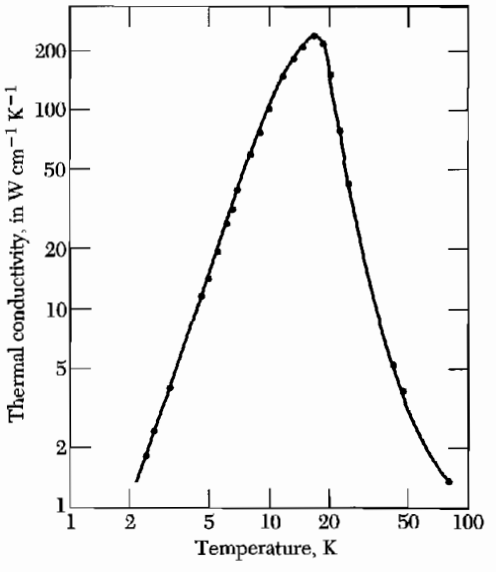


Figure 18 Thermal conductivity of a highly purified crystal of sodium fluoride, after H. E. Jackson, C. T. Walker, and T. F. McNelly.

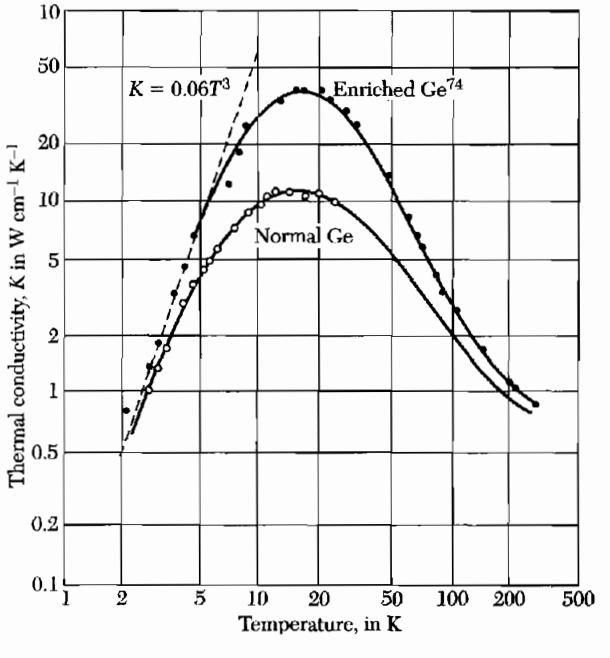


Figure 19 Isotope effect on thermal conduction in germanium, amounting to a factor of three at the conductivity maximum. The enriched specimen is 96 percent Ge^{74}, natural germanium is 20 percent Ge^{70}, 27 percent Ge^{72}, 8 percent Ge^{73}, 37 percent Ge^{74}, and 8 percent Ge^{76}. Below 5 K the enriched specimen has $K = 0.06\,T^3$, which agrees well with Casimir's theory for thermal resistance caused by boundary scattering. (After T. H. Geballe and G. W. Hull.)

The only temperature-dependent term on the right is C, the heat capacity, which varies as T^3 at low temperatures. We expect the thermal conductivity to vary as T^3 at low temperatures. The size effect enters whenever the phonon mean free path becomes comparable with the diameter of the specimen.

Dielectric crystals may have thermal conductivities as high as metals. Synthetic sapphire (Al_2O_3) has one of the highest values of the conductivity: nearly

200 W cm^{-1} K^{-1} at 30 K. The maximum of the thermal conductivity in sapphire is greater than the maximum of 100 W cm^{-1} K^{-1} in copper. Metallic gallium, however, has a conductivity of 845 W cm^{-1} K^{-1} at 1.8 K. The electronic contribution to the thermal conductivity of metals is treated in Chapter 6.

In an otherwise perfect crystal, the distribution of isotopes of the chemical elements often provides an important mechanism for phonon scattering. The random distribution of isotopic mass disturbs the periodicity of the density as seen by an elastic wave. In some substances scattering of phonons by isotopes is comparable in importance to scattering by other phonons. Results for germanium are shown in Fig. 19. Enhanced thermal conductivity has been observed also in isotopically pure silicon and diamond; the latter has device importance as a heat sink for laser sources.

Problems

1. ***Singularity in density of states.*** (a) From the dispersion relation derived in Chapter 4 for a monatomic linear lattice of N atoms with nearest-neighbor interactions, show that the density of modes is

$$D(\omega) = \frac{2N}{\pi} \cdot \frac{1}{(\omega_m^2 - \omega^2)^{1/2}} .$$

where ω_m is the maximum frequency. (b) Suppose that an optical phonon branch has the form $\omega(K) = \omega_0 - AK^2$, near $K = 0$ in three dimensions. Show that $D(\omega) = (L/2\pi)^3(2\pi/A^{3/2})(\omega_0 - \omega)^{1/2}$ for $\omega < \omega_0$ and $D(\omega) = 0$ for $\omega > \omega_0$. Here the density of modes is discontinuous.

2. ***Rms thermal dilation of crystal cell.*** (a) Estimate for 300 K the root mean square thermal dilation $\Delta V/V$ for a primitive cell of sodium. Take the bulk modulus as 7×10^{10} erg cm^{-3}. Note that the Debye temperature 158 K is less than 300 K, so that the thermal energy is of the order of $k_B T$. (b) Use this result to estimate the root mean square thermal fluctuation $\Delta a/a$ of the lattice parameter.

3. ***Zero point lattice displacement and strain.*** (a) In the Debye approximation, show that the mean square displacement of an atom at absolute zero is $\langle R^2 \rangle = 3\hbar\omega_D^2/8\pi^2\rho v^3$, where v is the velocity of sound. Start from the result (4.29) summed over the independent lattice modes: $\langle R^2 \rangle = (\hbar/2\rho V)\Sigma\omega^{-1}$. We have included a factor of $\frac{1}{2}$ to go from mean square amplitude to mean square displacement. (b) Show that $\Sigma\omega^{-1}$ and $\langle R^2 \rangle$ diverge for a one-dimensional lattice, but that the mean square strain is finite. Consider $\langle (\partial R/\partial x)^2 \rangle = \frac{1}{2}\Sigma K^2 u_0^2$ as the mean square strain, and show that it is equal to $\hbar\omega_D^2 L/4MNv^3$ for a line of N atoms each of mass M, counting longitudinal modes only. The divergence of R^2 is not significant for any physical measurement.

4. ***Heat capacity of layer lattice.*** (a) Consider a dielectric crystal made up of layers of atoms, with rigid coupling between layers so that the motion of the atoms is restricted to the plane of the layer. Show that the phonon heat capacity in the Debye approximation in the low temperature limit is proportional to T^2.

(b) Suppose instead, as in many layer structures, that adjacent layers are very weakly bound to each other. What form would you expect the phonon heat capacity to approach at extremely low temperatures?

*5. ***Grüneisen constant.*** (a) Show that the free energy of a phonon mode of frequency ω is $k_BT \ln [2 \sinh (\hbar\omega/2k_BT)]$. It is necessary to retain the zero-point energy $\frac{1}{2}\hbar\omega$ to obtain this result. (b) If Δ is the fractional volume change, then the free energy of the crystal may be written as

$$F(\Delta, T) = \tfrac{1}{2}B\Delta^2 + k_BT \sum \ln [2 \sinh (\hbar\omega_{\mathbf{K}}/2k_BT)] \ ,$$

where B is the bulk modulus. Assume that the volume dependence of $\omega_{\mathbf{K}}$ is $\delta\omega/\omega = -\gamma\Delta$, where γ is known as the Grüneisen constant. If γ is taken as independent of the mode $\mathbf{K}$, show that F is a minimum with respect to Δ when $B\Delta = \gamma\Sigma\frac{1}{2}\hbar\omega$ coth $(\hbar\omega/2k_BT)$, and show that this may be written in terms of the thermal energy density as $\Delta = \gamma U(T)/B$. (c) Show that on the Debye model $\gamma = -\partial \ln \theta/\partial \ln V$. Note: Many approximations are involved in this theory: the result (a) is valid only if ω is independent of temperature; γ may be quite different for different modes.

*This problem is somewhat difficult.

6

Free Electron Fermi Gas

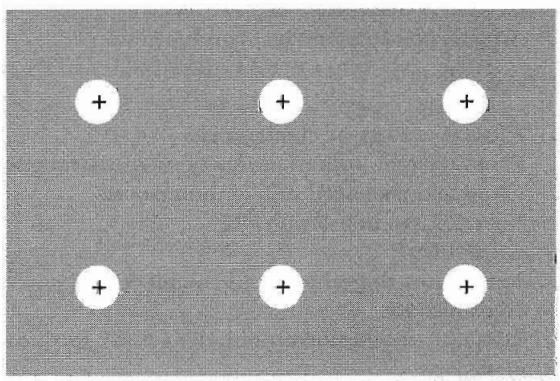


Figure 1 Schematic model of a crystal of sodium metal. The atomic cores are Na^+ ions: they are immersed in a sea of conduction electrons. The conduction electrons are derived from the 3*s* valence electrons of the free atoms. The atomic cores contain 10 electrons in the configuration $1s^2 2s^2 2p^6$. In an alkali metal the atomic cores occupy a relatively small part (~15 percent) of the total volume of the crystal, but in a noble metal (Cu, Ag, Au) the atomic cores are relatively larger and may be in contact with each other. The common crystal structure at room temperature is bcc for the alkali metals and fcc for the noble metals.

CHAPTER 6: FREE ELECTRON FERMI GAS

In a theory which has given results like these, there must certainly be a great deal of truth.
H. A. Lorentz

We can understand many physical properties of metals, and not only of the simple metals, in terms of the free electron model. According to this model, the valence electrons of the constituent atoms become conduction electrons and move about freely through the volume of the metal. Even in metals for which the free electron model works best, the charge distribution of the conduction electrons reflects the strong electrostatic potential of the ion cores. The utility of the free electron model is greatest for properties that depend essentially on the kinetic properties of the conduction electrons. The interaction of the conduction electrons with the ions of the lattice is treated in the next chapter.

The simplest metals are the alkali metals—lithium, sodium, potassium, cesium, and rubidium. In a free atom of sodium the valence electron is in a $3s$ state; in the metal this electron becomes a conduction electron in the $3s$ conduction band.

A monovalent crystal which contains N atoms will have N conduction electrons and N positive ion cores. The Na^+ ion core contains 10 electrons that occupy the $1s$, $2s$, and $2p$ shells of the free ion, with a spatial distribution that is essentially the same when in the metal as in the free ion. The ion cores fill only about 15 percent of the volume of a sodium crystal, as in Fig. 1. The radius of the free Na^+ ion is 0.98 Å, whereas one-half of the nearest-neighbor distance of the metal is 1.83 Å.

The interpretation of metallic properties in terms of the motion of free electrons was developed long before the invention of quantum mechanics. The classical theory had several conspicuous successes, notably the derivation of the form of Ohm's law and the relation between the electrical and thermal conductivity. The classical theory fails to explain the heat capacity and the magnetic susceptibility of the conduction electrons. (These are not failures of the free electron model, but failures of the classical Maxwell distribution function.)

There is a further difficulty with the classical model. From many types of experiments it is clear that a conduction electron in a metal can move freely in a straight path over many atomic distances, undeflected by collisions with other conduction electrons or by collisions with the atom cores. In a very pure specimen at low temperatures, the mean free path may be as long as 10^8 interatomic spacings (more than 1 cm).

Why is condensed matter so transparent to conduction electrons? The answer to the question contains two parts: (a) A conduction electron is not

deflected by ion cores arranged on a *periodic* lattice because matter waves can propagate freely in a periodic structure, as a consequence of the mathematics treated in the following chapter. (b) A conduction electron is scattered only infrequently by other conduction electrons. This property is a consequence of the Pauli exclusion principle. By a **free electron Fermi gas**, we shall mean a gas of free electrons subject to the Pauli principle.

ENERGY LEVELS IN ONE DIMENSION

Consider a free electron gas in one dimension, taking account of quantum theory and of the Pauli principle. An electron of mass m is confined to a length L by infinite barriers (Fig. 2). The wavefunction $\psi_n(x)$ of the electron is a solution of the Schrödinger equation $\mathcal{H}\psi = \epsilon\psi$; with the neglect of potential energy we have $\mathcal{H} = p^2/2m$, where p is the momentum. In quantum theory p may be represented by the operator $-i\hbar\, d/dx$, so that

$$\mathcal{H}\psi_n = -\frac{\hbar^2}{2m}\frac{d^2\psi_n}{dx^2} = \epsilon_n\psi_n \ , \tag{1}$$

where ϵ_n is the energy of the electron in the orbital.

We use the term **orbital** to denote a solution of the wave equation for a system of only one electron. The term allows us to distinguish between an exact quantum state of the wave equation of a system of N interacting electrons and an approximate quantum state which we construct by assigning the N electrons to N different orbitals, where each orbital is a solution of a wave equation for one electron. The orbital model is exact only if there are no interactions between electrons.

The boundary conditions are $\psi_n(0) = 0$; $\psi_n(L) = 0$, as imposed by the infinite potential energy barriers. They are satisfied if the wavefunction is sinelike with an integral number n of half-wavelengths between 0 and L:

$$\psi_n = A \sin\left(\frac{2\pi}{\lambda_n}x\right) \ ; \qquad \tfrac{1}{2}n\lambda_n = L \ , \tag{2}$$

where A is a constant. We see that (2) is a solution of (1), because

$$\frac{d\psi_n}{dx} = A\left(\frac{n\pi}{L}\right)\cos\left(\frac{n\pi}{L}x\right) \ ; \qquad \frac{d^2\psi_n}{dx^2} = -A\left(\frac{n\pi}{L}\right)^2 \sin\left(\frac{n\pi}{L}x\right) \ ,$$

whence the energy ϵ_n is given by

$$\epsilon_n = \frac{\hbar^2}{2m}\left(\frac{n\pi}{L}\right)^2 \ . \tag{3}$$

We want to accommodate N electrons on the line. According to the **Pauli exclusion principle**, no two electrons can have all their quantum numbers

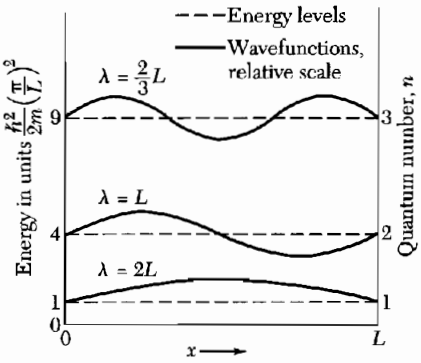


Figure 2 First three energy levels and wavefunctions of a free electron of mass m confined to a line of length L. The energy levels are labeled according to the quantum number n which gives the number of half-wavelengths in the wavefunction. The wavelengths are indicated on the wavefunctions. The energy ϵ_n of the level of quantum number n is equal to $(h^2/2m)(n/2L)^2$.

identical. That is, each orbital can be occupied by at most one electron. This applies to electrons in atoms, molecules, or solids.

In a linear solid the quantum numbers of a conduction electron orbital are n and m_s, where n is any positive integer and the magnetic quantum number $m_s = \pm\frac{1}{2}$, according to spin orientation. A pair of orbitals labeled by the quantum number n can accommodate two electrons, one with spin up and one with spin down.

If there are six electrons, then in the ground state of the system the filled orbitals are those given in the table:

n	m_s	Electron occupancy	n	m_s	Electron occupancy
1	↑	1	3	↑	1
1	↓	1	3	↓	1
2	↑	1	4	↑	0
2	↓	1	4	↓	0

More than one orbital may have the same energy. The number of orbitals with the same energy is called the **degeneracy.**

Let n_F denote the topmost filled energy level, where we start filling the levels from the bottom ($n = 1$) and continue filling higher levels with electrons until all N electrons are accommodated. It is convenient to suppose that N is an even number. The condition $2n_F = N$ determines n_F, the value of n for the uppermost filled level.

The **Fermi energy** ϵ_F is defined as the energy of the topmost filled level in the ground state of the N electron system. By (3) with $n = n_F$ we have in one dimension:

$$\epsilon_F = \frac{\hbar^2}{2m}\left(\frac{n_F\pi}{L}\right)^2 = \frac{\hbar^2}{2m}\left(\frac{N\pi}{2L}\right)^2 . \tag{4}$$

EFFECT OF TEMPERATURE ON THE FERMI-DIRAC DISTRIBUTION

The ground state is the state of the N electron system at absolute zero. What happens as the temperature is increased? This is a standard problem in elementary statistical mechanics, and the solution is given by the Fermi-Dirac distribution function (Appendix D and *TP*, Chapter 7).

The kinetic energy of the electron gas increases as the temperature is increased: some energy levels are occupied which were vacant at absolute zero, and some levels are vacant which were occupied at absolute zero (Fig. 3). The **Fermi-Dirac distribution** gives the probability that an orbital at energy ϵ will be occupied in an ideal electron gas in thermal equilibrium:

$$f(\epsilon) = \frac{1}{\exp[(\epsilon - \mu)/k_B T] + 1} . \tag{5}$$

The quantity μ is a function of the temperature; μ is to be chosen for the particular problem in such a way that the total number of particles in the system comes out correctly—that is, equal to N. At absolute zero $\mu = \epsilon_F$, because in the limit $T \to 0$ the function $f(\epsilon)$ changes discontinuously from the value 1 (filled) to the value 0 (empty) at $\epsilon = \epsilon_F = \mu$. *At all temperatures $f(\epsilon)$ is equal to $\frac{1}{2}$ when* $\epsilon = \mu$, for then the denominator of (5) has the value 2.

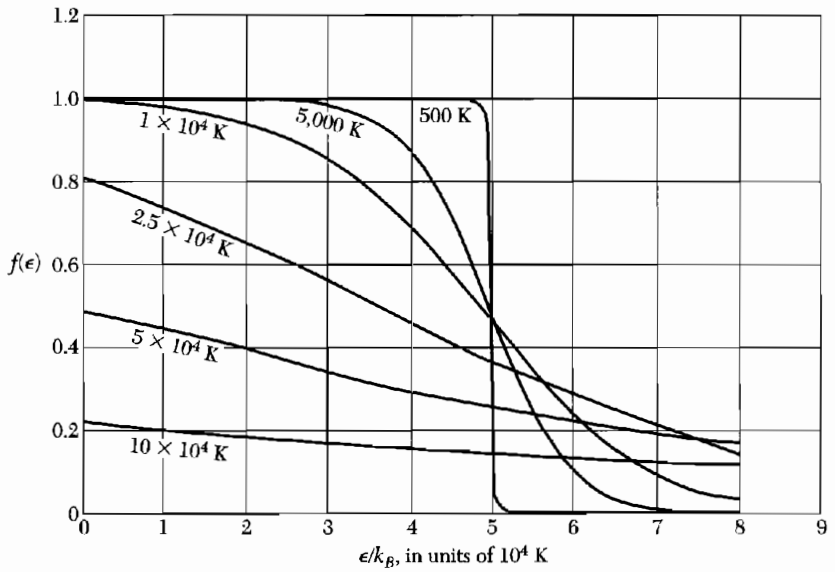


Figure 3 Fermi-Dirac distribution function (5) at the various labelled temperatures, for $T_F \equiv \epsilon_F/k_B = 50{,}000$ K. The results apply to a gas in three dimensions. The total number of particles is constant, independent of temperature. The chemical potential μ at each temperature may be read off the graph as the energy at which $f = 0.5$.

The quantity μ is the **chemical potential** (*TP*, Chapter 5), and we see that at absolute zero the chemical potential is equal to the Fermi energy, defined as the energy of the topmost filled orbital at absolute zero.

The high energy tail of the distribution is that part for which $\epsilon - \mu \gg k_B T$; here the exponential term is dominant in the denominator of (5), so that $f(\epsilon) \cong \exp[(\mu - \epsilon)/k_B T]$. This limit is called the Boltzmann or Maxwell distribution.

FREE ELECTRON GAS IN THREE DIMENSIONS

The free-particle Schrödinger equation in three dimensions is

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\psi_{\mathbf{k}}(\mathbf{r}) = \epsilon_{\mathbf{k}}\,\psi_{\mathbf{k}}(\mathbf{r}) \ . \tag{6}$$

If the electrons are confined to a cube of edge L, the wavefunction is the standing wave

$$\psi_n(\mathbf{r}) = A \sin(\pi n_x x/L) \sin(\pi n_y y/L) \sin(\pi n_z z/L) \ , \tag{7}$$

where n_x, n_y, n_z are positive integers. The origin is at one corner of the cube.

It is convenient to introduce wavefunctions that satisfy periodic boundary conditions, as we did for phonons in Chapter 5. We now require the wavefunctions to be periodic in x, y, z with period L. Thus

$$\psi(x + L, y, z) = \psi(x, y, z) \ , \tag{8}$$

and similarly for the y and z coordinates. Wavefunctions satisfying the free-particle Schrödinger equation and the periodicity condition are of the form of a traveling plane wave:

$$\boxed{\psi_{\mathbf{k}}(\mathbf{r}) = \exp(i\mathbf{k}\cdot\mathbf{r}) \ ,} \tag{9}$$

provided that the components of the wavevector $\mathbf{k}$ satisfy

$$k_x = 0 \ ; \quad \pm\frac{2\pi}{L} \ ; \quad \pm\frac{4\pi}{L} \ ; \ \ldots \ , \tag{10}$$

and similarly for k_y and k_z.

Any component of $\mathbf{k}$ of the form $2n\pi/L$ will satisfy the periodicity condition over a length L, where n is a positive or negative integer. The components of k are the quantum numbers of the problem, along with the quantum number m_s for the spin direction. We confirm that these values of k_x satisfy (8), for

$$\begin{aligned}\exp[ik_x(x + L)] &= \exp[i2n\pi(x + L)/L] \\ &= \exp(i2n\pi x/L)\exp(i2n\pi) = \exp(i2n\pi x/L) = \exp(ik_x x) \ .\end{aligned} \tag{11}$$

On substituting (9) in (6) we have the energy $\epsilon_{\mathbf{k}}$ of the orbital with wavevector $\mathbf{k}$:

$$\epsilon_{\mathbf{k}} = \frac{\hbar^2}{2m} k^2 = \frac{\hbar^2}{2m} (k_x^2 + k_y^2 + k_z^2) \ . \tag{12}$$

The magnitude k of the wavevector is related to the wavelength λ by $k = 2\pi/\lambda$.

The linear momentum $\mathbf{p}$ may be represented in quantum mechanics by the operator $\mathbf{p} = -i\hbar\nabla$, whence for the orbital (9)

$$\mathbf{p}\psi_{\mathbf{k}}(\mathbf{r}) = -i\hbar\nabla\psi_{\mathbf{k}}(\mathbf{r}) = \hbar\mathbf{k}\psi_{\mathbf{k}}(\mathbf{r}) \ , \tag{13}$$

so that the plane wave $\psi_{\mathbf{k}}$ is an eigenfunction of the linear momentum with the eigenvalue $\hbar\mathbf{k}$. The particle velocity in the orbital $\mathbf{k}$ is given by $\mathbf{v} = \hbar\mathbf{k}/m$.

In the ground state of a system of N free electrons, the occupied orbitals may be represented as points inside a sphere in $\mathbf{k}$ space. The energy at the surface of the sphere is the Fermi energy; the wavevectors at the Fermi surface have a magnitude k_F such that (Fig. 4):

$$\epsilon_F = \frac{\hbar^2}{2m} k_F^2 \ . \tag{14}$$

From (10) we see that there is one allowed wavevector—that is, one distinct triplet of quantum numbers k_x, k_y, k_z—for the volume element $(2\pi/L)^3$ of $\mathbf{k}$ space. Thus in the sphere of volume $4\pi k_F^3/3$ the total number of orbitals is

$$2 \cdot \frac{4\pi k_F^3/3}{(2\pi/L)^3} = \frac{V}{3\pi^2} k_F^3 = N \ , \tag{15}$$

where the factor 2 on the left comes from the two allowed values of the spin quantum number for each allowed value of $\mathbf{k}$. Then (15) gives

$$k_F = \left(\frac{3\pi^2 N}{V}\right)^{1/3} \ , \tag{16}$$

which depends only on the particle concentration.

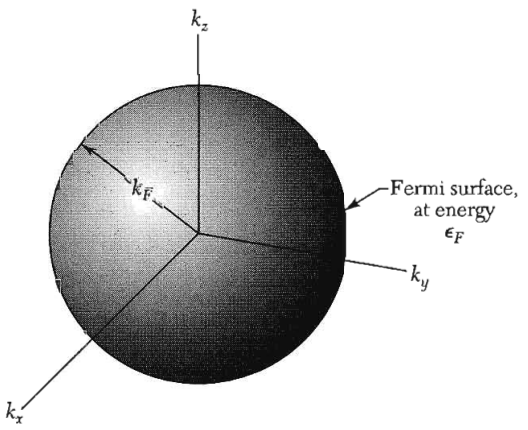


Figure 4 In the ground state of a system of N free electrons the occupied orbitals of the system fill a sphere of radius k_F, where $\epsilon_F = \hbar^2 k_F^2/2m$ is the energy of an electron having a wavevector k_F.

Table 1 Calculated free electron Fermi surface parameters for metals at room temperature

(Except for Na, K, Rb, Cs at 5 K and Li at 78 K)

Valency	Metal	Electron concentration, in cm^{-3}	Radius[a] parameter r_s	Fermi wavevector, in cm^{-1}	Fermi velocity, in cm s^{-1}	Fermi energy, in eV	Fermi temperature $T_F \equiv \epsilon_F/k_B$, in deg K
1	Li	4.70×10^{22}	3.25	1.11×10^8	1.29×10^8	4.72	5.48×10^4
	Na	2.65	3.93	0.92	1.07	3.23	3.75
	K	1.40	4.86	0.75	0.86	2.12	2.46
	Rb	1.15	5.20	0.70	0.81	1.85	2.15
	Cs	0.91	5.63	0.64	0.75	1.58	1.83
	Cu	8.45	2.67	1.36	1.57	7.00	8.12
	Ag	5.85	3.02	1.20	1.39	5.48	6.36
	Au	5.90	3.01	1.20	1.39	5.51	6.39
2	Be	24.2	1.88	1.93	2.23	14.14	16.41
	Mg	8.60	2.65	1.37	1.58	7.13	8.27
	Ca	4.60	3.27	1.11	1.28	4.68	5.43
	Sr	3.56	3.56	1.02	1.18	3.95	4.58
	Ba	3.20	3.69	0.98	1.13	3.65	4.24
	Zn	13.10	2.31	1.57	1.82	9.39	10.90
	Cd	9.28	2.59	1.40	1.62	7.46	8.66
3	Al	18.06	2.07	1.75	2.02	11.63	13.49
	Ga	15.30	2.19	1.65	1.91	10.35	12.01
	In	11.49	2.41	1.50	1.74	8.60	9.98
4	Pb	13.20	2.30	1.57	1.82	9.37	10.87
	Sn(w)	14.48	2.23	1.62	1.88	10.03	11.64

[a]The dimensionless radius parameter is defined as $r_s = r_0/a_H$, where a_H is the first Bohr radius and r_0 is the radius of a sphere that contains one electron.

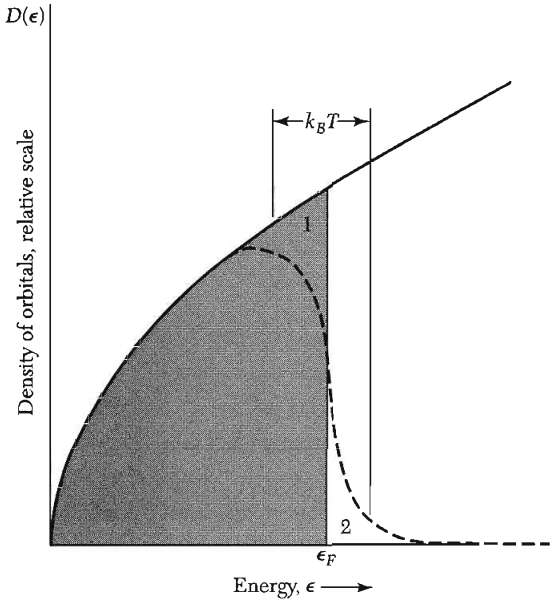


Figure 5 Density of single-particle states as a function of energy, for a free electron gas in three dimensions. The dashed curve represents the density $f(\epsilon, T)D(\epsilon)$ of filled orbitals at a finite temperature, but such that k_BT is small in comparison with ϵ_F. The shaded area represents the filled orbitals at absolute zero. The average energy is increased when the temperature is increased from 0 to T, for electrons are thermally excited from region 1 to region 2.

Using (14) and (16),

$$\boxed{\epsilon_F = \frac{\hbar^2}{2m}\left(\frac{3\pi^2 N}{V}\right)^{2/3}} \quad . \tag{17}$$

This relates the Fermi energy to the electron concentration N/V. The electron velocity v_F at the Fermi surface is

$$v_F = \left(\frac{\hbar k_F}{m}\right) = \left(\frac{\hbar}{m}\right)\left(\frac{3\pi^2 N}{V}\right)^{1/3} \quad . \tag{18}$$

Calculated values of k_F, v_F, and ϵ_F are given in Table 1 for selected metals; also given are values of the quantity T_F which is defined as ϵ_F/k_B. (The quantity T_F has nothing to do with the temperature of the electron gas!)

We now find an expression for the number of orbitals per unit energy range, $D(\epsilon)$, called the **density of states.**[1] We use (17) to obtain the total number of orbitals of energy $\leq\epsilon$:

$$N = \frac{V}{3\pi^2}\left(\frac{2m\epsilon}{\hbar^2}\right)^{3/2} \quad , \tag{19}$$

so that the density of states (Fig. 5) is

$$\boxed{D(\epsilon) \equiv \frac{dN}{d\epsilon} = \frac{V}{2\pi^2} \cdot \left(\frac{2m}{\hbar^2}\right)^{3/2} \cdot \epsilon^{1/2}} \quad . \tag{20}$$

[1]Strictly, $D(\epsilon)$ is the density of one-particle states, or density of orbitals.

This result may be expressed more simply by comparing (19) and (20) to obtain at ϵ

$$D(\epsilon) \equiv \frac{dN}{d\epsilon} = \frac{3N}{2\epsilon} \ . \tag{21}$$

Within a factor of the order of unity, the number of orbitals per unit energy range at the Fermi energy is the total number of conduction electrons divided by the Fermi energy, just as we would expect.

HEAT CAPACITY OF THE ELECTRON GAS

The question that caused the greatest difficulty in the early development of the electron theory of metals concerns the heat capacity of the conduction electrons. Classical statistical mechanics predicts that a free particle should have a heat capacity of $\frac{3}{2}k_B$, where k_B is the Boltzmann constant. If N atoms each give one valence electron to the electron gas, and the electrons are freely mobile, then the electronic contribution to the heat capacity should be $\frac{3}{2}Nk_B$, just as for the atoms of a monatomic gas. But the observed electronic contribution at room temperature is usually less than 0.01 of this value.

This important discrepancy distracted the early workers, such as Lorentz: How can the electrons participate in electrical conduction processes as if they were mobile, while not contributing to the heat capacity? The question was answered only upon the discovery of the Pauli exclusion principle and the Fermi distribution function. Fermi found the correct result and he wrote, "One recognizes that the specific heat vanishes at absolute zero and that at low temperatures it is proportional to the absolute temperature."

When we heat the specimen from absolute zero, not every electron gains an energy $\sim k_BT$ as expected classically, but only those electrons in orbitals within an energy range k_BT of the Fermi level are excited thermally, as in Fig. 5. This gives an immediate qualitative solution to the problem of the heat capacity of the conduction electron gas. If N is the total number of electrons, only a fraction of the order of T/T_F can be excited thermally at temperature T, because only these lie within an energy range of the order of k_BT of the top of the energy distribution.

Each of these NT/T_F electrons has a thermal energy of the order of k_BT. The total electronic thermal kinetic energy U is of the order of

$$U_{el} \approx (NT/T_F)k_BT \ . \tag{22}$$

The electronic heat capacity is given by

$$C_{el} = \partial U/\partial T \approx Nk_B(T/T_F) \tag{23}$$

and is directly proportional to T, in agreement with the experimental results discussed in the following section. At room temperature C_{el} is smaller

than the classical value $\frac{3}{2}Nk_B$ by a factor of the order of 0.01 or less, for $T_F \sim 5 \times 10^4$ K.

We now derive a quantitative expression for the electronic heat capacity valid at low temperatures $k_BT \ll \epsilon_F$. The increase $\Delta U \equiv U(T) - U(0)$ in the total energy (Fig. 5) of a system of N electrons when heated from 0 to T is

$$\Delta U = \int_0^\infty d\epsilon\, \epsilon D(\epsilon) f(\epsilon) - \int_0^{\epsilon_F} d\epsilon\, \epsilon D(\epsilon) \ . \tag{24}$$

Here $f(\epsilon)$ is the Fermi-Dirac function (5):

$$\boxed{f(\epsilon, T, \mu) = \frac{1}{\exp[(\epsilon - \mu)/k_BT + 1]}} \tag{24a}$$

and $D(\epsilon)$ is the number of orbitals per unit energy range. We multiply the identity

$$N = \int_0^\infty d\epsilon\, D(\epsilon) f(\epsilon) = \int_0^{\epsilon_F} d\epsilon\, D(\epsilon) \tag{25}$$

by ϵ_F to obtain

$$\left(\int_0^{\epsilon_F} + \int_{\epsilon_F}^\infty\right) d\epsilon\, \epsilon_F f(\epsilon) D(\epsilon) = \int_0^{\epsilon_F} d\epsilon\, \epsilon_F D(\epsilon) \ . \tag{26}$$

We use (26) to rewrite (24) as

$$\Delta U = \int_{\epsilon_F}^\infty d\epsilon(\epsilon - \epsilon_F) f(\epsilon) D(\epsilon) + \int_0^{\epsilon_F} d\epsilon(\epsilon_F - \epsilon)[1 - f(\epsilon)] D(\epsilon) \ . \tag{27}$$

The first integral on the right-hand side of (27) gives the energy needed to take electrons from ϵ_F to the orbitals of energy $\epsilon > \epsilon_F$, and the second integral gives the energy needed to bring the electrons to ϵ_F from orbitals below ϵ_F. Both contributions to the energy are positive.

The product $f(\epsilon)D(\epsilon)d\epsilon$ in the first integral of (27) is the number of electrons elevated to orbitals in the energy range $d\epsilon$ at an energy ϵ. The factor $[1 - f(\epsilon)]$ in the second integral is the probability that an electron has been removed from an orbital ϵ. The function ΔU is plotted in Fig. 6.

The heat capacity of the electron gas is found on differentiating ΔU with respect to T. The only temperature-dependent term in (27) is $f(\epsilon)$, whence we can group terms to obtain

$$C_{el} = \frac{dU}{dT} = \int_0^\infty d\epsilon(\epsilon - \epsilon_F)\frac{df}{dT} D(\epsilon) \ . \tag{28}$$

At the temperatures of interest in metals, $\tau/\epsilon_F < 0.01$, and we see from Fig. 3 that $(\epsilon - \epsilon_F)\, df/dT$ has large positive peaks at energies near ϵ_F. It is a

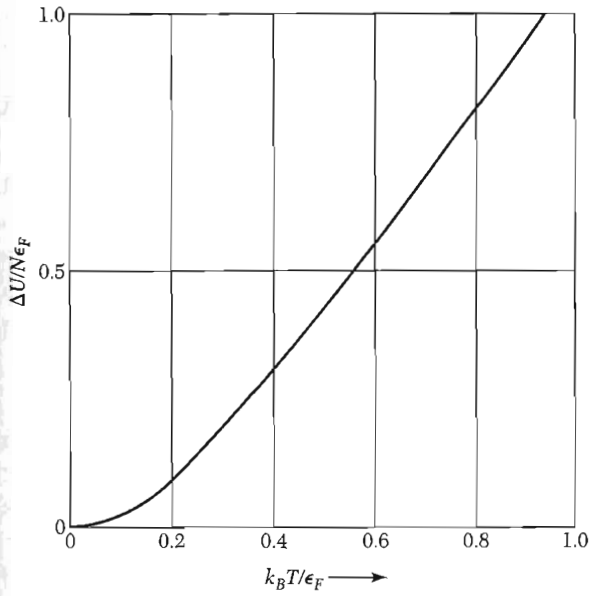


Figure 6 Temperature dependence of the energy of a noninteracting fermion gas in three dimensions. The energy is plotted in normalized form as $\Delta U/N\epsilon_F$, where N is the number of electrons. The temperature is plotted as k_BT/ϵ_F.

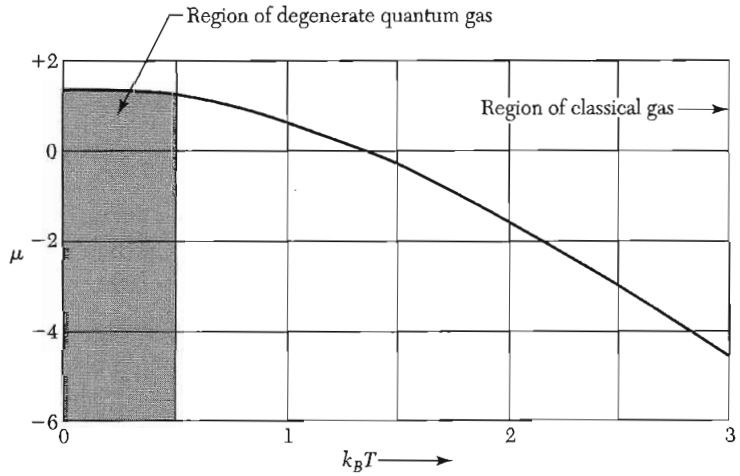


Figure 7 Plot of the chemical potential μ versus temperature as k_BT for a gas of noninteracting fermions in three dimensions. For convenience in plotting, the units of μ and k_BT are $0.763\epsilon_F$.

good approximation to evaluate the density of states $D(\epsilon)$ at ϵ_F and take it outside of the integral:

$$C_{el} \cong D(\epsilon_F) \int_0^\infty d\epsilon(\epsilon - \epsilon_F) \frac{df}{dT} . \tag{29}$$

Examination of the graphs in Figs. 7 and 8 of the variation of the chemical potential μ with T suggests that when $k_BT \ll \epsilon_F$ we ignore the temperature

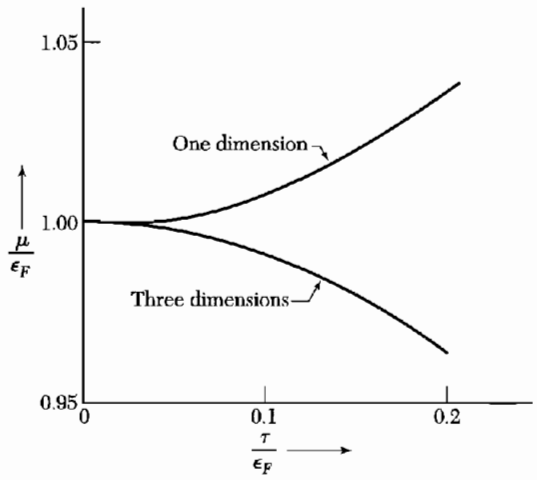


Figure 8 Variation with temperature of the chemical potential μ, for free electron Fermi gases in one and three dimensions. In common metals $\tau/\epsilon_F \simeq 0.01$ at room temperature, so that μ is closely equal to ϵ_F. These curves were calculated from series expansions of the integral for the number of particles in the system.

dependence of the chemical potential μ in the Fermi-Dirac distribution function and replace μ by the constant ϵ_F. We have then, with $\tau \equiv k_B T$,

$$\frac{df}{d\tau} = \frac{\epsilon - \epsilon_F}{\tau^2} \cdot \frac{\exp[(\epsilon - \epsilon_F)/\tau]}{\{\exp[(\epsilon - \epsilon_F)/\tau] + 1\}^2} \ . \tag{30}$$

We set

$$x \equiv (\epsilon - \epsilon_F)/\tau \ , \tag{31}$$

and it follows from (29) and (30) that

$$C_{el} = k_B^2 T \, D(\epsilon_F) \int_{-\epsilon_F/\tau}^{\infty} dx \, x^2 \frac{e^x}{(e^x + 1)^2} \ . \tag{32}$$

We may safely replace the lower limit by $-\infty$ because the factor e^x in the integrand is already negligible at $x = -\epsilon_F/\tau$ if we are concerned with low temperatures such that $\epsilon_F/\tau \sim 100$ or more. The integral in (32) then becomes

$$\int_{-\infty}^{\infty} dx \, x^2 \frac{e^x}{(e^x + 1)^2} = \frac{\pi^2}{3} \ , \tag{33}$$

whence the heat capacity of an electron gas is

$$C_{el} = \tfrac{1}{3}\pi^2 D(\epsilon_F) k_B^2 T \ . \tag{34}$$

From (21) we have

$$D(\epsilon_F) = 3N/2\epsilon_F = 3N/2k_B T_F \tag{35}$$

for a free electron gas, with $k_B T_F \equiv \epsilon_F$. Thus (34) becomes

$$C_{el} = \tfrac{1}{2}\pi^2 N k_B T/T_F \ . \tag{36}$$

Recall that although T_F is called the Fermi temperature, it is not the electron temperature, but only a convenient reference notation.

Experimental Heat Capacity of Metals

At temperatures much below both the Debye temperature θ and the Fermi temperature T_F, the heat capacity of metals may be written as the sum of electron and phonon contributions: $C = \gamma T + AT^3$, where γ and A are constants characteristic of the material. The electronic term is linear in T and is dominant at sufficiently low temperatures. It is convenient to exhibit the experimental values of C as a plot of C/T versus T^2:

$$C/T = \gamma + AT^2 \ , \tag{37}$$

for then the points should lie on a straight line with slope A and intercept γ. Such a plot for potassium is shown in Fig. 9. Observed values of γ, called the Sommerfeld parameter, are given in Table 2.

The observed values of the coefficient γ are of the expected magnitude, but often do not agree very closely with the value calculated for free electrons of mass m by use of (17) and (34). It is common practice to express the ratio of the observed to the free electron values of the electronic heat capacity as a ratio of a **thermal effective mass** m_{th} to the electron mass m, where m_{th} is defined by the relation

$$\frac{m_{\text{th}}}{m} \equiv \frac{\gamma(\text{observed})}{\gamma(\text{free})} \ . \tag{38}$$

This form arises in a natural way because ϵ_F is inversely proportional to the mass of the electron, whence $\gamma \propto m$. Values of the ratio are given in Table 2. The departure from unity involves three separate effects:

- The interaction of the conduction electrons with the periodic potential of the rigid crystal lattice. The effective mass of an electron in this potential is called the band effective mass.

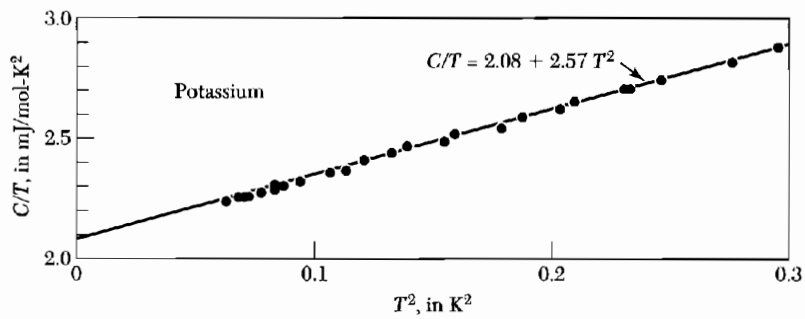


Figure 9 Experimental heat capacity values for potassium, plotted as C/T versus T^2. (After W. H. Lien and N. E. Phillips.)

Table 2 Experimental and free electron values of electronic heat capacity constant γ of metals

(From compilations kindly furnished by N. Phillips and N. Pearlman. The thermal effective mass is defined by Eq. (38).)

Observed γ in mJ mol^{-1} K^{-2}.
Calculated free electron γ in mJ mol^{-1} K^{-2}.
m_{th}/m = (observed γ)/(free electron γ).

Li	Be											B	C	N
1.63	0.17													
0.749	0.500													
2.18	0.34													
Na	Mg											Al	Si	P
1.38	1.3											1.35		
1.094	0.992											0.912		
1.26	1.3											1.48		
K	Ca	Sc	Ti	V	Cr	Mn(γ)	Fe	Co	Ni	Cu	Zn	Ga	Ge	As
2.08	2.9	10.7	3.35	9.26	1.40	9.20	4.98	4.73	7.02	0.695	0.64	0.596		0.19
1.668	1.511									0.505	0.753	1.025		
1.25	1.9									1.38	0.85	0.58		
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn (w)	Sb
2.41	3.6	10.2	2.80	7.79	2.0	—	3.3	4.9	9.42	0.646	0.688	1.69	1.78	0.11
1.911	1.790									0.645	0.948	1.233	1.410	
1.26	2.0									1.00	0.73	1.37	1.26	
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg(α)	Tl	Pb	Bi
3.20	2.7	10.	2.16	5.9	1.3	2.3	2.4	3.1	6.8	0.729	1.79	1.47	2.98	0.008
2.238	1.937									0.642	0.952	1.29	1.509	
1.43	1.4									1.14	1.88	1.14	1.97	

- The interaction of the conduction electrons with phonons. An electron tends to polarize or distort the lattice in its neighborhood, so that the moving electron tries to drag nearby ions along, thereby increasing the effective mass of the electron.
- The interaction of the conduction electrons with themselves. A moving electron causes an inertial reaction in the surrounding electron gas, thereby increasing the effective mass of the electron.

Heavy Fermions. Several metallic compounds have been discovered that have enormous values, two or three orders of magnitude higher than usual, of the electronic heat capacity constant γ. The heavy fermion compounds include UBe_{13}, $CeAl_3$, and $CeCu_2Si_2$. It has been suggested that f electrons in these compounds may have inertial masses as high as 1000 m, because of the weak overlap of wavefunctions of f electrons on neighboring ions (see Chapter 9, "tight binding").

ELECTRICAL CONDUCTIVITY AND OHM'S LAW

The momentum of a free electron is related to the wavevector by $m\mathbf{v} = \hbar\mathbf{k}$. In an electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$ the force $\mathbf{F}$ on an electron of charge $-e$ is $-e[\mathbf{E} + (1/c)\mathbf{v} \times \mathbf{B}]$, so that Newton's second law of motion becomes

$$\text{(CGS)} \qquad \boxed{\mathbf{F} = m\frac{d\mathbf{v}}{dt} = \hbar\frac{d\mathbf{k}}{dt} = -e\left(\mathbf{E} + \frac{1}{c}\mathbf{v} \times \mathbf{B}\right) .} \qquad (39)$$

In the absence of collisions the Fermi sphere (Fig. 10) moves in $\mathbf{k}$ space at a uniform rate by a constant applied electric field. We integrate (39) with $\mathbf{B} = 0$ to obtain

$$\mathbf{k}(t) - \mathbf{k}(0) = -e\mathbf{E}t/\hbar \ . \qquad (40)$$

If the force $\mathbf{F} = -e\mathbf{E}$ is applied at time $t = 0$ to an electron gas that fills the Fermi sphere centered at the origin of $\mathbf{k}$ space, then at a later time t the sphere will be displaced to a new center at

$$\delta\mathbf{k} = -e\mathbf{E}t/\hbar \ . \qquad (41)$$

Notice that the Fermi sphere is displaced as a whole because every electron is displaced by the same $\delta\mathbf{k}$.

Because of collisions of electrons with impurities, lattice imperfections, and phonons, the displaced sphere may be maintained in a steady state in an electric field. If the collision time is τ, the displacement of the Fermi sphere in the steady state is given by (41) with $t = \tau$. The incremental velocity is $\mathbf{v} = \hbar\delta\mathbf{k}/m = -e\mathbf{E}\tau/m$. If in a constant electric field $\mathbf{E}$ there are n electrons of charge $q = -e$ per unit volume, the electric current density is

$$\mathbf{j} = nq\mathbf{v} = ne^2\tau\mathbf{E}/m \ . \qquad (42)$$

This is **Ohm's law**.

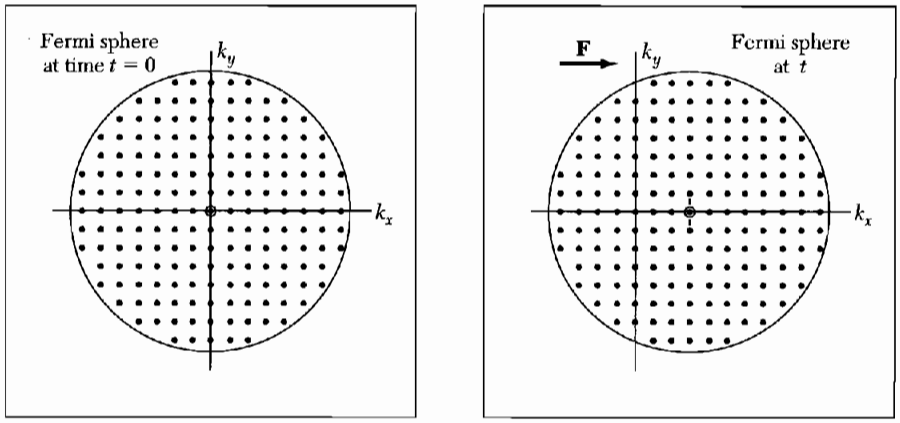


Figure 10 (a) The Fermi sphere encloses the occupied electron orbitals in **k** space in the ground state of the electron gas. The net momentum is zero, because for every orbital **k** there is an occupied orbital at $-\mathbf{k}$. (b) Under the influence of a constant force **F** acting for a time interval t every orbital has its **k** vector increased by $\delta\mathbf{k} = \mathbf{F}t/\hbar$. This is equivalent to a displacement of the whole Fermi sphere by $\delta\mathbf{k}$. The total momentum is $N\hbar\delta\mathbf{k}$, if there are N electrons present. The application of the force increases the energy of the system by $N(\hbar\delta\mathbf{k})^2/2m$.

The electrical conductivity σ is defined by $\mathbf{j} = \sigma\mathbf{E}$, so by (42)

$$\boxed{\sigma = \frac{ne^2\tau}{m}} \quad . \tag{43}$$

The electrical resistivity ρ is defined as the reciprocal of the conductivity, so that

$$\rho = m/ne^2\tau \quad . \tag{44}$$

Values of the electrical conductivity and resistivity of the elements are given in Table 3. In Gaussian units σ has the dimensions of frequency.

It is easy to understand the result (43) for the conductivity of a Fermi gas. We expect the charge transported to be proportional to the charge density ne; the factor e/m enters (43) because the acceleration in a given electric field is proportional to e and inversely proportional to the mass m. The time τ describes the free time during which the field acts on the carrier. Closely the same result for the electrical conductivity is obtained for a classical (Maxwellian) gas of electrons, as realized at low carrier concentration in many semiconductor problems.

Experimental Electrical Resistivity of Metals

The electrical resistivity of most metals is dominated at room temperature (300 K) by collisions of the conduction electrons with lattice phonons and at

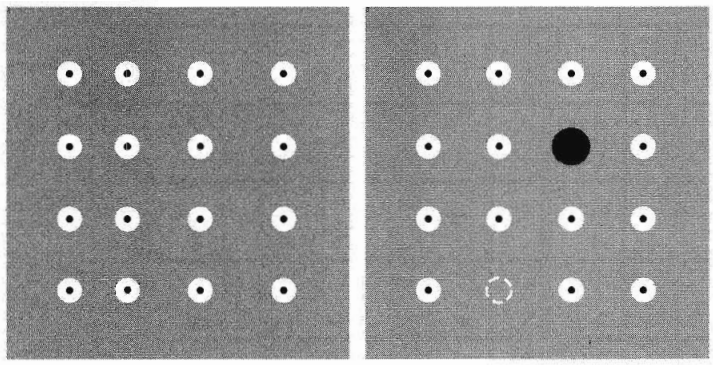
Table 3 Electrical conductivity and resistivity of metals at 295 K

(Resistivity values as given by G. T. Meaden, *Electrical resistance of metals*, Plenum, 1965; residual resistivities have been subtracted.)

Li	Be											B	C	N	O	F	Ne
1.07	3.08																
9.32	3.25																
Na	Mg											Al	Si	P	S	Cl	Ar
2.11	2.33	Conductivity in units of 10^5 $(ohm\text{-}cm)^{-1}$										3.65					
4.75	4.30	Resistivity in units of 10^{-6} ohm-cm.										2.74					
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
1.39	2.78	0.21	0.23	0.50	0.78	0.072	1.02	1.72	1.43	5.88	1.69	0.67					
7.19	3.6	46.8	43.1	19.9	12.9	139.	9.8	5.8	7.0	1.70	5.92	14.85					
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn (w)	Sb	Te	I	Xe
0.80	0.47	0.17	0.24	0.69	1.89	~0.7	1.35	2.08	0.95	6.21	1.38	1.14	0.91	0.24			
12.5	21.5	58.5	42.4	14.5	5.3	~14.	7.4	4.8	10.5	1.61	7.27	8.75	11.0	41.3			
Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg liq.	Tl	Pb	Bi	Po	At	Rn
0.50	0.26	0.13	0.33	0.76	1.89	0.54	1.10	1.96	0.96	4.55	0.10	0.61	0.48	0.086	0.22		
20.0	39.	79.	30.6	13.1	5.3	18.6	9.1	5.1	10.4	2.20	95.9	16.4	21.0	116.	46.		
Fr	Ra	Ac															

Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
0.12	0.15	0.17		0.10	0.11	0.070	0.090	0.11	0.13	0.12	0.16	0.38	0.19
81.	67.	59.		99.	89.	134.	111.	90.0	77.7	81.	62.	26.4	53.

Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr
0.66		0.39	0.085	0.070									
15.2		25.7	118.	143.									



(a) (b)

Figure 11 Electrical resistivity in most metals arises from collisions of electrons with irregularities in the lattice, as in (a) by phonons and in (b) by impurities and vacant lattice sites.

liquid helium temperature (4 K) by collisions with impurity atoms and mechanical imperfections in the lattice (Fig. 11). The rates of these collisions are often independent to a good approximation, so that if the electric field were switched off the momentum distribution would relax back to its ground state with the net relaxation rate

$$\frac{1}{\tau} = \frac{1}{\tau_L} + \frac{1}{\tau_i} , \tag{45}$$

where τ_L and τ_i are the collision times for scattering by phonons and by imperfections, respectively.

The net resistivity is given by

$$\rho = \rho_L + \rho_i , \tag{46}$$

where ρ_L is the resistivity caused by the thermal phonons, and ρ_i is the resistivity caused by scattering of the electron waves by static defects that disturb the periodicity of the lattice. Often ρ_L is independent of the number of defects when their concentration is small, and often ρ_i is independent of temperature. This empirical observation expresses **Matthiessen's rule,** which is convenient in analyzing experimental data (Fig. 12).

The residual resistivity, $\rho_i(0)$, is the extrapolated resistivity at 0 K because ρ_L vanishes as $T \to 0$. The lattice resistivity, $\rho_L(T) = \rho - \rho_i(0)$, is the same for different specimens of a metal, even though $\rho_i(0)$ may itself vary widely. The **resistivity ratio** of a specimen is usually defined as the ratio of its resistivity at room temperature to its residual resistivity. It is a convenient approximate indicator of sample purity: for many materials an impurity in solid solution creates a residual resistivity of about 1 μohm-cm (1×10^{-6} ohm-cm) per atomic percent of impurity. A copper specimen with a resistivity ratio of 1000 will have a residual resistivity of 1.7×10^{-3} μohm-cm, corresponding to an

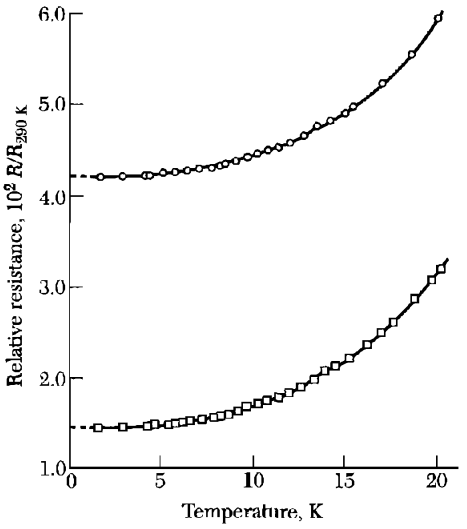


Figure 12 Resistance of potassium below 20 K, as measured on two specimens by D. MacDonald and K. Mendelssohn. The different intercepts at 0 K are attributed to different concentrations of impurities and static imperfections in the two specimens.

impurity concentration of about 20 ppm. In exceptionally pure specimens the resistivity ratio may be as high as 10^6, whereas in some alloys (e.g., manganin) it is as low as 1.1.

It is possible to obtain crystals of copper so pure that their conductivity at liquid helium temperatures (4 K) is nearly 10^5 times that at room temperature; for these conditions $\tau \approx 2 \times 10^{-9}$ s at 4 K. The mean free path ℓ of a conduction electron is defined as

$$\ell = v_F \tau \ , \tag{47}$$

where v_F is the velocity at the Fermi surface, because all collisions involve only electrons near the Fermi surface. From Table 1 we have $v_F = 1.57 \times 10^8$ cm s^{-1} for Cu, thus the mean free path is $\ell(4\ \text{K}) \approx 0.3$ cm. Mean free paths as long as 10 cm have been observed in very pure metals in the liquid helium temperature range.

The temperature-dependent part of the electrical resistivity is proportional to the rate at which an electron collides with thermal phonons and thermal electrons. The collision rate with phonons is proportional to the concentration of thermal phonons. One simple limit is at temperatures over the Debye temperature θ: here the phonon concentration is proportional to the temperature T, so that $\rho \propto T$ for $T > \theta$. A sketch of the theory is given in Appendix J.

Umklapp Scattering

Umklapp scattering of electrons by phonons (Chapter 5) accounts for most of the electrical resistivity of metals at low temperatures. These are electron-phonon scattering processes in which a reciprocal lattice vector **G** is involved, so that electron momentum change in the process may be much larger

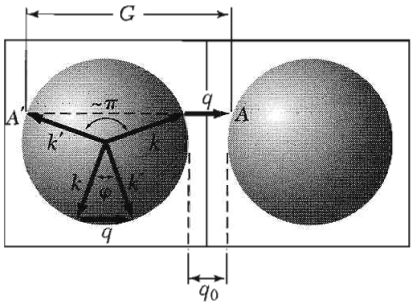


Figure 13 Two Fermi spheres in adjacent zones: a construction to show the role of phonon umklapp processes in electrical resistivity.

than in a normal electron-phonon scattering process at low temperatures. (In an umklapp process the wavevector of one particle may be "flipped over.")

Consider a section perpendicular to [100] through two adjacent Brillouin zones in bcc potassium, with the equivalent Fermi spheres inscribed within each (Fig. 13). The lower half of the figure shows the normal electron-phonon collision $\mathbf{k}' = \mathbf{k} + \mathbf{q}$, while the upper half shows a possible scattering process $\mathbf{k}' = \mathbf{k} + \mathbf{q} + \mathbf{G}$ involving the same phonon and terminating outside the first Brillouin zone, at the point A. This point is exactly equivalent to the point A' inside the original zone, where AA' is a reciprocal lattice vector $\mathbf{G}$. This scattering is an umklapp process, in analogy to phonons. Such collisions are strong scatterers because the scattering angle can be close to π.

When the Fermi surface does not intersect the zone boundary, there is some minimum phonon wavevector q_0 for umklapp scattering. At low enough temperatures the number of phonons available for umklapp scattering falls as $\exp(-\theta_U/T)$, where θ_U is a characteristic temperature calculable from the geometry of the Fermi surface inside the Brillouin zone. For a spherical Fermi surface with one electron orbital per atom inside the bcc Brillouin zone, one shows by geometry that $q_0 = 0.267\, k_F$.

The experimental data (Fig. 12) for potassium have the expected exponential form with $\theta_U = 23$ K compared with the Debye $\theta = 91$ K. At the very lowest temperatures (below about 2 K in potassium) the number of umklapp processes is negligible and the lattice resistivity is then caused only by small angle scattering, which is the normal (not umklapp) scattering.

MOTION IN MAGNETIC FIELDS

By the arguments of (39) and (41) we are led to the equation of motion for the displacement $\delta\mathbf{k}$ of a Fermi sphere of particles acted on by a force $\mathbf{F}$ and by friction as represented by collisions at a rate $1/\tau$:

$$\hbar\left(\frac{d}{dt} + \frac{1}{\tau}\right)\delta\mathbf{k} = \mathbf{F} \ . \tag{48}$$

The free particle acceleration term is $(\hbar d/dt)\,\delta\mathbf{k}$ and the effect of collisions (the friction) is represented by $\hbar\delta\mathbf{k}/\tau$, where τ is the collision time.

Consider now the motion of the system in a uniform magnetic field **B**. The Lorentz force on an electron is

(CGS) $$\mathbf{F} = -e\left(\mathbf{E} + \frac{1}{c}\mathbf{v}\times\mathbf{B}\right); \tag{49}$$

(SI) $$\mathbf{F} = -e(\mathbf{E} + \mathbf{v}\times\mathbf{B})$$

If $m\mathbf{v} = \hbar\delta\mathbf{k}$, then the equation of motion is

(CGS) $$m\left(\frac{d}{dt} + \frac{1}{\tau}\right)\mathbf{v} = -e\left(\mathbf{E} + \frac{1}{c}\mathbf{v}\times\mathbf{B}\right). \tag{50}$$

An important situation is the following: let a static magnetic field **B** lie along the z axis. Then the component equations of motion are

(CGS) $$m\left(\frac{d}{dt} + \frac{1}{\tau}\right)v_x = -e\left(E_x + \frac{B}{c}v_y\right);$$

$$m\left(\frac{d}{dt} + \frac{1}{\tau}\right)v_y = -e\left(E_y - \frac{B}{c}v_x\right); \tag{51}$$

$$m\left(\frac{d}{dt} + \frac{1}{\tau}\right)v_z = -eE_z.$$

The results in SI are obtained by replacing c by 1.

In the steady state in a static electric field the time derivatives are zero, so that the drift velocity is

$$v_x = -\frac{e\tau}{m}E_x - \omega_c\tau v_y; \qquad v_y = -\frac{e\tau}{m}E_y + \omega_c\tau v_x; \qquad v_z = -\frac{e\tau}{m}E_z, \tag{52}$$

where $\omega_c = eB/mc$ is the **cyclotron frequency,** as discussed in Chapter 8 for cyclotron resonance in semiconductors.

Hall Effect

The Hall field is the electric field developed across two faces of a conductor, in the direction $\mathbf{j}\times\mathbf{B}$, when a current **j** flows across a magnetic field **B**. Consider a rod-shaped specimen in a longitudinal electric field E_x and a transverse magnetic field, as in Fig. 14. If current cannot flow out of the rod in the y direction we must have $\delta v_y = 0$. From (52) this is possible only if there is a transverse electric field

(CGS) $$E_y = -\omega_c\tau E_x = -\frac{eB\tau}{mc}E_x; \tag{53}$$

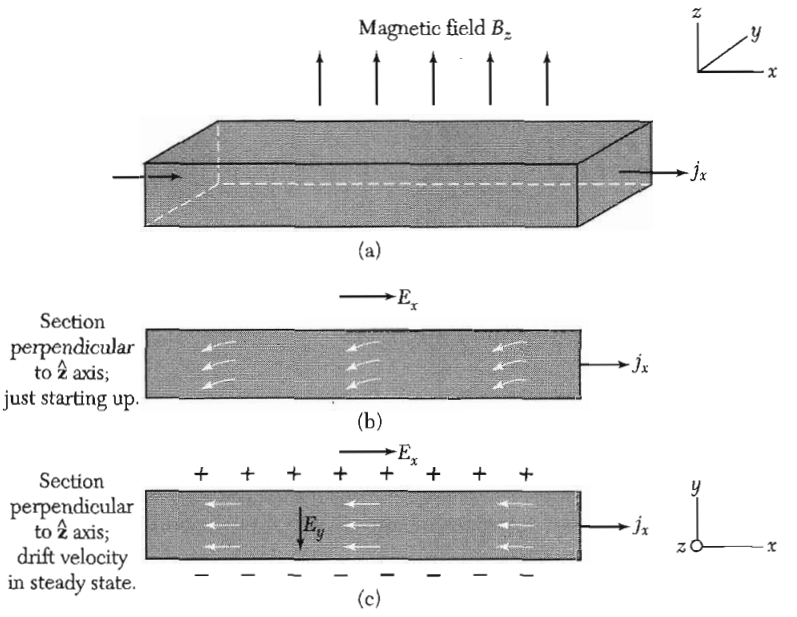


Figure 14 The standard geometry for the Hall effect: a rod-shaped specimen of rectangular cross-section is placed in a magnetic field B_z, as in (a). An electric field E_x applied across the end electrodes causes an electric current density j_x to flow down the rod. The drift velocity of the negatively-charged electrons immediately after the electric field is applied as shown in (b). The deflection in the $-y$ direction is caused by the magnetic field. Electrons accumulate on one face of the rod and a positive ion excess is established on the opposite face until, as in (c), the transverse electric field (Hall field) just cancels the Lorentz force due to the magnetic field.

(SI) $$E_y = -\omega_c \tau E_x = -\frac{eB\tau}{m} E_x \ .$$

The quantity defined by

$$\boxed{R_H = \frac{E_y}{j_x B}} \tag{54}$$

is called the **Hall coefficient**. To evaluate it on our simple model we use $j_x = ne^2\tau E_x/m$ and obtain

(CGS) $$R_H = -\frac{eB\tau E_x/mc}{ne^2\tau E_x B/m} = -\frac{1}{nec} \ ; \tag{55}$$

(SI) $$R_H = -\frac{1}{ne} \ .$$

This is negative for free electrons, for e is positive by definition.

Table 4 Comparison of observed Hall coefficients with free electron theory

[The experimental values of R_H as obtained by conventional methods are summarized from data at room temperature presented in the Landolt-Bornstein tables. The values obtained by the helicon wave method at 4 K are by J. M. Goodman. The values of the carrier concentration n are from Table 1.4 except for Na, K, Al, In, where Goodman's values are used. To convert the value of R_H in CGS units to the value in volt-cm/amp-gauss, multiply by 9×10^{11}; to convert R_H in CGS to m^3/coulomb, multiply by 9×10^{13}.]

Metal	Method	Experimental R_H, in 10^{-24} CGS units	Assumed carriers per atom	Calculated $-1/nec$, in 10^{-24} CGS units
Li	conv.	−1.89	1 electron	−1.48
Na	helicon	−2.619	1 electron	−2.603
	conv.	−2.3		
K	helicon	−4.946	1 electron	−4.944
	conv.	−4.7		
Rb	conv.	−5.6	1 electron	−6.04
Cu	conv.	−0.6	1 electron	−0.82
Ag	conv.	−1.0	1 electron	−1.19
Au	conv.	−0.8	1 electron	−1.18
Be	conv.	+2.7	—	—
Mg	conv.	−0.92	—	—
Al	helicon	+1.136	1 hole	+1.135
In	helicon	+1.774	1 hole	+1.780
As	conv.	+50.	—	—
Sb	conv.	−22.	—	—
Bi	conv.	−6000.	—	—

The lower the carrier concentration, the greater the magnitude of the Hall coefficient. Measuring R_H is an important way of measuring the carrier concentration. *Note:* The symbol R_H denotes the Hall coefficient (54), but the same symbol is sometimes used with a different meaning, that of Hall resistance in two-dimensional problems.

The simple result (55) follows from the assumption that all relaxation times are equal, independent of the velocity of the electron. A numerical factor of order unity enters if the relaxation time is a function of the velocity. The expression becomes somewhat more complicated if both electrons and holes contribute to the conductivity.

In Table 4 observed values of the Hall coefficient are compared with values calculated from the carrier concentration. The most accurate measurements are made by the method of helicon resonance which is treated as a problem in Chapter 14.

The accurate values for sodium and potassium are in excellent agreement with values calculated for one conduction electron per atom, using (55).

Notice, however, the experimental values for the trivalent elements aluminum and indium: these agree with values calculated for one *positive* charge carrier per atom and thus disagree in magnitude and sign with values calculated for the expected three negative charge carriers.

The problem of an apparent positive sign for the charge carriers arises also for Be and As, as seen in the table. The anomaly of the sign was explained by Peierls (1928). The motion of carriers of apparent positive sign, which Heisenberg later called "holes," cannot be explained by a free electron gas, but finds a natural explanation in terms of the energy band theory to be developed in Chapters 7–9. Band theory also accounts for the occurrence of very large values of the Hall coefficient, as for As, Sb, and Bi.

THERMAL CONDUCTIVITY OF METALS

In Chapter 5 we found an expression $K = \frac{1}{3}Cv\ell$ for the thermal conductivity of particles of velocity v, heat capacity C per unit volume, and mean free path ℓ. The thermal conductivity of a Fermi gas follows from (36) for the heat capacity, and with $\epsilon_F = \frac{1}{2}mv_F^2$:

$$K_{el} = \frac{\pi^2}{3} \cdot \frac{nk_B^2T}{mv_F^2} \cdot v_F \cdot \ell = \frac{\pi^2 nk_B^2 T\tau}{3m} \ . \tag{56}$$

Here $\ell = v_F\tau$; the electron concentration is n, and τ is the collision time.

Do the electrons or the phonons carry the greater part of the heat current in a metal? In pure metals the electronic contribution is dominant at all temperatures. In impure metals or in disordered alloys, the electron mean free path is reduced by collisions with impurities, and the phonon contribution may be comparable with the electronic contribution.

Ratio of Thermal to Electrical Conductivity

The **Wiedemann-Franz law** states that for metals at not too low temperatures the ratio of the thermal conductivity to the electrical conductivity is directly proportional to the temperature, with the value of the constant of proportionality independent of the particular metal. This result was important in the history of the theory of metals, for it supported the picture of an electron gas as the carrier of charge and energy. It can be explained by using (43) for σ and (56) for K:

$$\frac{K}{\sigma} = \frac{\pi^2 k_B^2 Tn\tau/3m}{ne\tau^2/m} = \frac{\pi^2}{3}\left(\frac{k_B}{e}\right)^2 T \ . \tag{57}$$

Table 5 Experimental Lorenz numbers

	$L \times 10^8$ watt-ohm/deg^2			$L \times 10^8$ watt-ohm/deg^2	
Metal	0°C	100°C	Metal	0°C	100°C
Ag	2.31	2.37	Pb	2.47	2.56
Au	2.35	2.40	Pt	2.51	2.60
Cd	2.42	2.43	Su	2.52	2.49
Cu	2.23	2.33	W	3.04	3.20
Mo	2.61	2.79	Zn	2.31	2.33

The **Lorenz number** L is defined as

$$L = K/\sigma T \ , \tag{58}$$

and according to (57) should have the value

$$L = \frac{\pi^2}{3}\left(\frac{k_B}{e}\right)^2 = 2.72 \times 10^{-13}\ (\text{erg/esu-deg})^2$$
$$= 2.45 \times 10^{-8}\ \text{watt-ohm/deg}^2 \ . \tag{59}$$

This remarkable result involves neither n nor m. Experimental values of L at 0°C and at 100°C as given in Table 5 are in good agreement with (59).

Problems

1. ***Kinetic energy of electron gas.*** Show that the kinetic energy of a three-dimensional gas of N free electrons at 0 K is

$$U_0 = \tfrac{3}{5} N\epsilon_F \ . \tag{60}$$

2. ***Pressure and bulk modulus of an electron gas.*** (a) Derive a relation connecting the pressure and volume of an electron gas at 0 K. Hint: Use the result of Problem 1 and the relation between ϵ_F and electron concentration. The result may be written as $p = \frac{2}{3}(U_0/V)$. (b) Show that the bulk modulus $B = -V(\partial p/\partial V)$ of an electron gas at 0 K is $B = 5p/3 = 10U_0/9V$. (c) Estimate for potassium, using Table 1, the value of the electron gas contribution to B.

3. ***Chemical potential in two dimensions.*** Show that the chemical potential of a Fermi gas in two dimensions is given by:

$$\mu(T) = k_B T \ln\left[\exp(\pi n\hbar^2/mk_B T) - 1\right] \ , \tag{61}$$

for n electrons per unit area. Note: The density of orbitals of a free electron gas in two dimensions is independent of energy: $D(\epsilon) = m/\pi\hbar^2$, per unit area of specimen.

4. ***Fermi gases in astrophysics.*** (a) Given $M_\odot = 2 \times 10^{33}$ g for the mass of the Sun, estimate the number of electrons in the Sun. In a white dwarf star this number of electrons may be ionized and contained in a sphere of radius 2×10^9 cm; find the Fermi energy of the electrons in electron volts. (b) The energy of an electron in the relativistic limit $\epsilon \gg mc^2$ is related to the wavevector as $\epsilon \cong pc = \hbar kc$. Show that the Fermi energy in this limit is $\epsilon_F \approx \hbar c\,(N/V)^{1/3}$, roughly. (c) If the above number of electrons were contained within a pulsar of radius 10 km, show that the Fermi energy would be $\approx 10^8$ eV. This value explains why pulsars are believed to be composed largely of neutrons rather than of protons and electrons, for the energy release in the reaction $n \rightarrow p + e^-$ is only 0.8×10^6 eV, which is not large enough to enable many electrons to form a Fermi sea. The neutron decay proceeds only until the electron concentration builds up enough to create a Fermi level of 0.8×10^6 eV, at which point the neutron, proton, and electron concentrations are in equilibrium.

5. ***Liquid He^3.*** The atom He^3 has spin $\frac{1}{2}$ and is a fermion. The density of liquid He^3 is 0.081 g cm^{-3} near absolute zero. Calculate the Fermi energy ϵ_F and the Fermi temperature T_F.

6. ***Frequency dependence of the electrical conductivity.*** Use the equation $m(dv/dt + v/\tau) = -eE$ for the electron drift velocity v to show that the conductivity at frequency ω is

$$\sigma(\omega) = \sigma(0)\left(\frac{1 + i\omega\tau}{1 + (\omega\tau)^2}\right), \tag{62}$$

where $\sigma(0) = ne^2\tau/m$.

*7. ***Dynamic magnetoconductivity tensor for free electrons.*** A metal with a concentration n of free electrons of charge $-e$ is in a static magnetic field $B\hat{\mathbf{z}}$. The electric current density in the xy plane is related to the electric field by

$$j_x = \sigma_{xx}E_x + \sigma_{xy}E_y \; ; \qquad j_y = \sigma_{yx}E_x + \sigma_{yy}E_y \; .$$

Assume that the frequency $\omega \gg \omega_c$ and $\omega \gg 1/\tau$, where $\omega_c = eB/mc$ and τ is the collision time. (a) Solve the drift velocity equation (51) to find the components of the magnetoconductivity tensor:

$$\sigma_{xx} = \sigma_{yy} = i\omega_p^2/4\pi\omega \; ; \qquad \sigma_{xy} = -\sigma_{yx} = \omega_c\omega_p^2/4\pi\omega^2 \; ,$$

where $\omega_p^2 = 4\pi ne^2/m$. (b) Note from a Maxwell equation that the dielectric function tensor of the medium is related to the conductivity tensor as $\boldsymbol{\epsilon} = \mathbf{1} + i(4\pi/\omega)\boldsymbol{\sigma}$. Consider an electromagnetic wave with wavevector $\mathbf{k} = k\hat{\mathbf{z}}$. Show that the dispersion relation for this wave in the medium is

$$c^2k^2 = \omega^2 - \omega_p^2 \pm \omega_c\omega_p^2/\omega \, . \tag{63}$$

At a given frequency there are two modes of propagation with different wavevectors and different velocities. The two modes correspond to circularly polarized

*This problem is somewhat difficult.

waves. Because a linearly polarized wave can be decomposed into two circularly polarized waves, it follows that the plane of polarization of a linearly polarized wave will be rotated by the magnetic field.

*8. ***Cohesive energy of free electron Fermi gas.*** We define the dimensionless length r_s, as r_0/a_H, where r_0 is the radius of a sphere that contains one electron, and a_H is the bohr radius $\hbar^2/e^2m$. (a) Show that the average kinetic energy per electron in a free electron Fermi gas at 0 K is $2.21/r_s^2$, where the energy is expressed in rydbergs, with 1 Ry $= me^4/2\hbar^2$. (b) Show that the coulomb energy of a point positive charge e interacting with the uniform electron distribution of one electron in the volume of radius r_0 is $-3e^2/2r_0$, or $-3/r_s$ in rydbergs. (c) Show that the coulomb self-energy of the electron distribution in the sphere is $3e^2/5r_0$, or $6/5r_s$ in rydbergs. (d) The sum of (b) and (c) gives $-1.80/r_s$ for the total coulomb energy per electron. Show that the equilibrium value of r_s is 2.45. Will such a metal be stable with respect to separated H atoms?

9. ***Static magnetoconductivity tensor.*** For the drift velocity theory of (51), show that the static current density can be written in matrix form as

$$\begin{pmatrix} j_x \\ j_y \\ j_z \end{pmatrix} = \frac{\sigma_0}{1+(\omega_c\tau)^2} \begin{pmatrix} 1 & -\omega_c\tau & 0 \\ \omega_c\tau & 1 & 0 \\ 0 & 0 & 1+(\omega_c\tau)^2 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} . \tag{64}$$

In the high magnetic field limit of $\omega_c\tau \gg 1$, show that

$$\sigma_{yx} = nec/B = -\sigma_{xy} . \tag{65}$$

In this limit $\sigma_{xx} = 0$, to order $1/\omega_c\tau$. The quantity σ_{yx} is called the **Hall conductivity**.

10. ***Maximum surface resistance.*** Consider a square sheet of side L, thickness d, and electrical resistivity ρ. The resistance measured between opposite edges of the sheet is called the surface resistance: $R_{sq} = \rho L/Ld = \rho/d$, which is independent of the area L^2 of the sheet. (R_{sq} is called the resistance per square and is expressed in ohms per square, because ρ/d has the dimensions of ohms.) If we express ρ by (44), then $R_{sq} = m/nde^2\tau$. Suppose now that the minimum value of the collision time is determined by scattering from the surfaces of the sheet, so that $\tau \approx d/v_F$, where v_F is the Fermi velocity. Thus the maximum surface resistivity is $R_{sq} \approx mv_F/nd^2e^2$. Show for a monatomic metal sheet one atom in thickness that $R_{sq} \approx \hbar/e^2 = 4.1\ \text{k}\Omega$.

*This problem is somewhat difficult.

7

Energy Bands

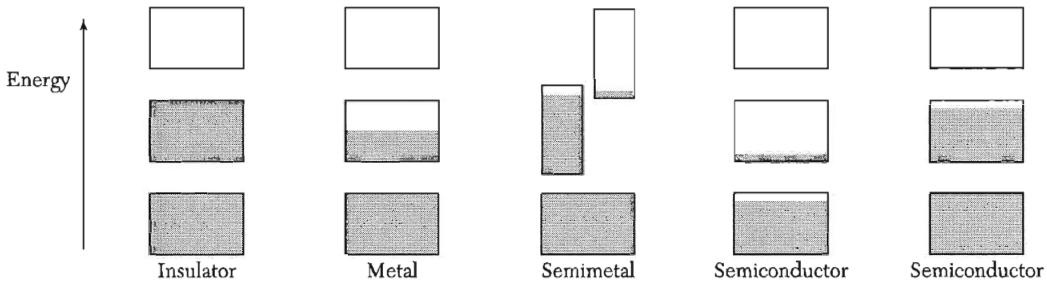


Figure 1 Schematic electron occupancy of allowed energy bands for an insulator, metal, semimetal, and semiconductor. The vertical extent of the boxes indicates the allowed energy regions; the shaded areas indicate the regions filled with electrons. In a **semimetal** (such as bismuth) one band is almost filled and another band is nearly empty at absolute zero, but a pure **semiconductor** (such as silicon) becomes an insulator at absolute zero. The left of the two semiconductors shown is at a finite temperature, with carriers excited thermally. The other semiconductor is electron-deficient because of impurities.

CHAPTER 7: ENERGY BANDS

When I started to think about it, I felt that the main problem was to explain how the electrons could sneak by all the ions in a metal. . . . By straight Fourier analysis I found to my delight that the wave differed from the plane wave of free electrons only by a periodic modulation.

F. Bloch

The free electron model of metals gives us good insight into the heat capacity, thermal conductivity, electrical conductivity, magnetic susceptibility, and electrodynamics of metals. But the model fails to help us with other large questions: the distinction between metals, semimetals, semiconductors, and insulators; the occurrence of positive values of the Hall coefficient; the relation of conduction electrons in the metal to the valence electrons of free atoms; and many transport properties, particularly magnetotransport. We need a less naïve theory, and fortunately it turns out that almost any simple attempt to improve upon the free electron model is enormously profitable.

The difference between a good conductor and a good insulator is striking. The electrical resistivity of a pure metal may be as low as 10^{-10} ohm-cm at a temperature of 1 K, apart from the possibility of superconductivity. The resistivity of a good insulator may be as high as 10^{22} ohm-cm. This range of 10^{32} may be the widest of any common physical property of solids.

Every solid contains electrons. The important question for electrical conductivity is how the electrons respond to an applied electric field. We shall see that electrons in crystals are arranged in **energy bands** (Fig. 1) separated by regions in energy for which no wavelike electron orbitals exist. Such forbidden regions are called **energy gaps** or **band gaps**, and result from the interaction of the conduction electron waves with the ion cores of the crystal.

The crystal behaves as an insulator if the allowed energy bands are either filled or empty, for then no electrons can move in an electric field. The crystal behaves as a metal if one or more bands are partly filled, say between 10 and 90 percent filled. The crystal is a semiconductor or a semimetal if one or two bands are slightly filled or slightly empty.

To understand the difference between insulators and conductors, we must extend the free electron model to take account of the periodic lattice of the solid. The possibility of a band gap is the most important new property that emerges.

We shall encounter other quite remarkable properties of electrons in crystals. For example, they respond to applied electric or magnetic fields as if the electrons were endowed with an effective mass m^*, which may be larger or smaller than the free electron mass, or may even be negative. Electrons in

crystals respond to applied fields as if endowed with negative or positive charges, $-e$ or $+e$, and herein lies the explanation of the negative and positive values of the Hall coefficient.

NEARLY FREE ELECTRON MODEL

On the free electron model the allowed energy values are distributed essentially continuously from zero to infinity. We saw in Chapter 6 that

$$\epsilon_{\mathbf{k}} = \frac{\hbar^2}{2m}(k_x^2 + k_y^2 + k_z^2) \ , \tag{1}$$

where, for periodic boundary conditions over a cube of side L,

$$k_x, k_y, k_z = 0 \ ; \quad \pm\frac{2\pi}{L} \ ; \quad \pm\frac{4\pi}{L} \ ; \ \ldots \ . \tag{2}$$

The free electron wavefunctions are of the form

$$\psi_{\mathbf{k}}(\mathbf{r}) = \exp(i\mathbf{k} \cdot \mathbf{r}) \ ; \tag{3}$$

they represent running waves and carry momentum $\mathbf{p} = \hbar\mathbf{k}$.

The band structure of a crystal can often be explained by the nearly free electron model for which the band electrons are treated as perturbed only weakly by the periodic potential of the ion cores. This model answers almost all the qualitative questions about the behavior of electrons in metals.

We know that Bragg reflection is a characteristic feature of wave propagation in crystals. Bragg reflection of electron waves in crystals is the cause of energy gaps. (At Bragg reflection wavelike solutions of the Schrödinger equation do not exist, as in Fig. 2.) These energy gaps are of decisive significance in determining whether a solid is an insulator or a conductor.

We explain physically the origin of energy gaps in the simple problem of a linear solid of lattice constant a. The low energy portions of the band structure

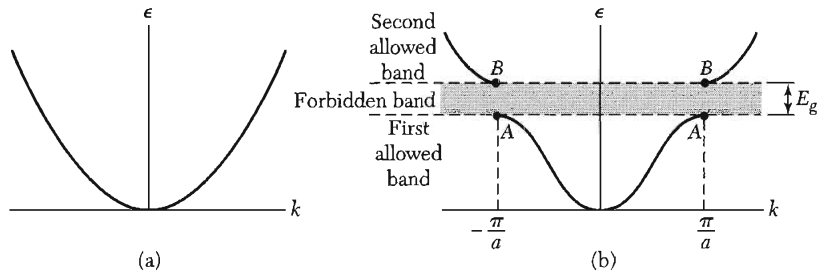


Figure 2 (a) Plot of energy ϵ versus wavevector k for a free electron. (b) Plot of energy versus wavevector for an electron in a monatomic linear lattice of lattice constant a. The energy gap E_g shown is associated with the first Bragg reflection at $k = \pm\pi/a$; other gaps are found at higher energies at $\pm n\pi/a$, for integral values of n.

are shown qualitatively in Fig. 2, in (a) for entirely free electrons and in (b) for electrons that are nearly free, but with an energy gap at $k = \pm\pi/a$. The Bragg condition $(\mathbf{k} + \mathbf{G})^2 = k^2$ for diffraction of a wave of wavevector $\mathbf{k}$ becomes in one dimension

$$k = \pm\tfrac{1}{2}G = \pm n\pi/a \ , \tag{4}$$

where $G = 2\pi n/a$ is a reciprocal lattice vector and n is an integer. The first reflections and the first energy gap occur at $k = \pm\pi/a$. The region in $\mathbf{k}$ space between $-\pi/a$ and π/a is the **first Brillouin zone** of this lattice. Other energy gaps occur for other values of the integer n.

The wavefunctions at $k = \pm\pi/a$ are not the traveling waves $\exp(i\pi x/a)$ or $\exp(-i\pi x/a)$ of free electrons. At these special values of k the wavefunctions are made up of equal parts of waves traveling to the right and to the left. When the Bragg reflection condition $k = \pm\pi/a$ is satisfied by the wavevector, a wave traveling to the right is Bragg-reflected to travel to the left, and vice versa. Each subsequent Bragg reflection will reverse the direction of travel of the wave. A wave that travels neither to the right nor to the left is a standing wave: it doesn't go anywhere.

The time-independent state is represented by standing waves. We can form two different standing waves from the two traveling waves

$$\exp(\pm i\pi x/a) = \cos(\pi x/a) \pm i \sin(\pi x/a),$$

so that the standing waves are

$$\begin{aligned} \psi(+) &= \exp(i\pi x/a) + \exp(-i\pi x/a) = 2\cos(\pi x/a) \ ; \\ \psi(-) &= \exp(i\pi x/a) - \exp(-i\pi x/a) = 2i \sin(\pi x/a) \ . \end{aligned} \tag{5}$$

The standing waves are labeled $(+)$ or $(-)$ according to whether or not they change sign when $-x$ is substituted for x. Both standing waves are composed of equal parts of right- and left-directed traveling waves.

Origin of the Energy Gap

The two standing waves $\psi(+)$ and $\psi(-)$ pile up electrons at different regions, and therefore the two waves have different values of the potential energy in the field of the ions of the lattice. This is the origin of the energy gap. The probability density ρ of a particle is $\psi^*\psi = |\psi|^2$. For a pure traveling wave $\exp(ikx)$, we have $\rho = \exp(-ikx)\exp(ikx) = 1$, so that the charge density is constant. The charge density is not constant for linear combinations of plane waves. Consider the standing wave $\psi(+)$ in (5); for this we have

$$\rho(+) = |\psi(+)|^2 \propto \cos^2 \pi x/a \ .$$

This function piles up electrons (negative charge) on the positive ions centered at $x = 0, a, 2a, \ldots$ in Fig. 3, where the potential energy is lowest.

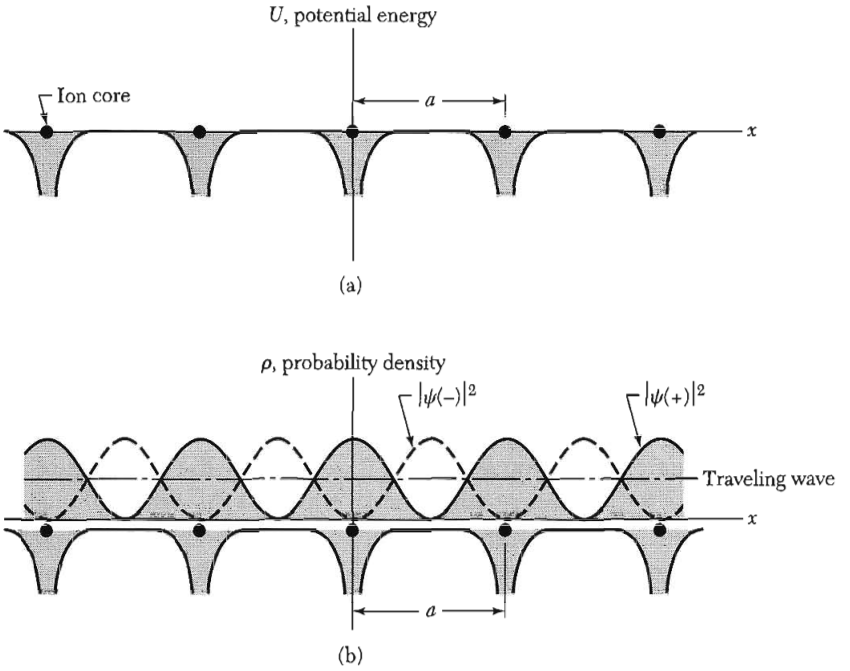


Figure 3 (a) Variation of potential energy of a conduction electron in the field of the ion cores of a linear lattice. (b) Distribution of probability density ρ in the lattice for $|\psi(-)|^2 \propto \sin^2 \pi x/a$; $|\psi(+)|^2 \propto \cos^2 \pi x/a$; and for a traveling wave. The wavefunction $\psi(+)$ piles up electronic charge on the cores of the positive ions, thereby lowering the potential energy in comparison with the average potential energy seen by a traveling wave. The wavefunction $\psi(-)$ piles up charge in the region between the ions, thereby raising the potential energy in comparison with that seen by a traveling wave. This figure is the key to understanding the origin of the energy gap.

Figure 3a pictures the variation of the electrostatic potential energy of a conduction electron in the field of the positive ion cores. The ion cores bear a net positive charge because the atoms are ionized in the metal, with the valence electrons taken off to form the conduction band. The potential energy of an electron in the field of a positive ion is negative, so that the force between them is attractive.

For the other standing wave $\psi(-)$ the probability density is

$$\rho(-) = |\psi(-)|^2 \propto \sin^2 \pi x/a \ ,$$

which concentrates electrons away from the ion cores. In Fig. 3b we show the electron concentration for the standing waves $\psi(+)$, $\psi(-)$, and for a traveling wave.

When we calculate the average or expectation values of the potential energy over these three charge distributions, we find that the potential energy of $\rho(+)$ is lower than that of the traveling wave, whereas the potential energy of $\rho(-)$ is higher than the traveling wave. We have an energy gap of width E_g if

the energies of $\rho(-)$ and $\rho(+)$ differ by E_g. Just below the energy gap at points A in Fig. 2 the wavefunction is $\psi(+)$, and just above the gap at points B the wavefunction is $\psi(-)$.

Magnitude of the Energy Gap

The wavefunctions at the Brillouin zone boundary $k = \pi/a$ are $\sqrt{2}\cos \pi x/a$ and $\sqrt{2}\sin \pi x/a$, normalized over unit length of line. Let us suppose that the potential energy of an electron in the crystal at point x is

$$U(x) = U \cos 2\pi x/a \ .$$

The first-order energy difference between the two standing wave states is

$$\begin{aligned} E_g &= \int_0^1 dx\, U(x)\,[|\psi(+)|^2 - |\psi(-)|^2] \\ &= 2\int dx\, U \cos(2\pi x/a)(\cos^2 \pi x/a - \sin^2 \pi x/a) = U \ . \end{aligned} \tag{6}$$

We see that the gap is equal to the Fourier component of the crystal potential.

BLOCH FUNCTIONS

F. Bloch proved the important theorem that the solutions of the Schrödinger equation for a periodic potential must be of a special form:

$$\boxed{\psi_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r}) \exp(i\mathbf{k} \cdot \mathbf{r}) \ ,} \tag{7}$$

where $u_{\mathbf{k}}(\mathbf{r})$ has the period of the crystal lattice with $u_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r} + \mathbf{T})$. Here $\mathbf{T}$ is a translation vector of the lattice. The result (7) expresses the Bloch theorem:

> The eigenfunctions of the wave equation for a periodic potential are the product of a plane wave $\exp(i\mathbf{k} \cdot \mathbf{r})$ times a function $u_{\mathbf{k}}(\mathbf{r})$ with the periodicity of the crystal lattice.

A one-electron wavefunction of the form (7) is called a Bloch function and can be decomposed into a sum of traveling waves, as we see later. Bloch functions can be assembled into wave packets to represent electrons that propagate freely through the potential field of the ion cores.

We give now a restricted proof of the Bloch theorem, valid when ψ_k is nondegenerate; that is, when there is no other wavefunction with the same energy and wavevector as ψ_k. The general case will be treated later. We consider N identical lattice points on a ring of length Na. The potential energy is periodic in a, with $U(x) = U(x + sa)$, where s is an integer.

Let us be guided by the symmetry of the ring to look for solutions of the wave equation such that

$$\psi(x + a) = C\psi(x) \ , \tag{8}$$

where C is a constant. Then, on going once around the ring,

$$\psi(x + Na) = \psi(x) = C^N \psi(x) \ ,$$

because $\psi(x)$ must be single-valued. It follows that C is one of the N roots of unity, or

$$C = \exp(i2\pi s/N) \ ; \qquad s = 0, 1, 2, \ . \ . \ . \ , N - 1 \ . \tag{9}$$

We use (9) to see that

$$\psi(x) = u_k(x) \exp(i2\pi sx/Na) \tag{10}$$

satisfies (8), provided that $u_k(x)$ has the periodicity a, so that $u_k(x) = u_k(x + a)$. This is the Bloch result (7).

KRONIG-PENNEY MODEL

A periodic potential for which the wave equation can be solved in terms of elementary functions is the square-well array of Fig. 4. The wave equation is

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + U(x)\psi = \epsilon\psi \ , \tag{11}$$

where $U(x)$ is the potential energy and ϵ is the energy eigenvalue.

In the region $0 < x < a$ in which $U = 0$, the eigenfunction is a linear combination,

$$\psi = Ae^{iKx} + Be^{-iKx} \ , \tag{12}$$

of plane waves traveling to the right and to the left, with energy

$$\epsilon = \hbar^2K^2/2m \ . \tag{13}$$

In the region $-b < x < 0$ within the barrier the solution is of the form

$$\psi = Ce^{Qx} + De^{-Qx} \ , \tag{14}$$

with

$$U_0 - \epsilon = \hbar^2Q^2/2m \ . \tag{15}$$

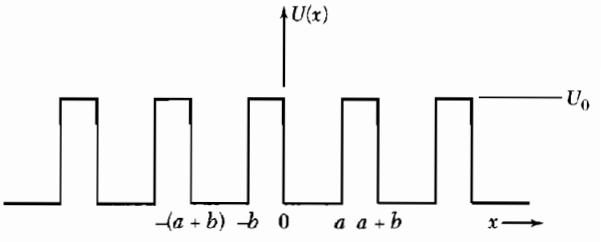


Figure 4 Square-well periodic potential as introduced by Kronig and Penney.

We want the complete solution to have the Bloch form (7). Thus the solution in the region $a < x < a + b$ must be related to the solution (14) in the region $-b < x < 0$ by the Bloch theorem:

$$\psi(a < x < a + b) = \psi(-b < x < 0)\, e^{ik(a+b)} \ , \tag{16}$$

which serves to define the wavevector k used as an index to label the solution.

The constants A, B, C, D are chosen so that ψ and $d\psi/dx$ are continuous at $x = 0$ and $x = a$. These are the usual quantum mechanical boundary conditions in problems that involve square potential wells. At $x = 0$,

$$A + B = C + D \ ; \tag{17}$$

$$iK(A - B) = Q(C - D) \ , \tag{18}$$

with Q from (14). At $x = a$, with the use of (16) for $\psi(a)$ under the barrier in terms of $\psi(-b)$,

$$Ae^{iKa} + Be^{-iKa} = (Ce^{-Qb} + De^{Qb})\, e^{ik(a+b)} \ ; \tag{19}$$

$$iK(Ae^{iKa} - Be^{-iKa}) = Q(Ce^{-Qb} - De^{Qb})\, e^{ik(a+b)} \ . \tag{20}$$

The four equations (17) to (20) have a solution only if the determinant of the coefficients of A, B, C, D vanishes, yielding

$$[(Q^2 - K^2)/2QK] \sinh Qb \sin Ka + \cosh Qb \cos Ka = \cos k(a + b) \ . \tag{21a}$$

It is rather tedious to obtain this equation.

The result is simplified if we represent the potential by the periodic delta function obtained when we pass to the limit $b = 0$ and $U_0 = \infty$ in such a way that $Q^2ba/2 = P$, a finite quantity. In this limit $Q \gg K$ and $Qb \ll 1$. Then (21a) reduces to

$$(P/Ka)\sin Ka + \cos Ka = \cos ka \ . \tag{21b}$$

The ranges of K for which this equation has solutions are plotted in Fig. 5, for the case $P = 3\pi/2$. The corresponding values of the energy are plotted in Fig. 6. Note the energy gaps at the zone boundaries. The wavevector k of the Bloch function is the important index, not the K in (12), which is related to the energy by (13). A treatment of this problem in wavevector space is given later in this chapter.

WAVE EQUATION OF ELECTRON IN A PERIODIC POTENTIAL

We considered in Fig. 3 the approximate form we expect for the solution of the Schrödinger equation if the wavevector is at a zone boundary, as at $k = \pm\pi/a$. We treat in detail the wave equation for a general potential, at general

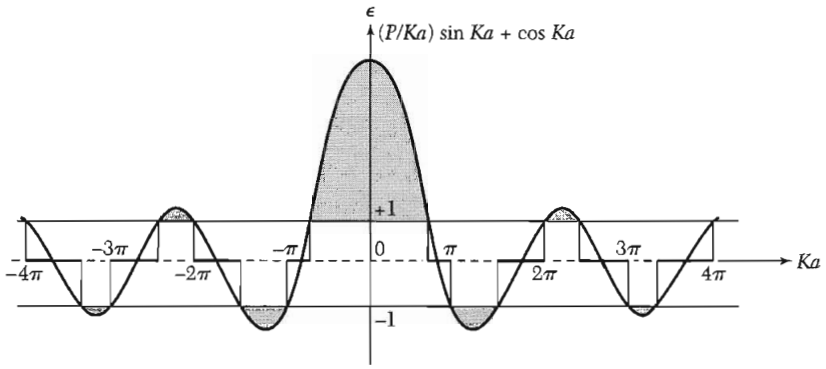


Figure 5 Plot of the function $(P/Ka)\sin Ka + \cos Ka$, for $P = 3\pi/2$. The allowed values of the energy e are given by those ranges of $Ka = (2m\epsilon/\hbar^2)^{1/2}a$ for which the function lies between ± 1. For other values of the energy there are no traveling wave or Bloch-like solutions to the wave equation, so that forbidden gaps in the energy spectrum are formed.

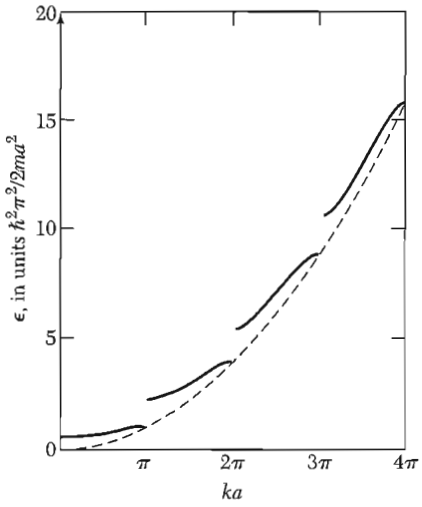


Figure 6 Plot of energy vs. wavenumber for the Kronig-Penney potential, with $P = 3\pi/2$. Notice the energy gaps at $ka = \pi, 2\pi, 3\pi \ldots$.

values of k. Let $U(x)$ denote the potential energy of an electron in a linear lattice of lattice constant a. We know that the potential energy is invariant under a crystal lattice translation: $U(x) = U(x + a)$. A function invariant under a crystal lattice translation may be expanded as a Fourier series in the reciprocal lattice vectors G. We write the Fourier series for the potential energy as

$$U(x) = \sum_G U_G e^{iGx} \quad . \tag{22}$$

The values of the coefficients U_G for actual crystal potentials tend to decrease rapidly with increasing magnitude of G. For a bare coulomb potential U_G decreases as $1/G^2$.

We want the potential energy $U(x)$ to be a real function:

$$U(x) = \sum_{G>0} U_G(e^{iGx} + e^{-iGx}) = 2\sum_{G>0} U_G \cos Gx \ . \tag{23}$$

For convenience we have assumed that the crystal is symmetric about $x = 0$ and that $U_0 = 0$.

The wave equation of an electron in the crystal is $\mathcal{H}\psi = \epsilon\psi$, where $\mathcal{H}$ is the hamiltonian and ϵ is the energy eigenvalue. The solutions ψ are called eigenfunctions or orbitals or Bloch functions. Explicitly, the wave equation is

$$\left(\frac{1}{2m}p^2 + U(x)\right)\psi(x) = \left(\frac{1}{2m}p^2 + \sum_G U_G\, e^{iGx}\right)\psi(x) = \epsilon\psi(x) \ . \tag{24}$$

Equation (24) is written in the one-electron approximation in which the orbital $\psi(x)$ describes the motion of one electron in the potential of the ion cores and in the average potential of the other conduction electrons.

The wavefunction $\psi(x)$ may be expressed as a Fourier series summed over all values of the wavevector permitted by the boundary conditions, so that

$$\psi = \sum_k C(k)\, e^{ikx} \ , \tag{25}$$

where k is real. (We could equally well write the index k as a subscript on C, as in C_k.)

The set of values of k has the form $2\pi n/L$, because these values satisfy periodic boundary conditions over length L. Here n is any integer, positive or negative. We do not assume, nor is it generally true, that $\psi(x)$ itself is periodic in the fundamental lattice translation a. The translational properties of $\psi(x)$ are determined by the Bloch theorem (7).

Not all wavevectors of the set $2\pi n/L$ enter the Fourier expansion of any one Bloch function. If one particular wavevector k is contained in a ψ, then all other wavevectors in the Fourier expansion of this ψ will have the form $k + G$, where G is any reciprocal lattice vector. We prove this result in (29) below.

We can label a wavefunction ψ that contains a component k as ψ_k or, equally well, as ψ_{k+G}, because if k enters the Fourier expansion then $k + G$ may enter. The wavevectors $k + G$ running over G are a restricted subset of the set $2\pi n/L$, as shown in Fig. 7.

We shall usually choose as a label for the Bloch function that k which lies within the first Brillouin zone. When other conventions are used, we shall say so. This situation differs from the phonon problem for a monatomic lattice where there are no components of the ion motion outside the first zone. The electron problem is like the x-ray diffraction problem because like the electron wavefunction the electromagnetic field exists everywhere within the crystal and not only at the ions.

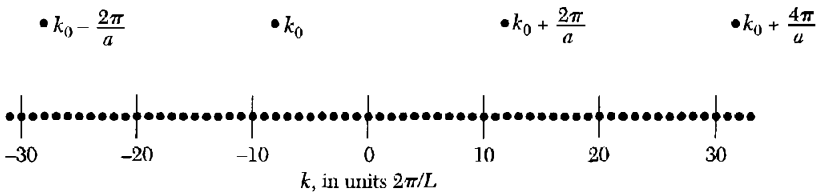


Figure 7 The lower points represent values of the wavevector $k = 2\pi n/L$ allowed by the periodic boundary condition on the wavefunction over a ring of circumference L composed of 20 primitive cells. The allowed values continue to $\pm\infty$. The upper points represent the first few wavevectors which may enter into the Fourier expansion of a wavefunction $\psi(x)$, starting from a particular wavevector $k = k_0 = -8(2\pi/L)$. The shortest reciprocal lattice vector is $2\pi/a = 20(2\pi/L)$.

To solve the wave equation, substitute (25) in (24) to obtain a set of linear algebraic equations for the Fourier coefficients. The kinetic energy term is

$$\frac{1}{2m}p^2\psi(x) = \frac{1}{2m}\left(-i\hbar\frac{d}{dx}\right)^2\psi(x) = -\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} = \frac{\hbar^2}{2m}\sum_k k^2C(k)\,e^{ikx} \ ;$$

and the potential energy term is

$$\left(\sum_G U_G e^{iGx}\right)\psi(x) = \sum_G\sum_k U_G e^{iGx}C(k)\,e^{ikx} \ .$$

The wave equation is obtained as the sum:

$$\sum_k \frac{\hbar^2}{2m}k^2C(k)\,e^{ikx} + \sum_G\sum_k U_G C(k)\,e^{i(k+G)x} = \epsilon\sum_k C(k)\,e^{ikx} \ . \tag{26}$$

Each Fourier component must have the same coefficient on both sides of the equation. Thus we have the **central equation**

$$\boxed{(\lambda_k - \epsilon)C(k) + \sum_G U_G C(k-G) = 0 \ .} \tag{27}$$

with the notation

$$\lambda_k = \hbar^2k^2/2m \ . \tag{28}$$

Equation (27) is a useful form of the wave equation in a periodic lattice, although unfamiliar because a set of algebraic equations has taken the place of the usual differential equation (24). The set appears unpleasant and formidable because there are, in principle, an infinite number of $C(k - G)$ to be determined. In practice a small number will often suffice, perhaps two or four. It takes some experience to appreciate the practical advantages of the algebraic approach.

Restatement of the Bloch Theorem

Once we determine the C's from (27), the wavefunction (25) is given as

$$\psi_k(x) = \sum_G C(k - G)\, e^{i(k-G)x} , \tag{29}$$

which may be rearranged as

$$\psi_k(x) = \left(\sum_G C(k - G)\, e^{-iGx}\right) e^{ikx} = e^{ikx} u_k(x) ,$$

with the definition

$$u_k(x) \equiv \sum_G C(k - G)\, e^{-iGx} .$$

Because $u_k(x)$ is a Fourier series over the reciprocal lattice vectors, it is invariant under a crystal lattice translation T, so that $u_k(x) = u_k(x + T)$. We verify this directly by evaluating $u_k(x + T)$:

$$u_k(x + T) = \sum C(k - G) e^{-iG(x+T)} = e^{-iGT}[\sum C(k - G)\, e^{-iGx}] = e^{-iGT} u_k(x) .$$

Because $\exp(-iGT) = 1$ by (2.17), it follows that $u_k(x + T) = u_k(x)$, thereby establishing the periodicity of u_k. This is an alternate and exact proof of the Bloch theorem and is valid even when the ψ_k are degenerate.

Crystal Momentum of an Electron

What is the significance of the wavevector $\mathbf{k}$ used to label the Bloch function? It has several properties:

- Under a crystal lattice translation which carries $\mathbf{r}$ to $\mathbf{r} + \mathbf{T}$ we have

$$\psi_{\mathbf{k}}(\mathbf{r} + \mathbf{T}) = e^{i\mathbf{k}\cdot\mathbf{T}} e^{i\mathbf{k}\cdot\mathbf{r}} u_{\mathbf{k}}(\mathbf{r} + \mathbf{T}) = e^{i\mathbf{k}\cdot\mathbf{T}} \psi_k(\mathbf{r}) , \tag{30}$$

  because $u_{\mathbf{k}}(\mathbf{r} + \mathbf{T}) = u_{\mathbf{k}}(\mathbf{r})$. Thus $\exp(i\mathbf{k} \cdot \mathbf{T})$ is the phase factor by which a Bloch function is multiplied when we make a crystal lattice translation $\mathbf{T}$.
- If the lattice potential vanishes, the central equation (27) reduces to $(\lambda_{\mathbf{k}} - \epsilon)C(\mathbf{k}) = 0$, so that all $C(\mathbf{k} - \mathbf{G})$ are zero except $C(\mathbf{k})$, and thus $u_{\mathbf{k}}(\mathbf{r})$ is constant. We have $\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}$, just as for a free electron. (This assumes we have had the foresight to pick the "right" $\mathbf{k}$ as the label. For many purposes other choices of $\mathbf{k}$, differing by a reciprocal lattice vector, will be more convenient.)
- The quantity $\mathbf{k}$ enters in the conservation laws that govern collision processes in crystals. (The conservation laws are really selection rules for transitions.) Thus $\hbar\mathbf{k}$ is called the **crystal momentum** of an electron. If an electron $\mathbf{k}$ absorbs in a collision a phonon of wavevector $\mathbf{q}$, the selection rule is $\mathbf{k} + \mathbf{q} = \mathbf{k}' + \mathbf{G}$. In this process the electron is scattered from a state $\mathbf{k}$ to a state $\mathbf{k}'$, with $\mathbf{G}$ a reciprocal lattice vector. Any arbitrariness in labeling the Bloch functions can be absorbed in the $\mathbf{G}$ without changing the physics of the process.

Solution of the Central Equation

The central equation (27),

$$(\lambda_k - \epsilon)C(k) + \sum_G U_G C(k - G) = 0 \ , \tag{31}$$

represents a set of simultaneous linear equations that connect the coefficients $C(k - G)$ for all reciprocal lattice vectors G. It is a set because there are as many equations as there are coefficients C. These equations are consistent if the determinant of the coefficients vanishes.

Let us write out the equations for an explicit problem. We let g denote the shortest G. We suppose that the potential energy $U(x)$ contains only a single Fourier component $U_g = U_{-g}$, denoted by U. Then a block of the determinant of the coefficients is given by:

$$\begin{vmatrix} \lambda_{k-2g} - \epsilon & U & 0 & 0 & 0 \\ U & \lambda_{k-g} - \epsilon & U & 0 & 0 \\ 0 & U & \lambda_k - \epsilon & U & 0 \\ 0 & 0 & U & \lambda_{k+g} - \epsilon & U \\ 0 & 0 & 0 & U & \lambda_{k+2g} - \epsilon \end{vmatrix} . \tag{32}$$

To see this, write out five successive equations of the set (31). The determinant in principle is infinite in extent, but it will often be sufficient to set equal to zero the portion we have shown.

At a given k, each root ϵ or ϵ_k lies on a different energy band, except in case of coincidence. The solution of the determinant (32) gives a set of energy eigenvalues ϵ_{nk}, where n is an index for ordering the energies and k is the wavevector that labels C_k.

Most often k will be taken in the first zone, to reduce possible confusion in the labeling. If we chose a k different from the original by some reciprocal lattice vector, we would have obtained the same set of equations in a different order—but having the same energy spectrum.

Kronig-Penney Model in Reciprocal Space

As an example of the use of the central equation (31) for a problem that is exactly solvable, we use the Kronig-Penney model of a periodic delta-function potential:

$$U(x) = 2\sum_{G>0} U_G \cos Gx = Aa\sum_s \delta(x - sa) \ , \tag{33}$$

where A is a constant and a the lattice spacing. The sum is over all integers s between 0 and $1/a$. The boundary conditions are periodic over a ring of unit

length, which means over $1/a$ atoms. Thus the Fourier coefficients of the potential are

$$\begin{aligned} U_G &= \int_0^1 dx\, U(x) \cos Gx = Aa\sum_s \int_0^1 dx\, \delta(x - sa) \cos Gx \\ &= Aa\sum_s \cos Gsa = A \ . \end{aligned} \tag{34}$$

All U_G are equal for the delta-function potential.

We write the central equation with k as the Bloch index. Thus (31) becomes

$$(\lambda_k - \epsilon)C(k) + A\sum_n C(k - 2\pi m/a) = 0 \ , \tag{35}$$

where $\lambda_k \equiv \hbar^2k^2/2m$ and the sum is over all integers n. We want to solve (35) for $\epsilon(k)$.

We define

$$f(k) = \sum_n C(k - 2\pi m/a) \ , \tag{36}$$

so that (35) becomes

$$C(k) = -\frac{(2mA/\hbar^2)f(k)}{k^2 - (2m\epsilon/\hbar^2)} \ . \tag{37}$$

Because the sum (36) is over all coefficients C, we have, for any n,

$$f(k) = f(k - 2\pi n/a) \ . \tag{38}$$

This relation lets us write

$$C(k - 2\pi n/a) = -(2mA/\hbar^2)f(k)[(k - 2\pi n/a)^2 - 2m\epsilon/\hbar^2)]^{-1} \ . \tag{39}$$

We sum both sides over all n to obtain, using (36) and cancelling $f(k)$ from both sides,

$$(\hbar^2/2mA) = -\sum_n [(k - 2\pi m/a)^2 - (2m\epsilon/\hbar^2)]^{-1} \ . \tag{40}$$

The sum can be carried out with the help of the standard relation

$$\operatorname{ctn} x = \sum_n \frac{1}{n\pi + x} \ . \tag{41}$$

After trigonometric manipulations in which we use relations for the difference of two cotangents and the product of two sines, the sum in (40) becomes

$$\frac{a^2 \sin Ka}{4Ka(\cos ka - \cos Ka)} \ , \tag{42}$$

where we write $K^2 = 2m\epsilon/\hbar^2$ as in (13).

The final result for (40) is

$$(mAa^2/2\hbar^2)(Ka)^{-1}\sin Ka + \cos Ka = \cos ka \ , \tag{43}$$

which agrees with the Kronig-Penney result (21b) with P written for $mAa^2/2\hbar^2$.

Empty Lattice Approximation

Actual band structures are usually exhibited as plots of energy versus wavevector in the first Brillouin zone. When wavevectors happen to be given outside the first zone, they are carried back into the first zone by subtracting a suitable reciprocal lattice vector. Such a translation can always be found. The operation is helpful in visualization.

When band energies are approximated fairly well by free electron energies $\epsilon_{\mathbf{k}} = \hbar^2\mathbf{k}^2/2m$, it is advisable to start a calculation by carrying the free electron energies back into the first zone. The procedure is simple enough once one gets the hang of it. We look for a $\mathbf{G}$ such that a $\mathbf{k}'$ in the first zone satisfies

$$\mathbf{k}' + \mathbf{G} = \mathbf{k} \ ,$$

where $\mathbf{k}$ is unrestricted and is the true free electron wavevector in the empty lattice. (Once the plane wave is modulated by the lattice, there is no single "true" wavevector for the state ψ.)

If we drop the prime on $\mathbf{k}'$ as unnecessary baggage, the free electron energy can always be written as

$$\begin{aligned}\epsilon(k_x,k_y,k_z) &= (\hbar^2/2m)(\mathbf{k}+\mathbf{G})^2 \\ &= (\hbar^2/2m)[(k_x+G_x)^2 + (k_y+G_y)^2 + (k_z+G_z)^2] \ ,\end{aligned}$$

with $\mathbf{k}$ in the first zone and $\mathbf{G}$ allowed to run over the appropriate reciprocal lattice points.

We consider as an example the low-lying free electron bands of a simple cubic lattice. Suppose we want to exhibit the energy as a function of $\mathbf{k}$ in the [100] direction. For convenience, choose units such that $\hbar^2/2m = 1$. We show several low-lying bands in this empty lattice approximation with their energies $\epsilon(000)$ at $\mathbf{k} = 0$ and $\epsilon(k_x00)$ along the k_x axis in the first zone:

Band	$Ga/2\pi$	$\epsilon(000)$	$\epsilon(k_x00)$
1	000	0	k_x^2
2,3	100,$\bar{1}$00	$(2\pi/a)^2$	$(k_x \pm 2\pi/a)^2$
4,5,6,7	010,0$\bar{1}$0,001,00$\bar{1}$	$(2\pi/a)^2$	$k_x^2 + (2\pi/a)^2$
8,9,10,11	110,101,1$\bar{1}$0,10$\bar{1}$	$2(2\pi/a)^2$	$(k_x + 2\pi/a)^2 + (2\pi/a)^2$
12,13,14,15	$\bar{1}$10,$\bar{1}$01,$\bar{1}$10,$\bar{1}$0$\bar{1}$	$2(2\pi/a)^2$	$(k_x - 2\pi/a)^2 + (2\pi/a)^2$
16,17,18,19	011,0$\bar{1}$1,01$\bar{1}$,0$\bar{1}\bar{1}$	$2(2\pi/a)^2$	$k_x^2 + 2(2\pi/a)^2$

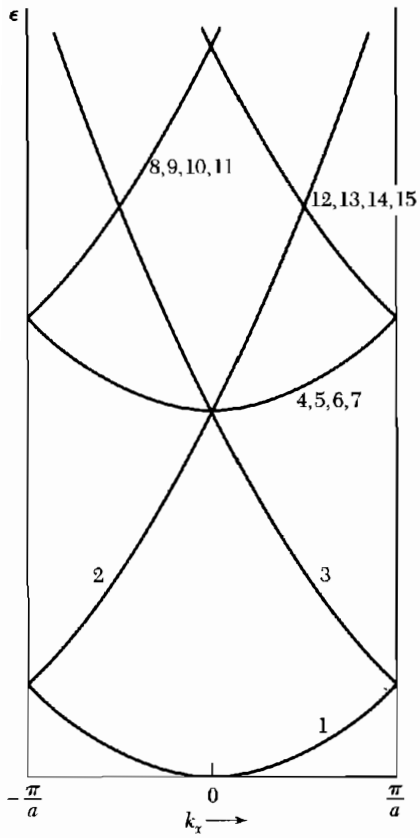


Figure 8 Low-lying free electron energy bands of the empty *sc* lattice, as transformed to the first Brillouin zone and plotted vs. $(k_x 00)$. The free electron energy is $\hbar^2(\mathbf{k}+\mathbf{G})^2/2m$, where the **G**'s are given in the second column of the table. The bold curves are in the first Brillouin zone, with $-\pi/a \leq k_x \leq \pi/a$. Energy bands drawn in this way are said to be in the reduced zone scheme.

These free electron bands are plotted in Fig. 8. It is a good exercise to plot the same bands for **k** parallel to the [111] direction of wavevector space.

Approximate Solution Near a Zone Boundary

We suppose that the Fourier components U_G of the potential energy are small in comparison with the kinetic energy of a free electron at the zone boundary. We first consider a wavevector exactly at the zone boundary at $\frac{1}{2}G$, that is, at π/a. Here

$$k^2 = (\tfrac{1}{2}G)^2 \;; \quad (k-G)^2 = (\tfrac{1}{2}G - G)^2 = (\tfrac{1}{2}G)^2 \;,$$

so that at the zone boundary the kinetic energy of the two component waves $k = \pm\frac{1}{2}G$ are equal.

If $C(\frac{1}{2}G)$ is an important coefficient in the orbital (29) at the zone boundary, then $C(-\frac{1}{2}G)$ is also an important coefficient. This result also follows from the discussion of (5). We retain only those equations in the central equation that contain both coefficients $C(\frac{1}{2}G)$ and $C(-\frac{1}{2}G)$, and neglect all other coefficients.

One equation of (31) becomes, with $k = \frac{1}{2}G$ and $\lambda = \hbar^2(\frac{1}{2}G)^2/2m$,

$$(\lambda - \epsilon)C(\tfrac{1}{2}G) + UC(-\tfrac{1}{2}G) = 0 \ . \tag{44}$$

Another equation of (31) becomes, with $k = \frac{1}{2}G$,

$$(\lambda - \epsilon)C(-\tfrac{1}{2}G) + UC(\tfrac{1}{2}G) = 0 \ . \tag{45}$$

These two equations have nontrivial solutions for the two coefficients if the energy ϵ satisfies

$$\begin{vmatrix} \lambda - \epsilon & U \\ U & \lambda - \epsilon \end{vmatrix} = 0 \ , \tag{46}$$

whence

$$(\lambda - \epsilon)^2 = U^2 \ ; \qquad \epsilon = \lambda \pm U = \frac{\hbar^2}{2m}(\tfrac{1}{2}G)^2 \pm U \ . \tag{47}$$

The energy has two roots, one lower than the free electron kinetic energy by U, and one higher by U. Thus the potential energy $2U \cos Gx$ has created an energy gap $2U$ at the zone boundary.

The ratio of the C's may be found from either (44) or (45):

$$\frac{C(-\frac{1}{2}G)}{C(\frac{1}{2}G)} = \frac{\epsilon - \lambda}{U} = \pm 1 \ , \tag{48}$$

where the last step uses (47). Thus the Fourier expansion of $\psi(x)$ at the zone boundary has the two solutions

$$\psi(x) = \exp(iGx/2) \pm \exp(-iGx/2) \ .$$

These orbitals are identical to (5).

One solution gives the wavefunction at the bottom of the energy gap; the other gives the wavefunction at the top of the gap. Which solution has the lower energy depends on the sign of U.

We now solve for orbitals with wavevector k near the zone boundary $\frac{1}{2}G$. We use the same two-component approximation, now with a wavefunction of the form

$$\psi(x) = C(k)\, e^{ikx} + C(k - G)\, e^{i(k-G)x} \ . \tag{49}$$

As directed by the central equation (31), we solve the pair of equations

$$(\lambda_k - \epsilon)C(k) + UC(k - G) = 0 \ ;$$
$$(\lambda_{k-G} - \epsilon)C(k - G) + UC(k) = 0 \ ,$$

with λ_k defined as $\hbar^2k^2/2m$. These equations have a solution if the energy ϵ satisfies

$$\begin{vmatrix} \lambda_k - \epsilon & U \\ U & \lambda_{k-G} - \epsilon \end{vmatrix} = 0 ,$$

whence $\epsilon^2 - \epsilon(\lambda_{k-G} + \lambda_k) + \lambda_{k-G}\lambda_k - U^2 = 0$.

The energy has two roots:

$$\epsilon = \tfrac{1}{2}(\lambda_{k-G} + \lambda_k) \pm [\tfrac{1}{4}(\lambda_{k-G} - \lambda_k)^2 + U^2]^{1/2}, \tag{50}$$

and each root describes an energy band, plotted in Fig. 9. It is convenient to expand the energy in terms of a quantity $\tilde{K}$ (the mark over the K is called a tilde), which measures the difference $\tilde{K} \equiv k - \frac{1}{2}G$ in wavevector between k and the zone boundary:

$$\begin{aligned} \epsilon_{\tilde{K}} &= (\hbar^2/2m)(\tfrac{1}{4}G^2 + \tilde{K}^2) \pm [4\lambda(\hbar^2\tilde{K}^2/2m) + U^2]^{1/2} \\ &\simeq (\hbar^2/2m)(\tfrac{1}{4}G^2 + \tilde{K}^2) \pm U[1 + 2(\lambda/U^2)(\hbar^2\tilde{K}^2/2m)] , \end{aligned} \tag{51}$$

in the region $\hbar^2G\tilde{K}/2m \ll |U|$. Here $\lambda = (\hbar^2/2m)(\frac{1}{2}G)^2$ as before.

Writing the two zone boundary roots of (47) as $\epsilon(\pm)$, we may write (51) as

$$\epsilon_{\tilde{K}}(\pm) = \epsilon(\pm) + \frac{\hbar^2\tilde{K}^2}{2m}\left(1 \pm \frac{2\lambda}{U}\right) . \tag{52}$$

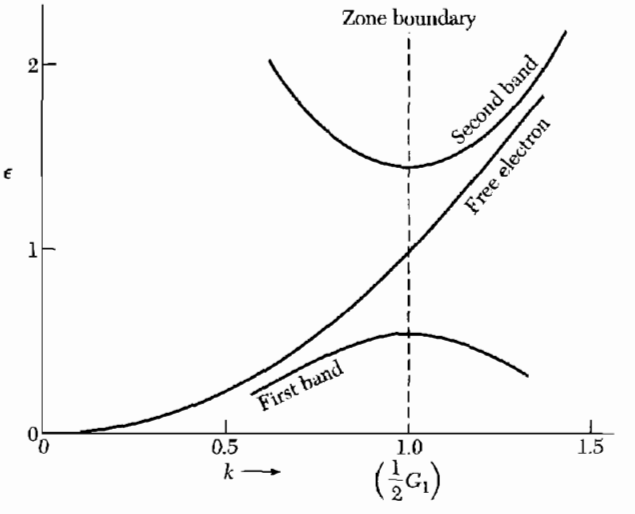


Figure 9 Solutions of (50) in the periodic zone scheme, in the region near a boundary of the first Brillouin zone. The units are such that $U = -0.45$, $G = 2$, and $\hbar^2/m = 1$. The free electron curve is drawn for comparison. The energy gap at the zone boundary is 0.90. The value of U has deliberately been chosen large for this illustration, too large for the two-term approximation to be accurate.

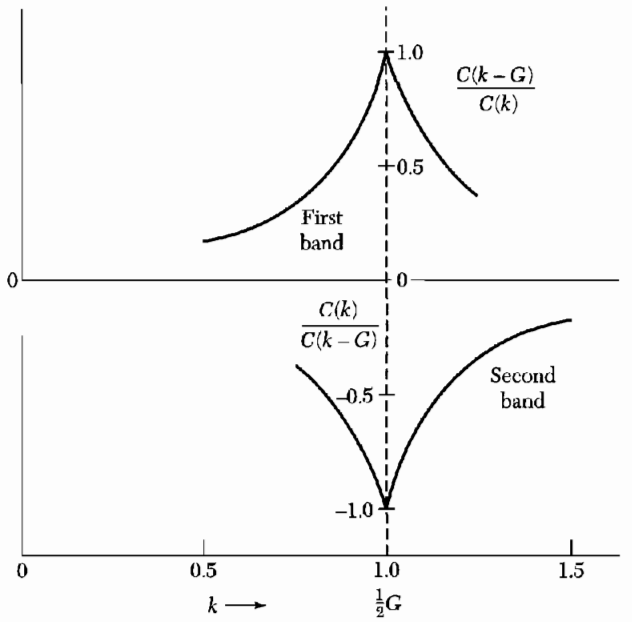


Figure 10 Ratio of the coefficients in $\psi(x) = C(k)\exp(ikx) + C(k - G)\exp[i(k - G)x]$ as calculated near the boundary of the first Brillouin zone. One component dominates as we move away from the boundary.

These are the roots for the energy when the wavevector is very close to the zone boundary at $\frac{1}{2}G$. Note the quadratic dependence of the energy on the wavevector $\bar{K}$. For U negative, the solution $\epsilon(-)$ corresponds to the upper of the two bands, and $\epsilon(+)$ to the lower of the two bands. The two C's are plotted in Fig. 10.

NUMBER OF ORBITALS IN A BAND

Consider a linear crystal constructed of an even number N of primitive cells of lattice constant a. In order to count states we apply periodic boundary conditions to the wavefunctions over the length of the crystal. The allowed values of the electron wavevector k in the first Brillouin zone are given by (2):

$$k = 0\ ;\quad \pm\frac{2\pi}{L}\ ;\quad \pm\frac{4\pi}{L}\ ;\ \ldots\ ;\ \frac{N\pi}{L}\ . \tag{53}$$

We cut the series off at $N\pi/L = \pi/a$, for this is the zone boundary. The point $-N\pi/L = -\pi/a$ is not to be counted as an independent point because it is connected by a reciprocal lattice vector with π/a. The total number of points is exactly N, the number of primitive cells.

Each primitive cell contributes exactly one independent value of k to each energy band. This result carries over into three dimensions. With account taken of the two independent orientations of the electron spin, **there**

are $2N$ independent orbitals in each energy band. If there is a single atom of valence, one in each primitive cell, the band can be half filled with electrons. If each atom contributes two valence electrons to the band, the band can be exactly filled. If there are two atoms of valence, one in each primitive cell, the band can also be exactly filled.

Metals and Insulators

If the valence electrons exactly fill one or more bands, leaving others empty, the crystal will be an insulator. An external electric field will not cause current flow in an insulator. (We suppose that the electric field is not strong enough to disrupt the electronic structure.) Provided that a filled band is separated by an energy gap from the next higher band, there is no continuous way to change the total momentum of the electrons if every accessible state is filled. Nothing changes when the field is applied. This is quite unlike the situation for free electrons for which $\mathbf{k}$ increases uniformly in a field (Chapter 6).

A crystal can be an insulator only if the number of valence electrons in a primitive cell of the crystal is an even integer. (An exception must be made for electrons in tightly bound inner shells which cannot be treated by band theory.) If a crystal has an even number of valence electrons per primitive cell, it is necessary to consider whether or not the bands overlap in energy. If the bands overlap in energy, then instead of one filled band giving an insulator, we can have two partly filled bands giving a metal (Fig. 11).

The alkali metals and the noble metals have one valence electron per primitive cell, so that they have to be metals. The alkaline earth metals have two valence electrons per primitive cell; they could be insulators, but the bands overlap in energy to give metals, but not very good metals. Diamond, silicon, and germanium each have two atoms of valence four, so that there are

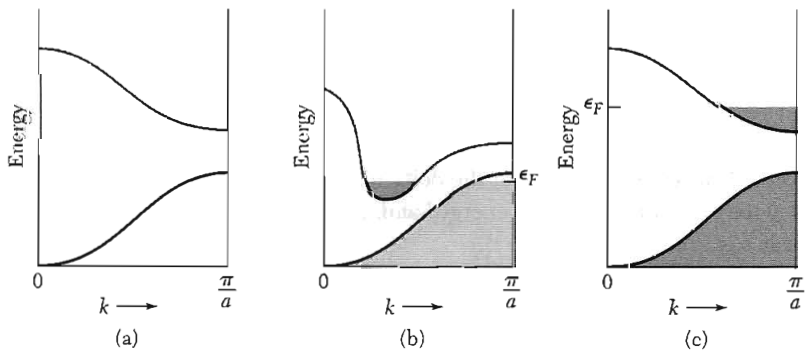


Figure 11 Occupied states and band structures giving (a) an insulator, (b) a metal or a semimetal because of band overlap, and (c) a metal because of electron concentration. In (b) the overlap need not occur along the same directions in the Brillouin zone. If the overlap is small, with relatively few states involved, we speak of a semimetal.

eight valence electrons per primitive cell; the bands do not overlap, and the pure crystals are insulators at absolute zero.

SUMMARY

- The solutions of the wave equation in a periodic lattice are of the Bloch form $\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}\, u_{\mathbf{k}}(\mathbf{r})$, where $u_{\mathbf{k}}(\mathbf{r})$ is invariant under a crystal lattice translation.
- There are regions of energy for which no Bloch function solutions of the wave equation exist (see Problem 5). These energies form forbidden regions in which the wavefunctions are damped in space and the values of the $\mathbf{k}$'s are complex, as pictured in Fig. 12. The existence of forbidden regions of energy is prerequisite to the existence of insulators.
- Energy bands may often be approximated by one or two plane waves: for example, $\psi_k(x) \cong C(k)e^{ikx} + C(k - G)e^{i(k-G)x}$ near the zone boundary at $\frac{1}{2}G$.
- The number of orbitals in a band is $2N$, where N is the number of primitive cells in the specimen.

Problems

1. ***Square lattice, free electron energies.*** (a) Show for a simple square lattice (two dimensions) that the kinetic energy of a free electron at a corner of the first zone is higher than that of an electron at midpoint of a side face of the zone by a factor of 2. (b) What is the corresponding factor for a simple cubic lattice (three dimensions)? (c) What bearing might the result of (b) have on the conductivity of divalent metals?

2. ***Free electron energies in reduced zone.*** Consider the free electron energy bands of an fcc crystal lattice in the approximation of an empty lattice, but in the reduced zone scheme in which all $\mathbf{k}$'s are transformed to lie in the first Brillouin zone. Plot roughly in the [111] direction the energies of all bands up to six times the lowest band energy at the zone boundary at $\mathbf{k} = (2\pi/a)(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. Let this be the unit of energy. This problem shows why band edges need not necessarily be at the zone center. Several of the degeneracies (band crossings) will be removed when account is taken of the crystal potential.

3. ***Kronig-Penney model.*** (a) For the delta-function potential and with $P \ll 1$, find at $k = 0$ the energy of the lowest energy band. (b) For the same problem find the band gap at $k = \pi/a$.

4. ***Potential energy in the diamond structure.*** (a) Show that for the diamond structure the Fourier component U_G of the crystal potential seen by an electron is equal to zero for $\mathbf{G} = 2\mathbf{A}$, where $\mathbf{A}$ is a basis vector in the reciprocal lattice referred to the conventional cubic cell. (b) Show that in the usual first-order approximation to the solutions of the wave equation in a periodic lattice the energy gap vanishes at the zone boundary plane normal to the end of the vector $\mathbf{A}$.

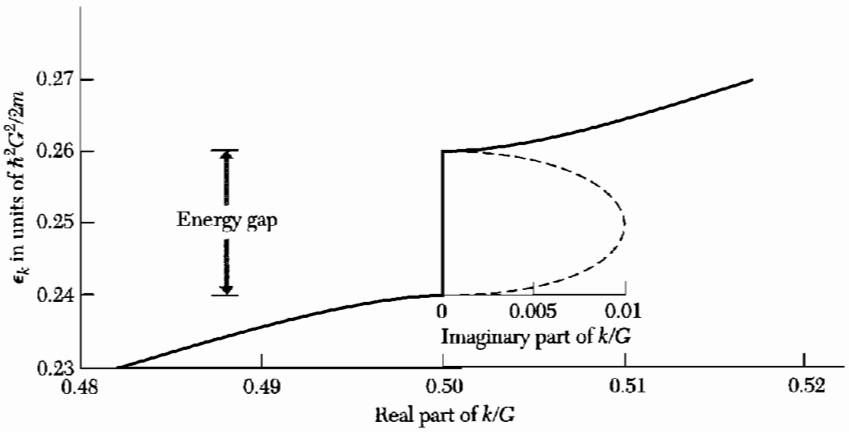


Figure 12 In the energy gap there exist solutions of the wave equation for complex values of the wavevector. At the boundary of the first zone the real part of the wavevector is $\frac{1}{2}G$. The imaginary part of k in the gap is plotted in the approximation of two plane waves, for $U = 0.01\ \hbar^2G^2/2m$. In an infinite unbounded crystal the wavevector must be real, or else the amplitude will increase without limit. But on a surface or at a junction there can exist solutions with complex wavevector.

*5. ***Complex wavevectors in the energy gap.*** Find an expression for the imaginary part of the wavevector in the energy gap at the boundary of the first Brillouin zone, in the approximation that led to Eq. (46). Give the result for the $\mathrm{Im}(k)$ at the center of the energy gap. The result for small $\mathrm{Im}(k)$ is

$$(\hbar^2/2m)[\mathrm{Im}(k)]^2 \approx 2mU^2/\hbar^2G^2 \ .$$

The form as plotted in Fig. 12 is of importance in the theory of Zener tunneling from one band to another in the presence of a strong electric field.

6. ***Square lattice.*** Consider a square lattice in two dimensions with the crystal potential

$$U(x,y) = -4U\cos(2\pi x/a)\cos(2\pi y/a) \ .$$

Apply the central equation to find approximately the energy gap at the corner point $(\pi/a, \pi/a)$ of the Brillouin zone. It will suffice to solve a 2×2 determinantal equation.

*This problem is somewhat difficult.

8

Semiconductor Crystals

NOTE: The discussion of carrier orbits in applied fields is continued in Chapter 9.

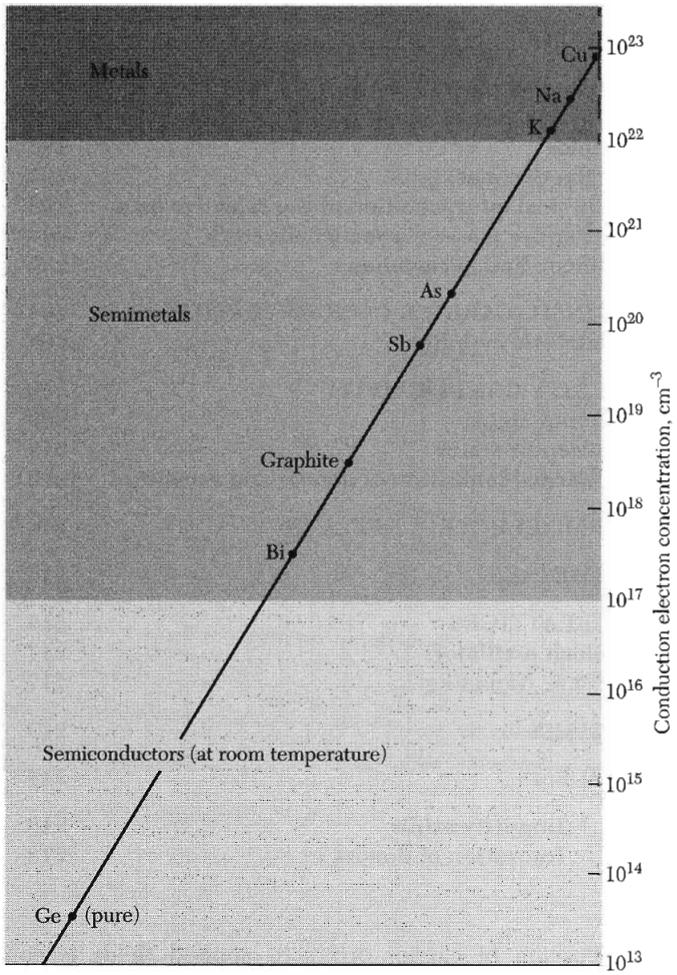


Figure 1 Carrier concentrations for metals, semimetals, and semiconductors. The semiconductor range may be extended upward by increasing the impurity concentration, and the range can be extended downward to merge eventually with the insulator range.

Carrier concentrations representative of metals, semimetals, and semiconductors are shown in Fig. 1. Semiconductors are generally classified by their electrical resistivity at room temperature, with values in the range of 10^{-2} to 10^{9} ohm-cm, and strongly dependent on temperature. At absolute zero a pure, perfect crystal of most semiconductors will be an insulator, if we arbitrarily define an insulator as having a resistivity above 10^{14} ohm-cm.

Devices based on semiconductors include transistors, switches, diodes, photovoltaic cells, detectors, and thermistors. These may be used as single circuit elements or as components of integrated circuits. We discuss in this chapter the central physical features of the classical semiconductor crystals, particularly silicon, germanium, and gallium arsenide.

Some useful nomenclature: the semiconductor compounds of chemical formula AB, where A is a trivalent element and B is a pentavalent element, are called III-V (three-five) compounds. Examples are indium antimonide and gallium arsenide. Where A is divalent and B is hexavalent, the compound is called a II-VI compound; examples are zinc sulfide and cadmium sulfide. Silicon and germanium are sometimes called diamond-type semiconductors, because they have the crystal structure of diamond. Diamond itself is more an insulator rather than a semiconductor. Silicon carbide SiC is a IV-IV compound.

A highly purified semiconductor exhibits intrinsic conductivity, as distinguished from the impurity conductivity of less pure specimens. In the **intrinsic temperature range** the electrical properties of a semiconductor are not essentially modified by impurities in the crystal. An electronic band scheme leading to intrinsic conductivity is indicated in Fig. 2. The conduction band is vacant at absolute zero and is separated by an energy gap E_g from the filled valence band.

The **band gap** is the difference in energy between the lowest point of the conduction band and the highest point of the valence band. The lowest point in the conduction band is called the **conduction band edge**; the highest point in the valence band is called the **valence band edge.**

As the temperature is increased, electrons are thermally excited from the valence band to the conduction band (Fig. 3). Both the electrons in the conduction band and the vacant orbitals or holes left behind in the valence band contribute to the electrical conductivity.

BAND GAP

The intrinsic conductivity and intrinsic carrier concentrations are largely controlled by E_g/k_BT, the ratio of the band gap to the temperature. When this ratio is large, the concentration of intrinsic carriers will be low and the

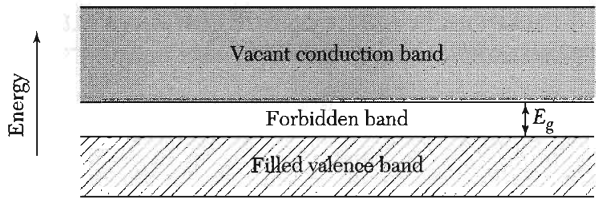


Figure 2 Band scheme for intrinsic conductivity in a semiconductor. At 0 K the conductivity is zero because all states in the valence band are filled and all states in the conduction band are vacant. As the temperature is increased, electrons are thermally excited from the valence band to the conduction band, where they become mobile. Such carriers are called "intrinsic."

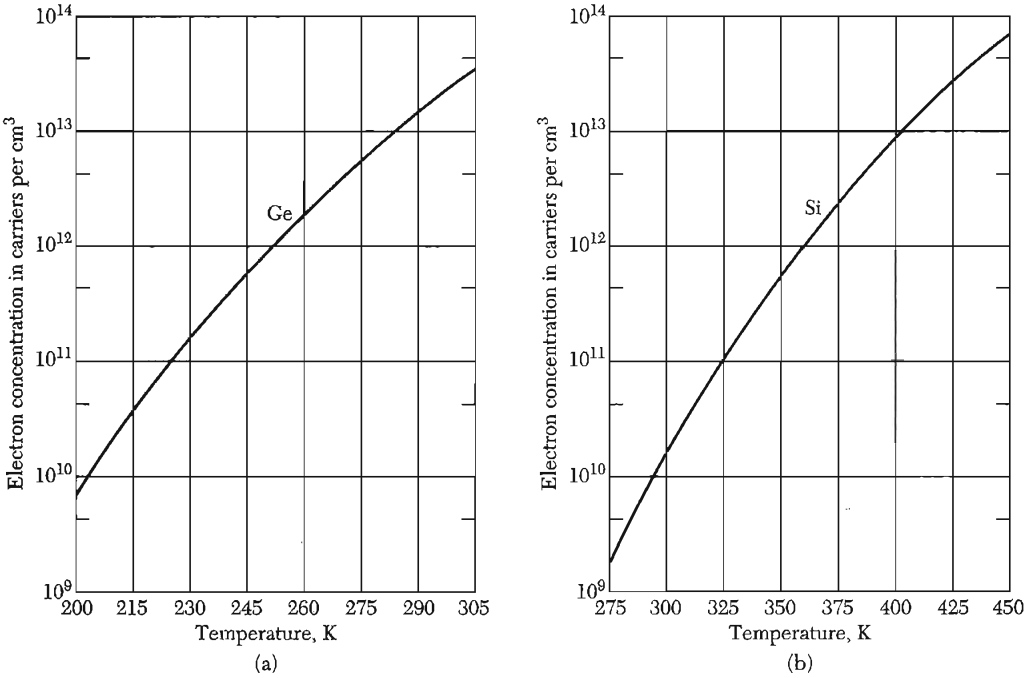


Figure 3 Intrinsic electron concentration as a function of temperature for (a) germanium and (b) silicon. Under intrinsic conditions the hole concentration is equal to the electron concentration. The intrinsic concentration at a given temperature is higher in Ge than in Si because the energy gap is narrower in Ge (0.66 eV) than in Si (1.11 eV). (After W. C. Dunlap.)

conductivity will be low. Band gaps of representative semiconductors are given in Table 1. The best values of the band gap are obtained by optical absorption.

In a **direct absorption process** the threshold of continuous optical absorption at frequency ω_g measures the band gap $E_g = \hbar\omega_g$ as shown in Figs. 4a and 5a. A photon is absorbed by the crystal with the creation of an electron and a hole.

In the **indirect absorption process** in Figs. 4b and 5b the minimum energy gap of the band structure involves electrons and holes separated by a

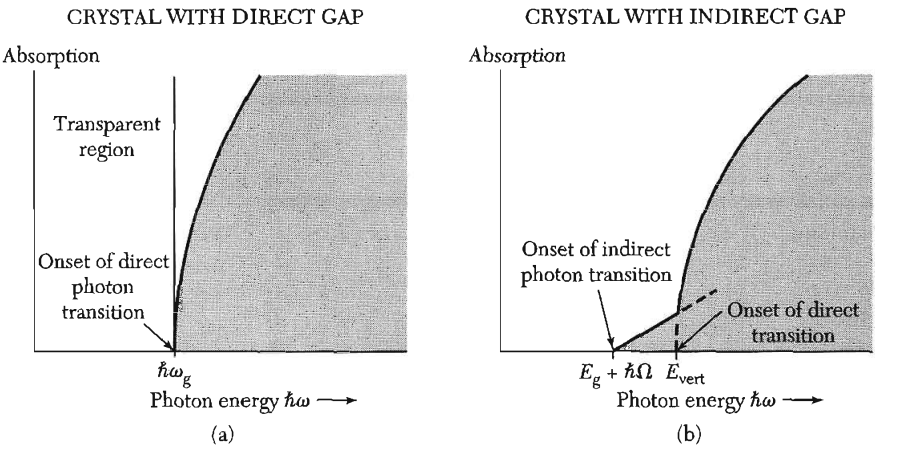


Figure 4 Optical absorption in pure insulators at absolute zero. In (a) the threshold determines the energy gap as $E_g = \hbar\omega_g$. In (b) the optical absorption is weaker near the threshold: at $\hbar\omega = E_g + \hbar\Omega$ a photon is absorbed with the creation of three particles: a free electron, a free hole, and a phonon of energy $\hbar\Omega$. In (b) the energy E_{vert} marks the threshold for the creation of a free electron and a free hole, with no phonon involved. Such a transition is called vertical; it is similar to the direct transition in (a). These plots do not show absorption lines that sometimes are seen lying just to the low energy side of the threshold. Such lines are due to the creation of a bound electron-hole pair, called an exciton.

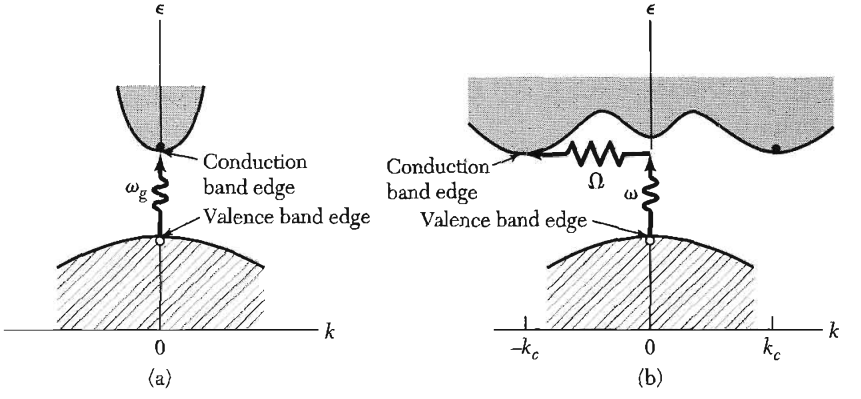


Figure 5 In (a) the lowest point of the conduction band occurs at the same value of $\mathbf{k}$ as the highest point of the valence band. A direct optical transition is drawn vertically with no significant change of $\mathbf{k}$, because the absorbed photon has a very small wavevector. The threshold frequency ω_g for absorption by the direct transition determines the energy gap $E_g = \hbar\omega_g$. The indirect transition in (b) involves both a photon and a phonon because the band edges of the conduction and valence bands are widely separated in $\mathbf{k}$ space. The threshold energy for the indirect process in (b) is greater than the true band gap. The absorption threshold for the indirect transition between the band edges is at $\hbar\omega = E_g + \hbar\Omega$, where Ω is the frequency of an emitted *phonon* of wavevector $\mathbf{K} \cong -\mathbf{k}_g$. At higher temperatures phonons are already present; if a phonon is absorbed along with a photon, the threshold energy is $\hbar\omega = E_g - \hbar\Omega$. *Note:* The figure shows only the threshold transitions. Transitions occur generally between almost all points of the two bands for which the wavevectors and energy can be conserved.

Table 1 Energy gap between the valence and conduction bands
(*i* = indirect gap; *d* = direct gap)

		E_g, eV				E_g, eV	
Crystal	Gap	0 K	300 K	Crystal	Gap	0 K	300 K
Diamond	*i*	5.4		SiC(hex)	*i*	3.0	—
Si	*i*	1.17	1.11	Te	*d*	0.33	—
Ge	*i*	0.744	0.66	HgTe[a]	*d*	−0.30	
αSn	*d*	0.00	0.00	PbS	*d*	0.286	0.34–0.37
InSb	*d*	0.23	0.17	PbSe	*i*	0.165	0.27
InAs	*d*	0.43	0.36	PbTe	*i*	0.190	0.29
InP	*d*	1.42	1.27	CdS	*d*	2.582	2.42
GaP	*i*	2.32	2.25	CdSe	*d*	1.840	1.74
GaAs	*d*	1.52	1.43	CdTe	*d*	1.607	1.44
GaSb	*d*	0.81	0.68	SnTe	*d*	0.3	0.18
AlSb	*i*	1.65	1.6	Cu_2O	*d*	2.172	—

[a]HgTe is a semimetal; the bands overlap.

substantial wavevector $\mathbf{k}_c$. Here a direct photon transition at the energy of the minimum gap cannot satisfy the requirement of conservation of wavevector, because photon wavevectors are negligible at the energy range of interest. But if a phonon of wavevector $\mathbf{K}$ and frequency Ω is created in the process, then we can have

$$\mathbf{k}(\text{photon}) = \mathbf{k}_c + \mathbf{K} \cong \mathbf{0} \; ; \qquad \hbar\omega = E_g + \hbar\Omega \; ,$$

as required by the conservation laws. The phonon energy $\hbar\Omega$ will generally be much less than E_g: a phonon even of high wavevector is an easily accessible source of crystal momentum because the phonon energies are characteristically small (~0.01 to 0.03 eV) in comparison with the energy gap. If the temperature is high enough that the necessary phonon is already thermally excited in the crystal, it is possible also to have a photon absorption process in which the phonon is absorbed.

The band gap may also be deduced from the temperature dependence of the conductivity or of the carrier concentration in the intrinsic range. The carrier concentration is obtained from measurements of the Hall voltage (Chapter 6), sometimes supplemented by conductivity measurements. Optical measurements determine whether the gap is direct or indirect. The band edges in Ge and in Si are connected by indirect transitions; the band edges in InSb and GaAs are connected by a direct transition (Fig. 6). The gap in αSn is

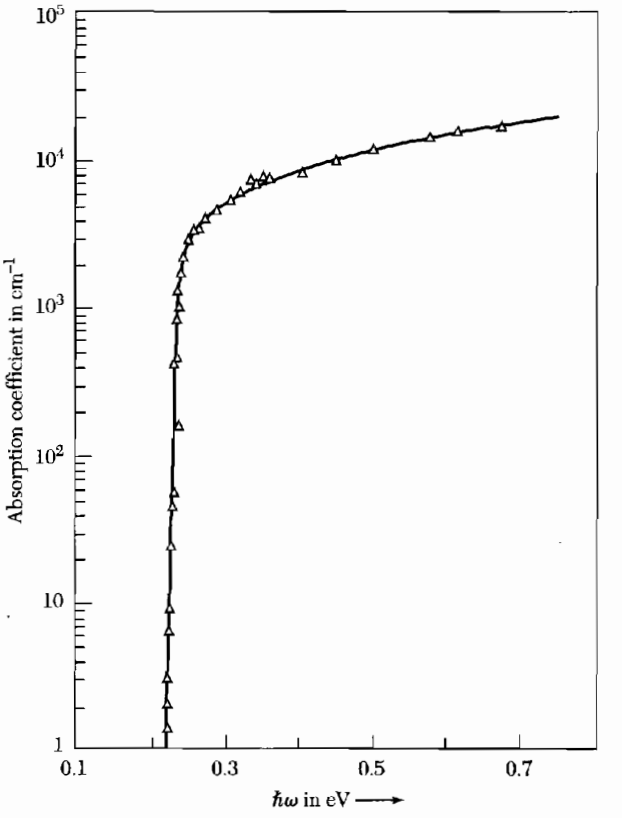


Figure 6 Optical absorption in pure indium antimonide, InSb. The transition is direct because both conduction and valence band edges are at the center of the Brillouin zone, $\mathbf{k} = 0$. Notice the sharp threshold. (After G. W. Gobeli and H. Y. Fan.)

direct and is exactly zero; HgTe and HgSe are semimetals and have negative gaps—the conduction and valence bands overlap.

EQUATIONS OF MOTION

We derive the equation of motion of an electron in an energy band. We look at the motion of a wave packet in an applied electric field. Suppose that the wave packet is made up of wavefunctions assembled near a particular wavevector k. The group velocity by definition is $v_g = d\omega/dk$. The frequency associated with a wavefunction of energy ϵ by quantum theory is $\omega = \epsilon/\hbar$, and so

$$v_g = \hbar^{-1}\, d\epsilon/dk \qquad \text{or} \qquad \mathbf{v} = \hbar^{-1} \nabla_{\mathbf{k}} \epsilon(\mathbf{k}) \ . \tag{1}$$

The effects of the crystal on the electron motion are contained in the dispersion relation $\epsilon(\mathbf{k})$.

The work $\delta\epsilon$ done on the electron by the electric field E in the time interval δt is

$$\delta\epsilon = -eEv_g\,\delta t \ . \tag{2}$$

We observe that

$$\delta\epsilon = (d\epsilon/dk)\delta k = \hbar v_g\,\delta k \ , \tag{3}$$

using (1). On comparing (2) with (3) we have

$$\delta k = -(eE/\hbar)\delta t \ , \tag{4}$$

whence $\hbar dk/dt = -eE$.

We may write (4) in terms of the external force $\mathbf{F}$ as

$$\boxed{\hbar\frac{d\mathbf{k}}{dt} = \mathbf{F} \ .} \tag{5}$$

This is an important relation: in a crystal $\hbar d\mathbf{k}/dt$ is equal to the external force on the electron. In free space $d(m\mathbf{v})/dt$ is equal to the force. We have not overthrown Newton's second law of motion: the electron in the crystal is subject to forces from the crystal lattice as well as from external sources.

The force term in (5) also includes the electric field and the Lorentz force on an electron in a magnetic field, under ordinary conditions where the magnetic field is not so strong that it breaks down the band structure. Thus the equation of motion of an electron of group velocity $\mathbf{v}$ in a constant magnetic field $\mathbf{B}$ is

$$\text{(CGS)} \quad \hbar\frac{d\mathbf{k}}{dt} = -\frac{e}{c}\mathbf{v}\times\mathbf{B} \ ; \qquad \text{(SI)} \quad \hbar\frac{d\mathbf{k}}{dt} = -e\mathbf{v}\times\mathbf{B} \tag{6}$$

where the right-hand side of each equation is the Lorentz force on the electron. With the group velocity $\mathbf{v} = \hbar^{-1}\text{grad}_{\mathbf{k}}\epsilon$, the rate of change of the wavevector is

$$\text{(CGS)} \quad \frac{d\mathbf{k}}{dt} = -\frac{e}{\hbar^2 c}\nabla_{\mathbf{k}}\epsilon\times\mathbf{B} \ ; \qquad \text{(SI)} \quad \frac{d\mathbf{k}}{dt} = -\frac{e}{\hbar^2}\nabla_{\mathbf{k}}\epsilon\times\mathbf{B} \tag{7}$$

where now both sides of the equation refer to the coordinates in $\mathbf{k}$ space.

We see from the vector cross-product in (7) that in a magnetic field an electron moves in $\mathbf{k}$ space in a direction normal to the direction of the gradient of the energy ϵ, so that **the electron moves on a surface of constant energy**. The value of the projection $k_{\mathbf{B}}$ of $\mathbf{k}$ on $\mathbf{B}$ is constant during the motion. The motion in $\mathbf{k}$ space is on a plane normal to the direction of $\mathbf{B}$, and the orbit is defined by the intersection of this plane with a surface of constant energy.

Physical Derivation of $\hbar\dot{\mathbf{k}} = \mathbf{F}$

We consider the Bloch eigenfunction $\psi_{\mathbf{k}}$ belonging to the energy eigenvalue $\epsilon_{\mathbf{k}}$ and wavevector $\mathbf{k}$:

$$\psi_{\mathbf{k}} = \sum_{\mathbf{G}} C(\mathbf{k} + \mathbf{G}) \exp[i(\mathbf{k} + \mathbf{G}) \cdot \mathbf{r}] \ . \tag{8}$$

The expectation value of the momentum of an electron in the Bloch state $\mathbf{k}$ is

$$\mathbf{P}_{\text{el}} = (\mathbf{k}|-i\hbar\nabla|\mathbf{k}) = \sum_{\mathbf{G}} \hbar(\mathbf{k} + \mathbf{G})|C(\mathbf{k} + \mathbf{G})|^2 = \hbar(\mathbf{k} + \sum_{\mathbf{G}} \mathbf{G}|C(\mathbf{k} + \mathbf{G})|^2) \ , \tag{9}$$

using $\Sigma|C(\mathbf{k} + \mathbf{G})|^2 = 1$.

We examine the transfer of momentum between the electron and the lattice when the state $\mathbf{k}$ of the electron is changed to $\mathbf{k} + \Delta\mathbf{k}$ by the application of an external force. We imagine an insulating crystal electrostatically neutral except for a single electron in the state $\mathbf{k}$ of an otherwise empty band.

We suppose that a weak external force is applied for a time interval such that the total impulse given to the entire crystal system is $\mathbf{J} = \int \mathbf{F}\, dt$. If the conduction electron were free ($m^* = m$), the total momentum imparted to the crystal system by the impulse would appear in the change of momentum of the conduction electron:

$$\mathbf{J} = \Delta\mathbf{p}_{\text{tot}} = \Delta\mathbf{p}_{\text{el}} = \hbar\Delta\mathbf{k} \ . \tag{10}$$

The neutral crystal suffers no net interaction with the electric field, either directly or indirectly through the free electron.

If the conduction electron interacts with the periodic potential of the crystal lattice, we must have

$$\mathbf{J} = \Delta\mathbf{p}_{\text{tot}} = \Delta\mathbf{p}_{\text{lat}} + \Delta\mathbf{p}_{\text{el}} \ . \tag{11}$$

From the result (9) for $\mathbf{p}_{\text{el}}$ we have

$$\Delta\mathbf{p}_{\text{el}} = \hbar\Delta\mathbf{k} + \sum_{\mathbf{G}} \hbar\mathbf{G}[(\nabla_{\mathbf{k}}|C(\mathbf{k} + \mathbf{G})|^2) \cdot \Delta\mathbf{k}] \ . \tag{12}$$

The change $\Delta\mathbf{p}_{\text{lat}}$ in the lattice momentum resulting from the change of state of the electron may be derived by an elementary physical consideration. An electron reflected by the lattice transfers momentum to the lattice. If an incident electron with plane wave component of momentum $\hbar\mathbf{k}$ is reflected with momentum $\hbar(\mathbf{k} + \mathbf{G})$, the lattice acquires the momentum $-\hbar\mathbf{G}$, as required by momentum conservation. The momentum transfer to the lattice when the state $\psi_{\mathbf{k}}$ goes over to $\psi_{\mathbf{k}+\Delta\mathbf{k}}$ is

$$\Delta\mathbf{p}_{\text{lat}} = -\hbar\sum_{\mathbf{G}} \mathbf{G}[(\nabla_{\mathbf{k}}|C(\mathbf{k} + \mathbf{G})|^2 \cdot \Delta\mathbf{k}] \ , \tag{13}$$

because the portion

$$\nabla_{\mathbf{k}}|C(\mathbf{k}+\mathbf{G})|^2 \cdot \Delta\mathbf{k} \tag{14}$$

of each individual component of the initial state is reflected during the state change $\Delta\mathbf{k}$.

The total momentum change is therefore

$$\Delta\mathbf{p}_{\text{el}} + \Delta\mathbf{p}_{\text{lat}} = \mathbf{J} = \hbar\Delta\mathbf{k} \ , \tag{15}$$

exactly as for free electrons, Eq. (10). Thus from the definition of $\mathbf{J}$, we have

$$\hbar d\mathbf{k}/dt = \mathbf{F} \ , \tag{16}$$

derived in (5) by a different method. A rigorous derivation of (16) by an entirely different method is given in Appendix E.

Holes

The properties of vacant orbitals in an otherwise filled band are important in semiconductor physics and in solid state electronics. Vacant orbitals in a band are commonly called holes, and without holes there would be no transistors. A hole acts in applied electric and magnetic fields as if it has a positive charge $+e$. The reason is given in five steps in the boxes that follow.

1. $$\mathbf{k}_h = -\mathbf{k}_e \ . \tag{17}$$

The total wavevector of the electrons in a filled band is zero: $\Sigma\mathbf{k} = 0$, where the sum is over all states in a Brillouin zone. This result follows from the geometrical symmetry of the Brillouin zone: every fundamental lattice type has symmetry under the inversion operation $\mathbf{r} \rightarrow -\mathbf{r}$ about any lattice point; it follows that the Brillouin zone of the lattice also has inversion symmetry. If the band is filled all pairs of orbitals $\mathbf{k}$ and $-\mathbf{k}$ are filled, and the total wavevector is zero.

If an electron is missing from an orbital of wavevector $\mathbf{k}_e$, the total wavevector of the system is $-\mathbf{k}_e$ and is attributed to the hole. This result is surprising: the electron is missing from $\mathbf{k}_e$ and the position of the hole is usually indicated graphically as situated at $\mathbf{k}_e$, as in Fig. 7. But the true wavevector $\mathbf{k}_h$ of the hole is $-\mathbf{k}_e$, which is the wavevector of the point G if the hole is at E. The wavevector $-\mathbf{k}_e$ enters into selection rules for photon absorption.

The hole is an alternate description of a band with one missing electron, and we either say that the hole has wavevector $-\mathbf{k}_e$ or that the band with one missing electron has total wavevector $-\mathbf{k}_e$.

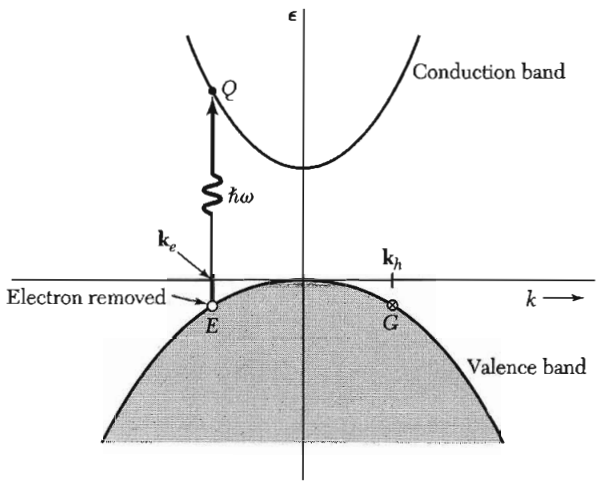


Figure 7 Absorption of a photon of energy $\hbar\omega$ and negligible wavevector takes an electron from E in the filled valence band to Q in the conduction band. If $\mathbf{k}_e$ was the wavevector of the electron at E, it becomes the wavevector of the electron at Q. The total wavevector of the valence band after the absorption is $-\mathbf{k}_e$, and this is the wavevector we must ascribe to the hole if we describe the valence band as occupied by one hole. Thus $\mathbf{k}_h = -\mathbf{k}_e$; the wavevector of the hole is the same as the wavevector of the electron which remains at G. For the entire system the total wavevector after the absorption of the photon is $\mathbf{k}_e + \mathbf{k}_h = 0$, so that the total wavevector is unchanged by the absorption of the photon and the creation of a free electron and free hole.

2. $$\epsilon_h(\mathbf{k}_h) = -\epsilon_e(\mathbf{k}_e) \ . \tag{18}$$

Here the zero of energy of the valence band is at the top of the band. The lower in the band the missing electron lies, the higher the energy of the system. The energy of the hole is opposite in sign to the energy of the missing electron, because it takes more work to remove an electron from a low orbital than from a high orbital. Thus if the band is symmetric,[1] $\epsilon_e(\mathbf{k}_e) = \epsilon_e(-\mathbf{k}_e) = -\epsilon_h(-\mathbf{k}_e) = -\epsilon_h(\mathbf{k}_h)$. We construct in Fig. 8 a band scheme to represent the properties of a hole. This hole band is a helpful representation because it appears right side up.

3. $$\mathbf{v}_h = \mathbf{v}_e \ . \tag{19}$$

The velocity of the hole is equal to the velocity of the missing electron. From Fig. 8 we see that $\nabla\epsilon_h(\boldsymbol{k}_h) = \nabla\epsilon_e(\boldsymbol{k}_e)$, so that $\boldsymbol{v}_h(\boldsymbol{k}_h) = \boldsymbol{v}_e(\boldsymbol{k}_e)$.

[1]Bands are always symmetric under the inversion $\mathbf{k} \rightarrow -\mathbf{k}$ if the spin-orbit interaction is neglected. Even with spin-orbit interaction, bands are always symmetric if the crystal structure permits the inversion operation. Without a center of symmetry, but with spin-orbit interaction, the bands are symmetric if we compare subbands for which the spin direction is reversed: $\epsilon(\boldsymbol{k}, \uparrow) = \epsilon(-\boldsymbol{k}, \downarrow)$. See *QTS*, Chapter 9.

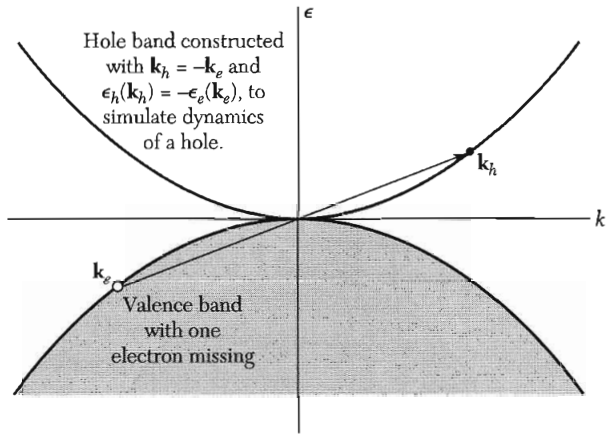


Figure 8 The upper half of the figure shows the hole band that simulates the dynamics of a hole, constructed by inversion of the valence band in the origin. The wavevector and energy of the hole are equal, but opposite in sign, to the wavevector and energy of the empty electron orbital in the valence band. We do not show the disposition of the electron removed from the valence band at $\mathbf{k}_e$.

4. $$m_h = -m_e \ . \tag{20}$$

We show below that the effective mass is inversely proportional to the curvature $d^2\epsilon/dk^2$, and for the hole band this has the opposite sign to that for an electron in the valence band. Near the top of the valence band m_e is negative, so that m_h is positive.

5. $$\hbar \frac{d\mathbf{k}_h}{dt} = e(\mathbf{E} + \frac{1}{c}\mathbf{v}_h \times \mathbf{B}) \ . \tag{21}$$

This comes from the equation of motion

(CGS) $$\hbar \frac{d\mathbf{k}_e}{dt} = -e(\mathbf{E} + \frac{1}{c}\mathbf{v}_e \times \mathbf{B}) \tag{22}$$

that applies to the missing electron when we substitute $-\mathbf{k}_h$ for $\mathbf{k}_e$ and $\mathbf{v}_h$ for $\mathbf{v}_e$. **The equation of motion for a hole is that of a particle of positive charge *e*.** The positive charge is consistent with the electric current carried by the valence band of Fig. 9: the current is carried by the unpaired electron in the orbital G:

$$\mathbf{j} = (-e)\mathbf{v}(G) = (-e)[-\mathbf{v}(E)] = e\mathbf{v}(E) \ , \tag{23}$$

which is just the current of a positive charge moving with the velocity ascribed to the missing electron at E. The current is shown in Fig. 10.

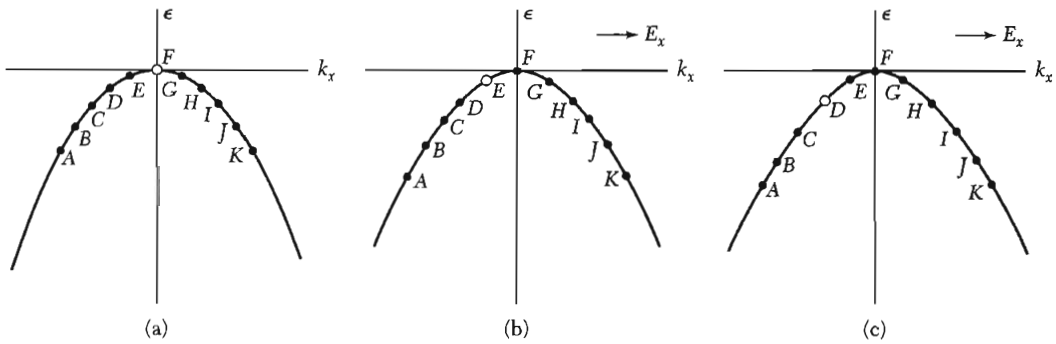


Figure 9 (a) At $t = 0$ all states are filled except F at the top of the band; the velocity v_x is zero at F because $d\epsilon/dk_x = 0$. (b) An electric field E_x is applied in the $+x$ direction. The force on the electrons is in the $-k_x$ direction and all electrons make transitions together in the $-k_x$ direction, moving the hole to the state E. (c) After a further interval the electrons move farther along in k space and the hole is now at D.

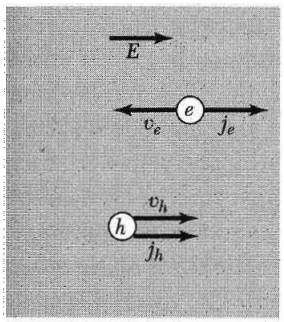


Figure 10 Motion of electrons in the conduction band and holes in the valence band in the electric field E. The hole and electron drift velocities are in opposite directions, but their electric currents are in the same direction, the direction of the electric field.

Effective Mass

When we look at the energy-wavevector relation $\epsilon = (\hbar^2/2m)k^2$ for free electrons, we see that the coefficient of k^2 determines the curvature of ϵ versus k. Turned about, we can say that $1/m$, the reciprocal mass, determines the curvature. For electrons in a band there can be regions of unusually high curvature near the band gap at the zone boundary, as we see from the solutions in Chapter 7 of the wave equation near the zone boundary. If the energy gap is small in comparison with the free electron energy λ at the boundary, the curvature is enhanced by the factor λ/E_g.

In semiconductors the band width, which is like the free electron energy, is of the order of 20 eV, while the band gap is of the order of 0.2 to 2 eV. Thus the reciprocal mass is enhanced by a factor 10 to 100, and the effective mass is reduced to 0.1–0.01 of the free electron mass. These values apply near the band gap; as we go away from the gap the curvatures and the masses are likely to approach those of free electrons.

To summarize the solutions of Chapter 7 for U positive, an electron near the lower edge of the second band has an energy that may be written as

$$\epsilon(K) = \epsilon_c + (\hbar^2/2m_e)K^2 \ ; \qquad m_e/m = 1/[(2\lambda/U)-1] \ . \tag{24}$$

Here K is the wavevector measured from the zone boundary, and m_e denotes the effective mass of the electron near the edge of the second band. An electron near the top of the first band has the energy

$$\epsilon(K) = \epsilon_v - (\hbar^2/2m_h)K^2 \ ; \qquad m_h/m = 1/[(2\lambda/U) + 1] \ . \tag{25}$$

The curvature and hence the mass will be negative near the top of the first band, but we have introduced a minus sign into (25) in order that the symbol m_h for the hole mass will have a positive value—see (20) above.

The crystal does not weigh any less if the effective mass of a carrier is less than the free electron mass, nor is Newton's second law violated for the crystal *taken as a whole*, ions plus carriers. The important point is that an electron in a periodic potential is accelerated relative to the lattice in an applied electric or magnetic field as if the mass of the electron were equal to an effective mass which we now define.

We differentiate the result (1) for the group velocity to obtain

$$\frac{dv_g}{dt} = \hbar^{-1}\frac{d^2\epsilon}{dk\,dt} = \hbar^{-1}\left(\frac{d^2\epsilon}{dk^2}\frac{dk}{dt}\right) \ . \tag{26}$$

We know from (5) that $dk/dt = F/\hbar$, whence

$$\frac{dv_g}{dt} = \left(\frac{1}{\hbar^2}\frac{d^2\epsilon}{dk^2}\right)F \ ; \qquad \text{or} \qquad F = \frac{\hbar^2}{d^2\epsilon/dk^2}\frac{dv_g}{dt} \ . \tag{27}$$

If we identify $\hbar^2/(d^2\epsilon/dk^2)$ as a mass, then (27) assumes the form of Newton's second law. We define the **effective mass** m^* by

$$\boxed{\frac{1}{m^*} = \frac{1}{\hbar^2}\frac{d^2\epsilon}{dk^2} \ .} \tag{28}$$

It is easy to generalize this to take account of an anisotropic electron energy surface, as for electrons in Si or Ge. We introduce the components of the reciprocal effective mass tensor

$$\left(\frac{1}{m^*}\right)_{\mu\nu} = \frac{1}{\hbar^2}\frac{d^2\,\epsilon_k}{dk_\mu\,dk_\nu} \ ; \qquad \frac{dv_\mu}{dt} = \left(\frac{1}{m^*}\right)_{\mu\nu}F_\nu \ , \tag{29}$$

where μ, ν are Cartesian coordinates.

Physical Interpretation of the Effective Mass

How can an electron of mass m when put into a crystal respond to applied fields as if the mass were m^*? It is helpful to think of the process of Bragg reflection of electron waves in a lattice. Consider the weak interaction approximation treated in Chapter 7. Near the bottom of the lower band the orbital is represented quite adequately by a plane wave $\exp(ikx)$ with momentum $\hbar k$; the wave component $\exp[i(k - G)x]$ with momentum $\hbar(k-G)$ is small and

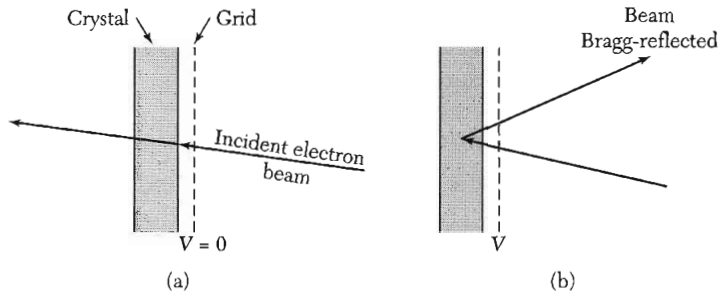


Figure 11 Explanation of negative effective masses which occur near, but below, a Brillouin zone boundary. In (a) the energy of the electron beam incident on a thin crystal is slightly too low to satisfy the condition for Bragg reflection and the beam is transmitted through the crystal. The application of a small voltage across the grid may, as in (b), cause the Bragg condition to be satisfied, and the electron beam will then be reflected from the appropriate set of crystal planes.

increases only slowly as k is increased, and in this region $m^* \simeq m$. An increase in the reflected component $\exp[i(k - G)x]$ as k is increased represents momentum transfer to the electron from the lattice.

Near the boundary the reflected component is quite large; at the boundary it becomes equal in amplitude to the forward component, at which point the eigenfunctions are standing waves, rather than running waves. Here the momentum component $\hbar(-\frac{1}{2}G)$ cancels the momentum component $\hbar(\frac{1}{2}G)$.

A single electron in an energy band may have positive or negative effective mass: the states of positive effective mass occur near the bottom of a band because positive effective mass means that the band has upward curvature ($d^2\epsilon/dk^2$ is positive). States of negative effective mass occur near the top of the band. A negative effective mass means that on going from state k to state $k + \Delta k$, the momentum transfer to the lattice from the electron is larger than the momentum transfer from the applied force to the electron. Although k is increased by Δk by the applied electric field, the approach to Bragg reflection can give an overall decrease in the forward momentum of the electron; when this happens the effective mass is negative (Fig. 11).

As we proceed in the second band away from the boundary, the amplitude of $\exp[i(k - G)x]$ decreases rapidly and m^* assumes a small positive value. Here the increase in electron velocity resulting from a given external impulse is larger than that which a free electron would experience. The lattice makes up the difference through the reduced recoil it experiences when the amplitude of $\exp[i(k - G)x]$ is diminished.

If the energy in a band depends only slightly on k, then the effective mass will be very large. That is, $m^*/m \gg 1$ when $d^2\epsilon/dk^2$ is very small. The tight-binding approximation discussed in Chapter 9 gives quick insight into the formation of narrow bands. If the wavefunctions centered on neighboring atoms overlap very little, then the overlap integral is small; the width of the band

narrow, and the effective mass large. The overlap of wavefunctions centered on neighboring atoms is small for the inner or core electrons. The *4f* electrons of the rare earth metals, for example, overlap very little.

Effective Masses in Semiconductors

In many semiconductors it has been possible to determine by cyclotron resonance the effective masses of carriers in the conduction and valence bands near the band edges. The determination of the energy surface is equivalent to a determination of the effective mass tensor (29). Cyclotron resonance in a semiconductor is carried out with centimeter wave or millimeter wave radiation at low carrier concentration.

The current carriers are accelerated in helical orbits about the axis of a static magnetic field. The angular rotation frequency ω_c is

$$\text{(CGS)} \quad \omega_c = \frac{eB}{m^*c} \,, \qquad \boxed{\text{(SI)} \quad \omega_c = \frac{eB}{m^*}} \qquad (30)$$

where m^* is the appropriate cyclotron effective mass. Resonant absorption of energy from an rf electric field perpendicular to the static magnetic field (Fig. 12) occurs when the rf frequency is equal to the cyclotron frequency. Holes and electrons rotate in opposite senses in a magnetic field.

We consider the experiment for $m^*/m = 0.1$. At $f_c = 24$ GHz, or $\omega_c = 1.5 \times 10^{11}\ \text{s}^{-1}$, we have $B = 860$ G at resonance. The line width is determined by the collision relaxation time τ, and to obtain a distinctive resonance it is necessary that $\omega_c\tau \geq 1$. The mean free path must be long enough to permit the average carrier to get one radian around a circle between collisions. The requirements are met with the use of higher frequency radiation and higher magnetic fields, with high purity crystals in liquid helium.

In direct-gap semiconductors with band edges at the center of the Brillouin zone, the bands have the structure shown in Fig. 13. The conduction band edge is spherical with the effective mass m_0:

$$\epsilon_c = E_g + \hbar^2k^2/2m_e \,, \qquad (31)$$

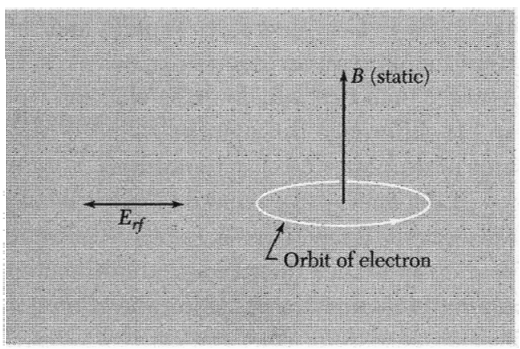


Figure 12 Arrangement of fields in a cyclotron resonance experiment in a semiconductor. The sense of the circulation is opposite for electrons and holes.

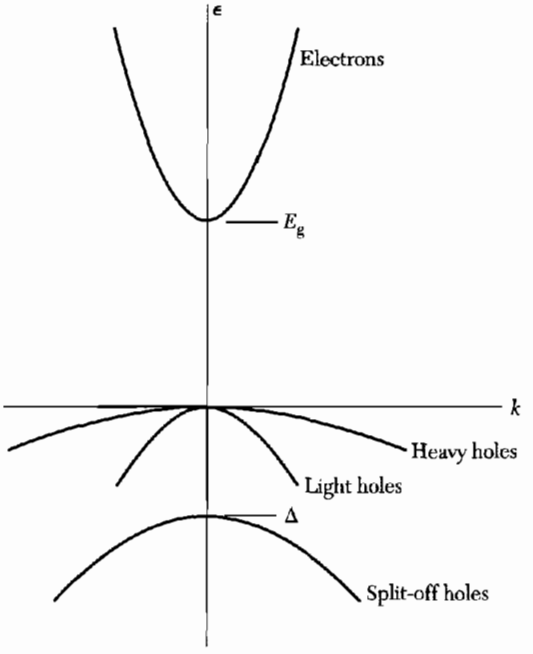


Figure 13 Simplified view of the band edge structure of a direct-gap semiconductor.

Table 2 Effective masses of electrons and holes in direct-gap semiconductors

Crystal	Electron m_e/m	Heavy hole m_{hh}/m	Light hole m_{lh}/m	Split-off hole m_{soh}/m	Spin-orbit Δ, eV
InSb	0.015	0.39	0.021	(0.11)	0.82
InAs	0.026	0.41	0.025	0.08	0.43
InP	0.073	0.4	(0.078)	(0.15)	0.11
GaSb	0.047	0.3	0.06	(0.14)	0.80
GaAs	0.066	0.5	0.082	0.17	0.34
Cu_2O	0.99	—	0.58	0.69	0.13

referred to the valence band edge. The valence bands are characteristically threefold near the edge, with the heavy hole *hh* and light hole *lh* bands degenerate at the center, and a band *soh* split off by the spin-orbit splitting Δ:

$$\epsilon_v(hh) \cong -\hbar^2k^2/2m_{hh} \ ; \qquad \epsilon_v(lh) \cong -\hbar^2k^2/2m_{lh} \ ;$$
$$\epsilon_v(soh) \cong -\Delta - \hbar^2k^2/2m_{soh} \ . \tag{32}$$

Values of the mass parameters are given in Table 2. The forms (32) are only approximate, because even close to $k = 0$ the heavy and light hole bands are not spherical—see the discussion below for Ge and Si.

The perturbation theory of band edges (Problem 9.8) suggests that the electron effective mass should be proportional to the band gap, approximately,

for a direct gap crystal. We use Tables 1 and 2 to find the fairly constant values $m_e/(mE_g) = 0.063$, 0.060, and 0.051 in $(\text{eV})^{-1}$ for the series InSb, InAs, and InP, in agreement with this suggestion.

Silicon and Germanium

The conduction and valence bands of germanium are shown in Fig. 14, based on a combination of theoretical and experimental results. The valence band edge in both Si and Ge is at $\mathbf{k} = 0$ and is derived from $p_{3/2}$ and $p_{1/2}$ states of the free atoms, as is clear from the tight-binding approximation (Chapter 9) to the wavefunctions.

The $p_{3/2}$ level is fourfold degenerate as in the atom; the four states correspond to m_J values $\pm\frac{3}{2}$ and $\pm\frac{1}{2}$. The $p_{1/2}$ level is doubly degenerate, with $m_J = \pm\frac{1}{2}$. The $p_{3/2}$ states are higher in energy than the $p_{1/2}$ states; the energy difference Δ is a measure of the spin-orbit interaction.

The valence band edges are not simple. Holes near the band edge are characterized by two effective masses, light and heavy. These arise from the two bands formed from the $p_{3/2}$ level of the atom. There is also a band formed from the $p_{1/2}$ level, split off from the $p_{3/2}$ level by the spin-orbit interaction. The energy surfaces are not spherical, but warped (*QTS*, p. 271):

$$\epsilon(\mathbf{k}) = Ak^2 \pm [B^2k^4 + C^2(k_x^2k_y^2 + k_y^2k_z^2 + k_z^2k_x^2)]^{1/2} \tag{33}$$

The choice of sign distinguishes the two masses. The split-off band has $\epsilon(k) = -\Delta + Ak^2$. The experiments give, in units $\hbar^2/2m$,

Si:	$A = -4.29$;	$\lvert B\rvert = 0.68$;	$\lvert C\rvert = 4.87$;	$\Delta = 0.044$ eV
Ge:	$A = -13.38$;	$\lvert B\rvert = 8.48$;	$\lvert C\rvert = 13.15$;	$\Delta = 0.29$ eV

Roughly, the light and heavy holes in germanium have masses 0.043 m and 0.34 m; in silicon 0.16 m and 0.52 m; in diamond 0.7 m and 2.12 m.

The conduction band edges in Ge are at the equivalent points L of the Brillouin zone, Fig. 15a. Each band edge has a spheroidal energy surface oriented along a $\langle 111\rangle$ crystal axis, with a longitudinal mass $m_l = 1.59\ m$ and a transverse mass $m_t = 0.082\ m$. For a static magnetic field at an angle θ with the longitudinal axis of a spheroid, the effective cyclotron mass m_c is

$$\frac{1}{m_c^2} = \frac{\cos^2\theta}{m_t^2} + \frac{\sin^2\theta}{m_t m_l} \tag{34}$$

Results for Ge are shown in Fig. 16.

In silicon the conduction band edges are spheroids oriented along the equivalent $\langle 100\rangle$ directions in the Brillouin zone, with mass parameters $m_l = 0.92\ m$ and $m_t = 0.19\ m$, as in Fig. 17a. The band edges lie along the lines labeled Δ in the zone of Fig. 15a, a little way in from the boundary points X.

In GaAs we have $A = -6.98$, $B = -4.5$, $|C| = 6.2$, $\Delta = 0.341$ eV. The band structure is shown in Fig. 17b. It has a direct band gap with an isotropic conduction electron mass of 0.067 m.

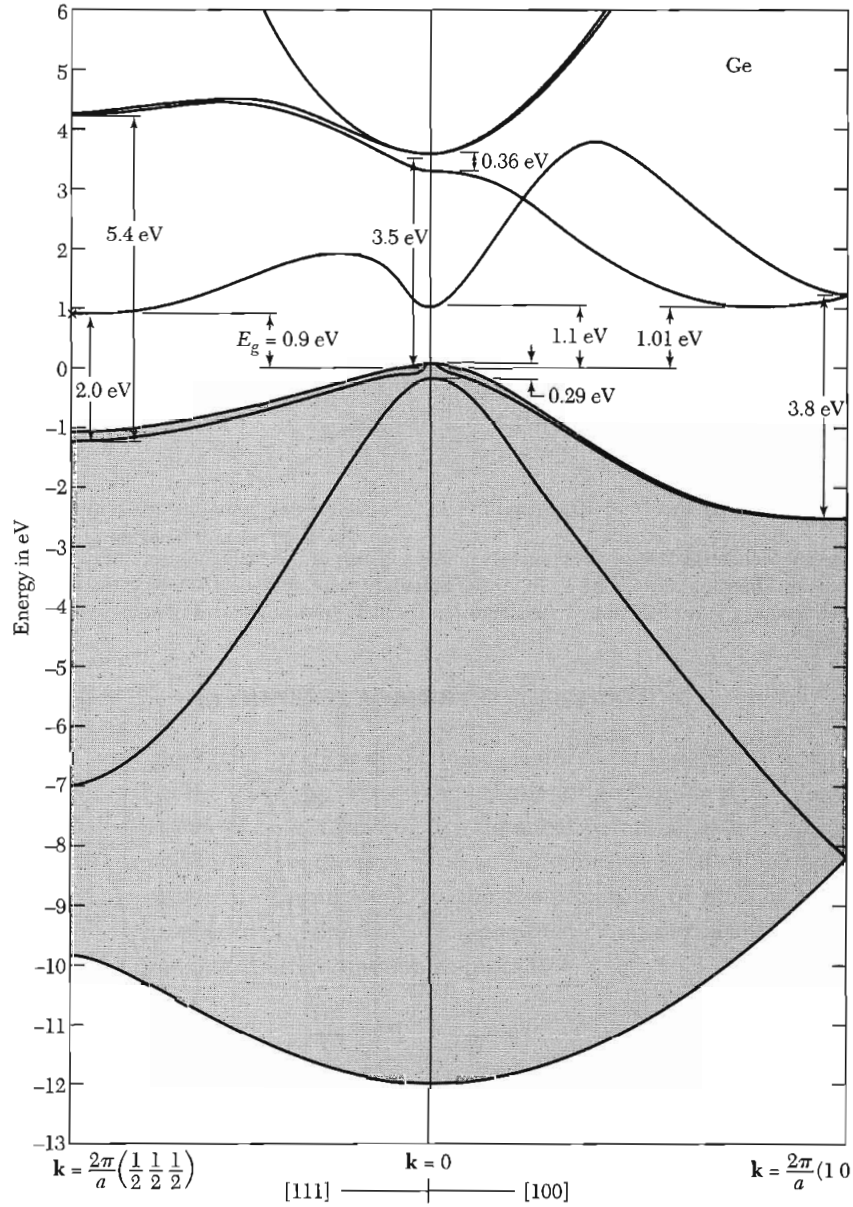


Figure 14 Calculated band structure of germanium, after C. Y. Fong. The general features are in good agreement with experiment. The four valence bands are shown in gray. The fine structure of the valence band edge is caused by spin-orbit splitting. The energy gap is indirect; the conduction band edge is at the point $(2\pi/a)(\frac{1}{2}\,\frac{1}{2}\,\frac{1}{2})$. The constant energy surfaces around this point are ellipsoidal.

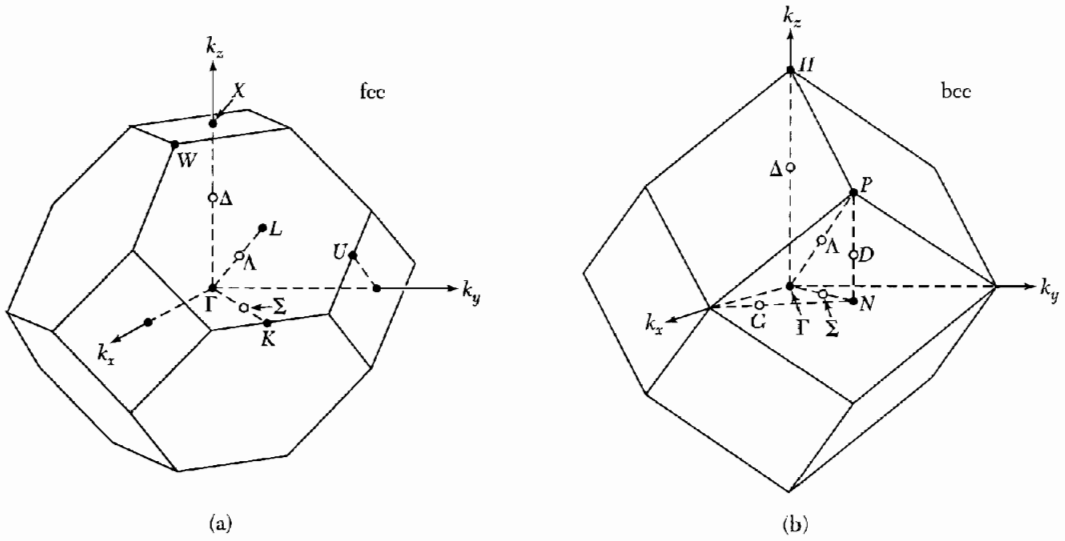


Figure 15 Standard labels of the symmetry points and axes of the Brillouin zones of the fcc and bcc lattices. The zone centers are Γ. In (a) the boundary point at $(2\pi/a)(100)$ is X; the boundary point at $(2\pi/a)(\frac{1}{2}\,\frac{1}{2}\,\frac{1}{2})$ is L; the line Δ runs between Γ and X. In (b) the corresponding symbols are H, P, and Δ.

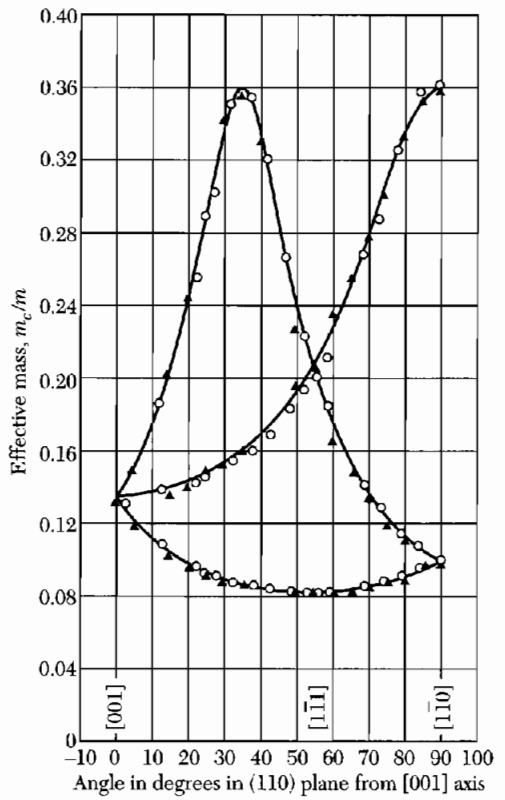
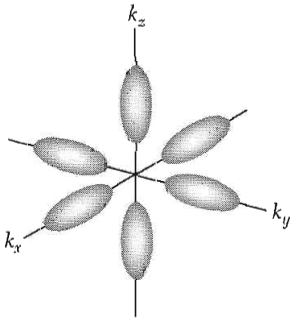
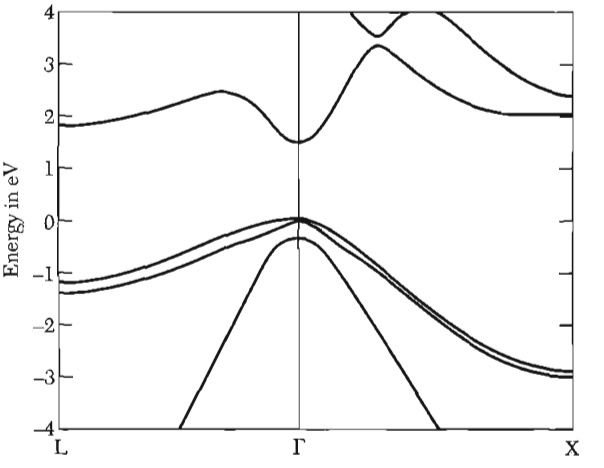


Figure 16 Effective cyclotron mass of electrons in germanium at 4 K for magnetic field directions in a (110) plane. There are four independent mass spheroids in Ge, one along each [111] axis, but viewed in the (110) plane two spheroids always appear equivalent. (After Dresselhaus, Kip, and Kittel.)



(a)

Figure 17a Constant energy ellipsoids for electrons in silicon, drawn for $m_l/m_t = 5$.



(b)

Figure 17b Band structure of GaAs, after S. G. Louie.

INTRINSIC CARRIER CONCENTRATION

We want the concentration of intrinsic carriers as a function of temperature, in terms of the band gap. We do the calculation for simple parabolic band edges. We first calculate in terms of the chemical potential μ the number of electrons excited to the conduction band at temperature T. In semiconductor physics μ is called the Fermi level. At the temperatures of interest we may suppose for the conduction band of a semiconductor that $\epsilon - \mu \gg k_B T$, so that the Fermi-Dirac distribution function reduces to

$$f_e \simeq \exp[(\mu - \epsilon)/k_B T] \ . \tag{35}$$

This is the probability that a conduction electron orbital is occupied, in an approximation valid when $f_e \ll 1$.

The energy of an electron in the conduction band is

$$\epsilon_k = E_c + \hbar^2 k^2 / 2m_e \ , \tag{36}$$

where E_c is the energy at the conduction band edge, as in Fig. 18. Here m_e is the effective mass of an electron. Thus from (6.20) the density of states at ϵ is

$$D_e(\epsilon) = \frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2}(\epsilon - E_c)^{1/2} \ . \tag{37}$$

The concentration of electrons in the conduction band is

$$n = \int_{E_c}^{\infty} D_e(\epsilon) f_e(\epsilon) d\epsilon = \frac{1}{2\pi^2}\left(\frac{2m_e}{\hbar^2}\right)^{3/2} \exp(\mu/k_B T) \times \int_{E_c}^{\infty} (\epsilon - E_c)^{1/2} \exp(-\epsilon/k_B T) d\epsilon \ , \tag{38}$$

which integrates to give

$$n = 2\left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{3/2} \exp[(\mu - E_c)/k_B T] \ . \tag{39}$$

The problem is solved for n when μ is known. It is useful to calculate the equilibrium concentration of holes p. The distribution function f_h for holes is related to the electron distribution function f_e by $f_h = 1 - f_e$, because a hole is the absence of an electron. We have

$$f_h = 1 - \frac{1}{\exp[(\epsilon - \mu)/k_B T] + 1} = \frac{1}{\exp[(\mu - \epsilon)/k_B T] + 1} \cong \exp[(\epsilon - \mu)/k_B T] \ , \tag{40}$$

provided $(\mu - \epsilon) \gg k_B T$.

If the holes near the top of the valence band behave as particles with effective mass m_h, the density of hole states is given by

$$D_h(\epsilon) = \frac{1}{2\pi^2}\left(\frac{2m_h}{\hbar^2}\right)^{3/2} (E_v - \epsilon)^{1/2} \ , \tag{41}$$

where E_v is the energy at the valence band edge. Proceeding as in (38) we obtain

$$p = \int_{-\infty}^{E_v} D_h(\epsilon) f_h(\epsilon) d\epsilon = 2\left(\frac{m_h k_B T}{2\pi\hbar^2}\right)^{3/2} \exp[(E_v - \mu)/k_B T] \tag{42}$$

for the concentration p of holes in the valence band.

We multiply together the expressions for n and p to obtain the equilibrium relation, with the energy gap $E_g = E_c - E_v$ as in Fig. 18,

$$np = 4\left(\frac{k_B T}{2\pi\hbar^2}\right)^3 (m_e m_h)^{3/2} \exp(-E_g/k_B T) \ . \tag{43}$$

This useful result does not involve the Fermi level μ. At 300 K the value of np is 2.10×10^{19} cm^{-6}, 2.89×10^{26} cm^{-6}, and 6.55×10^{12} cm^{-6}, for the actual band structures of Si, Ge, and GaAs, respectively.

We have nowhere assumed in the derivation that the material is intrinsic: the result holds for impurity ionization as well. The only assumption made is that the distance of the Fermi level from the edge of both bands is large in comparison with $k_B T$.

A simple kinetic argument shows why the product np is constant at a given temperature. Suppose that the equilibrium population of electrons and holes

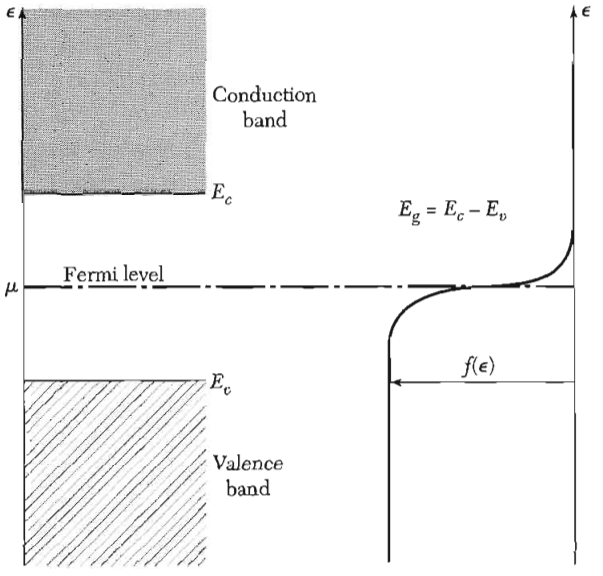


Figure 18 Energy scale for statistical calculations. The Fermi distribution function is shown on the same scale, for a temperature $k_BT \ll E_g$. The Fermi level μ is taken to lie well within the band gap, as for an intrinsic semiconductor. If $\epsilon = \mu$, then $f = \frac{1}{2}$.

is maintained by black-body photon radiation at temperature T. The photons generate electron-hole pairs at a rate $A(T)$, while $B(T)np$ is the rate of the recombination reaction $e + h =$ photon. Then

$$dn/dt = A(T) - B(T)np = dp/dt \ . \tag{44}$$

In equilibrium $dn/dt = 0$, $dp/dt = 0$, whence $np = A(T)/B(T)$.

Because the product of the electron and hole concentrations is a constant independent of impurity concentration at a given temperature, the introduction of a small proportion of a suitable impurity to increase n, say, must decrease p. This result is important in practice—we can reduce the total carrier concentration $n + p$ in an impure crystal, sometimes enormously, by the controlled introduction of suitable impurities. Such a reduction is called **compensation**.

In an intrinsic semiconductor the number of electrons is equal to the number of holes, because the thermal excitation of an electron leaves behind a hole in the valence band. Thus from (43) we have, letting the subscript i denote intrinsic and $E_g = E_c - E_v$,

$$n_i = p_i = 2\left(\frac{k_BT}{2\pi\hbar^2}\right)^{3/2}(m_em_h)^{3/4}\exp(-E_g/2k_BT) \ . \tag{45}$$

The intrinsic carrier concentration depends exponentially on $E_g/2k_BT$, where E_g is the energy gap. We set (39) equal to (42) to obtain, for the Fermi level as measured from the top of the valence band,

$$\exp(2\mu/k_BT) = (m_h/m_e)^{3/2}\exp(E_g/k_BT) \ ; \tag{46}$$

$$\mu = \tfrac{1}{2}E_g + \tfrac{3}{4}k_BT\ln(m_h/m_e) \ . \tag{47}$$

If $m_h = m_e$, then $\mu = \frac{1}{2}E_g$ and the Fermi level is in the middle of the forbidden gap.

Intrinsic Mobility

The mobility is the magnitude of the drift velocity of a charge carrier per unit electric field:

$$\mu = |v|/E \ . \tag{48}$$

The mobility is defined to be positive for both electrons and holes, although their drift velocities are opposite in a given field. By writing μ_e or μ_h with subscripts for the electron or hole mobility we can avoid any confusion between μ as the chemical potential and as the mobility.

The electrical conductivity is the sum of the electron and hole contributions:

$$\sigma = (ne\mu_e + pe\mu_h) \ , \tag{49}$$

where n and p are the concentrations of electrons and holes. In Chapter 6 the drift velocity of a charge q was found to be $v = q\tau E/m$, whence

$$\mu_e = e\tau_e/m_e \ ; \qquad \mu_h = e\tau_h/m_h \ , \tag{50}$$

where τ is the collision time.

The mobilities depend on temperature as a modest power law. The temperature dependence of the conductivity in the intrinsic region will be dominated by the exponential dependence $\exp(-E_g/2k_BT)$ of the carrier concentration, Eq. (45).

Table 3 gives experimental values of the mobility at room temperature. The mobility in SI units is expressed in m^2/V-s and is 10^{-4} of the mobility in practical units. For most substances the values quoted are limited by the scattering of carriers by thermal phonons. The hole mobilities typically are smaller than the electron mobilities because of the occurrence of band degeneracy at the valence band edge at the zone center, thereby making possible interband scattering processes that reduce the mobility considerably.

Table 3 Carrier mobilities at room temperature, in cm^2/V-s

Crystal	Electrons	Holes	Crystal	Electrons	Holes
Diamond	1800	1200	GaAs	8000	300
Si	1350	480	GaSb	5000	1000
Ge	3600	1800	PbS	550	600
InSb	800	450	PbSe	1020	930
InAs	30000	450	PbTe	2500	1000
InP	4500	100	AgCl	50	—
AlAs	280	—	KBr (100 K)	100	—
AlSb	900	400	SiC	100	10–20

In some crystals, particularly in ionic crystals, the holes are essentially immobile and get about only by thermally-activated hopping from ion to ion. The principal cause of this "self-trapping" is the lattice distortion associated with the Jahn-Teller effect of degenerate states. The orbital degeneracy necessary for self-trapping is much more frequent for holes than for electrons.

There is a tendency for crystals with small energy gaps at direct band edges to have high values of the electron mobility. Small gaps lead to small effective masses, which favor high mobilities. The highest mobility observed in a bulk semiconductor is 5×10^6 cm^2/V-s in PbTe at 4 K, where the gap is 0.19 eV.

IMPURITY CONDUCTIVITY

Certain impurities and imperfections drastically affect the electrical properties of a semiconductor. The addition of boron to silicon in the proportion of 1 boron atom to 10^5 silicon atoms increases the conductivity of pure silicon at room temperature by a factor of 10^3. In a compound semiconductor a stoichiometric deficiency of one constituent will act as an impurity; such semiconductors are known as **deficit semiconductors**. The deliberate addition of impurities to a semiconductor is called **doping**.

We consider the effect of impurities in silicon and germanium. These elements crystallize in the diamond structure. Each atom forms four covalent bonds, one with each of its nearest neighbors, corresponding to the chemical valence four. If an impurity atom of valence five, such as phosphorus, arsenic, or antimony, is substituted in the lattice in place of a normal atom, there will be one valence electron from the impurity atom left over after the four covalent bonds are established with the nearest neighbors, that is, after the impurity atom has been accommodated in the structure with as little disturbance as possible. Impurity atoms that can give up an electron are called **donors**.

Donor States. The structure in Fig. 19 has a positive charge on the impurity atom (which has lost one electron). Lattice constant studies have verified that the pentavalent impurities enter the lattice by substitution for normal atoms, and not in interstitial positions. The crystal as a whole remains neutral because the electron remains in the crystal.

The extra electron moves in the coulomb potential $e/\epsilon r$ of the impurity ion, where ϵ in a covalent crystal is the static dielectric constant of the medium. The factor $1/\epsilon$ takes account of the reduction in the coulomb force between charges caused by the electronic polarization of the medium. This treatment is valid for orbits large in comparison with the distance between atoms, and for slow motions of the electron such that the orbital frequency is low in comparison with the frequency ω_g corresponding to the energy gap. These conditions are satisfied quite well in Ge and Si by the donor electron of P, As, or Sb.

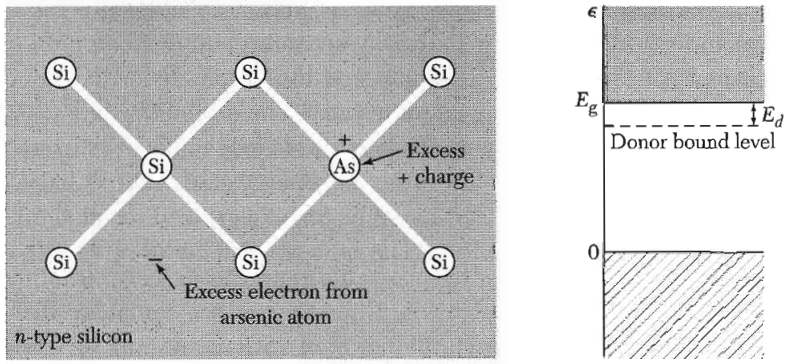


Figure 19 Charges associated with an arsenic impurity atom in silicon. Arsenic has five valence electrons, but silicon has only four valence electrons. Thus four electrons on the arsenic form tetrahedral covalent bonds similar to silicon, and the fifth electron is available for conduction. The arsenic atom is called a **donor** because when ionized it donates an electron to the conduction band.

We estimate the ionization energy of the donor impurity. The Bohr theory of the hydrogen atom may be modified to take into account the dielectric constant of the medium and the effective mass of an electron in the periodic potential of the crystal. The ionization energy of atomic hydrogen is $-e^4m/2\hbar^2$ in CGS and $-e^4m/2(4\pi\epsilon_0\hbar)^2$ in SI.

In the semiconductor with dielectric constant ϵ we replace e^2 by e^2/ϵ and m by the effective mass m_e to obtain

$$\text{(CGS)} \quad E_d = \frac{e^4 m_e}{2\epsilon^2\hbar^2} = \left(\frac{13.6}{\epsilon^2}\frac{m_e}{m}\right) eV \ ; \qquad \text{(SI)} \ E_d = \frac{e^4 m_e}{2(4\pi\epsilon\epsilon_0\hbar)^2} \tag{51}$$

as the donor ionization energy of the semiconductor.

The Bohr radius of the ground state of hydrogen is $\hbar^2/me^2$ in CGS or $4\pi\epsilon_0\hbar^2/me^2$ in SI. Thus the Bohr radius of the donor is

$$\text{(CGS)} \ a_d = \frac{\epsilon\hbar^2}{m_e e^2} = \left(\frac{0.53\epsilon}{m_e/m}\right) \text{Å} \ ; \qquad \text{(SI)} \ a_d = \frac{4\pi\epsilon\epsilon_0\hbar^2}{m_e e^2} \tag{52}$$

The application of impurity state theory to germanium and silicon is complicated by the anisotropic effective mass of the conduction electrons. But the dielectric constant has the more important effect on the donor energy because it enters as the square, whereas the effective mass enters only as the first power.

To obtain a general impression of the impurity levels we use $m_e \approx 0.1\ m$ for electrons in germanium and $m_e \approx 0.2\ m$ in silicon. The static dielectric constant is given in Table 4. The ionization energy of the free hydrogen atom is 13.6 eV. For germanium the donor ionization energy E_d on our model is 5 meV, reduced with respect to hydrogen by the factor $m_e/m\epsilon^2 = 4 \times 10^{-4}$. The corresponding result for silicon is 20 meV. Calculations using the correct

Table 4 Static relative dielectric constant of semiconductors

Crystal	ϵ	Crystal	ϵ
Diamond	5.5	GaSb	15.69
Si	11.7	GaAs	13.13
Ge	15.8	AlAs	10.1
InSb	17.88	AlSb	10.3
InAs	14.55	SiC	10.2
InP	12.37	Cu_2O	7.1

Table 5 Donor ionization energies E_d of pentavalent impurities in germanium and silicon, in meV

	P	As	Sb
Si	45.	49.	39.
Ge	12.0	12.7	9.6

anisotropic mass tensor predict 9.05 meV for germanium and 29.8 meV for silicon. Observed values of donor ionization energies in Si and Ge are given in Table 5. In GaAs donors have $E_d \simeq 6$ meV.

The radius of the first Bohr orbit is increased by $\epsilon m/m_e$ over the value 0.53 Å for the free hydrogen atom. The corresponding radius is (160)(0.53) $\simeq$ 80 Å in germanium and (60)(0.53) $\simeq$ 30 Å in silicon. These are large radii, so that donor orbits overlap at relatively low donor concentrations, compared to the number of host atoms. With appreciable orbit overlap, an "impurity band" is formed from the donor states: see the discussion of the metal-insulator transition in Chapter 14.

The semiconductor can conduct in the impurity band by electrons hopping from donor to donor. The process of impurity band conduction sets in at lower donor concentration levels if there are also some acceptor atoms present, so that some of the donors are always ionized. It is easier for a donor electron to hop to an ionized (unoccupied) donor than to an occupied donor atom, in order that two electrons will not have to occupy the same site during charge transport.

Acceptor States. A hole may be bound to a trivalent impurity in germanium or silicon (Fig. 20), just as an electron is bound to a pentavalent impurity. Trivalent impurities such as B, Al, Ga, and In are called **acceptors** because they accept electrons from the valence band in order to complete the covalent bonds with neighbor atoms, leaving holes in the band.

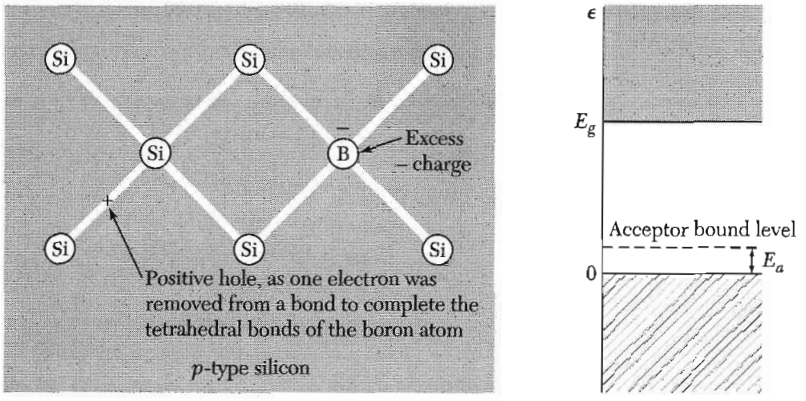


Figure 20 Boron has only three valence electrons; it can complete its tetrahedral bonds only by taking an electron from a Si-Si bond, leaving behind a hole in the silicon valence band. The positive hole is then available for conduction. The boron atom is called an **acceptor** because when ionized it accepts an electron from the valence band. At 0 K the hole is bound.

Table 6 Acceptor ionization energies E_a of trivalent impurities in germanium and silicon, in meV

	B	Al	Ga	In
Si	45.	57.	65.	157.
Ge	10.4	10.2	10.8	11.2

When an acceptor is ionized a hole is freed, which requires an input of energy. On the usual energy band diagram, an electron rises when it gains energy, whereas a hole sinks in gaining energy.

Experimental ionization energies of acceptors in germanium and silicon are given in Table 6. The Bohr model applies qualitatively for holes just as for electrons, but the degeneracy at the top of the valence band complicates the effective mass problem.

The tables show that donor and acceptor ionization energies in Si are comparable with k_BT at room temperature (26 meV), so that the thermal ionization of donors and acceptors is important in the electrical conductivity of silicon at room temperature. If donor atoms are present in considerably greater numbers than acceptors, the thermal ionization of donors will release electrons into the conduction band. The conductivity of the specimen then will be controlled by electrons (negative charges), and the material is said to be n type.

If acceptors are dominant, holes will be released into the valence band and the conductivity will be controlled by holes (positive charges): the material is p type. The sign of the Hall voltage (6.53) is a rough test for n or p type.

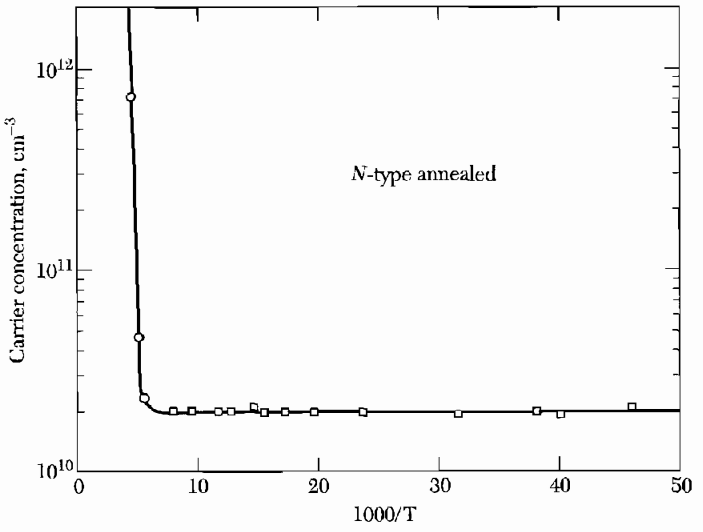


Figure 21 Temperature dependence of the free carrier concentration in ultrapure Ge, after R. N. Hall. The net concentration of electrically active impurities is 2×10^{10} cm^{-3}, as determined by Hall coefficient measurements. The rapid onset of intrinsic excitation as the temperature is increased is evident at low values of $1/T$. The carrier concentration is closely constant between 20 K and 200 K.

Another handy laboratory test is the sign of the thermoelectric potential, discussed below.

The numbers of holes and electrons are equal in the intrinsic regime. The intrinsic electron concentration n_i at 300 K is 1.7×10^{13} cm^{-3} in germanium and 4.6×10^{9} cm^{-3} in silicon. The electrical resistivity of intrinsic material is 43 ohm-cm for germanium and 2.6×10^{5} ohm-cm for silicon.

Germanium has 4.42×10^{22} atoms per cm^3. The purification of Ge has been carried further than any other element. The concentration of the common electrically active impurities—the shallow donor and acceptor impurities—has been reduced below 1 impurity atom in 10^{11} Ge atoms (Fig. 21). For example, the concentration of P in Ge can be reduced below 4×10^{10} cm^{-3}. There are impurities (H, O, Si, C) whose concentrations in Ge cannot usually be reduced below $10^{12}-10^{14}$ cm^{-3}, but these do not affect electrical measurements and therefore may be hard to detect.

Thermal Ionization of Donors and Acceptors

The calculation of the equilibrium concentration of conduction electrons from ionized donors is identical with the standard calculation in statistical mechanics of the thermal ionization of hydrogen atoms (*TP*, p. 369). If there are no acceptors present, the result in the low temperature limit $k_B T \ll E_d$ is

$$n \cong (n_0 N_d)^{1/2} \exp(-E_d/2k_B T) \ , \tag{53}$$

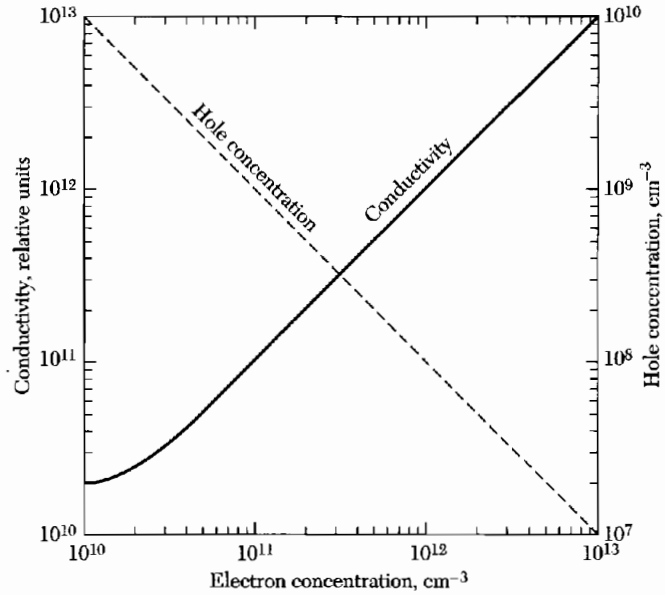


Figure 22 Electrical conductivity and hole concentration p calculated as a function of electron concentration n for a semiconductor at a temperature such that $np = 10^{20}$ cm^{-6}. The conductivity is symmetrical about $n = 10^{10}$ cm^{-3}. For $n > 10^{10}$, the specimen is n type; for $n < 10^{10}$, it is p type. We have taken $\mu_e = \mu_h$, for the mobilities.

with $n_0 \equiv 2(m_e k_B T/2\pi\hbar^2)^{3/2}$; here N_d is the concentration of donors. To obtain (53) we apply the laws of chemical equilibria to the concentration ratio $[e][N_d^+]/[N_d]$, and then set $[N_d^+] = [e] = n$. Identical results hold for acceptors, under the assumption of no donor atoms.

If the donor and acceptor concentrations are comparable, affairs are complicated and the equations are solved by numerical methods. However, the law of mass action (43) requires the np product to be constant at a given temperature. An excess of donors will increase the electron concentration and decrease the hole concentration; the sum $n + p$ will increase. The conductivity will increase as $n + p$ if the mobilities are equal, as in Fig. 22.

THERMOELECTRIC EFFECTS

Consider a semiconductor maintained at a constant temperature while an electric field drives through it an electric current density j_q. If the current is carried only by electrons, the charge flux is

$$j_q = n(-e)(-\mu_e)E = ne\mu_e E \ , \tag{54}$$

where μ_e is the electron mobility. The average energy transported by an electron is referred to the Fermi level μ,

$$(E_c - \mu) + \tfrac{3}{2}k_B T \ ,$$

where E_c is the energy at the conduction band edge. We refer the energy to the Fermi level because different conductors in contact have the same Fermi level. The energy flux that accompanies the charge flux is

$$j_U = n(E_c - \mu + \tfrac{3}{2}k_BT)(-\mu_e)E \ . \tag{55}$$

The **Peltier coefficient** Π is defined by $j_U = \Pi j_q$; or the energy carried per unit charge. For electrons,

$$\Pi_e = -(E_c - \mu + \tfrac{3}{2}k_BT)/e \tag{56}$$

and is negative because the energy flux is opposite to the charge flux. For holes

$$j_q = pe\mu_h E \ ; \qquad j_U = p(\mu - E_v + \tfrac{3}{2}k_BT)\mu_h E \ , \tag{57}$$

where E_v is the energy at the valence band edge. Thus

$$\Pi_h = (\mu - E_v + \tfrac{3}{2}k_BT)/e \tag{58}$$

and is positive. Equations (56) and (58) are the result of our simple drift velocity theory; a treatment by the Boltzmann transport equation gives minor numerical differences.[2]

The **absolute thermoelectric power** Q is defined from the open circuit electric field created by a temperature gradient:

$$E = Q \operatorname{grad} T \ . \tag{59}$$

The Peltier coefficient Π is related to the thermoelectric power Q by

$$\Pi = QT \ . \tag{60}$$

This is the famous Kelvin relation of irreversible thermodynamics. A measurement of the sign of the voltage across a semiconductor specimen, one end of which is heated, is a rough and ready way to tell if the specimen is n type or p type (Fig. 23).

SEMIMETALS

In semimetals the conduction band edge is very slightly lower in energy than the valence band edge. A small overlap in energy of the conduction and valence bands leads to small concentration of holes in the valence band and of electrons in the conduction band (Table 7). Three of the semimetals, arsenic, antimony, and bismuth, are in group V of the periodic table.

Their atoms associate in pairs in the crystal lattice, with two ions and ten valence electrons per primitive cell. The even number of valence electrons

[2]A simple discussion of Boltzmann transport theory is given in Appendix F.

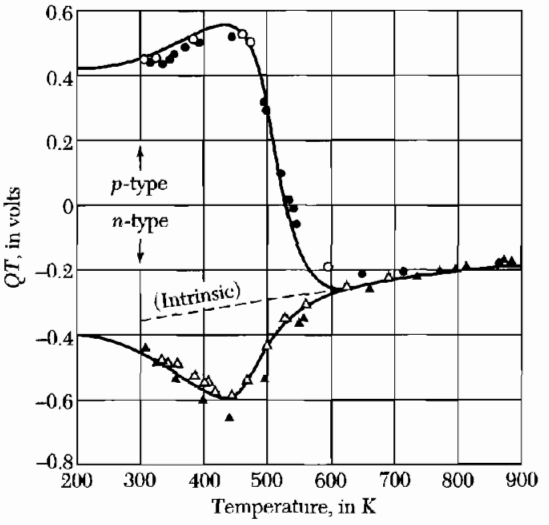


Figure 23 Peltier coefficient of n and p silicon as a function of temperature. Above 600 K the specimens act as intrinsic semiconductors. The curves are calculated and the points are experimental. (After T. H. Geballe and G. W. Hull.)

Table 7 Electron and hole concentrations in semimetals

Semimetal	n_e, in cm^{-3}	n_h, in cm^{-3}
Arsenic	$(2.12 \pm 0.01) \times 10^{20}$	$(2.12 \pm 0.01) \times 10^{20}$
Antimony	$(5.54 \pm 0.05) \times 10^{19}$	$(5.49 \pm 0.03) \times 10^{19}$
Bismuth	2.88×10^{17}	3.00×10^{17}
Graphite	2.72×10^{18}	2.04×10^{18}

could allow these elements to be insulators. Like semiconductors, the semimetals may be doped with suitable impurities to vary the relative numbers of holes and electrons. Their concentrations may also be varied with pressure, for the band edge overlap varies with pressure.

SUPERLATTICES

Consider a multilayer crystal of alternating thin layers of different compositions. Coherent layers on a nanometer thickness scale may be deposited by molecular-beam epitaxy or metal-organic vapor deposition, thus building up a superperiodic structure on a large scale. Systems of alternate layers of GaAs and GaAlAs have been studied to 50 periods or more, with lattice spacing A of perhaps 5 nm (50 Å). A superperiodic crystal potential arises from the superperiodic structure and acts on the conduction electrons and holes to create new (small) Brillouin zones and mini energy bands superposed on the band structures of the

constituent layers. Here we treat the motion of an electron in a superlattice in an applied electric field.

Bloch Oscillator

Consider a collisionless electron in a periodic lattice in one dimension, with motion normal to the planes of the superlattice. The equation of motion in a constant electric field parallel to k is $\hbar dk/dt = -eE$ or, for motion across a Brillouin zone with reciprocal lattice vector $G = 2\pi/A$, we have $\hbar G = \hbar 2\pi/A = eET$, where T is the period of the motion. The **Bloch frequency** of the motion is $\omega_B = 2\pi/T = eEA/\hbar$. The electron accelerates from $k = 0$ toward the zone boundary; when it reaches $k = \pi/A$ it reappears (as by an Umklapp process) at the zone boundary at the identical point $-\pi/A$, using the argument of Chapter 2.

We consider the motion in a model system in real space. We suppose that the electron lies in a simple energy band of width ϵ_0:

$$\epsilon = \epsilon_0(1 - \cos kA) . \tag{61}$$

The velocity in k-space (momentum space) is

$$v = \hbar^{-1} d\epsilon/dk = (A\epsilon_0/\hbar) \sin kA , \tag{62}$$

and the position of the electron in real space, with the initial condition $z = 0$ at $t = 0$, is given by

$$\begin{aligned} z &= \int v\, dt = \int dk\, v(k)(dt/dk) = (A\epsilon_0/\hbar) \int dk(-\hbar/eE) \sin kA \\ &= (-\epsilon_0/eE)(\cos kA - 1) = (-\epsilon_0/eE)(\cos(-eEAt/\hbar) - 1) . \end{aligned} \tag{63}$$

This confirms that the Bloch oscillation frequency in real space is $\omega_B = eEA/\hbar$. The motion in the periodic lattice is quite different from the motion in free space, for which the acceleration is constant.

Zener Tunneling

Thus far we have considered the effect of the electrostatic potential $-eEz$ (or $-eEnA$) on one energy band; the potential tilts the whole band. Higher bands will also be tilted similarly, creating the possibility of crossing between ladder levels of different bands. The interaction between different band levels at the same energy opens the possibility for an electron in one band at n to cross to another band at n'. This field-induced interband tunneling is an example of **Zener breakdown**, met most often at a single junction as in the Zener diode.

SUMMARY

- The motion of a wave packet centered at wavevector $\mathbf{k}$ is described by $\mathbf{F} = \hbar d\mathbf{k}/dt$, where $\mathbf{F}$ is the applied force. The motion in real space is obtained from the group velocity $\mathbf{v}_g = \hbar^{-1} \nabla_{\mathbf{k}} \epsilon(\mathbf{k})$.

- The smaller the energy gap, the smaller is the effective mass $|m^*|$ near the gap.
- A crystal with one hole has one empty electron state in an otherwise filled band. The properties of the hole are those of the $N - 1$ electrons in this band.
  (a) If the electron is missing from the state of wavevector $\mathbf{k}_e$, then the wavevector of the hole is $\mathbf{k}_h = -\mathbf{k}_e$.
  (b) The rate of change of $\mathbf{k}_h$ in an applied field requires the assignment of a positive charge to the hole: $e_h = e = -e_e$.
  (c) If $\boldsymbol{v}_e$ is the velocity an electron would have in the state $\mathbf{k}_e$, then the velocity to be ascribed to the hole of wavevector $\mathbf{k}_h = -\mathbf{k}_e$ is $\boldsymbol{v}_h = \boldsymbol{v}_e$.
  (d) The energy of the hole referred to zero for a filled band is positive and is $\epsilon_h(\mathbf{k}_h) = -\epsilon(\mathbf{k}_e)$.
  (e) The effective mass of a hole is opposite to the effective mass of an electron at the same point on the energy band: $m_h = -m_e$.

Problems

1. ***Impurity orbits.*** Indium antimonide has $E_g = 0.23$ eV; dielectric constant $\epsilon = 18$; electron effective mass $m_e = 0.015\ m$. Calculate (a) the donor ionization energy; (b) the radius of the ground state orbit. (c) At what minimum donor concentration will appreciable overlap effects between the orbits of adjacent impurity atoms occur? This overlap tends to produce an impurity band—a band of energy levels which permit conductivity presumably by a hopping mechanism in which electrons move from one impurity site to a neighboring ionized impurity site.

2. ***Ionization of donors.*** In a particular semiconductor there are 10^{13} donors/cm^3 with an ionization energy E_d of 1 meV and an effective mass $0.01\ m$. (a) Estimate the concentration of conduction electrons at 4 K. (b) What is the value of the Hall coefficient? Assume no acceptor atoms are present and that $E_g \gg k_B T$.

3. ***Hall effect with two carrier types.*** Assuming concentration n, p; relaxation times τ_e, τ_h; and masses m_e, m_h, show that the Hall coefficient in the drift velocity approximation is

(CGS)
$$R_H = \frac{1}{ec} \cdot \frac{p - nb^2}{(p + nb)^2},$$

where $b = \mu_e/\mu_h$ is the mobility ratio. In the derivation neglect terms of order B^2. In SI we drop the c. Hint: In the presence of a longitudinal electric field, find the transverse electric field such that the transverse current vanishes. The algebra may seem tedious, but the result is worth the trouble. Use (6.64), but for two carrier types; neglect $(\omega_c\tau)^2$ in comparison with $\omega_c\tau$.

4. ***Cyclotron resonance for a spheroidal energy surface.*** Consider the energy surface

$$\epsilon(\mathbf{k}) = \hbar^2\left(\frac{k_x^2 + k_y^2}{2m_t} + \frac{k_z^2}{2m_l}\right),$$

where m_t is the transverse mass parameter and m_l is the longitudinal mass parameter. A surface on which $\epsilon(\mathbf{k})$ is constant will be a spheroid. Use the equation of motion (6), with $\mathbf{v} = \hbar^{-1}\nabla_k\epsilon$, to show that $\omega_c = eB/(m_l m_t)^{1/2}c$ when the static magnetic field B lies in the xy plane. This result agrees with (34) when $\theta = \pi/2$. The result is in CGS; to obtain SI, omit the c.

5. ***Magnetoresistance with two carrier types.*** Problem 6.9 shows that in the drift velocity approximation the motion of charge carriers in electric and magnetic fields does not lead to transverse magnetoresistance. The result is different with two carrier types. Consider a conductor with a concentration n of electrons of effective mass m_e and relaxation time τ_e; and a concentration p of holes of effective mass m_h and relaxation time τ_h. Treat the limit of very strong magnetic fields, $\omega_c\tau \gg 1$. (a) Show in this limit that $\sigma_{yx} = (n - p)ec/B$. (b) Show that the Hall field is given by, with $Q \equiv \omega_c\tau$,

$$E_y = -(n - p)\left(\frac{n}{Q_e} + \frac{p}{Q_h}\right)^{-1} E_x,$$

which vanishes if $n = p$. (c) Show that the effective conductivity in the x direction is

$$\sigma_{\text{eff}} = \frac{ec}{B}\left[\left(\frac{n}{Q_e} + \frac{p}{Q_h}\right) + (n - p)^2\left(\frac{n}{Q_e} + \frac{p}{Q_h}\right)^{-1}\right].$$

If $n = p$, $\sigma \propto B^{-2}$. If $n \neq p$, σ saturates in strong fields; that is, it approaches a limit independent of B as $B \to \infty$.

9

Fermi Surfaces and Metals

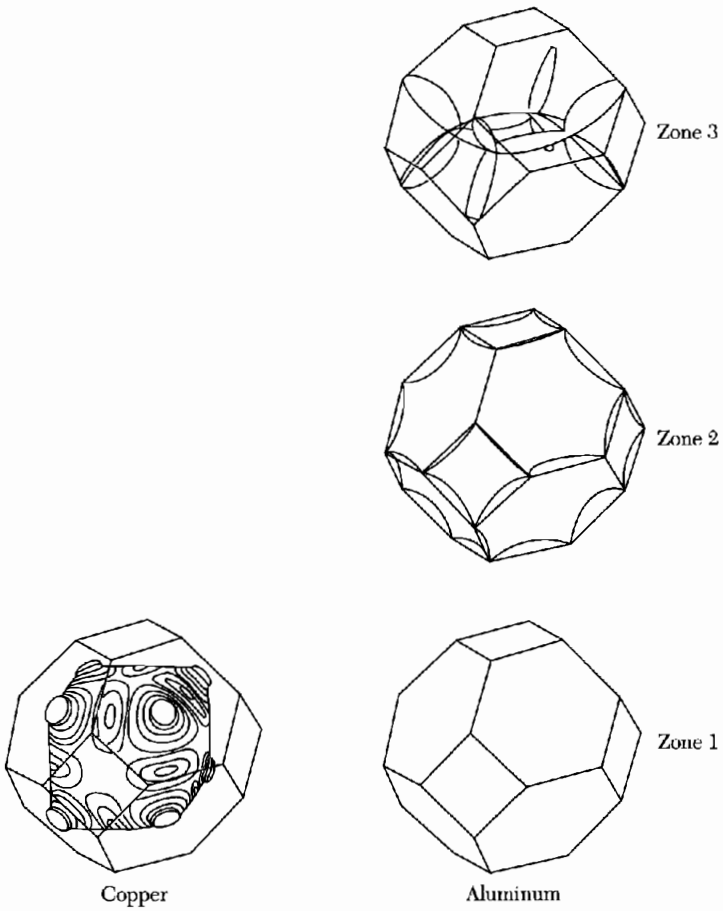


Figure 1 Free electron Fermi surfaces for fcc metals with one (Cu) and three (Al) valence electrons per primitive cell. The Fermi surface shown for copper has been deformed from a sphere to agree with the experimental results. The second zone of aluminum is nearly half-filled with electrons. (A. R. Mackintosh.)

Few people would define a metal as "a solid with a Fermi surface." This may nevertheless be the most meaningful definition of a metal one can give today; it represents a profound advance in the understanding of why metals behave as they do. The concept of the Fermi surface, as developed by quantum physics, provides a precise explanation of the main physical properties of metals.

A. R. Mackintosh

The Fermi surface is the surface of constant energy ϵ_F in $\mathbf{k}$ space. The Fermi surface separates the unfilled orbitals from the filled orbitals, at absolute zero. The electrical properties of the metal are determined by the volume and shape of the Fermi surface, because the current is due to changes in the occupancy of states near the Fermi surface.

The shape may be very intricate as viewed in the reduced zone scheme below and yet have a simple interpretation when reconstructed to lie near the surface of a sphere. We exhibit in Fig. 1 the free electron Fermi surfaces constructed for two metals that have the face-centered cubic crystal structure: copper, with one valence electron, and aluminum, with three. The free electron Fermi surfaces were developed from spheres of radius k_F determined by the valence electron concentration. The surface for copper is deformed by interaction with the lattice. How do we construct these surfaces from a sphere? The constructions require the reduced and also the periodic zone schemes.

Reduced Zone Scheme

It is always possible to select the wavevector index $\mathbf{k}$ of any Bloch function to lie within the first Brillouin zone. The procedure is known as mapping the band in the reduced zone scheme.

If we encounter a Bloch function written as $\psi_{\mathbf{k}'}(\mathbf{r}) = e^{i\mathbf{k}'\cdot\mathbf{r}}u_{\mathbf{k}'}(\mathbf{r})$, with $\mathbf{k}'$ outside the first zone, as in Fig. 2, we may always find a suitable reciprocal lattice vector $\mathbf{G}$ such that $\mathbf{k} = \mathbf{k}' + \mathbf{G}$ lies within the first Brillouin zone. Then

$$\begin{aligned}\psi_{\mathbf{k}'}(\mathbf{r}) &= e^{i\mathbf{k}'\cdot\mathbf{r}}u_{\mathbf{k}'}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}(e^{-i\mathbf{G}\cdot\mathbf{r}}u_{\mathbf{k}'}(\mathbf{r})) \\ &= e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r}) = \psi_{\mathbf{k}}(\mathbf{r}) \ ,\end{aligned} \tag{1}$$

where $u_{\mathbf{k}}(\mathbf{r}) = e^{-i\mathbf{G}\cdot\mathbf{r}}u_{\mathbf{k}'}(\mathbf{r})$. Both $e^{-i\mathbf{G}\cdot\mathbf{r}}$ and $u_{\mathbf{k}'}(\mathbf{r})$ are periodic in the crystal lattice, so $u_{\mathbf{k}}(\mathbf{r})$ is also, whence $\psi_k(\mathbf{r})$ is of the Bloch form.

Even with free electrons it is useful to work in the reduced zone scheme, as in Fig. 3. Any energy $\epsilon_{\mathbf{k}'}$ for $\mathbf{k}'$ outside the first zone is equal to an $\epsilon_{\mathbf{k}}$ in the first zone, where $\mathbf{k} = \mathbf{k}' + \mathbf{G}$. Thus we need solve for the energy only in the

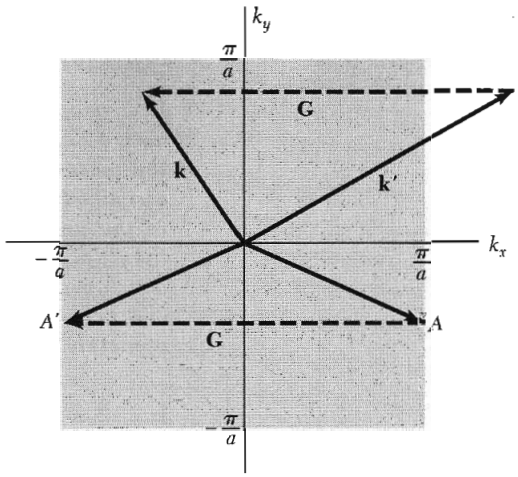


Figure 2 First Brillouin zone of a square lattice of side a. The wavevector $\mathbf{k}'$ can be carried into the first zone by forming $\mathbf{k}' + \mathbf{G}$. The wavevector at a point A on the zone boundary is carried by $\mathbf{G}$ to the point A' on the opposite boundary of the same zone. Shall we count both A and A' as lying in the first zone? Because they can be connected by a reciprocal lattice vector, we count them as *one* identical point in the zone.

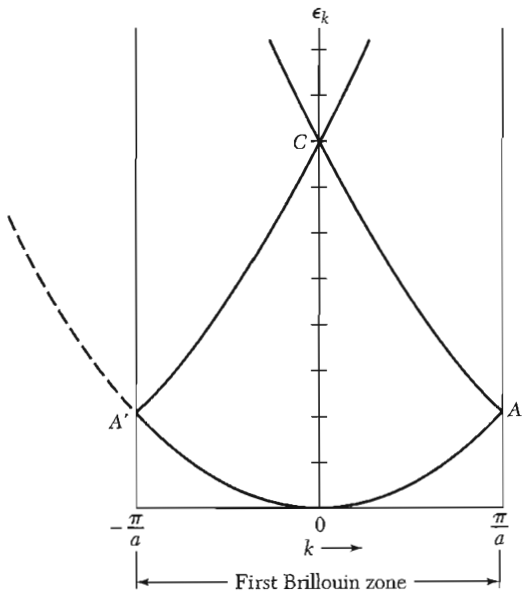


Figure 3 Energy-wavevector relation $\epsilon_k = \hbar^2k^2/2m$ for free electrons as drawn in the reduced zone scheme. This construction often gives a useful idea of the overall appearance of the band structure of a crystal. The branch AC if displaced by $-2\pi/a$ gives the usual free electron curve for negative k, as suggested by the dashed curve. The branch $A'C$ if displaced by $2\pi/a$ gives the usual curve for positive k. A crystal potential $U(x)$ will introduce band gaps at the edges of the zone (as at A and A') and at the center of the zone (as at C). The point C when viewed in the extended zone scheme falls at the edges of the second zone. The overall width and gross features of the band structure are often indicated properly by such free electron bands in the reduced zone scheme.

first Brillouin zone, for each band. An energy band is a single branch of the $\epsilon_\mathbf{k}$ versus $\mathbf{k}$ surface. In the reduced zone scheme we may find different energies at the same value of the wavevector. Each different energy characterizes a different band. Two bands are shown in Fig. 3.

Two wavefunctions at the same $\mathbf{k}$ but of different energies will be independent of each other: the wavefunctions will be made up of different combinations of the plane wave components $\exp[i(\mathbf{k} + \mathbf{G}) \cdot \mathbf{r}]$ in the expansion of (7.29). Because the values of the coefficients $C(\mathbf{k} + \mathbf{G})$ will differ for the

different bands, we should add a symbol, say n, to the C's to serve as a band index: $C_n(\mathbf{k} + \mathbf{G})$. Thus the Bloch function for a state of wavevector $\mathbf{k}$ in the band n can be written as

$$\psi_{n,\mathbf{k}} = \exp(i\mathbf{k} \cdot \mathbf{r}) u_{n,\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} C_n(\mathbf{k} + \mathbf{G}) \exp[i(\mathbf{k} + \mathbf{G}) \cdot \mathbf{r}] \ .$$

Periodic Zone Scheme

We can repeat a given Brillouin zone periodically through all of wavevector space. To repeat a zone, we translate the zone by a reciprocal lattice vector. If we can translate a band from other zones into the first zone, we can translate a band in the first zone into every other zone. In this scheme the energy $\epsilon_{\mathbf{k}}$ of a band is a periodic function in the reciprocal lattice:

$$\epsilon_{\mathbf{k}} = \epsilon_{\mathbf{k}+\mathbf{G}} \ . \tag{2}$$

Here $\epsilon_{\mathbf{k}+\mathbf{G}}$ is understood to refer to the same energy band as $\epsilon_{\mathbf{k}}$.

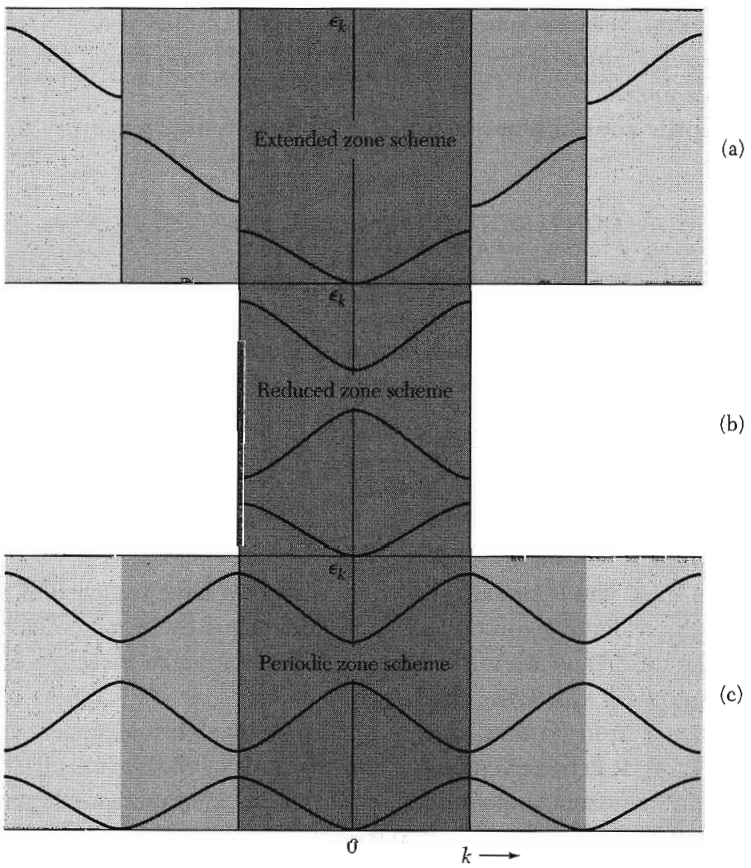


Figure 4 Three energy bands of a linear lattice plotted in (a) the extended (Brillouin), (b) reduced, and (c) periodic zone schemes.

The result of this construction is known as the **periodic zone scheme**. The periodic property of the energy also can be seen easily from the central equation (7.27).

Consider for example an energy band of a simple cubic lattice as calculated in the tight-binding approximation in (13) below:

$$\epsilon_k = -\alpha - 2\gamma\,(\cos k_x a + \cos k_y a + \cos k_z a)\ , \tag{3}$$

where α and γ are constants. One reciprocal lattice vector of the sc lattice is $\mathbf{G} = (2\pi/a)\hat{\mathbf{x}}$; if we add this vector to $\mathbf{k}$ the only change in (3) is

$$\cos k_x a \rightarrow \cos\,(k_x + 2\pi/a)a = \cos\,(k_x a + 2\pi)\ ,$$

but this is identically equal to $\cos k_x a$. The energy is unchanged when the wavevector is increased by a reciprocal lattice vector, so that the energy is a periodic function of the wavevector.

Three different zone schemes are useful (Fig. 4):

- The **extended zone scheme** in which different bands are drawn in different zones in wavevector space.
- The **reduced zone scheme** in which all bands are drawn in the first Brillouin zone.
- The **periodic zone scheme** in which every band is drawn in every zone.

CONSTRUCTION OF FERMI SURFACES

We consider in Fig. 5 the analysis for a square lattice. The equation of the zone boundaries is $2\mathbf{k}\cdot\mathbf{G} + G^2 = 0$ and is satisfied if $\mathbf{k}$ terminates on the plane normal to $\mathbf{G}$ at the midpoint of $\mathbf{G}$. The first Brillouin zone of the square lattice is the area enclosed by the perpendicular bisectors of $\mathbf{G}_1$ and of the three reciprocal lattice vectors equivalent by symmetry to $\mathbf{G}_1$ in Fig. 5a. These four reciprocal lattice vectors are $\pm(2\pi/a)\hat{\mathbf{k}}_x$ and $\pm(2\pi/a)\hat{\mathbf{k}}_y$.

The second zone is constructed from $\mathbf{G}_2$ and the three vectors equivalent to it by symmetry, and similarly for the third zone. The pieces of the second and third zones are drawn in Fig. 5b.

To determine the boundaries of some zones we have to consider sets of several nonequivalent reciprocal lattice vectors. Thus the boundaries of section 3_a of the third zone are formed from the perpendicular bisectors of three $\mathbf{G}$'s, namely $(2\pi/a)\hat{\mathbf{k}}_x$; $(4\pi/a)\hat{\mathbf{k}}_y$; and $(2\pi/a)(\hat{\mathbf{k}}_x + \hat{\mathbf{k}}_y)$.

The free electron Fermi surface for an arbitrary electron concentration is shown in Fig. 6. It is inconvenient to have sections of the Fermi surface that belong to the same zone appear detached from one another. The detachment can be repaired by a transformation to the reduced zone scheme.

We take the triangle labeled 2_a and move it by a reciprocal lattice vector $\mathbf{G} = -(2\pi/a)\hat{\mathbf{k}}_x$ such that the triangle reappears in the area of the first

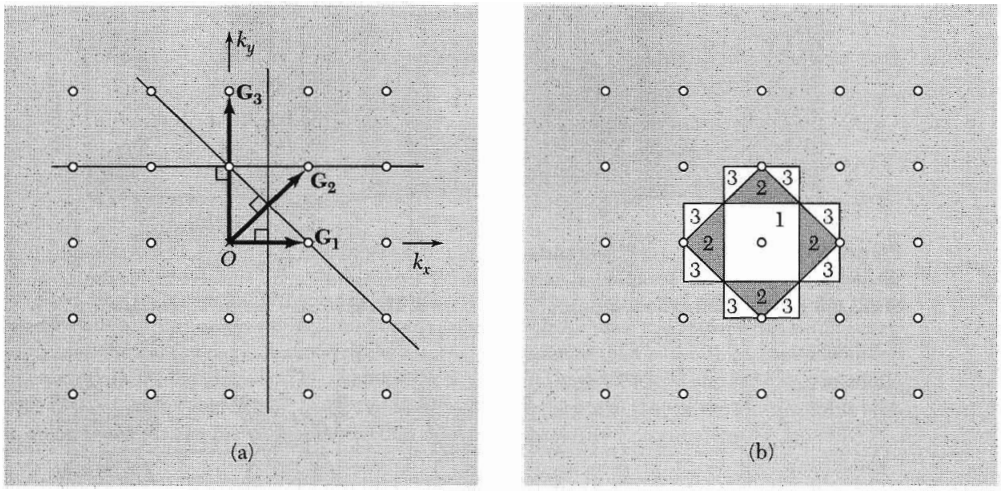


Figure 5 (a) Construction in **k** space of the first three Brillouin zones of a square lattice. The three shortest forms of the reciprocal lattice vectors are indicated as $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_3$. The lines drawn are the perpendicular bisectors of these **G**'s. (b) On constructing all lines equivalent by symmetry to the three lines in (a) we obtain the regions in **k** space which form the first three Brillouin zones. The numbers denote the zone to which the regions belong; the numbers here are ordered according to the length of the vector **G** involved in the construction of the outer boundary of the region.

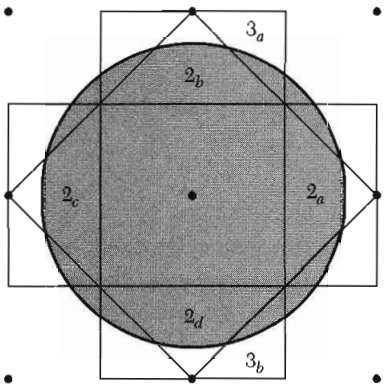


Figure 6 Brillouin zones of a square lattice in two dimensions. The circle shown is a surface of constant energy for free electrons; it will be the Fermi surface for some particular value of the electron concentration. The total area of the filled region in **k** space depends only on the electron concentration and is independent of the interaction of the electrons with the lattice. The shape of the Fermi surface depends on the lattice interaction, and the shape will not be an exact circle in an actual lattice. The labels within the sections of the second and third zones refer to Fig. 7.

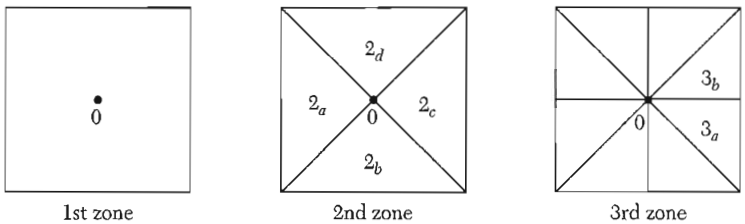


Figure 7 Mapping of the first, second, and third Brillouin zones in the reduced zone scheme. The sections of the second zone in Fig. 6 are put together into a square by translation through an appropriate reciprocal lattice vector. A different **G** is needed for each piece of a zone.

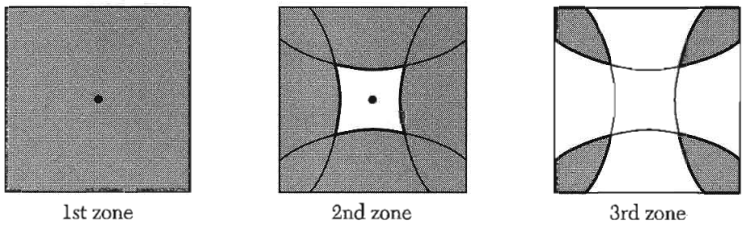


Figure 8 The free electron Fermi surface of Fig. 6, as viewed in the reduced zone scheme. The shaded areas represent occupied electron states. Parts of the Fermi surface fall in the second, third, and fourth zones. The fourth zone is not shown. The first zone is entirely occupied.

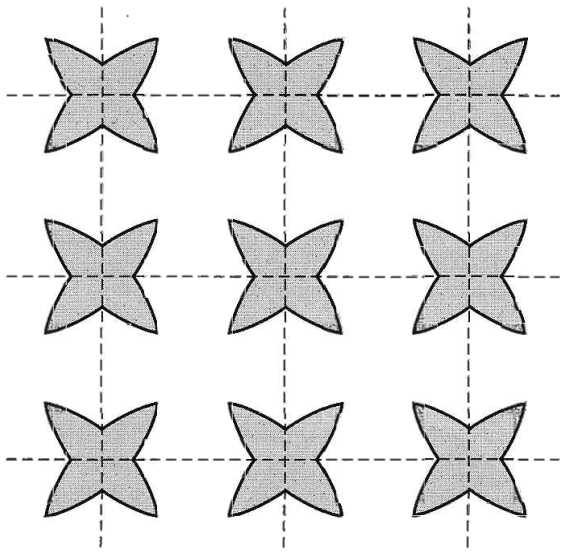


Figure 9 The Fermi surface in the third zone as drawn in the periodic zone scheme. The figure was constructed by repeating the third zone of Fig. 8.

Brillouin zone (Fig. 7). Other reciprocal lattice vectors will shift the triangles 2_b, 2_c, 2_d to other parts of the first zone, completing the mapping of the second zone into the reduced zone scheme. The parts of the Fermi surface falling in the second zone are now connected, as shown in Fig. 8.

A third zone is assembled into a square in Fig. 8, but the parts of the Fermi surface still appear disconnected. When we look at it in the periodic zone scheme (Fig. 9), the Fermi surface forms a lattice of rosettes.

Nearly Free Electrons

How do we go from Fermi surfaces for free electrons to Fermi surfaces for nearly free electrons? We can make approximate constructions freehand by the use of four facts:

- The interaction of the electron with the periodic potential of the crystal creates energy gaps at the zone boundaries.
- Almost always the Fermi surface will intersect zone boundaries perpendicularly.

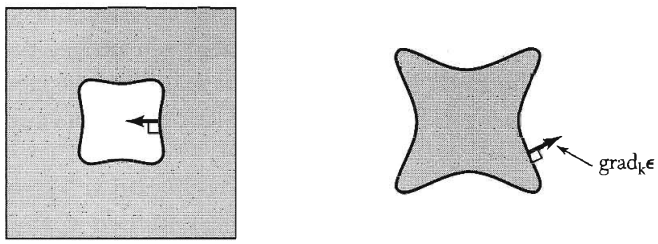


Figure 10 Qualitative impression of the effect of a weak periodic crystal potential on the Fermi surface of Fig. 8. At one point on each Fermi surface we have shown the vector grad$_k\epsilon$. In the second zone the energy increases toward the interior of the figure, and in the third zone the energy increases toward the exterior. The shaded regions are filled with electrons and are lower in energy than the unshaded regions. We shall see that a Fermi surface like that of the third zone is electronlike, whereas one like that of the second zone is holelike.

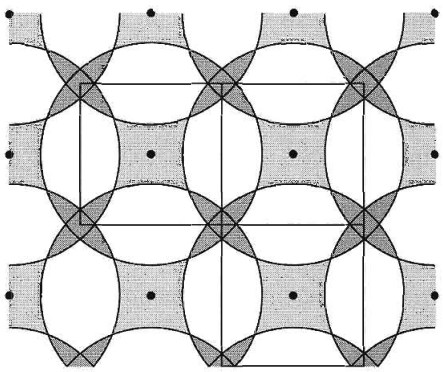


Figure 11 Harrison construction of free electron Fermi surfaces on the second, third, and fourth zones for a square lattice. The Fermi surface encloses the entire first zone, which therefore is filled with electrons.

- The crystal potential will round out sharp corners in the Fermi surfaces.
- The total volume enclosed by the Fermi surface depends only on the electron concentration and is independent of the details of the lattice interaction.

We cannot make quantitative statements without calculation, but qualitatively we expect the Fermi surfaces in the second and third zones of Fig. 8 to be changed as shown in Fig. 10.

Freehand impressions of the Fermi surfaces derived from free electron surfaces are useful. Fermi surfaces for free electrons are constructed by a procedure credited to Harrison, Fig. 11. The reciprocal lattice points are determined, and a free electron sphere of radius appropriate to the electron concentration is drawn around each point. Any point in **k** space that lies within at least one sphere corresponds to an occupied state in the first zone. Points within at least two spheres correspond to occupied states in the second zone, and similarly for points in three or more spheres.

We said earlier that the alkali metals are the simplest metals, with weak interactions between the conduction electrons and the lattice. Because the

alkalis have only one valence electron per atom, the first Brillouin zone boundaries are distant from the approximately spherical Fermi surface that fills one-half of the volume of the zone. It is known by calculation and experiment that the Fermi surface of Na is closely spherical, and that the Fermi surface for Cs is deformed by perhaps 10 percent from a sphere.

The divalent metals Be and Mg also have weak lattice interactions and nearly spherical Fermi surfaces. But because they have two valence electrons each, the Fermi surface encloses twice the volume in **k** space as for the alkalis. That is, the volume enclosed by the Fermi surface is exactly equal to that of a zone, but because the surface is spherical it extends out of the first zone and into the second zone.

ELECTRON ORBITS, HOLE ORBITS, AND OPEN ORBITS

We saw in Eq. (8.7) that electrons in a static magnetic field move on a curve of constant energy on a plane normal to **B**. An electron on the Fermi surface will move in a curve on the Fermi surface, because this is a surface of constant energy. Three types of orbits in a magnetic field are shown in Fig. 12.

The closed orbits of (a) and (b) are traversed in opposite senses. Because particles of opposite charge circulate in a magnetic field in opposite senses, we say that one orbit is **electronlike** and the other orbit is **holelike**. Electrons in holelike orbits move in a magnetic field as if endowed with a positive charge. This is consistent with the treatment of holes in Chapter 8.

In (c) the orbit is not closed: the particle on reaching the zone boundary at A is instantly folded back to B, where B is equivalent to B' because

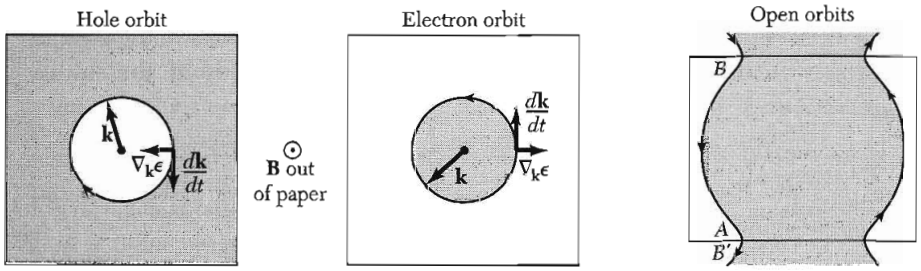


Figure 12 Motion in a magnetic field of the wavevector of an electron on the Fermi surface, in (a) and (b) for Fermi surfaces topologically equivalent to those of Fig. 10. In (a) the wavevector moves around the orbit in a clockwise direction; in (b) the wavevector moves around the orbit in a counter-clockwise direction. The direction in (b) is what we expect for a free electron of charge $-e$; the smaller **k** values have the lower energy, so that the filled electron states lie inside the Fermi surface. We call the orbit in (b) **electronlike**. The sense of the motion in a magnetic field is opposite in (a) to that in (b), so that we refer to the orbit in (a) as **holelike**. A hole moves as a particle of positive charge e. In (c) for a rectangular zone we show the motion on an **open orbit** in the periodic zone scheme. An open orbit is topologically intermediate between a hole orbit and an electron orbit.

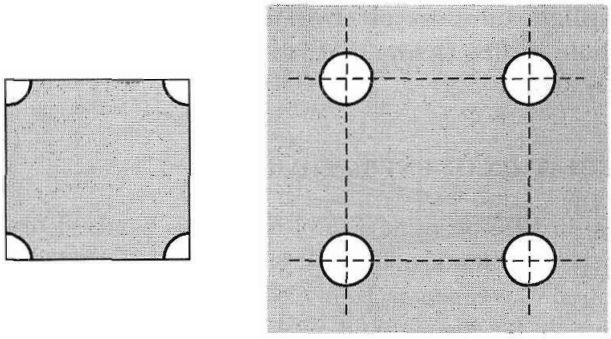


Figure 13 (a) Vacant states at the corners of an almost-filled band, drawn in the reduced zone scheme. (b) In the periodic zone scheme the various parts of the Fermi surface are connected. Each circle forms a holelike orbit. The different circles are entirely equivalent to each other, and the density of states is that of a single circle. (The orbits need not be true circles: for the lattice shown it is only required that the orbits have fourfold symmetry.)

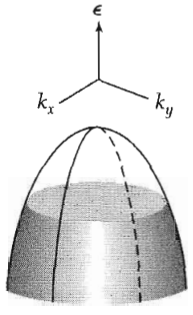


Figure 14 Vacant states near the top of an almost filled band in a two-dimensional crystal. This figure is equivalent to Fig. 12a.

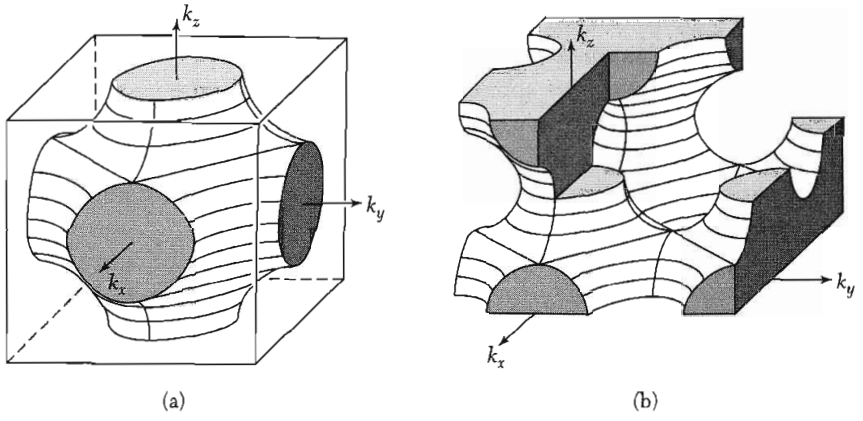


Figure 15 Constant energy surface in the Brillouin zone of a simple cubic lattice, for the assumed energy band $\epsilon_k = -\alpha - 2\gamma(\cos k_x a + \cos k_y a + \cos k_z a)$. (a) Constant energy surface $\epsilon = -\alpha$. The filled volume contains one electron per primitive cell. (b) The same surface exhibited in the periodic zone scheme. The connectivity of the orbits is clearly shown. Can you find electron, hole, and open orbits for motion in a magnetic field $B\hat{z}$? (A. Sommerfeld and H. A. Bethe.)

they are connected by a reciprocal lattice vector. Such an orbit is called an **open orbit**. *Open orbits have an important effect on the magnetoresistance.*

Vacant orbitals near the top of an otherwise filled band give rise to hole-like orbits, as in Figs. 13 and 14. A view of a possible energy surface in three dimensions is given in Fig. 15.

Orbits that enclose filled states are electron orbits. Orbits that enclose empty states are hole orbits. Orbits that move from zone to zone without closing are open orbits.

CALCULATION OF ENERGY BANDS

Wigner and Seitz, who in 1933 performed the first serious band calculations, refer to afternoons spent on the manual desk calculators of those days, using one afternoon for a trial wavefunction. Here we limit ourselves to three introductory methods: the tight-binding method, useful for interpolation; the Wigner-Seitz method, useful for the visualization and understanding of the alkali metals; and the pseudopotential method, utilizing the general theory of Chapter 7, which shows the simplicity of many problems.

Tight Binding Method for Energy Bands

Let us start with neutral separated atoms and watch the changes in the atomic energy levels as the charge distributions of adjacent atoms overlap when the atoms are brought together to form a crystal. Consider two hydrogen atoms, each with an electron in the 1*s* ground state. The wavefunctions ψ_A, ψ_B on the separated atoms are shown in Fig. 16a.

As the atoms are brought together, their wavefunctions overlap. We consider the two combinations $\psi_A \pm \psi_B$. Each combination shares an electron with the two protons, but an electron in the state $\psi_A + \psi_B$ will have a somewhat lower energy than in the state $\psi_A - \psi_B$.

In $\psi_A + \psi_B$ the electron spends part of the time in the region midway between the two protons, and in this region it is in the attractive potential of both protons at once, thereby increasing the binding energy. In $\psi_A - \psi_B$ the probability density vanishes midway between the nuclei; an extra binding does not appear.

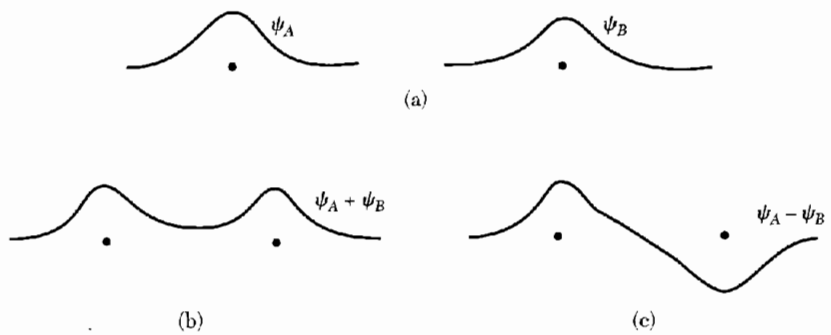


Figure 16 (a) Schematic drawing of wavefunctions of electrons on two hydrogen atoms at large separation. (b) Ground state wavefunction at closer separation. (c) Excited state wavefunction.

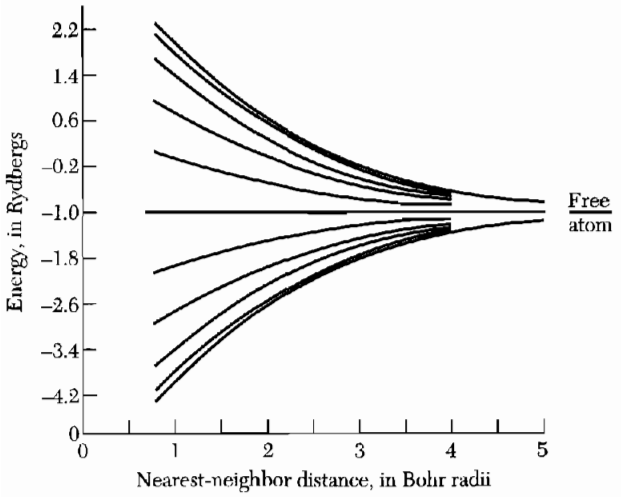


Figure 17 The 1*s* band of a ring of 20 hydrogen atoms; the one-electron energies are calculated in the tight-binding approximation with the nearest-neighbor overlap integral of Eq. (9).

As two atoms are brought together, two separated energy levels are formed for each level of the isolated atom. For N atoms, N orbitals are formed for each orbital of the isolated atom (Fig. 17).

As free atoms are brought together, the coulomb interaction between the atom cores and the electron splits the energy levels, spreading them into bands. Each state of given quantum number of the free atom is spread in the crystal into a band of energies. The width of the band is proportional to the strength of the overlap interaction between neighboring atoms.

There will also be bands formed from $p, d, \ldots$ states ($l = 1, 2, \ldots$) of the free atoms. States degenerate in the free atom will form different bands. Each will not have the same energy as any other band over any substantial range of the wavevector. Bands may coincide in energy at certain values of $\mathbf{k}$ in the Brillouin zone.

The approximation that starts out from the wavefunctions of the free atoms is known as the tight-binding approximation or the LCAO (linear combination of atomic orbitals) approximation. The approximation is quite good for the inner electrons of atoms, but it is not often a good description of the conduction electrons themselves. It is used to describe approximately the d bands of the transition metals and the valence bands of diamondlike and inert gas crystals.

Suppose that the ground state of an electron moving in the potential $U(\mathbf{r})$ of an isolated atom is $\varphi(\mathbf{r})$, an s state. The treatment of bands arising from degenerate ($p, d, \ldots$) atomic levels is more complicated. If the influence of one atom on another is small, we obtain an approximate wavefunction for one electron in the whole crystal by taking

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_j C_{\mathbf{k}j}\varphi(\mathbf{r} - \mathbf{r}_j) \ , \tag{4}$$

where the sum is over all lattice points. We assume the primitive basis contains one atom. This function is of the Bloch form (7.7) if $C_{\mathbf{k}j} = N^{-1/2}\, e^{i\mathbf{k}\cdot\mathbf{r}}$, which gives, for a crystal of N atoms,

$$\psi_{\mathbf{k}}(\mathbf{r}) = N^{-1/2} \sum_j \exp(i\mathbf{k}\cdot\mathbf{r}_j)\varphi(\mathbf{r}-\mathbf{r}_j) \ . \tag{5}$$

We prove (5) is of the Bloch form. Consider a translation $\mathbf{T}$ connecting two lattice points:

$$\begin{aligned}\psi_{\mathbf{k}}(\mathbf{r}+\mathbf{T}) &= N^{-1/2} \sum_j \exp(i\mathbf{k}\cdot\mathbf{r}_j)\varphi(\mathbf{r}+\mathbf{T}-\mathbf{r}_j) \\ &= \exp(i\mathbf{k}\cdot\mathbf{T})\, N^{-1/2} \sum_j \exp[i\mathbf{k}\cdot(\mathbf{r}_j-\mathbf{T})]\varphi[\mathbf{r}-(\mathbf{r}_j-\mathbf{T})] \\ &= \exp(i\mathbf{k}\cdot\mathbf{T})\, \psi_{\mathbf{k}}(\mathbf{r}) \ ,\end{aligned} \tag{6}$$

exactly the Bloch condition.

We find the first-order energy by calculating the diagonal matrix elements of the hamiltonian of the crystal:

$$\langle \mathbf{k}|H|\mathbf{k}\rangle = N^{-1} \sum_j \sum_m \exp[i\mathbf{k}\cdot(\mathbf{r}_j-\mathbf{r}_m)]\, \langle \varphi_m|H|\varphi_j\rangle \ , \tag{7}$$

where $\varphi_m \equiv \varphi(\mathbf{r}-\mathbf{r}_m)$. Writing $\boldsymbol{\rho}_m = \mathbf{r}_m - \mathbf{r}_j$,

$$\langle \mathbf{k}|H|\mathbf{k}\rangle = \sum_m \exp(-i\mathbf{k}\cdot\boldsymbol{\rho}_m) \int dV\, \varphi^*(\mathbf{r}-\boldsymbol{\rho}_m) H \varphi(\mathbf{r}) \ . \tag{8}$$

We now neglect all integrals in (8) except those on the same atom and those between nearest neighbors connected by $\boldsymbol{\rho}$. We write

$$\int dV\, \varphi^*(\mathbf{r}) H \varphi(\mathbf{r}) = -\alpha \ ; \quad \int dV\, \varphi^*(\mathbf{r}-\boldsymbol{\rho}) H \varphi(\mathbf{r}) = -\gamma \ ; \tag{9}$$

and we have the first-order energy, provided $\langle \mathbf{k}|\mathbf{k}\rangle = 1$:

$$\langle \mathbf{k}|H|\mathbf{k}\rangle = -\alpha - \gamma \sum_m \exp(-i\mathbf{k}\cdot\boldsymbol{\rho}_m) = \epsilon_{\mathbf{k}} \ . \tag{10}$$

The dependence of the overlap energy γ on the interatomic separation ρ can be evaluated explicitly for two hydrogen atoms in $1s$ states. In rydberg energy units, $\mathrm{Ry} = me^4/2\hbar^2$, we have

$$\gamma(\mathrm{Ry}) = 2(1+\rho/a_0)\exp(-\rho/a_0) \ , \tag{11}$$

where $a_0 = \hbar^2/me^2$. The overlap energy decreases exponentially with the separation.

For a simple cubic structure the nearest-neighbor atoms are at

$$\boldsymbol{\rho}_m = (\pm a,0,0) \ ; \quad (0,\pm a,0) \ ; \quad (0,0,\pm a) \ , \tag{12}$$

so that (10) becomes

$$\epsilon_{\mathbf{k}} = -\alpha - 2\gamma(\cos k_x a + \cos k_y a + \cos k_z a) \ . \tag{13}$$

Thus the energies are confined to a band of width 12γ. The weaker the overlap, the narrower is the energy band. A constant energy surface is shown in Fig. 15. For $ka \ll 1$, $\epsilon_k \simeq -\alpha - 6\gamma + \gamma k^2 a^2$. The effective mass is $m^* = \hbar^2/2\gamma a^2$. When the overlap integral γ is small, the band is narrow and the effective mass is high.

We considered one orbital of each free atom and obtained one band $\epsilon_{\mathbf{k}}$. The number of orbitals in the band that corresponds to a nondegenerate atomic level is $2N$, for N atoms. We see this directly: values of $\mathbf{k}$ within the first Brillouin zone define independent wavefunctions. The simple cubic zone has $-\pi/a < k_x < \pi/a$, etc. The zone volume is $8\pi^3/a^3$. The number of orbitals (counting both spin orientations) per unit volume of $\mathbf{k}$ space is $V/4\pi^3$, so that the number of orbitals is $2V/a^3$. Here V is the volume of the crystal, and $1/a^3$ is the number of atoms per unit volume. Thus there are $2N$ orbitals.

For the fcc structure with eight nearest neighbors,

$$\epsilon_{\mathbf{k}} = -\alpha - 8\gamma \cos \tfrac{1}{2} k_x a \cos \tfrac{1}{2} k_y a \cos \tfrac{1}{2} k_z a \ . \tag{14}$$

For the fcc structure with 12 nearest neighbors,

$$\epsilon_{\mathbf{k}} = -\alpha - 4\gamma(\cos \tfrac{1}{2} k_y a \cos \tfrac{1}{2} k_z a + \cos \tfrac{1}{2} k_z a \cos \tfrac{1}{2} k_x a + \cos \tfrac{1}{2} k_x a \cos \tfrac{1}{2} k_y a) \ . \tag{15}$$

A constant energy surface is shown in Fig. 18.

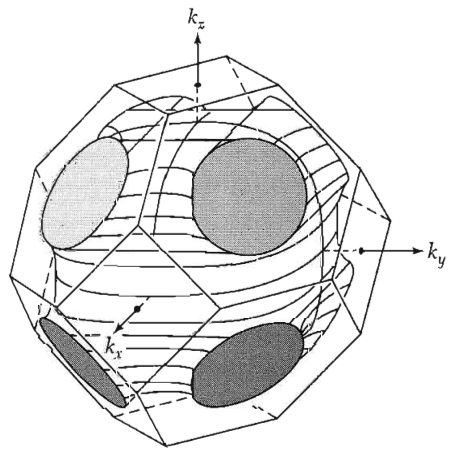


Figure 18 A constant energy surface of an fcc crystal structure, in the nearest-neighbor tight-binding approximation. The surface shown has $\epsilon = -\alpha + 2|\gamma|$.

Wigner-Seitz Method

Wigner and Seitz showed that for the alkali metals there is no inconsistency between the electron wavefunctions of free atoms and the nearly free electron model of the band structure of a crystal. Over most of a band the energy may depend on the wavevector nearly as for a free electron. However the Bloch wavefunction, unlike a plane wave, will pile up charge on the positive ion cores as in the atomic wavefunction.

A Bloch function satisfies the wave equation

$$\left(\frac{1}{2m}\mathbf{p}^2 + U(\mathbf{r})\right)e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r}) = \epsilon_{\mathbf{k}}e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r}) \ . \tag{16}$$

With $\mathbf{p} \equiv -i\hbar$ grad, we have

$$\mathbf{p}\, e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r}) = \hbar\mathbf{k}\, e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r}) + e^{i\mathbf{k}\cdot\mathbf{r}}\mathbf{p}u_{\mathbf{k}}(\mathbf{r}) \ ;$$

$$\mathbf{p}^2 e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r}) = (\hbar k)^2 e^{i\mathbf{k}\cdot\mathbf{r}}u_{\mathbf{k}}(\mathbf{r}) + e^{i\mathbf{k}\cdot\mathbf{r}}(2\hbar\mathbf{k}\cdot\mathbf{p})u_{\mathbf{k}}(\mathbf{r}) + e^{i\mathbf{k}\cdot\mathbf{r}}\mathbf{p}^2u_{\mathbf{k}}(\mathbf{r}) \ ;$$

thus the wave equation (16) may be written as an equation for $u_{\mathbf{k}}$:

$$\left(\frac{1}{2m}(\mathbf{p} + \hbar\mathbf{k})^2 + U(\mathbf{r})\right)u_{\mathbf{k}}(\mathbf{r}) = \epsilon_{\mathbf{k}}u_{\mathbf{k}}(\mathbf{r}) \ . \tag{17}$$

At $\mathbf{k} = 0$ we have $\psi_0 = u_0(\mathbf{r})$, where $u_0(\mathbf{r})$ has the periodicity of the lattice, sees the ion cores, and near them will look like the wavefunction of the free atom.

It is much easier to find a solution at $\mathbf{k} = 0$ than at a general $\mathbf{k}$, because at $\mathbf{k} = 0$ a nondegenerate solution will have the full symmetry of $U(\mathbf{r})$, that is, of the crystal. We can then use $u_0(\mathbf{r})$ to construct the approximate solution

$$\psi_{\mathbf{k}} = \exp(i\mathbf{k}\cdot\mathbf{r})u_0(r) \ . \tag{18}$$

This is of the Bloch form, but u_0 is not an exact solution of (17): it is a solution only if we drop the term in $\mathbf{k}\cdot\mathbf{p}$. Often this term is treated as a perturbation, as in Problem 8. The $\mathbf{k}\cdot\mathbf{p}$ perturbation theory developed there is especially useful in finding the effective mass m^* at a band edge.

Because it takes account of the ion core potential the function (18) is a much better approximation than a plane wave to the correct wavefunction. The energy of the approximate solution depends on $\mathbf{k}$ as $(\hbar k)^2/2m$, exactly as for the plane wave, even though the modulation represented by $u_0(\mathbf{r})$ may be very strong. Because u_0 is a solution of

$$\left(\frac{1}{2m}\mathbf{p}^2 + U(\mathbf{r})\right)u_0(\mathbf{r}) = \epsilon_0 u_0(\mathbf{r}) \ , \tag{19}$$

the function (18) has the energy expectation value $\epsilon_0 + (\hbar^2k^2/2m)$. The function $u_0(\mathbf{r})$ often will give us a good picture of the charge distribution within a cell.

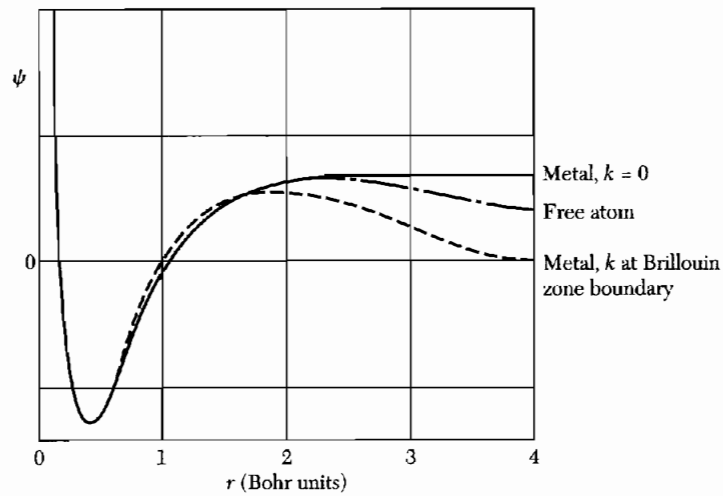


Figure 19 Radial wavefunctions for the 3*s* orbital of free sodium atom and for the 3*s* conduction band in sodium metal. The wavefunctions, which are not normalized here, are found by integrating the Schrödinger equation for an electron in the potential well of an Na^+ ion core. For the free atom the wavefunction is integrated subject to the usual Schrödinger boundary condition $\psi(r) \to 0$ as $r \to \infty$; the energy eigenvalue is -5.15 eV. The wavefunction for wavevector $k = 0$ in the metal is subject to the Wigner-Seitz boundary condition that $d\psi/dr = 0$ when r is midway between neighboring atoms; the energy of this orbital is -8.2 eV, considerably lower than for the free atom. The orbitals at the zone boundary are not filled in sodium; their energy is $+2.7$ eV. (After E. Wigner and F. Seitz.)

Wigner and Seitz developed a simple and fairly accurate method of calculating $u_0(\mathbf{r})$. Figure 19 shows the Wigner-Seitz wavefunction for $\mathbf{k} = 0$ in the 3*s* conduction band of metallic sodium. The function is practically constant over 0.9 of the atomic volume. To the extent that the solutions for higher $\mathbf{k}$ may be approximated by $\exp(i\mathbf{k} \cdot \mathbf{r})u_0(\mathbf{r})$, the wavefunctions in the conduction band will be similar to plane waves over most of the atomic volume, but increase markedly and oscillate within the ion core.

Cohesive Energy. The stability of the simple metals with respect to free atoms is caused by the lowering of the energy of the Bloch orbital with $\mathbf{k} = 0$ in the crystal compared to the ground valence orbital of the free atom. The effect is illustrated in Fig. 19 for sodium and in Fig. 20 for a linear periodic potential of attractive square wells. The ground orbital energy is much lower (because of lower kinetic energy) at the actual spacing in the metal than for isolated atoms.

A decrease in ground orbital energy will increase the binding. The decrease in ground orbital energy is a consequence of the change in the boundary condition on the wavefunction: The Schrödinger boundary condition for the free atom is $\psi(\mathbf{r}) \to 0$ as $r \to \infty$. In the crystal the $\mathbf{k} = 0$ wavefunction $u_0(\mathbf{r})$ has the symmetry of the lattice and is symmetric about r = 0. To have this, the normal derivative of ψ must vanish across every plane midway between adjacent atoms.

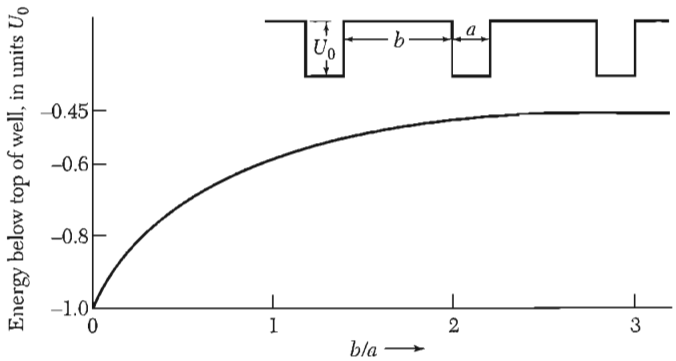


Figure 20 Ground orbital ($k = 0$) energy for an electron in a periodic square well potential of depth $|U_0| = 2\hbar^2/ma^2$. The energy is lowered as the wells come closer together. Here a is held constant and b is varied. Large b/a corresponds to separated atoms. (Courtesy of C. Y. Fong.)

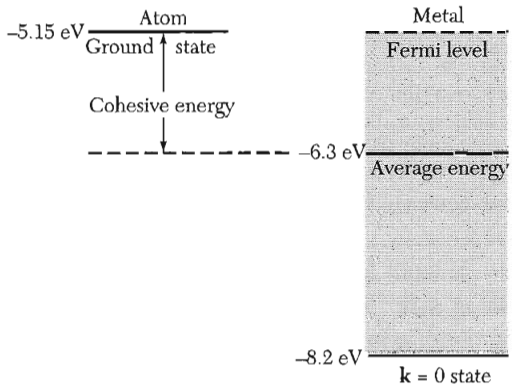


Figure 21 Cohesive energy of sodium metal is the difference between the average energy of an electron in the metal (−6.3 eV) and the ground state energy (−5.15 eV) of the valence 3*s* electron in the free atom, referred to an Na^+ ion plus free electron at infinite separation.

In a spherical approximation to the shape of the smallest Wigner-Seitz cell we use the Wigner-Seitz boundary condition

$$(d\psi/dr)_{r_0} = 0 \ , \tag{20}$$

where r_0 is the radius of a sphere equal in volume to a primitive cell of the lattice. In sodium, $r_0 = 3.95$ Bohr units, or 2.08 Å; the half distance to a nearest neighbor is 1.86 Å. The spherical approximation is not bad for fcc and bcc structures. The boundary condition allows the ground orbital wavefunction to have much less curvature than the free atom boundary condition. Much less curvature means much less kinetic energy.

In sodium the other filled orbitals in the conduction band can be represented in a rough approximation by wavefunctions of the form (18), with

$$\psi_{\mathbf{k}} = e^{i\mathbf{k}\cdot\mathbf{r}} u_0(\mathbf{r}) \ ; \quad \epsilon_{\mathbf{k}} = \epsilon_0 + \frac{\hbar^2 k^2}{2m} \ .$$

The Fermi energy is 3.1 eV, from Table 6.1. The average kinetic energy per electron is 0.6 of the Fermi energy, or 1.9 eV. Because $\epsilon_0 = -8.2$ eV at $\mathbf{k} = 0$, the average electron energy is $\langle \epsilon_k \rangle = -8.2 + 1.9 = -6.3$ eV, compared with -5.15 eV for the valence electron of the free atom, Fig. 21.

We therefore estimate that sodium metal is stable by about 1.1 eV with respect to the free atom. This result agrees well with the experimental value 1.13 eV.

Pseudopotential Methods

Conduction electron wavefunctions are usually smoothly varying in the region between the ion cores, but have a complicated nodal structure in the region of the cores. This behavior is illustrated by the ground orbital of sodium, Fig. 19. It is helpful to view the nodes in the conduction electron wavefunction in the core region as created by the requirement that the function be orthogonal to the wavefunctions of the core electrons. This all comes out of the Schrödinger equation, but we can see that we need the flexibility of two nodes in the 3*s* conduction orbital of Na in order to be orthogonal both to the 1*s* core orbital with no nodes and the 2*s* core orbital with one node.

Outside the core the potential energy that acts on the conduction electron is relatively weak: the potential energy is only the coulomb potential of the singly-charged positive ion cores and is reduced markedly by the electrostatic screening of the other conduction electrons, Chapter 14. In this outer region the conduction electron wavefunctions are as smoothly varying as plane waves.

If the conduction orbitals in this outer region are approximately plane waves, the energy must depend on the wavevector approximately as $\epsilon_k = \hbar^2 k^2/2m$ as for free electrons. But how do we treat the conduction orbitals in the core region where the orbitals are not at all like plane waves?

What goes on in the core is largely irrelevant to the dependence of ϵ on $\mathbf{k}$. Recall that we can calculate the energy by applying the hamiltonian operator to an orbital at any point in space. Applied in the outer region, this operation will give an energy nearly equal to the free electron energy.

This argument leads naturally to the idea that we might replace the actual potential energy (and filled shells) in the core region by an effective potential energy[1] that gives the same wavefunctions outside the core as are given by the actual ion cores. It is startling to find that the effective potential or

[1] J. C. Phillips and L. Kleinman, Phys. Rev. **116**, 287 (1959); E. Antoncik, J. Phys. Chem. Solids **10**, 314 (1959). The general theory of pseudopotentials is discussed by B. J. Austin, V. Heine, and L. J. Sham, Phys. Rev. **127**, 276 (1962); see also Vol. 24 of *Solid state physics*. The utility of the empty core model has been known for many years: it goes back to E. Fermi, Nuovo Cimento **2**, 157 (1934); H. Hellmann, Acta Physiochimica URSS 1, 913 (1935); and H. Hellmann and W. Kassatotschkin, J. Chem. Phys. **4**, 324 (1936), who wrote "Since the field of the ion determined in this way runs a rather flat course, it is sufficient in the first approximation to set the valence electron in the lattice equal to a plane wave."

pseudopotential that satisfies this requirement is nearly zero. This conclusion about pseudopotentials is supported by a large amount of empirical experience as well as by theoretical arguments. The result is referred to as the **cancellation theorem**.

The pseudopotential for a problem is not unique nor exact, but it may be very good. On the Empty Core Model (ECM) we can even take the unscreened pseudopotential to be zero inside some radius R_e:

$$U(r) = \begin{cases} 0 & , \quad \text{for } r < R_e \ ; \\ -e^2/r & , \quad \text{for } r > R_e \ . \end{cases} \tag{21}$$

This potential should now be screened as described in Chapter 10. Each component $U(\mathbf{K})$ of $U(r)$ is to be divided by the dielectric constant $\epsilon(\mathbf{K})$ of the electron gas. If, just as an example, we use the Thomas-Fermi dielectric function (14.33), we obtain the screened pseudopotential plotted in Fig. 22a.

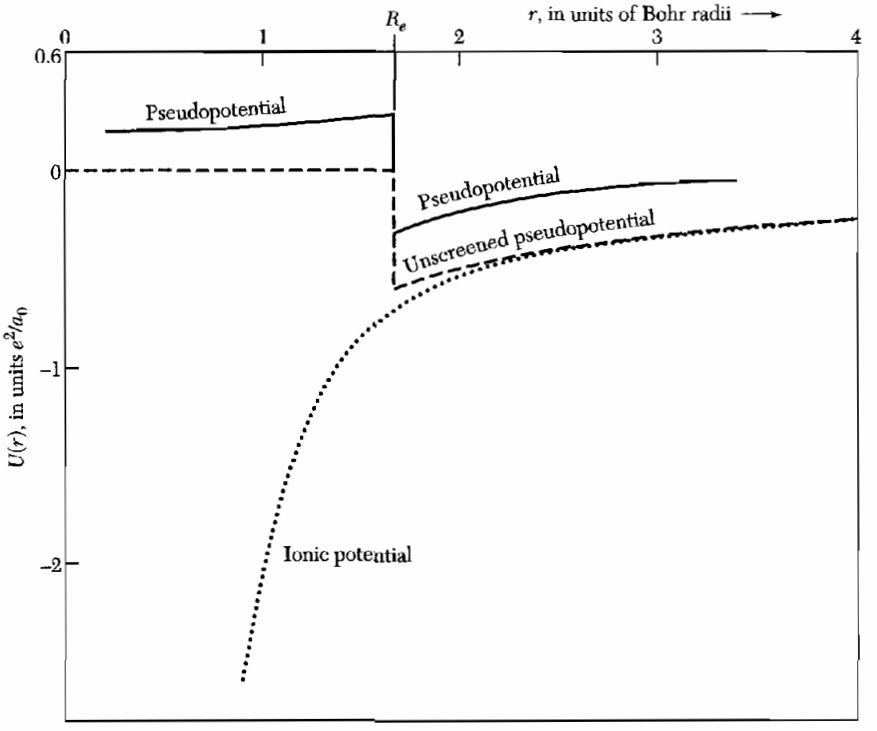


Figure 22a Pseudopotential for metallic sodium, based on the empty core model and screened by the Thomas-Fermi dielectric function. The calculations were made for an empty core radius $R_e = 1.66a_0$, where a_0 is the Bohr radius, and for a screening parameter $k_s a_0 = 0.79$. The dashed curve shows the assumed unscreened potential, as from (21). The dotted curve is the actual potential of the ion core; other values of $U(r)$ are -50.4, -11.6, and -4.6, for $r = 0.15$, 0.4, and 0.7, respectively. Thus the actual potential of the ion (chosen to fit the energy levels of the free atom) is very much larger than the pseudopotential, over 200 times larger at $r = 0.15$.

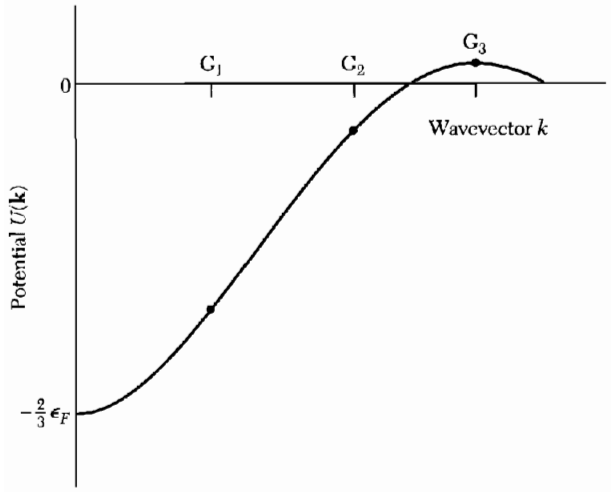


Figure 22b A typical reciprocal space pseudopotential. Values of $U(\mathbf{k})$ for wavevectors equal to the reciprocal lattice vectors, $\mathbf{G}$, are indicated by the dots. For very small $\mathbf{k}$ the potential approaches $(-2/3)$ times the Fermi energy, which is the screened-ion limit for metals. (After M. L. Cohen.)

The pseudopotential as drawn is much weaker than the true potential, but the pseudopotential was adjusted so that the wavefunction in the outer region is nearly identical to that for the true potential. In the language of scattering theory, we adjust the phase shifts of the pseudopotential to match those of the true potential.

Calculation of the band structure depends only on the Fourier components of the pseudopotential at the reciprocal lattice vectors. Usually only a few values of the coefficients $U(\mathbf{G})$ are needed to get a good band structure: see the $U(\mathbf{G})$ in Fig. 22b. These coefficients are sometimes calculated from model potentials, and sometimes they are obtained from fits of tentative band structures to the results of optical measurements. Good values of $U(0)$ can be estimated from first principles; it is shown in (14.43) that for a screened coulomb potential $U(0) = -\frac{2}{3}\epsilon_F$.

In the remarkably successful Empirical Pseudopotential Method (EPM) the band structure is calculated using a few coefficients $U(\mathbf{G})$ deduced from theoretical fits to measurements of the optical reflectance and absorption of crystals, as discussed in Chapter 15. Charge density maps can be plotted from the wavefunctions generated by the EPM—see Fig. 3.11. The results are in excellent agreement with x-ray diffraction determinations; such maps give an understanding of the bonding and have great predictive value for proposed new structures and compounds.

The EPM values of the coefficients $U(\mathbf{G})$ often are additive in the contributions of the several types of ions that are present. Thus it may be possible to construct the $U(\mathbf{G})$ for entirely new structures, starting from results on known structures. Further, the pressure dependence of a band structure may be determined when it is possible to estimate from the form of the $U(r)$ curve the dependence of $U(\mathbf{G})$ on small variations of $\mathbf{G}$.

It is often possible to calculate band structures, cohesive energy, lattice constants, and bulk moduli from first principles. In such *ab initio* pseudopotential calculations the basic inputs are the crystal structure type and the atomic number, along with well-tested theoretical approximations to exchange energy terms. This is not the same as calculating from atomic number alone, but it is the most reasonable basis for a first-principles calculation. The results of Yin and Cohen are compared with experiment in the table that follows.

	Lattice constant (Å)	Cohesive energy (eV)	Bulk modulus (Mbar)
Silicon			
Calculated	5.45	4.84	0.98
Experimental	5.43	4.63	0.99
Germanium			
Calculated	5.66	4.26	0.73
Experimental	5.65	3.85	0.77
Diamond			
Calculated	3.60	8.10	4.33
Experimental	3.57	7.35	4.43

EXPERIMENTAL METHODS IN FERMI SURFACE STUDIES

Powerful experimental methods have been developed for the determination of Fermi surfaces. The methods include magnetoresistance, anomalous skin effect, cyclotron resonance, magneto-acoustic geometric effects, the Shubnikow-de Haas effect, and the de Haas-van Alphen effect. Further information on the momentum distribution is given by positron annihilation, Compton scattering, and the Kohn effect.

We propose to study one method rather thoroughly. All the methods are useful, but need detailed theoretical analysis. We select the de Haas-van Alphen effect because it exhibits very well the characteristic periodicity in $1/B$ of the properties of a metal in a uniform magnetic field.

Quantization of Orbits in a Magnetic Field

The momentum $\mathbf{p}$ of a particle in a magnetic field is the sum (Appendix G) of two parts, the kinetic momentum $\mathbf{p}_{\text{kin}} = m\mathbf{v} = \hbar\mathbf{k}$ and the potential momentum or field momentum $\mathbf{p}_{\text{field}} = q\mathbf{A}/c$, where q is the charge. The vector potential is related to the magnetic field by $\mathbf{B} = \text{curl}\,\mathbf{A}$. The total momentum is

$$\text{(CGS)} \qquad \mathbf{p} = \mathbf{p}_{\text{kin}} + \mathbf{p}_{\text{field}} = \hbar\mathbf{k} + q\mathbf{A}/c \ . \qquad (22)$$

In SI the factor c^{-1} is omitted.

Following the semiclassical approach of Onsager and Lifshitz, we assume that the orbits in a magnetic field are quantized by the Bohr-Sommerfeld relation

$$\oint \mathbf{p} \cdot d\mathbf{r} = (n + \gamma)2\pi\hbar \ , \tag{23}$$

when n is an integer and γ is a phase correction that for free electrons has the value $\frac{1}{2}$. Then

$$\oint \mathbf{p} \cdot d\mathbf{r} = \oint \hbar\mathbf{k} \cdot d\mathbf{r} + \frac{q}{c}\oint \mathbf{A} \cdot d\mathbf{r} \ . \tag{24}$$

The equation of motion of a particle of charge q in a magnetic field is

$$\hbar\frac{d\mathbf{k}}{dt} = \frac{q}{c}\frac{d\mathbf{r}}{dt} \times \mathbf{B} \ . \tag{25a}$$

We integrate with respect to time to give

$$\hbar\mathbf{k} = \frac{q}{c}\mathbf{r} \times \mathbf{B} \ ,$$

apart from an additive constant which does not contribute to the final result.

Thus one of the path integrals in (24) is

$$\oint \hbar\mathbf{k} \cdot d\mathbf{r} = \frac{q}{c}\oint \mathbf{r} \times \mathbf{B} \cdot d\mathbf{r} = -\frac{q}{c}\mathbf{B} \cdot \oint \mathbf{r} \times d\mathbf{r} = -\frac{2q}{c}\Phi \ , \tag{25b}$$

where Φ is the magnetic flux contained within the orbit in real space. We have used the geometrical result that

$$\oint \mathbf{r} \times d\mathbf{r} = 2 \times (\text{area enclosed by the orbit}) \ .$$

The other path integral in (24) is

$$\frac{q}{c}\oint \mathbf{A} \cdot d\mathbf{r} = \frac{q}{c}\int \mathbf{curl}\ \mathbf{A} \cdot d\boldsymbol{\sigma} = \frac{q}{c}\int \mathbf{B} \cdot d\boldsymbol{\sigma} = \frac{q}{c}\Phi \ , \tag{25c}$$

by the Stokes theorem. Here $d\boldsymbol{\sigma}$ is the area element in real space. The momentum path integral is the sum of (25b) and (25c):

$$\oint \mathbf{p} \cdot d\mathbf{r} = -\frac{q}{c}\boldsymbol{\Phi} = (n + \gamma)2\pi\hbar \ . \tag{26}$$

It follows that the orbit of an electron is quantized in such a way that the flux through it is

$$\Phi_n = (n + \gamma)(2\pi\hbar c/e) \ . \tag{27}$$

The flux unit $2\pi\hbar c/e = 4.14 \times 10^{-7}$ gauss cm^2 or T m^2.

In the de Haas-van Alphen effect discussed below we need the area of the orbit in wavevector space. We obtained in (27) the flux through the orbit in

real space. By (25a) we know that a line element Δr in the plane normal to $\mathbf{B}$ is related to Δk by $\Delta r = (\hbar c/eB)\Delta k$, so that the area S_n in $\mathbf{k}$ space is related to the area A_n of the orbit in $\mathbf{r}$ space by

$$A_n = (\hbar c/eB)^2 S_n \ . \tag{28}$$

It follows that

$$\Phi_n = \left(\frac{\hbar c}{e}\right)^2 \frac{1}{B} S_n = (n + \gamma)\frac{2\pi\hbar c}{e} \ , \tag{29}$$

from (27), whence the area of an orbit in $\mathbf{k}$ space will satisfy

$$S_n = (n + \gamma)\frac{2\pi e}{\hbar c} B \ . \tag{30}$$

In Fermi surface experiments we may be interested in the increment ΔB for which two successive orbits, n and $n + 1$, have the same area in $\mathbf{k}$ space on the Fermi surface. The areas are equal when

$$S\left(\frac{1}{B_{n+1}} - \frac{1}{B_n}\right) = \frac{2\pi e}{\hbar c} \ , \tag{31}$$

from (30). We have the important result that equal increments of $1/B$ reproduce similar orbits—this periodicity in $1/B$ is a striking feature of the magneto-oscillatory effects in metals at low temperatures: resistivity, susceptibility, heat capacity.

The population of orbits on or near the Fermi surface oscillates as B is varied, causing a wide variety of effects. From the period of the oscillation we reconstruct the Fermi surface. The result (30) is independent of the gauge of the vector potential used in the expression (22) for momentum; that is, $\mathbf{p}$ is not gauge invariant, but S_n is. Gauge invariance is discussed further in Chapter 10 and in Appendix G.

De Haas-van Alphen Effect

The de Haas-van Alphen effect is the oscillation of the magnetic moment of a metal as a function of the static magnetic field intensity. The effect can be observed in pure specimens at low temperatures in strong magnetic fields: we do not want the quantization of the electron orbits to be blurred by collisions, and we do not want the population oscillations to be averaged out by thermal population of adjacent orbits.

The analysis of the dHvA effect is given for absolute zero in Fig. 23. The electron spin is neglected. The treatment is given for a two-dimensional (2D) system; in 3D we need only multiply the 2D wavefunction by plane wave factors $\exp(ik_z z)$, where the magnetic field is parallel to the z axis. The area of an orbit in k_x, k_y space is quantized as in (30). The area between successive orbits is

$$\Delta S = S_n - S_{n-1} = 2\pi eB/\hbar c \ . \tag{32}$$

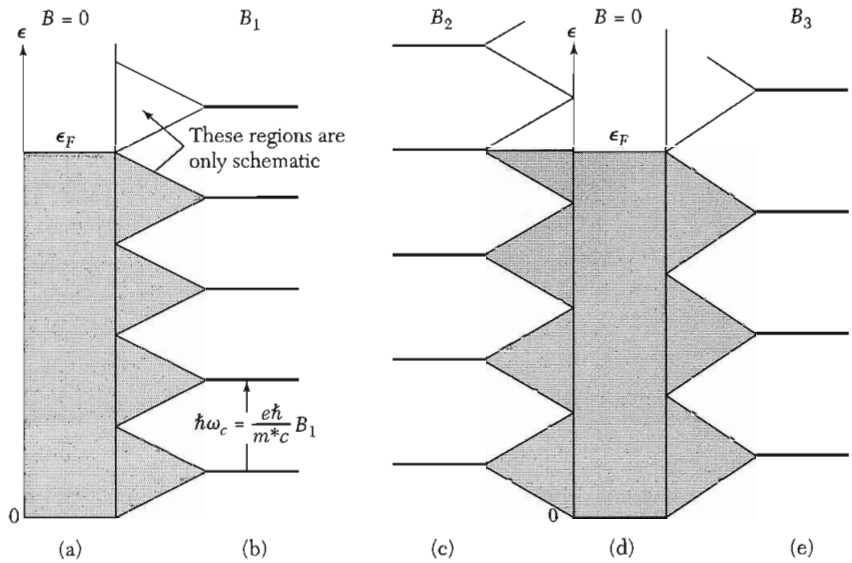


Figure 23 Explanation of the de Haas-van Alphen effect for a free electron gas in two dimensions in a magnetic field. The filled orbitals of the Fermi sea in the absence of a magnetic field are shaded in *a* and *d*. The energy levels in a magnetic field are shown in *b*, *c*, and *e*. In *b* the field has a value B_1 such that the total energy of the electrons is the same as in the absence of a magnetic field: as many electrons have their energy raised as lowered by the orbital quantization in the magnetic field B_1. When we increase the field to B_2 the total electron energy is increased, because the uppermost electrons have their energy raised. In *e* for field B_3 the energy is again equal to that for the field $B = 0$. The total energy is a minimum at points such as B_1, B_3, B_5, . . . , and a maximum near points such as B_2, B_4, ...

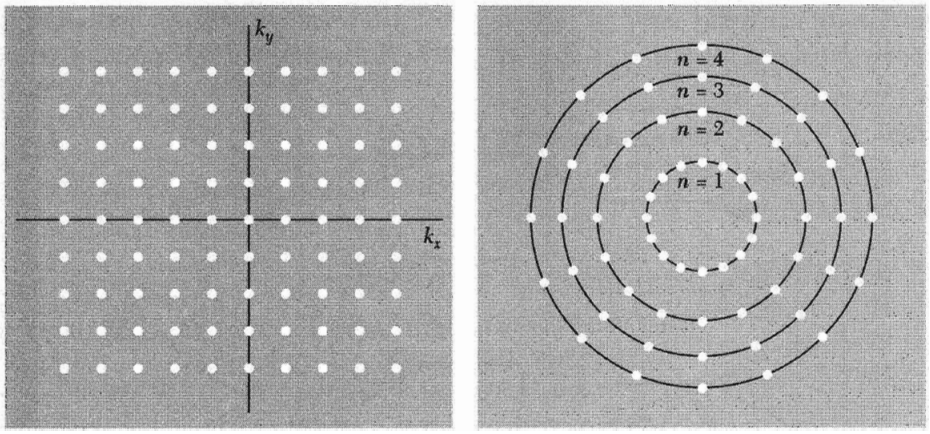
The area in **k** space occupied by a single orbital is $(2\pi/L)^2$, neglecting spin, for a square specimen of side L. Using (32) we find that the number of free electron orbitals that coalesce in a single magnetic level is

$$D = (2\pi eB/\hbar c)(L/2\pi)^2 = \rho B \ , \tag{33}$$

where $\rho = eL^2/2\pi\hbar c$, as in Fig. 24. Such a magnetic level is called a **Landau level**.

The dependence of the Fermi level on B is dramatic. For a system of N electrons at absolute zero the Landau levels are entirely filled up to a magnetic quantum number we identify by s, where s is a positive integer. Orbitals at the next higher level $s + 1$ will be partly filled to the extent needed to accommodate the electrons. The Fermi level will lie in the Landau level $s + 1$ if there are electrons in this level; as the magnetic field is increased the electrons move to lower levels. When $s + 1$ is vacated, the Fermi level moves down abruptly to the next lower level s.

The electron transfer to lower Landau levels can occur because their degeneracy D increases as B is increased, as shown in Fig. 25. As B is

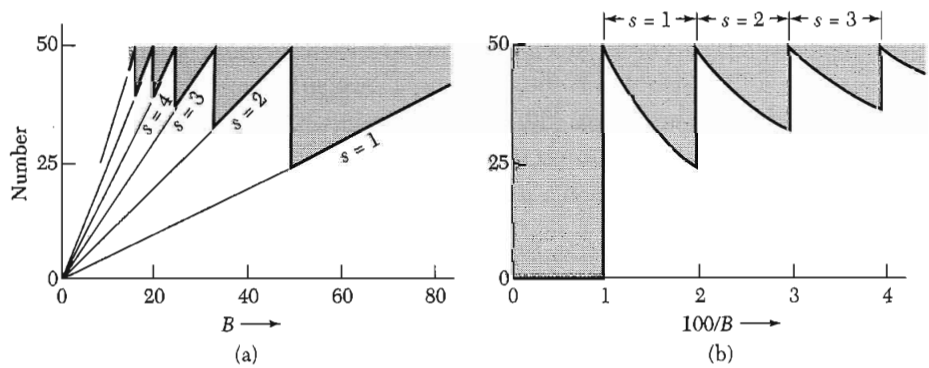


(a) (b)

Figure 24 (a) Allowed electron orbitals in two dimensions in absence of a magnetic field. (b) In a magnetic field the points which represent the orbitals of free electrons may be viewed as restricted to circles in the former $k_x k_y$ plane. The successive circles correspond to successive values of the quantum number n in the energy $(n - \frac{1}{2})k\omega_c$. The area between successive circles is

(CGS) $$\pi\Delta(k^2) = 2\pi k(\Delta k) = (2\pi m/\hbar^2)\,\Delta\epsilon = 2\pi m\omega_c/\hbar - 2\pi eB/\hbar c \ .$$

The angular position of the points has no significance. The number of orbitals on a circle is constant and is equal to the area between successive circles times the number of orbitals per unit area in (a), or $(2\pi eB/\hbar c)(L/2\pi)^2 = L^2eB/2\pi\hbar c$, neglecting electron spin.



(a) (b)

Figure 25 (a) The heavy line gives the number of particles in levels which are completely occupied in a magnetic field B, for a two-dimensional system with $N = 50$ and $\rho = 0.50$. The shaded area gives the number of particles in levels partially occupied. The value of s denotes the quantum number of the highest level which is completely filled. Thus at $B = 40$ we have $s = 2$; the levels $n = 1$ and $n = 2$ are filled and there are 10 particles in the level $n = 3$. At $B = 50$ the level $n = 3$ is empty. (b) The periodicity in $1/B$ is evident when the same points are plotted against $1/B$.

increased there occur values of B at which the quantum number of the uppermost filled level decreases abruptly by unity. At the critical magnetic fields labeled B_s no level is partly occupied at absolute zero, so that

$$s\rho B_s = N \ . \tag{34}$$

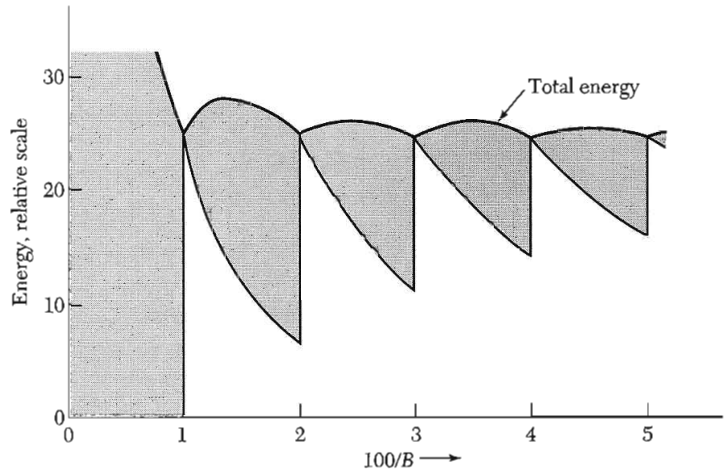


Figure 26 The upper curve is the total electronic energy versus $1/B$. The oscillations in the energy U may be detected by measurement of the magnetic moment, given by $-\partial U/\partial B$. The thermal and transport properties of the metal also oscillate as successive orbital levels cut through the Fermi level when the field is increased. The shaded region in the figure gives the contribution to the energy from levels that are only partly filled. The parameters for the figure are the same as for Fig. 25, and we have taken the units of B such that $B = \hbar\omega_c$.

The number of filled levels times the degeneracy at B_s must equal the number of electrons N.

To show the periodicity of the energy as B is varied, we use the result that the energy of the Landau level of magnetic quantum number n is $E_n = (n - \frac{1}{2})\hbar\omega_c$, where $\omega_c = eB/m^*c$ is the cyclotron frequency. The result for E_n follows from the analogy between the cyclotron resonance orbits and the simple harmonic oscillator, but now we have found it convenient to start counting at $n = 1$ instead of at $n = 0$.

The total energy of the electrons in levels that are fully occupied is

$$\sum_{n=1}^{s} D\hbar\omega_c(n - \tfrac{1}{2}) = \tfrac{1}{2}D\hbar\omega_c s^2 \ , \tag{35}$$

where D is the number of electrons in each level. The total energy of the electrons in the partly occupied level $s + 1$ is

$$\hbar\omega_c(s + \tfrac{1}{2})(N - sD) \ , \tag{36}$$

where sD is the number of electrons in the lower filled levels. The total energy of the N electrons is the sum of (35) and (36), as in Fig. 26.

The magnetic moment μ of a system at absolute zero is given by $\mu = -\partial U/\partial B$. The moment here is an oscillatory function of $1/B$, Fig. 27. This oscillatory magnetic moment of the Fermi gas at low temperatures is the de Haas-van Alphen effect. From (31) we see that the oscillations occur at equal intervals of $1/B$ such that

$$\Delta\left(\frac{1}{B}\right) = \frac{2\pi e}{\hbar c S} \ , \tag{37}$$

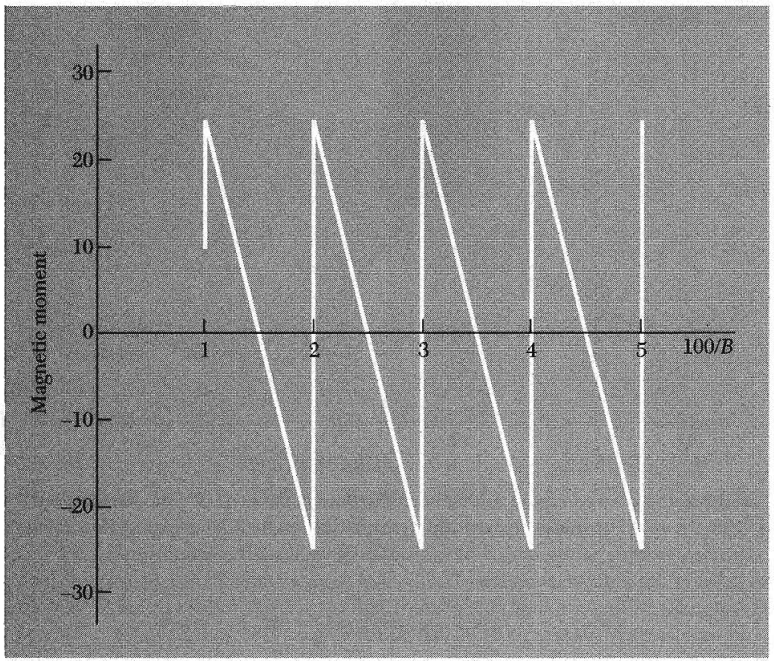


Figure 27 At absolute zero the magnetic moment is given by $-\partial U \partial B$. The energy plotted in Fig. 26 leads to the magnetic moment shown here, an oscillatory function of $1/B$. In impure specimens the oscillations are smudged out in part because the energy levels are no longer sharply defined.

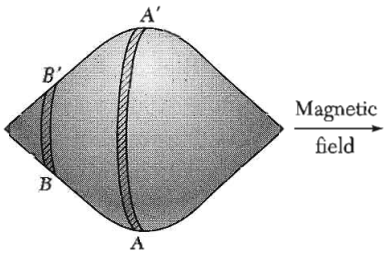


Figure 28 The orbits in the section AA' are extremal orbits: the cyclotron period is roughly constant over a reasonable section of the Fermi surface. Other sections such as BB' have orbits that vary in period along the section.

where S is the extremal area (see below) of the Fermi surface normal to the direction of **B**. From measurements of $\Delta(1/B)$, we deduce the corresponding extremal areas S; thereby much can be inferred about the shape and size of the Fermi surface.

Extremal Orbits. One point in the interpretation of the dHvA effect is subtle. For a Fermi surface of general shape the sections at different values of k_B will have different periods. Here k_B is the component of k along the direction of the magnetic field. The response will be the sum of contributions from all sections or all orbits. *But the dominant response of the system comes from orbits whose periods are stationary with respect to small changes in* k_B. Such

orbits are called **extremal orbits**. Thus, in Fig. 28 the section AA' dominates the observed cyclotron period.

The argument can be put in mathematical form, but we do not give the proof here (*QTS*, p. 223). Essentially it is a question of phase cancellation; the contributions of different nonextremal orbits cancel, but near the extrema the phase varies only slowly and there is a net signal from these orbits. Sharp resonances are obtained even from complicated Fermi surfaces because the experiment selects the extremal orbits.

Fermi Surface of Copper. The Fermi surface of copper (Fig. 29) is distinctly nonspherical: eight necks make contact with the hexagonal faces of the first Brillouin zone of the fcc lattice. The electron concentration in a monovalent metal with an fcc structure is $n = 4/a^3$; there are four electrons in a cube of volume a^3. The radius of a free electron Fermi sphere is

$$k_F = (3\pi^2 n)^{1/3} = (12\pi^2/a^3)^{1/3} \cong (4.90/a) \ , \tag{38}$$

and the diameter is $9.80/a$.

The shortest distance across the Brillouin zone (the distance between hexagonal faces) is $(2\pi/a)(3)^{1/2} = 10.88/a$, somewhat larger than the diameter of the free electron sphere. The sphere does not touch the zone boundary, but we know that the presence of a zone boundary tends to lower the band energy near the boundary. Thus it is plausible that the Fermi surface should neck out to meet the closest (hexagonal) faces of the zone (Figs. 18 and 29).

The square faces of the zone are more distant, with separation $12.57/a$, and the Fermi surface does not neck out to meet these faces.

EXAMPLE: *Fermi Surface of Gold.* In gold for quite a wide range of field directions Shoenberg finds the magnetic moment has a period of 2×10^{-9} gauss^{-1}. This period corresponds to an extremal orbit of area

$$S = \frac{2\pi e/\hbar c}{\Delta(1/B)} \cong \frac{9.55 \times 10^7}{2 \times 10^{-9}} \cong 4.8 \times 10^{16} \text{ cm}^{-2} \ .$$

From Table 6.1, we have $k_F = 1.2 \times 10^8$ cm^{-1} for a free electron Fermi sphere for gold, or an extremal area of 4.5×10^{16} cm^{-2}, in general agreement with the experimental value. The actual periods reported by Shoenberg are 2.05×10^{-9} gauss^{-1} and 1.95×10^{-9} gauss^{-1}. In the [111] direction in Au a large period of 6×10^{-8} gauss^{-1} is also found; the corresponding orbital area is 1.6×10^{15} cm^{-2}. This is the "neck" orbit N. Another extremal orbit, the "dog's bone," is shown in Fig. 30; its area in Au is about 0.4 of the belly area. Experimental results are shown in Fig. 31. To do the example in SI, drop c from the relation for S and use as the period 2×10^{-5} tesla^{-1}.

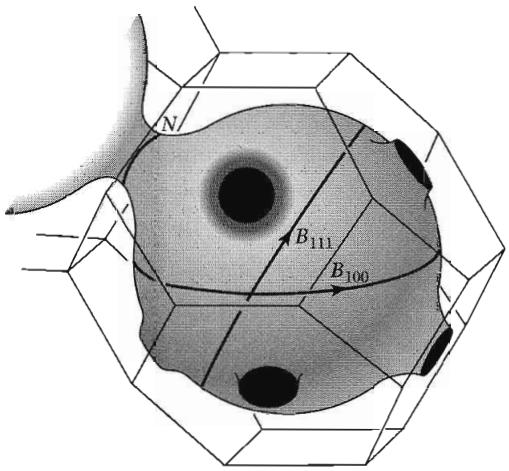


Figure 29 Fermi surface of copper, after Pippard. The Brillouin zone of the fcc structure is the truncated octahedron derived in Chapter 2. The Fermi surface makes contact with the boundary at the center of the hexagonal faces of the zone, in the [111] directions in **k** space. Two "belly" extremal orbits are shown, denoted by B; the extremal "neck" orbit is denoted by N.

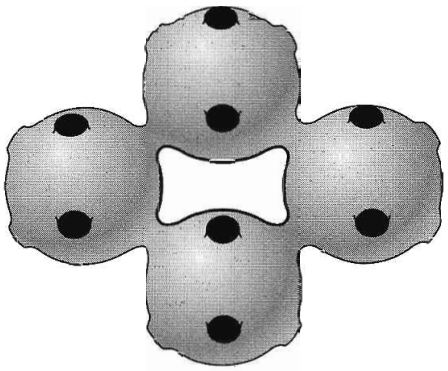


Figure 30 Dog's bone orbit of an electron on the Fermi surface of copper or gold in a magnetic field. This orbit is classified as holelike because the energy increases toward the interior of the orbit.

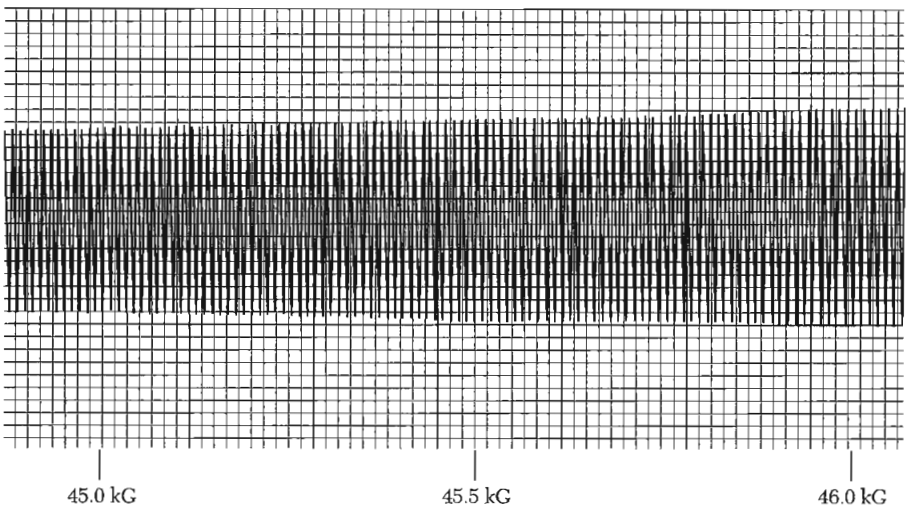


Figure 31 De Haas-van Alphen effect in gold with **B** || [110]. The oscillation is from the dog's bone orbit of Fig. 30. The signal is related to the second derivative of the magnetic moment with respect to field. The results were obtained by a field modulation technique in a high-homogeneity superconducting solenoid at about 1.2 K. (Courtesy of I. M. Templeton.)

The free electron Fermi sphere of aluminum fills the first zone entirely and has a large overlap into the second and third zones, Fig. 1. The third zone Fermi surface is quite complicated, even though it is just made up of certain pieces of the surface of the free electron sphere. The free electron model also gives small pockets of holes in the fourth zone, but when the lattice potential is

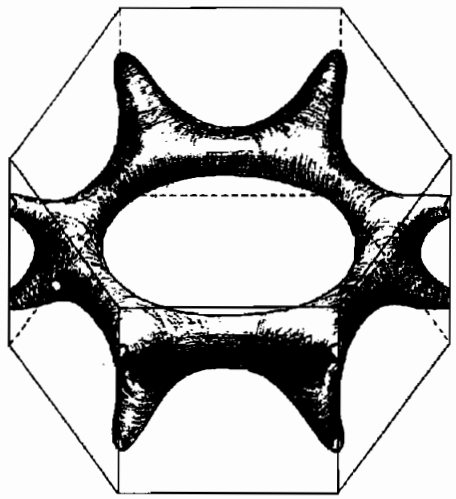


Figure 32 Multiply-connected hole surface of magnesium in bands 1 and 2, according to L. M. Falicov. (Drawing by Marta Puebla.)

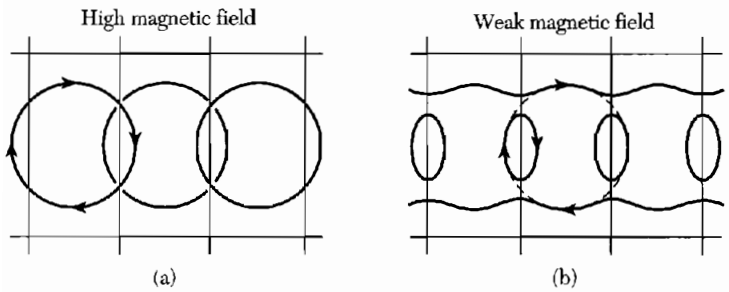


Figure 33 Breakdown of band structure by a strong magnetic field. Brillouin zone boundaries are the light lines. The free electron orbits (a) in a strong field change connectivity in a weak field (b) to become open orbits in the first band and electron orbits in the second band. Both bands are mapped together.

taken into account these empty out, the electrons being added to the third zone. The general features of the predicted Fermi surface of aluminum are quite well verified by experiment. Figure 32 shows part of the free electron Fermi surface of magnesium.

Magnetic Breakdown. Electrons in sufficiently high magnetic fields will move in free particle orbits, the circular cyclotron orbits of Fig. 33a. Here the magnetic forces are dominant, and the lattice potential is a slight perturbation. In this limit the classification of the orbitals into bands may have little importance. However, we know that at low magnetic fields the motion is described by (8.7) with the band structure $\epsilon_{\mathbf{k}}$ that obtains in the absence of a magnetic field.

The eventual breakdown of this description as the magnetic field is increased is called magnetic breakdown. The passage to strong magnetic fields may drastically change the connectivity of the orbits, as in the figure. The onset of magnetic breakdown will be revealed by physical properties such as

magnetoresistance that depend sensitively on the connectivity. The condition for magnetic breakdown is that $\hbar\omega_c\epsilon_F > E_g^2$, approximately. Here ϵ_F is the free electron Fermi energy and E_g is the energy gap. This condition is much milder, especially in metals with small gaps, than the naïve condition that the magnetic splitting $\hbar\omega_c$ exceed the gap.

Small gaps may be found in hcp metals where the gap across the hexagonal face of the zone would be zero except for a small splitting introduced by the spin-orbit interaction. In Mg the splitting is of the order of 10^{-3} eV; for this gap and $\epsilon_F \sim 10$ eV the breakdown condition is $\hbar\omega_c > 10^{-5}$ eV, or $B > 1000$ G.

SUMMARY

- A Fermi surface is the surface in $\mathbf{k}$ space of constant energy equal to ϵ_F. The Fermi surface separates filled states from empty states at absolute zero. The form of the Fermi surface is usually exhibited best in the reduced zone scheme, but the connectivity of the surfaces is clearest in the periodic zone scheme.
- An energy band is a single branch of the $\epsilon_{\mathbf{k}}$ versus $\mathbf{k}$ surface.
- The cohesion of simple metals is accounted for by the lowering of energy of the $\mathbf{k} = 0$ conduction band orbital when the boundary conditions on the wavefunction are changed from Schrödinger to Wigner-Seitz.
- The periodicity in the de Haas-van Alphen effect measures the extremal cross-section area S in $\mathbf{k}$ space of the Fermi surface, the cross section being taken perpendicular to $\mathbf{B}$:

(CGS) $$\Delta\left(\frac{1}{B}\right) = \frac{2\pi e}{\hbar c S} .$$

Problems

1. ***Brillouin zones of rectangular lattice.*** Make a plot of the first two Brillouin zones of a primitive rectangular two-dimensional lattice with axes $a, b = 3a$.

2. ***Brillouin zone, rectangular lattice.*** A two-dimensional metal has one atom of valency one in a simple rectangular primitive cell $a = 2$ Å; $b = 4$ Å. (a) Draw the first Brillouin zone. Give its dimensions, in cm^{-1}. (b) Calculate the radius of the free electron Fermi sphere, in cm^{-1}. (c) Draw this sphere to scale on a drawing of the first Brillouin zone. Make another sketch to show the first few periods of the free electron band in the periodic zone scheme, for both the first and second energy bands. Assume there is a small energy gap at the zone boundary.

3. ***Hexagonal close-packed structure.*** Consider the first Brillouin zone of a crystal with a simple hexagonal lattice in three dimensions with lattice constants a and c. Let $\mathbf{G}_c$ denote the shortest reciprocal lattice vector parallel to the c axis of the

crystal lattice. (a) Show that for a hexagonal-close-packed crystal structure the Fourier component $U(\mathbf{G}_c)$ of the crystal potential $U(\mathbf{r})$ is zero. (b) Is $U(2\mathbf{G}_c)$ also zero? (c) Why is it possible in principle to obtain an insulator made up of divalent atoms at the lattice points of a simple hexagonal lattice? (d) Why is it not possible to obtain an insulator made up of monovalent atoms in a hexagonal-close-packed structure?

4. ***Brillouin zones of two-dimensional divalent metal.*** A two-dimensional metal in the form of a square lattice has two conduction electrons per atom. In the almost free electron approximation, sketch carefully the electron and hole energy surfaces. For the electrons choose a zone scheme such that the Fermi surface is shown as closed.

5. ***Open orbits.*** An open orbit in a monovalent tetragonal metal connects opposite faces of the boundary of a Brillouin zone. The faces are separated by $G = 2 \times 10^8$ cm $^{-1}$. A magnetic field $B = 10^3$ gauss $= 10^{-1}$ tesla is normal to the plane of the open orbit. (a) What is the order of magnitude of the period of the motion in $\mathbf{k}$ space? Take $v \approx 10^8$ cm/sec. (b) Describe in real space the motion of an electron on this orbit in the presence of the magnetic field.

6. ***Cohesive energy for a square well potential.*** (a) Find an expression for the binding energy of an electron in one dimension in a single square well of depth U_0 and width a. (This is the standard first problem in elementary quantum mechanics.) Assume that the solution is symmetric about the midpoint of the well. (b) Find a numerical result for the binding energy in terms of U_0 for the special case $|U_0| = 2\hbar^2/ma^2$ and compare with the appropriate limit of Fig. 20. In this limit of widely separated wells the band width goes to zero, so the energy for $k = 0$ is the same as the energy for any other k in the lowest energy band. Other bands are formed from the excited states of the well, in this limit.

7. ***De Haas-van Alphen period of potassium.*** (a) Calculate the period $\Delta(1/B)$ expected for potassium on the free electron model. (b) What is the area in real space of the extremal orbit, for $B = 10$ kG $= 1$ T? The same period applies to oscillations in the electrical resistivity, known as the Shubnikow-de Haas effect.

*8. ***Band edge structure on $\mathbf{k} \cdot \mathbf{p}$ perturbation theory.*** Consider a nondegenerate orbital $\psi_{n\mathbf{k}}$ at $\mathbf{k} = 0$ in the band n of a cubic crystal. Use second-order perturbation theory to find the result

$$\epsilon_n(\mathbf{k}) = \epsilon_n(0) + \frac{\hbar^2k^2}{2m} + \frac{\hbar^2}{m^2}\sum_j{}' \frac{|\langle n0|\mathbf{k}\cdot\mathbf{p}|j0\rangle|^2}{\epsilon_n(0) - \epsilon_j(0)}\,, \tag{39}$$

where the sum is over all other orbitals $\psi_{j\mathbf{k}}$ at $\mathbf{k} = 0$. The effective mass at this point is

$$\frac{m}{m^*} \approx 1 + \frac{2}{m}\sum_j{}' \frac{|\langle n0|\mathbf{p}|j0\rangle|^2}{\epsilon_n(0) - \epsilon_j(0)}\,. \tag{40}$$

*This problem is somewhat difficult.

The mass at the conduction band edge in a narrow gap semiconductor is often dominated by the effect of the valence band edge, whence

$$\frac{m}{m^*} \approx \frac{2}{mE_g} \sum_v |\langle c|\, \mathbf{p}\, |v\rangle|^2 \ , \tag{41}$$

where the sum is over the valence bands; E_g is the energy gap. For given matrix elements, small gaps lead to small masses.

9. ***Wannier functions.*** The Wannier functions of a band are defined in terms of the Bloch functions of the same band by

$$w(\mathbf{r} - \mathbf{r}_n) = N^{-1/2} \sum_{\mathbf{k}} \exp(-i\mathbf{k} \cdot \mathbf{r}_n)\, \psi_{\mathbf{k}}(\mathbf{r}) \ , \tag{42}$$

where $\mathbf{r}_n$ is a lattice point. (a) Prove that Wannier functions about different lattice points n,m are orthogonal:

$$\int dV\, w^*(\mathbf{r} - \mathbf{r}_n) w(\mathbf{r} - \mathbf{r}_m) = 0 \ , \quad n \neq m \ . \tag{43}$$

This orthogonality property makes the functions often of greater use than atomic orbitals centered on different lattice sites, because the latter are not generally orthogonal. (b) The Wannier functions are peaked around the lattice sites. Show that for $\psi_k = N^{-1/2} e^{ikx} u_0(x)$ the Wannier function is

$$w(x - x_n) = u_0(x) \frac{\sin \pi(x - x_n)/a}{\pi(x - x_n)/a} \ ,$$

for N atoms on a line of lattice constant a.

10. ***Open orbits and magnetoresistance.*** We considered the transverse magnetoresistance of free electrons in Problem 6.9 and of electrons and holes in Problem 8.5. In some crystals the magnetoresistance saturates except in special crystal orientations. An open orbit carries current only in a single direction in the plane normal to the magnetic field; such carriers are not deflected by the field. In the arrangement of Fig. 6.14, let the open orbits be parallel to k_x; in real space these orbits carry current parallel to the y axis. Let $\sigma_{yy} = s\sigma_0$ be the conductivity of the open orbits; this defines the constant s. The magnetoconductivity tensor in the high field limit $\omega_c\tau \gg 1$ is

$$\sigma_0 \begin{pmatrix} Q^{-2} & -Q^{-1} & 0 \\ Q^{-1} & s & 0 \\ 0 & 0 & 1 \end{pmatrix} \ ,$$

with $Q = \omega_c\tau$. (a) Show that the Hall field is $E_y = -E_x/sQ$. (b) Show that the effective resistivity in the x direction is $\rho = (Q^2/\sigma_0)(s/s + 1)$, so that the resistivity does not saturate, but increases as B^2.

11. ***Landau levels.*** The vector potential of a uniform magnetic field $B\hat{\mathbf{z}}$ is $\mathbf{A} = -By\hat{\mathbf{x}}$ in the Landau gauge. The hamiltonian of a free electron without spin is

$$H = -(\hbar^2/2m)(\partial^2/\partial y^2 + \partial^2/\partial z^2) + (1/2m)[-i\hbar\partial/\partial x - eyB/c]^2 \ .$$

We will look for an eigenfunction of the wave equation $H\psi = \epsilon\psi$ in the form

$$\psi = \chi(y) \exp[i(k_x x + k_z z)] \ .$$

(a) Show that $\chi(y)$ satisfies the equation

$$(\hbar^2/2m)d^2\chi/dy^2 + [\epsilon - (\hbar^2 k_z^2/2m) - \tfrac{1}{2} m\omega_c^2(y - y_0)^2]\chi = 0 \ ,$$

where $\omega_c = eB/mc$ and $y_0 = c\hbar k_x/eB$. (b) Show that this is the wave equation of a harmonic oscillator with frequency ω_c, where

$$\epsilon_n = (n + \tfrac{1}{2})\hbar\omega_c + \hbar^2 k_z^2/2m \ .$$

10

Superconductivity

NOTATION: In this chapter B_a denotes the applied magnetic field. In the CGS system the critical value B_{ac} of the applied field will be denoted by the symbol H_c in accordance with the custom of workers in superconductivity. Values of B_{ac} are given in gauss in CGS units and in teslas in SI units, with $1\ \mathrm{T} = 10^4\ \mathrm{G}$. In SI we have $B_{ac} = \mu_0 H_c$.

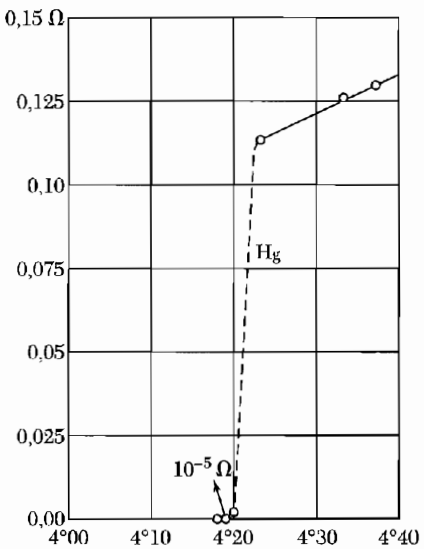


Figure 1 Resistance in ohms of a specimen of mercury versus absolute temperature. This plot by Kamerlingh Onnes marked the discovery of superconductivity.

CHAPTER 10: SUPERCONDUCTIVITY

The electrical resistivity of many metals and alloys drops suddenly to zero when the specimen is cooled to a sufficiently low temperature, often a temperature in the liquid helium range. This phenomenon, called superconductivity, was observed first by Kamerlingh Onnes in Leiden in 1911, three years after he first liquified helium. At a critical temperature T_c the specimen undergoes a phase transition from a state of normal electrical resistivity to a superconducting state, Fig. 1.

Superconductivity is now very well understood. It is a field with many practical and theoretical aspects. The length of this chapter and the relevant appendices reflect the richness and subtleties of the field.

EXPERIMENTAL SURVEY

In the superconducting state the dc electrical resistivity is zero, or so close to zero that persistent electrical currents have been observed to flow without attenuation in superconducting rings for more than a year, until at last the experimentalist wearied of the experiment.

The decay of supercurrents in a solenoid was studied by File and Mills using precision nuclear magnetic resonance methods to measure the magnetic field associated with the supercurrent. They concluded that the decay time of the supercurrent is not less than 100,000 years. We estimate the decay time below. In some superconducting materials, particularly those used for superconducting magnets, finite decay times are observed because of an irreversible redistribution of magnetic flux in the material.

The magnetic properties exhibited by superconductors are as dramatic as their electrical properties. The magnetic properties cannot be accounted for by the assumption that a superconductor is a normal conductor with zero electrical resistivity.

It is an experimental fact that a bulk superconductor in a weak magnetic field will act as a perfect diamagnet, with zero magnetic induction in the interior. When a specimen is placed in a magnetic field and is then cooled through the transition temperature for superconductivity, the magnetic flux originally present is ejected from the specimen. This is called the **Meissner effect**. The sequence of events is shown in Fig. 2. The unique magnetic properties of superconductors are central to the characterization of the superconducting state.

The superconducting state is an ordered state of the conduction electrons of the metal. The order is in the formation of loosely associated pairs of electrons. The electrons are ordered at temperatures below the transition temperature, and they are disordered above the transition temperature.

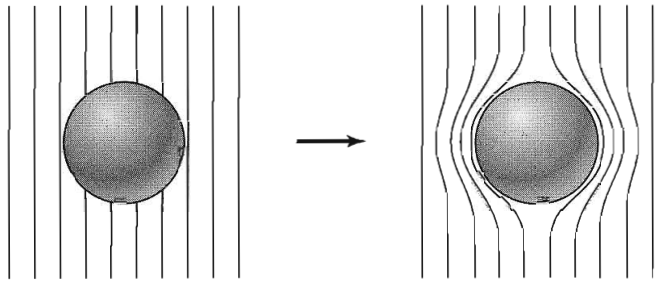


Figure 2 Meissner effect in a superconducting sphere cooled in a constant applied magnetic field; on passing below the transition temperature the lines of induction **B** are ejected from the sphere.

The nature and origin of the ordering was explained by Bardeen, Cooper, and Schrieffer.[1] In the present chapter we develop as far as we can in an elementary way the physics of the superconducting state. We shall also discuss the basic physics of the materials used for superconducting magnets, but not their technology. Appendices H and I give deeper treatments of the superconducting state.

Occurrence of Superconductivity

Superconductivity occurs in many metallic elements of the periodic system and also in alloys, intermetallic compounds, and doped semiconductors. The range of transition temperatures best confirmed at present extends from 90.0 K for the compound $YBa_2Cu_3O_7$ to below 0.001 K for the element Rh. Several *f*-band superconductors, also known as "exotic superconductors," are listed in Chapter 6. Several materials become superconducting only under high pressure; for example, Si has a superconducting form at 165 kbar, with $T_c = 8.3$ K. The elements known to be superconducting are displayed in Table 1, for zero pressure.

Will every nonmagnetic metallic element become a superconductor at sufficiently low temperatures? We do not know. In experimental searches for superconductors with ultralow transition temperatures it is important to eliminate from the specimen even trace quantities of foreign paramagnetic elements, because they can lower the transition temperature severely. One part of Fe in 10^4 will destroy the superconductivity of Mo, which when pure has $T_c = 0.92$ K; and 1 at. percent of gadolinium lowers the transition temperature of lanthanum from 5.6 K to 0.6 K. Nonmagnetic impurities have no very marked effect on the transition temperature. The transition temperatures of a number of interesting superconducting compounds are listed in Table 2. Several organic compounds show superconductivity at fairly low temperatures.

[1]J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Phys. Rev. **106**, 162 (1957); **108**, 1175 (1957).

Table 1 Superconductivity parameters of the elements

An asterisk denotes an element superconducting only in thin films or under high pressure in a crystal modification not normally stable. Data courtesy of B. T. Matthias, revised by T. Geballe.

Li	Be											B	C	N	O	F	Ne
	0.026																
Na	Mg											Al	Si*	P*	S*	Cl	Ar
		Transition temperature in K										1.140					
		Critical magnetic field at absolute zero in gauss (10^{-4} tesla)										105					
K	Ca	Sc	Ti	V	Cr*	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge*	As*	Se*	Br	Kr
			0.39	5.38							0.875	1.091					
			100	1420							53	51					
Rb	Sr	Y*	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn (w)	Sb*	Te*	I	Xe
			0.546	9.50	0.92	7.77	0.51	.0003			0.56	3.4035	3.722				
			47	1980	95	1410	70	.049			30	293	309				
Cs*	Ba*	La fcc	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg (α)	Tl	Pb	Bi*	Po	At	Rn
		6.00	0.12	4.483	0.012	1.4	0.655	0.14			4.153	2.39	7.193				
		1100		830	1.07	198	65	19			412	171	803				
Fr	Ra	Ac															

Ce*	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
													0.1
Th	Pa	U*(α)	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr
1.368	1.4												
1.62													

Table 2 Superconductivity of selected compounds

Compound	T_c, in K	Compound	T_c, in K
Nb_3Sn	18.05	V_3Ga	16.5
Nb_3Ge	23.2	V_3Si	17.1
Nb_3Al	17.5	$YBa_2Cu_3O_{6.9}$	90.0
NbN	16.0	Rb_2CsC_{60}	31.3
C_{60}	19.2	MgB_2	39.0

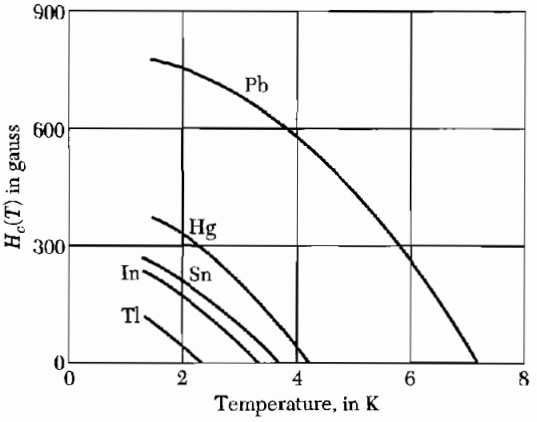


Figure 3 Experimental threshold curves of the critical field $H_c(T)$ versus temperature for several superconductors. A specimen is superconducting below the curve and *normal above the curve.*

Destruction of Superconductivity by Magnetic Fields

A sufficiently strong magnetic field will destroy superconductivity. The threshold or critical value of the applied magnetic field for the destruction of superconductivity is denoted by $H_c(T)$ and is a function of the temperature. At the critical temperature the critical field is zero: $H_c(T_c) = 0$. The variation of the critical field with temperature for several superconducting elements is shown in Fig. 3.

The threshold curves separate the superconducting state in the lower left of the figure from the normal state in the upper right. Note: We should denote the critical value of the applied magnetic field as B_{ac}, but this is not common practice among workers in superconductivity. In the CGS system we shall always understand that $H_c \equiv B_{ac}$, and in the SI we have $H_c \equiv B_{ac}/\mu_0$. The symbol B_a denotes the applied magnetic field.

Meissner Effect

Meissner and Ochsenfeld (1933) found that if a superconductor is cooled in a magnetic field to below the transition temperature, then at the transition the lines of induction B are pushed out (Fig. 2). The Meissner effect shows that a bulk superconductor behaves as if $B = 0$ inside the specimen.

We obtain a particularly useful form of this result if we limit ourselves to long thin specimens with long axes parallel to B_a: now the demagnetizing field contribution (see Chapter 16) to B will be negligible, whence:[2]

$$\text{(CGS)} \qquad B = B_a + 4\pi M = 0 \ ; \qquad \text{or} \qquad \frac{M}{B_a} = -\frac{1}{4\pi} \ ; \tag{1}$$

$$\text{(SI)} \qquad B = B_a + \mu_0 M = 0 \ ; \qquad \text{or} \qquad \frac{M}{B_a} = -\frac{1}{\mu_0} = -\epsilon_0 c^2 \ .$$

The result $B = 0$ cannot be derived from the characterization of a superconductor as a medium of zero resistivity. From Ohm's law, $\mathbf{E} = \rho\mathbf{j}$, we see that if the resistivity ρ goes to zero while $\mathbf{j}$ is held finite, then $\mathbf{E}$ must be zero. By a Maxwell equation $d\mathbf{B}/dt$ is proportional to curl $\mathbf{E}$, so that zero resistivity implies $d\mathbf{B}/dt = 0$, but not $\mathbf{B} = 0$. This argument is not entirely transparent, but the result predicts that the flux through the metal cannot change on cooling through the transition. The Meissner effect suggests that perfect diamagnetism is an essential property of the superconducting state.

We expect another difference between a superconductor and a perfect conductor, defined as a conductor in which the electrons have an infinite mean free path. When the problem is solved in detail, it turns out that a perfect conductor when placed in a magnetic field cannot produce a permanent eddy current screen: the field will penetrate about 1 cm in an hour.[3]

The magnetization curve expected for a superconductor under the conditions of the Meissner-Ochsenfeld experiment is sketched in Fig. 4a. This applies quantitatively to a specimen in the form of a long solid cylinder placed in a longitudinal magnetic field. Pure specimens of many materials exhibit this behavior; they are called **type I superconductors** or, formerly, soft superconductors. The values of H_c are always too low for type I superconductors to have application in coils for superconducting magnets.

Other materials exhibit a magnetization curve of the form of Fig. 4b and are known as **type II superconductors**. They tend to be alloys (as in Fig. 5a) or transition metals with high values of the electrical resistivity in the normal state: that is, the electronic mean free path in the normal state is short. We shall see later why the mean free path is involved in the "magnetization" of superconductors.

Type II superconductors have superconducting electrical properties up to a field denoted by H_{c2}. Between the lower critical field H_{c1} and the upper critical field H_{c2} the flux density $B \neq 0$ and the Meissner effect is said to be incomplete. The value of H_{c2} may be 100 times or more higher (Fig. 5b) than

[2]Diamagnetism, the magnetization M, and the magnetic susceptibility are defined in Chapter 14. The magnitude of the apparent diamagnetic susceptibility of bulk superconductors is very much larger than in typical diamagnetic substances. In (1), M is the magnetization equivalent to the superconducting currents in the specimen.

[3]A. B. Pippard, *Dynamics of conduction electrons*, Gordon and Breach, 1965.

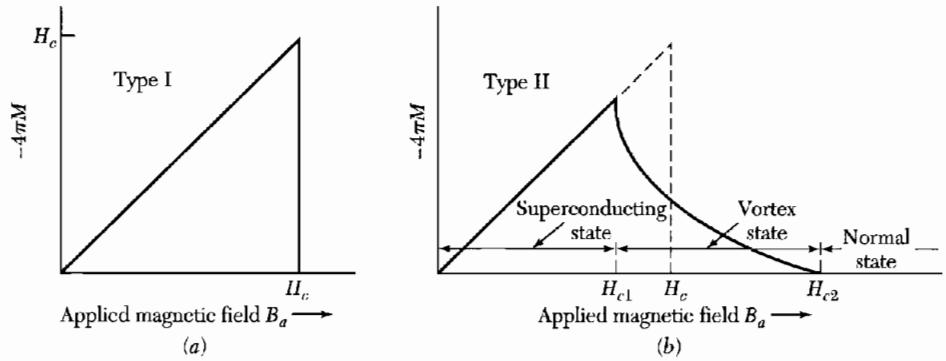


Figure 4 (a) Magnetization versus applied magnetic field for a bulk superconductor exhibiting a complete Meissner effect (perfect diamagnetism). A superconductor with this behavior is called a type I superconductor. Above the critical field H_c the specimen is a normal conductor and the magnetization is too small to be seen on this scale. Note that minus $4\pi M$ is plotted on the vertical scale: the negative value of M corresponds to diamagnetism. (b) Superconducting magnetization curve of a type II superconductor. The flux starts to penetrate the specimen at a field H_{c1} lower than the thermodynamic critical field H_c. The specimen is in a vortex state between H_{c1} and H_{c2}, and it has superconducting electrical properties up to H_{c2}. Above H_{c2} the specimen is a normal conductor in every respect, except for possible surface effects. For given H_c the area under the magnetization curve is the same for a type II superconductor as for a type I. (CGS units in all parts of this figure.)

the value of the critical field H_c calculated from the thermodynamics of the transition. In the region between H_{c1} and H_{c2} the superconductor is threaded by flux lines and is said to be in the **vortex state**. A field H_{c2} of 410 kG (41 teslas) has been attained in an alloy of Nb, Al, and Ge at the boiling point of helium, and 540 kG (54 teslas) has been reported for $PbMo_6S_8$.

Commercial solenoids wound with a hard superconductor produce high steady fields over 100 kG. A "hard superconductor" is a type II superconductor with a large magnetic hysteresis, usually induced by mechanical treatment. Such materials have an important medical application in magnetic resonance imaging (MRI).

Heat Capacity

In all superconductors the entropy decreases markedly on cooling below the critical temperature T_c. Measurements for aluminum are plotted in Fig. 6. The decrease in entropy between the normal state and the superconducting state tells us that the superconducting state is more ordered than the normal state, for the entropy is a measure of the disorder of a system. Some or all of the electrons thermally excited in the normal state are ordered in the superconducting state. The change in entropy is small, in aluminum of the order of 10^{-4} k_B per atom. The small entropy change must mean that only a small fraction (of the order of 10^{-4}) of the conduction electrons participate in the transition to the ordered superconducting state. The free energies of normal and superconducting states are compared in Fig. 7.

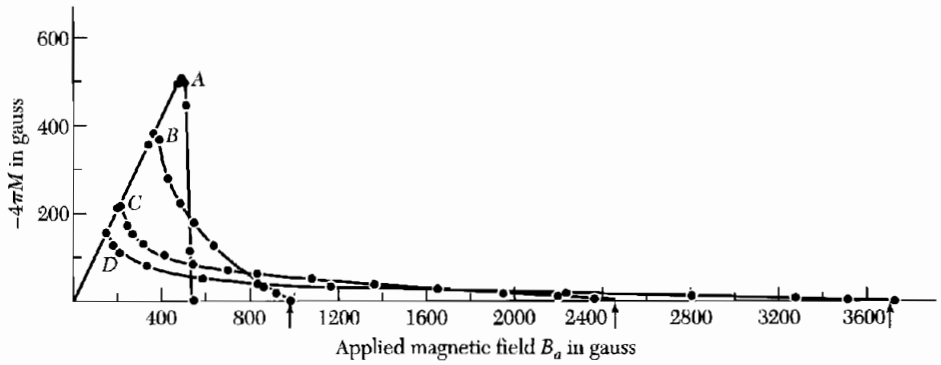


Figure 5a Superconducting magnetization curves of annealed polycrystalline lead and lead-indium alloys at 4.2 K. (*A*) lead; (*B*) lead–2.08 wt. percent indium; (*C*) lead–8.23 wt. percent indium; (*D*) lead–20.4 wt. percent indium. (After Livingston.)

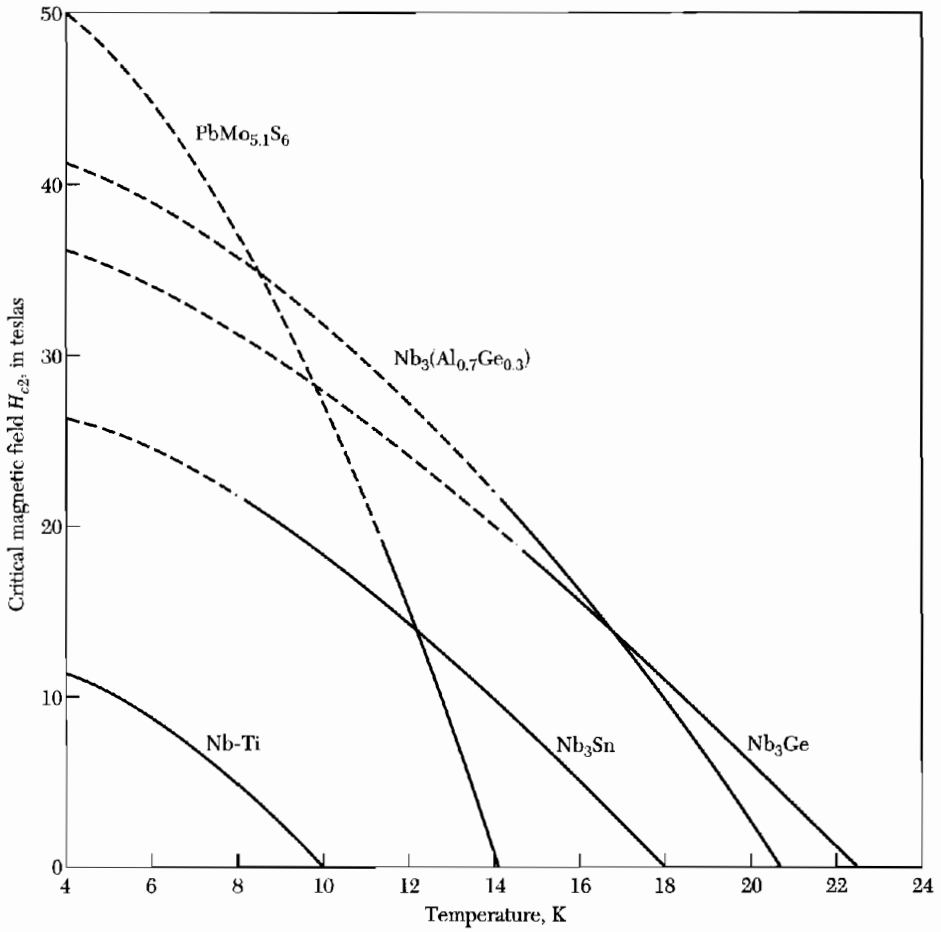


Figure 5b Strong magnetic fields are within the capability of certain Type II materials.

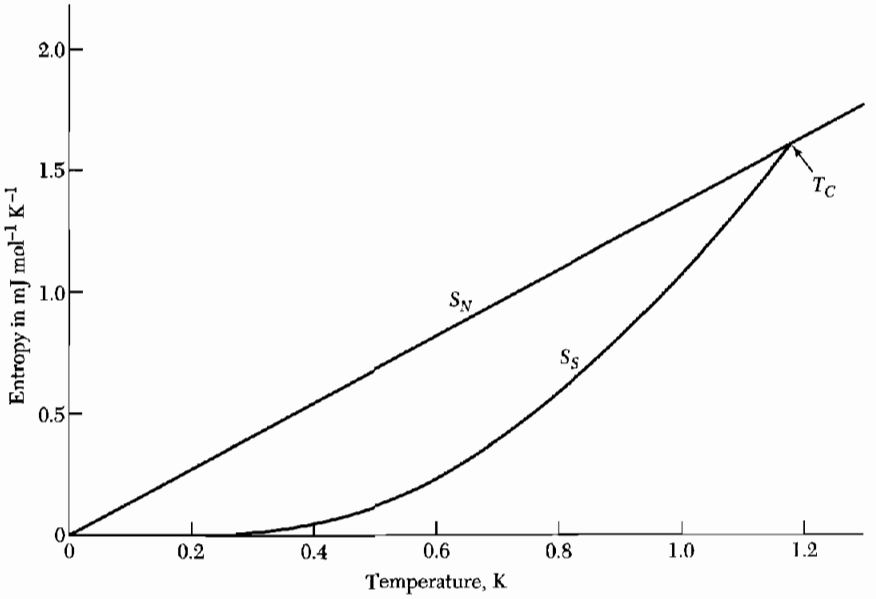


Figure 6 Entropy S of aluminum in the normal and superconducting states as a function of the temperature. The entropy is lower in the superconducting state because the electrons are more ordered here than in the normal state. At any temperature below the critical temperature T_c the specimen can be put in the normal state by application of a magnetic field stronger than the critical field.

The heat capacity of gallium is plotted in Fig. 8: (a) compares the normal and superconducting states; (b) shows that the electronic contribution to the heat capacity in the superconducting state is an exponential form with an argument proportional to $-1/T$, suggestive of excitation of electrons across an energy gap. An energy gap (Fig. 9) is a characteristic, but not universal, feature of the superconducting state. The gap is accounted for by the Bardeen-Cooper-Schrieffer (BCS) theory of superconductivity (see Appendix H).

Energy Gap

The energy gap of superconductors is of entirely different origin and nature than the energy gap of insulators. In an insulator the energy gap is caused by the electron-lattice interaction, Chapter 7. This interaction ties the electrons to the lattice. In a superconductor the important interaction is the electron-electron interaction which orders the electrons in **k** space with respect to the Fermi gas of electrons.

The argument of the exponential factor in the electronic heat capacity of a superconductor is found to be $-E_g/2k_BT$ and not $-E_g/k_BT$. This has been learnt from comparison with optical and electron tunneling determinations of the gap E_g. Values of the gap in several superconductors are given in Table 3.

The transition in zero magnetic field from the superconducting state to the normal state is observed to be a second-order phase transition. At a

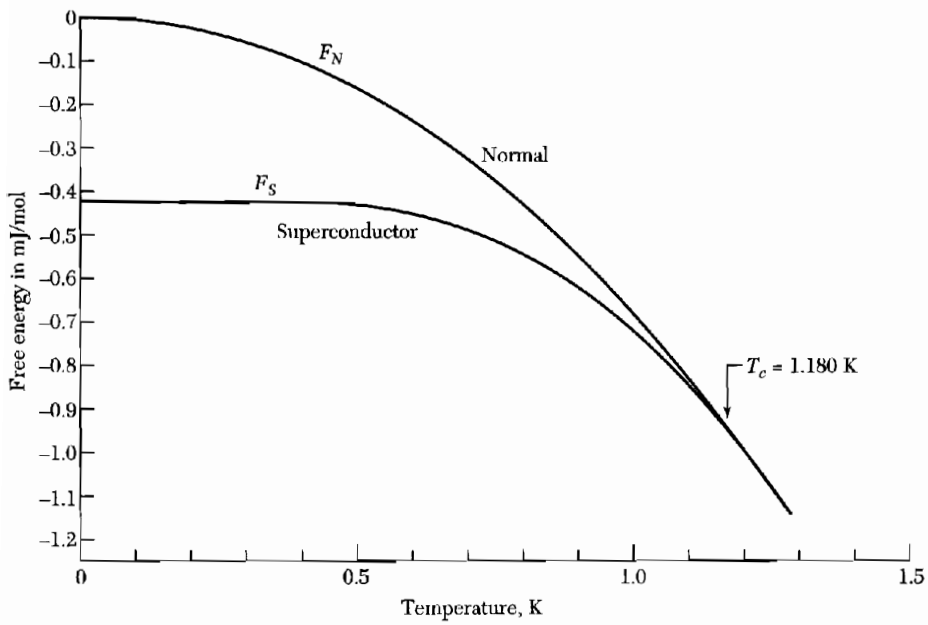


Figure 7 Experimental values of the free energy as a function of temperature for aluminum in the superconducting state and in the normal state. Below the transition temperature $T_c = 1.180$ K the free energy is lower in the superconducting state. The two curves merge at the transition temperature, so that the phase transition is second order (there is no latent heat of transition at T_c). The curve F_S is measured in zero magnetic field, and F_N is measured in a magnetic field sufficient to put the specimen in the normal state. (Courtesy of N. E. Phillips.)

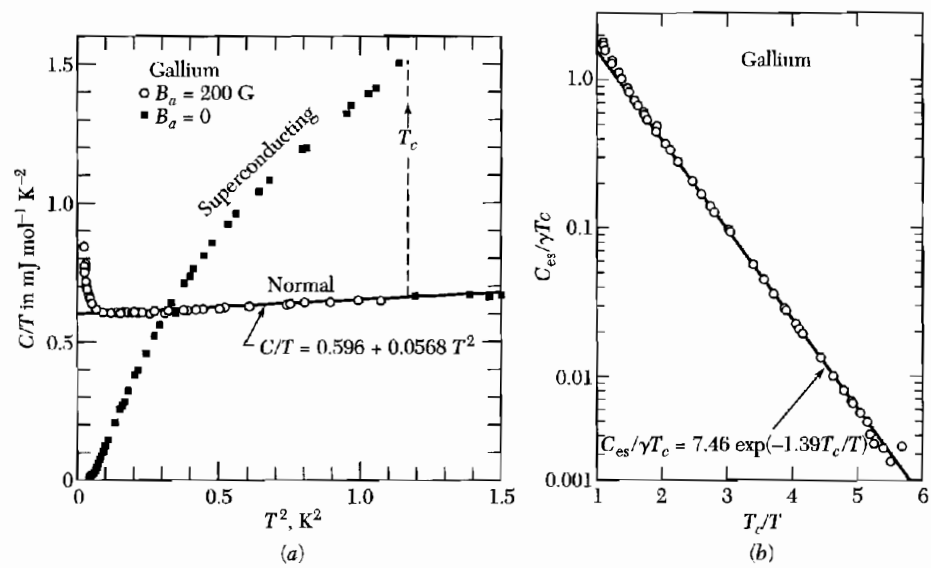


Figure 8 (a) The heat capacity of gallium in the normal and superconducting states. The normal state (which is restored by a 200 G field) has electronic, lattice, and (at low temperatures) nuclear quadrupole contributions. In (b) the electronic part C_{es} of the heat capacity in the superconducting state is plotted on a log scale versus T_c/T: the exponential dependence on $1/T$ is evident. Here $\gamma = 0.60$ mJ mol^{-1} deg^{-2}. (After N. E. Phillips.)

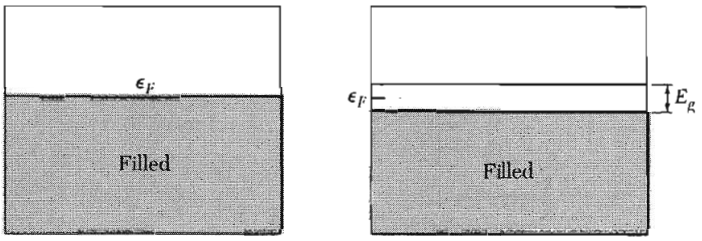


Figure 9 (a) Conduction band in the normal state: (b) energy gap at the Fermi level in the superconducting state. Electrons in excited states above the gap behave as normal electrons in rf fields: they cause resistance; at dc they are shorted out by the superconducting electrons. The gap E_g is exaggerated in the figure: typically $E_g \sim 10^{-4}\,\epsilon_F$.

Table 3 Energy gaps in superconductors, at $T = 0$

$E_g(0)$ in 10^{-4} eV.
$E_g(0)/k_B T_c$.

Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge
		16.							2.4	3.3	
		3.4							3.2	3.5	
Y	**Zr**	**Nb**	**Mo**	**Tc**	**Ru**	**Rh**	**Pd**	**Ag**	**Cd**	**In**	**Sn** (w)
		30.5	2.7						1.5	10.5	11.5
		3.80	3.4						3.2	3.6	3.5
La fcc	**Hf**	**Ta**	**W**	**Re**	**Os**	**Ir**	**Pt**	**Au**	**Hg** (α)	**Tl**	**Pb**
19.		14.							16.5	7.35	27.3
3.7		3.60							4.6	3.57	4.38

Al	Si
3.4	
3.3	

second-order transition there is no latent heat, but there is a discontinuity in the heat capacity, evident in Fig. 8a. Further, the energy gap decreases continuously to zero as the temperature is increased to the transition temperature T_c, as in Fig. 10. A first-order transition would be characterized by a latent heat and by a discontinuity in the energy gap.

Microwave and Infrared Properties

The existence of an energy gap means that photons of energy less than the gap energy are not absorbed. Nearly all the photons incident are reflected as for any metal because of the impedance mismatch at the boundary between vacuum and metal, but for a very thin (~20 Å) film more photons are transmitted in the superconducting state than in the normal state.

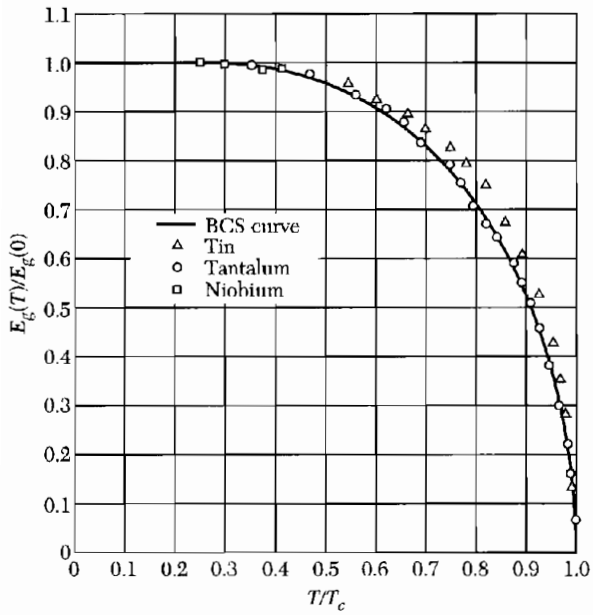


Figure 10 Reduced values of the observed energy gap $E_g(T)/E_g(0)$ as a function of the reduced temperature T/T_c, after Townsend and Sutton. The solid curve is drawn for the BCS theory.

For photon energies less than the energy gap, the resistivity of a superconductor vanishes at absolute zero. At $T \ll T_c$ the resistance in the superconducting state has a sharp threshold at the gap energy. Photons of lower energy see a resistanceless surface. Photons of higher energy than the energy gap see a resistance that approaches that of the normal state because such photons cause transitions to unoccupied normal energy levels above the gap.

As the temperature is increased not only does the gap decrease in energy, but the resistivity for photons with energy below the gap energy no longer vanishes, except at zero frequency. At zero frequency the superconducting electrons short-circuit any normal electrons that have been thermally excited above the gap. At finite frequencies the inertia of the superconducting electrons prevents them from completely screening the electric field, so that thermally excited normal electrons now can absorb energy (Problem 3).

Isotope Effect

It has been observed that the critical temperature of superconductors varies with isotopic mass. In mercury T_c varies from 4.185 K to 4.146 K as the average atomic mass M varies from 199.5 to 203.4 atomic mass units. The transition temperature changes smoothly when we mix different isotopes of the same element. The experimental results within each series of isotopes may be fitted by a relation of the form

$$M^{\alpha}T_c = \text{constant} \ . \tag{2}$$

Observed values of α are given in Table 4.

Table 4 Isotope effect in superconductors

Experimental values of α in $M^{\alpha}T_c$ = constant, where M is the isotopic mass.

Substance	α	Substance	α
Zn	0.45 ± 0.05	Ru	0.00 ± 0.05
Cd	0.32 ± 0.07	Os	0.15 ± 0.05
Sn	0.47 ± 0.02	Mo	0.33
Hg	0.50 ± 0.03	Nb_3Sn	0.08 ± 0.02
Pb	0.49 ± 0.02	Zr	0.00 ± 0.05

From the dependence of T_c on the isotopic mass we learn that lattice vibrations and hence electron-lattice interactions are deeply involved in superconductivity. This was a fundamental discovery: there is no other reason for the superconducting transition temperature to depend on the number of neutrons in the nucleus.

The original BCS model gave the result $T_c \propto \theta_{\text{Debye}} \propto M^{-1/2}$, so that $\alpha = \frac{1}{2}$ in (2), but the inclusion of coulomb interactions between the electrons changes the relation. Nothing is sacred about $\alpha = \frac{1}{2}$. The absence of an isotope effect in Ru and Zr has been accounted for in terms of the electron band structure of these metals.

THEORETICAL SURVEY

A theoretical understanding of the phenomena associated with superconductivity has been reached in several ways. Certain results follow directly from thermodynamics. Many important results can be described by phenomenological equations: the London equations and the Landau-Ginzburg equations (Appendix 1). A successful quantum theory of superconductivity was given by Bardeen, Cooper, and Schrieffer, and has provided the basis for subsequent work. Josephson and Anderson discovered the importance of the phase of the superconducting wavefunction.

Thermodynamics of the Superconducting Transition

The transition between the normal and superconducting states is thermodynamically reversible, just as the transition between liquid and vapor phases of a substance is reversible. Thus we may apply thermodynamics to the transition, and we thereby obtain an expression for the entropy difference between normal and superconducting states in terms of the critical field curve H_c versus T. This is analogous to the vapor pressure equation for the liquid-gas coexistence curve (*TP*, Chapter 10).

We treat a type **I** superconductor with a complete Meissner effect. so that $B = 0$ inside the superconductor. We shall see that the critical field H_c is a quantitative measure of the free energy difference between the superconducting and normal states at constant temperature. The symbol H_c will always refer to a bulk specimen, never to a thin film. For type II superconductors, H_c is understood to be the thermodynamic critical field related to the stabilization free energy.

The stabilization free energy of the superconducting state with respect to the normal state can be determined by calorimetric or magnetic measurements. In the calorimetric method the heat capacity is measured as a function of temperature for the superconductor and for the normal conductor, which means the superconductor in a magnetic field larger than H_c. From the difference of the heat capacities we can compute the free energy difference, which is the stabilization free energy of the superconducting state.

In the magnetic method the stabilization free energy is found from the value of the applied magnetic field that will destroy the superconducting state, at constant temperature. The argument follows. Consider the work done (Fig. 11) on a superconductor when it is brought reversibly at constant temperature from a position at infinity (where the applied field is zero) to a position **r** in the field of a permanent magnet:

$$W = -\int_0^{B_a} \mathbf{M} \cdot d\mathbf{B}_a \ , \tag{3}$$

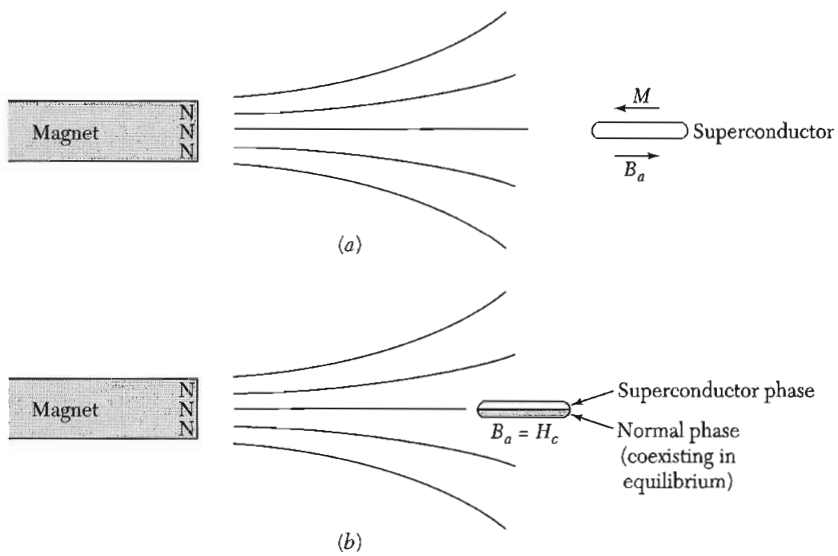


Figure 11 (a) A superconductor in which the Meissner effect is complete has $B = 0$, as if the magnetization were $M = -B_a/4\pi$, in CGS units. (b) When the applied field reaches the value B_{ac}, the normal state can coexist in equilibrium with the superconducting state. In coexistence the free energy densities are equal: $F_N(T, B_{ac}) = F_S(T, B_{ac})$.

per unit volume of specimen. This work appears in the energy of the magnetic field. The thermodynamic identity for the process is

$$dF = -\mathbf{M} \cdot d\mathbf{B}_a \ , \tag{4}$$

as in *TP*, Chapter 8.

For a superconductor with $\mathbf{M}$ related to $\mathbf{B}_a$ by (1) we have

(CGS) $$dF_S = \frac{1}{4\pi} B_a \, dB_a \ ; \tag{5}$$

(SI) $$dF_S = \frac{1}{\mu_0} B_a \, dB_a \ .$$

The increase in the free energy density of the superconductor is

(CGS) $$F_S(B_a) - F_S(0) = B_a^2/8\pi \ ; \tag{6}$$

(SI) $$F_s(B_a) - F_s(0) - B_a^2/2\mu_0 \ ,$$

on being brought from a position where the applied field is zero to a position where the applied field is B_a.

Now consider a normal nonmagnetic metal. If we neglect the small susceptibility[4] of a metal in the normal state, then $M = 0$ and the energy of the normal metal is independent of field. At the critical field we have

$$F_N(B_{ac}) = F_N(0) \ . \tag{7}$$

The results (6) and (7) are all we need to determine the stabilization energy of the superconducting state at absolute zero. At the critical value B_{ac} of the applied magnetic field the energies are equal in the normal and superconducting states:

(CGS) $$F_N(B_{ac}) = F_S(B_{ac}) = F_S(0) + B_{ac}^2/8\pi \ , \tag{8}$$

(SI) $$F_N(B_{ac}) = F_S(B_{ac}) = F_S(0) + B_{ac}^2/2\mu_0 \ ,$$

In SI units $H_c \equiv B_{ac}/\mu_0$, whereas in CGS units $H_c \equiv B_{ac}$.

The specimen is stable in either state when the applied field is equal to the critical field. Now by (7) it follows that

(CGS) $$\Delta F \equiv F_N(0) - F_S(0) = B_{ac}^2/8\pi \ , \tag{9}$$

[4]This is an adequate assumption for type I superconductors. In type II superconductors in high fields the change in spin paramagnetism of the conduction electrons lowers the energy of the normal phase significantly. In some, but not all, type II superconductors the upper critical field is limited by this effect. Clogston has suggested that $H_{c2}(\max) = 18{,}400\ T_c$, where H_{c2} is in gauss and T_c in K.

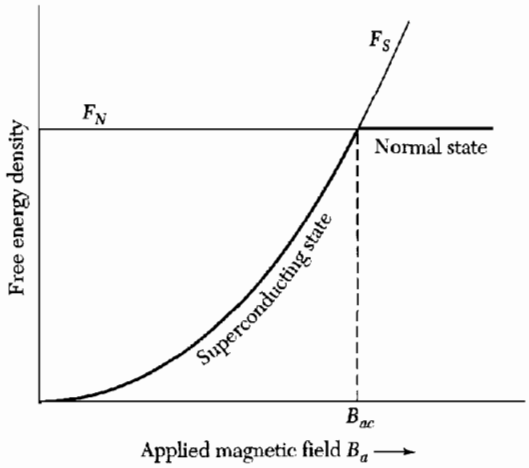


Figure 12 The free energy density F_N of a nonmagnetic normal metal is approximately independent of the intensity of the applied magnetic field B_a. At a temperature $T < T_c$ the metal is a superconductor in zero magnetic field, so that $F_S(T, 0)$ is lower than $F_N(T, 0)$. An applied magnetic field increases F, by $B_a^2/8\pi$, in CGS units, so that $F_S(T, B_a) = F_S(T, 0) + B_a^2/8\pi$. If B_a is larger than the critical field B_{ac} the free energy density is lower in the normal state than in the superconducting state, and now the normal state is the stable state. The origin of the vertical scale in the drawing is at $F_S(T, 0)$. The figure equally applies to U_S and U_N at $T = 0$.

where ΔF is the stabilization free energy density of the superconducting state. For aluminum, B_{ac} at absolute zero is 105 gauss, so that at absolute zero $\Delta F = (105)^2/8\pi = 439$ erg cm ^{3}, in excellent agreement with the result of thermal measurements, 430 erg cm^{-3}.

At a finite temperature the normal and superconducting phases are in equilibrium when the magnetic field is such that their free energies $F = U - TS$ are equal. The free energies of the two phases are sketched in Fig. 12 as a function of the magnetic field. Experimental curves of the free energies of the two phases for aluminum are shown in Fig. 7. Because the slopes dF/dT are equal at the transition temperature, there is no latent heat at T_c.

London Equation

We saw that the Meissner effect implies a magnetic susceptibility $\chi = -1/4\pi$ in CGS in the superconducting state or, in SI, $\chi = -1$. Can we modify a constitutive equation of electrodynamics (such as Ohm's law) in some way to obtain the Meissner effect? We do not want to modify the Maxwell equations themselves. Electrical conduction in the normal state of a metal is described by Ohm's law $\mathbf{j} = \sigma\mathbf{E}$. We need to modify this drastically to describe conduction and the Meissner effect in the superconducting state. Let us make a *postulate* and see what happens.

We postulate that in the superconducting state the current density is directly proportional to the vector potential $\mathbf{A}$ of the local magnetic field, where $\mathbf{B} = \text{curl } \mathbf{A}$. The gauge of $\mathbf{A}$ will be specified. In CGS units we write the constant of proportionality as $-c/4\pi\lambda_L^2$ for reasons that will become clear.

Here c is the speed of light and λ_L is a constant with the dimensions of length. In SI units we write $-1/\mu_0\lambda_L^2$. Thus

$$\text{(CGS)} \quad \mathbf{j} = -\frac{c}{4\pi\lambda_L^2}\mathbf{A} \; ; \qquad \text{(SI)} \quad \mathbf{j} = -\frac{1}{\mu_0\lambda_L^2}\mathbf{A} \; . \tag{10}$$

This is the London equation. We express it another way by taking the curl of both sides to obtain

$$\text{(CGS)} \quad \operatorname{curl}\mathbf{j} = -\frac{c}{4\pi\lambda_L^2}\mathbf{B} \; ; \qquad \text{(SI)} \quad \operatorname{curl}\mathbf{j} = -\frac{1}{\mu_0\lambda_L^2}\mathbf{B} \; . \tag{11}$$

The London equation (10) is understood to be written with the vector potential in the London gauge in which div $\mathbf{A} = 0$, and $\mathbf{A}_n = 0$ on any external surface through which no external current is fed. The subscript n denotes the component normal to the surface. Thus div $\mathbf{j} = 0$ and $\mathbf{j}_n = 0$, the actual physical boundary conditions. The form (10) applies to a simply connected superconductor; additional terms may be present in a ring or cylinder, but (11) holds true independent of geometry.

First we show that the London equation leads to the Meissner effect. By a Maxwell equation we know that

$$\text{(CGS)} \quad \operatorname{curl}\mathbf{B} = \frac{4\pi}{c}\mathbf{j} \; ; \qquad \text{(SI)} \quad \operatorname{curl}\mathbf{B} = \mu_0\mathbf{j} \; ; \tag{12}$$

under static conditions. We take the curl of both sides to obtain

$$\text{(CGS)} \qquad \operatorname{curl}\operatorname{curl}\mathbf{B} = -\nabla^2\mathbf{B} = \frac{4\pi}{c}\operatorname{curl}\mathbf{j} \; ;$$

$$\text{(SI)} \qquad \operatorname{curl}\operatorname{curl}\mathbf{B} = -\nabla^2\mathbf{B} = \mu_0 \operatorname{curl}\mathbf{j} \; ;$$

which may be combined with the London equation (11) to give for a superconductor

$$\nabla^2\mathbf{B} = \mathbf{B}/\lambda_L^2 \; . \tag{13}$$

This equation is seen to account for the Meissner effect because it does not allow a solution uniform in space, so that a uniform magnetic field cannot exist in a superconductor. That is, $\mathbf{B}(\mathbf{r}) = \mathbf{B}_0 =$ constant is not a solution of (13) unless the constant field $\mathbf{B}_0$ is identically zero. The result follows because $\nabla^2\mathbf{B}_0$ is always zero, but $\mathbf{B}_0/\lambda_L^2$ is not zero unless $\mathbf{B}_0$ is zero. Note further that (12) ensures that $\mathbf{j} = 0$ in a region where $\mathbf{B} = 0$.

In the pure superconducting state the only field allowed is exponentially damped as we go in from an external surface. Let a semi-infinite superconductor

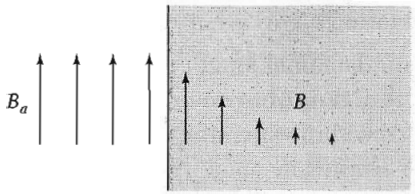


Figure 13 Penetration of an applied magnetic field into a semi-infinite superconductor. The penetration depth λ is defined as the distance in which the field decreases by the factor e^{-1}. Typically, $\lambda \approx 500$ Å in a pure superconductor.

occupy the space on the positive side of the x axis, as in Fig. 13. If $B(0)$ is the field at the plane boundary, then the field inside is

$$B(x) = B(0) \exp(-x/\lambda_L) \ , \tag{14}$$

for this is a solution of (13). In this example the magnetic field is assumed to be parallel to the boundary. Thus we see λ_L measures the depth of penetration of the magnetic field; it is known as the **London penetration depth**. Actual penetration depths are not described precisely by λ_L alone, for the London equation is now known to be somewhat oversimplified. It is shown by comparison of (22) with (11) that

$$\text{(CGS)} \quad \lambda_L = (mc^2/4\pi nq^2)^{1/2} \ ; \qquad \text{(SI)} \quad \lambda_L = (\epsilon_0 mc^2/nq^2)^{1/2} \tag{14a}$$

for particles of charge q and mass m in concentration n. Values are given in Table 5.

An applied magnetic field B_a will penetrate a thin film fairly uniformly if the thickness is much less than λ_L; thus in a thin film the Meissner effect is not complete. In a thin film the induced field is much less than B_a, and there is little effect of B_a on the energy density of the superconducting state, so that (6) does not apply. It follows that the critical field H_c of thin films in parallel magnetic fields will be very high.

Table 5 Calculated intrinsic coherence length and London penetration depth, at absolute zero

Metal	Intrinsic Pippard coherence length ξ_0, in 10^{-6} cm	London penetration depth λ_L, in 10^{-6} cm	λ_L/ξ_0
Sn	23.	3.4	0.16
Al	160.	1.6	0.010
Pb	8.3	3.7	0.45
Cd	76.	11.0	0.14
Nb	3.8	3.9	1.02

After R. Meservey and B. B. Schwartz.

Coherence Length

The London penetration depth λ_L is a fundamental length that characterizes a superconductor. An independent length is the **coherence length** ξ. The coherence length is a measure of the distance within which the superconducting electron concentration cannot change drastically in a spatially-varying magnetic field.

The London equation is a *local* equation: it relates the current density at a point $\mathbf{r}$ to the vector potential at the same point. So long as $\mathbf{j}(\mathbf{r})$ is given as a constant time $\mathbf{A}(\mathbf{r})$, the current is required to follow exactly any variation in the vector potential. But the coherence length ξ is a measure of the range over which we should average $\mathbf{A}$ to obtain $\mathbf{j}$. It is also a measure of the minimum spatial extent of a transition layer between normal and superconductor. The coherence length is best introduced into the theory through the Landau-Ginzburg equations, Appendix I. Now we give a plausibility argument for the energy required to modulate the superconducting electron concentration.

Any spatial variation in the state of an electronic system requires extra kinetic energy. A modulation of an eigenfunction increases the kinetic energy because the modulation will increase the integral of $d^2\varphi/dx^2$. It is reasonable to restrict the spatial variation of $\mathbf{j}(\mathbf{r})$ in such a way that the extra energy is less than the stabilization energy of the superconducting state.

We compare the plane wave $\psi(x) = e^{ikx}$ with the strongly modulated wavefunction:

$$\varphi(x) = 2^{-1/2}\,(e^{i(k+q)x} + e^{ikx}) \ . \tag{15a}$$

The probability density associated with the plane wave is uniform in space: $\psi^*\psi = e^{-ikx}\,e^{ikx} = 1$, whereas $\varphi^*\varphi$ is modulated with the wavevector q:

$$\begin{aligned}\varphi^*\varphi &= \tfrac{1}{2}(e^{-i(k+q)x} + e^{-ikx})(e^{i(k+q)x} + e^{ikx}) \\ &= \tfrac{1}{2}(2 + e^{iqx} + e^{-iqx}) = 1 + \cos qx \ .\end{aligned} \tag{15b}$$

The kinetic energy of the wave $\psi(x)$ is $\epsilon = \hbar^2k^2/2m$; the kinetic energy of the modulated density distribution is higher, for

$$\int dx\ \varphi^*\left(-\frac{\hbar^2}{2m}\frac{d^2}{dx^2}\right)\varphi = \frac{1}{2}\left(\frac{\hbar^2}{2m}\right)[(k+q)^2 + k^2] \cong \frac{\hbar^2}{2m}k^2 + \frac{\hbar^2}{2m}kq \ ,$$

where we neglect q^2 for $q \ll k$.

The increase of energy required to modulate is $\hbar^2kq/2m$. If this increase exceeds the energy gap E_g, superconductivity will be destroyed. The critical value q_0 of the modulation wavevector is given by

$$\frac{\hbar^2}{2m}k_Fq_0 = E_g \ . \tag{16a}$$

We define an **intrinsic coherence length** ξ_0 related to the critical modulation by $\xi_0 = 1/q_0$. We have

$$\xi_0 = \hbar^2 k_F / 2mE_g = \hbar v_F / 2E_g \ , \tag{16b}$$

where v_F is the electron velocity at the Fermi surface. On the BCS theory a similar result is found:

$$\boxed{\xi_0 = 2\hbar v_F / \pi E_g \ .} \tag{17}$$

Calculated values of ξ_0 from (17) are given in Table 5. The intrinsic coherence length ξ_0 is characteristic of a pure superconductor.

In impure materials and in alloys the coherence length ξ is shorter than ξ_0. This may be understood qualitatively: in impure material the electron eigenfunctions already have wiggles in them: we can construct a given localized variation of current density with less energy from wavefunctions with wiggles than from smooth wavefunctions.

The coherence length first appeared in the Landau-Ginzburg equations; these equations also follow from the BCS theory. They describe the structure of the transition layer between normal and superconducting phases in contact. The coherence length and the actual penetration depth λ depend on the mean free path ℓ of the electrons measured in the normal state; the relationships are indicated in Fig. 14. When the superconductor is very impure, with a very small ℓ, then $\xi \approx (\xi_0 \ell)^{1/2}$ and $\lambda \approx \lambda_L \, (\xi_0/\ell)^{1/2}$, so that $\lambda/\xi \approx \lambda_L/\ell$. This is the "dirty superconductor" limit. The ratio λ/ξ is denoted by κ.

BCS Theory of Superconductivity

The basis of a quantum theory of superconductivity was laid by the classic 1957 papers of Bardeen, Cooper, and Schrieffer. There is a "BCS theory of superconductivity" with a very wide range of applicability, from He^3 atoms in their condensed phase, to type I and type II metallic superconductors, and to high-temperature superconductors based on planes of cuprate ions. Further,

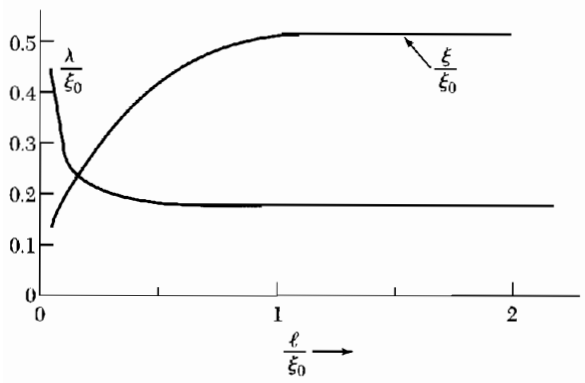


Figure 14 Penetration depth λ and the coherence length ξ as functions of the mean free path ℓ of the conduction electrons in the normal state. All lengths are in units of ξ_0, the intrinsic coherence length. The curves are sketched for $\xi_0 = 10\lambda_L$. For short mean free paths the coherence length becomes shorter and the penetration depth becomes longer. The increase in the ratio $\kappa\lambda/\xi$ favors type II superconductivity.

there is a "BCS wavefunction" composed of particle pairs $\mathbf{k}\uparrow$ and $-\mathbf{k}\downarrow$, which, when treated by the BCS theory, gives the familiar electronic superconductivity observed in metals and exhibits the energy gaps of Table 3. This pairing is known as s-wave pairing. There are other forms of particle pairing possible with the BCS theory, but we do not have to consider other than the BCS wavefunction here. In this chapter we treat the specific accomplishments of BCS theory with a BCS wavefunction, which include:

1. An attractive interaction between electrons can lead to a ground state separated from excited states by an energy gap. The critical field, the thermal properties, and most of the electromagnetic properties are consequences of the energy gap.

2. The electron-lattice-electron interaction leads to an energy gap of the observed magnitude. The indirect interaction proceeds when one electron interacts with the lattice and deforms it; a second electron sees the deformed lattice and adjusts itself to take advantage of the deformation to lower its energy. Thus the second electron interacts with the first electron via the lattice deformation.

3. The penetration depth and the coherence length emerge as natural consequences of the BCS theory. The London equation is obtained for magnetic fields that vary slowly in space. Thus the central phenomenon in superconductivity, the Meissner effect, is obtained in a natural way.

4. The criterion for the transition temperature of an element or alloy involves the electron density of orbitals $D(\epsilon_F)$ of one spin at the Fermi level and the electron-lattice interaction U, which can be estimated from the electrical resistivity because the resistivity at room temperature is a measure of the electron-phonon interaction. For $UD(\epsilon_F) \ll 1$ the BCS theory predicts

$$T_c = 1.14\theta \exp[-1/UD(\epsilon_F)] \ , \tag{18}$$

where θ is the Debye temperature and U is an attractive interaction. The result for T_c is satisfied at least qualitatively by the experimental data. There is an interesting apparent paradox: the higher the resistivity at room temperature the higher is U, and thus the more likely it is that the metal will be a superconductor when cooled.

5. Magnetic flux through a superconducting ring is quantized and the effective unit of charge is $2e$ rather than e. The BCS ground state involves pairs of electrons; thus flux quantization in terms of the pair charge $2e$ is a consequence of the theory.

BCS Ground State

The filled Fermi sea is the ground state of a Fermi gas of noninteracting electrons. This state allows arbitrarily small excitations—we can form an

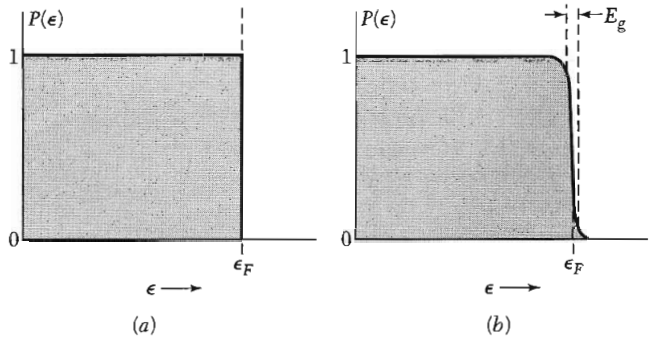


Figure 15 (a) Probability P that an orbital of kinetic energy ϵ is occupied in the ground state of the noninteracting Fermi gas; (b) the BCS ground state differs from the Fermi state in a region of width of the order of the energy gap E_g. Both curves are for absolute zero.

excited state by taking an electron from the Fermi surface and raising it just above the Fermi surface. The BCS theory shows that with an appropriate attractive interaction between electrons the new ground state is superconducting and is separated by a finite energy E_g from its lowest excited state.

The formation of the BCS ground state is suggested by Fig. 15. The BCS state in (b) contains admixtures of one-electron orbitals from above the Fermi energy ϵ_F. At first sight the BCS state appears to have a higher energy than the Fermi state: the comparison of (b) with (a) shows that the kinetic energy of the BCS state is higher than that of the Fermi state. But the attractive potential energy of the BCS state, although not represented in the figure, acts to lower the total energy of the BCS state with respect to the Fermi state.

When the BCS ground state of a many-electron system is described in terms of the occupancy of one-particle orbitals, those near ϵ_F are filled somewhat like a Fermi-Dirac distribution for some finite temperature.

The central feature of the BCS state is that the one-particle orbitals are occupied in pairs: if an orbital with wavevector $\mathbf{k}$ and spin up is occupied, then the orbital with wavevector $-\mathbf{k}$ and spin down is also occupied. If $\mathbf{k}\uparrow$ is vacant, then $-\mathbf{k}\downarrow$ is also vacant. The pairs are called **Cooper pairs**, treated in Appendix H. They have spin zero and have many attributes of bosons.

Flux Quantization in a Superconducting Ring

We prove that the total magnetic flux that passes through a superconducting ring may assume only quantized values, integral multiples of the flux quantum $2\pi\hbar c/q$, where by experiment $q = 2e$, the charge of an electron pair. Flux quantization is a beautiful example of a long-range quantum effect in which the coherence of the superconducting state extends over a ring or solenoid.

Let us first consider the electromagnetic field as an example of a similar boson field. The electric field intensity $E(\mathbf{r})$ acts qualitatively as a probability field amplitude. When the total number of photons is large, the energy density may be written as

$$E^*(\mathbf{r})E(\mathbf{r})/4\pi \cong n(\mathbf{r})\hbar\omega \ ,$$

where $n(\mathbf{r})$ is the number density of photons of frequency ω. Then we may write the electric field in a semiclassical approximation as

$$E(\mathbf{r}) \cong (4\pi\hbar\omega)^{1/2}\, n(\mathbf{r})^{1/2}\, e^{i\theta(\mathbf{r})} \qquad E^*(\mathbf{r}) \cong (4\pi\hbar\omega)^{1/2}\, n(\mathbf{r})^{1/2}\, e^{-i\theta(\mathbf{r})} \ ,$$

where $\theta(\mathbf{r})$ is the phase of the field. A similar probability amplitude describes Cooper pairs.

The arguments that follow apply to a boson gas with a large number of bosons in the same orbital. We then can treat the boson probability amplitude as a classical quantity, just as the electromagnetic field is used for photons. Both amplitude and phase are then meaningful and observable. The arguments do not apply to a metal in the normal state because an electron in the normal state acts as a single unpaired fermion that cannot be treated classically.

We first show that a charged boson gas obeys the London equation. Let $\psi(\mathbf{r})$ be the particle probability amplitude. We suppose that the pair concentration $n = \psi^*\psi$ = constant. At absolute zero n is one-half of the concentration of electrons in the conduction band, for n refers to pairs. Then we may write

$$\psi = n^{1/2}\, e^{i\,\theta(\mathbf{r})} \ ; \qquad \psi^* = n^{1/2}\, e^{-i\,\theta(\mathbf{r})} \ . \tag{19}$$

The phase $\theta(\mathbf{r})$ is important for what follows. In SI units, set $c = 1$ in the equations that follow.

The velocity of a particle is, from the Hamilton equations of mechanics,

$$\text{(CGS)} \qquad \mathbf{v} = \frac{1}{m}\left(\mathbf{p} - \frac{q}{c}\mathbf{A}\right) = \frac{1}{m}\left(-i\hbar\nabla - \frac{q}{c}\mathbf{A}\right) \ .$$

The particle flux is given by

$$\psi^*\mathbf{v}\psi = \frac{n}{m}\left(\hbar\nabla\theta - \frac{q}{c}\mathbf{A}\right) \ , \tag{20}$$

so that the electric current density is

$$\mathbf{j} = q\psi^*\mathbf{v}\psi \ = \frac{nq}{m}\left(\hbar\nabla\theta - \frac{q}{c}\mathbf{A}\right) \ . \tag{21}$$

We may take the curl of both sides to obtain the London equation:

$$\operatorname{curl}\mathbf{j} = -\frac{nq^2}{mc}\mathbf{B} \ , \tag{22}$$

with use of the fact that the curl of the gradient of a scalar is identically zero. The constant that multiplies $\mathbf{B}$ agrees with (14a). We recall that the Meissner effect is a consequence of the London equation, which we have here derived.

Quantization of the magnetic flux through a ring is a dramatic consequence of Eq. (21). Let us take a closed path C through the interior of the

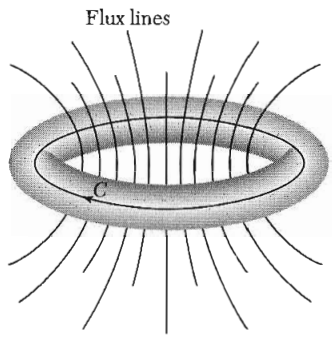


Figure 16 Path of integration C through the interior of a superconducting ring. The flux through the ring is the sum of the flux Φ_{ext} from external sources and the flux Φ_{sc} from the superconducting currents which flow in the surface of the ring; $\Phi = \Phi_{ext} + \Phi_{sc}$. The flux Φ is quantized. There is normally no quantization condition on the flux from external sources, so that Φ_{sc} must adjust itself appropriately in order that Φ assume a quantized value.

superconducting material, well away from the surface (Fig. 16). The Meissner effect tells us that **B** and **j** are zero in the interior. Now (21) is zero if

$$\hbar c \nabla\theta = q\mathbf{A} \ . \tag{23}$$

We form

$$\oint_C \nabla\theta \cdot dl = \theta_2 - \theta_1$$

for the change of phase on going once around the ring.

The probability amplitude ψ is measurable in the classical approximation, so that ψ must be single-valued and

$$\theta_2 - \theta_1 = 2\pi s \ , \tag{24}$$

where s is an integer. By the Stokes theorem,

$$\oint_C \mathbf{A} \cdot dl = \int_C (\text{curl } \mathbf{A}) \cdot d\boldsymbol{\sigma} = \int_C \mathbf{B} \cdot d\boldsymbol{\sigma} = \Phi \ , \tag{25}$$

where $d\boldsymbol{\sigma}$ is an element of area on a surface bounded by the curve C, and Φ is the magnetic flux through C. From (23), (24), and (25) we have $2\pi\hbar cs = q\Phi$, or

$$\Phi = (2\pi\hbar c/q)s \ . \tag{26}$$

Thus the flux through the ring is quantized in integral multiples of $2\pi\hbar c/q$.

By experiment $q = -2e$ as appropriate for electron pairs, so that the quantum of flux in a superconductor is

(CGS) $$\Phi_0 = 2\pi\hbar c/2e \cong 2.0678 \times 10^{-7} \text{ gauss cm}^2 \ ;$$

(SI) $$\Phi_0 = 2\pi\hbar/2e \cong 2.0678 \times 10^{-15} \text{ tesla m}^2 \ . \tag{27}$$

This flux quantum is called a **fluxoid** or **fluxon**.

The flux through the ring is the sum of the flux Φ_{ext} from external sources and the flux Φ_{sc} from the persistent superconducting currents which flow in

the surface of the ring: $\Phi = \Phi_{ext} + \Phi_{sc}$. The flux Φ is quantized. There is normally no quantization condition on the flux from external sources, so that Φ_{sc} must adjust itself appropriately in order that Φ assume a quantized value.

Duration of Persistent Currents

Consider a persistent current that flows in a ring of a type I superconductor of wire of length L and cross-sectional area A. The persistent current maintains a flux through the ring of some integral number of fluxoids (27). A fluxoid cannot leak out of the ring and thereby reduce the persistent current unless by a thermal fluctuation a minimum volume of the superconducting ring is momentarily in the normal state.

The probability per unit time that a fluxoid will leak out is the product

$$P = (\text{attempt frequency})(\text{activation barrier factor}) \ . \tag{28}$$

The activation barrier factor is $\exp(-\Delta F/k_B T)$, where the free energy of the barrier is

$$\Delta F \approx (\text{minimum volume})(\text{excess free energy density of normal state}) \ .$$

The minimum volume of the ring that must turn normal to allow a fluxoid to escape is of the order of $R\xi^2$, where ξ is the coherence length of the superconductor and R the wire thickness. The excess free energy density of the normal state is $H_c^2/8\pi$, whence the barrier free energy is

$$\Delta F \approx R\xi^2 H_c^2/8\pi \ . \tag{29}$$

Let the wire thickness be 10^{-4} cm, the coherence length $= 10^{-4}$ cm, and $H_c = 10^3$ G; then $\Delta F \approx 10^{-7}$ erg. As we approach the transition temperature from below, ΔF will decrease toward zero, but the value given is a fair estimate between absolute zero and 0.8 T_c. Thus the activation barrier factor is

$$\exp(-\Delta F/k_B T) \approx \exp(-10^8) \approx 10^{-(4.34\times 10^7)} \ .$$

The characteristic frequency with which the minimum volume can attempt to change its state must be of the order of $E_g/\hbar$. If $E_g = 10^{-15}$ erg, the attempt frequency is $\approx 10^{-15}/10^{-27} \approx 10^{12}\ \text{s}^{-1}$. The leakage probability (28) becomes

$$P \approx 10^{12} 10^{-4.34\times 10^7\ \text{s}^{-1}} = 10^{-4.34\times 10^7\ \text{s}^{-1}} \ .$$

The reciprocal of this is a measure of the time required for a fluxoid to leak out, $T = 1/P = 10^{4.34\times 10^7}$ s.

The age of the universe is only 10^{18} s, so that a fluxoid will not leak out in the age of the universe, under our assumed conditions. Accordingly, the current is maintained.

There are two circumstances in which the activation energy is much lower and a fluxoid can be observed to leak out of a ring—either very close to the critical temperature, where H_c is very small, or when the material of the ring is

a type II superconductor and already has fluxoids embedded in it. These special situations are discussed in the literature under the subject of fluctuations in superconductors.

Type II Superconductors

There is no difference in the mechanism of superconductivity in type I and type II superconductors. Both types have similar thermal properties at the superconductor-normal transition in zero magnetic field. But the Meissner effect is entirely different (Fig. 5).

A good type I superconductor excludes a magnetic field until superconductivity is destroyed suddenly, and then the field penetrates completely. A good type II superconductor excludes the field completely up to a field H_{c1}. Above H_{c1} the field is partially excluded, but the specimen remains electrically superconducting. At a much higher field, H_{c2}, the flux penetrates completely and superconductivity vanishes. (An outer surface layer of the specimen may remain superconducting up to a still higher field H_{c3}.)

An important difference in a type I and a type II superconductor is in the mean free path of the conduction electrons in the normal state. If the coherence length ξ is longer than the penetration depth λ, the superconductor will be type I. Most pure metals are type I, with $\lambda/\xi < 1$ (see Table 5 on p. 275).

But, when the mean free path is short, the coherence length is short and the penetration depth is great (Fig. 14). This is the situation when $\lambda/\xi > 1$, and the superconductor will be type II.

We can change some metals from type I to type II by a modest addition of an alloying element. In Figure 5 the addition of 2 wt. percent of indium changes lead from type I to type II, although the transition temperature is scarcely changed at all. Nothing fundamental has been done to the electronic structure of lead by this amount of alloying, but the magnetic behavior as a superconductor has changed drastically.

The theory of type II superconductors was developed by Ginzburg, Landau, Abrikosov, and Gorkov. Later Kunzler and co-workers observed that Nb_3Sn wires can carry large supercurrents in fields approaching 100 kG; this led to the commercial development of strong-field superconducting magnets.

Consider the interface between a region in the superconducting state and a region in the normal state. The interface has a surface energy that may be positive or negative and that decreases as the applied magnetic field is increased. A superconductor is type I if the surface energy is always positive as the magnetic field is increased, and type II if the surface energy becomes negative as the magnetic field is increased. The sign of the surface energy has no importance for the transition temperature.

The free energy of a bulk superconductor is increased when the magnetic field is expelled. However, a parallel field can penetrate a very thin film nearly uniformly (Fig. 17), only a part of the flux is expelled, and the energy of the

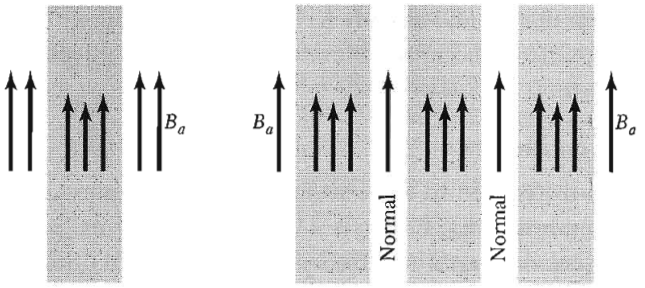


Figure 17 (a) Magnetic field penetration into a thin film of thickness equal to the penetration depth λ. The arrows indicate the intensity of the magnetic field. (b) Magnetic field penetration in a homogeneous bulk structure in the mixed or vortex state, with alternate layers in normal and superconducting states. The superconducting layers are thin in comparison with λ. The laminar structure is shown for convenience; the actual structure consists of rods of the normal state surrounded by the superconducting state. (The N regions in the vortex state are not exactly normal, but are described by low values of the stabilization energy density.)

superconducting film will increase only slowly as the external magnetic field is increased. This causes a large increase in the field intensity required for the destruction of superconductivity. The film has the usual energy gap and will be resistanceless. A thin film is not a type II superconductor, but the film results show that under suitable conditions superconductivity can exist in high magnetic fields.

Vortex State. The results for thin films suggest the question: Are there stable configurations of a superconductor in a magnetic field with regions (in the form of thin rods or plates) in the normal state, each normal region surrounded by a superconducting region? In such a mixed state, called the vortex state, the external magnetic field will penetrate the thin normal regions uniformly, and the field will also penetrate somewhat into the surrounding superconducting material, as in Fig. 18.

The term vortex state describes the circulation of superconducting currents in vortices throughout the bulk specimen, as in Fig. 19. There is no chemical or crystallographic difference between the normal and the superconducting regions in the vortex state. The vortex state is stable when the penetration of the applied field into the superconducting material causes the surface energy to become negative. A type II superconductor is characterized by a vortex state stable over a certain range of magnetic field strength; namely, between H_{c1} and H_{c2}.

Estimation of H_{c1} and H_{c2}. What is the condition for the onset of the vortex state as the applied magnetic field is increased? We estimate H_{c1} from the penetration depth λ. The field in the normal core of the fluxoid will be H_{c1} when the applied field is H_{c1}.

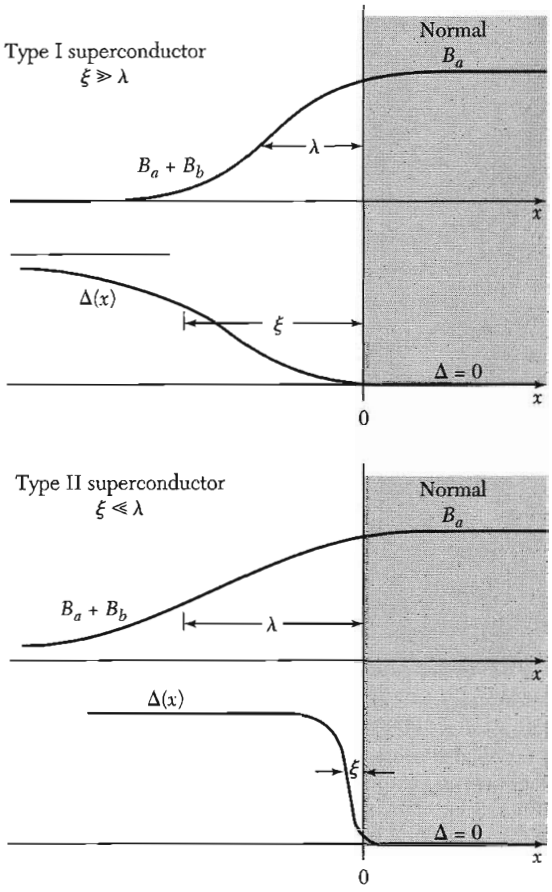


Figure 18 Variation of the magnetic field and energy gap parameter $\Delta(x)$ at the interface of superconducting and normal regions, for type I and type II superconductors. The energy gap parameter is a measure of the stabilization energy density of the superconducting state.

The field will extend out from the normal core a distance λ into the superconducting environment. The flux thus associated with a single core is $\pi\lambda^2 H_{c1}$, and this must be equal to the flux quantum Φ_0 defined by (27). Thus

$$H_{c1} \simeq \Phi_0/\pi\lambda^2 \ . \tag{30}$$

This is the field for nucleation of a single fluxoid.

At H_{c2} the fluxoids are packed together as tightly as possible, consistent with the preservation of the superconducting state. This means as densely as the coherence length ξ will allow. The external field penetrates the specimen almost uniformly, with small ripples on the scale of the fluxoid lattice. Each core is responsible for carrying a flux of the order of $\pi\xi^2 H_{c2}$, which also is quantized to Φ_0. Thus

$$H_{c2} \approx \Phi_0/\pi\xi^2 \tag{31}$$

gives the upper critical field. The larger the ratio λ/ξ, the larger is the ratio of H_{c2} to H_{c1}.

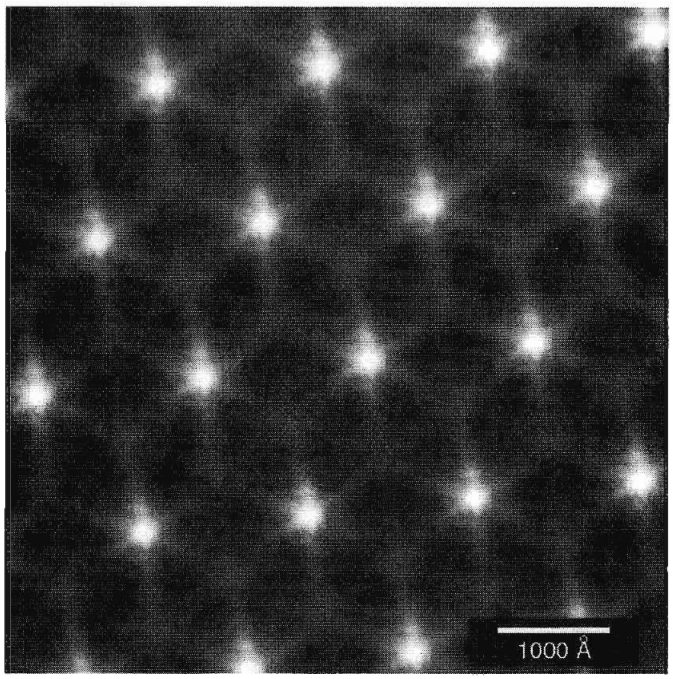


Figure 19 Flux lattice in $NbSe_2$ at 1,000 gauss at 0.2K, as viewed with a scanning tunneling microscope. The photo shows the density of states at the Fermi level, as in Figure 23. The vortex cores have a high density of states and are shaded white; the superconducting regions are dark, with no states at the Fermi level. The amplitude and spatial extent of these states is determined by a potential well formed by $\Delta(x)$ as in Fig. 18 for a Type II superconductor. The potential well confines the core state wavefunctions in the image here. The star shape is a finer feature, a result special to $NbSe_2$ of the sixfold disturbance of the charge density at the Fermi surface. Photo courtesy of H. F. Hess.

It remains to find a relation between these critical fields and the thermodynamic critical field H_c that measures the stabilization energy density of the superconducting state, which is known by (9) to be $H_c^2/8\pi$. In a type II superconductor we can determine H_c only indirectly by calorimetric measurement of the stabilization energy. To estimate H_{c1} in terms of H_c, we consider the stability of the vortex state at absolute zero in the impure limit $\xi < \lambda$; here $\kappa > 1$ and the coherence length is short in comparison with the penetration depth.

We estimate in the vortex state the stabilization energy of a fluxoid core viewed as a normal metal cylinder which carries an average magnetic field B_a. The radius is of the order of the coherence length, the thickness of the boundary between N and S phases. The energy of the normal core referred to the energy of a pure superconductor is given by the product of the stabilization energy times the area of the core:

(CGS)
$$f_{\text{core}} \approx \frac{1}{8\pi} H_c^2 \times \pi\xi^2 \ , \tag{32}$$

per unit length. But there is also a decrease in magnetic energy because of the penetration of the applied field B_a into the superconducting material around the core:

$$\text{(CGS)} \qquad f_{\text{mag}} \approx -\frac{1}{8\pi} B_a^2 \times \pi\lambda^2 \ . \tag{33}$$

For a single fluxoid we add these two contributions to obtain

$$\text{(CGS)} \qquad f = f_{\text{core}} + f_{\text{mag}} \approx \tfrac{1}{8}(H_c^2\xi^2 - B_a^2\lambda^2) \ . \tag{34}$$

The core is stable if $f < 0$. The threshold field for a stable fluxoid is at $f = 0$, or, with H_{c1} written for B_a,

$$H_{c1}/H_c \approx \xi/\lambda \ . \tag{35}$$

The threshold field divides the region of positive surface energy from the region of negative surface energy.

We can combine (30) and (35) to obtain a relation for H_c:

$$\pi\xi\lambda H_c \approx \Phi_0 \ . \tag{36}$$

We can combine (30), (31), and (35) to obtain

$$(H_{c1}H_{c2})^{1/2} \approx H_c \ , \tag{37a}$$

and

$$H_{c2} \approx (\lambda/\xi)H_c = \kappa H_c \ . \tag{37b}$$

Single Particle Tunneling

Consider two metals separated by an insulator, as in Fig. 20. The insulator normally acts as a barrier to the flow of conduction electrons from one metal to the other. If the barrier is sufficiently thin (less than 10 or 20 Å) there is a significant probability that an electron which impinges on the barrier will pass from one metal to the other: this is called **tunneling**. In many experiments the insulating layer is simply a thin oxide layer formed on one of two evaporated metal films, as in Fig. 21.

When both metals are normal conductors, the current-voltage relation of the sandwich or tunneling junction is ohmic at low voltages, with the current directly proportional to the applied voltage. Giaever (1960) discovered that if one of the metals becomes superconducting the current-voltage characteristic changes from the straight line of Fig. 22a to the curve shown in Fig. 22b.

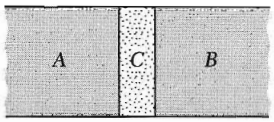


Figure 20 Two metals, *A* and *B*, separated by a thin layer of an insulator *C*.

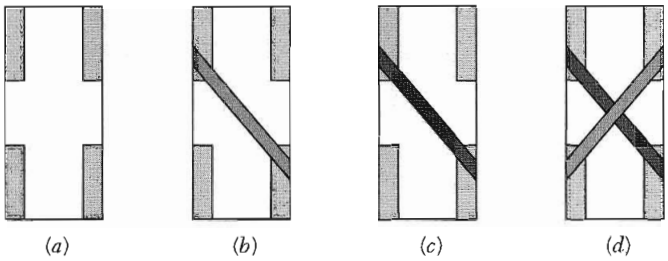


Figure 21 Preparation of an $Al/Al_2O_3/Sn$ sandwich. (a) Glass slide with indium contacts. (b) An aluminum strip 1 mm wide and 1000 to 3000 Å thick has been deposited across the contacts. (c) The aluminum strip has been oxidized to form an Al_2O_3 layer 10 to 20 Å in thickness. (d) A tin film has been deposited across the aluminum film, forming an $Al/Al_2O_3/Sn$ sandwich. The external leads are connected to the indium contacts; two contacts are used for the current measurement and two for the voltage measurement. (After Giaever and Megerle.)

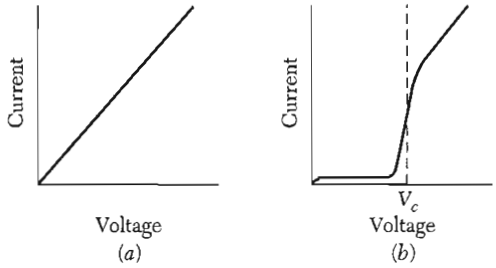


Figure 22 (a) Linear current-voltage relation for junction of normal metals separated by oxide layer; (b) current-voltage relation with one metal normal and the other metal superconducting.

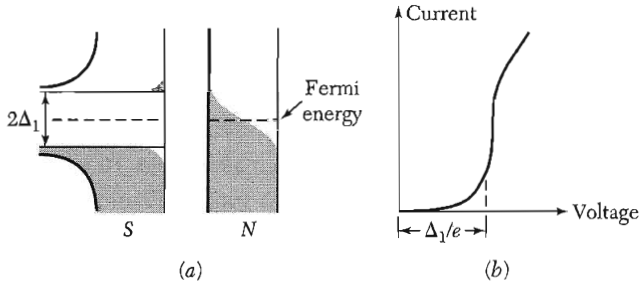


Figure 23 The density of orbitals and the current-voltage characteristic for a tunneling junction. In (a) the energy is plotted on the vertical scale and the density of orbitals on the horizontal scale. One metal is in the normal state and one in the superconducting state. (b) I versus V; the dashes indicate the expected break at $T = 0$. (After Giaever and Megerle.)

Figure 23a contrasts the electron density of orbitals in the superconductor with that in the normal metal. In the superconductor there is an energy gap centered at the Fermi level. At absolute zero no current can flow until the applied voltage is $V = E_g/2e = \Delta/e$.

The gap E_g corresponds to the break-up of a pair of electrons in the superconducting state, with the formation of two electrons, or an electron and

a hole, in the normal state. The current starts when $eV = \Delta$. At finite temperatures there is a small current flow even at low voltages, because of electrons in the superconductor that are thermally excited across the energy gap.

Josephson Superconductor Tunneling

Under suitable conditions we observe remarkable effects associated with the tunneling of superconducting electron pairs from a superconductor through a layer of an insulator into another superconductor. Such a junction is called a weak link. The effects of pair tunneling include:

Dc Josephson effect. A dc current flows across the junction in the absence of any electric or magnetic field.

Ac Josephson effect. A dc voltage applied across the junction causes rf current oscillations across the junction. This effect has been utilized in a precision determination of the value of $\hbar/e$. Further, an rf voltage applied with the dc voltage can then cause a dc current across the junction.

Macroscopic long-range quantum interference. A dc magnetic field applied through a superconducting circuit containing two junctions causes the maximum supercurrent to show interference effects as a function of magnetic field intensity. This effect can be utilized in sensitive magnetometers.

Dc Josephson Effect. Our discussion of Josephson junction phenomena follows the discussion of flux quantization. Let ψ_1 be the probability amplitude of electron pairs on one side of a junction, and let ψ_2 be the amplitude on the other side. For simplicity, let both superconductors be identical. For the present we suppose that they are both at zero potential.

The time-dependent Schrödinger equation $i\hbar\partial\psi/\partial t = \mathcal{H}\psi$ applied to the two amplitudes gives

$$i\hbar \frac{\partial \psi_1}{\partial t} = \hbar T \psi_2 \ ; \qquad i\hbar \frac{\partial \psi_2}{\partial t} = \hbar T \psi_1 \ . \tag{38}$$

Here $\hbar T$ represents the effect of the electron-pair coupling or transfer interaction across the insulator; T has the dimensions of a rate or frequency. It is a measure of the leakage of ψ_1 into the region 2, and of ψ_2 into the region 1. If the insulator is very thick, T is zero and there is no pair tunneling.

Let $\psi_1 = n_1^{1/2} e^{i\theta_1}$ and $\psi_2 = n_2^{1/2} e^{i\theta_2}$. Then

$$\frac{\partial \psi_1}{\partial t} = \tfrac{1}{2} n_1^{-1/2} e^{i\theta_1} \frac{\partial n_1}{\partial t} + i\psi_1 \frac{\partial \theta_1}{\partial t} = -iT\psi_2 \ ; \tag{39}$$

$$\frac{\partial \psi_2}{\partial t} = \tfrac{1}{2} n_2^{-1/2} e^{i\theta_2} \frac{\partial n_2}{\partial t} + i\psi_2 \frac{\partial \theta_2}{\partial t} = -iT\psi_1 \ . \tag{40}$$

We multiply (39) by $n_1^{1/2} e^{-i\theta_1}$ to obtain, with $\delta \equiv \theta_2 - \theta_1$,

$$\frac{1}{2}\frac{\partial n_1}{\partial t} + i n_1 \frac{\partial \theta_1}{\partial t} = -iT(n_1 n_2)^{1/2} e^{i\delta} \ . \tag{41}$$

We multiply (40) by $n_2^{1/2} e^{-i\theta_2}$ to obtain

$$\frac{1}{2}\frac{\partial n_2}{\partial t} + i n_2 \frac{\partial \theta_2}{\partial t} = -iT(n_1 n_2)^{1/2} e^{-i\delta} \ . \tag{42}$$

Now equate the real and imaginary parts of (41) and similarly of (42):

$$\frac{\partial n_1}{\partial t} = 2T(n_1 n_2)^{1/2} \sin\delta \ ; \qquad \frac{\partial n_2}{\partial t} = -2T(n_1 n_2)^{1/2} \sin\delta \ ; \tag{43}$$

$$\frac{\partial \theta_1}{\partial t} = -T\left(\frac{n_2}{n_1}\right)^{1/2} \cos\delta \ ; \qquad \frac{\partial \theta_2}{\partial t} = -T\left(\frac{n_1}{n_2}\right)^{1/2} \cos\delta \ . \tag{44}$$

If $n_1 \cong n_2$ as for identical superconductors 1 and 2, we have from (44) that

$$\frac{\partial \theta_1}{\partial t} = \frac{\partial \theta_2}{\partial t} \ ; \qquad \frac{\partial}{\partial t}(\theta_2 - \theta_1) = 0 \ . \tag{45}$$

From (43) we see that

$$\frac{\partial n_2}{\partial t} = -\frac{\partial n_1}{\partial t} \ . \tag{46}$$

The current flow from (1) to (2) is proportional to $\partial n_2/\partial t$ or, the same thing, $-\partial n_1/\partial t$. We therefore conclude from (43) that the current J of superconductor pairs across the junction depends on the phase difference δ as

$$\boxed{J = J_0 \sin\delta = J_0 \sin(\theta_2 - \theta_1) \ .} \tag{47}$$

where J_0 is proportional to the transfer interaction T. The current J_0 is the maximum zero-voltage current that can be passed by the junction. With no applied voltage a dc current will flow across the junction (Fig. 24), with a value between J_0 and $-J_0$ according to the value of the phase difference $\theta_2 - \theta_1$. This is the **dc Josephson effect**.

Ac Josephson Effect. Let a dc voltage V be applied across the junction. We can do this because the junction is an insulator. An electron pair experiences a potential energy difference qV on passing across the junction, where $q = -2e$. We can say that a pair on one side is at potential energy $-eV$ and a pair on the other side is at eV. The equations of motion that replace (38) are

$$i\hbar\, \partial\psi_1/\partial t = \hbar T\psi_2 - eV\psi_1 \ ; \qquad i\hbar\, \partial\psi_2/\partial t = \hbar T\psi_1 + eV\psi_2 \ . \tag{48}$$

We proceed as above to find in place of (41) the equation

$$\frac{1}{2}\frac{\partial n_1}{\partial t} + i n_1 \frac{\partial \theta_1}{\partial t} = ieVn_1\hbar^{-1} - iT(n_1 n_2)^{1/2} e^{i\delta} \ . \tag{49}$$

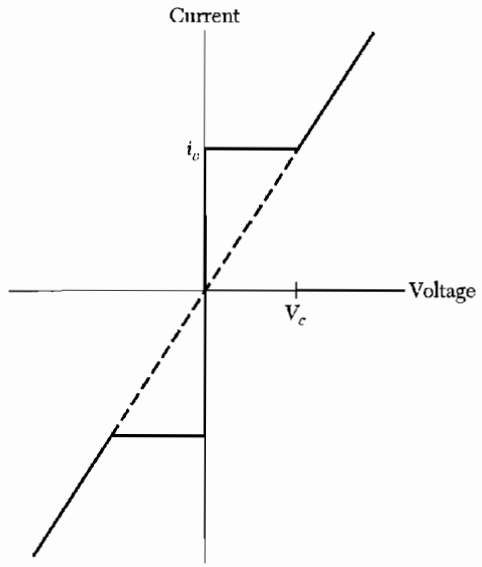


Figure 24 Current-voltage characteristic of a Josephson junction. Dc currents flow under zero applied voltage up to a critical current i_c: this is the dc Josephson effect. At voltages above V_c the junction has a finite resistance, but the current has an oscillatory component of frequency $\omega = 2eV/\hbar$: this is the ac Josephson effect.

This equation breaks up into the real part

$$\partial n_1/\partial t = 2T(n_1 n_2)^{1/2} \sin \delta \ , \tag{50}$$

exactly as without the voltage V, and the imaginary part

$$\partial \theta_1/\partial t = (eV/\hbar) - T(n_2/n_1)^{1/2} \cos \delta \ , \tag{51}$$

which differs from (44) by the term $eV/\hbar$.

Further, by extension of (42),

$$\frac{1}{2}\frac{\partial n_2}{\partial t} + in_2 \frac{\partial \theta_2}{\partial t} = -i\, eVn_2\hbar^{-1} - iT(n_1 n_2)^{1/2} e^{-i\delta} \ , \tag{52}$$

whence

$$\partial n_2/\partial t = -2T(n_1 n_2)^{1/2} \sin \delta \ ; \tag{53}$$

$$\partial \theta_2/\partial t = -(eV/\hbar) - T(n_1/n_2)^{1/2} \cos \delta \ . \tag{54}$$

From (51) and (54) with $n_1 \cong n_2$, we have

$$\partial(\theta_2 - \theta_1)/\partial t = \partial \delta/\partial t = -2eV/\hbar \ . \tag{55}$$

We see by integration of (55) that with a dc voltage across the junction the relative phase of the probability amplitudes varies as

$$\delta(t) = \delta(0) - (2eVt/\hbar) \ . \tag{56}$$

The superconducting current is given by (47) with (56) for the phase:

$$\boxed{J = J_0 \sin [\delta(0) - (2eVt/\hbar)] \ .} \tag{57}$$

The current oscillates with frequency

$$\omega = 2eV/\hbar \ . \tag{58}$$

This is the **ac Josephson effect**. A dc voltage of 1 μV produces a frequency of 483.6 MHz. The relation (58) says that a photon of energy $\hbar\omega = 2eV$ is emitted or absorbed when an electron pair crosses the barrier. By measuring the voltage and the frequency it is possible to obtain a very precise value of $e/\hbar$.

Macroscopic Quantum Interference. We saw in (24) and (26) that the phase difference $\theta_2 - \theta_1$ around a closed circuit which encompasses a total magnetic flux Φ is given by

$$\theta_2 - \theta_1 = (2e/\hbar c)\Phi \ . \tag{59}$$

The flux is the sum of that due to external fields and that due to currents in the circuit itself.

We consider two Josephson junctions in parallel, as in Fig. 25. No voltage is applied. Let the phase difference between points 1 and 2 taken on a path through junction a be δ_a. When taken on a path through junction b, the phase difference is δ_b. In the absence of a magnetic field these two phases must be equal.

Now let the flux Φ pass through the interior of the circuit. We do this with a straight solenoid normal to the plane of the paper and lying inside the circuit. By (59), $\delta_b - \delta_a = (2e/\hbar c)\Phi$, or

$$\delta_b = \delta_0 + \frac{e}{\hbar c}\Phi \ ; \qquad \delta_a = \delta_0 - \frac{e}{\hbar c}\Phi \ . \tag{60}$$

The total current is the sum of J_a and J_b. The current through each junction is of the form (47), so that

$$J_{\text{Total}} = J_0\left\{\sin\left(\delta_0 + \frac{e}{\hbar c}\Phi\right) + \sin\left(\delta_0 - \frac{e}{\hbar c}\Phi\right)\right\} = 2(J_0 \sin\delta_0)\cos\frac{e\Phi}{\hbar c} \ .$$

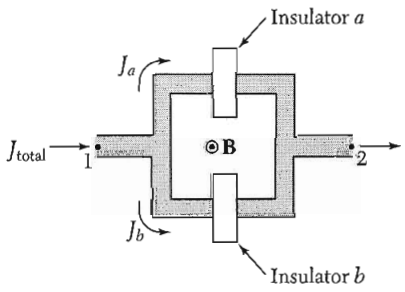


Figure 25 The arrangement for experiment on macroscopic quantum interference. A magnetic flux Φ passes through the interior of the loop.

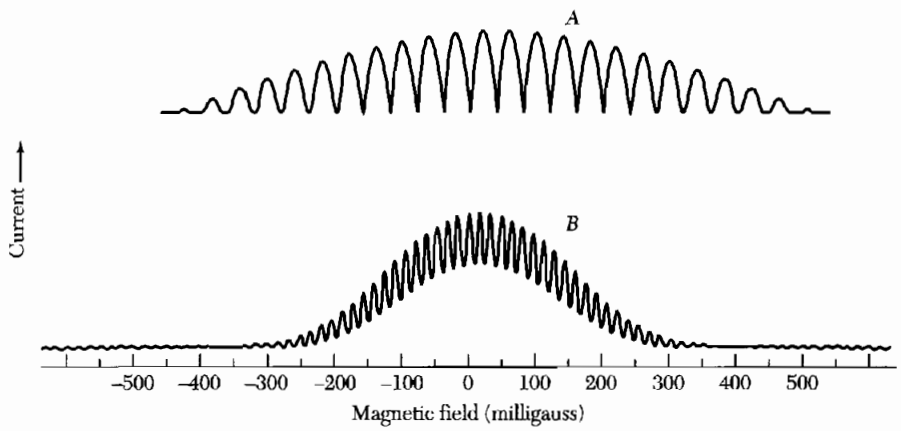


Figure 26 Experimental trace of J_{max} versus magnetic field showing interference and diffraction effects for two junctions A and B. The field periodicity is 39.5 and 16 mG for A and B, respectively. Approximate maximum currents are 1 mA (A) and 0.5 mA (B). The junction separation is 3 mm and junction width 0.5 mm for both cases. The zero offset of A is due to a background magnetic field. (After R. C. Jaklevic, J. Lambe, J. E. Mercereau, and A. H. Silver.)

The current varies with Φ and has maxima when

$$e\Phi/\hbar c = s\pi \ , \quad s = \text{integer} \ . \tag{61}$$

The periodicity of the current is shown in Fig. 26. The short period variation is produced by interference from the two junctions, as predicted by (61). The longer period variation is a diffraction effect and arises from the finite dimensions of each junction—this causes Φ to depend on the particular path of integration (Problem 6).

HIGH-TEMPERATURE SUPERCONDUCTORS

High T_c or HTS denotes superconductivity in materials, chiefly copper oxides, with high transition temperatures, accompanied by high critical currents and magnetic fields. By 1988 the long-standing 23 K ceiling of T_c in intermetallic compounds had been elevated to 125 K in bulk superconducting oxides; these passed the standard tests for superconductivity—the Meissner effect, ac Josephson effect, persistent currents of long duration, and substantially zero dc resistivity. Memorable steps in the advance include:

$BaPb_{0.75}Bi_{0.25}O_3$	$T_c = 12$ K	[BPBO]
$La_{1.85}Ba_{0.15}CuO_4$	$T_c = 36$ K	[LBCO]
$YBa_2Cu_3O_7$	$T_c = 90$ K	[YBCO]
$Tl_2Ba_2Ca_2Cu_3O_{10}$	$T_c = 120$ K	[TBCO]
$Hg_{0.8}Tl_{0.2}Ba_2Ca_2Cu_3O_{8.33}$	$T_c = 138$ K	

SUMMARY
(In CGS Units)

- A superconductor exhibits infinite conductivity.
- A bulk specimen of metal in the superconducting state exhibits perfect diamagnetism, with the magnetic induction $\mathbf{B} = 0$. This is the Meissner effect. The external magnetic field will penetrate the surface of the specimen over a distance determined by the penetration depth λ.
- There are two types of superconductors, I and II. In a bulk specimen of type I superconductor the superconducting state is destroyed and the normal state is restored by application of an external magnetic field in excess of a critical value H_c. A type II superconductor has two critical fields, $H_{c1} < H_c < H_{c2}$; a vortex state exists in the range between H_{c1} and H_{c2}. The stabilization energy density of the pure superconducting state is $H_c^2/8\pi$ in both type I and II superconductors.
- In the superconducting state an energy gap, $E_g \approx 4k_BT_c$, separates superconducting electrons below from normal electrons above the gap. The gap is detected in experiments on heat capacity, infrared absorption, and tunneling.
- Three important lengths enter the theory of superconductivity; the London penetration depth λ_L; the intrinsic coherence length ξ_0; and the normal electron mean free path ℓ.
- The London equation

$$\mathbf{j} = -\frac{c}{4\pi\lambda_L^2}\mathbf{A} \qquad \text{or} \qquad \text{curl}\,\mathbf{j} = -\frac{c}{4\pi\lambda_L^2}\mathbf{B}$$

  leads to the Meissner effect through the penetration equation $\nabla^2 B = B/\lambda_L^2$, where $\lambda_L \approx (mc^2/4\pi ne^2)^{1/2}$ is the London penetration depth.
- In the London equation $\mathbf{A}$ or $\mathbf{B}$ should be a weighted average over the coherence length ξ. The intrinsic coherence length $\xi_0 = 2\hbar v_F/\pi E_g$.
- The BCS theory accounts for a superconducting state formed from pairs of electrons $\mathbf{k}\uparrow$ and $-\mathbf{k}\downarrow$. These pairs act as bosons.
- Type II superconductors have $\xi < \lambda$. The critical fields are related by $H_{c1} \approx (\xi/\lambda)H_c$ and $H_{c2} \approx (\lambda/\xi)H_c$. The Ginzburg-Landau parameter κ is defined as λ/ξ.

Problems

1. ***Magnetic field penetration in a plate.*** The penetration equation may be written as $\lambda^2\nabla^2 B = B$, where λ is the penetration depth. (a) Show that $B(x)$ inside a superconducting plate perpendicular to the x axis and of thickness δ is given by

$$B(x) = B_a \frac{\cosh(x/\lambda)}{\cosh(\delta/2\lambda)},$$

where B_a is the field outside the plate and parallel to it; here $x = 0$ is at the center of the plate. (b) The effective magnetization $M(x)$ in the plate is defined by $B(x) - B_a = 4\pi M(x)$. Show that, in CGS, $4\pi M(x) = -B_a(1/8\lambda^2)(\delta^2 - 4x^2)$, for $\delta \ll \lambda$. In SI we replace the 4π by μ_0.

2. ***Critical field of thin films.*** (a) Using the result of Problem 1b, show that the free energy density at $T = 0$ K within a superconducting film of thickness δ in an external magnetic field B_a is given by, for $\delta \ll \lambda$,

(CGS) $$F_S(x, B_a) = U_S(0) + (\delta^2 - 4x^2)B_a^2/64\pi\lambda^2 \ .$$

In SI the factor π is replaced by $\frac{1}{4}\mu_0$. We neglect a kinetic energy contribution to the problem. (b) Show that the magnetic contribution to F_S when averaged over the thickness of the film is $B_a^2(\delta/\lambda)^2/96\pi$. (c) Show that the critical field of the thin film is proportional to $(\lambda/\delta)H_c$, where H_c is the bulk critical field, if we consider only the magnetic contribution to U_S.

3. ***Two-fluid model of a superconductor.*** On the two-fluid model of a superconductor we assume that at temperatures $0 < T < T_c$ the current density may be written as the sum of the contributions of normal and superconducting electrons: $\mathbf{j} = \mathbf{j}_N + \mathbf{j}_S$, where $\mathbf{j}_N = \sigma_0\mathbf{E}$ and $\mathbf{j}_S$ is given by the London equation. Here σ_0 is an ordinary normal conductivity, decreased by the reduction in the number of normal electrons at temperature T as compared to the normal state. Neglect inertial effects on both j_N and j_S. (a) Show from the Maxwell equations that the dispersion relation connecting wavevector $\mathbf{k}$ and frequency ω for electromagnetic waves in the superconductor is

(CGS) $$k^2c^2 = 4\pi\sigma_0\omega i - c^2\lambda_L^{-2} + \omega^2 \ ; \qquad \text{or}$$

(SI) $$k^2c^2 = (\sigma_0/\epsilon_0)\,\omega i - c^2\lambda_L^{-2} + \omega^2 \ ;$$

where λ_L^2 is given by (14a) with n replaced by n_S. Recall that curl curl $\mathbf{B} = -\nabla^2\mathbf{B}$. (b) If τ is the relaxation time of the normal electrons and n_N is their concentration, show by use of the expression $\sigma_0 = n_N e^2\tau/m$ that at frequencies $\omega \ll 1/\tau$ the dispersion relation does not involve the normal electrons in an important way, so that the motion of the electrons is described by the London equation alone. The supercurrent short-circuits the normal electrons. The London equation itself only holds true if $\hbar\omega$ is small in comparison with the energy gap. Note: The frequencies of interest are such that $\omega \ll \omega_p$, where ω_p is the plasma frequency.

*4. ***Structure of a vortex.*** (a) Find a solution to the London equation that has cylindrical symmetry and applies outside a line core. In cylindrical polar coordinates, we want a solution of

$$B - \lambda^2\nabla^2 B = 0$$

*This problem is somewhat difficult.

that is singular at the origin and for which the total flux is the flux quantum:

$$2\pi \int_0^\infty d\rho\, \rho B(\rho) = \Phi_0 \ .$$

The equation is in fact valid only outside the normal core of radius ξ. (b) Show that the solution has the limits

$$B(\rho) \simeq (\Phi_0/2\pi\lambda^2) \ln(\lambda/\rho) \ , \qquad (\xi \ll \rho \ll \lambda)$$

$$B(\rho) \simeq (\Phi_0/2\pi\lambda^2)(\pi\lambda/2\rho)^{1/2} \exp(-\rho/\lambda) \ . \qquad (\rho \gg \lambda)$$

5. ***London penetration depth.*** (a) Take the time derivative of the London equation (10) to show that $\partial \mathbf{j}/\partial t = (c^2/4\pi\lambda_L^2)\mathbf{E}$. (b) If $m d\mathbf{v}/dt = q\mathbf{E}$, as for free carriers of charge q and mass m, show that $\lambda_L^2 = mc^2/4\pi nq^2$.

6. ***Diffraction effect of Josephson junction.*** Consider a junction of rectangular cross section with a magnetic field B applied in the plane of the junction, normal to an edge of width w. Let the thickness of the junction be T. Assume for convenience that the phase difference of the two superconductors is $\pi/2$ when $B = 0$. Show that the dc current in the presence of the magnetic field is

$$J \approx J_0 \frac{\sin(wTBe/\hbar c)}{(wTBe/\hbar c)} \ .$$

7. ***Meissner effect in sphere.*** Consider a sphere of a type I superconductor with critical field H_c. (a) Show that in the Meissner regime the effective magnetization M within the sphere is given by $-8\pi M/3 = B_a$, the uniform applied magnetic field. (b) Show that the magnetic field at the surface of the sphere in the equatorial plane is $3B_a/2$. (It follows that the applied field at which the Meissner effect starts to break down is $2H_c/3$.) Reminder: The demagnetization field of a uniformly magnetized sphere is $-4\pi M/3$.

Reference

An excellent superconductor review is the website superconductors.org.

11

Diamagnetism and Paramagnetism

NOTATION: In the problems treated in this chapter the magnetic field B is always closely equal to the applied field B_a, so that we write B for B_a in most instances.

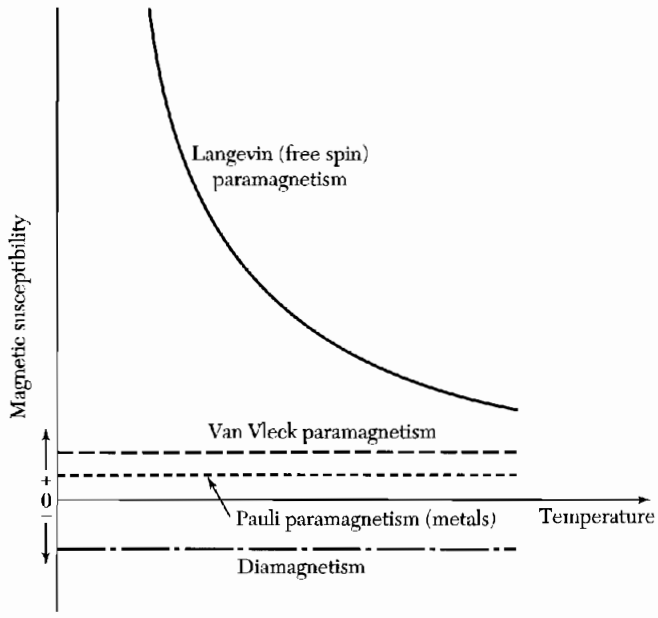


Figure 1 Characteristic magnetic susceptibilities of diamagnetic and paramagnetic substances.

Magnetism is inseparable from quantum mechanics, for a strictly classical system in thermal equilibrium can display no magnetic moment, even in a magnetic field. The magnetic moment of a free atom has three principal sources: the spin with which electrons are endowed; their orbital angular momentum about the nucleus; and the change in the orbital moment induced by an applied magnetic field.

The first two effects give paramagnetic contributions to the magnetization, and the third gives a diamagnetic contribution. In the ground $1s$ state of the hydrogen atom the orbital moment is zero, and the magnetic moment is that of the electron spin along with a small induced diamagnetic moment. In the $1s^2$ state of helium the spin and orbital moments are both zero, and there is only an induced moment. Atoms with all filled electron shells have zero spin and zero orbital moment: finite moments are associated with unfilled shells.

The magnetization M is defined as the magnetic moment per unit volume. The **magnetic susceptibility** per unit volume is defined as

$$\text{(CGS)} \quad \chi = \frac{M}{B}\ ; \qquad \text{(SI)} \quad \chi = \frac{\mu_0 M}{B} \tag{1}$$

where B is the macroscopic magnetic field intensity. In both systems of units χ is dimensionless. We shall sometimes for convenience refer to M/B as the susceptibility without specifying the system of units.

Quite frequently a susceptibility is defined referred to unit mass or to a mole of the substance. The molar susceptibility is written as χ_M; the magnetic moment per gram is sometimes written as σ. Substances with a negative magnetic susceptibility are called **diamagnetic**. Substances with a positive susceptibility are called **paramagnetic**, as in Fig. 1.

Ordered arrays of magnetic moments are discussed in Chapter 12; the arrays may be ferromagnetic, ferrimagnetic, antiferromagnetic, helical, or more complex in form. Nuclear magnetic moments give rise to **nuclear paramagnetism**. Magnetic moments of nuclei are of the order of 10^{-3} times smaller than the magnetic moment of the electron.

LANGEVIN DIAMAGNETISM EQUATION

Diamagnetism is associated with the tendency of electrical charges partially to shield the interior of a body from an applied magnetic field. In electromagnetism we are familiar with Lenz's law: when the flux through an electrical circuit is changed, an induced (diamagnetic) current is set up in such a direction as to oppose the flux change.

In a superconductor or in an electron orbit within an atom, the induced current persists as long as the field is present. The magnetic field of the induced current is opposite to the applied field, and the magnetic moment associated with the current is a diamagnetic moment. Even in a normal metal there is a diamagnetic contribution from the conduction electrons, and this diamagnetism is not destroyed by collisions of the electrons.

The usual treatment of the diamagnetism of atoms and ions employs the Larmor theorem: In a magnetic field the motion of the electrons around a central nucleus is, to the first order in B, the same as a possible motion in the absence of B except for the superposition of a precession of the electrons with angular frequency

$$\text{(CGS)} \quad \omega = eB/2mc \ ; \qquad \text{(SI)} \quad \omega = eB/2m. \tag{2}$$

If the field is applied slowly, the motion in the rotating reference system will be the same as the original motion in the rest system before the application of the field.

If the average electron current around the nucleus is zero initially, the application of the magnetic field will cause a finite current around the nucleus. The current is equivalent to a magnetic moment opposite to the applied field. It is assumed that the Larmor frequency (2) is much lower than the frequency of the original motion in the central field. This condition is not satisfied in free carrier cyclotron resonance, and the cyclotron frequency of the carriers is twice the frequency (2).

The Larmor precession of $\mathbf{Z}$ electrons is equivalent to an electric current

$$\text{(SI)} \qquad I = (\text{charge})(\text{revolutions per unit time}) = (-Ze)\left(\frac{1}{2\pi} \cdot \frac{eB}{2m}\right). \tag{3}$$

The magnetic moment μ of a current loop is given by the product (current) $\times$ (area of the loop). The area of the loop of radius ρ is $\pi\rho^2$. We have

$$\text{(SI)} \quad \mu = -\frac{Ze^2B}{4m}\langle\rho^2\rangle \ , \qquad \text{(CGS)} \quad \mu = -\frac{Ze^2B}{4mc^2}\langle\rho^2\rangle \ . \tag{4}$$

Here $\langle\rho^2\rangle = \langle x^2\rangle + \langle y^2\rangle$ is the mean square of the perpendicular distance of the electron from the field axis through the nucleus. The mean square distance of the electrons from the nucleus is $\langle r^2\rangle = \langle x^2\rangle + \langle y^2\rangle + \langle z^2\rangle$. For a spherically symmetrical distribution of charge we have $\langle x^2\rangle = \langle y^2\rangle = \langle z^2\rangle$, so that $\langle r^2\rangle = \frac{3}{2}\langle\rho^2\rangle$.

From (4) the diamagnetic susceptibility per unit volume is, if N is the number of atoms per unit volume,

(CGS) $$\chi = \frac{N\mu}{B} = -\frac{NZe^2}{6mc^2}\langle r^2\rangle \ ; \tag{5}$$

(SI) $$\chi = \frac{\mu_0 N\mu}{B} = -\frac{\mu_0 NZe^2}{6m}\langle r^2\rangle .$$

This is the classical Langevin result.

The problem of calculating the diamagnetic susceptibility of an isolated atom is reduced to the calculation of $\langle r^2\rangle$ for the electron distribution within the atom. The distribution can be calculated by quantum mechanics.

Experimental values for neutral atoms are most easily obtained for the inert gases. Typical experimental values of the molar susceptibilities are the following:

	He	Ne	Ar	Kr	Xe
χ_M in CGS in 10^{-6} cm^3/mole:	−1.9	−7.2	−19.4	−28.0	−43.0

In dielectric solids the diamagnetic contribution of the ion cores is described roughly by the Langevin result. The contribution of conduction electrons in metals is more complicated, as is evident from the de Haas-van Alphen effect discussed in Chapter 9.

QUANTUM THEORY OF DIAMAGNETISM OF MONONUCLEAR SYSTEMS

We give the quantum treatment of the classical Langevin result. From Appendix (G.18) the effect of a magnetic field is to add to the Hamiltonian the terms

$$\mathcal{H} = \frac{ie\hbar}{2mc}(\nabla\cdot\mathbf{A} + \mathbf{A}\cdot\nabla) + \frac{e^2}{2mc^2}A^2 \ ; \tag{6}$$

for an atomic electron these terms may usually be treated as a small perturbation. If the magnetic field is uniform and in the z direction, we may write

$$A_x = -\tfrac{1}{2}yB \ , \qquad A_y = \tfrac{1}{2}xB \ , \qquad A_z = 0 \ , \tag{7}$$

and (6) becomes

$$\mathcal{H} = \frac{ie\hbar B}{2mc}\left(x\frac{\partial}{\partial y} - y\frac{\partial}{\partial x}\right) + \frac{e^2B^2}{8mc^2}(x^2 + y^2) \ . \tag{8}$$

The first term on the right is proportional to the orbital angular momentum component L_z if $\mathbf{r}$ is measured from the nucleus. In mononuclear systems

this term gives rise only to paramagnetism. The second term gives for a spherically symmetric system a contribution

$$E' = \frac{e^2B^2}{12mc^2}\langle r^2\rangle \ , \tag{9}$$

by first-order perturbation theory. The associated magnetic moment is diamagnetic:

$$\mu = -\frac{\partial E'}{\partial B} = -\frac{e^2\langle r^2\rangle}{6mc^2}B \ , \tag{10}$$

in agreement with the classical result (5).

PARAMAGNETISM

Electronic paramagnetism (positive contribution to χ) is found in:

1. Atoms, molecules, and lattice defects possessing an odd number of electrons, as here the total spin of the system cannot be zero. Examples: free sodium atoms; gaseous nitric oxide (NO); organic free radicals such as triphenylmethyl, $C(C_6H_5)_3$; F centers in alkali halides.

2. Free atoms and ions with a partly filled inner shell: transition elements; ions isoelectronic with transition elements; rare earth and actinide elements. Examples: Mn^{2+}, Gd^{3+}, U^{4+}. Paramagnetism is exhibited by many of these ions even when incorporated into solids, but not invariably.

3. A few compounds with an even number of electrons, including molecular oxygen and organic biradicals.

4. Metals.

QUANTUM THEORY OF PARAMAGNETISM

The magnetic moment of an atom or ion in free space is given by

$$\boxed{\boldsymbol{\mu} = \gamma\hbar\mathbf{J} = -g\mu_B\mathbf{J} \ ,} \tag{11}$$

where the total angular momentum $\hbar\mathbf{J}$ is the sum of the orbital $\hbar\mathbf{L}$ and spin $\hbar\mathbf{S}$ angular momenta.

The constant γ is the ratio of the magnetic moment to the angular momentum; γ is called the **gyromagnetic ratio** or **magnetogyric ratio**. For electronic systems a quantity g called the g factor or the spectroscopic splitting factor is defined by

$$g\mu_B \equiv -\gamma\hbar \ . \tag{12}$$

For an electron spin $g = 2.0023$, usually taken as 2.00. For a free atom the g factor is given by the Landé equation

$$g = 1 + \frac{J(J+1) + S(S+1) - L(L+1)}{2J(J+1)} \ . \tag{13}$$

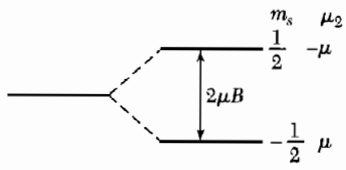


Figure 2 Energy level splitting for one electron in a magnetic field B directed along the positive z axis. For an electron the magnetic moment μ is opposite in sign to the spin S, so that $\mu = -g\mu_B S$. In the low-energy state the magnetic moment is parallel to the magnetic field.

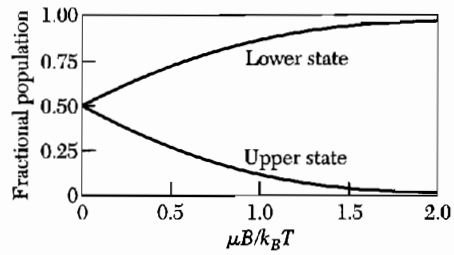


Figure 3 Fractional populations of a two-level system in thermal equilibrium at temperature T in a magnetic field B. The magnetic moment is proportional to the difference between the two curves.

The **Bohr magneton** μ_B is defined as $e\hbar/2mc$ in CGS and $e\hbar/2m$ in SI. It is closely equal to the spin magnetic moment of a free electron.

The energy levels of the system in a magnetic field are

$$U = -\boldsymbol{\mu} \cdot \mathbf{B} = m_J g \mu_B B \ , \tag{14}$$

where m_J is the azimuthal quantum number and has the values $J, J-1, \ldots, -J$. For a single spin with no orbital moment we have $m_J = \pm\frac{1}{2}$ and $g = 2$, whence $U = \pm\mu_B B$. This splitting is shown in Fig. 2.

If a system has only two levels the equilibrium populations are, with $\tau \equiv k_B T$,

$$\frac{N_1}{N} = \frac{\exp(\mu B/\tau)}{\exp(\mu B/\tau) + \exp(-\mu B/\tau)} \ ; \tag{15}$$

$$\frac{N_2}{N} = \frac{\exp(-\mu B/\tau)}{\exp(\mu B/\tau) + \exp(-\mu B/\tau)} \ ; \tag{16}$$

here N_1, N_2 are the populations of the lower and upper levels, and $N = N_1 + N_2$ is the total number of atoms. The fractional populations are plotted in Fig. 3.

The projection of the magnetic moment of the upper state along the field direction is $-\mu$ and of the lower state is μ. The resultant magnetization for N atoms per unit volume is, with $x = \mu B/k_B T$,

$$M = (N_1 - N_2)\mu = N\mu \cdot \frac{e^x - e^{-x}}{e^x + e^{-x}} = N\mu \tanh x \ . \tag{17}$$

For $x \ll 1$, $\tanh x \cong x$, and we have

$$M \cong N\mu(\mu B/k_B T) \ . \tag{18}$$

In a magnetic field an atom with angular momentum quantum number J has $2J + 1$ equally spaced energy levels. The magnetization (Fig. 4) is given by

$$M = NgJ\mu_B B_J(x) \ , \qquad (x = gJ\mu_B B/k_B T) \ , \tag{19}$$

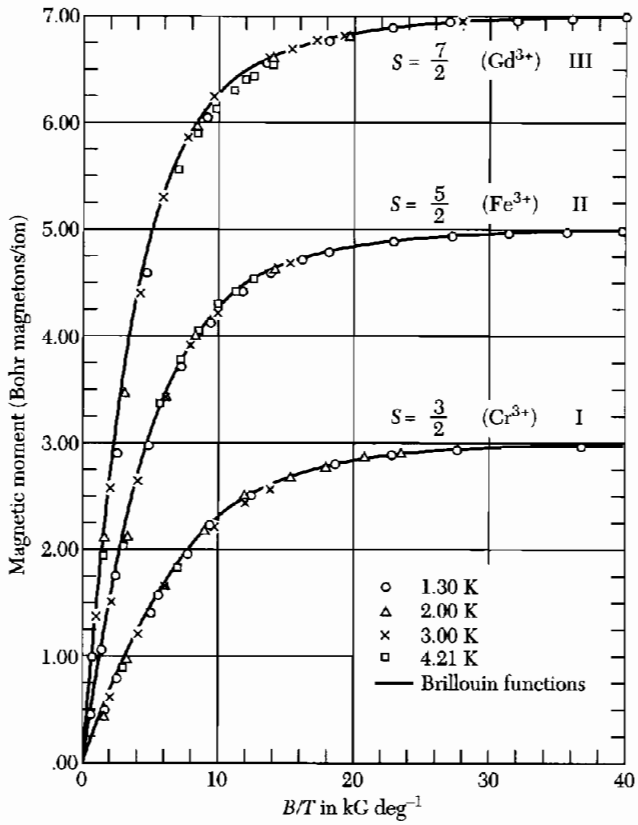


Figure 4 Plot of magnetic moment versus B/T for spherical samples of (I) potassium chromium alum, (II) ferric ammonium alum, and (III) gadolinium sulfate octahydrate. Over 99.5% magnetic saturation is achieved at 1.3 K and about 50,000 gauss (5T). After W. E. Henry.

where the **Brillouin function** B_J is defined by

$$B_J(x) = \frac{2J+1}{2J}\,\mathrm{ctnh}\left(\frac{(2J+1)x}{2J}\right) - \frac{1}{2J}\,\mathrm{ctnh}\left(\frac{x}{2J}\right) . \tag{20}$$

Equation (17) is a special case of (20) for $J = \frac{1}{2}$.

For $x = \mu B/k_B T \ll 1$, we have

$$\mathrm{ctnh}\, x = \frac{1}{x} + \frac{x}{3} - \frac{x^3}{45} + \cdots , \tag{21}$$

and the susceptibility is

$$\frac{M}{B} \cong \frac{NJ(J+1)g^2\mu_B^2}{3k_B T} = \frac{Np^2\mu_B^2}{3k_B T} = \frac{C}{T} . \tag{22}$$

Here p is the **effective number of Bohr magnetons**, defined as

$$p \equiv g[J(J+1)]^{1/2} . \tag{23}$$

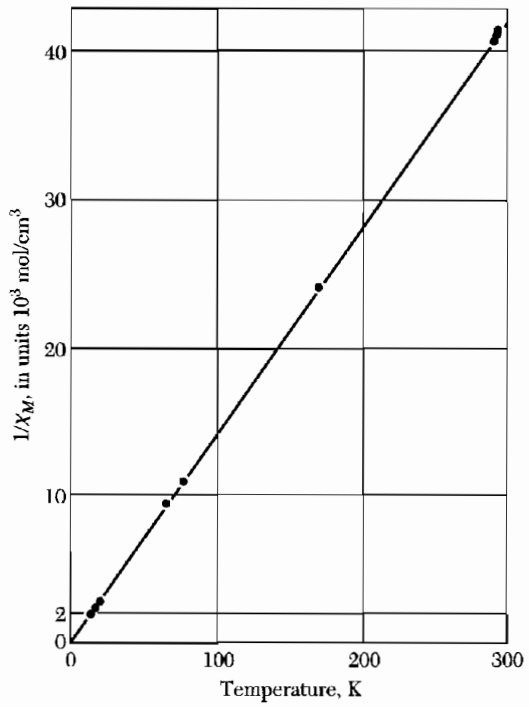


Figure 5 Plot of $1/\chi$ *vs* T for a gadolinium salt, $Gd(C_2H_3SO_4)_3 \cdot 9H_2O$. The straight line is the Curie law. (After L. C. Jackson and H. Kamerlingh Onnes.)

The constant C is known as the **Curie constant**. The form (19) is known as the Curie-Brillouin law, and (22) is known as the **Curie law**. Results for the paramagnetic ions in a gadolinium salt are shown in Fig. 5.

Rare Earth Ions

The ions of the rare earth elements (Table 1) have closely similar chemical properties, and their chemical separation in tolerably pure form was accomplished only long after their discovery. Their magnetic properties are fascinating: The ions exhibit a systematic variety and intelligible complexity. The chemical properties of the trivalent ions are similar because the outermost electron shells are identically in the $5s^25p^6$ configuration, like neutral xenon. In lanthanum, just before the rare earth group begins, the $4f$ shell is empty; at cerium there is one $4f$ electron, and the number of $4f$ electrons increases steadily through the group until we have $4f^{13}$ at ytterbium and the filled shell $4f^{14}$ at lutecium. The radii of the trivalent ions contract fairly smoothly as we go through the group from 1.11 Å at cerium to 0.94 Å at ytterbium. This is known as the "lanthanide contraction." What distinguishes the magnetic behavior of one ion species from another is the number of $4f$ electrons compacted in the inner shell with a radius of perhaps 0.3 Å. Even in the metals the

Table 1 Effective magneton numbers p for trivalent lanthanide group ions

		(Near room temperature)		
Ion	Configuration	Basic level	p(calc) = $g[J(J+1)]^{1/2}$	p(exp), approximate
Ce^{3+}	$4f^1 5s^2 p^6$	$^2F_{5/2}$	2.54	2.4
Pr^{3+}	$4f^2 5s^2 p^6$	$^3H_4$	3.58	3.5
Nd^{3+}	$4f^3 5s^2 p^6$	$^4I_{9/2}$	3.62	3.5
Pm^{3+}	$4f^4 5s^2 p^6$	$^5I_4$	2.68	—
Sm^{3+}	$4f^5 5s^2 p^6$	$^6H_{5/2}$	0.84	1.5
Eu^{3+}	$4f^6 5s^2 p^6$	$^7F_0$	0	3.4
Gd^{3+}	$4f^7 5s^2 p^6$	$^8S_{7/2}$	7.94	8.0
Tb^{3+}	$4f^8 5s^2 p^6$	$^7F_6$	9.72	9.5
Dy^{3+}	$4f^9 5s^2 p^6$	$^6H_{15/2}$	10.63	10.6
Ho^{3+}	$4f^{10} 5s^2 p^6$	$^5I_8$	10.60	10.4
Er^{3+}	$4f^{11} 5s^2 p^6$	$^4I_{15/2}$	9.59	9.5
Tm^{3+}	$4f^{12} 5s^2 p^6$	$^3H_6$	7.57	7.3
Yb^{3+}	$4f^{13} 5s^2 p^6$	$^2F_{7/2}$	4.54	4.5

$4f$ core retains its integrity and its atomic properties: no other group of elements in the periodic table is as interesting.

The preceding discussion of paramagnetism applies to atoms that have a $(2J + 1)$-fold degenerate ground state, the degeneracy being lifted by a magnetic field. The influence of all higher energy states of the system is neglected. These assumptions appear to be satisfied by a number of rare-earth ions, Table 1. The calculated magneton numbers are obtained with g values from the Landé result (13) and the ground-state level assignment predicted below by the Hund theory of spectral terms. The discrepancy between the experimental magneton numbers and those calculated on these assumptions is quite marked for Eu^{3+} and Sm^{3+} ions. For these ions it is necessary to consider the influence of the high states of the $L - S$ multiplet, as the intervals between successive states of the multiplet are not large compared to $k_B T$ at room temperature. A multiplet is the set of levels of different J values arising out of a given L and S. The levels of a multiplet are split by the spin-orbit interaction.

Hund Rules

The Hund rules as applied to electrons in a given shell of an atom affirm that electrons will occupy orbitals in such a way that the ground state is characterized by the following:

1. The maximum value of the total spin S allowed by the exclusion principle;

2. The maximum value of the orbital angular momentum L consistent with this value of S;

3. The value of the total angular momentum J is equal to $|L - S|$ when the shell is less than half full and to $L + S$ when the shell is more than half full. When the shell is just half full, the application of the first rule gives $L = 0$, so that $J = S$.

The first Hund rule has its origin in the exclusion principle and the coulomb repulsion between electrons. The exclusion principle prevents two electrons of the same spin from being at the same place at the same time. Thus electrons of the same spin are kept apart, further apart than electrons of opposite spin. Because of the coulomb interaction the energy of electrons of the same spin is lower—the average potential energy is less positive for parallel spin than for antiparallel spin. A good example is the ion Mn^{2+}. This ion has five electrons in the $3d$ shell, which is therefore half-filled. The spins can all be parallel if each electron enters a different orbital, and there are exactly five different orbitals available, characterized by the orbital quantum numbers $m_L = 2, 1, 0, -1, -2$. These will be occupied by one electron each. We expect $S = \frac{5}{2}$, and because $\Sigma m_L = 0$ the only possible value of L is 0, as observed.

The second Hund rule is best approached by model calculations. Pauling and Wilson,[1] for example, give a calculation of the spectral terms that arise from the configuration p^2. The third Hund rule is a consequence of the sign of the spin-orbit interaction: For a single electron the energy is lowest when the spin is antiparallel to the orbital angular momentum. But the low-energy pairs m_L, m_S are progressively used up as we add electrons to the shell; by the exclusion principle when the shell is more than half full the state of lowest energy necessarily has the spin parallel to the orbit.

Consider two examples of the Hund rules: The ion Ce^{3+} has a single f electron; an f electron has $l = 3$ and $s = \frac{1}{2}$. Because the f shell is less than half full, the J value by the preceding rule is $|L - S| = L - \frac{1}{2} = \frac{5}{2}$. The ion Pr^{3+} has two f electrons; one of the rules tells us that the spins add to give $S = 1$. Both f electrons cannot have $m_l = 3$ without violating the Pauli exclusion principle, so that the maximum L consistent with the Pauli principle is not 6, but 5. The J value is $|L - S| = 5 - 1 = 4$.

Iron Group Ions

Table 2 shows that the experimental magneton numbers for salts of the iron transition group of the periodic table are in poor agreement with (23). The values often agree quite well with magneton numbers $p = 2[S(S + 1)]^{1/2}$ calculated as if the orbital moment were not there at all.

Crystal Field Splitting

The difference in behavior of the rare earth and the iron group salts is that the $4f$ shell responsible for paramagnetism in the rare earth ions lies deep

[1] L. Pauling and E. B. Wilson, *Introduction to quantum mechanics*, McGraw-Hill, 1935, pp. 239–246. See also Dover Reprint.

Table 2 Effective magneton numbers for iron group ions

Ion	Configuration	Basic level	p(calc) = $g[J(J+1)]^{1/2}$	p(calc) = $2[S(S+1)]^{1/2}$	p(exp)[a]
Ti^{3+}, V^{4+}	$3d^1$	$^2D_{3/2}$	1.55	1.73	1.8
V^{3+}	$3d^2$	$^3F_2$	1.63	2.83	2.8
Cr^{3+}, V^{2+}	$3d^3$	$^4F_{3/2}$	0.77	3.87	3.8
Mn^{3+}, Cr^{2+}	$3d^4$	$^5D_0$	0	4.90	4.9
Fe^{3+}, Mn^{2+}	$3d^5$	$^6S_{5/2}$	5.92	5.92	5.9
Fe^{2+}	$3d^6$	$^5D_4$	6.70	4.90	5.4
Co^{2+}	$3d^7$	$^4F_{9/2}$	6.63	3.87	4.8
Ni^{2+}	$3d^8$	$^3F_4$	5.59	2.83	3.2
Cu^{2+}	$3d^9$	$^2D_{5/2}$	3.55	1.73	1.9

[a] Representative values.

inside the ions, within the 5*s* and 5*p* shells, whereas in the iron group ions the 3*d* shell responsible for paramagnetism is the outermost shell. The 3*d* shell experiences the intense inhomogeneous electric field produced by neighboring ions. This inhomogeneous electric field is called the **crystal field**. The interaction of the paramagnetic ions with the crystal field has two major effects: The coupling of **L** and **S** vectors is largely broken up, so that the states are no longer specified by their J values; further, the $2L + 1$ sublevels belonging to a given L which are degenerate in the free ion may now be split by the crystal field, as in Fig. 6. This splitting diminishes the contribution of the orbital motion to the magnetic moment.

Quenching of the Orbital Angular Momentum

In an electric field directed toward a fixed nucleus, the plane of a classical orbit is fixed in space, so that all the orbital angular momentum components L_x, L_y, L_z are constant. In quantum theory one angular momentum component, usually taken as L_z, and the square of the total orbital angular momentum L^2 are constant in a central field. In a noncentral field the plane of the orbit will move about; the angular momentum components are no longer constant and may average to zero. In a crystal L_z will no longer be a constant of the motion, although to a good approximation L^2 may continue to be constant. When L_z averages to zero, the orbital angular momentum is said to be quenched. The magnetic moment of a state is given by the average value of the magnetic moment operator $\mu_B(\mathbf{L} + 2\mathbf{S})$. In a magnetic field along the z direction the orbital contribution to the magnetic moment is proportional to the quantum expectation value of L_z; the orbital magnetic moment is quenched if the mechanical moment L_z is quenched.

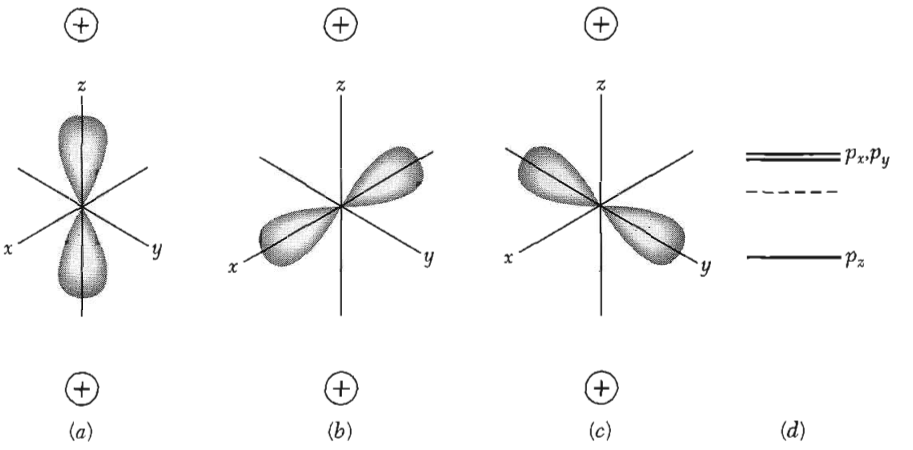


Figure 6 Consider an atom with orbital angular momentum $L = 1$ placed in the uniaxial crystalline electric field of the two positive ions along the z axis. In the free atom the states $m_L = \pm 1, 0$ have identical energies—they are degenerate. In the crystal the atom has a lower energy when the electron cloud is close to positive ions as in (a) than when it is oriented midway between them, as in (b) and (c). The wavefunctions that give rise to these charge densities are of the form $zf(r)$, $xf(r)$ and $yf(r)$ and are called the p_z, p_x, p_y orbitals, respectively. In an axially symmetric field, as shown, the p_x and p_y orbitals are degenerate. The energy levels referred to the free atom (dotted line) are shown in (d). If the electric field does not have axial symmetry, all three states will have different energies.

As an example, consider a single electron with orbital quantum number $L = 1$ moving about a nucleus, the whole being placed in an inhomogeneous crystalline electric field. We omit electron spin.

In a crystal of orthorhombic symmetry the charges on neighboring ions will produce an electrostatic potential φ about the nucleus of the form

$$e\varphi = Ax^2 + By^2 - (A + B)z^2 \ , \tag{24}$$

where A and B are constants. This expression is the lowest degree polynomial in x, y, z which is a solution of the Laplace equation $\nabla^2\varphi = 0$ and compatible with the symmetry of the crystal.

In free space the ground state is three-fold degenerate, with magnetic quantum numbers $m_L = 1, 0, -1$. In a magnetic field these levels are split by energies proportional to the field B, and it is this field-proportional splitting which is responsible for the normal paramagnetic susceptibility of the ion. In the crystal the picture may be different. We take as the three wavefunctions associated with the unperturbed ground state of the ion

$$U_x = xf(r) \ ; \qquad U_y = yf(r) \ ; \qquad U_z = zf(r) \ . \tag{25}$$

These wavefunctions are orthogonal, and we assume that they are normalized. Each of the U's can be shown to have the property

$$\mathcal{L}^2 U_i = L(L+1)U_i = 2U_i \ , \tag{26}$$

where $\mathcal{L}^2$ is the operator for the square of the orbital angular momentum, in units of $\hbar$. The result (26) confirms that the selected wavefunctions are in fact p functions, having $L = 1$.

We observe now that the U's are diagonal with respect to the perturbation, as by symmetry the nondiagonal elements vanish:

$$\langle U_x | e\varphi | U_y \rangle = \langle U_x | e\varphi | U_z \rangle = \langle U_y | e\varphi | U_z \rangle = 0 \ . \tag{27}$$

Consider for example,

$$\langle U_x | e\varphi | U_y \rangle = \int xy |f(r)|^2 \{Ax^2 + By^2 - (A+B)z^2\}\, dx\, dy\, dz \ ; \tag{28}$$

the integrand is an odd function of x (and also of y) and therefore the integral must be zero. The energy levels are then given by the diagonal matrix elements:

$$\begin{aligned} \langle U_x | e\varphi | U_x \rangle &= \int |f(r)|^2 \{Ax^4 + By^2x^2 - (A+B)z^2x^2\}\, dx\, dy\, dz \\ &= A(I_1 - I_2) \ , \end{aligned} \tag{29}$$

where

$$I_1 = \int |f(r)|^2 x^4\, dx\, dy\, dz \ ; \qquad I_2 = \int |f(r)|^2 x^2 y^2\, dx\, dy\, dz \ .$$

In addition,

$$\langle U_y | e\varphi | U_y \rangle = B(I_1 - I_2) \ ; \qquad \langle U_z | e\varphi | U_z \rangle = -(A+B)(I_1 - I_2) \ .$$

The three eigenstates in the crystal field are p functions with their angular lobes directed along each of the x, y, z axes, respectively.

The orbital moment of each of the levels is zero, because

$$\langle U_x | L_z | U_x \rangle = \langle U_y | L_z | U_y \rangle = \langle U_z | L_z | U_z \rangle = 0 \ .$$

The level still has a definite total angular momentum, since $\mathcal{L}^2$ is diagonal and gives $L = 1$, but the spatial components of the angular momentum are not constants of the motion and their time average is zero in the first approximation. Therefore the components of the orbital magnetic moment also vanish in the same approximation. The role of the crystal field in the quenching process is to split the originally degenerate levels into nonmagnetic levels separated by energies $\gg \mu H$, so that the magnetic field is a small perturbation in comparison with the crystal field.

At a lattice site of cubic symmetry there is no term in the potential of the form (24), that is, quadratic in the electron coordinates. Now the ground state of an ion with one p electron (or with one hole in a p shell) will be triply

degenerate. However, the energy of the ion will be lowered if the ion displaces itself with respect to the surroundings, thereby creating a noncubic potential such as (24). Such a spontaneous displacement is known as a **Jahn-Teller effect** and is often large and important, particularly with the Mn^{3+} and Cu^{2+} ions and with holes in alkali and silver halides.

Spectroscopic Splitting Factor

We suppose for convenience that the crystal field constants, A, B are such that $U_x = xf(r)$ is the orbital wave function of the ground state of the atom in the crystal. For a spin $S = \frac{1}{2}$ there are two possible spin states $S_z = \pm\frac{1}{2}$ represented by the spin functions α, β, which in the absence of a magnetic field are degenerate in the zeroth approximation. The problem is to take into account the spin-orbit interaction energy $\lambda \mathbf{L} \cdot \mathbf{S}$.

If the ground state function is $\psi_0 = U_x\alpha = xf(r)\alpha$ in the zeroth approximation, then in the first approximation, considering the $\lambda \mathbf{L} \cdot \mathbf{S}$ interaction by standard perturbation theory, we have

$$\psi = [U_x - i(\lambda/2\Delta_1)U_y]\alpha - i(\lambda/2\Delta_2)U_z\beta \ , \tag{30}$$

where Δ_1 is the energy difference between the U_x and U_y states, and Δ_2 is the difference between the U_x and U_z states. The term in $U_z\beta$ actually has only a second-order effect on the result and may be discarded. The expectation value of the orbital angular momentum to the first order is given directly by

$$(\psi|L_z|\psi) = -\lambda/\Delta_1 \ ,$$

and the magnetic moment of the state as measured in the z direction is

$$\mu_B(\psi|L_z + 2S_z|\psi) = [-(\lambda/\Delta_1) + 1]\mu_B \ .$$

As the separation between the levels $S_z = \pm\frac{1}{2}$ in a field H is

$$\Delta E \approx g\mu_B H = 2[1 - (\lambda/\Delta_1)]\mu_B H \ ,$$

the g value or spectroscopic splitting factor (12) in the z direction is

$$g = 2[1 - (\lambda/\Delta_1)] \ . \tag{31}$$

Van Vleck Temperature-Independent Paramagnetism

We consider an atomic or molecular system which has no magnetic moment in the ground state, by which we mean that the diagonal matrix element of the magnetic moment operator μ_z is zero.

Suppose that there is a nondiagonal matrix element $\langle s|\mu_z|0\rangle$ of the magnetic moment operator, connecting the ground state 0 with the excited state s of energy $\Delta = E_s - E_0$ above the ground state. Then by standard perturbation theory the wavefunction of the ground state in a weak field ($\mu_2 B \ll \Delta$) becomes

$$\psi_0' = \psi_0 + (B/\Delta)\langle s|\mu_z|0\rangle\psi_s \ , \tag{32}$$

and the wavefunction of the excited state becomes

$$\psi_s' = \psi_s - (B/\Delta)\langle 0|\mu_z|s\rangle\psi_0 \ . \tag{33}$$

The perturbed ground state now has a moment

$$\langle 0'|\mu_z|0'\rangle \cong 2B|\langle s|\mu_z|0\rangle|^2/\Delta \ , \tag{34}$$

and the upper state has a moment

$$\langle s'|\mu_z|s'\rangle \cong -2B|\langle s|\mu_z|0\rangle|^2/\Delta \ . \tag{35}$$

There are two interesting cases to consider:

Case (a). $\Delta \ll k_B T$. The surplus population in the ground state over the excited state is approximately equal to $N\Delta/2k_BT$, so that the resultant magnetization is

$$M = \frac{2B|\langle s|\mu_z|0\rangle|^2}{\Delta} \cdot \frac{N\Delta}{2k_BT} \ , \tag{36}$$

which gives for the susceptibility

$$\chi = N|\langle s|\mu_z|0\rangle|^2/k_BT \ . \tag{37}$$

Here N is the number of molecules per unit volume. This contribution is of the usual Curie form, although the mechanism of magnetization here is by polarization of the states of the system, whereas with free spins the mechanism of magnetization is the redistribution of ions among the spin states. We note that the splitting Δ does not enter in (37).

Case (b). $\Delta \gg k_B T$. Here the population is nearly all in the ground state, so that

$$M = \frac{2NB|\langle s|\mu_z|0\rangle|^2}{\Delta} \ . \tag{38}$$

The susceptibility is

$$\chi = \frac{2N|\langle s|\mu_z|0\rangle|^2}{\Delta} \ , \tag{39}$$

independent of temperature. This type of contribution is known as Van Vleck paramagnetism.

COOLING BY ISENTROPIC DEMAGNETIZATION

The first method for attaining temperatures much below 1 K was that of isentropic, or adiabatic, demagnetization of a paramagnetic salt. By its use, temperatures of 10^{-3} K and lower have been reached. The method rests on the fact that at a fixed temperature the entropy of a system of magnetic moments is lowered by the application of a magnetic field.

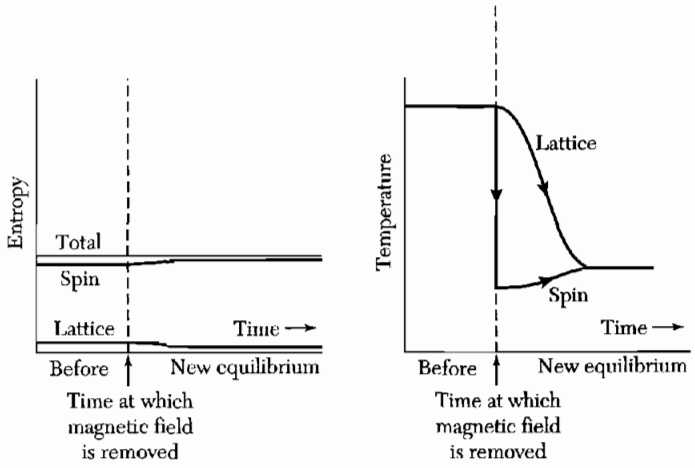


Figure 7 During isentropic demagnetization the total entropy of the specimen is constant. For effective cooling the initial entropy of the lattice should be small in comparison with the entropy of the spin system.

The entropy is a measure of the disorder of a system: the greater the disorder, the higher is the entropy. In the magnetic field the moments will be partly lined up (partly ordered), so that the entropy is lowered by the field. The entropy is also lowered if the temperature is lowered, as more of the moments line up.

If the magnetic field can then be removed without changing the entropy of the spin system, the order of the spin system will look like a lower temperature than the same degree of order in the presence of the field. When the specimen is demagnetized at constant entropy, entropy can flow into the spin system only from the system of lattice vibrations, as in Fig. 7. At the temperatures of interest the entropy of the lattice vibrations is usually negligible, thus the entropy of the spin system will be essentially constant during isentropic demagnetization of the specimen. Magnetic cooling is a one-shot operation, not cyclic.

We first find an expression for the spin entropy of a system of N ions, each of spin S, at a temperature sufficiently high that the spin system is entirely disordered. That is, T is supposed to be much higher than some temperature Δ which characterizes the energy of the interactions ($E_{\text{int}} \equiv k_B\Delta$) tending to orient the spins preferentially. Some of these interactions are discussed in Chapter 12. The definition of the entropy σ of a system of G accessible states is $\sigma = k_B \ln G$. At a temperature so high that all of the $2S + 1$ states of each ion are nearly equally populated, G is the number of ways of arranging N spins in $2S + 1$ states. Thus $G = (2S + 1)^N$, whence the spin entropy σ_S is:

$$\sigma_S = k_B \ln (2S + 1)^N = Nk_B \ln (2S + 1) \ . \tag{40}$$

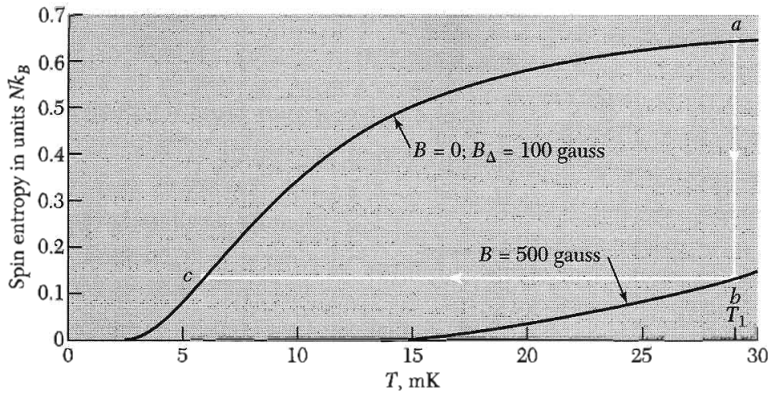


Figure 8 Entropy for a spin $\frac{1}{2}$ system as a function of temperature, assuming an internal random magnetic field B_Δ of 100 gauss. The specimen is magnetized isothermally along ab, and is then insulated thermally. The external magnetic field is turned off along bc. In order to keep the figure on a reasonable scale the initial temperature T_1 is lower than would be used in practice, and so is the external magnetic field.

This spin entropy is reduced by a magnetic field if the lower levels gain in population when the field separates the $2S + 1$ states in energy.

The steps carried out in the cooling process are shown in Fig. 8. The field is applied at temperature T_1 with the specimen in good thermal contact with the surroundings, giving the isothermal path ab. The specimen is then insulated ($\Delta\sigma = 0$) and the field removed, the specimen follows the constant entropy path bc, ending up at temperature T_2. The thermal contact at T_1 is provided by helium gas, and the thermal contact is broken by removing the gas with a pump.

The population of a magnetic sublevel is a function only of $\mu B/k_B T$, hence of B/T. The spin-system entropy is a function only of the population distribution; hence the spin entropy is a function only of B/T. If B_Δ is the effective field that corresponds to the local interactions, the final temperature T_2 reached in an isentropic demagnetization experiment is

$$\boxed{T_2 = T_1(B_\Delta/B)\ ,} \tag{41}$$

where B is the initial field and T_1 the initial temperature.

Nuclear Demagnetization

Because nuclear magnetic moments are weak, nuclear magnetic interactions are much weaker than similar electronic interactions. We expect to reach a temperature 100 times lower with a nuclear paramagnet than with an electron paramagnet. The initial temperature T_1 of the nuclear stage in a nuclear spin-cooling experiment must be lower than in an electron spin-cooling

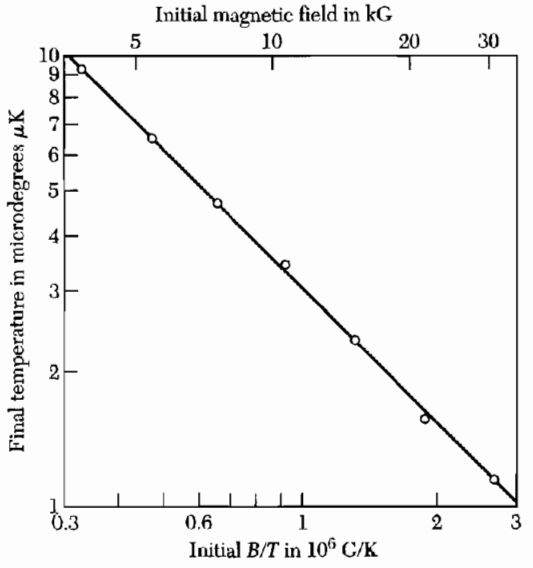


Figure 9 Nuclear demagnetizations of copper nuclei in the metal, starting from 0.012 K and various fields. (After M. V. Hobden and N. Kurti.)

experiment. If we start at $B = 50$ kG and $T_1 = 0.01$ K, then $\mu B/k_B T_1 \approx 0.5$, and the entropy decrease on magnetization is over 10 percent of the maximum spin entropy. This is sufficient to overwhelm the lattice and from (41) we estimate a final temperature $T_2 \approx 10^{-7}$ K. The first nuclear cooling experiment was carried out on Cu nuclei in the metal, starting from a first stage at about 0.02 K as attained by electronic cooling. The lowest temperature reached was 1.2×10^{-6} K.

The results in Fig. 9 fit a line of the form of (41): $T_2 = T_1(3.1/B)$ with B in gauss, so that $B_\Delta = 3.1$ gauss. This is the effective interaction field of the magnetic moments of the Cu nuclei. The motivation for using nuclei in a metal is that conduction electrons help ensure rapid thermal contact of lattice and nuclei at the temperature of the first stage.

PARAMAGNETIC SUSCEPTIBILITY OF CONDUCTION ELECTRONS

> We are going to try to show how on the basis of these statistics the fact that many metals are diamagnetic, or only weakly paramagnetic, can be brought into agreement with the existence of a magnetic moment of the electrons.
>
> W. Pauli, 1927

Classical free electron theory gives an unsatisfactory account of the paramagnetic susceptibility of the conduction electrons. An electron has associated with it a magnetic moment of one Bohr magneton, μ_B. One might expect that

the conduction electrons would make a Curie-type paramagnetic contribution (22) to the magnetization of the metal: $M = N\mu_B^2 B/k_B T$. Instead it is observed that the magnetization of most normal nonferromagnetic metals is independent of temperature.

Pauli showed that the application of the Fermi-Dirac distribution (Chapter 6) would correct the theory as required. We first give a qualitative explanation. The result (18) tells us that the probability an atom will be lined up parallel to the field B exceeds the probability of the antiparallel orientation by roughly $\mu B/k_B T$. For N atoms per unit volume, this gives a net magnetization $\approx N\mu^2 B/k_B T$, the standard result.

Most conduction electrons in a metal, however, have no possibility of turning over when a field is applied, because most orbitals in the Fermi sea with parallel spin are already occupied. Only the electrons within a range $k_B T$ of the top of the Fermi distribution have a chance to turn over in the field; thus only the fraction T/T_F of the total number of electrons contribute to the susceptibility. Hence

$$M \approx \frac{N\mu^2 B}{k_B T} \cdot \frac{T}{T_F} = \frac{N\mu^2}{k_B T_F} B \ , \tag{42}$$

which is independent of temperature and of the observed order of magnitude.

We now calculate the expression for the paramagnetic susceptibility of a free electron gas at $T \ll T_F$. We follow the method of calculation suggested by Fig. 10. An alternate derivation is the subject of Problem 5.

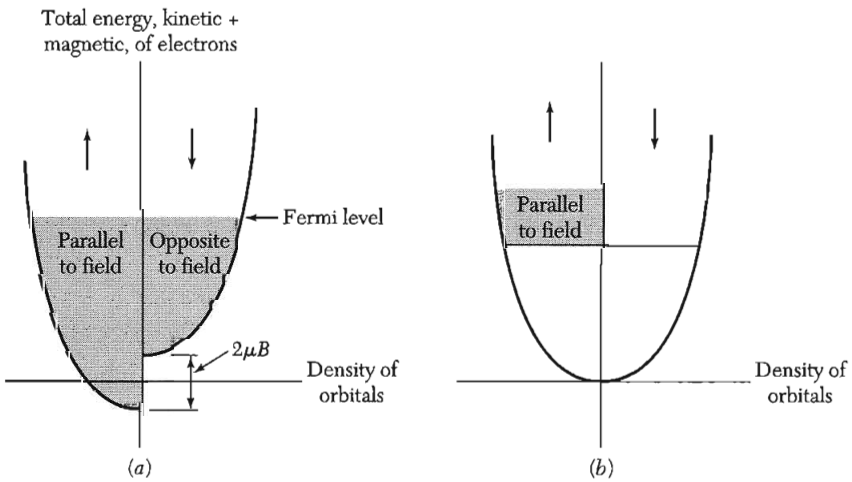


Figure 10 Pauli paramagnetism at absolute zero; the orbitals in the shaded regions in (a) are occupied. The numbers of electrons in the "up" and "down" band will adjust to make the energies equal at the Fermi level. The chemical potential (Fermi level) of the moment up electrons is equal to that of the moment down electrons. In (b) we show the excess of moment up electrons in the magnetic field.

The concentration of electrons with magnetic moments parallel to the magnetic field is

$$N_+ = \frac{1}{2}\int_{-\mu B}^{\epsilon_F} d\epsilon\, D(\epsilon + \mu B) \cong \frac{1}{2}\int_0^{\epsilon_F} d\epsilon\, D(\epsilon) + \frac{1}{2}\mu B\, D(\epsilon_F) \ , \tag{43}$$

written for absolute zero. Here $\frac{1}{2}D(\epsilon + \mu B)$ is the density of orbitals of one spin orientation, with allowance for the downward shift of energy by $-\mu B$. The approximation is written for $k_B T \ll \epsilon_F$.

The concentration of electrons with magnetic moments antiparallel to the magnetic field is

$$N_- = \frac{1}{2}\int_{\mu B}^{\epsilon_F} d\epsilon\, D(\epsilon - \mu B) \cong \frac{1}{2}\int_0^{\epsilon_F} d\epsilon\, D(\epsilon) - \frac{1}{2}\mu B\, D(\epsilon_F) \ . \tag{44}$$

The magnetization is given by $M = \mu(N_+ - N_-)$, so that

$$M = \mu^2 D(\epsilon_F) B = \frac{3N\mu^2}{2k_B T_F} B \ , \tag{45}$$

with $D(\epsilon_F) = 3N/2\epsilon_F = 3N/2k_B T_F$ from Chapter 6. The result (45) gives the **Pauli spin magnetization** of the conduction electrons, for $k_B T \ll \epsilon_F$.

In deriving the paramagnetic susceptibility, we have supposed that the spatial motion of the electrons is not affected by the magnetic field. But the wavefunctions are modified by the magnetic field; Landau has shown that for free electrons this causes a diamagnetic moment equal to $-\frac{1}{3}$ of the paramagnetic moment. Thus the total magnetization of a free electron gas is

$$M = \frac{N\mu_B^2}{k_B T_F} B \ . \tag{46}$$

Before comparing (46) with the experiment we must take account of the diamagnetism of the ionic cores, of band effects, and of electron-electron interactions. In sodium the interaction effects increase the spin susceptibility by perhaps 75 percent.

The magnetic susceptibility is considerably higher for most transition metals (with unfilled inner electron shells) than for the alkali metals. The high values suggest that the density of orbitals is unusually high for transition metals, in agreement with measurements of the electronic heat capacity. We saw in Chapter 9 how this arises from band theory.

SUMMARY

(In CGS Units)

- The diamagnetic susceptibility of N atoms of atomic number Z is $\chi = -Ze^2N\langle r^2\rangle/6mc^2$, where $\langle r^2\rangle$ is the mean square atomic radius. (Langevin)
- Atoms with a permanent magnetic moment $\boldsymbol{\mu}$ have a paramagnetic susceptibility $\chi = N\mu^2/3k_B T$, for $\mu B \ll k_B T$. (Curie-Langevin)

- For a system of spins $S = \frac{1}{2}$, the exact magnetization is $M = N\mu \tanh(\mu B/k_B T)$, where $\mu = \frac{1}{2}g\mu_B$. (Brillouin)
- The ground state of electrons in the same shell have the maximum value of S allowed by the Pauli principle and the maximum L consistent with this S. The J value is $L + S$ if the shell is more than half full and $|L - S|$ if the shell is less than half full.
- A cooling process operates by demagnetization of a paramagnetic salt at constant entropy. The final temperature reached is of the order of $(B_\Delta/B)T_{\text{initial}}$, where B_Δ is the effective local field and B is the initial applied magnetic field.
- The paramagnetic susceptibility of a Fermi gas of conduction electrons is $\chi = 3N\mu^2/2\epsilon_F$, independent of temperature for $k_B T \ll \epsilon_F$. (Pauli)

Problems

1. ***Diamagnetic susceptibility of atomic hydrogen.*** The wave function of the hydrogen atom in its ground state $(1s)$ is $\psi = (\pi a_0^3)^{-1/2}\exp(-r/a_0)$, where $a_0 = \hbar^2/me^2 = 0.529 \times 10^{-8}$ cm. The charge density is $\rho(x, y, z) = -e|\psi|^2$, according to the statistical interpretation of the wave function. Show that for this state $\langle r^2\rangle = 3a_0^2$, and calculate the molar diamagnetic susceptibility of atomic hydrogen $(-2.36 \times 10^{-6}\ \text{cm}^3/\text{mole})$.

2. ***Hund rules.*** Apply the Hund rules to find the ground state (the basic level in the notation of Table 1) of (a) Eu^{++}, in the configuration $4f^7\,5s^2p^6$; (b) Yb^{3+}; (c) Tb^{3+}. The results for (b) and (c) are in Table 1, but you should give the separate steps in applying the rules.

3. ***Triplet excited states.*** Some organic molecules have a triplet $(S = 1)$ excited state at an energy $k_B\Delta$ above a singlet $(S = 0)$ ground state. (a) Find an expression for the magnetic moment $\langle\mu\rangle$ in a field B. (b) Show that the susceptibility for $T \gg \Delta$ is approximately independent of Δ. (c) With the help of a diagram of energy levels versus field and a rough sketch of entropy versus field, explain how this system might be cooled by isentropic magnetization (not demagnetization).

4. ***Heat capacity from internal degrees of freedom.*** (a) Consider a two-level system with an energy splitting $k_B\Delta$ between upper and lower states; the splitting may arise from a magnetic field or in other ways. Show that the heat capacity per system is

$$C = \left(\frac{\partial U}{\partial T}\right)_\Delta = k_B \frac{(\Delta/T)^2 e^{\Delta/T}}{(1 + e^{\Delta/T})^2}\ .$$

The function is plotted in Fig. 11. Peaks of this type in the heat capacity are often known as **Schottky anomalies**. The maximum heat capacity is quite high, but for $T \ll \Delta$ and for $T \gg \Delta$ the heat capacity is low. (b) Show that for $T \gg \Delta$ we have $C \cong k_B(\Delta/2T)^2 + \ldots$. The hyperfine interaction between nuclear and electronic magnetic moments in paramagnetic salts (and in systems having electron spin order) causes

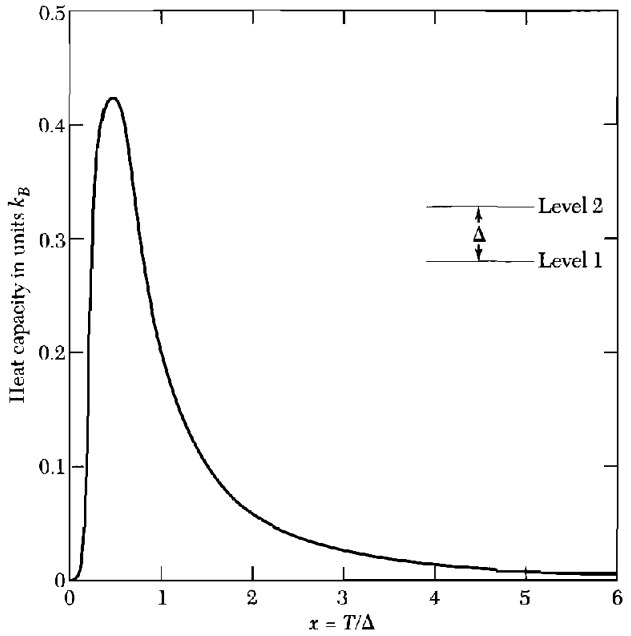


Figure 11 Heat capacity of a two-level system as a function of T/Δ, where Δ is the level splitting. The Schottky anomaly in the heat capacity is a very useful tool for determining energy level splittings of ions in rare-earth and transition-group metals, compounds, and alloys.

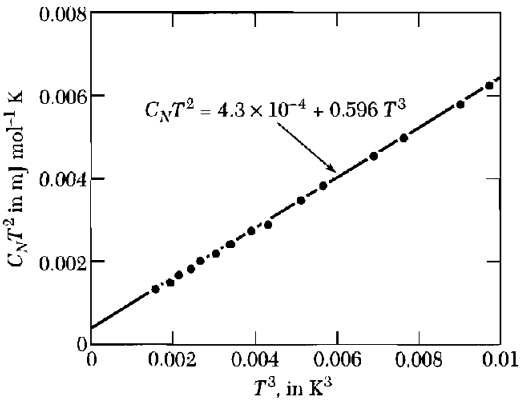


Figure 12 The normal-state heat capacity of gallium at $T < 0.21$ K. The nuclear quadrupole ($C \propto T^{-2}$) and conduction electron ($C \propto T$) contributions dominate the heat capacity at very low temperatures. (After N. E. Phillips.)

splittings with $\Delta \approx 1$ to 100 mK. These splittings are often detected experimentally by the presence of a term in $1/T^2$ in the heat capacity in the region $T \gg \Delta$. Nuclear electric quadrupole interactions with crystal fields also cause splittings, as in Fig. 12.

5. ***Pauli spin susceptibility.*** The spin susceptibility of a conduction electron gas at absolute zero may be approached by another method. Let

$$N^+ = \tfrac{1}{2}N(1 + \zeta) \; ; \qquad N^- = \tfrac{1}{2}N(1 - \zeta)$$

be the concentrations of spin-up and spin-down electrons. (a) Show that in a magnetic field B the total energy of the spin-up band in a free electron gas is

$$E^{\uparrow} = E_0(1+\zeta)^{5/3} - \tfrac{1}{2}N\mu B(1+\zeta) \ ,$$

where $E_0 = \frac{3}{10}N\epsilon_F$, in terms of the Fermi energy ϵ_F in zero magnetic field. Find a similar expression for E^{-}. (b) Minimize $E_{\text{total}} = E^{+} + E^{-}$ with respect to ζ and solve for the equilibrium value of ζ in the approximation $\zeta \ll 1$. Go on to show that the magnetization is $M = 3N\mu^2 B/2\epsilon_F$, in agreement with Eq. (45).

6. ***Conduction electron ferromagnetism.*** We approximate the effect of exchange interactions among the conduction electrons if we assume that electrons with parallel spins interact with each other with energy $-V$, and V is positive, while electrons with antiparallel spins do not interact with each other. (a) Show with the help of Problem 5 that the total energy of the spin-up band is

$$E^{+} = E_0(1+\zeta)^{5/3} - \tfrac{1}{8}VN^2(1+\zeta)^2 - \tfrac{1}{2}N\mu B(1+\zeta) \ ;$$

find a similar expression for E^{-}. (b) Minimize the total energy and solve for ζ in the limit $\zeta \ll 1$. Show that the magnetization is

$$M = \frac{3N\mu^2}{2\epsilon_F - \frac{3}{2}VN}B \ ,$$

so that the exchange interaction enhances the susceptibility. (c) Show that with $B = 0$ the total energy is unstable at $\zeta = 0$ when $V > 4\epsilon_F/3N$. If this is satisfied, a ferromagnetic state ($\zeta \neq 0$) will have a lower energy than the paramagnetic state. Because of the assumption $\zeta \ll 1$, this is a sufficient condition for ferromagnetism, but it may not be a necessary condition. It is known as the **Stoner condition**.

7. ***Two-level system.*** The result of Problem 4 is often seen in another form. (a) If the two energy levels are at Δ and $-\Delta$, show that the energy and heat capacity are

$$U = -\Delta \tanh(\Delta/k_BT) \ ; \qquad C = k_B(\Delta/k_BT)^2 \operatorname{sech}^2(\Delta/k_BT) \ .$$

(b) If the system has a random composition such that all values of Δ are equally likely up to some limit Δ_0, show that the heat capacity is linearly proportional to the temperature, provided $k_BT \ll \Delta_0$. This result was applied to the heat capacity of dilute magnetic alloys by W. Marshall, Phys. Rev. **118**, 1519 (1960). It is also used in the theory of glasses.

8. ***Paramagnetism of $S = 1$ system.*** (a) Find the magnetization as a function of magnetic field and temperature for a system of spins with $S = 1$, moment μ, and concentration n. (b) Show that in the limit $\mu B \ll kT$ the result is $M \cong (2n\mu^2/3kT)B$.

12

Ferromagnetism and Antiferromagnetism

NOTATION: (CGS) $B = H + 4\pi M$; (SI) $B = \mu_0(H + M)$. We call B_a the applied magnetic field in both systems of units: in CGS we have $B_a = H_a$ and in SI we have $B_a = \mu_0 H_a$. The susceptibility is $\chi = M/B_a$ in CGS and $\chi = M/H_a = \mu_0 M/B_a$ in SI. One tesla = 10^4 gauss.

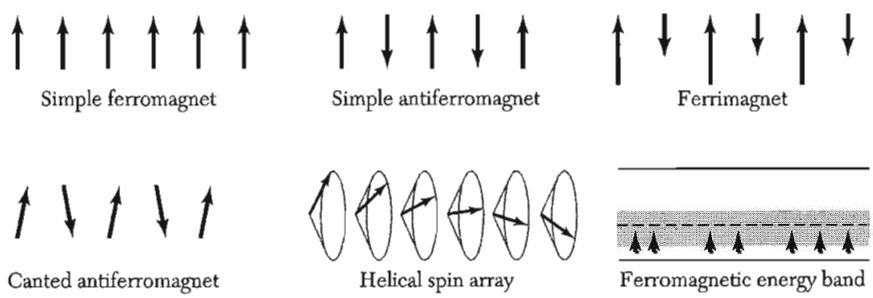


Figure 1 Ordered arrangements of electron spins.

FERROMAGNETIC ORDER

A ferromagnet has a spontaneous magnetic moment—a magnetic moment even in zero applied magnetic field. The existence of a spontaneous moment suggests that electron spins and magnetic moments are arranged in a regular manner. The order need not be simple: all of the spin arrangements sketched in Fig. 1 except the simple antiferromagnet have a spontaneous magnetic moment, called the **saturation moment**.

Curie Point and the Exchange Integral

Consider a paramagnet with a concentration of N ions of spin S. Given an internal interaction tending to line up the magnetic moments parallel to each other, we shall have a ferromagnet. Let us postulate such an interaction and call it the **exchange field**.[1] The orienting effect of the exchange field is opposed by thermal agitation, and at elevated temperatures the spin order is destroyed.

We treat the exchange field as equivalent to a magnetic field $\mathbf{B}_E$. The magnitude of the exchange field may be as high as 10^7 gauss (10^3 tesla). We assume that $\mathbf{B}_E$ is proportional to the magnetization $\mathbf{M}$.

The magnetization is defined as the magnetic moment per unit volume; unless otherwise specified it is understood to be the value in thermal equilibrium at the temperature T. If domains (regions magnetized in different directions) are present, the magnetization refers to the value within a domain.

In the **mean-field approximation** we assume each magnetic atom experiences a field proportional to the magnetization:

$$\mathbf{B}_E = \lambda \mathbf{M} \ , \tag{1}$$

where λ is a constant, independent of temperature. According to (1), each spin sees the average magnetization of all the other spins. In truth, it may see only near neighbors, but our simplification is good for a first look at the problem.

The **Curie temperature** T_c is the temperature above which the spontaneous magnetization vanishes; it separates the disordered paramagnetic phase at $T > T_c$ from the ordered ferromagnetic phase at $T < T_c$. We can find T_c in terms of the constant λ in (1).

[1]Also called the molecular field or the Weiss field, after Pierre Weiss who was the first to imagine such a field. The exchange field B_E simulates a real magnetic field in the expressions for the energy $-\boldsymbol{\mu} \cdot \mathbf{B}_E$ and the torque $\boldsymbol{\mu} \times \mathbf{B}_E$ on a magnetic moment $\boldsymbol{\mu}$. But B_E is not really a magnetic field and therefore does not enter into the Maxwell equations; for example, there is no current density $\mathbf{j}$ related to $\mathbf{B}_E$ by curl $\mathbf{H} = 4\pi\mathbf{j}/c$. The magnitude of B_E is typically 10^4 larger than the average magnetic field of the magnetic dipoles of the ferromagnet.

Consider the paramagnetic phase: an applied field B_a will cause a finite magnetization and this in turn will cause a finite exchange field B_E. If χ_p is the paramagnetic susceptibility,

$$\text{(CGS)} \quad M = \chi_p(B_a + B_E) \; ; \qquad \text{(SI)} \quad \mu_0 M = \chi_p(B_a + B_E) \; . \tag{2}$$

The magnetization is equal to a constant susceptibility times a field only if the fractional alignment is small: this is where the assumption enters that the specimen is in the paramagnetic phase.

The paramagnetic susceptibility (Chapt. 11) is given by the Curie law $\chi_p = C/T$, where C is the Curie constant. Substitute (1) in (2); we find $MT = C(B_a + \lambda M)$ and

$$\text{(CGS)} \qquad \chi = \frac{M}{B_a} = \frac{C}{(T - C\lambda)} \; . \tag{3}$$

The susceptibility (3) has a singularity at $T = C\lambda$. At this temperature (and below) there exists a spontaneous magnetization, because if χ is infinite we can have a finite M for zero B_a. From (3) we have the **Curie-Weiss law**

$$\text{(CGS)} \qquad \chi = \frac{C}{T - T_c} \; ; \qquad T_c = C\lambda \; . \tag{4}$$

This expression describes fairly well the observed susceptibility variation in the paramagnetic region above the Curie point. The reciprocal susceptibility of nickel is plotted in Fig. 2.

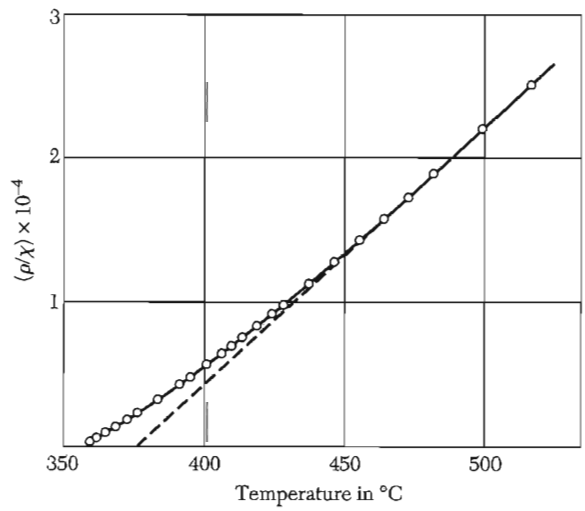


Figure 2 Reciprocal of the susceptibility per gram of nickel in the neighborhood of the Curie temperature (358°C). The density is ρ. The dashed line is a linear extrapolation from high temperatures. (After P. Weiss and R. Forrer.)

From (4) and the definition (11.22) of the Curie constant C we may determine the value of the mean field constant λ in (1):

$$\text{(CGS)} \qquad \lambda = \frac{T_c}{C} = \frac{3k_B T_c}{Ng^2 S(S+1)\mu_B^2} \ . \tag{5}$$

For iron $T_c \approx 1000$ K, $g \approx 2$, and S ≈ 1; from (5) we have $\lambda \approx 5000$. With $M_s \approx 1700$ we have $B_E \approx \lambda M \approx (5000)(1700) \approx 10^7$ G $= 10^3$ T. The exchange field in iron is very much stronger than the real magnetic field due to the other magnetic ions in the crystal: a magnetic ion produces a field $\approx \mu_B/a^3$ or about 10^3 G $= 0.1$ T at a neighboring lattice point.

The exchange field gives an approximate representation of the quantum-mechanical exchange interaction. On certain assumptions it is shown in texts on quantum theory that the energy of interaction of atoms i, j bearing electron spins S_i, S_j contains a term

$$U = -2J\mathbf{S}_i \cdot \mathbf{S}_j \ , \tag{6}$$

where J is the exchange integral and is related to the overlap of the charge distributions of the atoms i, j. Equation (6) is called the **Heisenberg model.**

The charge distribution of a system of two spins depends on whether the spins are parallel or antiparallel[2] for the Pauli principle excludes two electrons of the same spin from being at the same place at the same time. It does not exclude two electrons of opposite spin. Thus the electrostatic energy of a system will depend on the relative orientation of the spins: the difference in energy defines the **exchange energy**.

The exchange energy of two electrons may be written in the form $-2J\mathbf{s}_1 \cdot \mathbf{s}_2$ as in (6), just as if there were a direct coupling between the directions of the two spins. For many purposes in ferromagnetism it is a good approximation to treat the spins as classical angular momentum vectors.

We can establish an approximate connection between the exchange integral J and the Curie temperature T_c. Suppose that the atom under consideration has z nearest neighbors, each connected with the central atom by the interaction J. For more distant neighbors we take J as zero. The mean field theory result is

$$J = \frac{3k_B T_c}{2zS(S+1)} \ . \tag{7}$$

Better statistical approximations give somewhat different results. For the sc, bcc, and fcc structures with $S = \frac{1}{2}$, Rushbrooke and Wood give

[2]If two spins are antiparallel, the wavefunctions of the two electrons must be symmetric, as in the combination $u(\mathbf{r}_1)v(\mathbf{r}_2) + u(\mathbf{r}_2)v(\mathbf{r}_1)$. If the two spins are parallel, the Pauli principle requires that the orbital part of the wavefunction be antisymmetric, as in $u(\mathbf{r}_1)v(\mathbf{r}_2) - u(\mathbf{r}_2)v(\mathbf{r}_1)$, for here if we interchange the coordinates $\mathbf{r}_1$, $\mathbf{r}_2$ the wavefunction changes sign. If we set the positions equal so that $\mathbf{r}_1 = \mathbf{r}_2$ then the antisymmetric function vanishes: for parallel spins there is zero probability of finding the two electrons at the same position.

$k_B T_c/zJ = 0.28$; 0.325; and 0.346, respectively, as compared with 0.500 from (7) for all three structures. If iron is represented by the Heisenberg model with $S = 1$, then the observed Curie temperature corresponds to $J = 11.9$ meV.

Temperature Dependence of the Saturation Magnetization

We can also use the mean field approximation below the Curie temperature to find the magnetization as a function of temperature. We proceed as before, but instead of the Curie law we use the complete Brillouin expression for the magnetization. For spin $\frac{1}{2}$ this is $M = N\mu \tanh(\mu B/k_B T)$.

If we omit the applied magnetic field and replace B by the molecular field $B_E = \lambda M$, then

$$M = N\mu \tanh(\mu\lambda M/k_B T) \ . \tag{8}$$

We shall see that solutions of this equation with nonzero M exist in the temperature range between 0 and T_c.

To solve (8) we write it in terms of the reduced magnetization $m \equiv M/N\mu$ and the reduced temperature $t = k_B T/N\mu^2\lambda$, whence

$$m = \tanh(m/t) \ . \tag{9}$$

We then plot the right and left sides of this equation separately as functions of m, as in Fig. 3. The intercept of the two curves gives the value of m at the temperature of interest. The critical temperature is $t = 1$, or $T_c = N\mu^2\lambda/k_B$.

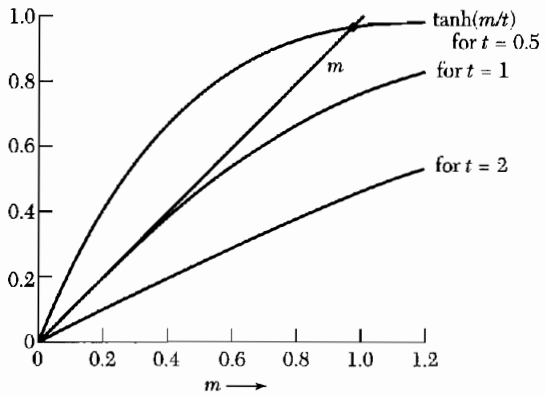


Figure 3 Graphical solution of Eq. (9) for the reduced magnetization m as a function of temperature. The reduced magnetization is defined as $m = M/N\mu$. The left-hand side of Eq. (9) is plotted as a straight line m with unit slope. The right-hand side is $\tanh(m/t)$ and is plotted vs. m for three different values of the reduced temperature $t = k_B T/N\mu^2\lambda = T/T_c$. The three curves correspond to the temperatures $2T_c$, T_c, and $0.5T_c$. The curve for $t = 2$ intersects the straight line m only at $m = 0$, as appropriate for the paramagnetic region (there is no external applied magnetic field). The curve for $t = 1$ (or $T = T_c$) is tangent to the straight line m at the origin; this temperature marks the onset of ferromagnetism. The curve for $t = 0.5$ is in the ferromagnetic region and intersects the straight line m at about $m = 0.94N\mu$. As $t \to 0$ the intercept moves up to $m = 1$, so that all magnetic moments are lined up at absolute zero.

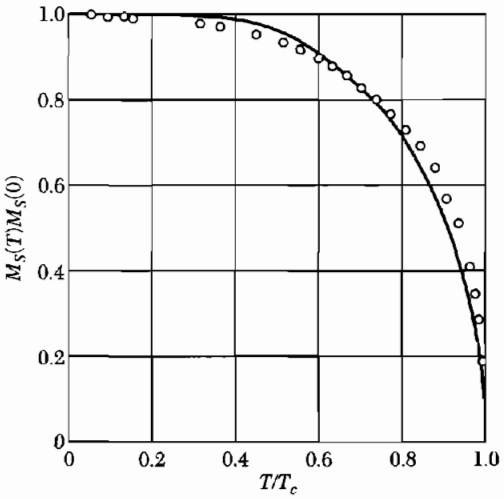


Figure 4 Saturation magnetization of nickel as a function of temperature, together with the theoretical curve for $S = \frac{1}{2}$ on the mean field theory. Experimental values by P. Weiss and R. Forrer.

The curves of M versus T obtained in this way reproduce roughly the features of the experimental results, as shown in Fig. 4 for nickel. As T increases, the magnetization decreases smoothly to zero at $T = T_c$. This behavior classifies the usual ferromagnetic/paramagnetic transition as a second-order transition.

The mean-field theory does not give a good description of the variation of M at low temperatures. For $T \ll T_c$ the argument of tanh in (9) is large, and

$$\tanh \xi \cong 1 - 2e^{-2\xi} \ldots .$$

To lowest order the magnetization deviation $\Delta M = M(0) - M(T)$ is

$$\Delta M \cong 2N\mu \exp(-2\lambda N\mu^2/k_B T) \ . \tag{10}$$

The argument of the exponential is equal to $-2T_c/T$. For $T = 0.1T_c$ we have $\Delta M/N\mu \cong 4 \times 10^{-9}$.

The experimental results show a much more rapid dependence of ΔM on temperature at low temperatures. At $T = 0.1T_c$ we have $\Delta M/M \cong 2 \times 10^{-3}$ from the data of Fig. 5. The leading term in ΔM is observed from experiment to have the form

$$\Delta M/M(0) = AT^{3/2} \ , \tag{11}$$

where the constant A has the experimental value $(7.5 \pm 0.2) \times 10^{-6}$ deg$^{-3/2}$ for Ni and $(3.4 \pm 0.2) \times 10^{-6}$ deg$^{-3/2}$ for Fe. The result (11) finds a natural explanation in terms of spin wave theory.

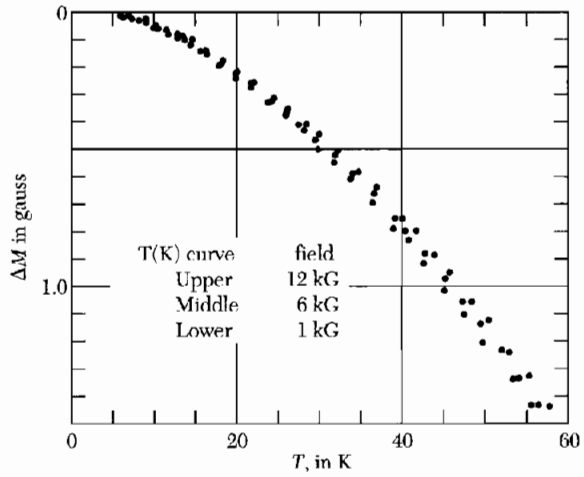


Figure 5 Decrease in magnetization of nickel with temperature, after Argyle, Charap, and Pugh. In the plot $\Delta M \equiv 0$ at 4.2 K.

Saturation Magnetization at Absolute Zero

Table 1 gives representative values of the saturation magnetization M_s, the ferromagnetic Curie temperature, and the effective magneton number defined by $M_s(0) = n_B N \mu_B$, where N is the number of formula units per unit volume. Do not confuse n_B with the paramagnetic effective magneton number p defined by (11.23).

Table 1 Ferromagnetic crystals

Substance	Magnetization M_s, in gauss		n_B(0 K), per formula unit	Curie temperature, in K
	Room temperature	0 K		
Fe	1707	1740	2.22	1043
Co	1400	1446	1.72	1388
Ni	485	510	0.606	627
Gd	—	2060	7.63	292
Dy	—	2920	10.2	88
MnAs	670	870	3.4	318
MnBi	620	680	3.52	630
MnSb	710	—	3.5	587
CrO_2	515	—	2.03	386
$MnOFe_2O_3$	410	—	5.0	573
$FeOFe_2O_3$	480	—	4.1	858
$NiOFe_2O_3$	270	—	2.4	(858)
$CuOFe_2O_3$	135	—	1.3	728
$MgOFe_2O_3$	110	—	1.1	713
EuO	—	1920	6.8	69
$Y_3Fe_5O_{12}$	130	200	5.0	560

Observed values of n_B are often nonintegral. There are several possible causes. One is the spin-orbit interaction which adds or subtracts some orbital magnetic moment. Another cause in ferromagnetic metals is the conduction electron magnetization induced locally about a paramagnetic ion core. A third cause is suggested by the drawing in Fig. 1 of the spin arrangement in a ferrimagnet: if there is one atom of spin projection $-S$ for every two atoms $+S$, the average spin is $\frac{1}{3}S$.

Are there in fact any simple ferromagnetic insulators, with all ionic spins parallel in the ground state? The few simple ferromagnets known at present include $CrBr_3$, EuO, and EuS.

A band or itinerant electron model accounts for the ferromagnetism of the transition metals Fe, Co, Ni. The approach is indicated in Figs. 6 and 7. The relationship of 4*s* and 3*d* bands is shown in Fig. 6 for copper, which is not ferromagnetic. If we remove one electron from copper, we obtain nickel which has the possibility of a hole in the 3*d* band. In the band structure of nickel shown in Fig. 7a for $T > T_c$ we have taken $2 \times 0.27 = 0.54$ of an electron away from the 3*d* band and 0.46 away from the 4*s* band, as compared with copper.

The band structure of nickel at absolute zero is shown in Fig. 7b. Nickel is ferromagnetic, and at absolute zero $n_B = 0.60$ Bohr magnetons per atom. After allowance for the magnetic moment contribution of orbital electronic motion, nickel has an excess of 0.54 electron per atom having spin preferentially oriented in one direction. The exchange enhancement of the susceptibility of metals was the subject of Problem 11.6.

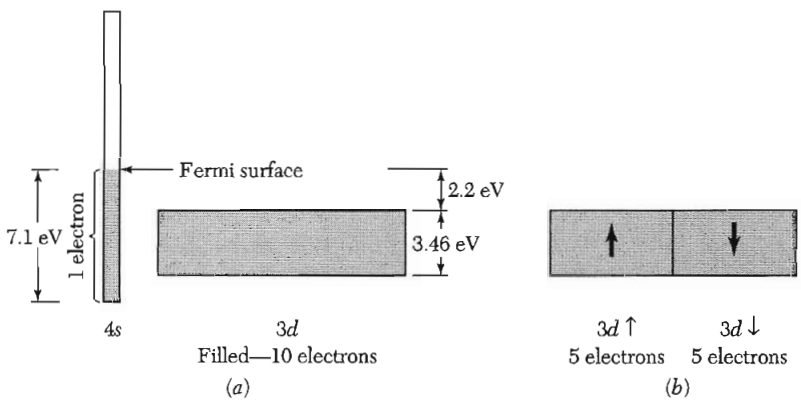


Figure 6a Schematic relationship of 4*s* and 3*d* bands in metallic copper. The 3*d* band holds 10 electrons per atom and is filled. The 4*s* band can hold two electrons per atom; it is shown half-filled, as copper has one valence electron outside the filled 3*d* shell.

Figure 6b The filled 3*d* band of copper shown as two separate sub-bands of opposite electron spin orientation, each band holding five electrons. With both sub-bands filled as shown, the net spin (and hence the net magnetization) of the *d* band is zero.

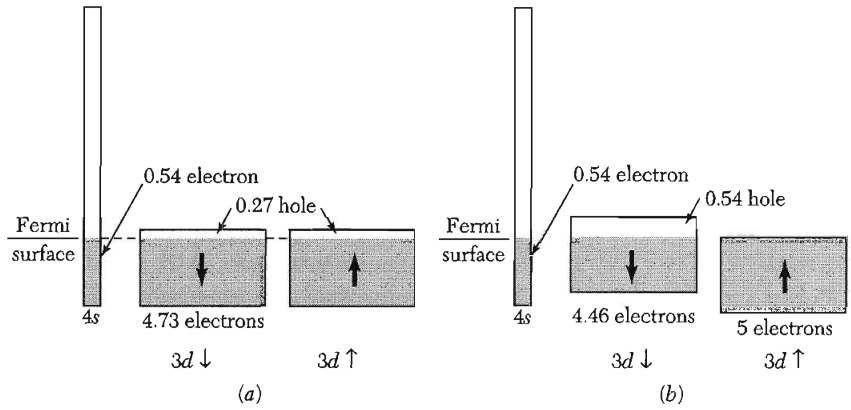


Figure 7a Band relationships in nickel above the Curie temperature. The net magnetic moment is zero, as there are equal numbers of holes in the $3d \downarrow$ and $3d \uparrow$ bands.

Figure 7b Schematic relationship of bands in nickel at absolute zero. The energies of the $3d \uparrow$ and $3d \downarrow$ sub-bands are separated by an exchange interaction. The $3d \uparrow$ band is filled; the $3d \downarrow$ band contains 4.46 electrons and 0.54 hole. The $4s$ band is usually thought to contain approximately equal numbers of electrons in both spin directions, and so we have not troubled to divide it into sub-bands. The net magnetic moment of 0.54 μ_B per atom arises from the excess population of the $3d \uparrow$ band over the $3d \downarrow$ band. It is often convenient to speak of the magnetization as arising from the 0.54 hole in the $3d \downarrow$ band.

MAGNONS

A magnon is a quantized spin wave. We use a classical argument, just as we did for phonons, to find the magnon dispersion relation for ω versus k. We then quantize the magnon energy and interpret the quantization in terms of spin reversal.

The ground state of a simple ferromagnet has all spins parallel, as in Fig. 8a. Consider N spins each of magnitude S on a line or a ring, with nearest-neighbor spins coupled by the Heisenberg interaction:

$$U = -2J \sum_{p=1}^{N} \mathbf{S}_p \cdot \mathbf{S}_{p+1} \ . \tag{12}$$

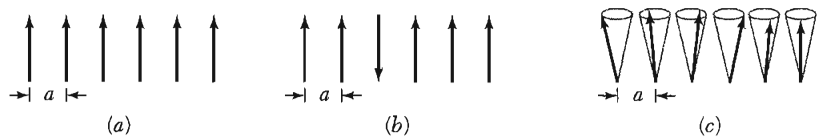


Figure 8 (a) Classical picture of the ground state of a simple ferromagnet; all spins are parallel. (b) A possible excitation; one spin is reversed. (c) The low-lying elementary excitations are spin waves. The ends of the spin vectors precess on the surfaces of cones, with successive spins advanced in phase by a constant angle.

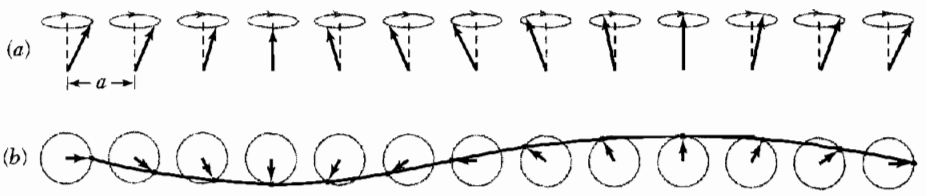


Figure 9 A spin wave on a line of spins. (a) The spins viewed in perspective. (b) Spins viewed from above, showing one wavelength. The wave is drawn through the ends of the spin vectors.

Here J is the exchange integral and $\hbar\mathbf{S}_p$ is the angular momentum of the spin at site p. If we treat the spins $\mathbf{S}_p$ as classical vectors, then in the ground state $\mathbf{S}_p \cdot \mathbf{S}_{p+1} = S^2$ and the exchange energy of the system is $U_0 = -2NJS^2$.

What is the energy of the first excited state? Consider an excited state with one particular spin reversed, as in Fig. 8b. We see from (12) that this increases the energy by $8JS^2$, so that $U_1 = U_0 + 8JS^2$. But we can form an excitation of much lower energy if we let all the spins share the reversal, as in Fig. 8c. The elementary excitations of a spin system have a wavelike form and are called magnons (Fig. 9). These are analogous to lattice vibrations or phonons. Spin waves are oscillations in the relative orientations of spins on a lattice; lattice vibrations are oscillations in the relative positions of atoms on a lattice.

We now give a classical derivation of the magnon dispersion relation. The terms in (12) which involve the pth spin are

$$-2J\mathbf{S}_p \cdot (\mathbf{S}_{p-1} + \mathbf{S}_{p+1}) \ . \tag{13}$$

We write magnetic moment at site p as $\boldsymbol{\mu}_p = -g\mu_B\mathbf{S}_p$. Then (13) becomes

$$-\boldsymbol{\mu}_p \cdot [(-2J/g\mu_B)(\mathbf{S}_{p-1} + \mathbf{S}_{p+1})] \ , \tag{14}$$

which is of the form $-\boldsymbol{\mu}_p \cdot \mathbf{B}_p$, where the effective magnetic field or exchange field that acts on the pth spin is

$$\mathbf{B}_p = (-2J/g\mu_B)(\mathbf{S}_{p-1} + \mathbf{S}_{p+1}) \ . \tag{15}$$

From mechanics the rate of change of the angular momentum $\hbar\mathbf{S}_p$ is equal to the torque $\boldsymbol{\mu}_p \times \mathbf{B}_p$ which acts on the spin: $\hbar\, d\mathbf{S}_p/dt = \boldsymbol{\mu}_p \times \mathbf{B}_p$, or

$$d\mathbf{S}_p/dt = (-g\mu_B/\hbar)\, \mathbf{S}_p \times \mathbf{B}_p = (2J/\hbar)(\mathbf{S}_p \times \mathbf{S}_{p-1} + \mathbf{S}_p \times \mathbf{S}_{p+1}) \ . \tag{16}$$

In Cartesian components

$$dS_p^x/dt = (2J/\hbar)[S_p^y(S_{p-1}^z + S_{p+1}^z) - S_p^z(S_{p-1}^y + S_{p+1}^y)] \ , \tag{17}$$

and similarly for dS_p^y/dt and dS_p^z/dt. These equations involve products of spin components and are nonlinear.

If the amplitude of the excitation is small (if $S_p^x, S_p^y \ll S$), we may obtain an approximate set of linear equations by taking all $S_p^z = S$ and by neglecting terms in the product of S^x and S^y which appear in the equation for dS^z/dt. The linearized equations are

$$dS_p^x/dt = (2JS/\hbar)(2S_p^y - S_{p-1}^y - S_{p+1}^y) \ ; \tag{18a}$$

$$dS_p^y/dt = -(2JS/\hbar)(2S_p^x - S_{p-1}^x - S_{p+1}^x) \ ; \tag{18b}$$

$$dS_p^z/dt = 0 \ . \tag{19}$$

By analogy with phonon problems we look for traveling wave solutions of (18) of the form

$$S_p^x = u \exp[i(pka - \omega t)] \ ; \qquad S_p^y = v \exp[i(pka - \omega t)] \ , \tag{20}$$

where u, v are constants, p is an integer, and a is the lattice constant. On substitution into (18) we have

$$-i\omega u = (2JS/\hbar)(2 - e^{-ika} - e^{ika})\, v = (4JS/\hbar)(1 - \cos ka)v \ ;$$

$$-i\omega v = -(2JS/\hbar)(2 - e^{-ika} - e^{ika})u = -(4JS/\hbar)(1 - \cos ka)u \ .$$

These equations have a solution for u and v if the determinant of the coefficients is equal to zero:

$$\begin{vmatrix} i\omega & (4JS/\hbar)(1 - \cos ka) \\ -(4JS/\hbar)(1 - \cos ka) & i\omega \end{vmatrix} = 0 \ , \tag{21}$$

whence

$$\hbar\omega = 4JS(1 - \cos ka) \ . \tag{22}$$

This result is plotted in Fig. 10. With this solution we find that $v = -iu$, corresponding to circular precession of each spin about the z axis. We see this on taking real parts of (20), with v set equal to $-iu$. Then

$$S_p^x = u \cos(pka - \omega t) \ ; \qquad S_p^y = u \sin(pka - \omega t) \ .$$

Equation (22) is the dispersion relation for spin waves in one dimension with nearest-neighbor interactions. Precisely the same result is obtained from the quantum-mechanical solution; see *QTS*, Chapter 4. At long wavelengths $ka \ll 1$, so that $(1 - \cos ka) \cong \frac{1}{2}(ka)^2$ and

$$\hbar\omega \cong (2JSa^2)k^2 \ . \tag{23}$$

The frequency is proportional to k^2; in the same limit the frequency of a phonon is directly proportional to k.

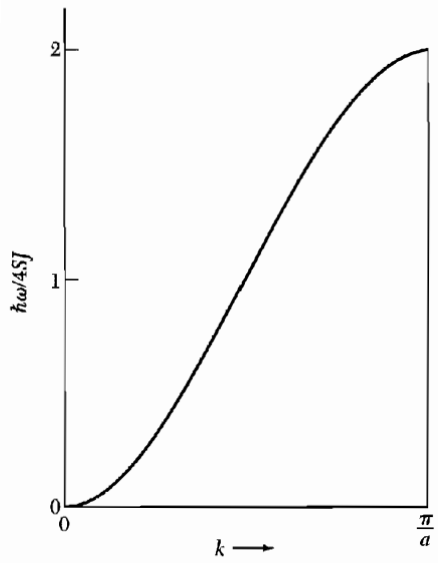


Figure 10 Dispersion relation for magnons in a ferromagnet in one dimension with nearest-neighbor interactions.

The dispersion relation for a ferromagnetic cubic lattice with nearest-neighbor interactions

$$\hbar\omega = 2JS[z - \sum_{\delta} \cos(\mathbf{k} \cdot \boldsymbol{\delta})] \ , \tag{24}$$

where the summation is over the z vectors denoted by $\boldsymbol{\delta}$ which join the central atom to its nearest neighbors. For $ka \ll 1$,

$$\hbar\omega = (2JSa^2)k^2 \tag{25}$$

for all three cubic lattices, where a is the lattice constant.

The coefficient of k^2 often may be determined accurately by neutron scattering or by spin wave resonance in thin films, Chapter 13. By neutron scattering G. Shirane and coworkers find, in the equation $\hbar\omega = Dk^2$, the values 281, 500, and 364 meV Å^2 for D at 295 K in Fe, Co, and Ni, respectively.

Quantization of Spin Waves. The quantization of spin waves proceeds as for photons and phonons. The energy of a mode of frequency ω_k with n_k magnons is given by

$$\epsilon_k = (n_k + \tfrac{1}{2})\hbar\omega_k \ . \tag{26}$$

The excitation of a magnon corresponds to the reversal of one spin $\frac{1}{2}$.

Thermal Excitation of Magnons

In thermal equilibrium the average value of the number of magnons excited in the mode $\mathbf{k}$ is given by the Planck distribution[3]

$$\langle n_k \rangle = \frac{1}{\exp(\hbar\omega_k/k_BT) - 1} \ . \tag{27}$$

The total number of magnons excited at a temperature T is

$$\sum_{\mathbf{k}} n_{\mathbf{k}} = \int d\omega \, D(\omega)\langle n(\omega)\rangle \ , \tag{28}$$

where $D(\omega)$ is the number of magnon modes per unit frequency range. The integral is taken over the allowed range of $\mathbf{k}$, which is the first Brillouin zone. At sufficiently low temperatures we may evaluate the integral between 0 and ∞ because $\langle n(\omega)\rangle \to 0$ exponentially as $\omega \to \infty$.

Magnons have a single polarization for each value of $\mathbf{k}$. In three dimensions the number of modes of wavevector less than k is $(1/2\pi)^3(4\pi k^3/3)$ per unit volume, whence the number of magnons $D(\omega)d\omega$ with frequency in $d\omega$ at ω is $(1/2\pi)^3(4\pi k^2)(dk/d\omega)\, d\omega$. In the approximation (25),

$$\frac{d\omega}{dk} = \frac{4JSa^2k}{\hbar} = 2\left(\frac{2JSa^2}{\hbar}\right)^{1/2}\omega^{1/2} \ .$$

Thus the density of modes for magnons is

$$D(\omega) = \frac{1}{4\pi^2}\left(\frac{\hbar}{2JSa^2}\right)^{3/2}\omega^{1/2} \ , \tag{29}$$

so that the total number of magnons is, from (28),

$$\sum_{\mathbf{k}} n_{\mathbf{k}} = \frac{1}{4\pi^2}\left(\frac{\hbar}{2JSa^2}\right)^{3/2}\int_0^\infty d\omega \frac{\omega^{1/2}}{e^{\beta\hbar\omega} - 1} = \frac{1}{4\pi^2}\left(\frac{k_BT}{2JSa^2}\right)^{3/2}\int_0^\infty dx \frac{x^{1/2}}{e^x - 1} \ .$$

The definite integral is found in tables and has the value $(0.0587)(4\pi^2)$.

The number N of atoms per unit volume is Q/a^3, where $Q = 1, 2, 4$ for sc, bcc, fcc lattices, respectively. Now $(\Sigma n_{\mathbf{k}})/NS$ is equal to the fractional change of magnetization $\Delta M/M(0)$, whence

$$\boxed{\frac{\Delta M}{M(0)} = \frac{0.0587}{SQ}\cdot\left(\frac{k_BT}{2JS}\right)^{3/2}} \ . \tag{30}$$

[3]The argument is exactly as for phonons or photons. The Planck distribution follows for any problem where the energy levels are identical with those of a harmonic oscillator or collection of harmonic oscillators.

This result is the **Bloch $T^{3/2}$ law** and has been confirmed experimentally. In neutron scattering experiments spin waves have been observed up to temperatures near the Curie temperature and even above the Curie temperature.

NEUTRON MAGNETIC SCATTERING

An x-ray photon sees the spatial distribution of electronic charge, whether the charge density is magnetized or unmagnetized. A neutron sees two aspects of a crystal: the distribution of nuclei and the distribution of electronic magnetization. The neutron diffraction pattern for iron is shown in Fig. 11.

The magnetic moment of the neutron interacts with the magnetic moment of the electron. The cross section for the neutron-electron interaction is of the same order of magnitude as for the neutron-nuclear interaction. Diffraction of neutrons by a magnetic crystal allows the determination of the distribution, direction, and order of the magnetic moments.

A neutron can be inelastically scattered by the magnetic structure, with the creation or annihilation of a magnon (Fig. 12); such events make possible the experimental determination of magnon spectra. If the incident neutron has wavevector $\mathbf{k}_n$ and is scattered to $\mathbf{k}_n'$ with the creation of a magnon of wavevector $\mathbf{k}$, then by conservation of crystal momentum $\mathbf{k}_n = \mathbf{k}_n' + \mathbf{k} + \mathbf{G}$, where $\mathbf{G}$ is a reciprocal lattice vector. By conservation of energy

$$\frac{\hbar^2 k_n^2}{2M_n} = \frac{\hbar^2 k_n'^2}{2M_n} + \hbar\omega_{\mathbf{k}} \ , \tag{31}$$

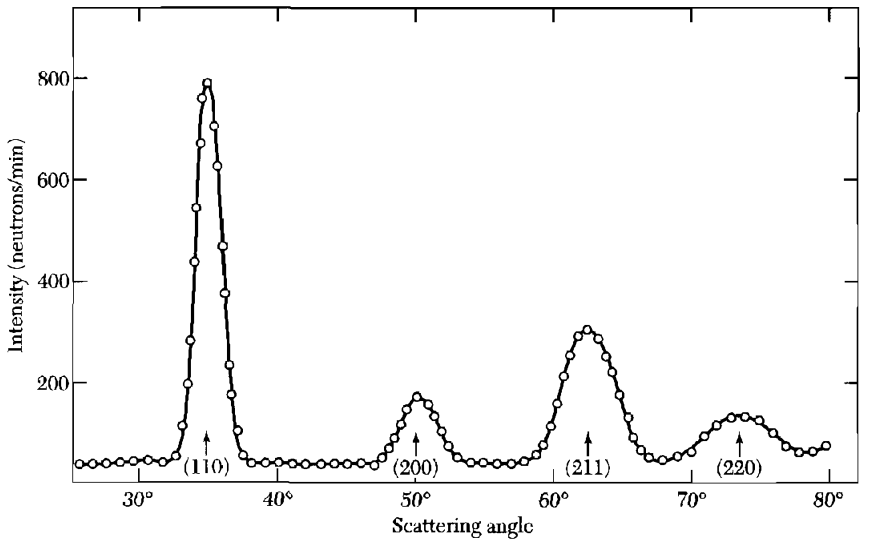


Figure 11 Neutron diffraction pattern for iron. (After C. G. Shull, E. O. Wollan, and W. C. Koehler.)

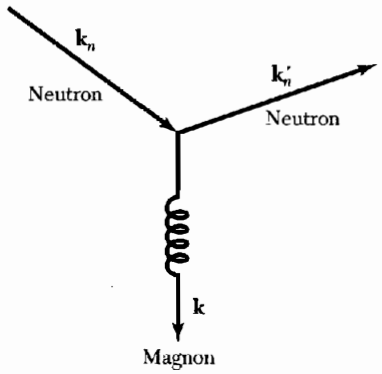


Figure 12 Scattering of a neutron by an ordered magnetic structure, with creation of a magnon.

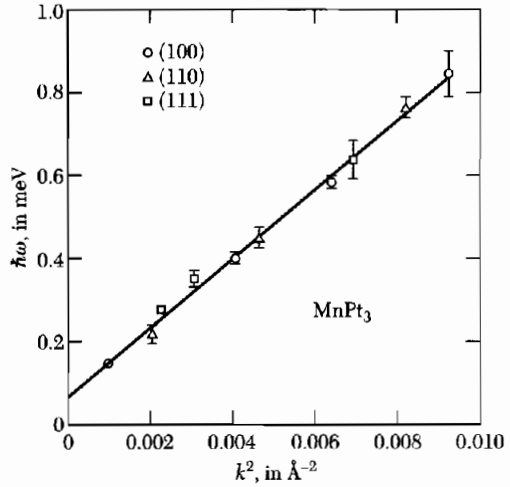


Figure 13 Magnon energy as a function of the square of the wavevector, for the ferromagnet $MnPt_3$. (After B. Antonini and V. J. Minkiewicz.)

where $\hbar\omega_{\mathbf{k}}$ is the energy of the magnon created in the process. The observed magnon spectrum for $MnPt_3$ is shown in Fig. 13.

FERRIMAGNETIC ORDER

In many ferromagnetic crystals the saturation magnetization at $T = 0$ K does not correspond to parallel alignment of the magnetic moments of the constituent paramagnetic ions, even in crystals for which the individual paramagnetic ions have their normal magnetic moments.

The most familiar example is magnetite, Fe_3O_4 or $FeO \cdot Fe_2O_3$. From Table 11.2 we see that ferric (Fe^{3+}) ions are in a state with spin $S = \frac{5}{2}$ and

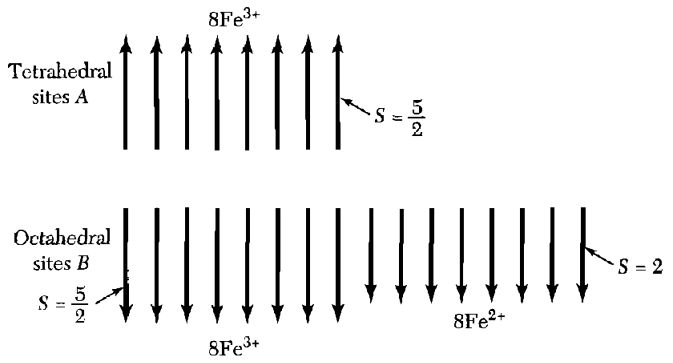


Figure 14 Spin arrangements in magnetite, $FeO \cdot Fe_2O_3$, showing how the moments of the Fe^{3+} ions cancel out, leaving only the moments of the Fe^{2+} ions.

zero orbital moment. Thus each ion should contribute $5\mu_B$ to the saturation moment. The ferrous (Fe^{2+}) ions have a spin of 2 and should contribute $4\mu_B$, apart from any residual orbital moment contribution. Thus the effective number of Bohr magnetons per Fe_3O_4 formula unit should be about $2 \times 5 + 4 = 14$ if all spins were parallel.

The observed value (Table 1) is 4.1. The discrepancy is accounted for if the moments of the Fe^{3+} ions are antiparallel to each other: then the observed moment arises only from the Fe^{2+} ion, as in Fig. 14. Neutron diffraction results agree with this model.

A systematic discussion of the consequences of this type of spin order was given by L. Néel with reference to an important class of magnetic oxides known as ferrites. The usual chemical formula of a ferrite is $MO \cdot Fe_2O_3$, where M is a divalent cation, often Zn, Cd, Fe, Ni, Cu, Co, or Mg. The term **ferrimagnetic** was coined originally to describe the ferrite-type ferromagnetic spin order such as Fig. 14, and by extension the term covers almost any compound in which some ions have a moment antiparallel to other ions. Many ferrimagnets are poor conductors of electricity, a quality exploited in applications such as rf transformer cores.

The cubic ferrites have the **spinel** crystal structure shown in Fig. 15. There are eight occupied tetrahedral (or A) sites and 16 occupied octahedral (or B) sites in a unit cube. The lattice constant is about 8 Å. A remarkable feature of the spinels is that all exchange integrals J_{AA}, J_{AB}, and J_{BB} are negative and favor *antiparallel* alignment of the spins connected by the interaction. But the AB interaction is the strongest, so that the A spins are parallel to each other and the B spins are parallel to each other, just in order that the A spins may be antiparallel to the B spins. If J in $U = -2J\mathbf{S}_i \cdot \mathbf{S}_j$ is positive, we say that the exchange integral is ferromagnetic; if J is negative, the exchange integral is antiferromagnetic.

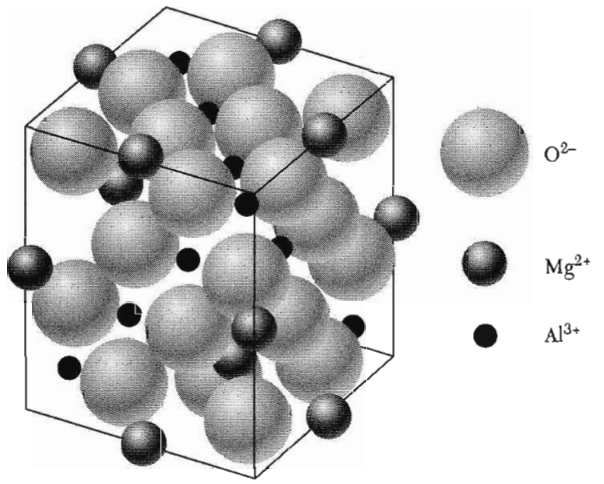


Figure 15 Crystal structure of the mineral spinel $MgAl_2O_4$; the Mg^{2+} ions occupy tetrahedral sites, each surrounded by four oxygen ions; the Al^{3+} occupy octahedral sites, each surrounded by six oxygen ions. This is a normal spinel arrangement: the divalent metal ions occupy the tetrahedral sites. In the inverse spinel arrangment the tetrahedral sites are occupied by trivalent metal ions, while the octahedral sites are occupied half by divalent and half by trivalent metal ions.

We now prove that three antiferromagnetic interactions can result in ferrimagnetism. The mean exchange fields acting on the A and B spin lattices may be written

$$\mathbf{B}_A = -\lambda \mathbf{M}_A - \mu \mathbf{M}_B \; ; \quad \mathbf{B}_B = -\mu \mathbf{M}_A - \nu \mathbf{M}_B \; ; \tag{32}$$

taking all mean field constants λ, μ, ν to be positive. The minus sign then corresponds to an antiparallel interaction. The interaction energy density is

$$U = -\tfrac{1}{2}(\mathbf{B}_A \cdot \mathbf{M}_A + \mathbf{B}_B \cdot \mathbf{M}_B) = \tfrac{1}{2}\lambda M_A^2 + \mu \mathbf{M}_A \cdot \mathbf{M}_B + \tfrac{1}{2}\nu M_B^2 \; ; \tag{33}$$

this is lower when $\mathbf{M}_A$ is antiparallel to $\mathbf{M}_B$ than when $\mathbf{M}_A$ is parallel to $\mathbf{M}_B$. The energy when antiparallel should be compared with zero, because a possible solution is $M_A = M_B = 0$. Thus when

$$\mu M_A M_B > \tfrac{1}{2}(\lambda M_A^2 + \nu M_B^2) \; , \tag{34}$$

the ground state will have M_A directed oppositely to M_B. (Under certain conditions there may be noncollinear spin arrays of still lower energy.)

Curie Temperature and Susceptibility of Ferrimagnets

We define separate Curie constants C_A and C_B for the ions on the A and B sites. For simplicity, let all interactions be zero except for an antiparallel interaction between the A and B sites: $\mathbf{B}_A = -\mu \mathbf{M}_B$; $\mathbf{B}_B = -\mu \mathbf{M}_A$, where μ is positive. The same constant μ is involved in both expressions because of the form of (33).

We have in the mean field approximation

$$\text{(CGS)} \qquad M_A T = C_A(B_a - \mu M_B) \; ; \quad M_B T = C_B(B_a - \mu M_A) \; , \tag{35}$$

where B_a is the applied field. These equations have a nonzero solution for M_A and M_B in zero applied field if

$$\begin{vmatrix} T & \mu C_A \\ \mu C_B & T \end{vmatrix} = 0 \; , \tag{36}$$

so that the ferrimagnetic Curie temperature is given by $T_c = \mu(C_A C_B)^{1/2}$.

We solve (35) for M_A and M_B to obtain the susceptibility at $T > T_c$:

$$\text{(CGS)} \qquad \chi = \frac{M_A + M_B}{B_a} = \frac{(C_A + C_B)T - 2\mu C_A C_B}{T^2 - T_c^2} \; , \tag{37}$$

a result more complicated than (4). Experimental values for Fe_3O_4 are plotted in Fig. 16. The curvature of the plot of $1/\chi$ versus T is a characteristic feature of a ferrimagnet. We consider below the antiferromagnetic limit $C_A = C_B$.

Iron Garnets. The iron garnets are cubic ferrimagnetic insulators with the general formula $M_3Fe_5O_{12}$, where M is a trivalent metal ion and the Fe is the trivalent ferric ion ($S = \frac{5}{2}$, $L = 0$). An example is yttrium iron garnet $Y_3Fe_5O_{12}$, known as YIG. Here Y^{3+} is diamagnetic.

The net magnetization of YIG is due to the resultant of two oppositely magnetized lattices of Fe^{3+} ions. At absolute zero each ferric ion contributes $\pm 5\mu_B$ to the magnetization, but in each formula unit the three Fe^{3+} ions on sites denoted as d sites are magnetized in one sense and the two Fe^{3+} ions on a sites are magnetized in the opposite sense, giving a resultant of $5\mu_B$ per formula unit in good agreement with the measurements of Geller *et al.*

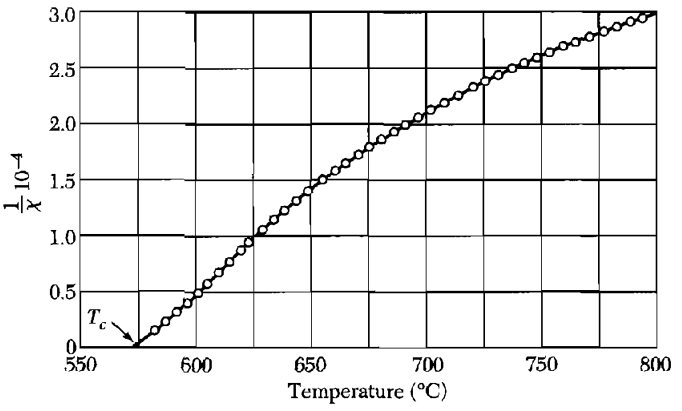


Figure 16 Reciprocal susceptibility of magnetite, $FeO \cdot Fe_2O_3$.

The mean field at an a site due to the ions on the d sites is $B_a = -(1.5 \times 10^4)M_d$. The observed Curie temperature 559 K of YIG is due to the a-d interaction. The only magnetic ions in YIG are the ferric ions. Because these are in an $L = 0$ state with a spherical charge distribution, their interaction with lattice deformations and phonons is weak. As a result YIG is characterized by very narrow linewidths in ferromagnetic resonance experiments.

ANTIFERROMAGNETIC ORDER

A classical example of magnetic structure determination by neutrons is shown in Fig. 17 for MnO, which has the NaCl structure. At 80 K there are extra neutron reflections not present at 293 K. The reflections at 80 K may be classified in terms of a cubic unit cell of lattice constant 8.85 Å. At 293 K the reflections correspond to an fcc unit cell of lattice constant 4.43 Å.

But the lattice constant determined by x-ray reflection is 4.43 Å at *both* temperatures, 80 K and 293 K. We conclude that the chemical unit cell has the 4.43 Å lattice parameter, but that at 80 K the electronic magnetic moments of

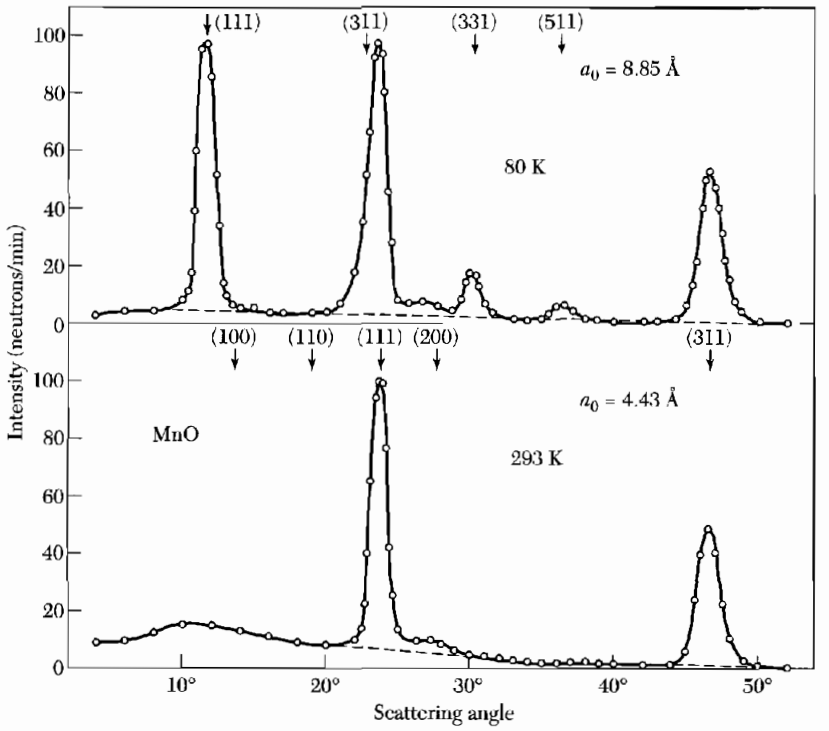


Figure 17 Neutron diffraction patterns for MnO below and above the spin-ordering temperature of 120 K, after C. G. Shull, W. A. Strauser, and E. O. Wollan. The reflection indices are based on an 8.85 Å cell at 80 K and on a 4.43 Å cell at 293 K. At the higher temperature the Mn^{2+} ions are still magnetic, but they are no longer ordered.

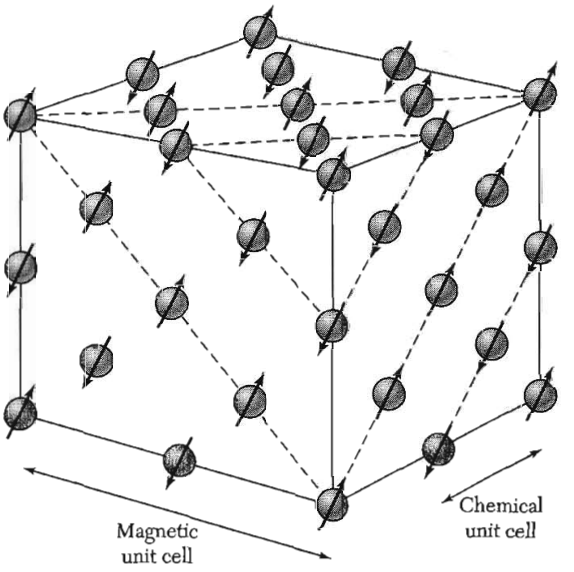


Figure 18 Ordered arrangements of spins of the Mn^{2+} ions in manganese oxide, MnO, as determined by neutron diffraction. The O^{2+} ions are not shown.

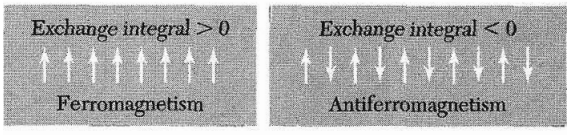


Figure 19 Spin ordering in ferromagnets ($J > 0$) and antiferromagnets ($J < 0$).

the Mn^{2+} ions are ordered in some nonferromagnetic arrangement. If the ordering were ferromagnetic, the chemical and magnetic cells would give the same reflections.

The spin arrangement shown in Fig. 18 is consistent with the neutron diffraction results and with magnetic measurements. The spins in a single [111] plane are parallel, but spins in adjacent [111] planes are antiparallel. Thus MnO is an antiferromagnet, as in Fig. 19.

In an **antiferromagnet** the spins are ordered in an antiparallel arrangement with zero net moment at temperatures below the ordering or **Néel temperature** (Table 2). The susceptibility of an antiferromagnet is not infinite at $T = T_N$, but has a weak cusp, as in Fig. 20.

An antiferromagnet is a special case of a ferrimagnet for which both sublattices A and B have equal saturation magnetizations. Thus $C_A = C_B$ in (37), and the Néel temperature in the mean field approximation is given by

$$T_N = \mu C \ , \tag{38}$$

Table 2 Antiferromagnetic crystals

Substance	Paramagnetic ion lattice	Transition temperature, T_N, in K	Curie-Weiss θ, in K	$\frac{\theta}{T_N}$	$\frac{\chi(0)}{\chi(T_N)}$
MnO	fcc	116	610	5.3	$\frac{2}{3}$
MnS	fcc	160	528	3.3	0.82
MnTe	hex. layer	307	690	2.25	
MnF_2	bc tetr.	67	82	1.24	0.76
FeF_2	bc tetr.	79	117	1.48	0.72
$FeCl_2$	hex. layer	24	48	2.0	<0.2
FeO	fcc	198	570	2.9	0.8
$CoCl_2$	hex. layer	25	38.1	1.53	
CoO	fcc	291	330	1.14	
$NiCl_2$	hex. layer	50	68.2	1.37	
NiO	fcc	525	~2000	~4	
Cr	bcc	308			

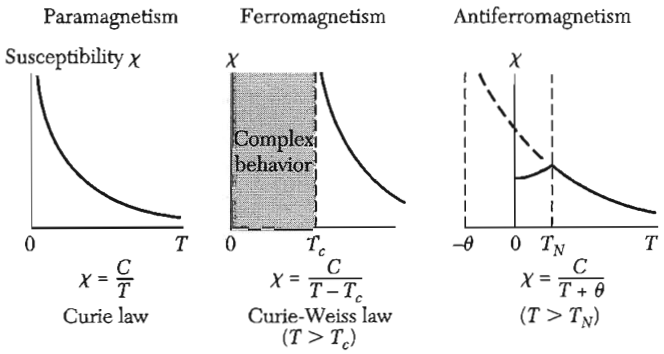


Figure 20 Temperature dependence of the magnetic susceptibility in paramagnets, ferromagnets, and antiferromagnets. Below the Néel temperature of an antiferromagnet the spins have antiparallel orientations; the susceptibility attains its maximum value at T_N where there is a well-defined kink in the curve of χ versus T. The transition is also marked by peaks in the heat capacity and the thermal expansion coefficient.

where C refers to a single sublattice. The susceptibility in the paramagnetic region $T > T_N$ is obtained from (37):

$$\chi = \frac{2CT - 2\mu C^2}{T^2 - (\mu C)^2} = \frac{2C}{T + \mu C} = \frac{2C}{T + T_N} \,. \tag{39}$$

The experimental results at $T > T_N$ are of the form

(CGS) $$\chi = \frac{2C}{T + \theta} \,. \tag{40}$$

Experimental values of θ/T_N listed in Table 2 often differ substantially from the value unity expected from (39). Values of θ/T_N of the observed magnitude may be obtained when next-nearest-neighbor interactions are provided for, and when possible sublattice arrangements are considered. If a mean field constant $-\epsilon$ is introduced to describe interactions within a sublattice, then $\theta/T_N = (\mu + \epsilon)/(\mu - \epsilon)$.

Susceptibility Below the Néel Temperature

There are two situations: with the applied magnetic field perpendicular to the axis of the spins; and with the field parallel to the axis of the spins. At and above the Néel temperature the susceptibility is nearly independent of the direction of the field relative to the spin axis.

For $\mathbf{B}_a$ perpendicular to the axis of the spins we can calculate the susceptibility by elementary considerations. The energy density in the presence of the field is, with $M = |M_A| = |M_B|$,

$$U = \mu \mathbf{M}_A \cdot \mathbf{M}_B - \mathbf{B}_a \cdot (\mathbf{M}_A + \mathbf{M}_B) \cong -\mu M^2 [1 - \tfrac{1}{2}(2\varphi)^2] - 2B_a M\varphi \ , \qquad (41)$$

where 2φ is the angle the spins make with each other (Fig. 21a). The energy is a minimum when

$$dU/d\varphi = 0 = 4\mu M^2 \varphi - 2B_a M \ ; \quad \varphi = B_a/2\mu M \ , \qquad (42)$$

so that

$$\text{(CGS)} \qquad \chi_\perp = 2M\varphi/B_a = 1/\mu \ . \qquad (43)$$

In the parallel orientation (Fig. 21b) the magnetic energy is not changed if the spin systems A and B make equal angles with the field. Thus the susceptibility at $T = 0$ K is zero:

$$\chi_\parallel(0) = 0 \ . \qquad (44)$$

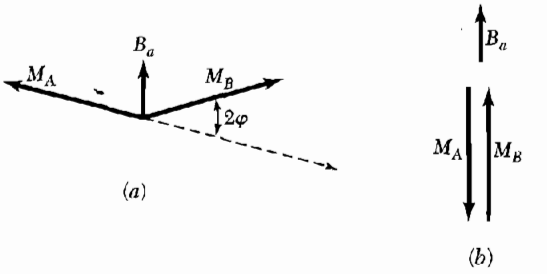


Figure 21 Calculation of (a) perpendicular and (b) parallel susceptibilities at 0 K, in the mean field approximation.

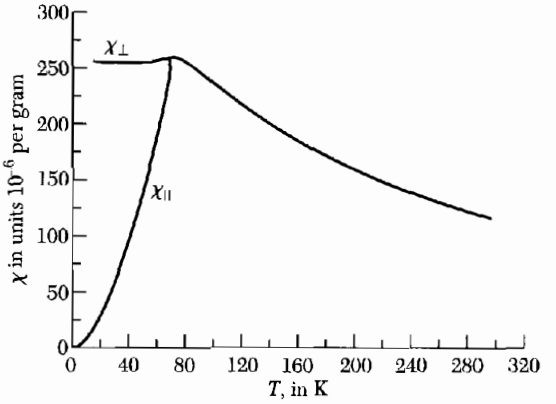


Figure 22 Magnetic susceptibility of manganese fluoride, MnF_2, parallel and perpendicular to the tetragonal axis. (After S. Foner.)

The parallel susceptibility increases smoothly with temperature up to T_N. Measurements on MnF_2 are shown in Fig. 22. In very strong fields the spin systems will turn discontinuously from the parallel orientation to the perpendicular orientation where the energy is lower.

Antiferromagnetic Magnons

We obtain the dispersion relation of magnons in a one-dimensional antiferromagnet by making the appropriate substitutions in the treatment (16)–(22) of the ferromagnetic line. Let spins with even indices $2p$ compose sublattice A, that with spins up ($S^z = S$); and let spins with odd indices $2p + 1$ compose sublattice B, that with spins down ($S^z = -S$).

We consider only nearest-neighbor interactions, with J negative. Then (18) written for A becomes, with a careful look at (17),

$$dS^x_{2p}/dt = (2JS/\hbar)(-2S^y_{2p} - S^y_{2p-1} - S^y_{2p+1}) \ ; \tag{45a}$$

$$dS^y_{2p}/dt = -(2JS/\hbar)(-2S^x_{2p} - S^x_{2p-1} - S^x_{2p+1}) \ . \tag{45b}$$

The corresponding equations for a spin on B are

$$dS^x_{2p+1}/dt = (2JS/\hbar)(2S^y_{2p+1} + S^y_{2p} + S^y_{2p+2}) \ ; \tag{46a}$$

$$dS^y_{2p+1}/dt = -(2JS/\hbar)(2S^x_{2p+1} + S^x_{2p} + S^x_{2p+2}) \ . \tag{46b}$$

We form $S^+ = S^x + iS^y$; then

$$dS^+_{2p}/dt = (2iJS/\hbar)(2S^+_{2p} + S^+_{2p-1} + S^+_{2p+1}) \ ; \tag{47}$$

$$dS^+_{2p+1}/dt = -(2iJS/\hbar)(2S^+_{2p+1} + S^+_{2p} + S^+_{2p+2}) \ . \tag{48}$$

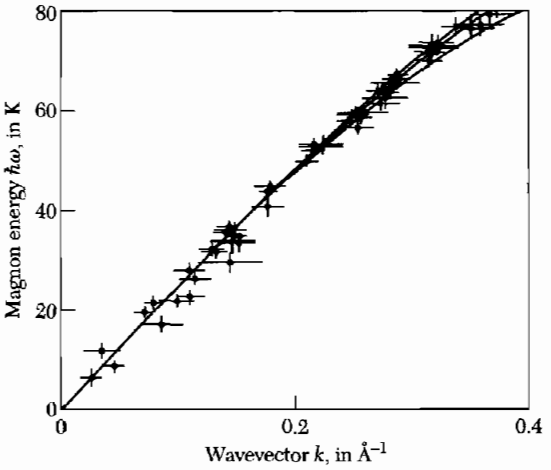


Figure 23 Magnon dispersion relation in the simple cubic antiferromagnet $RbMnF_3$ as determined at 4.2 K by inelastic neutron scattering. (After C. G. Windsor and R. W. H. Stevenson.)

We look for solutions of the form

$$S_{2p}^{+} = u \exp[i2pka - iwt] \; ; \quad S_{2p+1}^{+} = v \exp[i(2p+1)\, ka - iwt] \; , \tag{49}$$

so that (47) and (48) become, with $\omega_{ex} \equiv -4JS/\hbar = 4|J|S/\hbar$,

$$\omega u = \tfrac{1}{2}\omega_{ex}(2u + ve^{-ika} + ve^{ika}) \; ; \tag{50a}$$

$$-\omega v = \tfrac{1}{2}\omega_{ex}(2v + ue^{-ika} + ue^{ika}) \; . \tag{50b}$$

Equations (50) have a solution if

$$\begin{vmatrix} \omega_{ex} - \omega & \omega_{ex}\cos ka \\ \omega_{ex}\cos ka & \omega_{ex} + \omega \end{vmatrix} = 0 \; ; \tag{51}$$

thus

$$\omega^2 = \omega_{ex}^2(1 - \cos^2 ka) \; ; \quad \omega = \omega_{ex}|\sin ka| \; . \tag{52}$$

The dispersion relation for magnons in an antiferromagnet is quite different from (22) for magnons in a ferromagnet. For $ka \ll 1$ we see that (52) is linear in k: $\omega \cong \omega_{ex}|ka|$. The magnon spectrum of $RbMnF_3$ is shown in Fig. 23, as determined by inelastic neutron scattering experiments. There is a large region in which the magnon frequency is linear in the wavevector.

Well-resolved magnons have been observed in MnF_2 at specimen temperatures up to 0.93 of the Néel temperature. Thus even at high temperatures the magnon approximation is useful.

FERROMAGNETIC DOMAINS

At temperatures well below the Curie point the electronic magnetic moments of a ferromagnet are essentially parallel when regarded on a microscopic scale. Yet, looking at a specimen as a whole, the magnetic moment may be very much less than the saturation moment, and the application of an external magnetic field may be required to saturate the specimen. The behavior observed in polycrystalline specimens is similar to that in single crystals.

Actual specimens are composed of small regions called domains, within each of which the local magnetization is saturated. The directions of magnetization of different domains need not be parallel. An arrangement of domains with approximately zero resultant magnetic moment is shown in Fig. 24. Domains form also in antiferromagnetics, ferroelectrics, antiferroelectrics, ferroelastics, superconductors, and sometimes in metals under conditions of a strong de Haas-van Alphen effect. The increase in the gross magnetic moment of a ferromagnetic specimen in an applied magnetic field takes place by two independent processes:

- In weak applied fields the volume of domains (Fig. 25) favorably oriented with respect to the field increases at the expense of unfavorably oriented domains;

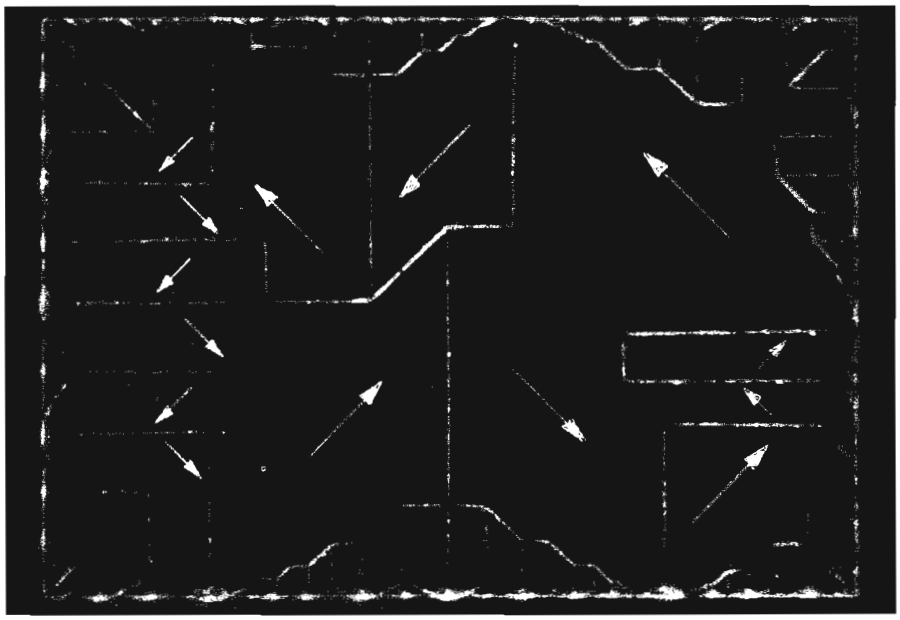


Figure 24 Ferromagnetic domain pattern on a single crystal platelet of nickel. The domain boundaries are made visible by the Bitter magnetic powder pattern technique. The direction of magnetization within a domain is determined by observing growth or contraction of the domain in a magnetic field. (After R. W. De Blois.)

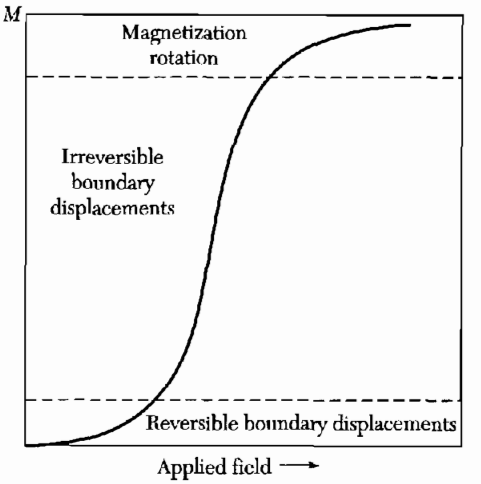


Figure 25 Representative magnetization curve, showing the dominant magnetization processes in the different regions of the curve.

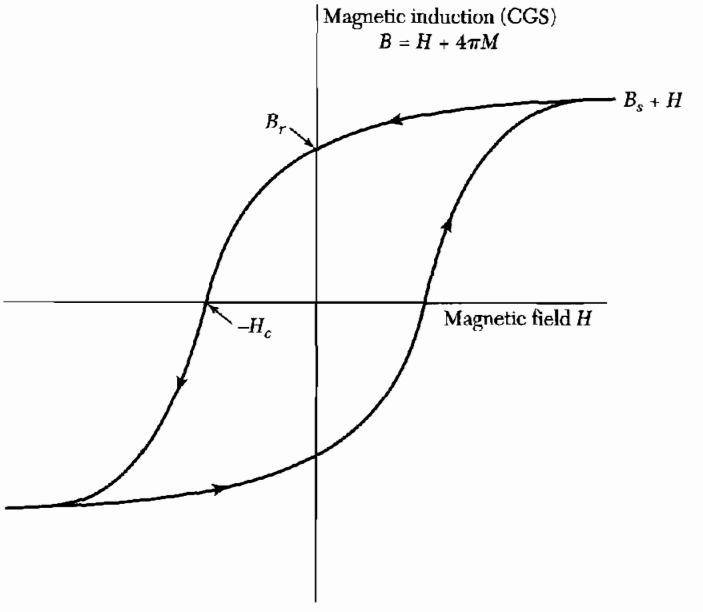


Figure 26 The technical magnetization curve (or hysteresis loop). The coercivity H_c is the reverse field that reduces B to zero; a related coercivity H_{ci} reduces M or $B - H$ to zero. The **remanence** B_r is the value of B at $H = 0$. The saturation induction B_s is the limit of $B - H$ at large H, and the saturation magnetization $M_s = B_s/4\pi$. In SI the vertical axis is $B = \mu_0(H + M)$.

- In strong applied fields the domain magnetization rotates toward the direction of the field.

Technical terms defined by the hysteresis loop are shown in Fig. 26. The coercivity is usually defined as the reverse field H_c that reduces the induction

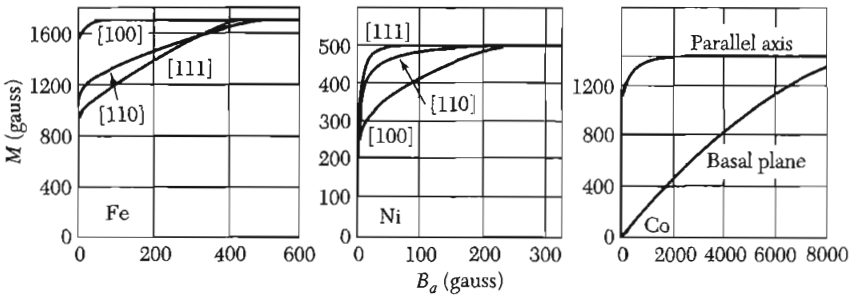


Figure 27 Magnetization curves for single crystals of iron, nickel, and cobalt. From the curves for iron we see that the [100] directions are easy directions of magnetization and the [111] directions are hard directions. The applied field is B_a. (After Honda and Kaya.)

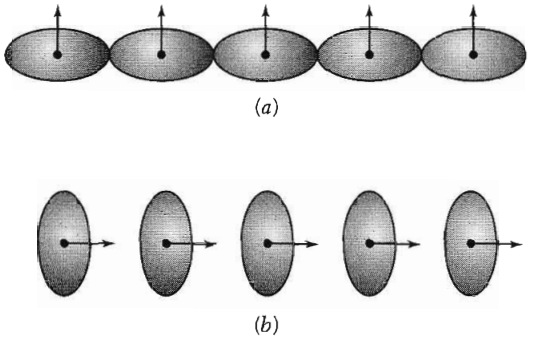


Figure 28 Asymmetry of the overlap of electron distributions on neighboring ions provides one mechanism of magnetocrystalline anisotropy. Because of spin-orbit interaction the charge distribution is spheroidal and not spherical. The asymmetry is tied to the direction of the spin, so that a rotation of the spin directions relative to the crystal axes changes the exchange energy and also changes the electrostatic interaction energy of the charge distributions on pairs of atoms. Both effects give rise to an anisotropy energy. The energy of (a) is not the same as the energy of (b).

B to zero, starting from saturation. In high coercivity materials the coercivity H_{ci} is defined as the reverse field that reduces the magnetization M to zero.

Anisotropy Energy

There is an energy in a ferromagnetic crystal which directs the magnetization along certain crystallographic axes called directions of easy magnetization. This energy is called the **magnetocrystalline** or **anisotropy energy**. It does not come about from the pure isotropic exchange interaction considered thus far.

Cobalt is a hexagonal crystal. The hexagonal axis is the direction of easy magnetization at room temperature, as shown in Fig. 27. One origin of the anisotropy energy is illustrated by Fig. 28. The magnetization of the crystal sees the crystal lattice through orbital overlap of the electrons: the spin interacts

with the orbital motion by means of the spin-orbit coupling. In cobalt the anisotropy energy density is given by

$$U_K = K_1' \sin^2\theta + K_2' \sin^4\theta \ , \tag{53}$$

where θ is the angle the magnetization makes with the hexagonal axis. At room temperature $K_1' = 4.1 \times 10^6$ erg/cm^3; $K_2' = 1.0 \times 10^6$ erg/cm^3.

Iron is a cubic crystal, and the cube edges are the directions of easy magnetization. To represent the anisotropy energy of iron magnetized in an arbitrary direction with direction cosines α_1, α_2, α_3 referred to the cube edges, we are guided by cubic symmetry. The expression for the anisotropy energy must be an even power of each α_i, provided opposite ends of a crystal axis are equivalent magnetically, and it must be invariant under interchanges of the α_i among themselves. The lowest order combination satisfying the symmetry requirements is $\alpha_1^2 + \alpha_2^2 + \alpha_3^2$, but this is identically equal to unity and does not describe anisotropy effects. The next combination is of the fourth degree: $\alpha_1^2\alpha_2^2 + \alpha_1^2\alpha_3^2 + \alpha_3^2\alpha_2^2$, and then of the sixth degree: $\alpha_1^2\alpha_2^2\alpha_3^2$. Thus

$$U_K = K_1(\alpha_1^2\alpha_2^2 + \alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2) + K_2\alpha_1^2\alpha_2^2\alpha_3^2 \ . \tag{54}$$

At room temperature in iron $K_1 = 4.2 \times 10^5$ erg/cm^3 and $K_2 = 1.5 \times 10^5$ erg/cm^3.

Transition Region Between Domains

A **Bloch wall** in a crystal is the transition layer that separates adjacent regions (domains) magnetized in different directions. The entire change in spin direction between domains does not occur in one discontinuous jump across a single atomic plane, but takes place in a gradual way over many atomic planes (Fig. 29). The exchange energy is lower when the change is distributed over many spins. This behavior may be understood by interpreting the Heisenberg equation (6) classically. We replace cos φ by $1 - \frac{1}{2}\varphi^2$; then $w_{ex} = JS^2\varphi^2$ is the exchange energy between two spins making a small angle φ with each other. Here J is the exchange integral and S is the spin quantum number; w_{ex} is referred to the energy for parallel spins.

If a total change of π occurs in N equal steps, the angle between neighboring spins is π/N, and the exchange energy per pair of neighboring atoms is $w_{ex} = JS^2(\pi/N)^2$. The total exchange energy of a line of $N + 1$ atoms is

$$Nw_{ex} = JS^2\pi^2/N \ . \tag{55}$$

The wall would thicken without limit were it not for the anisotropy energy, which acts to limit the width of the transition layer. The spins contained within the wall are largely directed away from the axes of easy magnetization, so there is an anisotropy energy associated with the wall, roughly proportional to the wall thickness.

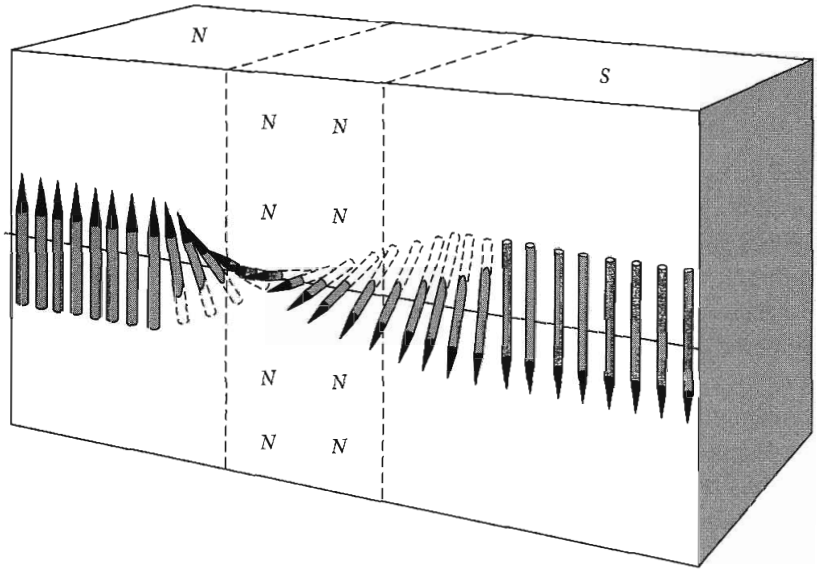


Figure 29 The structure of the Bloch wall separating domains. In iron the thickness of the transition region is about 300 lattice constants.

Consider a wall parallel to the cube face of a simple cubic lattice and separating domains magnetized in opposite directions. We wish to determine the number N of atomic planes contained within the wall. The energy per unit area of wall is the sum of contributions from exchange and anisotropy energies: $\sigma_w = \sigma_{ex} + \sigma_{anis}$.

The exchange energy is given approximately by (55) for each line of atoms normal to the plane of the wall. There are $1/a^2$ such lines per unit area, where a is the lattice constant. Thus $\sigma_{ex} = \pi^2 JS^2/Na^2$ per unit area of wall.

The anisotropy energy is of the order of the anisotropy constant times the thickness Na, or $\sigma_{anis} \approx KNa$; therefore

$$\sigma_w \approx (\pi^2 JS^2/Na^2) + KNa \ . \tag{56}$$

This is a minimum with respect to N when

$$\partial\sigma_w/\partial N = 0 = -(\pi^2 JS^2/N^2a^2) + Ka \ ; \tag{57}$$

or

$$N = (\pi^2 JS^2/Ka^3)^{1/2} \ . \tag{58}$$

For order of magnitude, $N \approx 300$ in iron.

The total wall energy per unit area on our model is

$$\sigma_w = 2\pi(KJS^2/a)^{1/2} \ ; \tag{59}$$

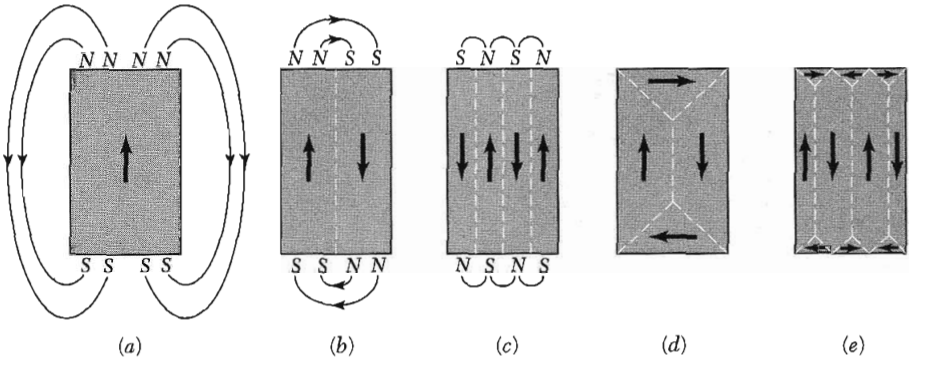


Figure 30 The origin of domains.

in iron $\sigma_w \approx 1$ erg/cm^2. Accurate calculation for a 180° wall in a (100) plane gives $\sigma_w = 2(2K_1 JS^2/a)^{1/2}$.

Origin of Domains

Landau and Lifshitz showed that domain structure is a natural consequence of the various contributions to the energy—exchange, anisotropy, and magnetic—of a ferromagnetic body.

Direct evidence of domain structure is furnished by photomicrographs of domain boundaries obtained by the technique of magnetic powder patterns and by optical studies using Faraday rotation. The powder pattern method developed by F. Bitter consists in placing a drop of a colloidal suspension of finely divided ferromagnetic material, such as magnetite, on the surface of the ferromagnetic crystal. The colloid particles in the suspension concentrate strongly about the boundaries between domains where strong local magnetic fields exist which attract the magnetic particles. The discovery of transparent ferromagnetic compounds has encouraged the use also of optical rotation for domain studies.

We may understand the origin of domains by considering the structures shown in Fig. 30, each representing a cross section through a ferromagnetic single crystal. In (a) we have a single domain; as a consequence of the magnetic "poles" formed on the surfaces of the crystal this configuration will have a high value of the magnetic energy $(1/8\pi) \int B^2 \, dV$. The magnetic energy density for the configuration shown will be of the order of $M_s^2 \simeq 10^6$ erg/cm^3; here M_s denotes the saturation magnetization, and the units are CGS.

In (b) the magnetic energy is reduced by roughly one-half by dividing the crystal into two domains magnetized in opposite directions. In (c) with N domains the magnetic energy is reduced to approximately $1/N$ of the magnetic energy of (a), because of the reduced spatial extension of the field.

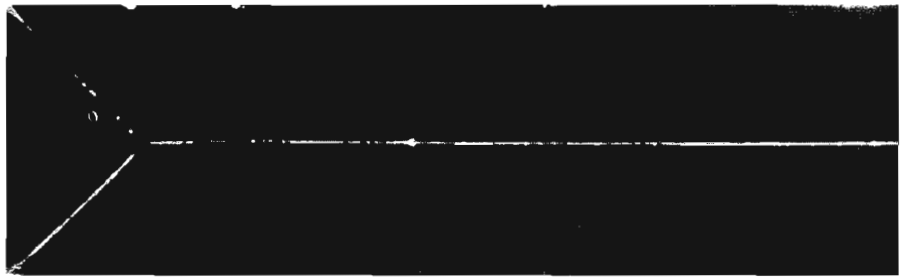


Figure 31 Domain of closure at the end of a single crystal iron whisker. The face is a (100) plane; the whisker axis is [001]. (Courtesy of R. V. Coleman, C. G. Scott, and A. Isin.)

In domain arrangements such as (d) and (e) the magnetic energy is zero. Here the boundaries of the triangular prism domains near the end faces of the crystal make equal angles (45°) with the magnetization in the rectangular domains and with the magnetization in the domains of closure. The component of magnetization normal to the boundary is continuous across the boundary and there is no magnetic field associated with the magnetization. The flux circuit is completed within the crystal—thus giving rise to the term **domains of closure** for surface domains that complete the flux circuit, as in Fig. 31.

Domain structures are often more complicated than our simple examples, but *domain structure always has its origin in the possibility of lowering the energy of a system by going from a saturated configuration with high magnetic energy to a domain configuration with a lower energy.*

Coercivity and Hysteresis

The coercivity is the magnetic field H_c required to reduce the magnetization or the induction B to zero (Fig. 26). The value of the coercivity ranges over seven orders of magnitude; it is the most sensitive property of ferromagnetic materials which is subject to control. The coercivity may vary from 600 G in a loudspeaker permanent magnet (Alnico V) and 10,000 G in a special high stability magnet ($SmCo_5$) to 0.5 G in a commercial power transformer (Fe-Si 4 wt. pet.) and 0.002 G in a pulse transformer (Supermalloy). Low coercivity is desired in a transformer, for this means low hysteresis loss per cycle of operation. Materials with low coercivity are called **soft**; those with high coercivity are called **hard**, although there is not necessarily a 1 : 1 relationship of magnetic hardness with mechanical hardness.

The coercivity decreases as the impurity content decreases and also as internal strains are removed by annealing (slow cooling). Amorphous ferromagnetic alloys may have low coercivity, low hysteresis losses, and high permeability. Alloys that contain a precipitated phase may have a high coercivity, as in Alnico V (Fig. 32).

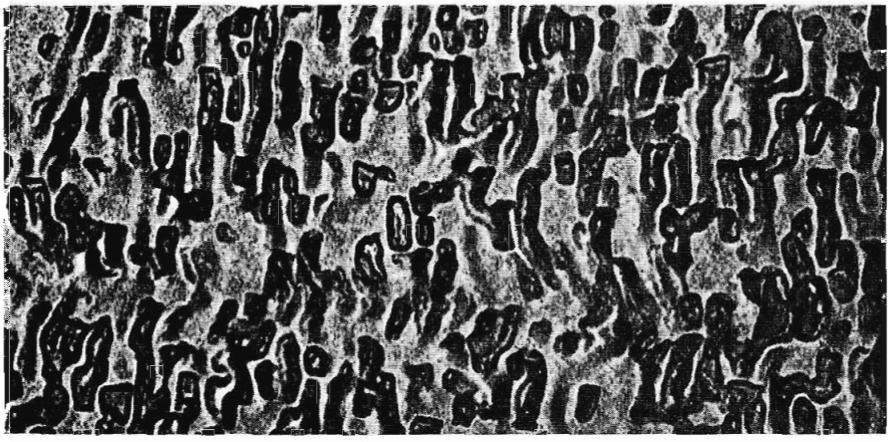


Figure 32 Microstructure of Alnico V in its optimum state as a permanent magnet. The composition of Alnico V is, by weight percent, 8 Al, 14 Ni, 24 Co, 3 Cu, 51 Fe. As a permanent magnet it is a two-phase system, with fine particles of one phase embedded in the other phase. The precipitation is carried out in a magnetic field, and the particles are oriented with their long axis parallel to the field direction. The width shown is 1.1 μm. (Courtesy of F. E. Luborsky.)

Soft magnetic materials are used to concentrate and shape magnetic flux, as in motors, generators, transformers, and sensors. Useful soft materials include electrical steels (usually alloyed with several percent of silicon to increase electrical resistivity and to decrease anisotropy); various alloys of Fe-Co-Mn, starting with permalloys of composition near $Ni_{78}Fe_{22}$, which have near-zero anisotropy energy and near-zero magnetostriction; NiZn and MnZn ferrites; and metallic glasses produced by rapid solidification. A commercial metallic glass (METGLAS 2605S-2) with composition $Fe_{79}B_{13}Si_9$ has a hysteresis loss per cycle much lower than the best grain-oriented silicon steel.

The high coercivity of materials composed of very small grains or fine powders is well understood. A sufficiently small particle, with diameter less than 10^{-5} or 10^{-6} cm, is always magnetized to saturation as a single domain because the formation of a flux-closure configuration is energetically unfavorable. In a single domain particle it is not possible for magnetization reversal to take place by means of the process of boundary displacement, which usually requires relatively weak fields. Instead the magnetization of the particle must rotate as a whole, a process that may require large fields depending on the anisotropy energy of the material and the anisotropy of the shape of the particle.

The coercivity of fine iron particles is expected theoretically to be about 500 gauss on the basis of rotation opposed by the crystalline anisotropy energy, and this is of the order of the observed value. Higher coercivities have been

reported for elongated iron particles, the rotation here being opposed by the shape anisotropy of the demagnetization energy.

Rare earth metals in alloys with Mn, Fe, Co, and Ni have very large crystal anisotropies K and correspondingly large coercivities, of the order of $2K/M$. These alloys are exceptionally good permanent magnets. For example, the hexagonal compound $SmCo_5$ has an anisotropy energy 1.1×10^8 erg cm^{-3}, equivalent to a coercivity $2K/M$ of 290 kG (29 T). Magnets of $Nd_2Fe_{14}B$ have energy products as high as 50 MGOe, exceeding all other commercially available magnets.

SINGLE-DOMAIN PARTICLES

The dominant industrial and commercial applications of ferromagnetism are in magnetic recording devices, where the magnetic material is in the form of single-domain particles or regions. The total value of the production of magnetic devices for recording may be comparable with the total value of semiconductor device production and greatly exceeds the value of superconducting device production, the latter being held back by low critical temperatures, as compared with magnetic Curie temperatures. The magnetic recording devices or memories typically are in the form of hard disks in computers and tape in video and audio recorders.

An ideal single-domain particle is a fine particle, usually elongated, that has its magnetic moment directed toward one end or the other of the particle. The alternative orientations may be labeled as N or S; + or −; in digital recording, as 0 or 1. To have digital properties a ferromagnetic particle should be fine enough, typically 10–100 nm, so that only one domain is within the particle. If the fine particle is elongated (acicular) or has uniaxial crystal symmetry, only two values of the magnetic moment of the single domain are permitted, which is what one wants for digital properties. The first successful recording material was acicular τ-Fe_2O_3 with length-to-width ratio of about 5 : 1, coercivity near 200 Oe and a length $<1\,\mu$m; chromium dioxide CrO_2 is the basis of a better material, in a form highly acicular (20 : 1) with coercivity near 500 Oe.

Effective elongation can be attained with spheres by making a chain, like a string of beads. An ensemble of such chains or of elongated single domain particles is said to exhibit **superparamagnetism** if the magnetic moment of a unit is constant. If μ is the magnetic moment in a magnetic field B, then the net magnetization of the ensemble will follow the Curie-Brillouin-Langevin law of Chapter 11 if the particles are embedded in a liquid so that they are each free to rotate as a whole. If the particles are frozen in a solid, there will be a remanent magnetization (Fig. 26) after removal of an applied field.

Geomagnetism and Biomagnetism

Single domain ferromagnetic properties are of special geological interest in sedimentary rocks because the rocks through their remanent magnetization carry a memory of the direction of the earth's magnetic field at the time that they were laid down, and thus of the geographical location of the rocks at that epoch. The magnetic record is perhaps the most important basis of the theory of the drift of continents. Annually, layers of sediment are deposited in stream beds, layers that may bear some magnetic particles in single domain form. This record persists over at least 500 million years of geological time and can tell us where on the surface of the earth the deposit was laid down at a given time. Lava flows also record magnetic field directions.

The change in magnetization from layer to layer gives a superb historical record of the drift of the continental plates on the earth's surface. The paleomagnetic record is one basis of the branch of geology called plate tectonics. The original interpretation of the record was made more difficult, or more exciting, by the associated discovery (Brunhes, 1906) that the magnetic field of the earth itself can show reversals in direction, an effect contained within the standard dynamo theory of the earth's magnetism. Reversals have taken place once every 1×10^4 to 25×10^6 years. When a reversal occurs, it is relatively sudden.

Fine single domain particles, often of magnetite Fe_3O_4, are even of importance in biology. A direction-seeking effect known as **magnetotaxis** often controls, possibly sometimes along with an astronomical guide system, the motion of bacteria, the migration of birds, and the movements of homing pigeons and bees. The effect is due to the interaction of a single domain particle (or cluster of such particles, Fig. 33) in the organism with the external magnetic field of the earth.

Magnetic Force Microscopy

The success of the scanning tunneling microscope (STM) stimulated the development of related scanning probe devices, of which the scanning magnetic force microscope is one of the most effective. A sharp tip of a magnetic material, such as nickel, is mounted on a cantilever lever (Fig. 34). Ideally, but not yet, the tip is a single domain particle. Forces from the magnetic sample act on the tip and cause a change, such as a deflection, in the cantilever status, and an image is formed by scanning the sample relative to the tip. The magnetic force microscope (MFM) is the only magnetic imaging technique that can provide high resolution (10–100 nm) with little surface preparation. One can, for example, observe and image the magnetic flux that exits from the surface at the intersection of a Bloch wall with the surface (Fig. 29). An important application is to the study of magnetic recording media—Figure 35 shows the magnetic signal from a test pattern of 2 μm bits magnetized in the plane of

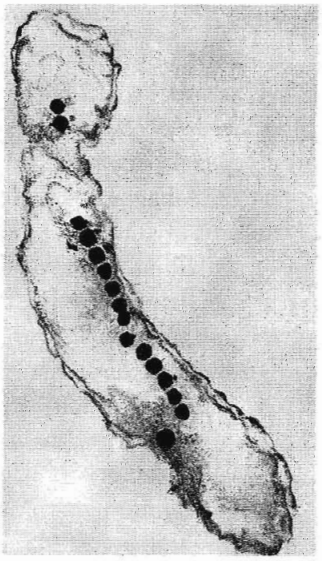


Figure 33 Thin section of a cell of a magnetotactic bacterium showing a chain of 50 nm particles of Fe_3O_4. Drawing by Marta Puebla from a photograph by R. B. Frankel and others.

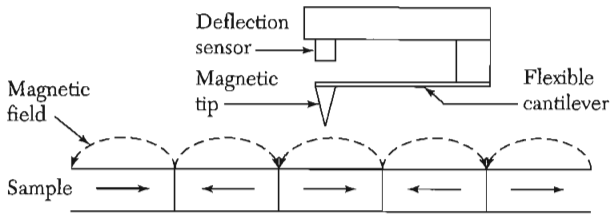


Figure 34 Basic concept of magnetic force microscopy. A magnetic tip attached to a flexible cantilever is used to detect the magnetic field produced by the regions of alternating magnetization in the plane of the sample. (After Gruetter, Mamin, and Rugar, 1992.)

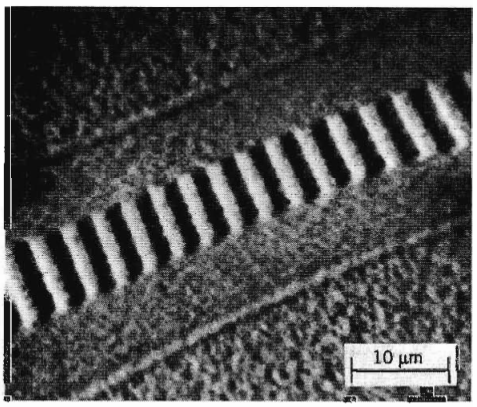


Figure 35 Test strip magnetization in the plane of a Co-alloy disk in 2 μm bits, as detected by MFM close above the plane of the disk. (After Rugar et al.)

a Co-alloy disk; the parallel component of the field seen by the sensor tip is what the photo shows.

SUMMARY

(In CGS Units)

- The susceptibility of a ferromagnet above the Curie temperature has the form $\chi = C/(T - T_c)$ in the mean field approximation.
- In the mean field approximation the effective magnetic field seen by a magnetic moment in a ferromagnet is $\mathbf{B}_a + \lambda\mathbf{M}$, when $\lambda = T_c/C$ and $\mathbf{B}_a$ is the applied magnetic field.
- The elementary excitations in a ferromagnet are magnons. Their dispersion relation for $ka \ll 1$ has the form $\hbar\omega \approx Jk^2a^2$ in zero external magnetic field. The thermal excitation of magnons leads at low temperatures to a heat capacity and to a fractional magnetization change both proportional to $T^{3/2}$.
- In an antiferromagnet two spin lattices are equal, but antiparallel. In a ferrimagnet two lattices are antiparallel, but the magnetic moment of one is larger than the magnetic moment of the other.
- In an antiferromagnet the susceptibility above the Néel temperature has the form $\chi = 2C/(T + \theta)$.
- The magnon dispersion relation in an antiferromagnet has the form $\hbar\omega \approx Jka$. The thermal excitation of magnons leads at low temperatures to a term in T^3 in the heat capacity, in addition to the phonon term in T^3.
- A Bloch wall separates domains magnetized in different directions. The thickness of a wall is $\approx(J/Ka^3)^{1/2}$ lattice constants, and the energy per unit area is $\approx(KJ/a)^{1/2}$, where K is the anisotropy energy density.

Problems

1. ***Magnon dispersion relation.*** Derive the magnon dispersion relation (24) for a spin S on a simple cubic lattice, $z = 6$. Hint: Show first that (18a) is replaced by

$$dS_{\boldsymbol{\rho}}^x/dt = (2JS/\hbar)(6S_{\boldsymbol{\rho}}^y - \sum_{\boldsymbol{\delta}} S_{\boldsymbol{\rho}+\boldsymbol{\delta}}^y) \ ,$$

where the central atom is at $\boldsymbol{\rho}$ and the six nearest neighbors are connected to it by six vectors $\boldsymbol{\delta}$. Look for solutions of the equations for $dS_{\boldsymbol{\rho}}^x/dt$ and $dS_{\boldsymbol{\rho}}^y/dt$ of the form $\exp(i\mathbf{k} \cdot \boldsymbol{\rho} - i\omega t)$.

2. ***Heat capacity of magnons.*** Use the approximate magnon dispersion relation $\omega = Ak^2$ to find the leading term in the heat capacity of a three-dimensional ferromagnet at low temperatures $k_BT \ll J$. The result is $0.113\ k_B(k_BT/\hbar A)^{3/2}$, per unit

volume. The zeta function that enters the result may be estimated numerically; it is tabulated in Jahnke-Emde.

3. ***Néel temperature.*** Taking the effective fields on the two-sublattice model of an antiferromagnetic as

$$B_A = B_a - \mu M_B - \epsilon M_A \ ; \quad B_B = B_a - \mu M_A - \epsilon M_B \ ,$$

show that

$$\frac{\theta}{T_N} = \frac{\mu + \epsilon}{\mu - \epsilon} \ .$$

4. ***Magnetoelastic coupling.*** In a cubic crystal the elastic energy density in terms of the usual strain components e_{ij} is (Chapter 3)

$$U_{el} = \tfrac{1}{2}C_{11}(e_{xx}^2 + e_{yy}^2 + e_{zz}^2) + \tfrac{1}{2}C_{44}(e_{xy}^2 + e_{yz}^2 + e_{zx}^2) + C_{12}(e_{yy}e_{zz} + e_{xx}e_{zz} + e_{xx}e_{yy}) \ ,$$

and the leading term in the magnetic anisotropy energy density is, from (54),

$$U_K = K_1(\alpha_1^2\alpha_2^2 + \alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2) \ .$$

Coupling between elastic strain and magnetization direction may be taken formally into account by including in the total energy density a term

$$U_c = B_1(\alpha_1^2 e_{xx} + \alpha_2^2 e_{yy} + \alpha_3^2 e_{zz}) + B_2(\alpha_1\alpha_2 e_{xy} + \alpha_2\alpha_3 e_{yz} + \alpha_3\alpha_1 e_{zx})$$

arising from the strain dependence of U_K; here B_1 and B_2 are called magnetoelastic coupling constants. Show that the total energy is a minimum when

$$e_{ii} = \frac{B_1[C_{12} - \alpha_i^2(C_{11} + 2C_{12})]}{[(C_{11} - C_{12})(C_{11} + 2C_{12})]} \ ; \quad e_{ij} = -\frac{B_2\alpha_i\alpha_j}{C_{44}} \quad (i \neq j) \ .$$

This explains the origin of magnetostriction, the change of length on magnetization.

5. ***Coercive force of a small particle.*** (a) Consider a small spherical single-domain particle of a uniaxial ferromagnet. Show that the reverse field along the axis required to reverse the magnetization is $B_a = 2K/M_s$, in CGS units. The coercive force of single-domain particles is observed to be of this magnitude. Take $U_K = K \sin^2\theta$ as the anisotropy energy density and $U_M = -B_a M \cos\theta$ as the interaction energy density with the external field; here θ is the angle between $\mathbf{B}_a$ and $\mathbf{M}$. Hint: Expand the energies for small angles about $\theta = \pi$, and find the value of B_a for which $U_K + U_M$ does not have a minimum near $\theta = \pi$. (b) Show that the magnetic energy of a saturated sphere of diameter d is $\approx M_s^2 d^3$. An arrangement with appreciably less magnetic energy has a single wall in an equatorial plane. The domain wall energy will be $\pi\sigma_w d^2/4$, where σ_w is the wall energy per unit area. *Estimate* for cobalt the critical radius below which the particles are stable as single domains, taking the value of JS^2/a as for iron.

6. ***Saturation magnetization near T_c.*** Show that in the mean field approximation the saturation magnetization just below the Curie temperature has the dominant

temperature dependence $(T_c - T)^{1/2}$. Assume the spin is $\frac{1}{2}$. The result is the same as that for a second-order transition in a ferroelectric crystal, as discussed in Chapter 16. The experimental data for ferromagnets (Table 1) suggest that the exponent is closer to 0.33.

7. ***Néel wall.*** The direction of magnetization change in a domain wall goes from that of the Bloch wall to that of a Néel wall (Fig. 36) in thin films of material of negligible crystalline anisotropy energy, such as Permalloy. The intercept of the Bloch wall with the surface of the film creates a surface region of high demagnetization energy. The Néel wall avoids this intercept contribution, but at the expense of a demagnetization contribution throughout the volume of the wall. The Néel wall becomes energetically favorable when the film becomes sufficiently thin. Consider, however, the energetics of the Néel wall in bulk material of negligible crystalline anisotropy energy. There is now a demagnetization contribution to the wall energy density. By a qualitative argument similar to (56), show that $\sigma_w \approx (\pi^2 JS^2/Na^2) + (2\pi M_s^2 Na)$. Find N for which σ_w is a minimum. Estimate the order of magnitude of σ_w for typical values of J, M_s, and a.

8. ***Giant magnetoresistance.*** In a ferromagnetic metal, the conductivity σ_p for electrons whose magnetic moments are oriented parallel to the magnetization is typically larger than σ_a for those antiparallel to the magnetization. Consider a ferromagnetic conductor consisting of two separate regions of identical dimensions in series whose magnetizations can be independently controlled. Electrons of a given spin flow first through one region and then through the other. It is observed that the resistance when both magnetizations point upwards, $R_{\uparrow\uparrow}$, is lower than the resistance when they point opposite, $R_{\uparrow\downarrow}$. This resistance change can be large for $\sigma_p/\sigma_a \gg 1$, and the phenomenon is called giant magnetoresistance (GMR). A small external magnetic field can switch the resistance from $R_{\uparrow\downarrow}$ to $R_{\uparrow\uparrow}$ by reorienting the magnetization of the second layer. This effect is increasingly used in magnetic storage applications such as the magnetic bit readout in hard drives. The giant magnetoresistance ratio is defined as:

$$GMRR = \frac{R_{\uparrow\downarrow} - R_{\uparrow\uparrow}}{R_{\uparrow\uparrow}} .$$

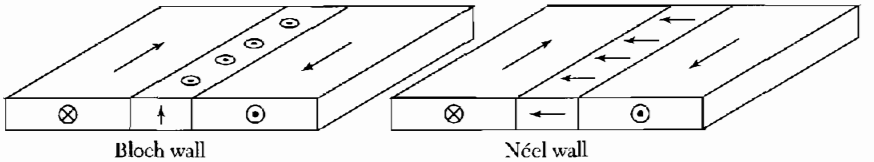


Figure 36 A Bloch wall and a Néel wall in a thin film. The magnetization in the Bloch wall is normal to the plane of the film and adds to the wall energy a demagnetization energy $\sim M_S^2 \delta d$ per unit length of wall, where δ is the wall thickness and d the film thickness. In the Néel wall the magnetization is parallel to the surface; the addition to the wall energy is negligible when $d \ll \delta$. The addition to the Néel wall energy when $d \gg \delta$ is the subject of Problem 7. (After S. Middelhoek.)

(a) If there is no spin-flip scattering for the conduction electrons, show that

$$GMRR = (\sigma_p/\sigma_a + \sigma_a/\sigma_p - 2)/4$$

(Hint: Treat the spin-up and spin-down conduction electrons as independent conducting channels in parallel.) (b) If $\sigma_a \to 0$, explain physically why the resistance in the $\uparrow\downarrow$ magnetization configuration is infinite.

13

Magnetic Resonance

NOTATION: In this chapter the symbols B_a and B_0 refer to the applied field, and B_i is the applied field plus the demagnetizing field. In particular we write $\mathbf{B}_a = B_0\hat{\mathbf{z}}$. For CGS readers it may be simpler to read H for B whenever it occurs in this chapter.

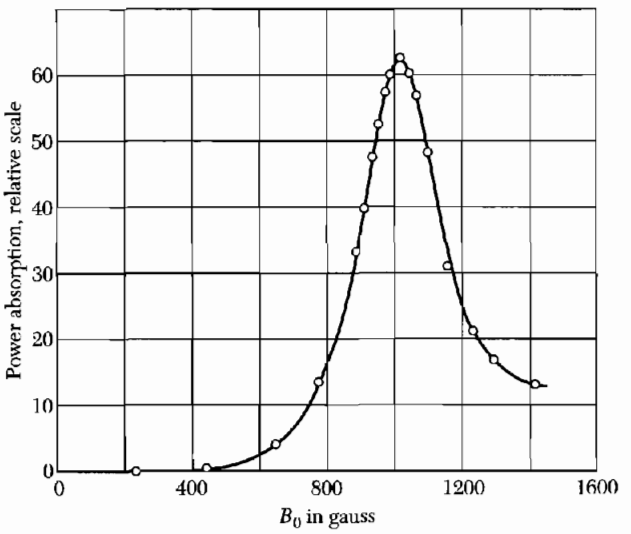


Figure 1 Electron spin resonance absorption in $MnSO_4$ at 298 K at 2.75 GHz, after Zavoisky.

CHAPTER 13: MAGNETIC RESONANCE

In this chapter we discuss dynamical magnetic effects associated with the spin angular momentum of nuclei and of electrons. The principal phenomena are often identified in the literature by their initial letters, such as

NMR: nuclear magnetic resonance
NQR: nuclear quadrupole resonance
EPR or ESR: electron paramagnetic or spin resonance (Fig. 1)
FMR: ferromagnetic resonance
SWR: spin wave resonance (ferromagnetic films)
AFMR: antiferromagnetic resonance
CESR: conduction electron spin resonance

The information that can be obtained about solids by resonance studies may be categorized:

- Electronic structure of single defects, as revealed by the fine structure of the absorption.
- Motion of the spin or of the surroundings, as revealed by changes in the line width.
- Internal magnetic fields sampled by the spin, as revealed by the position of the resonance line (chemical shift; Knight shift).
- Collective spin excitations.

It is best to discuss NMR as a basis for a brief account of the other resonance experiments. A great impact of NMR has been in organic chemistry and biochemistry, where NMR provides a powerful tool for the identification and the structure determination of complex molecules. This success is due to the extremely high resolution attainable in diamagnetic liquids. A major medical application of NMR is magnetic resonance imaging (MRI), which allows the resolution in 3D of abnormal growths, configurations, and reactions in the whole body.

NUCLEAR MAGNETIC RESONANCE

We consider a nucleus that possesses a magnetic moment $\boldsymbol{\mu}$ and an angular momentum $\hbar\mathbf{I}$. The two quantities are parallel, and we may write

$$\boldsymbol{\mu} = \gamma\hbar\mathbf{I} \ ; \tag{1}$$

the magnetogyric ratio γ is constant. By convention $\mathbf{I}$ denotes the nuclear angular momentum measured in units of $\hbar$.

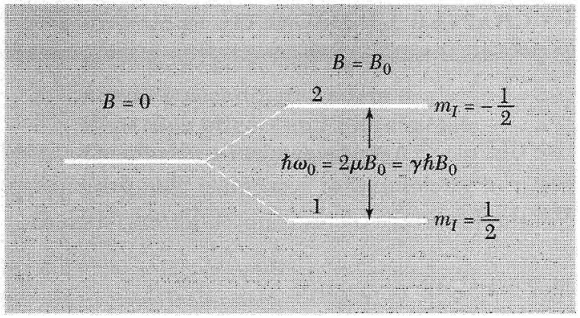


Figure 2 Energy level splitting of a nucleus of spin $I = \frac{1}{2}$ in a static magnetic field B_0.

The energy of interaction with the applied magnetic field is

$$U = -\boldsymbol{\mu} \cdot \mathbf{B}_a \ ; \tag{2}$$

if $\mathbf{B}_a = B_0\hat{\mathbf{z}}$, then

$$U = -\mu_z B_0 = -\gamma\hbar B_0 I_z \ . \tag{3}$$

The allowed values of I_z are $m_I = I, I - 1, \ldots, -I$, and $U = -m_I\gamma\hbar B_0$.

In a magnetic field a nucleus with $I = \frac{1}{2}$ has two energy levels corresponding to $m_I = \pm\frac{1}{2}$, as in Fig. 2. If $\hbar\omega_0$ denotes the energy difference between the two levels, then $\hbar\omega_0 = \gamma\hbar B_0$ or

$$\omega_0 = \gamma B_0 \ . \tag{4}$$

This is the fundamental condition for magnetic resonance absorption.

For the proton[1] $\gamma = 2.675 \times 10^4\ \mathrm{s^{-1}\ gauss^{-1}} = 2.675 \times 10^8\ \mathrm{s^{-1}\ tesla^{-1}}$, so that

$$\boxed{\nu(\mathrm{MHz}) = 4.258\ B_0(\mathrm{kilogauss}) = 42.58\ B_0(\mathrm{tesla}) \ ,} \tag{4a}$$

where ν is the frequency. One tesla is precisely 10^4 gauss. Magnetic data for selected nuclei are given in Table 1. For the electron spin,

$$\boxed{\nu(\mathrm{GHz}) = 2.80\ B_0(\mathrm{kilogauss}) = 28.0\ B_0(\mathrm{tesla}) \ .} \tag{4b}$$

[1]The magnetic moment μ_p of the proton is 1.4106×10^{-23} erg $\mathrm{G^{-1}}$ or 1.4106×10^{-26} J $\mathrm{T^{-1}}$, and $\gamma = 2\mu_p/\hbar$. The **nuclear magneton** μ_n is defined as $e\hbar/2M_pc$ and is equal to 5.0509×10^{-24} erg $\mathrm{G^{-1}}$ or 5.0509×10^{-27} J $\mathrm{T^{-1}}$; thus $\mu_p = 2.793$ nuclear magnetons.

Table 1 Nuclear magnetic resonance data

For every element the most abundant magnetic isotope is shown. After Varian Associates NMR Table.

H^{1}																	He^{3}
1/2																	1/2
99.98																	10^{-6}
2.792																	−2.127
Li^{7}	Be^{9}											B^{11}	C^{13}	N^{14}	O^{17}	F^{19}	Ne^{21}
3/2	3/2											3/2	1/2	1	5/2	1/2	3/2
92.57	100.											81.17	1.108	99.64	0.04	100.	0.257
3.256	−1.177											2.688	0.702	0.404	−1.893	2.627	−0.662
Na^{23}	Mg^{25}	Most abundant isotope with nonzero nuclear spin										Al^{27}	Si^{29}	P^{31}	S^{33}	Cl^{35}	Ar
3/2	5/2	Nuclear spin, in units of										5/2	1/2	1/2	3/2	3/2	
100.	10.05	Natural abundance of isotope, in percent										100.	4.70	100.	0.74	75.4	
2.216	0.855	Nuclear magnetic moment, in units of $e\hbar/2M_pc$										3.639	0.555	1.131	0.643	0.821	
K^{39}	Ca^{43}	Sc^{45}	Ti^{47}	V^{51}	Cr^{53}	Mn^{55}	Fe^{57}	Co^{59}	Ni^{61}	Cu^{63}	Zn^{67}	Ga^{69}	Ge^{73}	As^{75}	Se^{77}	Br^{79}	Kr^{83}
3/2	7/2	7/2	5/2	7/2	3/2	5/2	1/2	7/2	3/2	3/2	5/2	3/2	9/2	3/2	1/2	3/2	9/2
93.08	0.13	100.	7.75	~100	9.54	100.	2.245	100.	1.25	69.09	4.12	60.2	7.61	100.	7.50	50.57	11.55
0.391	−1.315	4.749	0.787	5.139	0.474	3.461	0.090	4.639	0.746	2.221	0.874	2.011	0.877	1.435	0.533	2.099	−0.967
Rb^{85}	Sr^{87}	Y^{89}	Zr^{91}	Nb^{93}	Mo^{95}	Tc	Ru^{101}	Rh^{103}	Pd^{105}	Ag^{107}	Cd^{111}	In^{115}	Sn^{119}	Sb^{121}	Te^{125}	I^{127}	Xe^{129}
5/2	9/2	1/2	5/2	9/2	5/2		5/2	1/2	5/2	1/2	1/2	9/2	1/2	5/2	1/2	5/2	1/2
72.8	7.02	100.	11.23	100.	15.78		16.98	100.	22.23	51.35	12.86	95.84	8.68	57.25	7.03	100.	26.24
1.348	1.089	0.137	1.298	6.144	0.910		−0.69	0.088	−0.57	−0.113	−0.592	5.507	−1.041	3.342	−0.882	2.794	−0.773
Cs^{133}	Ba^{137}	La^{139}	Hf^{177}	Ta^{181}	W^{183}	Re^{187}	Os^{189}	Ir^{193}	Pt^{195}	Au^{197}	Hg^{199}	Tl^{205}	Pb^{207}	Bi^{209}	Po	At	Rn
7/2	3/2	7/2	7/2	7/2	1/2	5/2	3/2	3/2	1/2	3/2	1/2	1/2	1/2	9/2			
100.	11.32	99.9	18.39	100.	14.28	62.93	16.1	61.5	33.7	100.	16.86	70.48	21.11	100.			
2.564	0.931	2.761	0.61	2.340	0.115	3.176	0.651	0.17	0.600	0.144	0.498	1.612	0.584	4.039			
Fr	Ra	Ac															

Ce^{141} ÷	Pr^{141}	Nd^{143}	Pm	Sm^{147}	Eu^{153}	Gd^{157}	Tb^{159}	Dy^{163}	Ho^{165}	Er^{167}	Tm^{169}	Yb^{173}	Lu^{175}
7/2	5/2	7/2		7/2	5/2	3/2	3/2	5/2	7/2	7/2	1/2	5/2	7/2
—	100.	12.20		15.07	52.23	15.64	100.	24.97	100.	22.82	100.	16.08	97.40
0.16	3.92	−1.25		−0.68	1.521	−0.34	1.52	−0.53	3.31	0.48	−0.20	−0.677	2.9
Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

Equations of Motion

The rate of change of angular momentum of a system is equal to the torque that acts on the system. The torque on a magnetic moment $\boldsymbol{\mu}$ in a magnetic field $\mathbf{B}$ is $\boldsymbol{\mu} \times \mathbf{B}$, so that we have the gyroscopic equation

$$\hbar d\mathbf{I}/dt = \boldsymbol{\mu} \times \mathbf{B}_a \ ; \tag{5}$$

or

$$d\boldsymbol{\mu}/dt = \gamma\boldsymbol{\mu} \times \mathbf{B}_a \ . \tag{6}$$

The nuclear magnetization $\mathbf{M}$ is the sum $\Sigma\boldsymbol{\mu}_i$ over all the nuclei in a unit volume. If only a single isotope is important, we consider only a single value of γ, so that

$$d\mathbf{M}/dt = \gamma\mathbf{M} \times \mathbf{B}_a \ . \tag{7}$$

We place the nuclei in a static field $\mathbf{B}_a = B_0\hat{\mathbf{z}}$. In thermal equilibrium at temperature T the magnetization will be along $\hat{\mathbf{z}}$:

$$M_x = 0 \ ; \quad M_y = 0 \ ; \quad M_z = M_0 = \chi_0 B_0 = CB_0/T \ , \tag{8}$$

where the susceptibility is χ_0 and the Curie constant $C = N\mu^2/3k_B$, as in Chapter 11.

The magnetization of a system of spins with $I = \frac{1}{2}$ is related to the population difference $N_1 - N_2$ of the lower and upper levels in Fig. 2: $M_z = (N_1 - N_2)\mu$, where the N's refer to a unit volume. The population ratio in thermal equilibrium is just given by the Boltzmann factor for the energy difference $2\mu B_0$:

$$(N_2/N_1)_0 = \exp(-2\mu B_0/k_B T) \ . \tag{9}$$

The equilibrium magnetization is $M_0 = N\mu \tanh(\mu B/k_B T)$.

When the magnetization component M_z is not in thermal equilibrium, we suppose that it approaches equilibrium at a rate proportional to the departure from the equilibrium value M_0:

$$\frac{dM_z}{dt} = \frac{M_0 - M_z}{T_1} \ . \tag{10}$$

In the standard notation T_1 is called the **longitudinal relaxation time** or the **spin-lattice relaxation time**.

If at $t = 0$ an unmagnetized specimen is placed in a magnetic field $B_0\hat{\mathbf{z}}$, the magnetization will increase from the initial value $M_z = 0$ to a final value $M_z = M_0$. Before and just after the specimen is placed in the field, the population N_1 will be equal to N_2, as appropriate to thermal equilibrium in zero magnetic field. It is necessary to reverse some spins to establish the new equilibrium distribution in the field B_0. On integrating (10):

$$\int_0^{M_z} \frac{dM_z}{M_0 - M_z} = \frac{1}{T_1}\int_0^t dt \ , \tag{11}$$

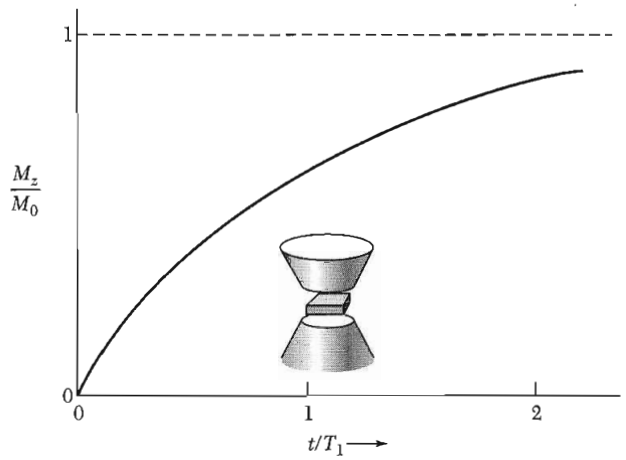


Figure 3 At time $t = 0$ an unmagnetized specimen $M_z(0) = 0$ is placed in a static magnetic field B_0. The magnetization increases with time and approaches the new equilibrium value $M_0 = \chi_0 B_0$. This experiment defines the longitudinal relaxation time T_1. The magnetic energy density $-\mathbf{M} \cdot \mathbf{B}$ decreases as part of the spin population moves into the lower level. The asymptotic value at $t \gg T_1$ is $-M_0 B_0$. The energy flows from the spin system to the system of lattice vibrations; thus T_1 is also called the spin-lattice relaxation time.

or

$$\log \frac{M_0}{M_0 - M_z} = \frac{t}{T_1} \; ; \quad M_z(t) = M_0[1 - \exp(-t/T_1)] \; , \tag{12}$$

as in Fig. 3. The magnetic energy $-\mathbf{M} \cdot \mathbf{B}_a$ decreases as M_z approaches its new equilibrium value.

Typical processes whereby the magnetization approaches equilibrium are indicated in Fig. 4. The dominant spin-lattice interaction of paramagnetic ions in crystals is by the phonon modulation of the crystalline electric field. Relaxation proceeds by three principal processes (Fig. 4b): direct (emission or absorption of a phonon); Raman (scattering of a phonon); and Orbach (intervention of a third state).

Taking account of (10), the z component of the equation of motion (7) becomes

$$\frac{dM_z}{dt} = \gamma(\mathbf{M} \times \mathbf{B}_a)_z + \frac{M_0 - M_z}{T_1} \; , \tag{13a}$$

where $(M_0 - M_z)/T_1$ is an extra term in the equation of motion, arising from the spin-lattice interactions not included in (7). That is, besides precessing about the magnetic field, $\mathbf{M}$ will relax to the equilibrium value $\mathbf{M}_0$.

If in a static field $B_0\hat{\mathbf{z}}$ the transverse magnetization component M_x is not zero, then M_x will decay to zero, and similarly for M_y. The decay occurs

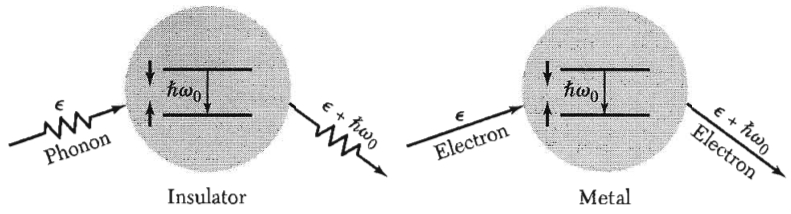


Figure 4a Some important processes that contribute to longitudinal magnetization relaxation in an insulator and in a metal. For the insulator we show a phonon scattered inelastically by the spin system. The spin system moves to a lower energy state, and the emitted phonon has higher energy by $\hbar\omega_0$ than the absorbed phonon. For the metal we show a similar inelastic scattering process in which a conduction electron is scattered.

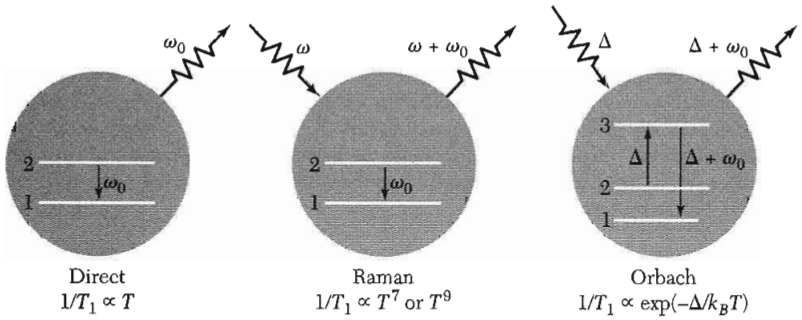


Figure 4b Spin relaxation from 2 → 1 by phonon emission, phonon scattering, and a two-stage phonon process. The temperature dependence of the longitudinal relaxation time T_1 is shown for the several processes.

because in thermal equilibrium the transverse components are zero. We can provide for transverse relaxation:

$$dM_x/dt = \gamma(\mathbf{M} \times \mathbf{B}_a)_x - M_x/T_2 \ ; \tag{13b}$$

$$dM_y/dt = \gamma(\mathbf{M} \times \mathbf{B}_a)_y - M_y/T_2 \ , \tag{13c}$$

where T_2 is called the **transverse relaxation time**.

The magnetic energy $-\mathbf{M} \cdot \mathbf{B}_a$ does not change as M_x or M_y changes, provided that $\mathbf{B}_a$ is along $\hat{\mathbf{z}}$. No energy need flow out of the spin system during relaxation of M_x or M_y, so that the conditions that determine T_2 may be less strict than for T_1. Sometimes the two times are nearly equal, and sometimes $T_1 \gg T_2$, depending on local conditions.

The time T_2 is a measure of the time during which the individual moments that contribute to M_x, M_y remain in phase with each other. Different local magnetic fields at the different spins will cause them to precess at different frequencies. If initially the spins have a common phase, the phases will become random in the course of time and the values of M_x, M_y will become zero. We can think of T_2 as a dephasing time.

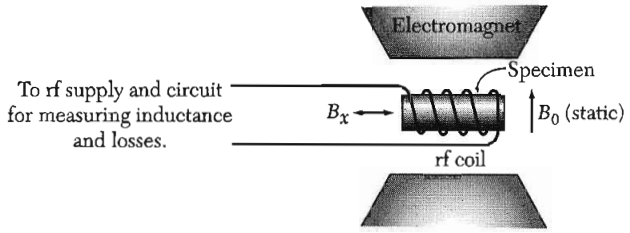


Figure 5 Schematic arrangement for magnetic resonance experiments.

The set of equations (13) are called the **Bloch equations**. They are not symmetrical in x, y, and z because we have biased the system with a static magnetic field along $\hat{\mathbf{z}}$. In experiments an rf magnetic field is usually applied along the $\hat{\mathbf{x}}$ or $\hat{\mathbf{y}}$ axis. Our main interest is in the behavior of the magnetization in the combined rf and static fields, as in Fig. 5. The Bloch equations are plausible, but not exact; they do not describe all spin phenomena, particularly not those in solids.

We determine the frequency of free precession of the spin system in a static field $\mathbf{B}_a = B_0\hat{\mathbf{z}}$ and with $M_z = M_0$. The Bloch equations reduce to

$$\frac{dM_x}{dt} = \gamma B_0 M_y - \frac{M_x}{T_2} \; ; \quad \frac{dM_y}{dt} = -\gamma B_0 M_x - \frac{M_y}{T_2} \; ; \quad \frac{dM_z}{dt} = 0 \; . \tag{14}$$

We look for damped oscillatory solutions of the form

$$M_x = m \exp(-t/T') \cos \omega t \; ; \quad M_y = -m \exp(-t/T') \sin \omega t \; . \tag{15}$$

On substitution in (14) we have for the left-hand equation

$$-\omega \sin \omega t - \frac{1}{T'} \cos \omega t = -\gamma B_0 \sin \omega t - \frac{1}{T_2} \cos \omega t \; , \tag{16}$$

so that the free precession is characterized by

$$\omega_0 = \gamma B_0 \; ; \quad T' = T_2 \; . \tag{17}$$

The motion (15) is similar to that of a damped harmonic oscillator in two dimensions. The analogy suggests correctly that the spin system will show resonance absorption of energy from a driving field near the frequency $\omega_0 = \gamma B_0$, and the frequency width of the response of the system to the driving field will be $\Delta\omega \approx 1/T_2$. Figure 6 shows the resonance of protons in water.

The Bloch equations may be solved to give the power absorption from a rotating magnetic field of amplitude B_1:

$$B_x = B_1 \cos \omega t \; ; \quad B_y = -B_1 \sin \omega t \; . \tag{18}$$

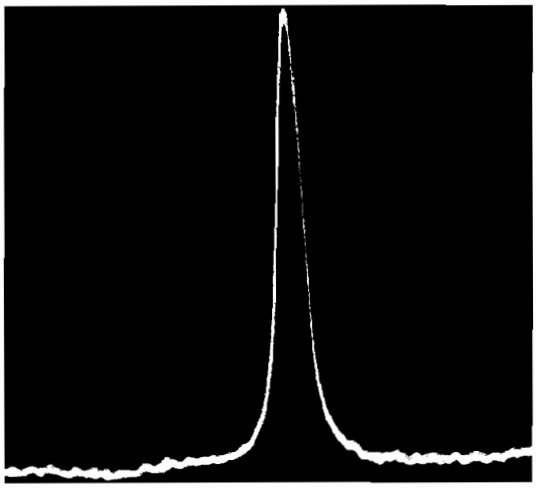


Figure 6 Proton resonance absorption in water. (E. L. Hahn.)

After a routine calculation one finds that the power absorption is

$$\text{(CGS)} \qquad \mathcal{P}(\omega) = \frac{\omega\gamma M_z T_2}{1 + (\omega_0 - \omega)^2 T_2^2} B_1^2 \ . \tag{19}$$

The half-width of the resonance at half-maximum power is

$$(\Delta\omega)_{1/2} = 1/T_2 \ . \tag{20}$$

LINE WIDTH

The magnetic dipolar interaction is usually the most important cause of line broadening in a rigid lattice of magnetic dipoles. The magnetic field $\Delta\mathbf{B}$ seen by a magnetic dipole $\boldsymbol{\mu}_1$ due to a magnetic dipole $\boldsymbol{\mu}_2$ at a point $\mathbf{r}_{12}$ from the first dipole is

$$\text{(CGS)} \qquad \Delta\mathbf{B} = \frac{3(\boldsymbol{\mu}_2 \cdot \mathbf{r}_{12})\mathbf{r}_{12} - \boldsymbol{\mu}_2 r_{12}^2}{r_{12}^5} \ , \tag{21}$$

by a fundamental result of magnetostatics.

The order of magnitude of the interaction is, with B_i written for ΔB,

$$\text{(CGS)} \qquad B_i \approx \mu/r^3 \ . \tag{22}$$

The strong dependence on r suggests that close-neighbor interactions will be dominant, so that

$$\text{(CGS)} \qquad B_i \approx \mu/a^3 \ , \tag{23}$$

where a is the separation of nearest neighbors. This result gives us a measure of the width of the spin resonance line, assuming random orientation of the neighbors. For protons at 2Å separation,

$$B_i \approx \frac{1.4 \times 10^{-23}\ \text{G cm}^3}{8 \times 10^{-24}\ \text{cm}^3} \approx 2\ \text{gauss} = 2 \times 10^{-4}\ \text{tesla} \ . \tag{24}$$

To express (21), (22), and (23) in SI, multiply the right-hand sides by $\mu_0/4\pi$.

Motional Narrowing

The line width decreases for nuclei in rapid relative motion. The effect in solids is illustrated by Fig. 7: diffusion resembles a random walk as atoms jump from one crystal site to another. An atom remains in one site for an average time τ that decreases markedly as the temperature increases.

The motional effects on the line width are even more spectacular in normal liquids, because the molecules are highly mobile. The width of the proton resonance line in water is only 10^{-5} of the width expected for water molecules frozen in position.

The effect of nuclear motion on T_2 and on the line width is subtle, but can be understood by an elementary argument. We know from the Bloch equations that T_2 is a measure of the time in which an individual spin becomes dephased by one radian because of a local perturbation in the magnetic field

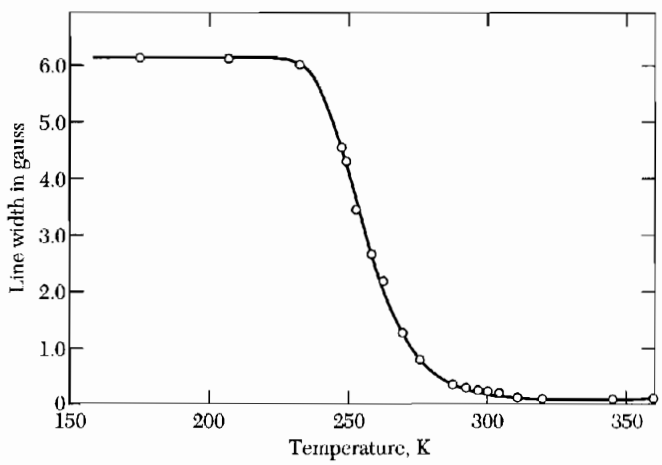


Figure 7 Effect of diffusion of nuclei on the Li^7 NMR line width in metallic lithium. At low temperatures the width agrees with the theoretical value for a rigid lattice. As the temperature increases, the diffusion rate increases and the line width decreases. The abrupt decrease in line width above T = 230 K occurs when the diffusion hopping time τ becomes shorter than $1/\gamma B_i$. Thus the experiment gives a direct measure of the hopping time for an atom to change lattice sites. (After H. S. Gutowsky and B. R. McGarvey.)

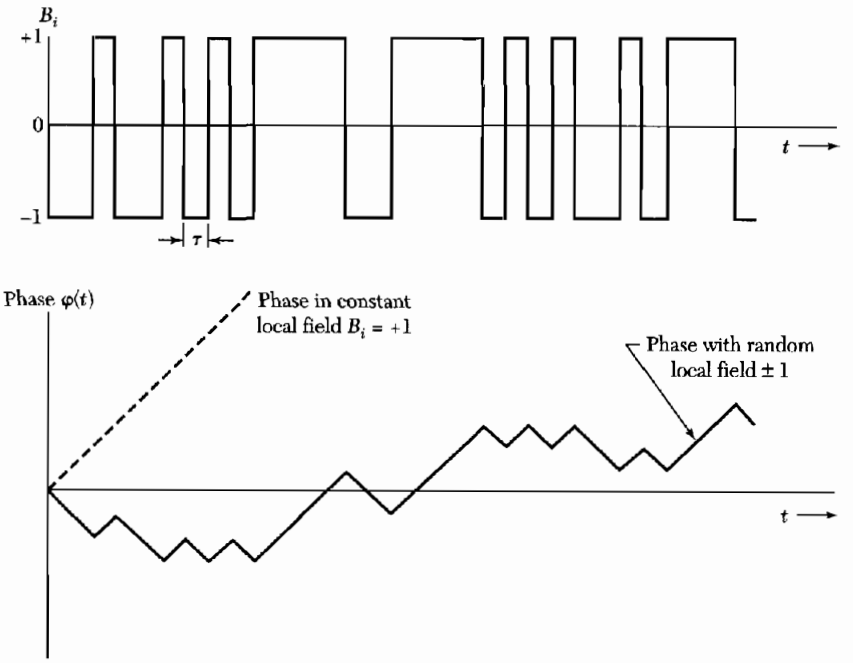


Figure 8 Phase of a spin in a constant local field, as compared with dephasing of a spin which after fixed time intervals τ hops at random among sites having local fields ± 1.

intensity. Let $(\Delta\omega)_0 \approx \gamma B_i$ denote the local frequency deviation due to a perturbation B_i. The local field may be caused by dipolar interactions with other spins.

If the atoms are in rapid relative motion, the local field B_i seen by a given spin will fluctuate rapidly in time. We suppose that the local field has a value $+B_i$ for an average time τ and then changes to $-B_i$, as in Fig. 8a. Such a random change could be caused by a change of the angle between $\boldsymbol{\mu}$ and $\mathbf{r}$ in (21). In the time τ the spin will precess by an extra phase angle $\delta\varphi = \pm\gamma B_i\tau$ relative to the phase angle of the steady precession in the applied field B_0.

The motional narrowing effect arises for short τ such that $\delta\varphi \ll 1$. After n intervals of duration τ the mean square dephasing angle in the field B_0 will be

$$\langle\varphi^2\rangle = n(\delta\varphi)^2 = n\gamma^2 B_i^2\tau^2 \ , \tag{25}$$

by analogy with a random walk process: the mean square displacement from the initial position after n steps of length ℓ in random directions is $\langle r^2\rangle = n\ell^2$.

The average number of steps necessary to dephase a spin by one radian is $n = 1/\gamma^2 B_i^2\tau^2$. (Spins dephased by much more than one radian do not contribute to the absorption signal.) This number of steps takes place in a time

$$T_2 = n\tau = 1/\gamma^2 B_i^2\tau \ , \tag{26}$$

quite different from the rigid lattice result $T_2 \cong 1/\gamma B_i$. From (26) we obtain as the line width for rapid motion with a characteristic time τ:

$$\Delta\omega = 1/T_2 = (\gamma B_i)^2 \tau \ , \tag{27}$$

or

$$\Delta\omega = 1/T_2 = (\Delta\omega)_0^2 \tau \ , \tag{28}$$

where $(\Delta\omega)_0$ is the line width in the rigid lattice.

The argument assumes that $(\Delta\omega)_0 \tau \ll 1$, as otherwise $\delta\varphi$ will not be $\ll 1$. Thus $\Delta\omega \ll (\Delta\omega)_0$. The shorter is τ, the narrower is the resonance line! This remarkable effect is known as **motional narrowing**.[2] The rotational relaxation time of water molecules at room temperature is known from dielectric constant measurements to be of the order of 10^{-10} s; if $(\Delta\omega)_0 \approx 10^5\ \text{s}^{-1}$, then $(\Delta\omega)_0 \tau \approx 10^{-5}$ and $\Delta\omega \approx (\Delta\omega)_0^2 \tau \approx 1\ \text{s}^{-1}$. Thus the motion narrows the proton resonance line to about 10^{-5} of the static width.

HYPERFINE SPLITTING

The hyperfine interaction is the magnetic interaction between the magnetic moment of a nucleus and the magnetic moment of an electron. To an observer stationed on the nucleus, the interaction is caused by the magnetic field produced by the magnetic moment of the electron and by the motion of the electron about the nucleus. There is an electron current about the nucleus if the electron is in a state with orbital angular momentum about the nucleus. But even if the electron is in a state of zero orbital angular momentum, there is an electron spin current about the nucleus, and this current gives rise to the **contact hyperfine interaction**, of particular importance in solids. We can understand the origin of the contact interaction by a qualitative physical argument, given in CGS.

The results of the Dirac theory of the electron suggest that the magnetic moment of $\mu_B = e\hbar/2mc$ of the electron arises from the circulation of an electron with velocity c in a current loop of radius approximately the electron Compton wavelength, $\lambda\!\!\!^{-}_e = \hbar/mc \sim 10^{-11}$ cm. The electric current associated with the circulation is

$$I \sim e \times (\text{turns per unit time}) \sim ec/\lambda\!\!\!^{-}_e \ , \tag{29}$$

[2]The physical ideas are due to N. Bloembergen, E. M. Purcell, and R. V. Pound, Phys. Rev. **73**, 679 (1948). The result differs from the theory of optical line width caused by *strong* collisions between atoms (as in a gas discharge), where a short τ gives a broad line. In the nuclear spin problem the collisions are weak. In most optical problems the collisions of atoms are strong enough to interrupt the phase of the oscillation. In nuclear resonance the phase may vary smoothly in a collision, although the frequency may vary suddenly from one value to another nearby value.

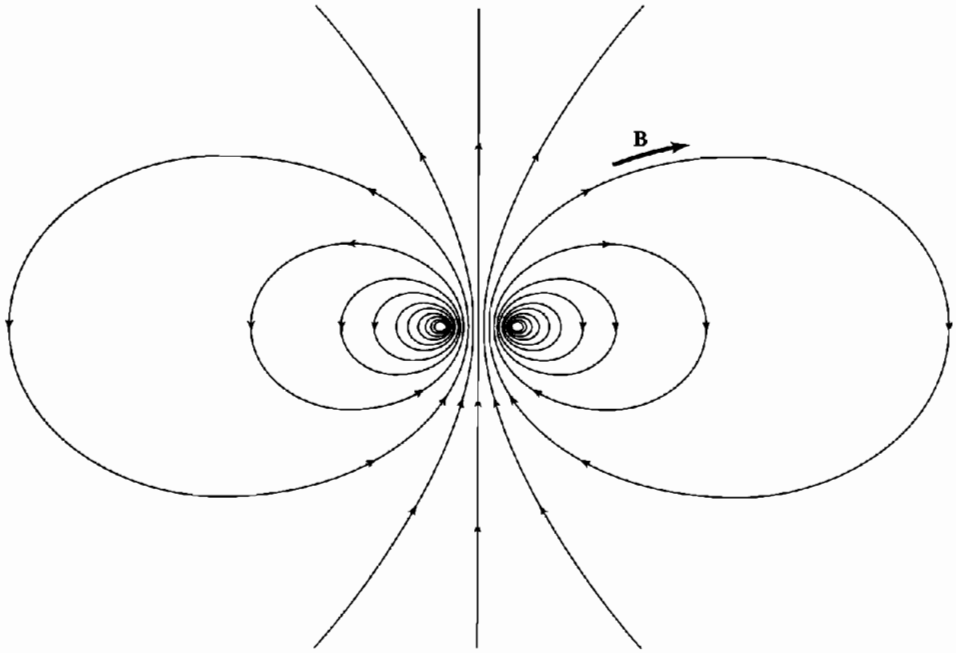


Figure 9 Magnetic field **B** produced by a charge moving in a circular loop. The contact part of the hyperfine interaction with a nuclear magnetic moment arises from the region within or near to the current loop. The field averaged over a spherical shell that encloses the loop is zero. Thus for an s electron ($L = 0$) only the contact part contributes to the interaction.

and the magnetic field (Fig. 9) produced by the current is

(CGS) $$B \sim I/\lambda\!\!\!^{-}_e c \sim e/\lambda\!\!\!^{-2}_e \ . \tag{30}$$

The observer on the nucleus has the probability

$$P \approx |\psi(0)|^2 \lambda\!\!\!^{-3}_e \tag{31}$$

of finding himself inside the electron, that is, within a sphere of volume $\lambda\!\!\!^{-3}_e$ about the electron. Here $\psi(0)$ is the value of the electron wavefunction at the nucleus. Thus the average value of the magnetic field seen by the nucleus is

$$\bar{B} \approx e|\psi(0)|^2 \lambda\!\!\!^{-}_e \approx \mu_B |\psi(0)|^2 \ , \tag{32}$$

where $\mu_B = e\hbar/2mc = \frac{1}{2} e\lambda\!\!\!^{-}_e$ is the Bohr magneton.

The contact part of the hyperfine interaction energy is

$$U = -\boldsymbol{\mu}_I \cdot \bar{\mathbf{B}} \approx -\boldsymbol{\mu}_I \cdot \boldsymbol{\mu}_B |\psi(0)|^2 \approx \gamma \hbar \mu_B |\psi(0)|^2 \mathbf{I} \cdot \mathbf{S} \ , \tag{33}$$

where I is the nuclear spin in units of $\hbar$. The contact interaction in an atom has the form

$$U = a\mathbf{I} \cdot \mathbf{S} \ . \tag{34}$$

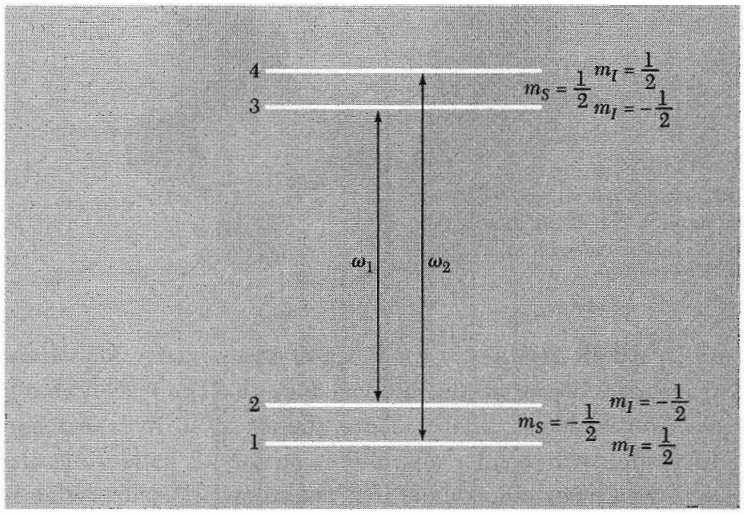


Figure 10 Energy levels in a magnetic field of a system with $S = \frac{1}{2}$, $I = \frac{1}{2}$. The diagram is drawn for the strong field approximation $\mu_B B \gg a$, where a is the hyperfine coupling constant, taken to be positive. The four levels are labeled by the magnetic quantum numbers m_S, m_I. The strong electronic transitions have $\Delta m_I = 0$, $\Delta m_S = \pm 1$.

Values of the hyperfine constant a for the ground states of several free atoms are:

nucleus	H^1	Li^7	Na^{23}	K^{39}	K^{41}
I	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$	$\frac{3}{2}$
a in gauss	507	144	310	83	85
a in MHz	1420	402	886	231	127

In a strong magnetic field the energy level scheme of a free atom or ion is dominated by the Zeeman energy splitting of the electron levels; the hyperfine interaction gives an additional splitting that in strong fields is $U' \cong a m_S m_I$, where m_S, m_I are the magnetic quantum numbers.

For the energy level diagram of Fig. 10 the two electronic transitions have the selection rules $\Delta m_S = \pm 1$, $\Delta m_I = 0$, the frequencies are $\omega = \gamma H_0 \pm a/2\hbar$. The nuclear transitions are not marked; they have $\Delta m_S = 0$, so that $\omega_{\text{nuc}} = a/2\hbar$. The frequency of the nuclear transition $1 \rightarrow 2$ is equal to that of $3 \rightarrow 4$.

The hyperfine interaction in a magnetic atom may split the ground energy level. The splitting in hydrogen is 1420 MHz; this is the radio frequency line of interstellar atomic hydrogen.

Examples: Paramagnetic Point Defects

The hyperfine splitting of the electron spin resonance furnishes valuable structural information about paramagnetic point defects, such as the F centers in alkali halide crystals and the donor impurity atoms in semiconductor crystals.

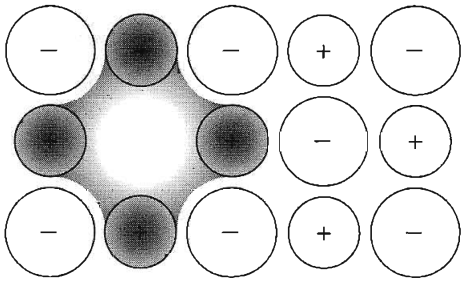


Figure 11 An F center is a negative ion vacancy with one excess electron bound at the vacancy. The distribution of the excess electron is largely on the positive metal ions adjacent to the vacant lattice site.

***F* Centers in Alkali Halides.** An F center is a negative ion vacancy with one excess electron bound at the vacancy (Fig. 11). The wavefunction of the trapped electron is shared chiefly among the six alkali ions adjacent to the vacant lattice site, with smaller amplitudes on the 12 halide ions that form the shell of second nearest neighbors. The counting applies to crystals with the NaCl structure. If $\varphi(\mathbf{r})$ is the wavefunction of the valence electron on a single alkali ion, then in the first (or LCAO) approximation

$$\psi(\mathbf{r}) = C \sum_p \varphi(\mathbf{r} - \mathbf{r}_p) \ , \tag{35}$$

where in the NaCl structure the six values of $\mathbf{r}_p$ mark the alkali ion sites that bound the lattice vacancy.

The width of the electron spin resonance line of an F center is determined essentially by the hyperfine interaction of the trapped electron with the nuclear magnetic moments of the alkali ions adjacent to the vacant lattice site. The observed line width is evidence for the simple picture of the wavefunction of the electron. By line width we mean the width of the envelope of the possible hyperfine structure components.

As an example, consider an F center in KCl. Natural potassium is 93 percent K^{39} with nuclear spin $I = \frac{3}{2}$. The total spin of the six potassium nuclei at the F center is $I_{max} = 6 \times \frac{3}{2} = 9$, so that the number of hyperfine components is $2I_{max} + 1 = 19$; this is the number of possible values of the quantum number m_I. There are $(2I + 1)^6 = 4^6 = 4096$ independent arrangements of the six spins distributed into the 19 components, as in Fig. 12. Often we observe only the envelope of the absorption line of an F center.

Donor Atoms in Silicon. Phosphorus is a donor when present in silicon. Each donor atom has five outer electrons, of which four enter diamagnetically into the covalent bond network of the crystal, and the fifth acts as a paramagnetic center of spin $S = \frac{1}{2}$. The experimental hyperfine splitting in the strong field limit is shown in Fig. 13.

When the concentration exceeds about 1×10^{18} donors cm^{-3}, the split line is replaced by a single narrow line. This is a motional narrowing effect (Eq. 28) of the rapid hopping of the donor electrons among many donor atoms. The

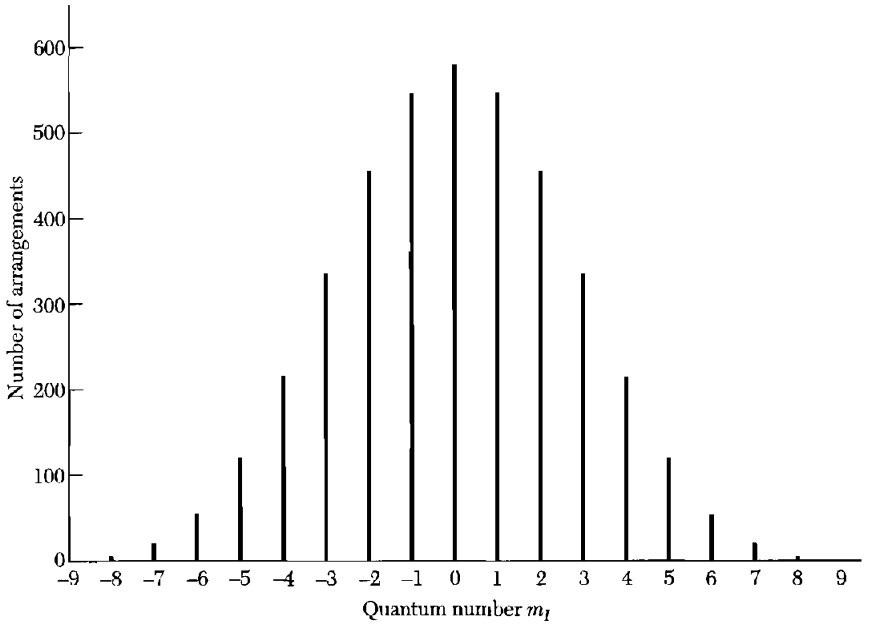


Figure 12 The 4096 arrangements of the six nuclear spins of K^{39} as distributed into 19 hyperfine components. Each component will be split further into a very large number of components by virtue of the residual hyperfine interaction with the 12 neighboring Cl nuclei, which may be Cl^{35} (75 percent) or Cl^{37} (25 percent). The envelope of the pattern is approximately gaussian in form.

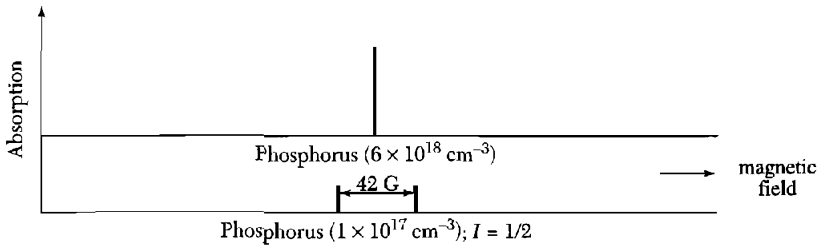


Figure 13 Electron spin resonance lines of P donor atoms in silicon. At the higher donor concentration near the metal-insulator transition, a donor electron can hop from site to site so rapidly that the hyperfine structure is suppressed. (After R. C. Fletcher, W. A. Yager, G. L. Pearson, and F. R. Merritt.)

rapid hopping averages out the hyperfine splitting. The hopping rate increases at the higher concentrations as the overlap of the donor electron wavefunctions is increased, a view supported by conductivity measurements (Chapter 14).

Knight Shift

At a fixed frequency the resonance of a nuclear spin is observed at a slightly different magnetic field in a metal than in a diamagnetic solid. The effect is known as the **Knight shift** or **metallic shift** and is valuable as a tool for the study of conduction electrons.

The interaction energy of a nucleus of spin **I** and magnetogyric ratio γ_I is

$$U = (-\gamma_I \hbar B_0 + a\langle S_z \rangle) I_z \ , \tag{36}$$

where the first term is the interaction with the applied magnetic field B_0 and the second is the average hyperfine interaction of the nucleus with the conduction electrons. The average conduction electron spin $\langle S_z \rangle$ is related to the Pauli spin susceptibility χ_s of the conduction electrons: $M_z = gN\mu_B \langle S_z \rangle = \chi_s B_0$, whence the interaction may be written as

$$U = \left(-\gamma_I \hbar + \frac{a\chi_s}{gN\mu_B} \right) B_0 I_z = -\gamma_I \hbar B_0 \left(1 + \frac{\Delta B}{B_0} \right) I_z \ . \tag{37}$$

The Knight shift is defined as

$$K = -\frac{\Delta B}{B_0} = \frac{a\chi_s}{gN\mu_B \gamma_I \hbar} \tag{38}$$

and simulates a fractional change in the magnetogyric ratio. By the definition (34) of the hyperfine contact energy, the Knight shift is given approximately by $K \approx \chi_s |\psi(0)|^2 / N$; that is, by the Pauli spin susceptibility increased in the ratio of the conduction electron concentration at the nucleus to the average conduction electron concentration.

Experimental values are given in Table 2. The value of the hyperfine coupling constant a is somewhat different in the metal than in the free atom because the wave functions at the nucleus are different. From the Knight shift of metallic Li it is deduced that the value of $|\psi(0)|^2$ in the metal is 0.44 of the value in the free atom; a calculated value of the ratio using theoretical wave functions is 0.49.

Table 2 Knight shifts in NMR in metallic elements

(At room temperature)

Nucleus	Knight shift in percent	Nucleus	Knight shift in percent
Li^7	0.0261	Cu^{63}	0.237
Na^{23}	0.112	Rb^{87}	0.653
Al^{27}	0.162	Pd^{105}	−3.0
K^{39}	0.265	Pt^{195}	−3.533
V^{51}	0.580	Au^{197}	1.4
Cr^{53}	0.69	Pb^{207}	1.47

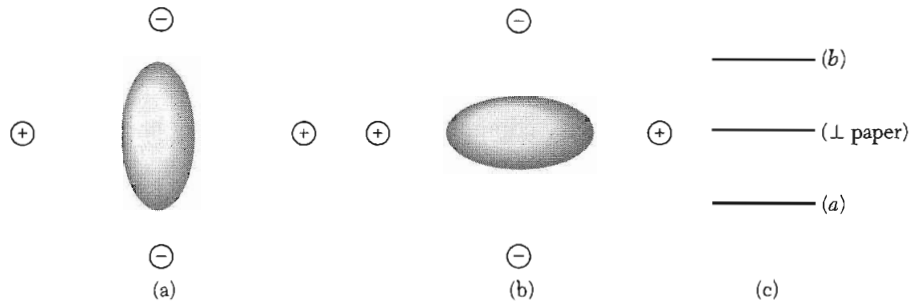


Figure 14 (a) Lowest-energy orientation of a nuclear electric quadrupole moment ($Q > 0$) in the local electric field of the four ions shown. The electrons of the ion itself are not shown. (b) Highest energy orientation. (c) The energy level splitting for $I = 1$.

NUCLEAR QUADRUPOLE RESONANCE

Nuclei of spin $I \geq 1$ have an electric quadrupole moment. The quadrupole moment Q is a measure of the ellipticity of the distribution of charge in the nucleus. The quantity of interest is defined classically by

$$eQ = \tfrac{1}{2}\int (3z^2 - r^2)\rho(\mathbf{r})d^3x \ , \tag{39}$$

where $\rho(\mathbf{r})$ is the charge density. An egg-shaped nucleus has Q positive; a saucer-shaped nucleus has Q negative. The nucleus when placed in a crystal will see the electrostatic field of its environment, as in Fig. 14. If the symmetry of this field is lower than cubic, then the nuclear quadrupole moment will lead to a set of energy levels split by the interaction of the quadrupole moment with the local electric field.

The states that are split are the $2I + 1$ states of a spin I. The quadrupole splittings can often be observed directly because an rf magnetic field of the appropriate frequency can cause transitions between the levels. The term **nuclear quadrupole resonance** refers to observations of nuclear quadrupole splittings in the absence of a static magnetic field. The quadrupole splittings are particularly large in covalently bonded molecules such as Cl_2, Br_2, and I_2; the splittings are of the order 10^7 or 10^8 Hz.

FERROMAGNETIC RESONANCE

Spin resonance at microwave frequencies in ferromagnets is similar in principle to nuclear spin resonance. The total electron magnetic moment of the specimen precesses about the direction of the static magnetic field, and energy is absorbed strongly from the rf transverse field when its frequency is equal to the precessional frequency. We may think of the macroscopic vector **S** representing the total spin of the ferromagnet as quantized in the static magnetic field, with energy levels separated by the usual Zeeman frequencies; the

magnetic selection rule $\Delta m_S = \pm 1$ allows transitions only between adjacent levels.

The unusual features of ferromagnetic resonance include:

- The transverse susceptibility components χ' and χ'' are very large because the magnetization of a ferromagnet in a given static field is very much larger than the magnetization of electronic or nuclear paramagnets in the same field.
- The shape of the specimen plays an important role. Because the magnetization is large, the demagnetization field is large.
- The strong exchange coupling between the ferromagnetic electrons tends to suppress the dipolar contribution to the line width, so that the ferromagnetic resonance lines can be quite sharp (<1 G) under favorable conditions.
- Saturation effects occur at low rf power levels. It is not possible, as it is with nuclear spin systems, to drive a ferromagnetic spin system so hard that the magnetization M_z is reduced to zero or reversed. The ferromagnetic resonance excitation breaks down into spin wave modes before the magnetization vector can be rotated appreciably from its initial direction.

Shape Effects in FMR

We treat the effects of specimen shape on the resonance frequency. Consider a specimen of a cubic ferromagnetic insulator in the form of an ellipsoid with principal axes parallel to x, y, z axes of a cartesian coordinate system. The **demagnetization factors** N_x, N_y, N_z are identical with the depolarization factors to be defined in Chapter 16. The components of the internal magnetic field $\mathbf{B}_i$ in the ellipsoid are related to the applied field by

$$B_x^i = B_x^0 - N_x M_x \; ; \quad B_y^i = B_y^0 - N_y M_y \; ; \quad B_z^i = B_z^0 - N_z M_z \; .$$

The Lorentz field $(4\pi/3)\mathbf{M}$ and the exchange field $\lambda\mathbf{M}$ do not contribute to the torque because their vector product with $\mathbf{M}$ vanishes identically. In SI we replace the components of $\mathbf{M}$ by $\mu_0\mathbf{M}$, with the appropriate redefinition of the N's.

The components of the spin equation of motion $\dot{\mathbf{M}} = \gamma(\mathbf{M} \times \mathbf{B}^i)$ become, for an applied static field $B_0\hat{\mathbf{z}}$,

$$\frac{dM_x}{dt} = \gamma(M_y B_z^i - M_z B_y^i) = \gamma[B_0 + (N_y - N_z)M]M_y \; ; \tag{40}$$

$$\frac{dM_y}{dt} = \gamma[M(-N_x M_x) - M_x(B_0 - N_z M)] = -\gamma[B_0 + (N_x - N_z)M]M_x \; .$$

To first order we may set $dM_z/dt = 0$ and $M_z = M$. Solutions of (40) with time dependence $\exp(-i\omega t)$ exist if

$$\begin{vmatrix} i\omega & \gamma[B_0 + (N_y - N_z)M] \\ -\gamma[B_0 + (N_x - N_z)M] & i\omega \end{vmatrix} = 0 \; ,$$

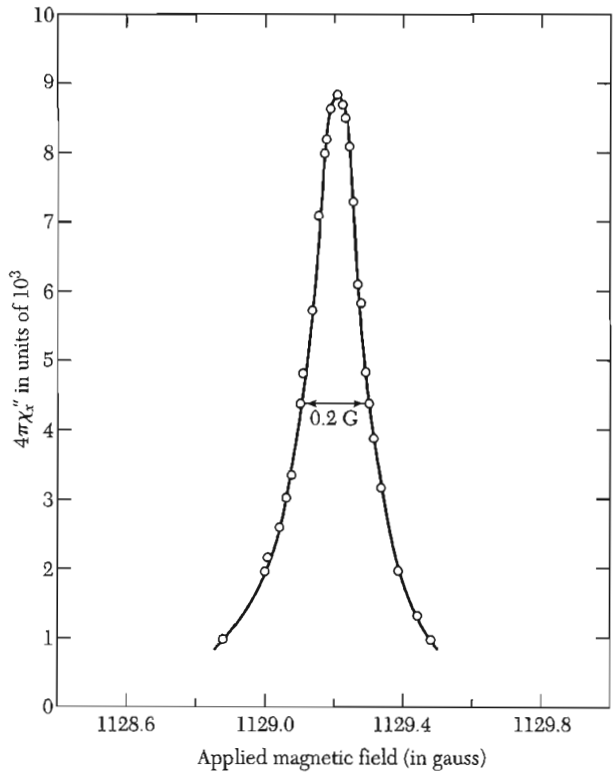


Figure 15 FMR in a polished sphere of the ferromagnet yttrium iron garnet at 3.33 GHz and 300 K for $B_0 \parallel [111]$. The total line width at half-power is only 0.2 G. (After R. C. LeCraw and E. Spencer.)

so that the ferromagnetic resonance frequency in the applied field B_0 is

$$\text{(CGS)} \qquad \omega_0^2 = \gamma^2[B_0 + (N_y - N_z)M][B_0 + (N_x - N_z)M] \ ; \tag{41}$$

$$\text{(SI)} \qquad \omega_0^2 = \gamma^2[B_0 + (N_y - N_z)\mu_0 M][B_0 + (N_x - N_z)\mu_0 M] \ .$$

The frequency ω_0 is called the frequency of the **uniform mode**, in distinction to the frequencies of magnon and other nonuniform modes. In the uniform mode all the moments precess together in phase with the same amplitude.

For a sphere $N_x = N_y = N_z$, so that $\omega_0 = \gamma B_0$. A very sharp resonance line in this geometry is shown in Fig. 15. For a flat plate with B_0 perpendicular to the plate $N_x = N_y = 0$; $N_z = 4\pi$, whence the ferromagnetic resonance frequency is

$$\text{(CGS)} \quad \omega_0 = \gamma(B_0 - 4\pi M) \ ; \qquad \text{(SI)} \quad \omega_0 = \gamma(B_0 - \mu_0 M) \ . \tag{42}$$

If B_0 is parallel to the plane of the plate, the xz plane, then $N_x = N_z = 0$; $N_y = 4\pi$, and

$$\text{(CGS)} \quad \omega_0 = \gamma[B_0(B_0 + 4\pi M)]^{1/2} \ ; \qquad \text{(SI)} \quad \omega_0 = \gamma[B_0(B_0 + \mu_0 M)]^{1/2} \ . \tag{43}$$

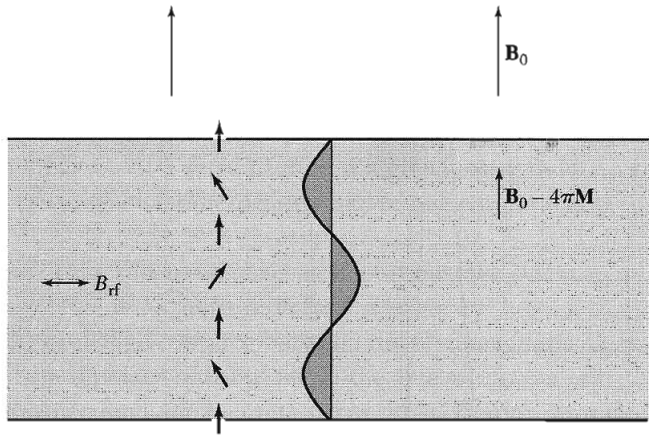


Figure 16 Spin wave resonance in a thin film. The plane of the film is normal to the applied magnetic field B_0. A cross section of the film is shown here. The internal magnetic field is $B_0 - 4\pi M$. The spins on the surfaces of the film are assumed to be held fixed in direction by surface anisotropy forces. A uniform rf field will excite spin wave modes having an odd number of half-wavelengths. The wave shown is for $n = 3$ half-wavelengths.

The experiments determine γ, which is related to the spectroscopic splitting factor g by $-\gamma = g\mu_B/\hbar$. Values of g for metallic Fe, Co, Ni at room temperature are 2.10, 2.18, and 2.21, respectively.

Spin Wave Resonance

Uniform rf magnetic fields can excite long-wavelength spin waves in thin ferromagnetic films if the electron spins on the surfaces of the film see different anisotropy fields than the spins within the films. In effect, the surface spins may be pinned by surface anisotropy interactions, as shown in Fig. 16. If the rf field is uniform, it can excite waves with an odd number of half-wavelengths within the thickness of the film. Waves with an even number of half-wavelengths have no net interaction energy with the field.

The condition for **spin wave resonance** (SWR) with the applied magnetic field normal to the film is obtained from (42) by adding to the right-hand side the exchange contribution to the frequency. The exchange contribution may be written as Dk^2, where D is the spin wave exchange constant. The assumption $ka \ll 1$ is valid for the SWR experiments. Thus in an applied field B_0 the spin wave resonance frequencies are:

$$\text{(CGS)} \qquad \omega_0 = \gamma(B_0 - 4\pi M) + Dk^2 = \gamma(B_0 - 4\pi M) + D(n\pi/L)^2 \ , \qquad (44)$$

where the wavevector for a mode of n half-wavelengths in a film of thickness L is $k = n\pi/L$. An experimental spectrum is shown in Fig. 17.

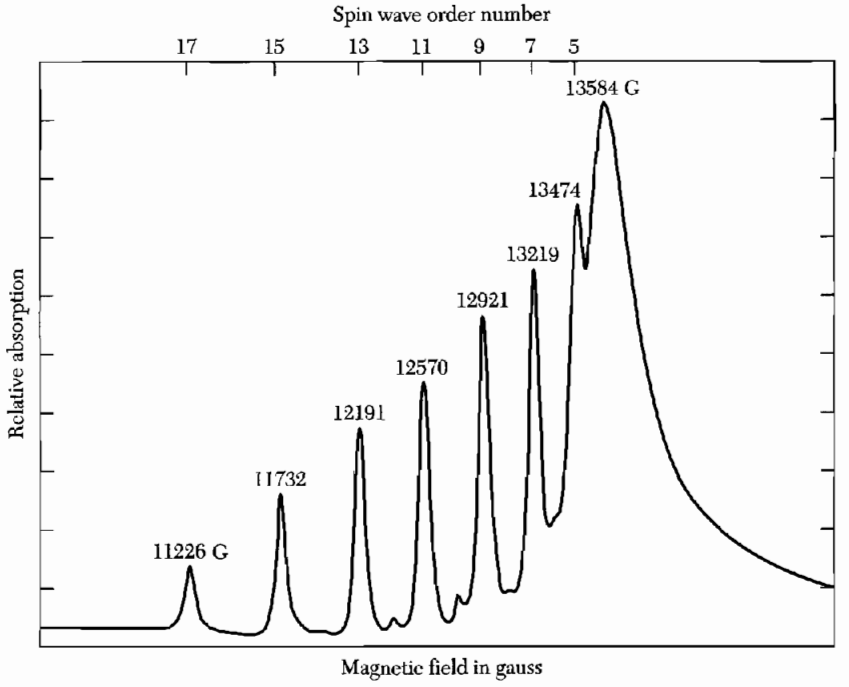


Figure 17 Spin wave resonance spectrum in a Permalloy (80Ni20Fe) film at 9 GHz. The order number is the number of half-wavelengths in the thickness of the film. (After R. Weber.)

ANTIFERROMAGNETIC RESONANCE

We consider a uniaxial antiferromagnet with spins on two sublattices, 1 and 2. We suppose that the magnetization $\mathbf{M}_1$ on sublattice 1 is directed along the $+z$ direction by an anisotropy field $B_A\hat{\mathbf{z}}$; the anisotropy field (Chapter 12) results from an anisotropy energy density $U_K(\theta_1) = K \sin^2 \theta_1$. Here θ_1 is the angle between $\mathbf{M}_1$ and the z axis, whence $B_A = 2K/M$, with $M = |\mathbf{M}_1| = |\mathbf{M}_2|$. The magnetization $\mathbf{M}_2$ is directed along the $-z$ direction by an anisotropy field $-B_A\hat{\mathbf{z}}$. If $+z$ is an easy direction of magnetization, so is $-z$. If one sublattice is directed along $+z$, the other will be directed along $-z$.

The exchange interaction between $\mathbf{M}_1$ and $\mathbf{M}_2$ is treated in the mean field approximation. The exchange fields are

$$\mathbf{B}_1(\text{ex}) = -\lambda\mathbf{M}_2\ ; \quad \mathbf{B}_2(\text{ex}) = -\lambda\mathbf{M}_1\ , \tag{45}$$

where λ is positive. Here $\mathbf{B}_1$ is the field that acts on the spins of sublattice 1, and $\mathbf{B}_2$ acts on sublattice 2. In the absence of an external magnetic field the total field acting on $\mathbf{M}_1$ is $\mathbf{B}_1 = -\lambda\mathbf{M}_2 + B_A\hat{\mathbf{z}}$; the total field on $\mathbf{M}_2$ is $\mathbf{B}_2 = -\lambda\mathbf{M}_1 - \mathrm{B}_A\hat{\mathbf{z}}$, as in Fig. 18.

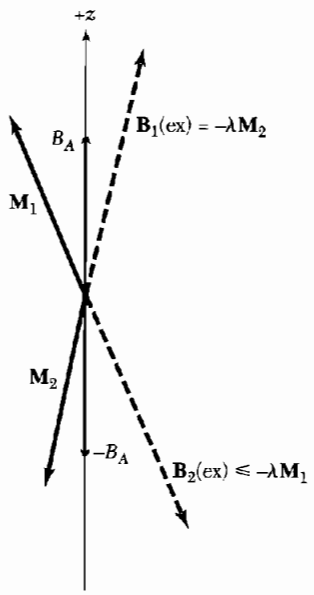


Figure 18 Effective fields in antiferromagnetic resonance. The magnetization $\mathbf{M}_1$ of sublattice 1 sees a field $-\lambda\mathbf{M}_2 + B_A\hat{\mathbf{z}}$; the magnetization $\mathbf{M}_2$ sees $-\lambda\mathbf{M}_1 - B_A\hat{\mathbf{z}}$. Both ends of the crystal axis are "easy axes" of magnetization.

In what follows we set $M_1^z = M$; $M_2^z = -M$. The linearized equations of motion are

$$dM_1^x/dt = \gamma[M_1^y(\lambda M + B_A) - M(-\lambda M_2^y)] \ ;$$

$$dM_1^y/dt = \gamma[M(-\lambda M_2^x) - M_1^x(\lambda M + B_A)] \ ; \tag{46}$$

$$dM_2^x/dt = \gamma[M_2^y(-\lambda M - B_A) - (-M)(-\lambda M_1^y)] \ ;$$

$$dM_2^y/dt = \gamma[(-M)(-\lambda M_1^x) - M_2^x(-\lambda M - B_A)] \ . \tag{47}$$

We define $M_1^+ = M_1^x + iM_1^y$; $M_2^+ = M_2^x + iM_2^y$. Then (46) and (47) become, for time dependence $\exp(-i\omega t)$,

$$-i\omega M_1^+ = -i\gamma[M_1^+(B_A + \lambda M) + M_2^+(\lambda M)] \ ;$$

$$-i\omega M_2^+ = i\gamma[M_2^+(B_A + \lambda M) + M_1^+(\lambda M)] \ .$$

These equations have a solution if, with the exchange field $B_E \equiv \lambda M$,

$$\begin{vmatrix} \gamma(B_A + B_E) - \omega & \gamma B_E \\ \gamma B_E & \gamma(B_A + B_E) + \omega \end{vmatrix} = 0 \ .$$

Thus the antiferromagnetic resonance frequency is given by

$$\omega_0^2 = \gamma^2 B_A(B_A + 2B_E) \ . \tag{48}$$

MnF_2 is an extensively studied antiferromagnet. The structure is shown in Fig. 19. The observed variation of ω_0 with temperature is shown in Fig. 20. Careful estimates were made by Keffer of B_A and B_E for MnF_2. He estimated

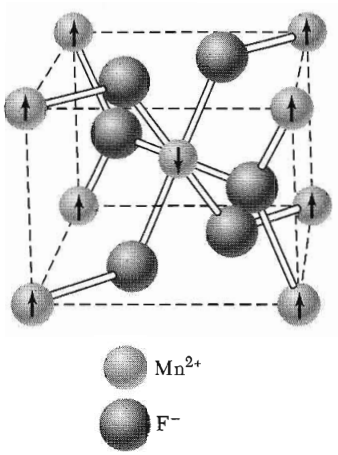


Figure 19 Chemical and magnetic structure of MnF_2. The arrows indicate the direction and arrangement of the magnetic moments assigned to the manganese atoms.

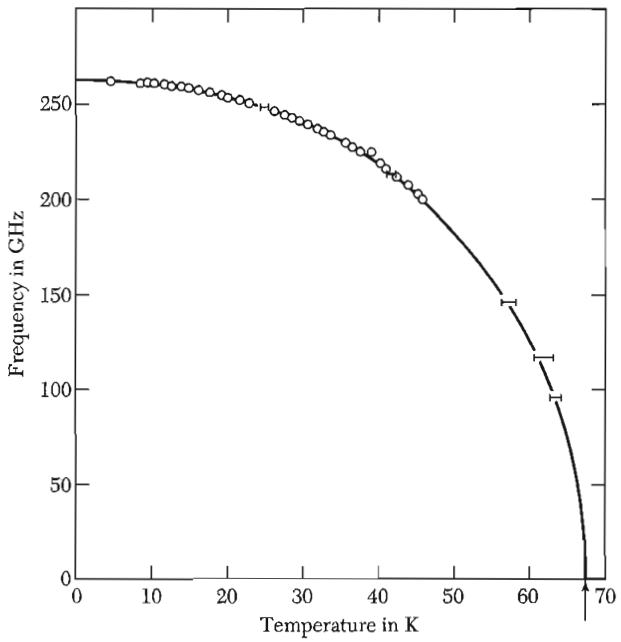


Figure 20 Antiferromagnetic resonance frequency for MnF_2 versus temperature. (After Johnson and Nethercot.)

$B_E = 540$ kG and $B_A = 8.8$ kG at 0 K, whence $(2B_AB_E)^{1/2} = 100$ kG. The observed value is 93 kG.

Richards has made a compilation of AFMR frequencies as extrapolated to 0 K:

Crystal	CoF_2	NiF_2	MnF_2	FeF_2	MnO	NiO
Frequency in 10^{10} Hz	85.5	93.3	26.0	158.	82.8	109

Exchange Narrowing

We consider a paramagnet with an exchange interaction J among nearest-neighbor electron spins. The temperature is assumed to be well above any spin-ordering temperature T_c. Under these conditions the width of the spin resonance line is usually much narrower than expected for the dipole-dipole interaction. The effect is called **exchange narrowing**; there is a close analogy with motional narrowing. We interpret the exchange frequency $\omega_{ex} \approx J/\hbar$ as a hopping frequency $1/\tau$. Then by generalization of the motional-narrowing result (28) we have for the width of the exchange-narrowed line:

$$\Delta\omega \approx (\Delta\omega)_0^2/\omega_{ex} \ , \tag{49}$$

where $(\Delta\omega)^2_0 = \gamma^2\langle B_i^2\rangle$ is the square of the static dipolar width in the absence of exchange.

A useful and striking example of exchange narrowing is the paramagnetic organic crystal known as the g marker or DPPH, diphenyl picryl hydrazyl, often used for magnetic field calibration. This free radical has a 1.35 G half-width of the resonance line at half-power, only a few percent of the pure dipole width.

Zero-Field Splitting

A number of paramagnetic ions have crystal field splittings of their magnetic ground state energy levels in the range of $10^{10} - 10^{11}$ Hz, conveniently accessible by microwave techniques. The Mn^{2+} ion has been studied in many crystals as an additive impurity. A ground state splitting in the range $10^7 - 10^9$ Hz is observed, according to the environment.

PRINCIPLE OF MASER ACTION

Crystals can be used as microwave and light amplifiers and as sources of coherent radiation. A **maser** amplifies microwaves by the stimulated emission of radiation; a **laser** amplifies light by the same method. The principle, due to Townes, may be understood from the two-level magnetic system of Fig. 21 relevant for masers. There are n_u atoms in the upper state and n_ℓ atoms in the lower state. We immerse the system in radiation at frequency ω; the amplitude of the magnetic component of the radiation field is B_{rf}. The probability per atom per unit time of a transition between the upper and lower states is

$$P = \left(\frac{\mu B_{rf}}{\hbar}\right)^2 \frac{1}{\Delta\omega} \ ; \tag{50}$$

here μ is the magnetic moment, and $\Delta\omega$ is the combined width of the two levels. The result (50) is from a standard result of quantum mechanics, called Fermi's golden rule.

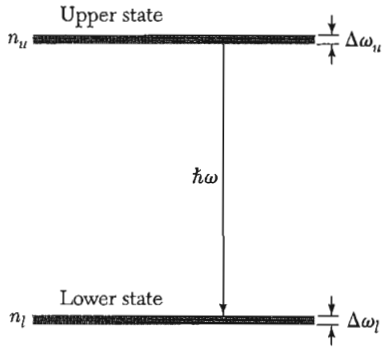


Figure 21 A two-level system, to explain maser operation. The populations of the upper and lower states are n_u and n_l, respectively. The frequency of the emitted radiation is ω; the combined width of the states is $\Delta\omega = \Delta\omega_u + \Delta\omega_l$.

The net energy emitted from atoms in both upper and lower states is

$$\mathcal{P} = \left(\frac{\mu B_{\text{rf}}}{\hbar}\right)^2 \frac{1}{\Delta\omega} \cdot \hbar\omega \cdot (n_u - n_l) \ , \tag{51}$$

per unit time. Here $\mathcal{P}$ denotes the power out; $\hbar\omega$ is the energy per photon; and $n_u - n_l$ is the excess of the number of atoms n_u initially able to emit a photon over the number of atoms n_l able to absorb a photon.

In thermal equilibrium $n_u < n_l$, so there is no net emission of radiation, but in a nonequilibrium condition with $n_u > n_l$ there will be emission. If we start with $n_u > n_l$ and reflect the emitted radiation back onto the system, we increase B_{rf} and thereby stimulate a higher rate of emission. The enhanced stimulation continues until the population in the upper state decreases and becomes equal to the population in the lower state.

We can build up the intensity of the radiation field by placing the crystal in an electromagnetic cavity. This is like multiple reflection from the walls of the cavity. There will be some power loss in the walls of the cavity: the rate of power loss is

$$\text{(CGS)} \quad \mathcal{P}_L = \frac{B_{\text{rf}}^2 V}{8\pi} \cdot \frac{\omega}{Q} \ ; \qquad \text{(SI)} \quad \mathcal{P}_L = \frac{B_{\text{rf}}^2}{2\mu_0} \cdot \frac{\omega}{Q} \ , \tag{52}$$

where V is the volume and Q is the Q factor of the cavity. We understand B_{rf}^2 to be a volume average.

The condition for maser action is that the emitted power $\mathcal{P}$ exceed the power loss $\mathcal{P}_L$. Both quantities involve B_{rf}^2. The maser condition can now be expressed in terms of the population excess in the upper state:

$$\text{(CGS)} \quad n_u - n_l > \frac{V\Delta B}{8\pi\mu Q} \ , \qquad \text{(SI)} \quad n_u - n_l > \frac{V\Delta B}{2\mu_0\mu Q} \ , \tag{53}$$

where μ is the magnetic moment. The line width ΔB is defined in terms of the combined line width $\Delta\omega$ of the upper and lower states as $\mu\Delta B = \hbar\Delta\omega$. The

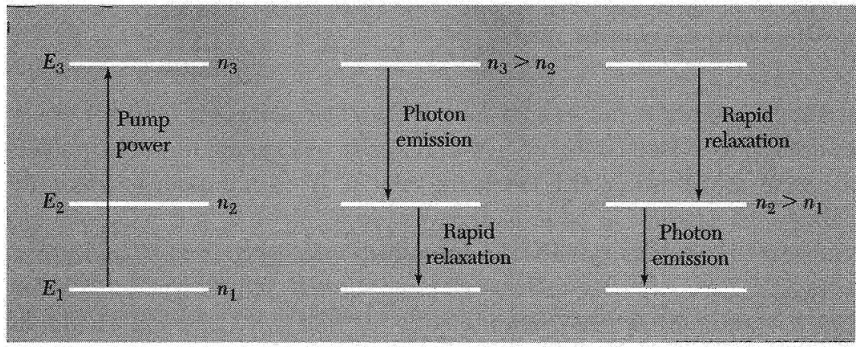


Figure 22 Three-level maser system. Two possible modes of operation are shown, starting from rf saturation of the states 3 and 1 to obtain $n_3 = n_1$.

central problem of the maser or laser is to obtain a suitable excess population in the upper state. This is accomplished in various ways in various devices.

Three-Level Maser

The three-level maser system (Fig. 22) is a clever solution to the excess population problem. Such a system may derive its energy levels from magnetic ions in a crystal, as Bloembergen showed. Rf power is applied at the pump frequency $\hbar\omega_p = E_3 - E_1$ in sufficient intensity to maintain the population of level 3 substantially equal to the population of level 1. This is called saturation—see Problem 6. Now consider the rate of change of the population n_2 of level 2 owing to normal thermal relaxation processes. In terms of the indicated transition rates P,

$$dn_2/dt = -n_2P(2\ 1\ \ 1) - n_2P(2\ 1\ \ 3) + n_3P(3\ 1\ \ 2) + n_1P(1\ 1\ \ 2)\ . \quad (54)$$

In the steady state $dn_2/dt = 0$, and by virtue of the saturation rf power we have $n_3 = n_1$, whence

$$\frac{n_2}{n_1} = \frac{P(3\ 1\ \ 2) + P(1\ 1\ \ 2)}{P(2\ 1\ \ 1) + P(2\ 1\ \ 3)}\ . \quad (55)$$

The transition rates are affected by many details of the paramagnetic ion and its environment, but one can hardly fail with this system, for either $n_2 > n_1$ and we get maser action between levels 2 and 1, or $n_2 < n_1 = n_3$ and we get maser action between levels 3 and 2. The energy levels of the Er^{3+} ion are used in communication fiber optics amplifiers. The ion is optically pumped from level 1 to level 3; there is fast nonradiative decay from level 3 to level 2. The signal at a wavelength of 1.55 μm is amplified by stimulated emission from level 2 to level 1. The wavelength is favorable for long-distance propagation in the optical fiber. The bandwidth is of the order of 4×10^{12} Hz.

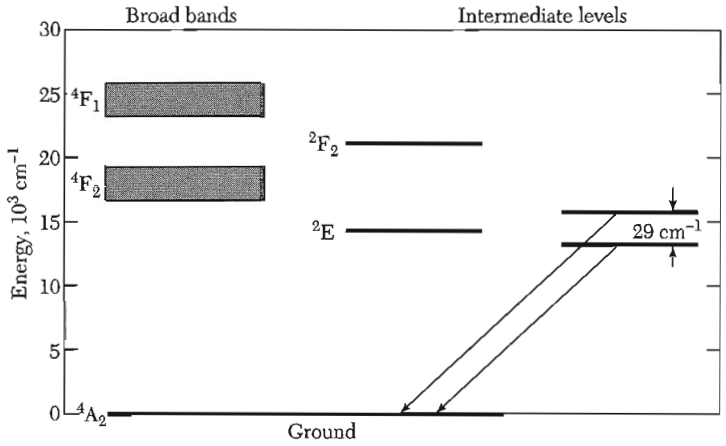


Figure 23 Energy level diagram of Cr^{3+} in ruby, as used in laser operation. The initial excitation takes place to the broad bands; they decay to the intermediate levels by the emission of phonons, and the intermediate levels radiate photons as the ion makes the transition to the ground level.

Lasers

The same crystal, ruby, used in the microwave maser was also the first crystal to exhibit optical maser action, but a different set of energy levels of Cr^{3+} are involved (Fig. 23). About 15,000 cm^{-1} above the ground state there lie a pair of states labeled ^2E, spaced 29 cm^{-1} apart. Above ^2E lie two broad bands of states, labeled $^4F_1$ and $^4F_2$. Because the bands are broad they can be populated efficiently by optical absorption from broadband light sources such as xenon flash lamps.

In operation of a ruby laser both of the broad ^4F bands are populated by broadband light. Atoms thus excited will decay in 10^{-7} sec by radiationless processes with the emission of phonons to the states ^2E. Photon emission from the lower of the states ^2E to the ground state occurs slowly, in about 5×10^{-3} sec, so that a large excited population can pile up in ^2E. For laser action this population must exceed that in the ground state.

The stored energy in ruby is 10^8 erg cm^{-3} if 10^{20} Cr^{3+} ions cm^{-3} are in an excited state. The ruby laser can emit at a very high power level if all this stored energy comes out in a short burst. The overall efficiency of conversion of a ruby laser from input electrical energy to output laser light is about one percent. Another popular solid state laser is the neodymium glass laser, made of calcium tungstate glass doped with Nd^{3+} ions. This operates as a four level system (Fig. 24). Here it is not necessary to empty out the ground state before laser action can occur.

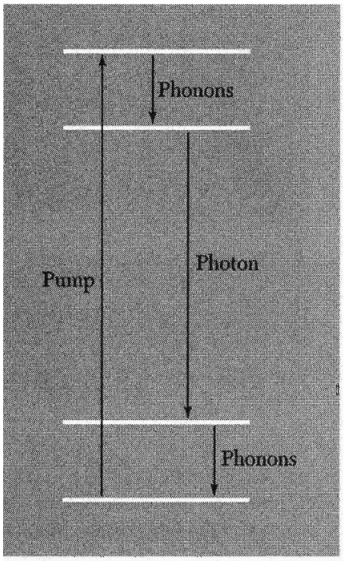


Figure 24 Four-level laser system, as in the neodymium glass laser.

SUMMARY

(In CGS Units)

- The resonance frequency of a free spin is $\omega_0 = \gamma B_0$, where $\gamma = \mu/\hbar I$ is the magnetogyric ratio.
- The Bloch equations are

$$dM_x/dt = \gamma(\mathbf{M} \times \mathbf{B})_x - M_x/T_2 \ ;$$
$$dM_y/dt = \gamma(\mathbf{M} \times \mathbf{B})_y - M_y/T_2 \ ;$$
$$dM_z/dt = \gamma(\mathbf{M} \times \mathbf{B})_z + (M_0 - M_z)/T_1 \ .$$

- The half-width of the resonance at half-power is $(\Delta\omega)_{1/2} = 1/T_2$.
- The dipolar line width in a rigid lattice is $(\Delta B)_0 \approx \mu/a^3$.
- If the magnetic moments are ambulatory, with a characteristic time $\tau \ll 1/(\Delta\omega)_0$, the line width is reduced by the factor $(\Delta\omega)_0\tau$. In this limit $1/T_1 \approx 1/T_2 \approx (\Delta\omega)_0^2\tau$. With exchange coupling in a paramagnet the line width becomes $\approx(\Delta\omega)_0^2\omega_{ex}$.
- The ferromagnetic resonance frequency in an ellipsoid of demagnetization factors N_x, N_y, N_z is $\omega_0^2 = \gamma^2[B_0 + (N_y - N_z)M][B_0 + (N_x - N_z)M]$.
- The antiferromagnetic resonance frequency is $\omega_0^2 = \gamma^2 B_A(B_A + 2B_E)$, in a spherical specimen with zero applied field. Here B_A is the anisotropy field and B_E is the exchange field.
- The condition for maser action is that $n_u - n_l > V\Delta B/8\pi\mu Q$.

Problems

1. ***Equivalent electrical circuit.*** Consider an empty coil of inductance L_0 in a series with a resistance R_0; show if the coil is completely filled with a spin system characterized by the susceptibility components $\chi'(\omega)$ and $\chi''(\omega)$ that the inductance at frequency ω becomes $L = [1 + 4\pi\chi'(\omega)]L_0$, in series with an effective resistance $R = 4\pi\omega\chi''(\omega)L_0 + R_0$. In this problem $\chi = \chi' + i\chi''$ is defined for a linearly polarized rf field. Hint: Consider the impedance of the circuit. (CGS units.)

2. ***Rotating coordinate system.*** We define the vector $\mathbf{F}(t) = F_x(t)\hat{\mathbf{x}} + F_y(t)\hat{\mathbf{y}} + F_z(t)\hat{\mathbf{z}}$. Let the coordinate system of the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ rotate with an instantaneous angular velocity Ω, so that $d\hat{\mathbf{x}}/dt = \Omega_y\hat{\mathbf{z}} - \Omega_z\hat{\mathbf{y}}$, etc. (a) Show that $d\mathbf{F}/dt = (d\mathbf{F}/dt)_R + \Omega \times \mathbf{F}$, where $(d\mathbf{F}/dt)_R$ is the time derivative of $\mathbf{F}$ as viewed in the rotating frame R. (b) Show that (7) may be written $(d\mathbf{M}/dt)_R = \gamma\mathbf{M} \times (\mathbf{B}_a + \Omega/\gamma)$. This is the equation of motion of $\mathbf{M}$ in a rotating coordinate system. The transformation to a rotating system is extraordinarily useful; it is exploited widely in the literature. (c) Let $\Omega = -\gamma B_0\hat{\mathbf{z}}$; thus in the rotating frame there is no static magnetic field. Still in the rotating frame, we now apply a dc pulse $B_1\hat{\mathbf{x}}$ for a time t. If the magnetization is initially along $\hat{\mathbf{z}}$, find an expression for the pulse length t such that the magnetization will be directed along $-\hat{\mathbf{z}}$ at the end of the pulse. (Neglect relaxation effects.) (d) Decribe this pulse as viewed from the laboratory frame of reference.

3. ***Hyperfine effects on ESR in metals.*** We suppose that the electron spin of a conduction electron in a metal sees an effective magnetic field from the hyperfine interaction of the electron spin with the nuclear spin. Let the z component of the field seen by the conduction electron be written

$$B_i = \left(\frac{a}{N}\right)\sum_{j=1}^{N} I_j^z \ ,$$

where I_j^z is equally likely to be $\pm\frac{1}{2}$. (a) Show that $\langle B_i^2\rangle = (a/2N)^2N$. (b) Show that $\langle B_i^4\rangle = 3(a/2N)^4N^2$, for $N \gg 1$.

4. ***FMR in the anisotropy field.*** Consider a spherical specimen of a uniaxial ferromagnetic crystal with an anisotropy energy density of the form $U_K = K\sin^2\theta$, where θ is the angle between the magnetization and the z axis. We assume that K is positive. Show that the ferromagnetic resonance frequency in an external magnetic field $B_0\hat{\mathbf{z}}$ is $\omega_0 = \gamma(B_0 + B_A)$, where $B_A = 2K/M_s$.

5. ***Exchange frequency resonance.*** Consider a ferrimagnet with two sublattices A and B of magnetizations $\mathbf{M}_A$ and $\mathbf{M}_B$, where $\mathbf{M}_B$ is opposite to $\mathbf{M}_A$ when the spin system is at rest. The gyromagnetic ratios are γ_A, γ_B and the molecular fields are $\mathbf{B}_A = -\lambda\mathbf{M}_B$; $\mathbf{B}_B = -\lambda\mathbf{M}_A$. Show that there is a resonance at

$$\omega_0^2 = \lambda^2(\gamma_A|M_B| - \gamma_B|M_A|)^2 \ .$$

This is called the exchange frequency resonance.

6. ***Rf saturation.*** Given, at equilibrium for temperature T, a two-level spin system in a magnetic field $H_0\hat{\mathbf{z}}$, with populations N_1, N_2 and transition rates W_{12}, W_{21}. We

apply an rf signal that gives a transition rate W_{rf}. (a) Derive the equation for dM_z/dt and show that in the steady state

$$M_z = M_0/(1 + 2W_{rf}T_1) \ ,$$

where $1/T_1 = W_{12} + W_{21}$. It will be helpful to write $N = N_1 + N_2$; $n = N_1 - N_2$; and $n_0 = N(W_{21} - W_{12})/(W_{21} + W_{12})$. We see that as long as $2W_{rf}T_1 \ll 1$ the absorption of energy from the rf field does not substantially alter the population distribution from its thermal equilibrium value. (b) Using the expression for n, write down the rate at which energy is absorbed from the rf field. What happens as W_{rf} approaches $1/2T_1$? This effect is called saturation, and its onset may be used to measure T_1.

14

Plasmons, Polaritons, and Polarons

NOTE: The text and problems of this chapter assume facility in the use of electromagnetic theory at the level of a good senior course.

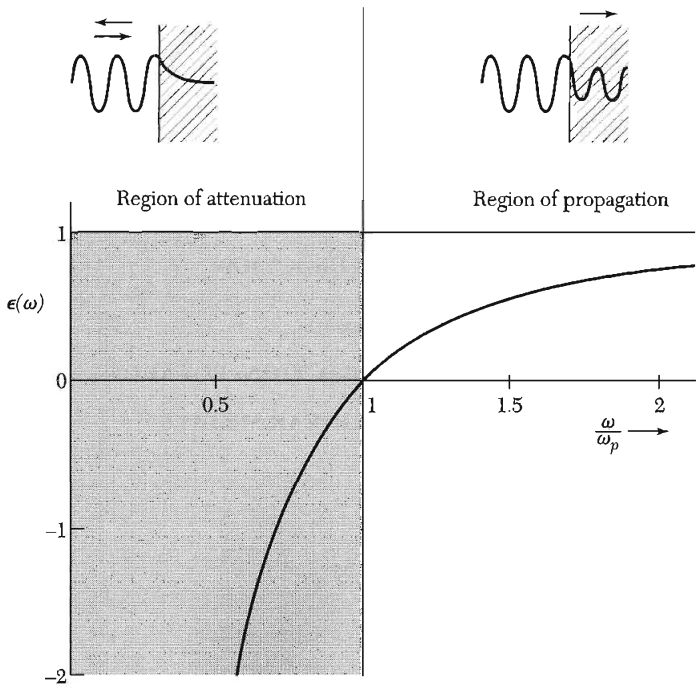


Figure 1 Dielectric function $\epsilon(\omega)$ or $\epsilon(\omega, 0)$ of a free-electron gas versus frequency in units of the plasma frequency ω_p. Electromagnetic waves propagate without damping only when ϵ is positive and real. Electromagnetic waves are totally reflected from the medium when ϵ is negative.

DIELECTRIC FUNCTION OF THE ELECTRON GAS

The dielectric function $\epsilon(\omega,\mathbf{K})$ of the electron gas, with its strong dependence on frequency and wavevector, has significant consequences for the physical properties of solids. In one limit, $\epsilon(\omega,0)$ describes the collective excitations of the Fermi sea—the volume and surface plasmons. In another limit, $\epsilon(0,\mathbf{K})$ describes the electrostatic screening of the electron-electron, electron-lattice, and electron-impurity interactions in crystals.

We will also use the dielectric function of an ionic crystal to derive the polariton spectrum. Later we discuss the properties of polarons. But first we are concerned with the electron gas in metals.

Definitions of the Dielectric Function. The dielectric constant ϵ of electrostatics is defined in terms of the electric field $\mathbf{E}$ and the polarization $\mathbf{P}$, the dipole moment density:

$$\text{(CGS)} \quad \mathbf{D} = \mathbf{E} + 4\pi\mathbf{P} = \epsilon\mathbf{E} \ ; \qquad \text{(SI)} \quad \mathbf{D} = \epsilon_0\mathbf{E} + \mathbf{P} = \epsilon\epsilon_0\mathbf{E} \ . \tag{1}$$

Thus defined, ϵ is also known as the relative permittivity.

The introduction of the displacement $\mathbf{D}$ is motivated by the usefulness of this vector related to the external applied charge density ρ_{ext} in the same way as $\mathbf{E}$ is related to the total charge density $\rho = \rho_{\text{ext}} + \rho_{\text{ind}}$, where ρ_{ind} is the charge density induced in the system by ρ_{ext}.

Thus the divergence relation of the electric field is

(CGS)	(SI)	
$\operatorname{div}\mathbf{D} = \operatorname{div}\epsilon\mathbf{E} = 4\pi\rho_{\text{ext}}$	$\operatorname{div}\mathbf{D} = \operatorname{div}\epsilon\epsilon_0\mathbf{E} = \rho_{\text{ext}}$	(2)
$\operatorname{div}\mathbf{E} = 4\pi\rho = 4\pi(\rho_{\text{ext}} + \rho_{\text{ind}})$	$\operatorname{div}\mathbf{E} = \rho/\epsilon_0 = (\rho_{\text{ext}} + \rho_{\text{ind}})/\epsilon_0$.	(3)

Parts of this chapter will be written in CGS; to obtain results in SI, write $1/\epsilon_0$ for 4π.

We need relations between the Fourier components of $\mathbf{D}$, $\mathbf{E}$, ρ, and the electrostatic potential φ. For brevity we do not exhibit here the frequency dependence. Define $\epsilon(\mathbf{K})$ such that

$$\mathbf{D}(\mathbf{K}) = \epsilon(\mathbf{K})\mathbf{E}(\mathbf{K}) \ ; \tag{3a}$$

then (3) becomes

$$\text{div } \mathbf{E} = \text{div} \sum \mathbf{E}(\mathbf{K}) \exp(i\mathbf{K} \cdot \mathbf{r}) = 4\pi \sum \rho(\mathbf{K}) \exp(i\mathbf{K} \cdot \mathbf{r}) \ , \tag{3b}$$

and (2) becomes

$$\text{div } \mathbf{D} = \text{div} \sum \epsilon(\mathbf{K})\mathbf{E}(\mathbf{K}) \exp(i\mathbf{K}\cdot\mathbf{r}) = 4\pi \sum \rho_{\text{ext}}(\mathbf{K}) \exp(i\mathbf{K}\cdot\mathbf{r}) \ . \tag{3c}$$

Each of the equations must be satisfied term by term; we divide one by the other to obtain

$$\epsilon(\mathbf{K}) = \frac{\rho_{\text{ext}}(\mathbf{K})}{\rho(\mathbf{K})} = 1 - \frac{\rho_{\text{ind}}(\mathbf{K})}{\rho(\mathbf{K})} \ . \tag{3d}$$

The electrostatic potential φ_{ext} defined by $-\nabla\varphi_{\text{ext}} = \mathbf{D}$ satisfies the Poisson equation $\nabla^2\varphi_{\text{ext}} = -4\pi\rho_{\text{ext}}$; and the electrostatic potential φ defined by $-\nabla\varphi = \mathbf{E}$ satisfies $\nabla^2\varphi = -4\pi\rho$. The Fourier components of the potentials must therefore satisfy

$$\frac{\varphi_{\text{ext}}(\mathbf{K})}{\varphi(\mathbf{K})} = \frac{\rho_{\text{ext}}(\mathbf{K})}{\rho(\mathbf{K})} = \epsilon(\mathbf{K}) \ , \tag{3e}$$

by (3d). We use this relation in the treatment of the screened coulomb potential.

Plasma Optics

The long wavelength dielectric response $\epsilon(\omega,0)$ or $\epsilon(\omega)$ of an electron gas is obtained from the equation of motion of a free electron in an electric field:

$$m\frac{d^2x}{dt^2} = -eE \ . \tag{4}$$

If x and E have the time dependence $e^{-i\omega t}$, then

$$-\omega^2 mx = -eE \ ; \qquad x = eE/m\omega^2 \ . \tag{5}$$

The dipole moment of one electron is $-ex = -e^2E/m\omega^2$, and the polarization, defined as the dipole moment per unit volume, is

$$P = -nex = -\frac{ne^2}{m\omega^2}E \ , \tag{6}$$

where n is the electron concentration.

The dielectric function at frequency ω is

(CGS)
$$\epsilon(\omega) \equiv \frac{D(\omega)}{E(\omega)} \equiv 1 + 4\pi\frac{P(\omega)}{E(\omega)} \ ; \tag{7}$$

(SI)
$$\epsilon(\omega) = \frac{D(\omega)}{\epsilon_0 E(\omega)} = 1 + \frac{P(\omega)}{\epsilon_0 E(\omega)} \ .$$

The dielectric function of the free electron gas follows from (6) and (7):

$$\text{(CGS)} \quad \epsilon(\omega) = 1 - \frac{4\pi ne^2}{m\omega^2} \; ; \qquad \text{(SI)} \quad \epsilon(\omega) = 1 - \frac{ne^2}{\epsilon_0 m\omega^2} \; . \tag{8}$$

The **plasma frequency** ω_p is defined by the relation

$$\text{(CGS)} \quad \omega_p^2 = 4\pi ne^2/m \; ; \qquad \text{(SI)} \quad \omega_p^2 = ne^2/\epsilon_0 m \; . \tag{9}$$

A plasma is a medium with equal concentration of positive and negative charges, of which at least one charge type is mobile. In a solid the negative charges of the conduction electrons are balanced by an equal concentration of positive charge of the ion cores. We write the dielectric function (8) as

$$\boxed{\epsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \; ,} \tag{10}$$

plotted in Fig. 1.

If the positive ion core background has a dielectric constant labeled $\epsilon(\infty)$ essentially constant up to frequencies well above ω_p, then (8) becomes

$$\epsilon(\omega) = \epsilon(\infty) - 4\pi ne^2/m\omega^2 = \epsilon(\infty)[1 - \bar{\omega}_p^2/\omega^2] \; , \tag{11}$$

where $\bar{\omega}_p$ is defined as

$$\bar{\omega}_p^2 = 4\pi ne^2/\epsilon(\infty)m \; . \tag{12}$$

Notice that $\epsilon = 0$ at $\omega = \bar{\omega}_p$.

Dispersion Relation for Electromagnetic Waves

In a nonmagnetic isotropic medium the electromagnetic wave equation is

$$\text{(CGS)} \quad \partial^2\mathbf{D}/\partial t^2 = c^2\nabla^2\mathbf{E} \; ; \qquad \text{(SI)} \quad \mu_0\partial^2\mathbf{D}/\partial t^2 = \nabla^2\mathbf{E} \; . \tag{13}$$

We look for a solution with $\mathbf{E} \propto \exp(-i\omega t)\exp(i\mathbf{K}\cdot\mathbf{r})$ and $\mathbf{D} = \epsilon(\omega,\mathbf{K})\mathbf{E}$; then we have the dispersion relation for electromagnetic waves:

$$\text{(CGS)} \quad \epsilon(\omega,\mathbf{K})\omega^2 = c^2K^2 \; ; \qquad \text{(SI)} \quad \epsilon(\omega,\mathbf{K})\epsilon_0\mu_0\omega^2 = K^2 \; . \tag{14}$$

This relation tells us a great deal. Consider

- ϵ real and > 0. For ω real, K is real and a transverse electromagnetic wave propagates with the phase velocity $c/\epsilon^{1/2}$.
- ϵ real and < 0. For ω real, K is imaginary and the wave is damped with a characteristic length $1/|\mathbf{K}|$.
- ϵ complex. For ω real, $\mathbf{K}$ is complex and the waves are damped in space.

- $\epsilon = \infty$. This means the system has a finite response in the absence of an applied force; thus the poles of $\epsilon(\omega,\mathbf{K})$ define the frequencies of the free oscillations of the medium.
- $\epsilon = 0$. We shall see that longitudinally polarized waves are possible only at the zeros of ϵ.

Transverse Optical Modes in a Plasma

The dispersion relation (14) becomes, with (11) for $\epsilon(\omega)$,

$$\text{(CGS)} \qquad \epsilon(\omega)\omega^2 = \epsilon(\infty)(\omega^2 - \bar{\omega}_p^2) = c^2K^2 \ . \tag{15}$$

For $\omega < \bar{\omega}_p$ we have $K^2 < 0$, so that K is imaginary. The solutions of the wave equation are of the form $\exp(-|K|x)$ in the frequency region $0 < \omega \leq \bar{\omega}_p$. Waves incident on the medium in this frequency region do not propagate, but will be totally reflected.

An electron gas is transparent when $\omega > \bar{\omega}_p$, for here the dielectric function is positive real. The dispersion relation in this region may be written as

$$\text{(CGS)} \qquad \omega^2 = \bar{\omega}_p^2 + c^2K^2/\epsilon(\infty) \ ; \tag{16}$$

this describes transverse electromagnetic waves in a plasma (Fig. 2).

Values of the plasma frequency ω_p and of the free space wavelength $\lambda_p \equiv 2\pi c/\omega_p$ for electron concentrations of interest are given below. A wave will propagate if its free space wavelength is less than λ_p; otherwise the wave is reflected.

n, electrons/cm^3	10^{22}	10^{18}	10^{14}	10^{10}
ω_p, s^{-1}	5.7×10^{15}	5.7×10^{13}	5.7×10^{11}	5.7×10^{9}
λ_p, cm	3.3×10^{-5}	3.3×10^{-3}	0.33	33

Transparency of Metals in the Ultraviolet. From the preceding discussion of the dielectric function we conclude that simple metals should reflect light in the visible region and be transparent to light at high frequencies. A comparison of calculated and observed cutoff wavelengths is given in Table 1. The reflection of light from a metal is entirely similar to the reflection of radio waves from the ionosphere, for the free electrons in the ionosphere make the dielectric constant negative at low frequencies. Experimental results for InSb with $n = 4 \times 10^{18}$ cm^{-3} are shown in Fig. 3, where the plasma frequency is near 0.09 eV.

Longitudinal Plasma Oscillations

The zeros of the dielectric function determine the frequencies of the longitudinal modes of oscillation. That is, the condition

$$\epsilon(\omega_L) = 0 \tag{17}$$

determines the longitudinal frequency ω_L near $K = 0$.

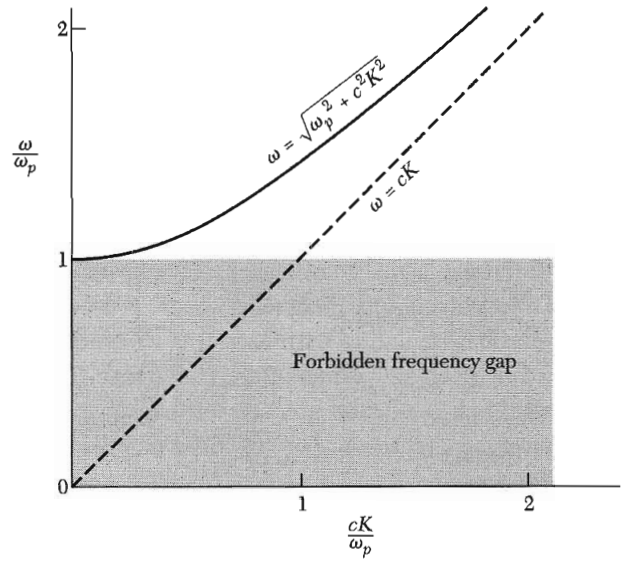


Figure 2 Dispersion relation for transverse electromagnetic waves in a plasma. The group velocity $v_g = d\omega/dK$ is the slope of the dispersion curve. Although the dielectric function is between zero and one, the group velocity is less than the velocity of light in vacuum.

Table 1 Ultraviolet transmission limits of alkali metals, in Å

	Li	Na	K	Rb	Cs
λ_p, calculated	1550	2090	2870	3220	3620
λ_p, observed	1550	2100	3150	3400	—

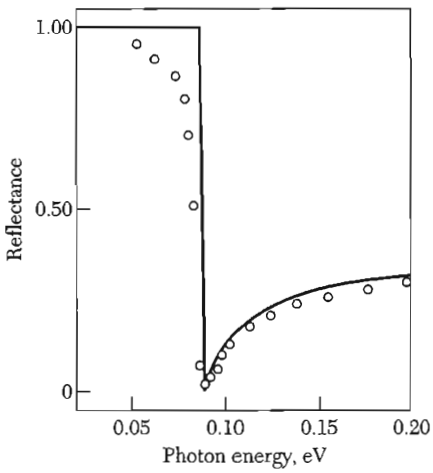


Figure 3 Reflectance of indium antimonide with $n = 4 \times 10^{18}$ cm^{-3}. (After J. N. Hodgson.)

By the geometry of a longitudinal polarization wave there is a depolarization field $\mathbf{E} = -4\pi\mathbf{P}$, discussed below. Thus $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P} = 0$ for a longitudinal wave in a plasma or more generally in a crystal. In SI units, $\mathbf{D} = \epsilon_0\mathbf{E} + \mathbf{P} = 0$.

For an electron gas, at the zero (17) of the dielectric function (10)

$$\epsilon(\omega_L) = 1 - \omega_p^2/\omega_L^2 = 0 \ , \tag{18}$$

whence $\omega_L = \omega_p$. Thus there is a free longitudinal oscillation mode (Fig. 4) of an electron gas at the plasma frequency described by (15) as the low-frequency cutoff of transverse electromagnetic waves.

A longitudinal plasma oscillation with $K = 0$ is shown in Fig. 5 as a uniform displacement of an electron gas in a thin metallic slab. The electron gas is moved as a whole with respect to the positive ion background. The displacement u of the electron gas creates an electric field $E = 4\pi neu$ that acts as a restoring force on the gas.

The equation of motion of a unit volume of the electron gas of concentration n is

$$\text{(CGS)} \qquad nm\frac{d^2u}{dt^2} = -neE = -4\pi n^2e^2u \ , \tag{19}$$

or

$$\text{(CGS)} \qquad \frac{d^2u}{dt^2} + \omega_p^2u = 0 \ ; \qquad \omega_p = \left(\frac{4\pi ne^2}{m}\right)^{1/2} \ . \tag{20}$$

This is the equation of motion of a simple harmonic oscillator of frequency ω_p, the plasma frequency. The expression for ω_p is identical with (9), which arose in a different connection. In SI, the displacement u creates the electric field $E = neu/\epsilon_0$, whence $\omega_p = (ne^2/\epsilon_0 m)^{1/2}$.

A plasma oscillation of small wavevector has approximately the frequency ω_p. The wavevector dependence of the dispersion relation for longitudinal oscillations in a Fermi gas is given by

$$\omega \cong \omega_p\,(1 + 3k^2v_F^2/10\omega_p^2 + \cdots) \ , \tag{21}$$

where v_F is the electron velocity at the Fermi energy.

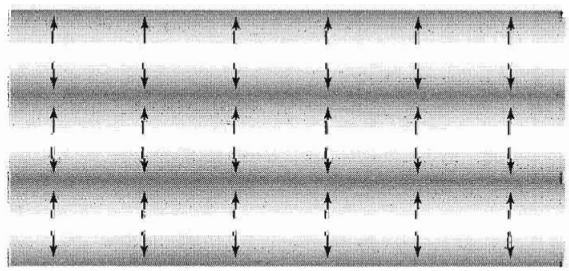


Figure 4 A plasma oscillation. The arrows indicate the direction of displacement of the electrons.

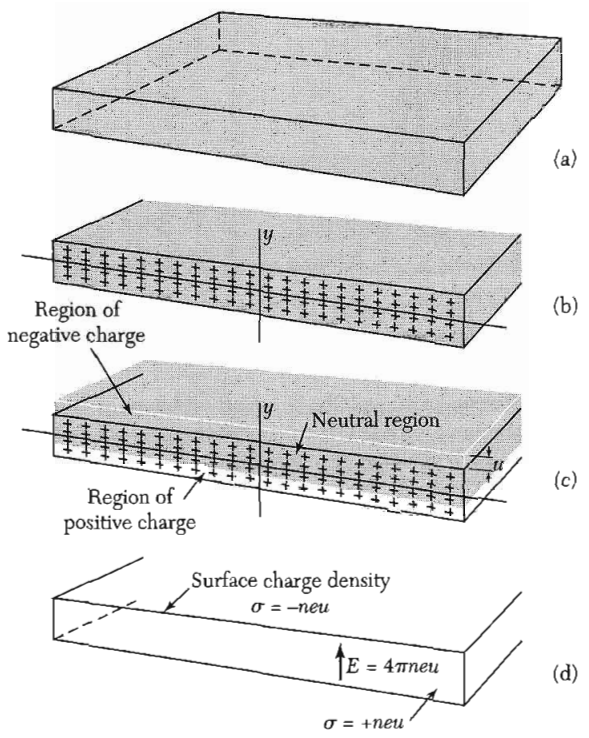


Figure 5 In (a) is shown a thin slab or film of a metal. A cross section is shown in (b), with the positive ion cores indicated by + signs and the electron sea indicated by the gray background. The slab is electrically neutral. In (c) the negative charge has been displaced upward uniformly by a small distance u, shown exaggerated in the figure. As in (d), this displacement establishes a surface charge density $-neu$ on the upper surface of the slab and $+neu$ on the lower surface, where n is the electron concentration. An electric field $E = 4\pi neu$ is produced inside the slab. This field tends to restore the electron sea to its equilibrium position (b). In SI units, $E = neu/\epsilon_0$.

PLASMONS

A plasma oscillation in a metal is a collective longitudinal excitation of the conduction electron gas. A **plasmon** is a quantum of a plasma oscillation; we may excite a plasmon by passing an electron through a thin metallic film (Figs. 6 and 7) or by reflecting an electron or a photon from a film. The charge of the electron couples with the electrostatic field fluctuations of the plasma oscillations. The reflected or transmitted electron will show an energy loss equal to integral multiples of the plasmon energy.

Experimental excitation spectra for Al and Mg are shown in Fig. 8. A comparison of observed and calculated values of plasmon energies is given in Table 2; further data are given in the reviews by Raether and by Daniels. Recall that $\overline{\omega}_p$ as defined by (12) includes the ion core effects by use of $\epsilon(\infty)$.

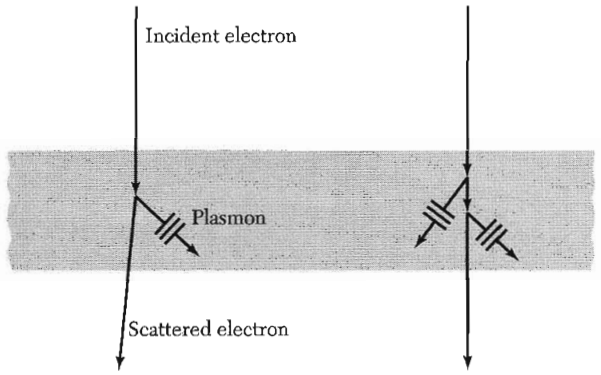


Figure 6 Creation of a plasmon in a metal film by inelastic scattering of an electron. The incident electron typically has an energy 1 to 10 keV; the plasmon energy may be of the order of 10 eV. An event is also shown in which two plasmons are created.

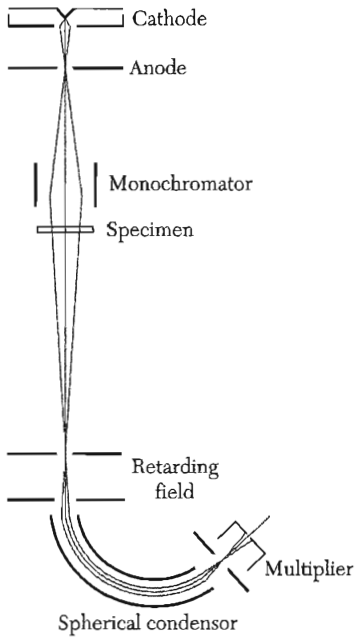


Figure 7 A spectrometer with electrostatic analyzer for the study of plasmon excitation by electrons. (After J. Daniels et al.)

It is equally possible to excite collective plasma oscillations in dielectric films; results for several dielectrics are included. The calculated plasma energies of Si, Ge, and InSb are based on four valence electrons per atom. In a dielectric the plasma oscillation is physically the same as in a metal; the entire valence electron gas oscillates back and forth with respect to the ion cores.

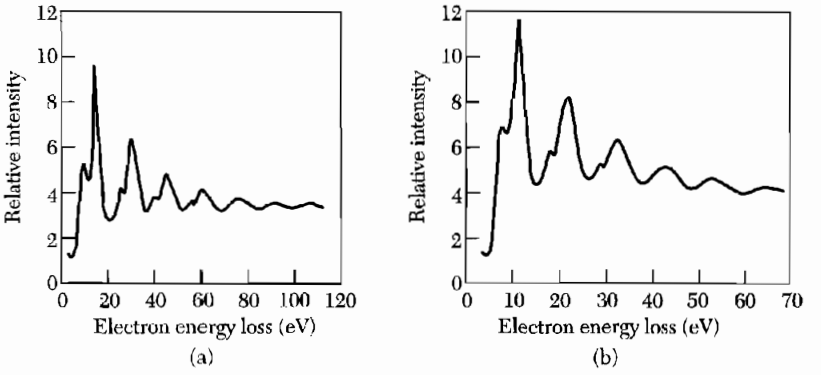


Figure 8 Energy loss spectra for electrons reflected from films of (a) aluminum and (b) magnesium, for primary electron energies of 2020 eV. The 12 loss peaks observed in Al are made up of combinations of 10.3 and 15.3 eV losses, where the 10.3 eV loss is due to surface plasmons and the 15.3 eV loss is due to volume plasmons. The ten loss peaks observed in Mg are made up of combinations of 7.1 eV surface plasmons and 10.6 eV volume plasmons. Surface plasmons are the subject of Problem 1. (After C. J. Powell and J. B. Swan.)

Table 2 Volume plasmon energies, in eV

		Calculated	
Material	Observed	$\hbar\omega_p$	$\hbar\bar{\omega}_p$
Metals			
Li	7.12	8.02	7.96
Na	5.71	5.95	5.58
K	3.72	4.29	3.86
Mg	10.6	10.9	
Al	15.3	15.8	
Dielectrics			
Si	16.4–16.9	16.0	
Ge	16.0–16.4	16.0	
InSb	12.0–13.0	12.0	

ELECTROSTATIC SCREENING

The electric field of a positive charge embedded in an electron gas falls off with increasing r faster than $1/r$, because the electron gas tends to gather around and thus to screen the positive charge. The static screening can be described by the wavevector dependence of the static dielectric function $\epsilon(0,\mathbf{K})$. We consider the response of the electrons to an applied external electrostatic field. We start with a uniform gas of electrons of charge concentration $-n_0e$ superimposed on a background of positive charge of concentration n_0e. Let the

positive charge background be deformed mechanically to produce a sinusoidal variation of positive charge density in the x direction:

$$\rho^+(x) = n_0 e + \rho_{ext}(K) \sin Kx \ . \tag{22}$$

The term $\rho_{ext}(K) \sin Kx$ gives rise to an electrostatic field that we call the external field applied to the electron gas.

The electrostatic potential φ of a charge distribution is found from the Poisson equation $\nabla^2\varphi = -4\pi\rho$, by (3) with $\mathbf{E} = -\nabla\varphi$. For the positive charge we have

$$\varphi = \varphi_{ext}(K) \sin Kx \ ; \quad \rho = \rho_{ext}(K) \sin Kx \ . \tag{23}$$

The Poisson equation gives the relation

$$K^2\varphi_{ext}(K) = 4\pi\rho_{ext}(K) \ . \tag{24}$$

The electron gas will be deformed by the combined influences of the electrostatic potential $\varphi_{ext}(K)$ of the positive charge distribution and of the as yet unknown induced electrostatic potential $\varphi_{ind}(K) \sin Kx$ of the deformation of the electron gas itself. The electron charge density is

$$\rho^-(x) = -n_0 e + \rho_{ind}(K) \sin Kx \ , \tag{25}$$

where $\rho_{ind}(K)$ is the amplitude of the charge density variation induced in the electron gas. We want to find $\rho_{ind}(K)$ in terms of $\rho_{ext}(K)$.

The amplitude of the total electrostatic potential $\varphi(K) = \varphi_{ext}(K) + \varphi_{ind}(K)$ of the positive and negative charge distributions is related to the total charge density variation $\rho(K) = \rho_{ext}(K) + \rho_{ind}(K)$ by the Poisson equation. Then, as in Eq. (24),

$$K^2\varphi(K) = 4\pi\rho(K) \ . \tag{26}$$

To go further we need another equation that relates the electron concentration to the electrostatic potential. We develop this connection in what is called the Thomas-Fermi approximation. The approximation consists in assuming that a local internal chemical potential can be defined as a function of the electron concentration at that point. Now the total chemical potential of the electron gas is constant in equilibrium, independent of position. In a region where there is no electrostatic contribution to the chemical potential we have

$$\mu = \epsilon_F^0 = \frac{\hbar^2}{2m}(3\pi^2 n_0)^{2/3} \tag{27}$$

at absolute zero, according to (6.17). In a region where the electrostatic potential is $\varphi(x)$, the total chemical potential (Fig. 9) is constant and equal to

$$\mu = \epsilon_F(x) - e\varphi(x) \simeq \frac{\hbar^2}{2m}[3\pi^2 n(x)]^{2/3} - e\varphi(x) \cong \frac{\hbar^2}{2m}[3\pi^2 n_0]^{2/3} \ , \tag{28}$$

where $\epsilon_F(x)$ is the local value of the Fermi energy.

The expression (28) is valid for static electrostatic potentials that vary slowly compared with the wavelength of an electron at the Fermi level;

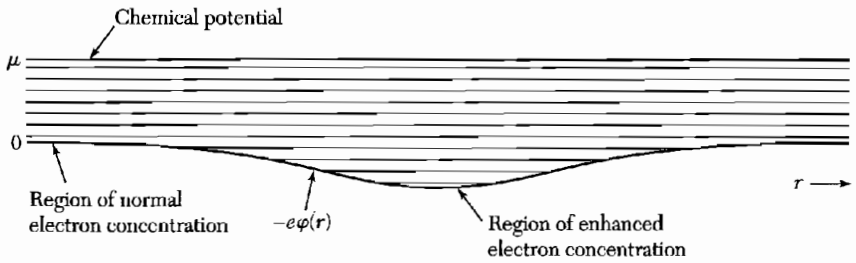


Figure 9 In thermal and diffusive equilibrium the chemical potential is constant; to maintain it constant we increase the electron concentration in regions of space where the potential energy is low, and we decrease the concentration where the potential is high.

specifically, the approximation is $q \ll k_F$. By a Taylor series expansion of ϵ_F, Eq. (28) may be written as

$$\frac{d\epsilon_F}{dn_0}[n(x) - n_0] \cong e\varphi(x) \ . \tag{29}$$

From (27) we have $d\epsilon_F/dn_0 = 2\epsilon_F/3n_0$, whence

$$n(x) - n_0 \cong \frac{3}{2} n_0 \frac{e\varphi(x)}{\epsilon_F} \ . \tag{30}$$

The left-hand side is the induced part of the electron concentration; thus the Fourier components of this equation are

$$\rho_{\text{ind}}(K) = -(3n_0e^2/2\epsilon_F)\varphi(K) \ . \tag{31}$$

By (26) this becomes

$$\rho_{\text{ind}}(K) = -(6\pi n_0e^2/\epsilon_F K^2)\rho(K) \ . \tag{32}$$

By (3d) we have

$$\epsilon(0,K) = 1 - \frac{\rho_{\text{ind}}(K)}{\rho(K)} = 1 + k_s^2/K^2 \ ; \tag{33}$$

here, after some rearrangement,

$$k_s^2 = 6\pi n_0e^2/\epsilon_F = 4(3/\pi)^{1/3}\, n_0^{1/3}/a_0 = 4\pi e^2 D(\epsilon_F) \ , \tag{34}$$

where a_0 is the Bohr radius and $D(\epsilon_F)$ is the density of states for a free electron gas. The approximation (33) for $\epsilon(0,K)$ is called the Thomas-Fermi dielectric function, and $1/k_s$ is the Thomas-Fermi screening length, as in (40) below. For copper with $n_0 = 8.5 \times 10^{22}\ \text{cm}^{-3}$, the screening length is 0.55 Å.

We have derived two limiting expressions for the dielectric function of an electron gas:

$$\epsilon(0,K) = 1 + \frac{k_s^2}{K^2} \ ; \qquad \epsilon(\omega,0) = 1 - \frac{\omega_p^2}{\omega^2} \ . \tag{35}$$

We notice that $\epsilon(0,K)$ as $K \to 0$ does not approach the same limit as $\epsilon(\omega,0)$ as $\omega \to 0$. This means that great care must be taken with the dielectric function

near the origin of the ω-K plane. The full theory for the general function $\epsilon(\omega,K)$ is due to Lindhard.[1]

Screened Coulomb Potential. We consider a point charge q placed in a sea of conduction electrons. The Poisson equation for the unscreened coulomb potential is

$$\nabla^2\varphi_0 = -4\pi q\delta(\mathbf{r}) \ , \tag{36}$$

and we know that $\varphi_0 = q/r$. Let us write

$$\varphi_0(\mathbf{r}) = (2\pi)^{-3} \int d\mathbf{K}\, \varphi_0(\mathbf{K}) \exp(i\mathbf{K}\cdot\mathbf{r}) \ . \tag{37}$$

We use in (36) the Fourier representation of the delta function:

$$\delta(\mathbf{r}) = (2\pi)^{-3} \int d\mathbf{K} \exp(i\mathbf{K}\cdot\mathbf{r}) \ , \tag{38}$$

whence $K^2\varphi_0(K) = 4\pi q$.

By (3e),

$$\varphi_0(K)/\varphi(K) = \epsilon(K) \ ,$$

where $\varphi(K)$ is the total or screened potential. We use $\epsilon(K)$ in the Thomas-Fermi form (33) to find

$$\varphi(\mathbf{K}) = \frac{4\pi q}{K^2 + k_s^2} \ . \tag{39}$$

The screened coulomb potential is the transform of $\varphi(\mathbf{K})$:

$$\begin{aligned}\varphi(r) &= \frac{4\pi q}{(2\pi)^3} \int_0^\infty dK \frac{2\pi K^2}{K^2 + k_s^2} \int_{-1}^1 d(\cos\theta) \exp(iKr\cos\theta) \\ &= \frac{2q}{\pi r} \int_0^\infty dK \frac{K\sin Kr}{K^2 + k_s^2} = \frac{q}{r}\exp(-k_s r)\end{aligned} \tag{40}$$

as in Fig. 10a. The screening parameter k_s is defined by (34). The exponential factor reduces the range of the coulomb potential. The bare potential q/r is obtained on letting the charge concentration $n_0 \to 0$, for then $k_s \to 0$. In the vacuum limit $\varphi(K) = 4\pi q/K^2$.

One application of the screened interaction is to the resistivity of certain alloys. The atoms of the series Cu, Zn, Ga, Ge, As have valences 1, 2, 3, 4, 5. An atom of Zn, Ga, Ge, or As added substitutionally to metallic Cu has an excess charge, referred to Cu, of 1, 2, 3, or 4 if all the valence electrons join the conduction band of the host metal. The foreign atom scatters the conduction electrons, with an interaction given by the screened coulomb potential. This scattering contributes to the residual electrical resistivity, and calculations by Mott of the resistivity increase are in fair agreement with experiment.

[1]A good discussion of the Lindhard dielectric function is given by J. Ziman, *Principles of the theory of solids*, 2nd ed., Cambridge, 1972, Chapter 5. The algebraic steps in the evaluation of Ziman's equation (5.16) are given in detail by C. Kittel, Solid state physics **22**, 1 (1968), Section 6.

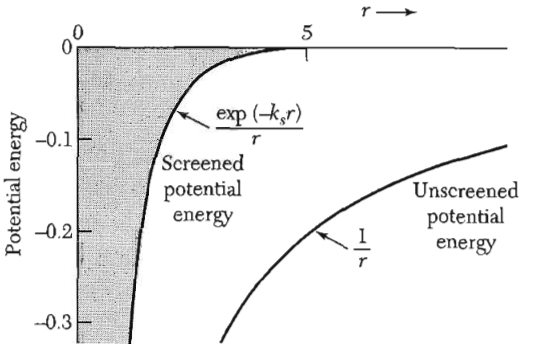


Figure 10a *Comparison of screened and unscreened* coulomb potentials of a static unit positive charge. The screening length $1/k_s$ is set equal to unity. The static screened interaction is taken in the Thomas-Fermi approximation, which holds for low wavevectors $q \ll k_F$; More complete calculations with all wavevectors included exhibit spatial oscillations, called **Friedel oscillations**, in $2k_F r$ and are plotted in QTS, p. 114.

Pseudopotential Component $U(0)$. In the legend to Fig. 9.22b we stated a result that is important in pseudopotential theory: "For very small k the potential approaches $-\frac{2}{3}$ times the Fermi energy." The result, which is known as the screened ion limit for metals, can be derived from Eq. (39). When converted to the potential energy of an electron of charge e in a metal of valency z with n_0 ions per unit volume, the potential energy component at $k = 0$ becomes

$$U(0) = -ezn_0\varphi(0) = -4\pi zn_0e^2/k_s^2 \ . \tag{41}$$

The result (34) for k_s^2 in this situation reads

$$k_s^2 = 6\pi zn_0e^2/\epsilon_F \ , \tag{42}$$

whence

$$U(0) = -\tfrac{2}{3}\epsilon_F \ . \tag{43}$$

Mott Metal-Insulator Transition

A crystal composed of one hydrogen atom per primitive cell should always be a metal, according to the independent-electron model, because there will always be a half-filled energy band within which charge transport can take place. A crystal with one hydrogen molecule per primitive cell is a different matter, because the two electrons can fill a band. Under extreme high pressure, as in the planet Jupiter, it is possible that hydrogen occurs in a metallic form.

But let us imagine a lattice of hydrogen atoms at absolute zero: will this be a metal or an insulator? The answer depends on the lattice constant, with small values of a giving a metal and large values giving an insulator. Mott made an early estimate of the critical value a_c of the lattice constant that separates the metallic state from the insulating state: $a_c = 4.5a_0$, where $a_0 = \hbar^2/me^2$ is the radius of the first Bohr orbit of a hydrogen atom.

On one approach to the problem, we start in the metallic state where a conduction electron sees a screened coulomb interaction from each proton:

$$U(r) = -(e^2/r)\exp(-k_s r) \ , \tag{44}$$

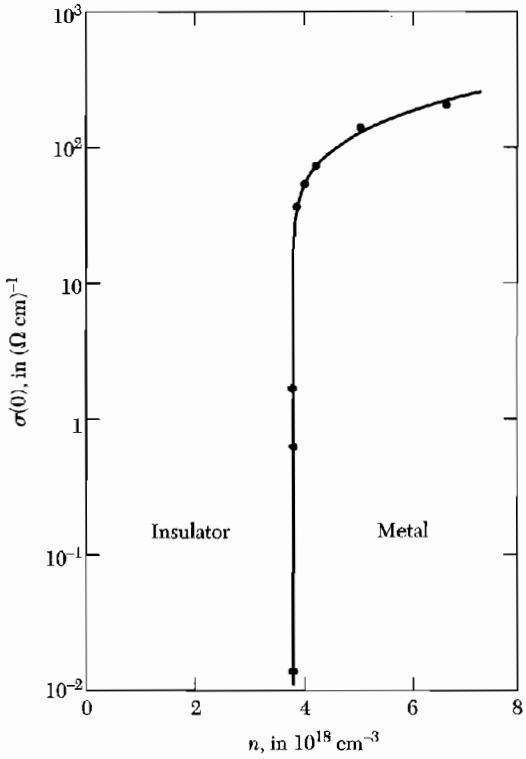


Figure 10b Semilog plot of observed "zero temperature" conductivity $\sigma(0)$ versus donor concentration n for phosphorous donors in silicon. (After T. F. Rosenbaum, et al.)

where $k_s^2 = 3.939 n_0^{1/3}/a_0$, as in (34), where n_0 is the electron concentration. At high concentrations k_s is large and the potential has no bound state, so that we must have a metal.

The potential is known to have a bound state when k_s is smaller than $1.19/a_0$. With a bound state possible the electrons may condense about the protons to form an insulator. The inequality may be written in terms of n_0 as

$$3.939 n_0^{1/3}/a_0 < 1.42/a_0^2 \ . \tag{45}$$

With $n_0 = 1/a^3$ for a simple cubic lattice, we may have an insulator when $a_c > 2.78 a_0$, which is not far from the Mott result $4.5 a_0$ found in a different way.

The term *metal-insulator transition* has come to denote situations where the electrical conductivity of a material changes from metal to insulator as a function of some external parameter, which may be composition, pressure, strain, or magnetic field. The metallic phase may usually be pictured in terms of an independent-electron model; the insulator phase may suggest important electron-electron interactions. Sites randomly occupied introduce new and interesting aspects to the problem, aspects that lie within percolation theory. The percolation transition is beyond the scope of our book.

When a semiconductor is doped with increasing concentrations of donor (or acceptor) atoms, a transition will occur to a conducting metallic phase.

Experimental results for P atoms in silicon are shown in Fig. 10b. Here the insulator-metal transition takes place when the concentration is so high that the ground state wavefunctions of electrons on neighboring impurity atoms overlap significantly.

The observed value of the critical concentration in the Si : P alloy system is $n_c = 3.74 \times 10^{18}$ cm^{-3}, as in the figure. If we take 32×10^{-8} cm as the radius of the ground state of a donor in Si in the spherical approximation, then by the Mott criterion $a_c = 1.44 \times 10^{-6}$ cm. The P atoms are believed to occupy lattice sites at random, but if instead their lattice were simple cubic, the critical Mott concentration would be

$$n_c = 1/a_c^3 = 0.33 \times 10^{18}\ \text{cm}^{-3}\ , \tag{46}$$

appreciably less than the observed value. It is usual in the semiconductor literature to refer to a heavily-doped semiconductor in the metallic range as a **degenerate semiconductor**.

Screening and Phonons in Metals

An interesting application of our two limiting forms of the dielectric function is to longitudinal acoustic phonons in metals. For longitudinal modes the total dielectric function, ions plus electrons, must be zero, by (17). Provided the sound velocity is less than the Fermi velocity of the electrons, we may use for the electrons the Thomas-Fermi dielectric function

$$\epsilon_{el}(\omega,K) = 1 + k_s^2/K^2\ . \tag{47}$$

Provided also that the ions are well-spaced and move independently, we may use for them the plasmon $\epsilon(\omega,0)$ limit with the approximate mass M.

The total dielectric function, lattice plus electrons, but without the electronic polarizability of the ion cores, is

$$\epsilon(\omega,K) = 1 - \frac{4\pi ne^2}{M\omega^2} + \frac{k_s^2}{K^2}\ . \tag{48}$$

At low K and ω we neglect the term 1. At a zero of $\epsilon(\omega,K)$ we have, with $\epsilon_F \equiv \frac{1}{2}mv_F^2$,

$$\omega^2 \frac{4\pi ne^2}{Mk_s^2}K^2 = \frac{4\pi ne^2}{M} \cdot \frac{\epsilon_F}{6\pi ne^2}K^2 = \frac{m}{3M}v_F^2K^2\ , \tag{49}$$

or

$$\omega = vK\ ; \qquad v = (m/3M)^{1/2}v_F\ . \tag{50}$$

This describes long wavelength longitudinal acoustic phonons.

In the alkali metals the result is in quite good agreement with the observed longitudinal wave velocity. For potassium we calculate $v = 1.8 \times 10^5$ cm s^{-1}; the observed longitudinal sound velocity at 4 K in the [100] direction is 2.2×10^5 cm s^{-1}.

There is another zero of $\epsilon(\omega,K)$ for positive ions imbedded in an electron sea. For high frequencies we use the dielectric contribution $-\omega_p^2/\omega^2$ of the electron gas:

$$\epsilon(\omega,0) = 1 - \frac{4\pi ne^2}{M\omega^2} - \frac{4\pi ne^2}{m\omega^2} , \tag{51}$$

and this function has a zero when

$$\omega^2 = \frac{4\pi ne^2}{\mu} ; \quad \frac{1}{\mu} = \frac{1}{M} + \frac{1}{m} . \tag{52}$$

This is the electron plasma frequency (20), but with the reduced mass correction for the motion of the positive ions.

POLARITONS

Longitudinal optical phonons and transverse optical phonons were discussed in Chapter 4, but we deferred treatment of the interaction of transverse optical phonons with transverse electromagnetic waves. When the two waves are at resonance the phonon-photon coupling entirely changes the character of the propagation, and a forbidden band is established for reasons that have nothing to do with the periodicity of the lattice.

By resonance we mean a condition in which the frequencies and wavevectors of both waves are approximately equal. The region of the crossover of the two dashed curves in Fig. 11 is the resonance region; the two dashed curves are the dispersion relations for photons and transverse optical phonons in the absence of any coupling between them. In reality, however, there always is coupling implicit in Maxwell's equations and expressed by the dielectric function. The quantum of the coupled phonon-photon transverse wave field is called a **polariton**.

In this section we see how the coupling is responsible for the dispersion relations shown as solid curves in the figure. All takes place at very low values of the wavevector in comparison with a zone boundary, because at crossover $\omega(\text{photon}) = ck(\text{photon}) = \omega(\text{phonon}) \approx 10^{13}\,\text{s}^{-1}$; thus $k \approx 300\,\text{cm}^{-1}$.

An early warning: although the symbol ω_L will necessarily arise in the theory, the effects do not concern longitudinal optical phonons. Longitudinal phonons do not couple to transverse photons in the bulk of a crystal.

The coupling of the electric field E of the photon with the dielectric polarization P of the TO phonon is described by the electromagnetic wave equation:

(CGS) $$c^2K^2E = \omega^2(E + 4\pi P) . \tag{53}$$

At low wavevectors the TO phonon frequency ω_T is independent of K. The polarization is proportional to the displacement of the positive ions relative to the negative ions, so that the equation of motion of the polarization is like that of an oscillator and may be written as, with $P = Nqu$,

$$-\omega^2P + \omega_T^2P = (Nq^2/M)E , \tag{54}$$

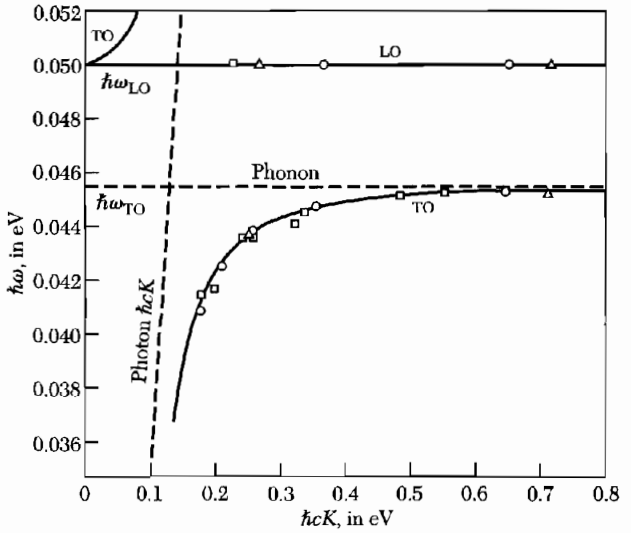


Figure 11 A plot of the observed energies and wavevectors of the polaritons and of the LO phonons in GaP. The theoretical dispersion curves are shown by the solid lines. The dispersion curves for the uncoupled phonons and photons are shown by the short, dashed lines. (After C. H. Henry and J. J. Hopfield.)

where there are N ion pairs of effective charge q and reduced mass M, per unit volume. For simplicity we neglect the electronic contribution to the polarization.

The equations (53) and (54) have a solution when

$$\begin{vmatrix} \omega^2 - c^2K^2 & 4\pi\omega^2 \\ Nq^2/M & \omega^2 - \omega_T^2 \end{vmatrix} = 0 \ . \tag{55}$$

This gives the polariton dispersion relation, similar to that plotted in Figs. 11 and 12. At $K = 0$ there are two roots, $\omega = 0$ for the photon and

$$\omega^2 = \omega_T^2 + 4\pi Nq^2/M \tag{56}$$

for the polariton. Here ω_T is the TO phonon frequency in the absence of coupling with photons.

The dielectric function obtained from (54) is:

$$\epsilon(\omega) = 1 + 4\pi P/E = 1 + \frac{4\pi Nq^2/M}{\omega_T^2 - \omega^2} \ . \tag{57}$$

If there is an optical electronic contribution to the polarization from the ion cores, this should be included. In the frequency range from zero up through the infrared, we write

$$\epsilon(\omega) = \epsilon(\infty) + \frac{4\pi Nq^2/M}{\omega_T^2 - \omega^2} \tag{58}$$

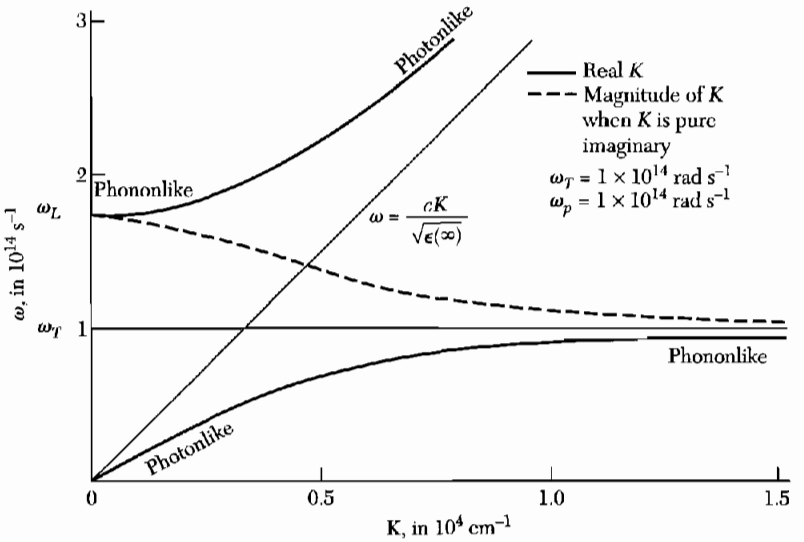


Figure 12 Coupled modes of photons and transverse optical phonons in an ionic crystal. The fine horizontal line represents oscillators of frequency ω_T in the absence of coupling to the electromagnetic field, and the fine line labeled $\omega = cK/\sqrt{\epsilon(\infty)}$ corresponds to electromagnetic waves in the crystal, but uncoupled to the lattice oscillators ω_T. The heavy lines are the dispersion relations in the presence of coupling between the lattice oscillators and the electromagnetic wave. One effect of the coupling is to create the frequency gap between ω_L and ω_T: within this gap the wavevector is pure imaginary of magnitude given by the broken line in the figure. In the gap the wave attenuates as $\exp(-|K|x)$, and we see from the plot that the attenuation is much stronger near ω_T than near ω_L. The character of the branches varies with K; there is a region of mixed electric-mechanical aspects near the nominal crossover. Note, finally, it is intuitively obvious that the group velocity of light in the medium is always $<c$, because the slope $\partial\omega/\partial K$ for the actual dispersion relations (heavy lines) is everywhere less than the slope c for the uncoupled photon in free space.

in accord with the definition of $\epsilon(\infty)$ as the optical dielectric constant, obtained as the square of the optical refractive index.

We set $\omega = 0$ to obtain the static dielectric function:

$$\epsilon(0) = \epsilon(\infty) + 4\pi Nq^2/M\omega_T^2 \ , \tag{59}$$

which is combined with (58) to obtain $\epsilon(\omega)$ in terms of accessible parameters:

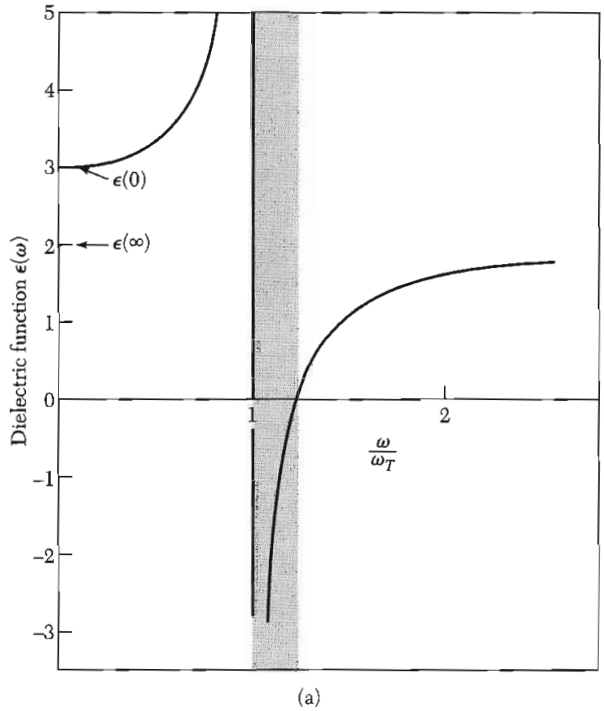
$$\epsilon(\omega) = \epsilon(\infty) + [\epsilon(0) - \epsilon(\infty)]\frac{\omega_T^2}{\omega_T^2 - \omega^2}$$

or

$$\epsilon(\omega) = \frac{\omega_T^2\epsilon(0) - \omega^2\epsilon(\infty)}{\omega_T^2 - \omega^2} = \epsilon(\infty)\left(\frac{\omega_L^2 - \omega^2}{\omega_T^2 - \omega^2}\right) \ . \tag{60}$$

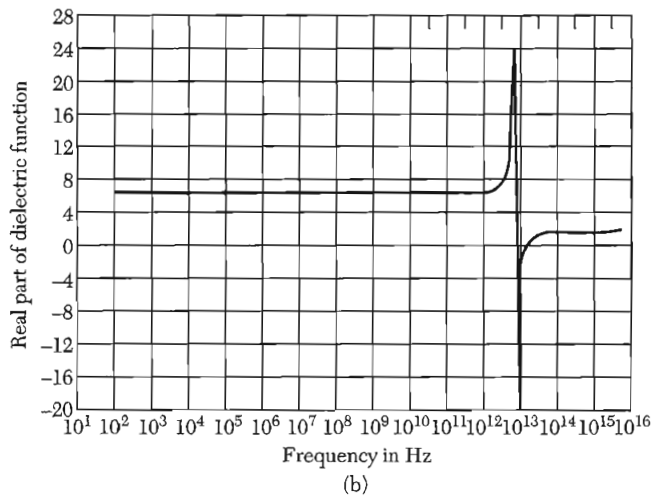
The zero of $\epsilon(\omega)$ defines the longitudinal optical phonon frequency ω_L, as the pole of $\epsilon(\omega)$ defines ω_T. The zero gives

$$\epsilon(\infty)\omega_L^2 = \epsilon(0)\omega_T^2 \ . \tag{61}$$



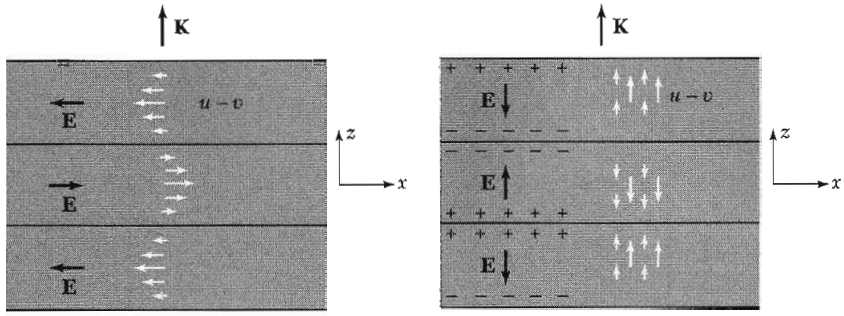
(a)

Figure 13a Plot of $\epsilon(\omega)$ from (60) for $\epsilon(\infty) = 2$ and $\epsilon(0) = 3$. The dielectric constant is negative between $\omega = \omega_T$ and $\omega_L = (3/2)^{1/2}\omega_T$; that is, between the pole (infinity) of $\epsilon(\omega)$ and the zero of $\epsilon(\omega)$. Incident electromagnetic waves of frequencies in the shaded regions $\omega_T < \omega < \omega_L$ will not propagate in the medium, but will be reflected at the boundary.



(b)

Figure 13b Dielectric function (real part) of SrF_2 measured over a wide frequency range, exhibiting the decrease of the ionic polarizability at high frequencies. (A. von Hippel.)



TO phonon LO phonon

Figure 14 Relative displacements of the positive and negative ions at one instant of time for a wave in an optical mode traveling along the z axis. The planes of nodes (zero displacement) are shown; for long wavelength phonons the nodal planes are separated by many planes of atoms. In the transverse optical phonon mode the particle displacement is perpendicular to the wavevector $\mathbf{K}$; the macroscopic electric field in an infinite medium will lie only in the $\pm x$ direction for the mode shown, and by the symmetry of the problem $\partial E_x/\partial x = 0$. It follows that div $\mathbf{E} = 0$ for a TO phonon. In the longitudinal optical phonon mode the particle displacements and hence the dielectric polarization $\mathbf{P}$ are parallel to the wavevector. The macroscopic electric field $\mathbf{E}$ satisfies $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P} = 0$ in CGS or $\epsilon_0\mathbf{E} + \mathbf{P} = 0$ in SI; by symmetry $\mathbf{E}$ and $\mathbf{P}$ are parallel to the z axis, and $\partial E_z/\partial z \neq 0$. Thus div $\mathbf{E} \neq 0$ for an LO phonon, and $\epsilon(\omega)$ div $\mathbf{E}$ is zero only if $\epsilon(\omega) = 0$.

Waves do not propagate in the frequency region for which $\epsilon(\omega)$ is negative, between its pole at $\omega = \omega_T$ and its zero at $\omega = \omega_L$, as in Fig. 13. For negative ϵ, waves do not propagate because then K is imaginary for real ω, and $\exp(iKx) \rightarrow \exp(-|K|x)$, damped in space. The zero of $\epsilon(\omega)$, by our earlier argument, is the LO frequency at low K, Fig. 14. Just as with the plasma frequency ω_p, the frequency ω_L has two meanings, one as the LO frequency at low K and the other as the upper cutoff frequency of the forbidden band for propagation of an electromagnetic wave. The value of ω_L is identical at both frequencies.

LST Relation

We write (61) as

$$\boxed{\frac{\omega_L^2}{\omega_T^2} = \frac{\epsilon(0)}{\epsilon(\infty)}}, \tag{62}$$

where $\epsilon(0)$ is the static dielectric constant and $\epsilon(\infty)$ is the high-frequency limit of the dielectric function, defined to include the core electron contribution. This result is the Lyddane-Sachs-Teller relation. The derivation assumed a cubic crystal with two atoms per primitive cell. For soft modes with $\omega_T \rightarrow 0$ we see that $\epsilon(0) \rightarrow \infty$, a characteristic of ferroelectricity.

Undamped electromagnetic waves with frequencies within the gap cannot propagate in a thick crystal. The reflectivity of a crystal surface is expected to be high in this frequency region, as in Fig. 15.

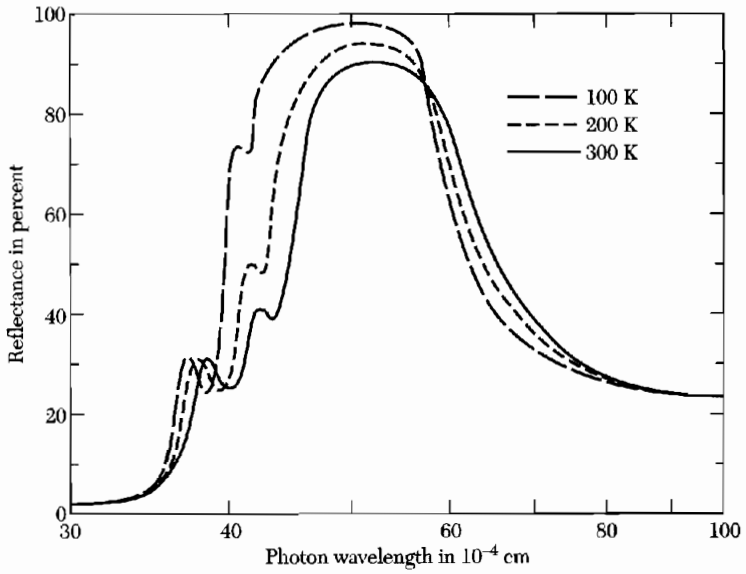


Figure 15 Reflectance of a crystal of NaCl at several temperatures, versus wavelength. The nominal values of ω_L and ω_T at room temperature correspond to wavelengths of 38 and 61 $\times$ 10^{-4} cm, respectively. (After A. Mitsuishi et al.)

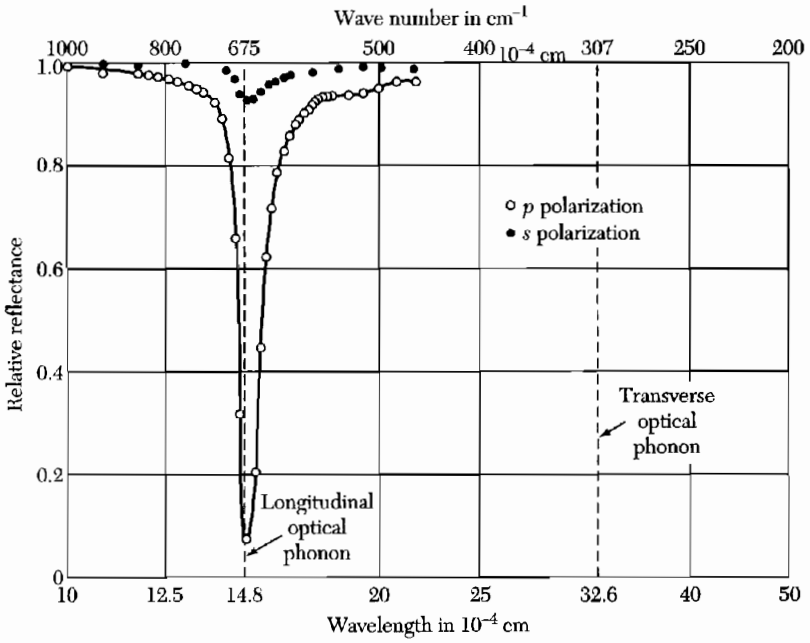


Figure 16 Reflectance versus wavelength of a LiF film backed by silver, for radiation incident near 30°. The longitudinal optical phonon absorbs strongly the radiation polarized (p) in the plane normal to the film, but absorbs hardly at all the radiation polarized (s) parallel to the film. (After D. W. Berreman.)

For films of thickness less than a wavelength the situation is changed. Because for frequencies in the gap the wave attenuates as $\exp(-|K|x)$, it is possible for the radiation to be transmitted through a film for the small values of $|K|$ near ω_L, but for the large values of $|K|$ near ω_T the wave will be reflected. By reflection at nonnormal incidence the frequency ω_L of longitudinal optical phonons can be observed, as in Fig. 16.

Experimental values of $\epsilon(0)$, $\epsilon(\infty)$, and ω_T are given in Table 3, with values of ω_L calculated using the LST relation, Eq. (62). We compare values of

Table 3 Lattice parameters, chiefly at 300 K

Crystal	Static dielectric constant $\epsilon(0)$	Optical dielectric constant $\epsilon(\infty)$	ω_T, in $10^{13}\,s^{-1}$ experimental	ω_L, in $10^{13}\,s^{-1}$ LST relation
LiH	12.9	3.6	11.	21.
LiF	8.9	1.9	5.8	12.
LiCl	12.0	2.7	3.6	7.5
LiBr	13.2	3.2	3.0	6.1
NaF	5.1	1.7	4.5	7.8
NaCl	5.9	2.25	3.1	5.0
NaBr	6.4	2.6	2.5	3.9
KF	5.5	1.5	3.6	6.1
KCl	4.85	2.1	2.7	4.0
KI	5.1	2.7	1.9	2.6
RbF	6.5	1.9	2.9	5.4
RbI	5.5	2.6	1.4	1.9
CsCl	7.2	2.6	1.9	3.1
CsI	5.65	3.0	1.2	1.6
TlCl	31.9	5.1	1.2	3.0
TlBr	29.8	5.4	0.81	1.9
AgCl	12.3	4.0	1.9	3.4
AgBr	13.1	4.6	1.5	2.5
MgO	9.8	2.95	7.5	14.
GaP	10.7	8.5	6.9	7.6
GaAs	13.13	10.9	5.1	5.5
GaSb	15.69	14.4	4.3	4.6
InP	12.37	9.6	5.7	6.5
InAs	14.55	12.3	4.1	4.5
InSb	17.88	15.6	3.5	3.7
SiC	9.6	6.7	14.9	17.9
C	5.5	5.5	25.1	25.1
Si	11.7	11.7	9.9	9.9
Ge	15.8	15.8	5.7	5.7

ω_L/ω_T obtained by inelastic neutron scattering with experimental values of $[\epsilon(0)/\epsilon(\infty)]^{1/2}$ obtained by dielectric measurements:

	NaI	KBr	GaAs
ω_L/ω_T	1.44 ± 0.05	1.39 ± 0.02	1.07 ± 0.02
$[\epsilon(0)/\epsilon(\infty)]^{1/2}$	1.45 ± 0.03	1.38 ± 0.03	1.08

The agreement with the LST relation is excellent.

ELECTRON-ELECTRON INTERACTION

Fermi Liquid

Because of the interaction of the conduction electrons with each other through their electrostatic interaction, the electrons suffer collisions. Further, a moving electron causes an inertial reaction in the surrounding electron gas, thereby increasing the effective mass of the electron. The effects of electron-electron interactions are usually described within the framework of the Landau theory of a Fermi liquid. The object of the theory is to give a unified account of the effect of interactions. A Fermi gas is a system of noninteracting fermions; the same system with interactions is a Fermi liquid.

Landau's theory gives a good account of the low-lying single particle excitations of the system of interacting electrons. These single particle excitations are called **quasiparticles**; they have a one-to-one correspondence with the single particle excitations of the free-electron gas. A quasiparticle may be thought of as a single particle accompanied by a distortion cloud in the electron gas. One effect of the coulomb interactions between electrons is to change the effective mass of the electron; in the alkali metals the increase is roughly of the order of 25 percent.

Electron-Electron Collisions. It is an astonishing property of metals that conduction electrons, although crowded together only 2 Å apart, travel long distances between collisions with each other. The mean free paths for electron-electron collisions are longer than 10^4 Å at room temperature and longer than 10 cm at 1 K.

Two factors are responsible for these long mean free paths, without which the free-electron model of metals would have little value. The most powerful factor is the exclusion principle (Fig. 17), and the second factor is the screening of the coulomb interaction between two electrons.

We show how the exclusion principle reduces the collision frequency of an electron that has a low excitation energy ϵ_1 outside a filled Fermi sphere (Fig. 18). We estimate the effect of the exclusion principle on the two-body collision $\mathbf{1} + \mathbf{2} \rightarrow \mathbf{3} + \mathbf{4}$ between an electron in the excited orbital **1** and an electron in the filled orbital **2** in the Fermi sea. It is convenient to refer all energies to the Fermi level μ taken as the zero of energy; thus ϵ_1 will be

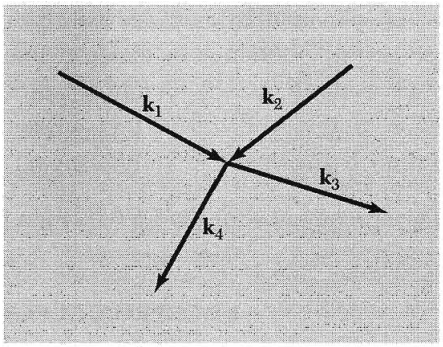


Figure 17 A collision between two electrons of wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$. After the collision the particles have wavevectors $\mathbf{k}_3$ and $\mathbf{k}_4$. The Pauli exclusion principle allows collisions only to final states $\mathbf{k}_3$, $\mathbf{k}_4$ which were vacant before the collision.

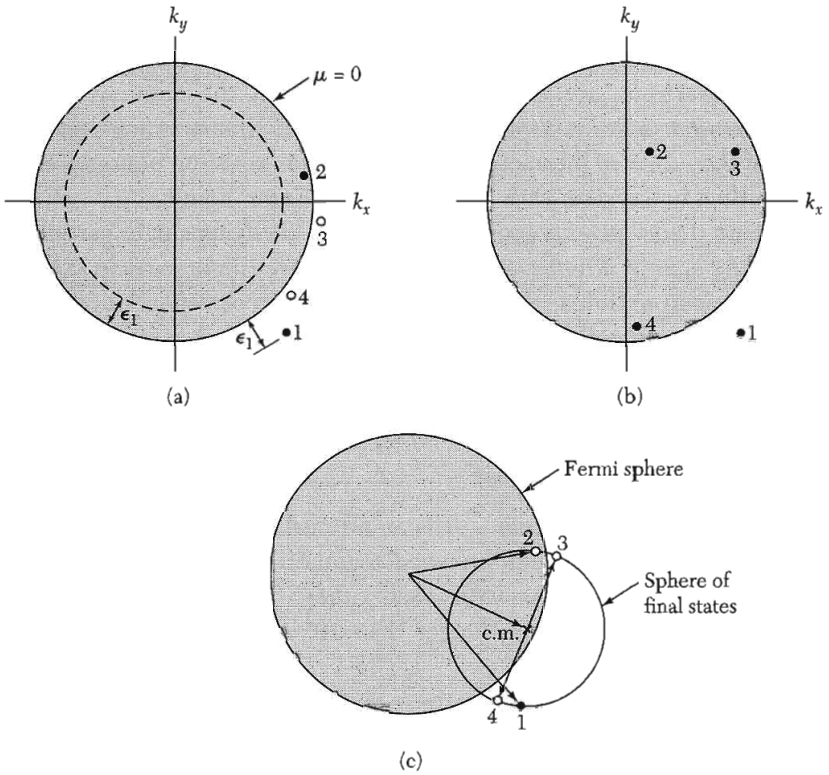


Figure 18 In (a) the electrons in initial orbitals 1 and 2 collide. If the orbitals 3 and 4 are initially vacant, the electrons 1 and 2 can occupy orbitals 3 and 4 after the collision. Energy and momentum are conserved. In (b) the electrons in initial orbitals 1 and 2 have no vacant final orbitals available that allow energy to be conserved in the collision. Orbitals such as 3 and 4 would conserve energy and momentum, but they are already filled with other electrons. In (c) we have denoted with $\times$ the wavevector of the center of mass of **1** and **2**. All pairs of orbitals 3 and 4 conserve momentum and energy if they lie at opposite ends of a diameter of the small sphere. The small sphere was drawn from the center of mass to pass through 1 and 2. But not all pairs of points 3, 4 are allowed by the exclusion principle, for both 3, 4 must lie outside the Fermi sphere; the fraction allowed is $\approx \epsilon_1/\epsilon_F$.

positive and ϵ_2 will be negative. Because of the exclusion principle the orbitals **3** and **4** of the electrons after collision must lie outside the Fermi sphere, all orbitals within the sphere being already occupied; thus both energies ϵ_3, ϵ_4 must be positive referred to zero on the Fermi sphere.

The conservation of energy requires that $|\epsilon_2| < \epsilon_1$, for otherwise $\epsilon_3 + \epsilon_4 = \epsilon_1 + \epsilon_2$ could not be positive. This means that collisions are possible only if the orbital **2** lies within a shell of thickness ϵ_1 within the Fermi surface, as in Fig. 18a. Thus the fraction $\approx \epsilon_1/\epsilon_F$ of the electrons in filled orbitals provides a suitable target for electron **1**. But even if the target electron **2** is in the suitable energy shell, only a small fraction of the final orbitals compatible with conservation of energy and momentum are allowed by the exclusion principle. This gives a second factor of ϵ_1/ϵ_F.

In Fig. 18c we show a small sphere on which all pairs of orbitals **3**, **4** at opposite ends of a diameter satisfy the conservation laws, but collisions can occur only if both orbitals **3**, **4** lie outside the Fermi sea. The product of the two fractions is $(\epsilon_1/\epsilon_F)^2$. If ϵ_1 corresponds to 1 K and ϵ_F to 5×10^4 K, we have $(\epsilon_1/\epsilon_F)^2 \approx 4 \times 10^{-10}$, the factor by which the exclusion principle reduces the collision rate.

The argument is not changed for a thermal distribution of electrons at a low temperature such that $k_BT \ll \epsilon_F$. We replace ϵ_1 by the thermal energy $\approx k_BT$, and now the rate at which electron-electron collisions take place is reduced below the classical value by $(k_BT/\epsilon_F)^2$, so that the effective collision cross section σ is

$$\sigma \approx (k_BT/\epsilon_F)^2\sigma_0 \ , \tag{63}$$

where σ_0 is the cross section for the electron-electron interaction.

The interaction of one electron with another has a range of the order of the screening length $1/k_s$ as in (34). Numerical calculations give the effective cross section with screening for collisions between electrons as of the order of 10^{-15} cm^2 or 10 Å^2 in typical metals. The effect of the electron-gas background in electron-electron collisions is to reduce σ_0 below the value expected from the Rutherford scattering equation for the unscreened coulomb potential. However, much the greater reduction in the cross section is caused by the Pauli factor $(k_BT/\epsilon_F)^2$.

At room temperature in a typical metal k_BT/ϵ_F is $\sim 10^{-2}$, so that $\sigma \sim 10^{-4}\sigma_0 \sim 10^{-19}$ cm^2. The mean free path for electron-electron collisions is $\ell \approx 1/n\sigma \sim 10^{-4}$ cm at room temperature. This is longer than the mean free path due to electron-phonon collisions by at least a factor of 10, so that at room temperature collisions with phonons are likely to be dominant. At liquid helium temperatures a contribution proportional to T^2 has been found in the resistivity of a number of metals, consistent with the form of the electron-electron scattering cross section (63). The mean free path of electrons in indium at 2 K is of the order of 30 cm, as expected from (63). Thus the Pauli

principle explains one of the central questions of the theory of metals: how do the electrons travel long distances without colliding with each other?

ELECTRON-PHONON INTERACTION: POLARONS

The most common effect of the electron-phonon interaction is seen in the temperature dependence of the electrical resistivity, which for pure copper is 1.55 microhm-cm at 0°C and 2.28 microhm-cm at 100°C. The electrons are scattered by the phonons, and the higher the temperature, the more phonons there are and hence more scattering. Above the Debye temperature the number of thermal phonons is roughly proportional to the absolute temperature, and we find that the resistivity increases approximately as the absolute temperature in any reasonably pure metal in this temperature region.

A more subtle effect of the electron-phonon interaction is the apparent increase in electron mass that occurs because the electron drags the heavy ion cores along with it. In an insulator the combination of the electron and its strain field is known as a **polaron**, Fig. 19. The effect is large in ionic crystals because of the strong coulomb interaction between ions and electrons. In covalent crystals the effect is weak because neutral atoms have only a weak interaction with electrons.

The strength of the electron-lattice interaction is measured by the dimensionless coupling constant α given by

$$\frac{1}{2}\alpha = \frac{\text{deformation energy}}{\hbar\omega_L} , \tag{64}$$

where ω_L is the longitudinal optical phonon frequency near zero wavevector. We view $\frac{1}{2}\alpha$ as "the number of phonons which surround a slow-moving electron in a crystal."

Values of α deduced from diverse experiments and theory are given in Table 4, after F. C. Brown. The values of α are high in ionic crystals and low in covalent crystals. The values of the effective mass m^*_{pol} of the polaron are from cyclotron resonance experiments. The values given for the band effective mass m^* were calculated from m^*_{pol}. The last row in the table gives the factor m^*_{pol}/m^* by which the band mass is increased by the deformation of the lattice.

Theory relates the effective mass of the polaron m^*_{pol} to the effective band mass m^* of the electron in the undeformed lattice by the relation

$$m^*_{\text{pol}} \cong m^* \left(\frac{1 - 0.0008\alpha^2}{1 - \frac{1}{6}\alpha + 0.0034\alpha^2} \right) ; \tag{65}$$

for $\alpha \ll 1$ this is approximately $m^*(1 + \frac{1}{6}\alpha)$. Because the coupling constant α is always positive, the polaron mass is greater than the bare mass, as we expect from the inertia of the ions.

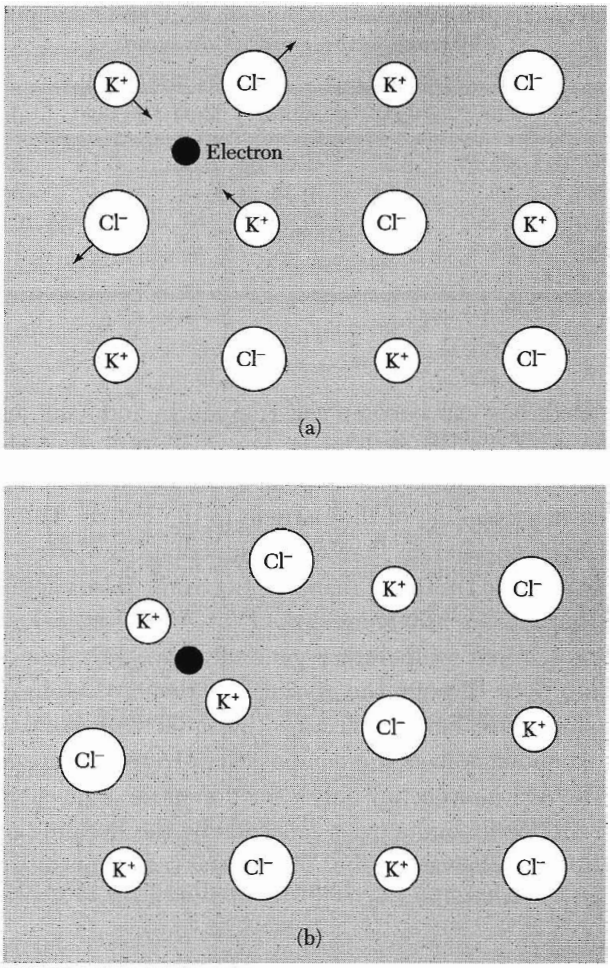


Figure 19 The formation of a polaron. (a) A conduction electron is shown in a rigid lattice of an ionic crystal, KCl. The forces on the ions adjacent to the electron are shown. (b) The electron is shown in an elastic or deformable lattice. The electron plus the associated strain field is called a polaron. The displacement of the ions increases the effective inertia and, hence, the effective mass of the electron; in KCl the mass is increased by a factor of 2.5 with respect to the band theory mass in a rigid lattice. In extreme situations, often with holes, the particle can become self-trapped (localized) in the lattice. In covalent crystals the forces on the atoms from the electron are weaker than in ionic crystals, so that polaron deformations are small in covalent crystals.

It is common to speak of large and small polarons. The electron associated with a large polaron moves in a band, but the mass is slightly enhanced; these are the polarons we have discussed above. The electron associated with a small polaron spends most of its time trapped on a single ion. At high temperatures the electron moves from site to site by thermally activated hopping; at low temperatures the electron tunnels slowly through the crystal, as if in a band of large effective mass.

Table 4 Polaron coupling constants α, masses m^*_{pol}, and band masses m^* for electrons in the conduction band

Crystal	KCl	KBr	AgCl	AgBr	ZnO	PbS	InSb	CaAs
α	3.97	3.52	2.00	1.69	0.85	0.16	0.014	0.06
m^*_{pol}/m	1.25	0.93	0.51	0.33	—	—	0.014	—
m^*/m	0.50	0.43	0.35	0.24	—	—	0.014	—
m^*_{pol}/m^*	2.5	2.2	1.5	1.4	—	—	1.0	—

Holes or electrons can become self-trapped by inducing an asymmetric local deformation of the lattice. This is most likely to occur when the band edge is degenerate and the crystal is polar (such as an alkali halide or silver halide), with strong coupling of the particle to the lattice. The valence band edge is more often degenerate than the conduction band edge, so that holes are more likely to be self-trapped than are electrons. Holes appear to be self-trapped in all the alkali and silver halides.

Ionic solids at room temperature generally have very low conductivities for the motion of ions through the crystal, less than 10^{-6} (ohm-cm)$^{-1}$, but a family of compounds has been reported with conductivities of 0.2 (ohm-cm)$^{-1}$ at 20°C. The compounds have the composition MAg_4I_5, where M denotes K, Rb, or NH_4. The Ag^+ ions occupy only a fraction of the equivalent lattice sites available, and the ionic conductivity proceeds by the hopping of a silver ion from one site to a nearby vacant site. The crystal structures also have parallel open channels.

PEIERLS INSTABILITY OF LINEAR METALS

Consider a one-dimensional metal with an electron gas filling all conduction band orbitals out to the wavevector k_F, at absolute zero of temperature. Peierls suggested that such a linear metal is unstable with respect to a static lattice deformation of wavevector $G = 2k_F$. Such a deformation creates an energy gap at the Fermi surface, thereby lowering the energy of electrons below the energy gap, Fig. 20. The deformation proceeds until limited by the increase of elastic energy; the equilibrium deformation Δ is given by the root of

$$\frac{d}{d\Delta}(E_{\text{electronic}} + E_{\text{elastic}}) = 0 \ . \tag{66}$$

Consider the elastic strain $\Delta \cos 2k_F x$. The spatial-average elastic energy per unit length is $E_{\text{elastic}} = \frac{1}{2}C\Delta^2\langle\cos^2 2k_F x\rangle = \frac{1}{4}C\Delta^2$, where C is the force constant of the linear metal. We next calculate $E_{\text{electronic}}$. Suppose that the ion

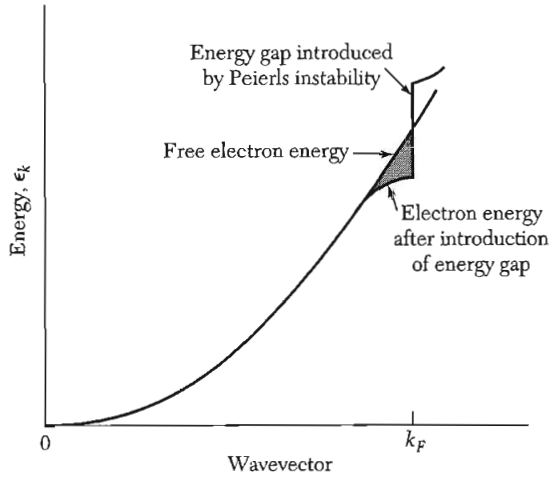


Figure 20 Peierls instability. Electrons with wavevectors near the Fermi surface have their energy lowered by a lattice deformation.

contribution to the lattice potential seen by a conduction electron is proportional to the deformation: $U(x) = 2A\Delta \cos 2k_F x$. From (7.51) we have

$$\epsilon_K = (\hbar^2/2m)(k_F^2 + K^2) \pm [4(\hbar^2 k_F^2/2m)(\hbar^2 K^2/2m) + A^2\Delta^2]^{1/2} \ . \tag{67}$$

It is convenient to define

$$x_K \equiv \hbar^2 K^2/m \ ; \quad x_F \equiv \hbar^2 k_F^2/m \ ; \quad x \equiv \hbar^2 K k_F/m \ .$$

We retain the $-$ sign in (67) and form

$$\frac{d\epsilon_K}{d\Delta} = \frac{-A^2\Delta}{(x_F x_K + A^2\Delta^2)^{1/2}} \ ,$$

whence, with dK/π as the number of orbitals per unit length,

$$\frac{dE_{\text{electronic}}}{d\Delta} = \frac{2}{\pi}\int_0^{k_F} dK \frac{d\epsilon_K}{d\Delta} = -(2A^2\Delta/\pi)\int_0^{k_F} \frac{dK}{(x_F x_K + A^2\Delta^2)^{1/2}}$$

$$= -(2A^2\Delta/\pi)(k_F/x_F)\int_0^{x_F} \frac{dx}{(x^2 + A^2\Delta^2)^{1/2}} = -(2A^2\Delta/\pi)(k_F/x_F)\sinh^{-1}(x_F/A\Delta) \ .$$

We put it all together. The equilibrium deformation is the root of

$$\tfrac{1}{2}C\Delta - (2A^2 m\Delta/\pi\hbar^2 k_F)\sinh^{-1}(\hbar^2 k_F^2/mA\Delta) = 0 \ .$$

The root Δ that corresponds to the minimum energy is given by

$$\hbar^2 k_F^2/mA\Delta = \sinh(-\hbar^2 k_F \pi C/4mA^2) \ , \tag{68}$$

whence

$$|A|\Delta \simeq (2\hbar^2 k_F^2/m)\exp(-\hbar^2 k_F \pi C/4mA^2) \ , \tag{69}$$

if the argument of the sinh in (68) is $\gg 1$. We assume $k_F \leqslant \frac{1}{2}k_{\max}$.

The result is of the form of the energy gap equation in the BCS theory of superconductivity, Chapter 10. The deformation Δ is a collective effect of all the electrons. If we set $W = \hbar^2 k_F^2/2m$ = conduction band width; $N(0) = 2m/\pi\hbar^2 k_F$ = density of orbitals at Fermi level; $V = 2A^2/C$ = effective electron-electron interaction energy, then we can write (69) as

$$|A|\Delta \simeq 4W \exp[-1/N(0)V] \ , \tag{70}$$

which is analogous to the BCS energy gap equation. An example of a Peierls insulator is TaS_3.

SUMMARY

(In CGS units)

- The dielectric function may be defined as

$$\epsilon(\omega,\mathbf{K}) = \frac{\rho_{\text{ext}}(\omega,\mathbf{K})}{\rho_{\text{ext}}(\omega,\mathbf{K}) + \rho_{\text{ind}}(\omega,\mathbf{K})} \ ,$$

  in terms of the applied and induced charge density components at $\omega,\mathbf{K}$.

- The plasma frequency $\bar{\omega}_p = [4\pi ne^2/\epsilon(\infty)m]^{1/2}$ is the frequency of the uniform collective longitudinal oscillation of the electron gas against a background of fixed positive ions. It is also the low frequency cutoff for propagation of transverse electromagnetic waves in the plasma.
- The poles of the dielectric function define ω_T and the zeroes define ω_L.
- In a plasma the coulomb interaction is screened; it becomes $(q/r)\exp(-k_s r)$, where the screening length $1/k_s = (\epsilon_F/6\pi n_0 e^2)^{1/2}$.
- A metal-insulator transition may occur when the nearest-neighbor separation a becomes of the order of $4a_0$, where a_0 is the radius of the first Bohr orbit in the insulator. The metallic phase exists at smaller values of a.
- A polariton is a quantum of the coupled TO phonon-photon fields. The coupling is assured by the Maxwell equations. The spectral region $\omega_T < \omega < \omega_L$ is forbidden to electromagnetic wave propagation.
- The Lyddane-Sachs-Teller relation is $\omega_L^2/\omega_T^2 = \epsilon(0)/\epsilon(\infty)$.

Problems

1. ***Surface plasmons.*** Consider a semi-infinite plasma on the positive side of the plane $z = 0$. A solution of Laplace's equation $\nabla^2\varphi = 0$ in the plasma is $\varphi_i(x,z) = A \cos kx\, e^{-kz}$, whence $E_{zi} = kA \cos kx\, e^{-kz}$; $E_{xi} = kA \sin kx\, e^{-kz}$. (a)Show that in the vacuum $\varphi_0(x,z) = A \cos kx\, e^{kz}$ for $z < 0$ satisfies the boundary condition that the tangential component of $\mathbf{E}$ be continuous at the boundary; that is, find E_{x0}. (b) Note that $\mathbf{D}_i = \epsilon(\omega)\mathbf{E}_i$; $\mathbf{D}_o = \mathbf{E}_o$. Show that the boundary condition that the

normal component of $\mathbf{D}$ be continuous at the boundary requires that $\epsilon(\omega) = -1$, whence from (10) we have the Stern-Ferrell result:

$$\omega_s^2 = \tfrac{1}{2}\omega_p^2 \tag{71}$$

for the frequency ω_s of a surface plasma oscillation.

2. ***Interface plasmons.*** We consider the plane interface $z = 0$ between metal 1 at $z > 0$ and metal 2 at $z < 0$. Metal 1 has bulk plasmon frequency ω_{p1}; metal 2 has ω_{p2}. The dielectric constants in both metals are those of free-electron gases. Show that surface plasmons associated with the interface have the frequency

$$\omega = [\tfrac{1}{2}(\omega_{p1}^2 + \omega_{p2}^2)]^{1/2} .$$

3. ***Alfvén waves.*** Consider a solid with an equal concentration n of electrons of mass m_e and holes of mass m_h. This situation may arise in a semimetal or in a compensated semiconductor. Place the solid in a uniform magnetic field $\mathbf{B} = B\hat{\mathbf{z}}$. Introduce the coordinate $\xi = x + iy$ appropriate for circularly polarized motion, with ξ having time dependence $e^{-i\omega t}$. Let $\omega_e = eB/m_e c$ and $\omega_h = eB/m_h c$. (a) In CGS units, show that $\xi_e = eE^+/m_e\omega(\omega + \omega_e)$; $\xi_h = -eE^+/m_h\omega(\omega - \omega_h)$ are the displacements of the electrons and holes in the electric field $E^- e^{-i\omega t} = (E_x + iE_y) e^{-i\omega t}$. (b) Show that the dielectric polarization $P^+ = ne(\xi_h - \xi_e)$ in the regime $\omega \ll \omega_e, \omega_h$ may be written as $P^+ = nc^2(m_h + m_e)E^+/B^2$, and the dielectric function $\epsilon(\omega) = \epsilon_l + 4\pi P^+/E^+ = \epsilon_l + 4\pi c^2\rho/B^2$, where ϵ_l is the dielectric constant of the host lattice and $\rho = n(m_e + m_h)$ is the mass density of the carriers. If ϵ_l may be neglected, the dispersion relation $\omega^2\epsilon(\omega) = c^2K^2$ becomes, for electromagnetic waves propagating in the z direction, $\omega^2 = (B^2/4\pi\rho)K^2$. Such waves are known as Alfvén waves; they propagate with the constant velocity $B/(4\pi\rho)^{1/2}$. If $B = 10$ kG; $n = 10^{18}$ cm^{-3}; $m = 10^{-27}$ g, the velocity is $\sim 10^8$ cm s^{-1}. Alfvén waves have been observed in semimetals and in electron-hole drops in germanium (Chapter 15).

4. ***Helicon waves.*** (a) Employ the method of Problem 3 to treat a specimen with only one carrier type, say holes in concentration p, and in the limit $\omega \ll \omega_h = eB/m_h c$. Show that $\epsilon(\omega) = 4\pi pe^2/m_h\omega\omega_h$, where $D^+(\omega) = \epsilon(\omega)E^+(\omega)$. The term ϵ_l in ϵ has been neglected. (b) Show further that the dispersion relation becomes $\omega = (Bc/4\pi pe)K^2$, the helicon dispersion relation, in CGS. For $K = 1$ cm^{-1} and $B = 1000$ G, estimate the helicon frequency in sodium metal. (The frequency is negative; with circular-polarized modes the sign of the frequency refers to the sense of the rotation.)

5. ***Plasmon mode of a sphere.*** The frequency of the uniform plasmon mode of a sphere is determined by the depolarization field $\mathbf{E} = -4\pi\mathbf{P}/3$ of a sphere, where the polarization $\mathbf{P} = -ne\mathbf{r}$, with $\mathbf{r}$ as the average displacement of the electrons of concentration n. Show from $\mathbf{F} = m\boldsymbol{a}$ that the resonance frequency of the electron gas is $\omega_0^2 = 4\pi ne^2/3m$. Because all electrons participate in the oscillation, such an excitation is called a collective excitation or collective mode of the electron gas.

6. ***Magnetoplasma frequency.*** Use the method of Problem 5 to find the frequency of the uniform plasmon mode of a sphere placed in a constant uniform magnetic field $\mathbf{B}$. Let $\mathbf{B}$ be along the z axis. The solution should go to the cyclotron

frequency $\omega_c = eB/mc$ in one limit and to $\omega_0 = (4\pi ne^2/3m)^{1/2}$ in another limit. Take the motion in the xy plane.

7. ***Photon branch at low wavevector.*** (a) Find what (56) becomes when $\epsilon(\infty)$ is taken into account. (b) Show that there is a solution of (55) which at low wavevector is $\omega = cK/\sqrt{\epsilon(0)}$, as expected for a photon in a crystal of refractive index $n^2 = \epsilon$.

8. ***Plasma frequency and electrical conductivity.*** An organic conductor has been found by optical studies to have $\omega_p = 1.80 \times 10^{15}\ s^{-1}$ for the plasma frequency, and $\tau = 2.83 \times 10^{-15}\ s$ for the electron relaxation time at room temperature. (a) Calculate the electrical conductivity from these data. The carrier mass is not known and is not needed here. Take $\epsilon(\infty) = 1$. Convert the result to units $(\Omega\ \text{cm})^{-1}$. (b) From the crystal and chemical structure, the conduction electron concentration is $4.7 \times 10^{21}\ \text{cm}^{-3}$. Calculate the electron effective mass m^*.

9. ***Bulk modulus of the Fermi gas.*** Show that the contribution of the kinetic energy to the bulk modulus of the electron gas at absolute zero is $B = \frac{1}{3}nmv_F^2$. It is convenient to use (6.60). We can use our result for B to find the velocity of sound, which in a compressible fluid is $v = (B/\rho)^{1/2}$, where $v = (m/3M)^{1/2}v_F$, in agreement with (46). These estimates neglect attractive interactions.

10. ***Response of electron gas.*** It is sometimes stated erroneously in books on electromagnetism that the static conductivity σ, which in gaussian units has the dimensions of a frequency, measures the response frequency of a metal to an electric field suddenly applied. Criticize this statement as it might apply to copper at room temperature. The resistivity is ~1μohm-cm; the electron concentration is $8 \times 10^{22}\ \text{cm}^{-3}$; the mean free path is ~400 Å; the Fermi velocity is $1.6 \times 10^8\ \text{cm}\ s^{-1}$. You will not necessarily need all these data. Give the order of magnitude of the three frequencies σ, ω_p, and $1/\tau$ that might be relevant in the problem. Set up and solve the problem of the response $x(t)$ of the system to an electric field $E(t < 0) = 0$; $E(t > 0) = 1$. The system is a sheet of copper; the field is applied normal to the sheet. Include the damping. Solve the differential equation by elementary methods.

*11. ***Gap plasmons and the van der Waals interaction.*** Consider two semi-infinite media with plane surfaces $z = 0, d$. The dielectric function of the identical media is $\epsilon(\omega)$. Show that for surface plasmons symmetrical with respect to the gap the frequency must satisfy $\epsilon(\omega) = -\tanh(Kd/2)$, where $K^2 = k_x^2 + k_y^2$. The electric potential will have the form

$$\varphi = f(z)\exp(ik_x x + ik_y y - i\omega t)\ .$$

Look for nonretarded solution—that is, solutions of the Laplace equation rather than of the wave equation. The sum of the zero-point energy of all gap modes is the nonretarded part of the van der Waals attraction between the two specimens—see N. G. van Kampen, B. R. A. Nijboer, and K. Schram, Physics Letters **26A**, 307 (1968).

*This problem is somewhat difficult.

15

Optical Processes and Excitons

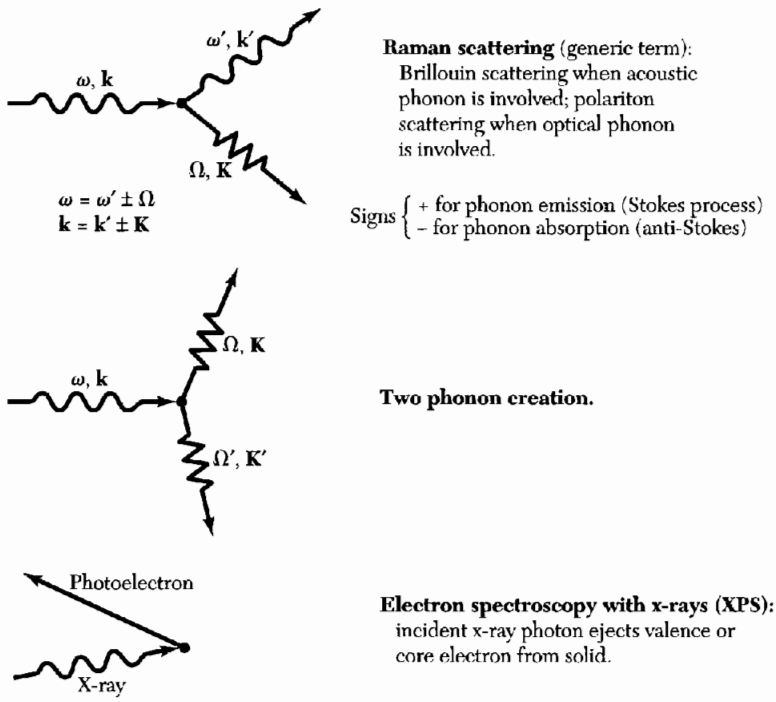


Figure 1 There are many types of experiments in which light interacts with wavelike excitations in a crystal. Several absorption processes are illustrated here.

The dielectric function $\epsilon(\omega,\mathbf{K})$ was introduced in the preceding chapter to describe the response of a crystal to an electromagnetic field (Fig. 1). The dielectric function depends sensitively on the electronic band structure of a crystal, and studies of the dielectric function by optical spectroscopy are very useful in the determination of the overall band structure of a crystal. Indeed, optical spectroscopy has developed into the most important experimental tool for band structure determination.

In the infrared, visible, and ultraviolet spectral regions the wavevector of the radiation is very small compared with the shortest reciprocal lattice vector, and therefore it may usually be taken as zero. We are concerned then with the real ϵ' and imaginary ϵ'' parts of the dielectric function at zero wavevector; $\epsilon(\omega) = \epsilon'(\omega) + i\epsilon''(\omega)$, also written as $\epsilon_1(\omega) + i\epsilon_2(\omega)$.

However, the dielectric function is not directly accessible experimentally from optical measurements: the directly accessible functions are the reflectance $R(\omega)$, the refractive index $n(\omega)$, and the extinction coefficient $K(\omega)$. Our first objective is to relate the experimentally observable quantities to the real and imaginary parts of the dielectric function.

OPTICAL REFLECTANCE

The optical measurements that give the fullest information on the electronic system are measurements of the reflectivity of light at normal incidence on single crystals. The **reflectivity coefficient** $r(\omega)$ is a complex function defined at the crystal surface as the ratio of the reflected electric field $E(\text{refl})$ to the incident electric field $E(\text{inc})$:

$$E(\text{refl})/E(\text{inc}) \equiv r(\omega) \equiv \rho(\omega) \exp[i\theta(\omega)] \ , \tag{1}$$

where we have separated the amplitude $\rho(\omega)$ and phase $\theta(\omega)$ components of the reflectivity coefficient.

The **refractive index** $n(\omega)$ and the **extinction coefficient** $K(\omega)$ in the crystal are related to the reflectivity at normal incidence by

$$r(\omega) = \frac{n + iK - 1}{n + iK + 1} \ , \tag{2}$$

as derived in Problem 3 by elementary consideration of the continuity of the components of $\mathbf{E}$ and $\mathbf{B}$ parallel to the crystal surface. By definition, $n(\omega)$ and $K(\omega)$ are related to the dielectric function $\epsilon(\omega)$ by

$$\sqrt{\epsilon(\omega)} \equiv n(\omega) + iK(\omega) \equiv N(\omega) \ , \tag{3}$$

where $N(\omega)$ is the complex refractive index. Do not confuse $K(\omega)$ as used here with the symbol for a wavevector.

If the incident traveling wave has the wavevector k, then the y component of a wave traveling in the x direction is

$$E_y(\text{inc}) = E_{y0} \exp[i(kx - \omega t)] \ . \tag{4}$$

The transmitted wave in the medium is attenuated because, by the dispersion relation for electromagnetic waves, the wavevector in the medium is related to the incident k in vacuum by $(n + iK)k$:

$$E_y(\text{trans}) \propto \exp\{[i[(n + iK)kx - wt]\} = \exp(-Kkx) \exp[i(nkx - \omega t)] \ . \tag{5}$$

One quantity measured in experiments is the **reflectance** R, defined as the ratio of the reflected intensity to the incident intensity:

$$R = E^*(\text{refl})E(\text{refl})/E^*(\text{inc})E(\text{inc}) = r^*r = \rho^2 \ . \tag{6}$$

It is difficult to measure the phase $\theta(\omega)$ of the reflected wave, but we show below that it can be calculated from the measured reflectance $R(\omega)$ if this is known at all frequencies.

Once we know both $R(\omega)$ and $\theta(\omega)$, we can proceed by (2) to obtain $n(\omega)$ and $K(\omega)$. We use these in (3) to obtain $\epsilon(\omega) = \epsilon'(\omega) + i\epsilon''(\omega)$, where $\epsilon'(\omega)$ and $\epsilon''(\omega)$ are the real and imaginary parts of the dielectric function. The inversion of (3) gives

$$\epsilon'(\omega) = n^2 - K^2 \ ; \qquad \epsilon''(\omega) = 2nK \ . \tag{7}$$

We now show how to find the phase $\theta(\omega)$ as an integral over the reflectance $R(\omega)$; by a similar method we relate the real and imaginary parts of the dielectric function. In this way we can find everything from the experimental $R(\omega)$.

Kramers-Kronig Relations

The Kramers-Kronig relations enable us to find the real part of the response of a linear passive system if we know the imaginary part of the response at all frequencies, and vice versa. They are central to the analysis of optical experiments on solids.

The response of any linear passive system can be represented as the superposition of the responses of a collection of damped harmonic oscillators with masses M_j. Let the response function $\alpha(\omega) = \alpha'(\omega) + i\alpha''(\omega)$ of the collection of oscillators be defined by

$$x_\omega = \alpha(\omega)F_\omega \ , \tag{8}$$

where the applied force field is the real part of $F_\omega \exp(-i\omega t)$ and the total displacement $x = \sum_j x_j$ is the real part of $x_\omega \exp(-i\omega t)$. From the equation of motion,

$$M_j(d^2/dt^2 + \rho_j d/dt + \omega_j^2)x_j = F \ ,$$

we have the complex response function of the oscillator system:

$$\alpha(\omega) = \sum_j \frac{f_j}{\omega_j^2 - \omega^2 - i\omega\rho_j} = \sum f_j \frac{\omega_j^2 - \omega^2 + i\omega\rho_j}{(\omega_j^2 - \omega^2)^2 + \omega^2\rho_j^2} , \tag{9}$$

where the constants $f_j = 1/M_j$ and relaxation frequencies ρ_j are all positive for a passive system.

If $\alpha(\omega)$ is the dielectric polarizability of atoms in concentration n, then f has the form of an oscillator strength times ne^2/m; such a dielectric response function is said to be of the Kramers-Heisenberg form. The relations we develop also apply to the electrical conductivity $\sigma(\omega)$ in Ohm's law, $j_\omega = \sigma(\omega)E_\omega$.

We need not assume the specific form (9), but we make use of three properties of the response function viewed as a function of the complex variable ω. Any function with the following properties will satisfy the Kramers-Kronig relations (11):

(a) The poles of $\alpha(\omega)$ are all below the real axis.

(b) The integral of $\alpha(\omega)/\omega$ vanishes when taken around an infinite semicircle in the upper half of the complex ω-plane. It suffices that $\alpha(\omega) \to 0$ uniformly as $|\omega| \to \infty$.

(c) The function $\alpha'(\omega)$ is even and $\alpha''(\omega)$ is odd with respect to real ω.

Consider the Cauchy integral in the form

$$\alpha(\omega) = \frac{1}{\pi i} \mathrm{P} \int_{-\infty}^{\infty} \frac{\alpha(s)}{s - \omega} ds , \tag{10}$$

where P denotes the principal part of the integral, as discussed in the mathematical note that follows. The right-hand side is to be completed by an integral over the semicircle at infinity in the upper half-plane, but we have seen in (b) that this integral vanishes.

We equate the real parts of (10) to obtain

$$\alpha'(\omega) = \frac{1}{\pi} \mathrm{P} \int_{-\infty}^{\infty} \frac{\alpha''(s)}{s - \omega} ds = \frac{1}{\pi} \mathrm{P} \left[\int_0^{\infty} \frac{\alpha''(s)}{s - \omega} ds + \int_{-\infty}^{0} \frac{\alpha''(p)}{p - \omega} dp \right] .$$

In the last integral we substitute s for $-p$ and use property (c) that $\alpha''(-s) = -\alpha''(s)$; this integral then becomes

$$\int_0^{\infty} \frac{\alpha''(s)}{s + \omega} ds ,$$

and we have, with

$$\frac{1}{s - \omega} + \frac{1}{s + \omega} = \frac{2s}{s^2 - \omega^2} ,$$

the result

$$\boxed{\alpha'(\omega) = \frac{2}{\pi} \mathrm{P} \int_0^\infty \frac{s\alpha''(s)}{s^2 - \omega^2} ds \ .} \tag{11a}$$

This is one of the Kramers-Kronig relations. The other relation follows on equating the imaginary parts of Eq. (10):

$$\alpha''(\omega) = -\frac{1}{\pi} \mathrm{P} \int_{-\infty}^{\infty} \frac{\alpha'(s)}{s - \omega} ds = -\frac{1}{\pi} \mathrm{P} \left[\int_0^\infty \frac{\alpha'(s)}{s - \omega} ds - \int_0^\infty \frac{\alpha'(s)}{s + \omega} ds \right] ,$$

whence

$$\boxed{\alpha''(\omega) = -\frac{2\omega}{\pi} \mathrm{P} \int_0^\infty \frac{\alpha'(s)}{s^2 - \omega^2} ds \ .} \tag{11b}$$

These relations are applied below to the analysis of optical reflectance data; this is their most important application.

Let us apply the Kramers-Kronig relations to $r(\omega)$ viewed as a response function between the incident and reflected waves in (1) and (6). We apply (11) to

$$\ln r(\omega) = \ln R^{1/2}(\omega) + i\theta(\omega) \tag{12}$$

to obtain the phase in terms of the reflectance:

$$\theta(\omega) = -\frac{\omega}{\pi} \mathrm{P} \int_0^\infty \frac{\ln R(s)}{s^2 - \omega^2} ds \ . \tag{13}$$

We integrate by parts to obtain a form that gives insight into the contributions to the phase angle:

$$\theta(\omega) = -\frac{1}{2\pi} \int_0^\infty \ln \left| \frac{s + \omega}{s - \omega} \right| \frac{d \ln R(s)}{ds} ds \ . \tag{14}$$

Spectral regions in which the reflectance is constant do not contribute to the integral; further, spectral regions $s \gg \omega$ and $s \ll \omega$ do not contribute much because the function $\ln |(s + \omega)/(s - \omega)|$ is small in these regions.

Mathematical Note. To obtain the Cauchy integral (10) we take the integral $\int \alpha(s)(s - \omega)^{-1} ds$ over the contour in Fig. 2. The function $\alpha(s)$ is analytic in the upper half-plane, so that the value of the integral is zero. The contribution of segment (4) to the integral vanishes if the integrand $\alpha(s)/s \to 0$ is faster than $|s|^{-1}$ as $|s| \to \infty$. For the response function (9) the integrand $\to 0$ as $|s|^{-3}$; and for the conductivity $\sigma(s)$ the integrand $\to 0$ as $|s|^{-2}$. The segment (2) contributes, in the limit as $u \to 0$,

$$\int_{(2)} \frac{\alpha(s)}{s - \omega} ds \to \alpha(\omega) \int_\pi^0 \frac{iu\, e^{i\theta} d\theta}{u\, e^{i\theta}} = -\pi i \alpha(\omega)$$

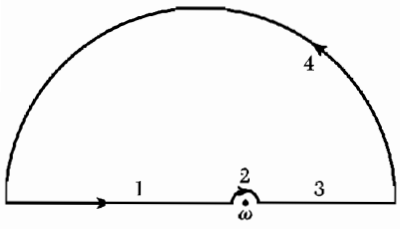


Figure 2 Contour for the Cauchy principal value integral.

to the integral, where $s = \omega + u\, e^{i\theta}$. The segments (1) and (3) are by definition the principal part of the integral between $-\infty$ and ∞. Because the integral over (1) + (2) + (3) + (4) must vanish,

$$\int_{(1)} + \int_{(3)} \equiv \mathrm{P} \int_{-\infty}^{\infty} \frac{\alpha(s)}{s-\omega}\, ds = \pi i \alpha(\omega) \ , \tag{15}$$

as in (10).

EXAMPLE: ***Conductivity of Collisionless Electron Gas.*** Consider a gas of free electrons in the limit as the collision frequency goes to zero. From (9) the response function is, with $f = 1/m$,

$$\alpha(\omega) = -\frac{1}{m\omega} \lim_{\rho \to 0} \frac{1}{\omega + i\rho} = -\frac{1}{m\omega}\left[\frac{1}{\omega} - i\pi\delta(\omega)\right] , \tag{16}$$

by the Dirac identity. We confirm that the delta function in (16) satisfies the Kramers-Kronig relation (11a), by which

$$\alpha'(\omega) = \frac{2}{m} \int_0^{\infty} \frac{\delta(s)}{s^2 - \omega^2}\, ds = -\frac{1}{m\omega^2} \ , \tag{17}$$

in agreement with (16).

We obtain the electrical conductivity $\sigma(\omega)$ from the dielectric function

$$\epsilon(\omega) - 1 = 4\pi P_\omega / E_\omega = -4\pi n e x_\omega / E_\omega = 4\pi n e^2 \alpha(\omega) \ , \tag{18}$$

where $\alpha(\omega) = x_\omega/(-e)E_\omega$ is the response function. We use the equivalence

$$\text{(CGS)} \qquad \sigma(\omega) = (-i\omega/4\pi)[\epsilon(\omega) - 1] \ , \tag{19}$$

for the Maxwell equation can be written either as c curl $\mathbf{H} = 4\pi\sigma(\omega)\mathbf{E} - i\omega\mathbf{E}$ or as c curl $\mathbf{H} = -i\omega\epsilon(\omega)\mathbf{E}$. We combine (16), (18), and (19) to find the conductivity of a collisionless electron gas:

$$\sigma'(\omega) + i\sigma''(\omega) = \frac{ne^2}{m}\left[\pi\delta(\omega) + \frac{i}{\omega}\right] . \tag{20}$$

For collisionless electrons the real part of the conductivity has a delta function at $\omega = 0$.

Electronic Interband Transitions

It came as a surprise that optical spectroscopy developed as an important experimental tool for the determination of band structure. First, the absorption and reflection bands of crystals are broad and apparently featureless functions of the photon energy when this is greater than the band gap. Second, direct interband absorption of a photon $\hbar\omega$ will occur at all points in the Brillouin zone for which energy is conserved:

$$\hbar\omega = \epsilon_c(\mathbf{k}) - \epsilon_v(\mathbf{k}) \ , \tag{21}$$

where c is an empty band and v is a filled band. The total absorption at given ω is an integral over all transitions in the zone that satisfy (21).

Three factors unraveled the spectra:

- The broad bands are not like a spectral line greatly broadened by damping, but the bands convey much intelligence which emerges when derivatives are taken of the reflectance (Fig. 3); derivatives with respect to wavelength, electric field, temperature, pressure, or uniaxial stress, for example. The spectroscopy of derivatives is called modulation spectroscopy.

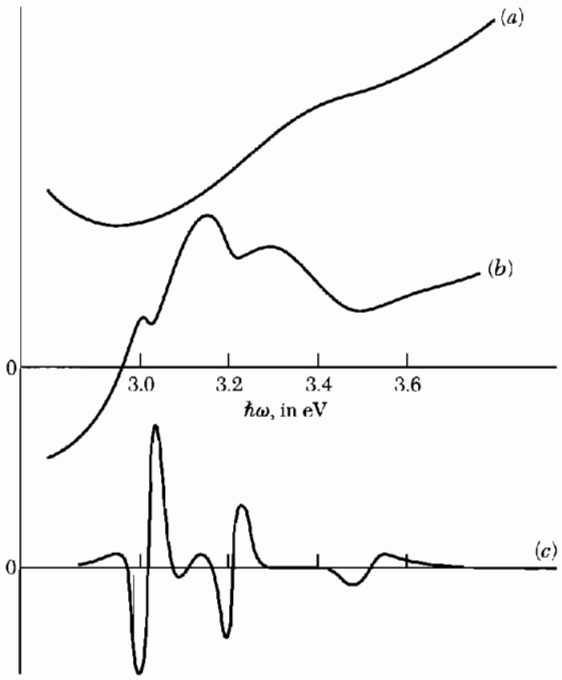


Figure 3 Comparison of (a) reflectance, (b) wavelength derivative reflectance (first derivative), and (c) electroreflectance (third derivative), of the spectral region in germanium between 3.0 and 3.6 eV. (After data by D. D. Sell, E. O. Kane, and D. E. Aspnes.)

- The relation (21) does not exclude spectral structure in a crystal, because transitions accumulate at frequencies for which the bands c, v are parallel—that is, at frequencies where

$$\nabla_{\mathbf{k}}[\epsilon_c(\mathbf{k}) - \epsilon_v(\mathbf{k})] = 0 \ . \tag{22}$$

  At these critical points in $\mathbf{k}$ space the joint density of states $D_c(\epsilon_v + \hbar\omega)D_v(\epsilon_v)$ is singular, according to the same argument we used in (5.37) to show that the density of phonon modes $D(\omega)$ is singular when $\nabla_{\mathbf{k}}\omega$ is zero.
- The pseudopotential method for calculating energy bands helps identify the positions in the Brillouin zone of the critical points found in modulation spectra. Band-band energy differences can be calculated with an accuracy as good as 0.1 eV. The experimental results can then be fed back to give improvements in the pseudopotential calculations.

EXCITONS

Reflectance and absorption spectra often show structure for photon energies just below the energy gap, where we might expect the crystal to be transparent. This structure is caused by the absorption of a photon with the creation of a bound electron-hole pair. An electron and a hole may be bound together by their attractive coulomb interaction, just as an electron is bound to a proton to form a neutral hydrogen atom.

The bound electron-hole pair is called an **exciton**, Fig. 4. An exciton can move through the crystal and transport energy; it does not transport charge

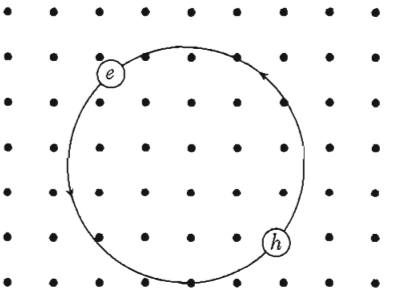


Figure 4a An exciton is a bound electron-hole pair, usually free to move together through the crystal. In some respects it is similar to an atom of positronium, formed from a positron and an electron. The exciton shown is a Mott-Wannier exciton: it is weakly bound, with an average electron-hole distance large in comparison with the lattice constant.

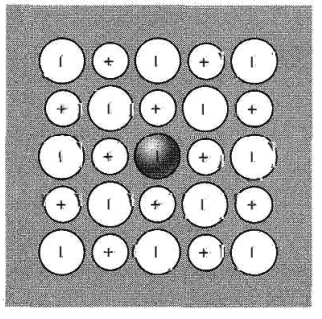


Figure 4b A tightly-bound or Frenkel exciton shown localized on one atom in an alkali halide crystal. An ideal Frenkel exciton will travel as a wave throughout the crystal, but the electron is always close to the hole.

because it is electrically neutral. It is similar to positronium, which is formed from an electron and a positron.

Excitons can be formed in every insulating crystal. When the band gap is indirect, excitons near the direct gap may be unstable with respect to decay into a free electron and free hole. All excitons are unstable with respect to the ultimate recombination process in which the electron drops into the hole. Excitons can also form complexes, such as a biexciton from two excitons.

We have seen that a free electron and free hole are created whenever a photon of energy greater than the energy gap is absorbed in a crystal. The threshold for this process is $\hbar\omega > E_g$ in a direct process. In the indirect phonon-assisted process of Chapter 8 the threshold is lower by the phonon energy $\hbar\Omega$. But in the formation of excitons the energy is lowered with respect to these thresholds by the binding energy of the exciton, which may be in the range 1 meV to 1 eV, as in Table 1.

Excitons can be formed by photon absorption at any critical point (22), for if $\nabla_k \epsilon_v = \nabla_k \epsilon_c$ the group velocities of electron and hole are equal and the particles may be bound by their coulomb attraction. Transitions leading to the formation of excitons below the energy gap are indicated in Figs. 5 and 6.

The binding energy of the exciton can be measured in three ways:

- In optical transitions from the valence band, by the difference between the energy required to create an exciton and the energy to create a free electron and free hole, Fig. 7.
- In recombination luminescence, by comparison of the energy of the free electron-hole recombination line with the energy of the exciton recombination line.
- By photo-ionization of excitons, to form free carriers. This experiment requires a high concentration of excitons.

We discuss excitons in two different limiting approximations, one by Frenkel in which the exciton is small and tightly bound, and the other by Mott and Wannier in which the exciton is weakly bound, with an electron-hole separation large in comparison with a lattice constant. Intermediate examples are known.

Table 1 Binding energy of excitons, in meV

Si	14.7	BaO	56.	RbCl	440.
Ge	4.15	InP	4.0	LiF	(1000)
GaAs	4.2	InSb	(0.4)	AgBr	20.
GaP	3.5	KI	480.	AgCl	30.
CdS	29.	KCl	400.	TlCl	11.
CdSe	15.	KBr	400.	TlBr	6.

Data assembled by Frederick C. Brown and Arnold Schmidt.

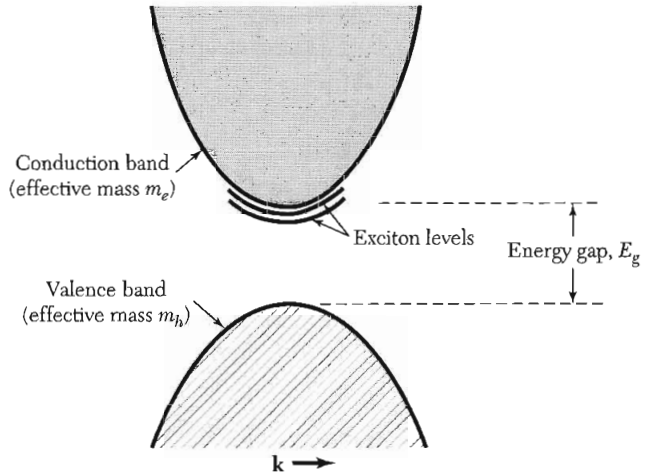


Figure 5 Exciton levels in relation to the conduction band edge, for a simple band structure with both conduction and valence band edges at $k = 0$. An exciton can have translational kinetic energy. Excitons are unstable with respect to radiative recombination in which the electron drops into the hole in the valence band, accompanied by the emission of a **photon** or **phonons**.

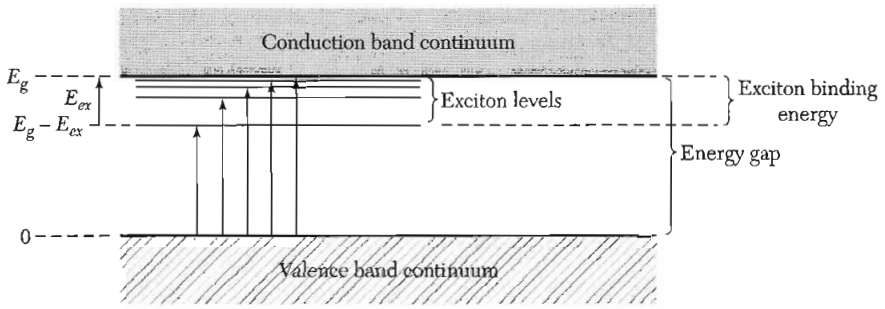


Figure 6 Energy levels of an exciton created in a direct process. Optical transitions from the top of the valence band are shown by the arrows; the longest arrow corresponds to the energy gap. The binding energy of the exciton is E_{ex}, referred to a free electron and free hole. The lowest frequency absorption line of the crystal at absolute zero is not E_{ex}, but is $E_g - E_{ex}$.

Frenkel Excitons

In a tightly bound exciton (Fig. 4b) the excitation is localized on or near a single atom: the hole is usually on the same atom as the electron although the pair may be anywhere in the crystal. A Frenkel exciton is essentially an excited state of a single atom, but the excitation can hop from one atom to another by virtue of the coupling between neighbors. The excitation wave travels through the crystal much as the reversed spin of a magnon travels through the crystal.

The crystalline inert gases have excitons which in their ground states correspond somewhat to the Frenkel model. Atomic krypton has its lowest strong

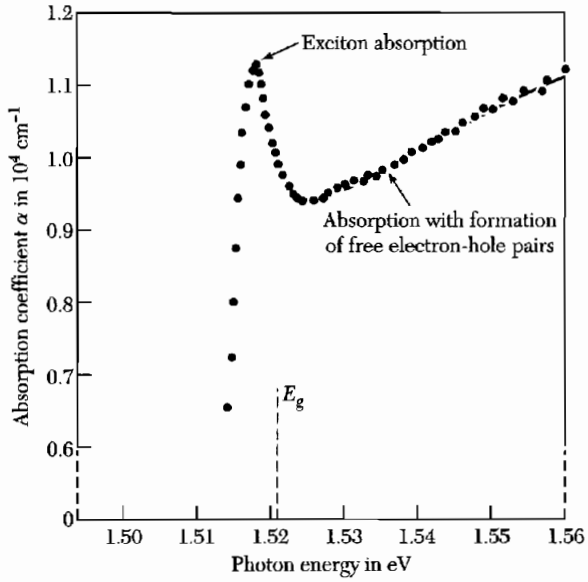


Figure 7 Effect of an exciton level on the optical absorption of a semiconductor for photons of energy near the band gap E_g in gallium arsenide at 21 K. The vertical scale is the intensity absorption coefficient α, as in $I(x) = I_0 \exp(-\alpha x)$. The energy gap and exciton binding energy are deduced from the shape of the absorption curve: the gap E_g is 1.521 eV and the exciton binding energy is 0.0034 eV. (After M. D. Sturge.)

atomic transition at 9.99 eV. The corresponding transition in the crystal is closely equal and is at 10.17 eV, Fig. 8. The energy gap in the crystal is 11.7 eV, so the exciton ground state energy is 11.7 − 10.17 = 1.5 eV, referred to a free electron and free hole separated and at rest in the crystal.

The translational states of Frenkel excitons have the form of propagating waves, like all other excitations in a periodic structure. Consider a crystal of N atoms on a line or ring. If u_j is the ground state of atom j, the ground state of the crystal is

$$\psi_g = u_1 u_2 \cdots u_{N-1} u_N \ , \tag{23}$$

if interactions between the atoms are neglected. If a single atom j is in an excited state v_j, the system is described by

$$\varphi_j = u_1 u_2 \cdots u_{j-1} v_j u_{j+1} \cdots u_N \ . \tag{24}$$

This function has the same energy as the function φ_l with any other atom l excited. However, the functions φ that describe a single excited atom and $N - 1$ atoms in their ground state are not the stationary quantum states of the problem. If there is any interaction between an excited atom and a nearby atom in its ground state, the excitation energy will be passed from atom to atom. The eigenstates will have a wavelike form, as we now show.

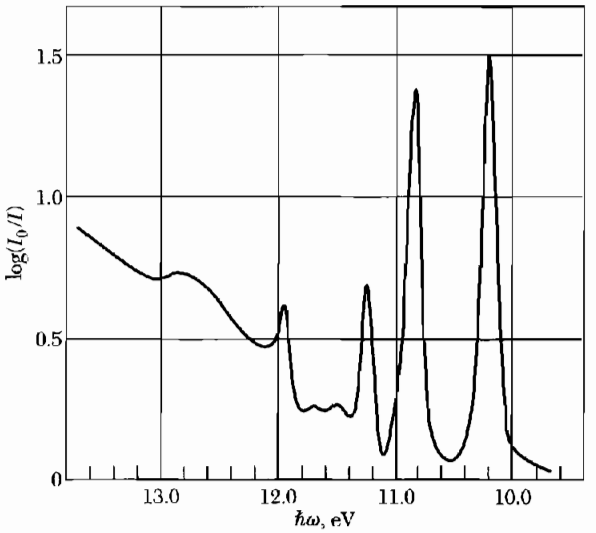


Figure 8 Absorption spectrum of solid krypton at 20 K. (After G. Baldini.)

When the hamiltonian of the system operates on the function φ_j with the jth atom excited, we obtain

$$\mathcal{H}\varphi_j = \epsilon\varphi_j + T(\varphi_{j-1} + \varphi_{j+1}) \ , \tag{25}$$

where ϵ is the free atom excitation energy; the interaction T measures the rate of transfer of the excitation from j to its nearest neighbors, $j - 1$ and $j + 1$. The solutions of (25) are waves of the Bloch form:

$$\psi_k = \sum_j \exp(ijka)\, \varphi_j \ . \tag{26}$$

To see this we let $\mathcal{H}$ operate on ψ_k:

$$\mathcal{H}\psi_k = \sum_j e^{ijka}\, \mathcal{H}\varphi_j = \sum_j e^{ijka}\, [\epsilon\varphi_j + T(\varphi_{j-1} + \varphi_{j+1})] \ , \tag{27}$$

from (25). We rearrange the right-hand side to obtain

$$\mathcal{H}\psi_k = \sum_j e^{ijka}[\epsilon + T(e^{ika} + e^{-ika})]\varphi_j = (\epsilon + 2T\cos ka)\psi_k \ , \tag{28}$$

so that the energy eigenvalues of the problem are

$$E_k = \epsilon + 2T\cos ka \ , \tag{29}$$

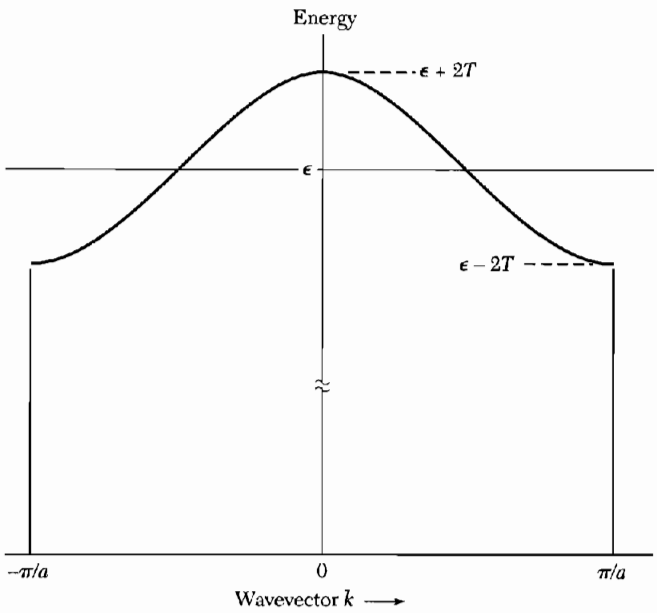


Figure 9 Energy versus wavevector for a Frenkel exciton, calculated with positive nearest-neighbor transfer interaction T.

as in Fig. 9. The application of periodic boundary conditions determines the allowed values of the wavevector k:

$$k = 2\pi s/Na; \quad s = -\tfrac{1}{2}N, -\tfrac{1}{2}N + 1, \cdots, \tfrac{1}{2}N - 1 \ . \tag{30}$$

Alkali Halides. In alkali halide crystals the lowest-energy excitons are localized on the negative halogen ions, as in Fig. 4b. The negative ions have lower electronic excitation levels than do the positive ions. Pure alkali halide crystals are transparent in the visible spectral region, which means that the exciton energies do not lie in the visible, but the crystals show considerable excitonic absorption structure in the vacuum ultraviolet.

A doublet structure is particularly evident in sodium bromide, a structure similar to that of the lowest excited state of the krypton atom—which is isoelectronic with the Br^- ion of KBr. The splitting is caused by the spin-orbit interaction. These excitons are Frenkel excitons.

Molecular Crystals. In molecular crystals the covalent binding within a molecule is strong in comparison with the van der Waals binding between molecules, so that the excitons are Frenkel excitons. Electronic excitation lines of an individual molecule appear in the crystalline solid as an exciton, often with little shift in frequency. At low temperatures the lines in the solid are quite sharp, although there may be more structure to the lines in the solid than in the molecule because of the Davydov splitting, as discussed in Problem 7.

Weakly Bound (Mott-Wannier) Excitons

Consider an electron in the conduction band and a hole in the valence band. The electron and hole attract each other by the coulomb potential

(CGS) $$U(r) = -e^2/\epsilon r \ , \tag{31}$$

where r is the distance between the particles and ϵ is the appropriate dielectric constant. (The lattice polarization contribution to the dielectric constant should not be included if the frequency of motion of the exciton is higher than the optical phonon frequencies.) There will be bound states of the exciton system having total energies lower than the bottom of the conduction band.

The problem is the hydrogen atom problem if the energy surfaces for the electron and hole are spherical and nondegenerate. The energy levels referred to the top of the valence band are given by a modified Rydberg equation

(CGS) $$E_n = E_g - \frac{\mu e^4}{2\hbar^2\epsilon^2 n^2} \ . \tag{32}$$

Here n is the principal quantum number and μ is the reduced mass:

$$\frac{1}{\mu} = \frac{1}{m_e} + \frac{1}{m_h} \ , \tag{33}$$

formed from the effective masses m_e, m_h of the electron and hole.

The exciton ground state energy is obtained on setting $n = 1$ in (32); this is the ionization energy of the exciton. Studies of the optical absorption lines in cuprous oxide, Cu_2O, at low temperatures give results for the exciton level spacing in good agreement with the Rydberg equation (32) except for transitions to the state $n = 1$. An empirical fit to the lines of Fig. 10 is obtained with the relation $\nu(\text{cm}^{-1}) = 17{,}508 - (800/n^2)$. Taking $\epsilon = 10$, we find $\mu \simeq 0.7\ m$ from the coefficient of $1/n^2$. The constant term 17,508 cm^{-1} corresponds to an energy gap $E_g = 2.17$ eV.

Exciton Condensation into Electron-Hole Drops (EHD)

A condensed phase of an electron-hole plasma forms in Ge and Si when maintained at a low temperature and irradiated by light. The following sequence of events takes place when an electron-hole drop (EHD) is formed in Ge: The absorption of a photon of energy $\hbar\omega > E_g$ produces a free electron and free hole, with high efficiency. These combine rapidly, perhaps in 1 ns, to form an exciton. The exciton may decay with annihilation of the e-h pair with a lifetime of 8 μs.

If the exciton concentration is sufficiently high—over 10^{13} cm^{-3} at 2 K—most of the excitons will condense into a drop. The drop lifetime is 40 μs, but in strained Ge may be as long as 600 μs. Within the drop the excitons dissolve into a degenerate Fermi gas of electrons and holes, with metallic properties: this state was predicted by L. V. Keldysh.

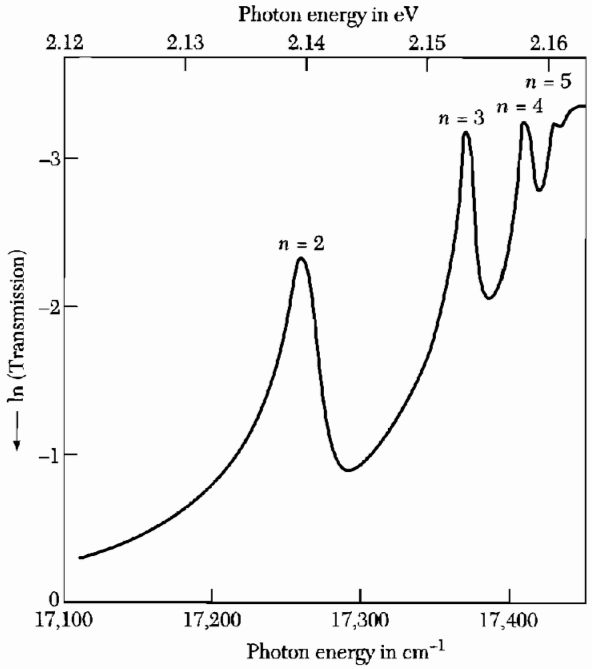


Figure 10 Logarithm of the optical transmission versus photon energy in cuprous oxide at 77 K, showing a series of exciton lines. Note that on the vertical axis the logarithm is plotted decreasing upward; thus a peak corresponds to absorption. The band gap E_g is 2.17 eV. (After P. W. Baumeister.)

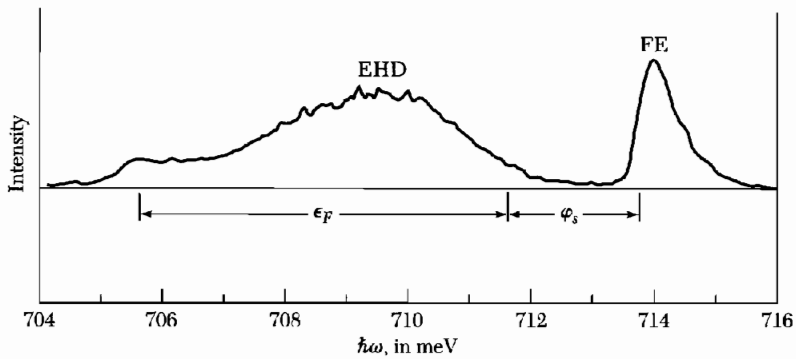


Figure 11 Recombination radiation of free electrons with holes and of electron-hole drops in Ge at 3.04 K. The Fermi energy in the drop is ϵ_F and the cohesive energy of the drop with respect to a free exciton is ϕ_s. (After T. K. Lo.)

Figure 11 shows the recombination radiation in Ge from free excitons (714 meV) and from the EHD phase (709 meV). The width of the 714 meV line is accounted for by Doppler broadening, and the width of the 709 meV line is compatible with the kinetic energy distribution of electrons and holes in a Fermi gas of concentration 2×10^{17} cm^{-3}. Figure 12 is a photograph of a large EHD.

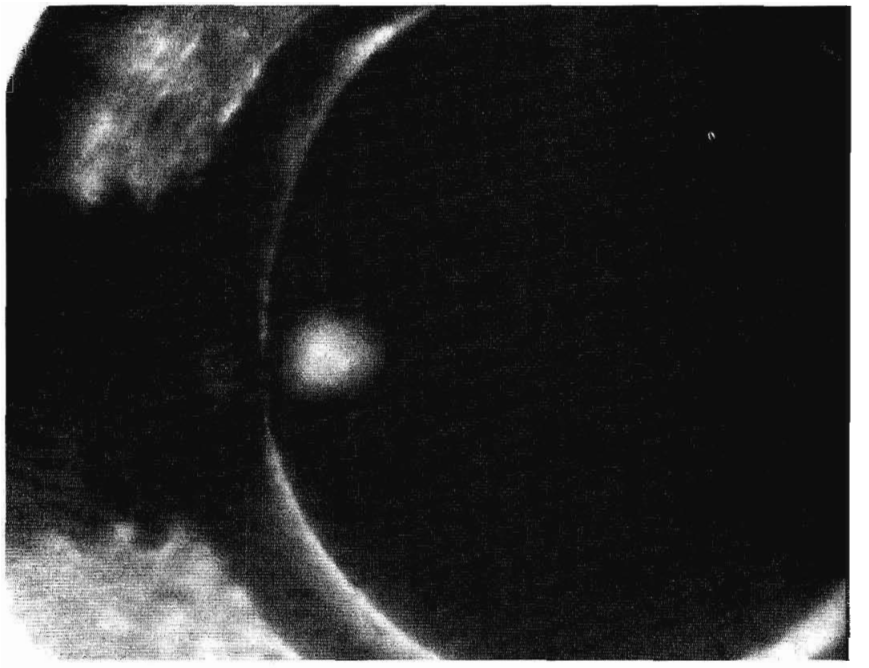


Figure 12 Photograph of an electron-hole drop in a 4 mm disk of pure germanium. The drop is the intense spot adjacent to the set screw on the left of the disk. The photograph is the image of the drop obtained by focusing its electron-hole recombination luminescence onto the surface of an infrared-sensitive camera. (After J. P. Wolfe et al.)

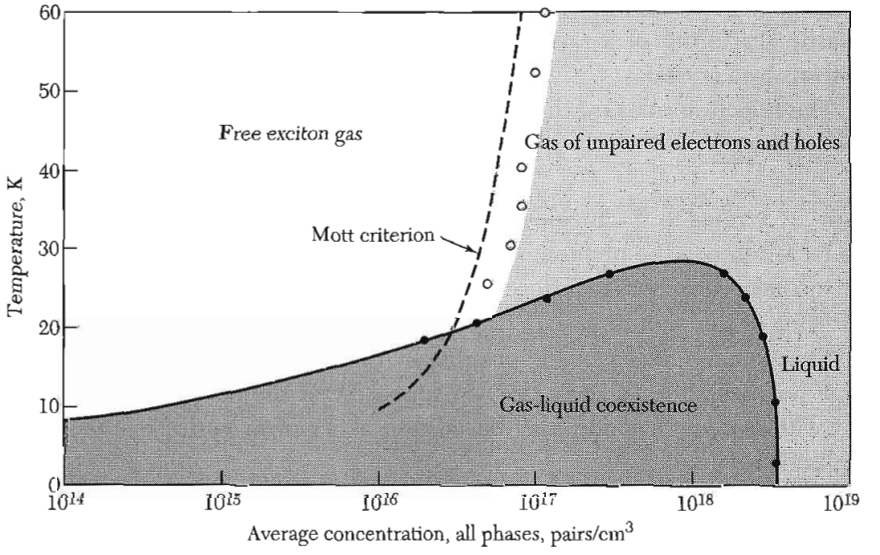


Figure 13 Phase diagram for photoexcited electrons and holes in unstressed silicon. The diagram shows, for example, that with an average concentration near 10^{17} cm^{-3} at 15 K, a free-exciton gas with saturated-gas concentration of 10^{16} cm^{-3} coexists with a (variable) volume of liquid droplets, each with a density of 3×10^{18} cm^{-3}. The liquid critical temperature is about 23 K. Theoretical and experimental values for the metal-insulator transition for excitons are also shown. (From J. P. Wolfe.)

Table 2 Electron-hole liquid parameters

Crystal (unstressed)	Binding energy relative to free exciton, in meV	Concentration, n or p, in cm^{-3}	Critical temperature in K
Ge	1.8	2.57×10^{17}	6.7
Si	9.3	3.5×10^{18}	23.
GaP	17.5	8.6×10^{18}	45.
3C-SiC	17.	$10. \times 10^{18}$	41.

Courtesy of D. Bimberg.

The exciton phase diagram for silicon is plotted in the temperature-concentration plane in Fig. 13. The exciton gas is insulating at low pressures. At high pressures (at the right of the diagram) the exciton gas breaks up into a conducting plasma of unpaired electrons and holes. The transition from excitons to the plasma is an example of the Mott transition, Chapter 14. Further data are given in Table 2.

RAMAN EFFECT IN CRYSTALS

Raman scattering involves two photons—one in, one out—and is one step more complex than the one photon processes treated earlier in this chapter. In the Raman effect a photon is scattered inelastically by a crystal, with creation or annihilation of a phonon or magnon (Fig. 14). The process is identical to the inelastic scattering of x-rays and it is similar to the inelastic scattering of neutrons by a crystal.

The selection rules for the first-order Raman effect are

$$\omega = \omega' \pm \Omega \ ; \quad \mathbf{k} = \mathbf{k}' \pm \mathbf{K} \ , \tag{34}$$

where ω, $\mathbf{k}$ refer to the incident photon; ω', $\mathbf{k}'$ refer to the scattered photon; and Ω, $\mathbf{K}$ refer to the phonon created or destroyed in the scattering event. In the second-order Raman effect, two phonons are involved in the inelastic scattering of the photon.

The Raman effect is made possible by the strain-dependence of the electronic polarizability. To show this, we suppose that the polarizability α associated with a phonon mode may be written as a power series in the phonon amplitude u:

$$\alpha = \alpha_0 + \alpha_1 u + \alpha_2 u^2 + \cdots \ . \tag{35}$$

If $u(t) = u_0 \cos \Omega t$ and the incident electric field is $E(t) = E_0 \cos \omega t$, then the induced electric dipole moment has a component

$$\alpha_1 E_0 u_0 \cos \omega t \cos \Omega t = \tfrac{1}{2}\alpha_1 E_0 u_0 [\cos(\omega + \Omega)t + \cos(\omega - \Omega)t] \ . \tag{36}$$

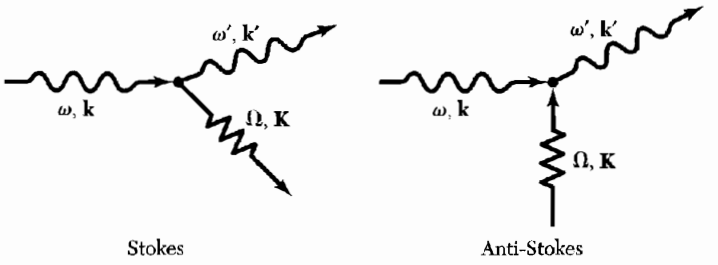


Figure 14 Raman scattering of a photon with emission or absorption of a phonon. Similar processes occur with magnons (spin waves).

Thus photons at frequencies $\omega + \Omega$ and $\omega - \Omega$ can be emitted, accompanied by absorption or emission of a phonon of frequency Ω.

The photon at $\omega - \Omega$ is called the Stokes line and that at $\omega + \Omega$ is the anti-Stokes line. The intensity of the Stokes line involves the matrix element for phonon creation, which is just the matrix element for the harmonic oscillator, as in Appendix C:

$$I(\omega - \Omega) \propto |\langle n_{\mathbf{K}} + 1|u|n_{\mathbf{K}}\rangle|^2 \propto n_{\mathbf{K}} + 1 \ , \tag{37}$$

where $n_{\mathbf{K}}$ is the initial population of phonon mode $\mathbf{K}$.

The anti-Stokes line involves phonon annihilation, with a photon intensity proportional to

$$I(\omega + \Omega) \propto |\langle n_{\mathbf{K}} - 1|u|n_{\mathbf{K}}\rangle|^2 \propto n_{\mathbf{K}} \ . \tag{38}$$

If the phonon population is initially in thermal equilibrium at temperature T, the intensity ratio of the two lines is

$$\frac{I(\omega + \Omega)}{I(\omega - \Omega)} = \frac{\langle n_{\mathbf{K}}\rangle}{\langle n_{\mathbf{K}}\rangle + 1} = \exp(-\hbar\Omega/k_B T) \ , \tag{39}$$

with $\langle n_{\mathbf{K}}\rangle$ given by the Planck distribution function $1/[\exp(\hbar\Omega/k_B T) - 1]$. We see that the relative intensity of the anti-Stokes lines vanishes as $T \to 0$, because here there are no thermal phonons available to be annihilated.

Observations on the $\mathbf{K} = 0$ optical phonon in silicon are shown in Figs. 15 and 16. Silicon has two identical atoms in the primitive cell, and there is no electric dipole moment associated with the primitive cell in the absence of deformation by phonons. But $\alpha_1 u$ does not vanish for silicon at $\mathbf{K} = 0$, so that we can observe the mode by first-order Raman scattering of light.

The second-order Raman effect arises from the term $\alpha_2 u^2$ in the polarizability. Inelastic scattering of light in this order is accompanied by the creation of two phonons, or the absorption of two phonons, or the creation of one and the absorption of another phonon. The phonons may have different frequencies. The intensity distribution in the scattered photon spectrum may be quite complicated if there are several atoms in the primitive cell because of the

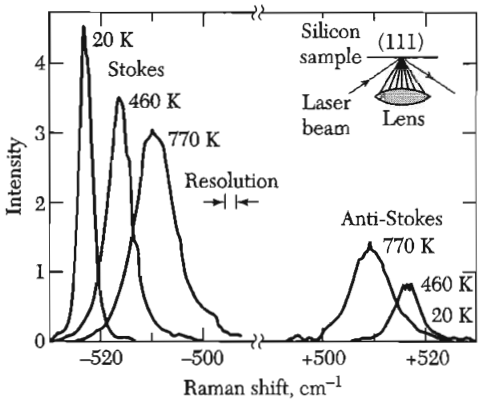


Figure 15 First-order Raman spectra of the $K \simeq 0$ optical mode of a silicon crystal observed at three temperatures. The incident photon has a wavelength of 5145 Å. The optical phonon frequency is equal to the frequency shift; it depends slightly on the temperature. (After T. R. Hart, R. L. Aggarwal, and B. Lax.)

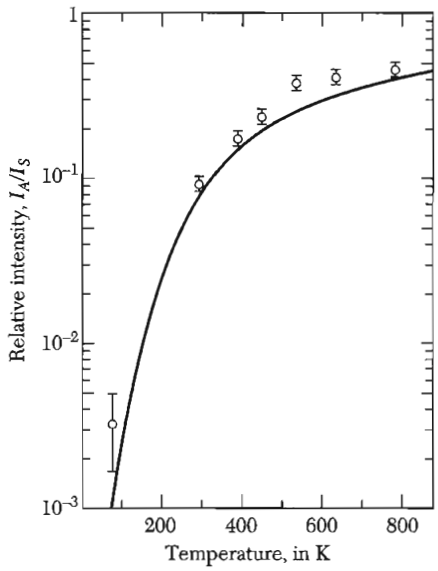


Figure 16 Intensity ratio of anti-Stokes to Stokes lines as a function of temperature, for the observations of Fig. 15 on the optical mode of silicon. The observed temperature dependence is in good agreement with the prediction of Eq. (39): the solid curve is a plot of the function $\exp(-\hbar\Omega/k_B T)$.

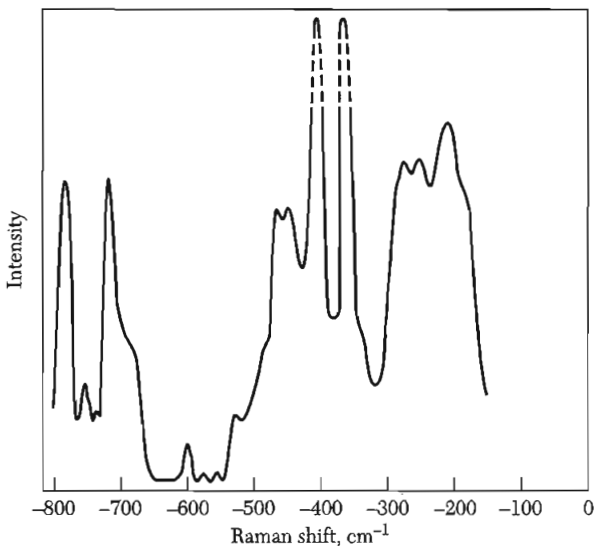


Figure 17 Raman spectrum of GaP at 20 K. The two highest peaks are the first-order Raman lines associated with the excitation of an LO phonon at 404 cm^{-1} and a TO phonon at 366 cm^{-1}. All the other peaks involve two phonons. (After M. V. Hobden and J. P. Russell.)

corresponding number of optical phonon modes. Second-order Raman spectra have been observed and analyzed in numerous crystals. Measurements on gallium phosphide (GaP) are shown in Fig. 17.

Electron Spectroscopy with X-Rays

The next degree of complexity in optical processes involves a photon in and an electron out of the solid, as in Fig. 1. Two important techniques are x-ray photoemission from solids (XPS) and ultraviolet photoemission (UPS). In solid state physics they are used in band structure studies and surface physics, including catalysis and adsorption.

XPS and UPS spectra can be compared directly with valence band densities of states $D(\epsilon)$. The specimen is irradiated with highly monochromatic x-rays or ultraviolet photons. The photon is absorbed, with the emission of a photoelectron whose kinetic energy is equal to the energy of the incident photon minus the binding energy of the electron in the solid. The electrons come from a thin layer near the surface, typically 50 Å in depth. The resolution of the best XPS spectrometer systems is less than 10 meV, which permits refined studies of band structure.

The valence band structure of silver is shown by Fig. 18, with the zero of energy set at the Fermi level. Electrons in the first 3 eV below the Fermi level come from the 5*s* conduction band. The strong peak with structure below 3 eV is from the 4*d* valence electrons.

Excitations are also seen from deeper levels, often accompanied by excitation of plasmons. For example, in silicon the 2*p* electron with a binding energy close to 99.2 eV is observed in replica at 117 eV with single plasmon excitation and at 134.7 eV with two plasmon excitation. The plasmon energy is 18 eV.

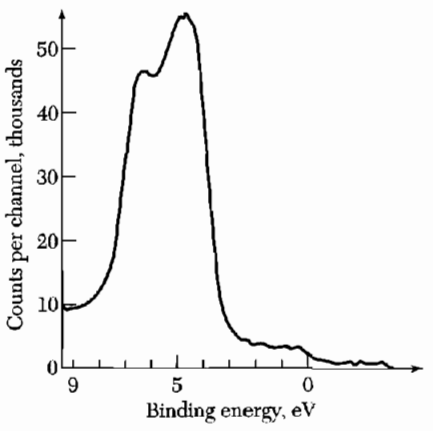


Figure 18 Valence-band electron emission from silver, after Siegbahn and co-workers.

ENERGY LOSS OF FAST PARTICLES IN A SOLID

So far we have used photons as probes of the electronic structure of solids. We can also use electron beams for the same purpose. The results also involve the dielectric function, now through the imaginary part of $1/\epsilon(\omega)$. The dielectric function enters as $\mathrm{Im}\{\epsilon(\omega)\}$ into the energy loss by an electromagnetic wave in a solid, but as $-\mathrm{Im}\{1/\epsilon(\omega)\}$ into the energy loss by a charged particle that penetrates a solid.

Consider this difference. The general result from electromagnetic theory for the power dissipation density by dielectric losses is

$$\text{(CGS)} \qquad \mathcal{P} = \frac{1}{4\pi}\,\mathbf{E}\cdot(\partial\mathbf{D}/\partial t)\ , \tag{40}$$

per unit volume. With a transverse electromagnetic wave $Ee^{-i\omega t}$ in the crystal, we have $dD/dt = -i\omega\,\epsilon(\omega)Ee^{-i\omega t}$, whence the time-average power is

$$\begin{aligned}\mathcal{P} &= \frac{1}{4\pi}\langle \mathrm{Re}\{Ee^{-i\omega t}\}\mathrm{Re}\{-i\omega\epsilon(\omega)Ee^{-i\omega t}\}\rangle \\ &= \frac{1}{4\pi}\,\omega E^2\langle(\epsilon''\cos\omega t - \epsilon'\sin\omega t)\cos\omega t\rangle = \frac{1}{8\pi}\,\omega\epsilon''(\omega)E^2\ ,\end{aligned} \tag{41}$$

proportional to $\epsilon''(\omega)$. The tangential component of E is continuous across the boundary of the solid.

If a particle of charge e and velocity v enters a crystal, the dielectric displacement is, by standard texts,

$$\text{(CGS)} \qquad \mathbf{D}(\mathbf{r},t) = -\mathrm{grad}\,\frac{e}{|\mathbf{r}-\mathbf{v}t|}\ , \tag{42}$$

because by the Poisson equation it is **D**, and not **E**, that is related to the free charge. In an isotropic medium the Fourier component $E(\omega,\mathbf{k})$ is related to the Fourier component $D(\omega,\mathbf{k})$ of $D(\mathbf{r},t)$ by $E(\omega,\mathbf{k}) = D(\omega,\mathbf{k})/\epsilon(\omega,\mathbf{k})$.

The time-average power dissipation associated with this Fourier component is

$$\begin{aligned}\mathcal{P}(\omega,\mathbf{k}) &= \frac{1}{4\pi}\langle \mathrm{Re}\{\epsilon^{-1}(\omega,\mathbf{k})D(\omega,\mathbf{k})e^{-i\omega t}\}\mathrm{Re}\{-i\omega\,D(\omega,\mathbf{k})\,e^{-i\omega t}\}\rangle \\ &= \frac{1}{4\pi}\,\omega\,D^2(\omega,\mathbf{k})\left\langle\left[\left(\frac{1}{\epsilon}\right)'\cos\omega t + \left(\frac{1}{\epsilon}\right)''\sin\omega t\right][-\sin\omega t]\right\rangle\ ,\end{aligned}$$

whence

$$\mathcal{P}(\omega,\mathbf{k}) = -\frac{1}{8\pi}\omega\left(\frac{1}{\epsilon}\right)''D^2(\omega,\mathbf{k}) = \frac{1}{8\pi}\,\omega\,\frac{\epsilon''(\omega,\mathbf{k})}{|\epsilon|^2}\,D^2(\omega,\mathbf{k})\ . \tag{43}$$

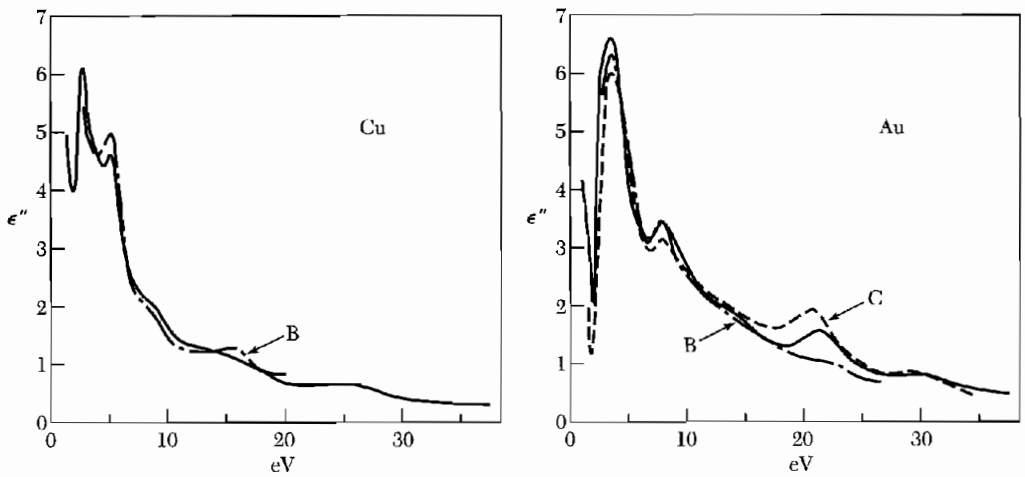


Figure 19 $\epsilon''(\omega)$ for Cu and Au; the bold lines are from energy loss measurements by J. Daniels, and the other lines were calculated from optical measurements by D. Beaglehole, and L. R. Canfield et al.

The result is the motivation for the introduction of the **energy loss function** $-\mathrm{Im}\{1/\epsilon(\omega,\mathbf{k})\}$ and it is also a motivation for experiments on energy losses by fast electrons in thin films.

If the dielectric function is independent of $\mathbf{k}$, the power loss is

$$\mathcal{P}(\omega) = -\frac{2}{\pi}\frac{e^2}{\hbar v}\,\mathrm{Im}\{1/\epsilon(\omega)\}\ln(k_0 v/\omega) \ , \tag{44}$$

where $\hbar k_0$ is the maximum possible momentum transfer from the primary particle to an electron of the crystal. Figure 19 shows the excellent experimental agreement between values of $\epsilon''(\omega)$ deduced from optical reflectivity measurements with values deduced from electron energy loss measurements.

SUMMARY

- The Kramers-Kronig relations connect the real and imaginary parts of a response function:

$$\alpha'(\omega) = \frac{2}{\pi}\mathrm{P}\int_0^\infty \frac{s\alpha''(s)}{s^2-\omega^2}ds \ ; \qquad \alpha''(\omega) = -\frac{2\omega}{\pi}\mathrm{P}\int_0^\infty \frac{\alpha'(s)}{s^2-\omega^2}ds \ .$$

- The complex refractive index $N(\omega) = n(\omega) + iK(\omega)$, where n is the refractive index and K is the extinction coefficient; further, $\epsilon(\omega) = N^2(\omega)$, whence $\epsilon'(\omega) = n^2 - K^2$ and $\epsilon''(\omega) = 2nK$.

- The reflectance at normal incidence is

$$R = \frac{(n-1)^2 + K^2}{(n+1)^2 + K^2} \ .$$

- The energy loss function $-\mathrm{Im}\{1/\epsilon(\omega)\}$ gives the energy loss by a charged particle moving in a solid.

Problems

1. ***Causality and the response function.*** The Kramers-Kronig relations are consistent with the principle that an effect not precede its cause. Consider a delta-function force applied at time $t = 0$:

$$F(t) = \delta(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega t}\, d\omega \ ,$$

whence $F_\omega = 1/2\pi$. (a) Show by direct integration or by use of the KK relations that the oscillator response function

$$\alpha(\omega) = (\omega_0^2 - \omega^2 - i\omega\rho)^{-1}$$

gives zero displacement, $x(t) = 0$, for $t < 0$ under the above force. For $t < 0$ the contour integral may be completed by a semicircle in the upper half-plane. (b) Evaluate $x(t)$ for $t > 0$. Note that $\alpha(\omega)$ has poles at $\pm(\omega_0^2 - \frac{1}{4}\rho^2)^{1/2} - \frac{1}{2}i\rho$, both in the lower half-plane.

2. ***Dissipation sum rule.*** By comparison of $\alpha'(\omega)$ from (9) and from (11a) in the limit $\omega \to \infty$, show that the following sum rule for the oscillator strengths must hold:

$$\Sigma f_j = \frac{2}{\pi}\int_0^{\infty} s\alpha''(s)\, ds \ .$$

3. ***Reflection at normal incidence.*** Consider an electromagnetic wave in vacuum, with field components of the form

$$E_y(\mathrm{inc}) = B_z(\mathrm{inc}) = Ae^{i(kx-\omega t)} \ .$$

Let the wave be incident upon a medium of dielectric constant ϵ and permeability $\mu = 1$ that fills the half-space $x > 0$. Show that the reflectivity coefficient $r(\omega)$ as defined by $E(\mathrm{refl}) = r(\omega)E(\mathrm{inc})$ is given by

$$r(\omega) = \frac{n + iK - 1}{n + iK + 1} \ ,$$

where $n + iK = \epsilon^{1/2}$, with n and K real. Show further that the reflectance is

$$R(\omega) = \frac{(n-1)^2 + K^2}{(n+1)^2 + K^2} \ .$$

*4. ***Conductivity sum rule and superconductivity.*** We write the electrical conductivity as $\sigma(\omega) = \sigma'(\omega) + i\sigma''(\omega)$, where σ', σ'' are real. (a) Show by a Kramers-Kronig relation that

$$\lim_{\omega\to\infty} \omega\, \sigma''(\omega) = \frac{2}{\pi}\int_0^{\infty} \sigma'(s)\, ds \ . \tag{45}$$

*This problem is somewhat difficult.

This result is used in the theory of superconductivity. If at very high frequencies (such as x-ray frequencies) $\sigma''(\omega)$ is identical for the superconducting and normal states, then we must have

$$\int_0^\infty \sigma_s'(\omega)\, d\omega = \int_0^\infty \sigma_n'(\omega)\, d\omega \ .$$

But at frequencies $0 < \omega < \omega_g$ within the superconducting energy gap the real part of the conductivity of a superconductor vanishes, so that in this region the integral on the left-hand side is lower by $\approx \sigma_n'\omega_g$. There must be an additional contribution to σ_s' to balance this deficiency. (b) Show that if $\sigma_s'(\omega < \omega_g) < \sigma_n'(\omega < \omega_g)$, as is observed experimentally, then $\sigma_s'(\omega)$ can have a delta function contribution at $\omega = 0$, and from the delta function there is a contribution $\sigma_s''(\omega) \approx \sigma_n' \omega_g/\omega$. The delta function corresponds to infinite conductivity at zero frequency. (c) By elementary consideration of the classical motion of conduction electrons at very high frequencies, show that

(CGS) $$\int_0^\infty \sigma'(\omega)\, d\omega = \pi n e^2/2m \ ,$$

a result found by Ferrell and Glover.

5. ***Dielectric constant and the semiconductor energy gap.*** The effect on $\epsilon''(\omega)$ of an energy gap ω_g in a semiconductor may be approximated very roughly by substituting $\frac{1}{2}\delta(\omega - \omega_g)$ for $\delta(\omega)$ in the response function (16); that is, we take $\epsilon''(\omega) = (2\pi ne^2/m\omega)\pi\delta(\omega - \omega_g)$. This is crude because it puts all the absorption at the gap frequency. The factor 1/2 enters as soon as we move the delta function away from the origin, because the integral in the sum rule of Problem 2 starts at the origin. Show that the real part of the dielectric constant on this model is

$$\epsilon'(\omega) = 1 + \omega_p^2/(\omega_g^2 - \omega^2) \ , \qquad \omega_p^2 \equiv 4\pi ne^2/m \ .$$

It follows that the static dielectric constant is $\epsilon'(0) = 1 + \omega_p^2/\omega_g^2$, widely used as a rule of thumb.

6. ***Hagen-Rubens relation for infrared reflectivity of metals.*** The complex refractive index $n + iK$ of a metal for $\omega\tau \ll 1$ is given by

(CGS) $$\epsilon(\omega) \equiv (n + iK)^2 = 1 + 4\pi i\sigma_0/\omega \ ,$$

where σ_0 is the conductivity for static fields. We assume here that intraband currents are dominant; interband transitions are neglected. Using the result of Problem 3 for the reflection coefficient at normal incidence, show that

(CGS) $$R \simeq 1 - (2\omega/\pi\sigma_0)^{1/2} \ ,$$

provided that $\sigma_0 \gg \omega$. This is the Hagen-Rubens relation. For sodium at room temperature, $\sigma_0 \simeq 2.1 \times 10^{17}\ \mathrm{s}^{-1}$ in CGS and $\tau = 3.1 \times 10^{-14}$ s, as deduced from

$\tau = \sigma_0 m/ne^2$. Radiation of 10 μm has $\omega = 1.88 \times 10^{14}\ s^{-1}$, so that the Hagen-Rubens result should apply: $R = 0.976$. The result calculated from experimental values of n and K is 0.987. Hint: If $\sigma_0 \gg \omega$, then $n^2 \simeq K^2$. This simplifies the algebra.

*7. ***Davydov splitting of exciton lines.*** The Frenkel exciton band of Fig. 9 is doubled when there are two atoms A, B in a primitive cell. Extend the theory of Eqs. (25) to (29) to a linear crystal AB.AB.AB.AB. with transfer integrals T_1 between AB and T_2 between B.A. Find an equation for the two bands as functions of the wavevector. The splitting between the bands at $k = 0$ is called the Davydov splitting.

*This problem is somewhat difficult.

16

Dielectrics and Ferroelectrics

NOTATION:

(CGS) $\epsilon_0 = 10^7/4\pi c^2$;

$D = E + 4\pi P = \epsilon E = (1 + 4\pi\chi)E$; $\alpha = p/E_{\text{local}}$;

(SI) $D = \epsilon_0 E + P = \epsilon\epsilon_0 E = (1 + \chi)\epsilon_0 E$; $\alpha = p/E_{\text{local}}$;

$\epsilon_{\text{CGS}} = \epsilon_{\text{SI}}$; $4\pi\chi_{\text{CGS}} = \chi_{\text{SI}}$; $\alpha_{\text{SI}} = 4\pi\epsilon_0\alpha_{\text{CGS}}$.

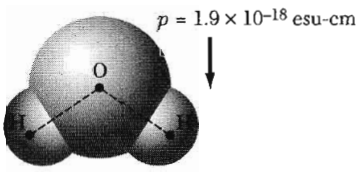


Figure 1 The permanent dipole moment of a molecule of water has the magnitude 1.9×10^{-18} esu-cm and is directed from the O^{2-} ion toward the midpoint of the line connecting the H^+ ions. (To convert to SI units, multiply p by $\frac{1}{3} \times 10^{11}$.)

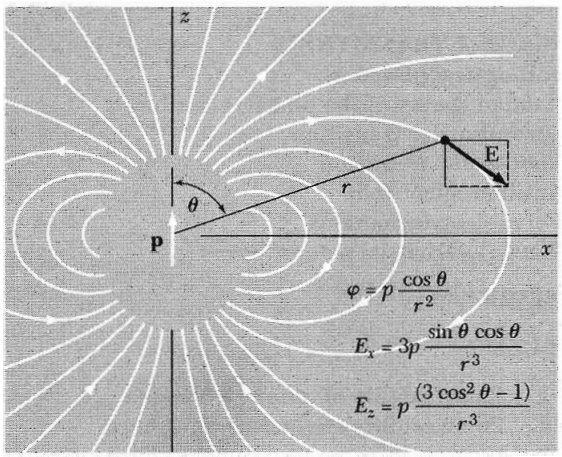


Figure 2 Electrostatic potential and field components in CGS at position r, θ for a dipole **p** directed along the z axis. For $\theta = 0$, we have $E_x = E_y = 0$ and $E_z = 2p/r^3$; for $\theta = \pi/2$ we have $E_x = E_y = 0$ and $E_z = -p/r^3$. To convert to SI, replace p by $p/4\pi\epsilon_0$. (After E. M. Purcell.)

First we relate the applied electric field to the internal electric field in a dielectric crystal. The study of the electric field within dielectric matter arises when we ask:

- What is the relation in the material between the dielectric polarization **P** and the macroscopic electric field **E** in the Maxwell equations?
- What is the relation between the dielectric polarization and the local electric field which acts at the site of an atom in the lattice? The local field determines the dipole moment of the atom.

Maxwell Equations

(CGS)	(SI)
$\text{curl}\,\mathbf{H} = \frac{4\pi}{c}\mathbf{j} + \frac{1}{c}\frac{\partial}{\partial t}(\mathbf{E} + 4\pi\mathbf{P})$;	$\text{curl}\,\mathbf{H} = \mathbf{j} + \frac{\partial}{\partial t}(\epsilon_0\mathbf{E} + \mathbf{P})$;
$\text{curl}\,\mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}$;	$\text{curl}\,\mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$;
$\text{div}\,\mathbf{E} = 4\pi\rho$;	$\text{div}\,\epsilon_0\mathbf{E} = \rho$;
$\text{div}\,\mathbf{B} = 0$;	$\text{div}\,\mathbf{B} = 0$.

Polarization

The **polarization P** is defined as the **dipole moment per unit volume**, averaged over the volume of a cell. The total dipole moment is defined as

$$\mathbf{p} = \Sigma q_n \mathbf{r}_n \ , \tag{1}$$

where $\mathbf{r}_n$ is the position vector of the charge q_n. The value of the sum will be independent of the origin chosen for the position vectors, provided that the system is neutral: Let $\mathbf{r}_n' = \mathbf{r}_n + \mathbf{R}$; then $\mathbf{p} = \Sigma q_n \mathbf{r}_n' = \mathbf{R}\Sigma q_n + \Sigma q_n \mathbf{r}_n = \Sigma q_n \mathbf{r}_n$. The dipole moment of a water molecule is shown in Fig. 1.

The electric field at a point **r** from a dipole moment **p** is given by a standard result of elementary electrostatics:

$$\text{(CGS)}\quad \mathbf{E}(\mathbf{r}) = \frac{3(\mathbf{p}\cdot\mathbf{r})\mathbf{r} - r^2\mathbf{p}}{r^5} \ ; \qquad \text{(SI)}\quad \mathbf{E}(\mathbf{r}) = \frac{3(\mathbf{p}\cdot\mathbf{r})\mathbf{r} - r^2\mathbf{p}}{4\pi\epsilon_0 r^5} \ . \tag{2}$$

The lines of force of a dipole pointing along the z axis are shown in Fig. 2.

MACROSCOPIC ELECTRIC FIELD

One contribution to the electric field inside a body is that of the applied electric field, defined as

$$\boxed{\mathbf{E}_0 \equiv \text{field produced by fixed charges external to the body .}} \tag{3}$$

The other contribution to the electric field is the sum of the fields of all charges that constitute the body. If the body is neutral, the contribution to the average field may be expressed in terms of the sum of the fields of atomic dipoles.

We define the average electric field $\mathbf{E}(\mathbf{r}_0)$ as the **average field over the volume of the crystal cell** that contains the lattice point $\mathbf{r}_0$:

$$\mathbf{E}(\mathbf{r}_0) = \frac{1}{V_c}\int dV\, \mathbf{e}(\mathbf{r}) \ , \tag{4}$$

where $\mathbf{e}(\mathbf{r})$ is the microscopic electric field at the point $\mathbf{r}$. The field $\mathbf{E}$ is a much smoother quantity than the microscopic field $\mathbf{e}$. We could well have written the dipole field (2) as $\mathbf{e}(\mathbf{r})$ because it is a microscopic unsmoothed field.

We call $\mathbf{E}$ the **macroscopic electric field**. It is adequate for all problems in the electrodynamics of crystals provided that we know the connection between $\mathbf{E}$, the polarization $\mathbf{P}$, and the current density $\mathbf{j}$, and provided that the wavelengths of interest are long in comparison with the lattice spacing.[1]

To find the contribution of the polarization to the macroscopic field, we can simplify the sum over all the dipoles in the specimen. By a famous theorem of electrostatics[2] the macroscopic electric field caused by a uniform polarization is equal to the electric field in vacuum of a fictitious surface charge

[1]A detailed derivation of the Maxwell equations for the macroscopic fields $\mathbf{E}$ and $\mathbf{B}$, starting from the Maxwell equations in terms of the microscopic fields $\mathbf{e}$ and $\mathbf{h}$, is given by E. M. Purcell, *Electricity and magnetism*, 2nd ed., McGraw-Hill, 1985.

[2]The electrostatic potential in CGS units of a dipole $\mathbf{p}$ is $\varphi(\mathbf{r}) = \mathbf{p} \cdot \text{grad}(1/r)$. For a volume distribution of polarization $\mathbf{P}$ we have

$$\varphi(\mathbf{r}) = \int dV \left(\mathbf{P} \cdot \text{grad}\,\frac{1}{r}\right) ,$$

which by a vector identity becomes

$$\varphi(\mathbf{r}) = \int dV \left(-\frac{1}{r}\,\text{div}\,\mathbf{P} + \text{div}\,\frac{\mathbf{P}}{r}\right) .$$

If $\mathbf{P}$ is constant, then div $\mathbf{P} = 0$ and by the Gauss theorem we have

$$\varphi(\mathbf{r}) = \int dS\,\frac{P_n}{r} = \int dS\,\frac{\sigma}{r} \ ,$$

where σdS is an element of charge on the surface of the body. This completes the proof.

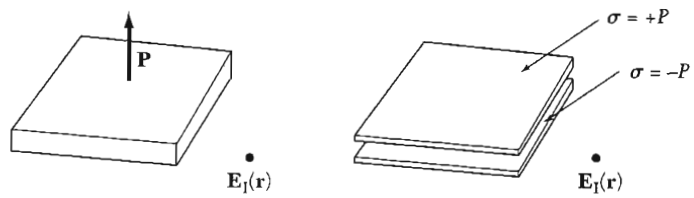


Figure 3 (a) A uniformly polarized dielectric slab, with the polarization vector **P** normal to the plane of the slab. (b) A pair of uniformly charged parallel plates which give rise to the identical electric field $\mathbf{E}_1$ as in (a). The upper plate has the surface charge density $\sigma = +P$, and the lower plate has $\sigma = -P$.

density $\sigma = \hat{\mathbf{n}} \cdot \mathbf{P}$ on the surface of the body. Here $\hat{\mathbf{n}}$ is the unit normal to the surface, drawn outward from the polarized matter.

We apply the result to a thin dielectric slab (Fig. 3a) with a uniform volume polarization **P**. The electric field $\mathbf{E}_1(\mathbf{r})$ produced by the polarization is equal to the field produced by the fictitious surface charge density $\sigma = \hat{\mathbf{n}} \cdot \mathbf{P}$ on the surface of the slab. On the upper boundary the unit vector $\hat{\mathbf{n}}$ is directed upward and on the lower boundary $\hat{\mathbf{n}}$ is directed downward. The upper boundary bears the fictitious charge $\sigma = \hat{\mathbf{n}} \cdot \mathbf{P} = P$ per unit area, and the lower boundary bears $-P$ per unit area.

The electric field $\mathbf{E}_1$ due to these charges has a simple form at any point between the plates, but comfortably removed from their edges. By Gauss's law

$$\text{(CGS)} \quad E_1 = -4\pi|\sigma| = -4\pi P \ ; \qquad \text{(SI)} \quad E_1 = -\frac{|\sigma|}{\epsilon_0} = \frac{P}{\epsilon_0} \ . \tag{4a}$$

We add $\mathbf{E}_1$ to the applied field $\mathbf{E}_0$ to obtain the total macroscopic field inside the slab, with $\hat{\mathbf{z}}$ the unit vector normal to the plane of the slab:

$$\text{(CGS)} \qquad \mathbf{E} = \mathbf{E}_0 + \mathbf{E}_1 = \mathbf{E}_0 - 4\pi P\hat{\mathbf{z}} \ ; \tag{5}$$

$$\text{(SI)} \qquad \mathbf{E} = \mathbf{E}_0 + \mathbf{E}_1 = \mathbf{E}_0 - \frac{P}{\epsilon_0}\hat{\mathbf{z}} \ .$$

We define

$$\boxed{\mathbf{E}_1 \equiv \text{field of the surface charge denisty } \hat{\mathbf{n}} \cdot \mathbf{P} \text{ on the boundary} \ .} \tag{6}$$

This field is smoothly varying in space inside and outside the body and satisfies the Maxwell equations as written for the macroscopic field **E**. The reason $\mathbf{E}_1$ is a smooth function when viewed on an atomic scale is that we have replaced the discrete lattice of dipoles $\mathbf{p}_j$ with the smoothed polarization **P**.

Depolarization Field, $\mathbf{E}_1$

If the polarization is uniform within the body, the only contributions to the macroscopic field are from $\mathbf{E}_0$ and $\mathbf{E}_1$:

$$\boxed{\mathbf{E} = \mathbf{E}_0 + \mathbf{E}_1 \ .} \tag{7}$$

Here $\mathbf{E}_0$ is the applied field and $\mathbf{E}_1$ is the field due to the uniform polarization.

The field $\mathbf{E}_1$ is called the **depolarization field**, for within the body it tends to oppose the applied field $\mathbf{E}_0$ as in Fig. 4. Specimens in the shape of ellipsoids, a class that includes spheres, cylinders, and discs as limiting forms, have an advantageous property: a uniform polarization produces a uniform depolarization field inside the body. This is a famous mathematical result demonstrated in classic texts on electricity and magnetism.[3]

If P_x, P_y, P_z are the components of the polarization $\mathbf{P}$ referred to the principal axes of an ellipsoid, then the components of the depolarization field are written

$$\text{(CGS)} \qquad E_{1x} = -N_x P_x \ ; \qquad E_{1y} = -N_y P_y \ ; \qquad E_{1z} = -N_z P_z \ ; \tag{8}$$

$$\text{(SI)} \qquad E_{1x} = -\frac{N_x P_x}{\epsilon_0} \ ; \qquad E_{1y} = -\frac{N_y P_y}{\epsilon_0} \ ; \qquad E_{1z} = -\frac{N_z P_z}{\epsilon_0} \ .$$

Here N_x, N_y, N_z are the **depolarization factors**; their values depend on the ratios of the principal axes of the ellipsoid. The N's are positive and satisfy the sum rule $N_x + N_y + N_z = 4\pi$ in CGS, and $N_x + N_y + N_z = 1$ in SI.

Values of N parallel to the figure axis of ellipsoids of revolution are plotted in Fig. 5. Additional cases have been calculated by Osborn[4] and by Stoner. In limiting cases N has the values:

Shape	Axis	N (CGS)	N (SI)
Sphere	any	$4\pi/3$	1/3
Thin slab	normal	4π	1
Thin slab	in plane	0	0
Long circular cylinder	longitudinal	0	0
Long circular cylinder	transverse	2π	1/2

We can reduce the depolarization field to zero in two ways, either by working parallel to the axis of a long fine specimen or by making an electrical connection between electrodes deposited on the opposite surfaces of a thin slab.

[3]R. Becker, *Electromagnetic fields and interactions*, Blaisdell, 1964, pp. 102–107.

[4]J. A. Osborn, Phys. Rev. **67**, 351 (1945); E. C. Stoner, Philosophical Magazine **36**, 803 (1945).

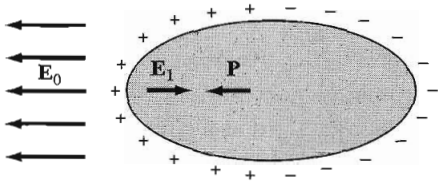


Figure 4 The depolarization field $\mathbf{E}_1$ is opposite to **P**. The fictitious surface charges are indicated: the field of these charges is $\mathbf{E}_1$ within the ellipsoid.

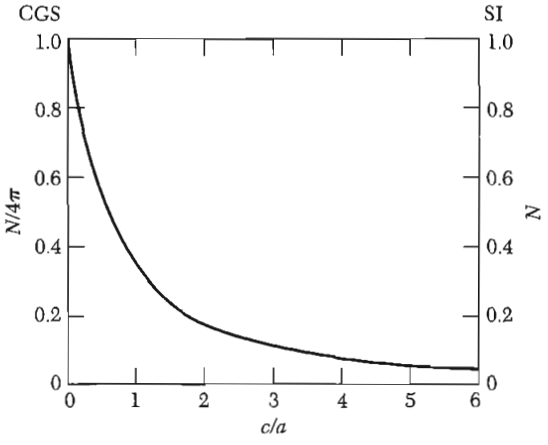


Figure 5 Depolarization factor N parallel to the figure axis of ellipsoids of revolution, as a function of the axial ratio c/a.

A uniform applied field $\mathbf{E}_0$ will induce uniform polarization in an ellipsoid. We introduce the **dielectric susceptibility** χ such that the relations

$$\text{(CGS)} \quad \mathbf{P} = \chi \mathbf{E} \ ; \qquad \text{(SI)} \quad \mathbf{P} = \epsilon_0 \chi \mathbf{E} \ , \tag{9}$$

connect the macroscopic field **E** inside the ellipsoid with the polarization **P**. Here $\chi_{\text{SI}} = 4\pi\chi_{\text{CGS}}$.

If $\mathbf{E}_0$ is uniform and parallel to a principal axis of the ellipsoid, then

$$\text{(CGS)} \quad E = E_0 + E_1 = E_0 - NP \ ; \qquad \text{(SI)} \ E = E_0 - \frac{NP}{\epsilon_0} \ , \tag{10}$$

by (8), whence

$$\text{(CGS)} \qquad P = \chi(E_0 - NP) \ ; \qquad P = \frac{\chi}{1 + N\chi} E_0 \ ; \tag{11}$$

$$\text{(SI)} \qquad P = \chi(\epsilon_0 E_0 - NP) \ ; \qquad P = \frac{\chi \epsilon_0}{1 + N\chi} E_0 \ .$$

The value of the polarization depends on the depolarization factor N.

LOCAL ELECTRIC FIELD AT AN ATOM

The value of the local electric field that acts at the site of an atom is significantly different from the value of the macroscopic electric field. We can convince ourselves of this by consideration of the local field at a site with a cubic arrangement of neighbors[5] in a crystal of spherical shape. The macroscopic electric field in a sphere is

(CGS) $$\mathbf{E} = \mathbf{E}_0 + \mathbf{E}_1 = \mathbf{E}_0 - \frac{4\pi}{3}\mathbf{P} \; ; \tag{12}$$

(SI) $$\mathbf{E} = \mathbf{E}_0 + \mathbf{E}_1 = \mathbf{E}_0 - \frac{1}{3\epsilon_0}\mathbf{P} \; ,$$

by (10).

But consider the field that acts on the atom at the center of the sphere (this atom is not unrepresentative). If all dipoles are parallel to the z axis and have magnitude p, the z component of the field at the center due to all other dipoles is, from (2),

(CGS) $$E_{\text{dipole}} = p \sum_i \frac{3z_i^2 - r_i^2}{r_i^5} = p \sum_i \frac{2z_i^2 - x_i^2 - y_i^2}{r_i^5} \; . \tag{13}$$

In SI we replace p by $p/4\pi\epsilon_0$. The x, y, z directions are equivalent because of the symmetry of the lattice and of the sphere; thus

$$\sum_i \frac{z_i^2}{r_i^5} = \sum_i \frac{x_i^2}{r_i^5} = \sum_i \frac{y_i^2}{r_i^5} \; ,$$

whence $E_{\text{dipole}} = 0$.

The correct local field is just equal to the applied field, $\mathbf{E}_{\text{local}} = \mathbf{E}_0$, for an atom site with a cubic environment in a spherical specimen. Thus the local field is not the same as the macroscopic average field $\mathbf{E}$.

We now develop an expression for the local field at a general lattice site, not necessarily of cubic symmetry. The local field at an atom is the sum of the electric field $\mathbf{E}_0$ from external sources and of the field from the dipoles within the specimen. It is convenient to decompose the dipole field so that part of the summation over dipoles may be replaced by integration.

We write

$$\boxed{\mathbf{E}_{\text{local}} = \mathbf{E}_0 + \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 \; .} \tag{14}$$

[5]Atom sites in a cubic crystal do not necessarily have cubic symmetry: thus the O^{2-} sites in the barium titanate structure of Fig. 10 do not have a cubic environment. However, the Na^+ and Cl^- sites in the NaCl structure and the Cs^+ and Cl^- sites in the CsCl structure have cubic symmetry.

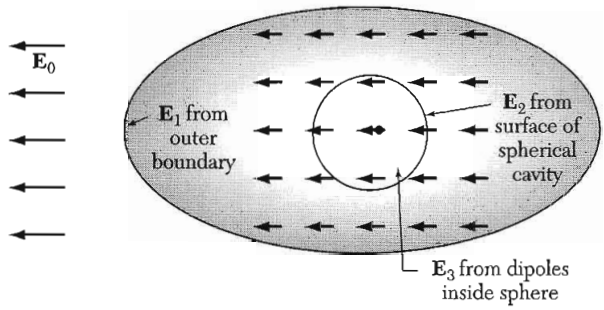


Figure 6 The internal electric field on an atom in a crystal is the sum of the external applied field $\mathbf{E}_0$ and of the field due to the other atoms in the crystal. The standard method of summing the dipole fields of the other atoms is first to sum individually over a moderate number of neighboring atoms inside an imaginary sphere concentric with the reference atom: this defines the field $\mathbf{E}_3$, which vanishes at a reference site with cubic symmetry. The atoms outside the sphere can be treated as a uniformly polarized dielectric. Their contribution to the field at the reference point is $\mathbf{E}_1 + \mathbf{E}_2$, where $\mathbf{E}_1$ is the depolarization field associated with the outer boundary and $\mathbf{E}_2$ is the field associated with the surface of the spherical cavity.

Here

$\mathbf{E}_0$ = field produced by fixed charges external to the body;

$\mathbf{E}_1$ = depolarization field, from a surface charge density $\hat{\mathbf{n}} \cdot \mathbf{P}$ on the outer surface of the specimen;

$\mathbf{E}_2$ = Lorentz cavity field: field from polarization charges on inside of a spherical cavity cut (as a mathematical fiction) out of the specimen with the reference atom as center, as in Fig. 6; $\mathbf{E}_1 + \mathbf{E}_2$ is the field due to uniform polarization of the body in which a hole has been created;

$\mathbf{E}_3$ = field of atoms inside cavity.

The contribution $\mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3$ to the local field is the total field at one atom caused by the dipole moments of all the other atoms in the specimen:

$$\text{(CGS)} \qquad \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 = \sum_i \frac{3(\mathbf{p}_i \cdot \mathbf{r}_i)\mathbf{r}_i - r_i^2 \mathbf{p}_i}{r_i^5} , \qquad (15)$$

and in SI we replace $\mathbf{p}_i$ by $\mathbf{p}_i/4\pi\epsilon_0$.

Dipoles at distances greater than perhaps ten lattice constants from the reference site make a smoothly varying contribution to this sum, a contribution which may be replaced by two surface integrals. One surface integral is taken over the outer surface of the ellipsoidal specimen and defines $\mathbf{E}_1$, as in Eq. (6). The second surface integral defines $\mathbf{E}_2$ and may be taken over any interior surface that is a suitable distance (say 50 Å) from the reference site. We count in $\mathbf{E}_3$ any dipoles not included in the volume bounded by the inner and outer surfaces. It is convenient to let the interior surface be spherical.

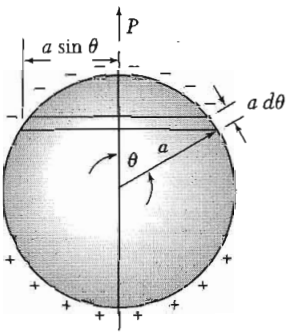


Figure 7 Calculation of the field in a spherical cavity in a uniformly polarized medium.

Charge on ring = $2\pi a \sin\theta \cdot a\, d\theta \cdot P \cos\theta$

Lorentz Field, $\mathbf{E}_2$

The field $\mathbf{E}_2$ due to the polarization charges on the surface of the fictitious cavity was calculated by Lorentz. If θ is the polar angle (Fig. 7) referred to the polarization direction, the surface charge density on the surface of the cavity is $-P \cos\theta$. The electric field at the center of the spherical cavity of radius a is

$$\text{(CGS)} \qquad \mathbf{E}_2 = \int_0^{\pi} (a^{-2})(2\pi a \sin\theta)(a\, d\theta)(\mathbf{P}\cos\theta)(\cos\theta) = \frac{4\pi}{3}\mathbf{P}\ ; \tag{16}$$

$$\text{(SI)} \qquad \mathbf{E}_2 = \frac{1}{3\epsilon_0}\mathbf{P}\ .$$

This is the negative of the depolarization field $\mathbf{E}_1$ in a polarized sphere, so that $\mathbf{E}_1 + \mathbf{E}_2 = 0$ for a sphere.

Field of Dipoles Inside Cavity, $\mathbf{E}_3$

The field $\mathbf{E}_3$ due to the dipoles within the spherical cavity is the only term that depends on the crystal structure. We showed for a reference site with cubic surroundings in a sphere that $\mathbf{E}_3 = 0$ if all the atoms may be replaced by point dipoles parallel to each other. The total local field at a cubic site is, from (14) and (16),

$$\text{(CGS)} \qquad \mathbf{E}_{\text{local}} = \mathbf{E}_0 + \mathbf{E}_1 + \frac{4\pi}{3}\mathbf{P} = \mathbf{E} + \frac{4\pi}{3}\mathbf{P}\ ; \tag{17}$$

$$\text{(SI)} \qquad \mathbf{E}_{\text{local}} = \mathbf{E} + \frac{1}{3\epsilon_0}\mathbf{P}\ .$$

This is the **Lorentz relation**: the field acting at an atom in a cubic site is the macroscopic field $\mathbf{E}$ of Eq. (7) plus $4\pi\mathbf{P}/3$ or $\mathbf{P}/3\epsilon_0$ from the polarization of the other atoms in the specimen. Experimental data for cubic ionic crystals support the Lorentz relation.

DIELECTRIC CONSTANT AND POLARIZABILITY

The **dielectric constant** ϵ of an isotropic or cubic medium relative to vacuum is defined in terms of the macroscopic field E:

(CGS) $$\epsilon = \frac{E + 4\pi P}{E} = 1 + 4\pi\chi \ ; \tag{18}$$

(SI) $$\epsilon = \frac{\epsilon_0 E + P}{\epsilon_0 E} = 1 + \chi \ .$$

Remember that $\chi_{SI} = 4\pi\chi_{CGS}$, by definition, but $\epsilon_{SI} = \epsilon_{CGS}$.

The susceptibility (9) is related to the dielectric constant by

(CGS) $$\chi = \frac{P}{E} = \frac{\epsilon - 1}{4\pi} \ ; \qquad \text{(SI)} \ \chi = \frac{P}{\epsilon_0 E} = \epsilon - 1 \ . \tag{19}$$

In a noncubic crystal the dielectric response is described by the components of the susceptibility tensor or of the dielectric constant tensor:

(CGS) $$P_\mu = \chi_{\mu\nu} E_\nu \ ; \quad \epsilon_{\mu\nu} = \delta_{\mu\nu} + 4\pi\chi_{\mu\nu} \ . \tag{20}$$

(SI) $$P_\mu = \chi_{\mu\nu}\epsilon_0 E_\nu \ ; \quad \epsilon_{\mu\nu} = \delta_{\mu\nu} + \chi_{\mu\nu} \ .$$

The **polarizability** α of an atom is defined in terms of the local electric field at the atom:

$$p = \alpha E_{\text{local}} \ , \tag{21}$$

where p is the dipole moment. This definition applies in CGS and in SI, but $\alpha_{SI} = 4\pi\epsilon_0\alpha_{CGS}$. The polarizability is an atomic property, but the dielectric constant will depend on the manner in which the atoms are assembled to form a crystal. For a nonspherical atom α will be a tensor.

The polarization of a crystal may be expressed approximately as the product of the polarizabilities of the atoms times the local electric field:

$$P = \sum_j N_j p_j = \sum_j N_j \alpha_j E_{\text{loc}}(j) \ , \tag{22}$$

where N_j is the concentration and α_j the polarizability of atoms j, and $E_{\text{loc}}(j)$ is the local field at atom sites j.

We want to relate the dielectric constant to the polarizabilities; the result will depend on the relation that holds between the macroscopic electric field and the local electric field. We give the derivation in CGS units and state the result in both systems of units.

If the local field is given by the Lorentz relation (17), then

$$\text{(CGS)} \qquad P = (\Sigma N_j \alpha_j)\left(E + \frac{4\pi}{3}P\right) ;$$

and we solve for P to find the susceptibility

$$\text{(CGS)} \qquad \chi = \frac{P}{E} = \frac{\Sigma N_j \alpha_j}{1 - \frac{4\pi}{3}\Sigma N_j \alpha_j} . \tag{23}$$

By definition $\epsilon = 1 + 4\pi\chi$ in CGS; we may rearrange (23) to obtain

$$\text{(CGS)} \quad \frac{\epsilon - 1}{\epsilon + 2} = \frac{4\pi}{3}\Sigma N_j \alpha_j ; \qquad \text{(SI)} \quad \frac{\epsilon - 1}{\epsilon + 2} = \frac{1}{3\epsilon_0}\Sigma N_j \alpha_j , \tag{24}$$

the **Clausius-Mossotti relation**. This relates the dielectric constant to the electronic polarizability, but only for crystal structures for which the Lorentz local field (17) obtains.

Electronic Polarizability

The total polarizability may usually be separated into three parts: electronic, ionic, and dipolar, as in Fig. 8. The electronic contribution arises from the displacement of the electron shell relative to a nucleus. The ionic contribution comes from the displacement of a charged ion with respect to other ions. The dipolar polarizability arises from molecules with a permanent electric dipole moment that can change orientation in an applied electric field.

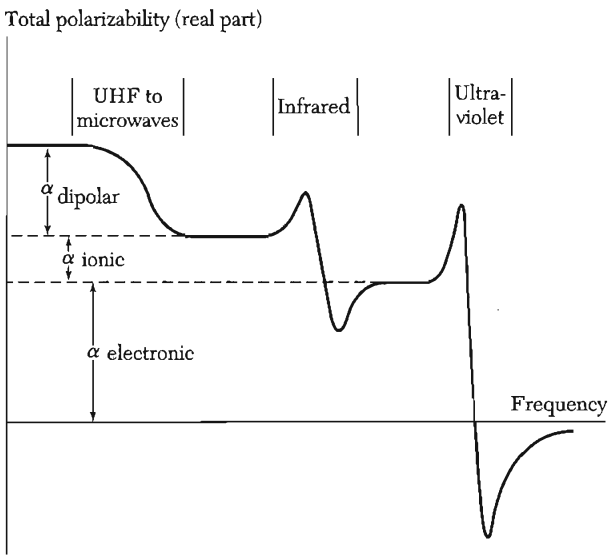


Figure 8 Frequency dependence of the several contributions to the polarizability.

In heterogeneous materials there is usually also an interfacial polarization arising from the accumulation of charge at structural interfaces. This is of little fundamental interest, but it is of considerable practical interest because commercial insulating materials are usually heterogeneous.[6]

The dielectric constant at optical frequencies arises almost entirely from the electronic polarizability. The dipolar and ionic contributions are small at high frequencies because of the inertia of the molecules and ions. In the optical range (24) reduces to

$$\text{(CGS)} \qquad \frac{n^2-1}{n^2+2} = \frac{4\pi}{3}\Sigma N_j\alpha_j(\text{electronic}) \ ; \qquad (25)$$

here we have used the relation $n^2 = \epsilon$, where n is the refractive index.

By applying (25) to large numbers of crystals we determine in Table 1 empirical values of the electronic polarizabilities that are reasonably consistent with the observed values of the refractive index. The scheme is not entirely self-consistent, because the electronic polarizability of an ion depends somewhat on the environment in which it is placed. The negative ions are highly polarizable because they are large.

Table 1 Electronic polarizabilities of atoms and ions, in 10^{-24} cm^3

			Hc	Li^+	Be^{2+}	B^{3+}	C^{4+}
Pauling			0.201	0.029	0.008	0.003	0.0013
JS				0.029			
	O^{2-}	F^-	Ne	Na^+	Mg^{2+}	Al^{3+}	Si^{4+}
Pauling	3.88	1.04	0.390	0.179	0.094	0.052	0.0165
JS-(TKS)	(2.4)	0.858		0.290			
	S^{2-}	Cl^-	Ar	K^+	Ca^{2+}	Se^{3+}	Ti^{4+}
Pauling	10.2	3.66	1.62	0.83	0.47	0.286	0.185
JS-(TKS)	(5.5)	2.947		1.133	(1.1)		(0.19)
	Se^{2-}	Br^-	Kr	Rb^+	Sr^{2+}	Y^{3+}	Zr^{4+}
Pauling	10.5	4.77	2.46	1.40	0.86	0.55	0.37
JS-(TKS)	(7.)	4.091		1.679	(1.6)		
	Te^{2-}	I^-	Xe	Cs^+	Ba^{2+}	La^{3+}	Ce^{4+}
Pauling	14.0	7.10	3.99	2.42	1.55	1.04	0.73
JS-(TKS)	(9.)	6.116		2.743	(2.5)		

Values from L. Pauling, Proc. R. Soc. London **A114**, 181 (1927); S. S. Jaswal and T. P. Sharma, J. Phys. Chem. Solids **34**, 509 (1973); and J. Tessman, A. Kahn, and W. Shockley, Phys. Rev. **92**, 890 (1953). The TKS polarizabilities are at the frequency of the D lines of sodium. The values are in CGS; to convert to SI, multiply by 9×10^{-15}.

[6]For references see D. E. Aspnes, Am. J. Phys. **50**, 704 (1982).

Classical Theory of Electronic Polarizability. An electron bound harmonically to an atom will show resonance absorption at a frequency $\omega_0 = (\beta/m)^{1/2}$, where β is the force constant. The displacement x of the electron occasioned by the application of a field E_{loc} is given by

$$-eE_{loc} = \beta x = m\omega_0^2 x \ , \tag{26}$$

so that the static electronic polarizability is

$$\alpha(\text{electronic}) = p/E_{loc} = -ex/E_{loc} = e^2/m\omega_0^2 \ . \tag{27}$$

The electronic polarizability will depend on frequency, and it is shown in the following example that for frequency ω

$$\text{(CGS)} \qquad \alpha(\text{electronic}) = \frac{e^2/m}{\omega_0^2 - \omega^2} \ ; \tag{28}$$

but in the visible region the frequency dependence (dispersion) is not usually very important in most transparent materials.

EXAMPLE: *Frequency dependence.* Find the frequency dependence of the electronic polarizability of an electron having the resonance frequency ω_0, treating the system as a simple harmonic oscillator.

The equation of motion in the local electric field $E_{loc} \sin \omega t$ is

$$m\frac{d^2x}{dt^2} + m\omega_0^2 x = -eE_{loc} \sin \omega t \ ,$$

so that, for $x = x_0 \sin \omega t$,

$$m(-\omega^2 + \omega_0^2)x_0 = -eE_{loc} \ .$$

The dipole moment has the amplitude

$$p_0 = -ex_0 = \frac{e^2 E_{loc}}{m(\omega_0^2 - \omega^2)} \ ,$$

from which (28) follows.

In quantum theory the expression corresponding to (28) is

$$\text{(CGS)} \qquad \alpha(\text{electronic}) = \frac{e^2}{m} \sum_j \frac{f_{ij}}{\omega_{ij}^2 - \omega^2} \ , \tag{29}$$

where f_{ij} is called the **oscillator strength** of the electric dipole transition between the atomic states i and j. Near a transition the polarizability changes sign (Fig. 8).

STRUCTURAL PHASE TRANSITIONS

It is not uncommon for crystals to transform from one crystal structure to another as the temperature or pressure is varied. The stable structure A at absolute zero generally has the lowest accessible internal energy of all the possible structures. Even this selection of a structure A can be varied with application of pressure, because a low atomic volume will favor closest-packed or even metallic structures. Hydrogen and xenon, for example, become metallic under extreme pressure.

Some other structure B may have a softer or lower frequency phonon spectrum than A. As the temperature is increased the phonons in B will be more highly excited (higher thermal average occupancies) than the phonons in A. Because the entropy increases with the occupancy, the entropy of B will become higher than the entropy of A as the temperature is increased.

It is thereby possible for the stable structure to transform from A to B as the temperature is increased. The stable structure at a temperature T is determined by the minimum of the free energy $F = U - TS$. There will be a transition from A to B if a temperature T_c exists (below the melting point) such that $F_A(T_c) = F_B(T_c)$.

Often several structures have nearly the same internal energy at absolute zero. The phonon dispersion relations for the structures may, however, be rather different. The phonon energies are sensitive to the number and arrangement of nearby atoms; these are the quantities that change as the structure is changed.

Some structural phase transitions have only small effects on the macroscopic physical properties of the material. However, if the transition is influenced by an applied stress, the crystal may yield mechanically quite easily near the transition temperature because the relative proportions in the two phases will change under stress. Some other structural phase transitions may have spectacular effects on the macroscopic electrical properties.

Ferroelectric transitions are a subgroup of structural phase transitions, a subgroup marked by the appearance of a spontaneous dielectric polarization in the crystal. Ferroelectrics are of theoretical and technical interest because they often have unusually high and unusually temperature-dependent values of the dielectric constant, the piezoelectric effect, the pyroelectric effect, and electro-optical effects, including optical frequency doubling.

FERROELECTRIC CRYSTALS

A ferroelectric crystal exhibits an electric dipole moment even in the absence of an external electric field. In the ferroelectric state the center of positive charge of the crystal does not coincide with the center of negative charge.

The plot of polarization versus electric field for the ferroelectric state shows a hysteresis loop. A crystal in a normal dielectric state usually does not

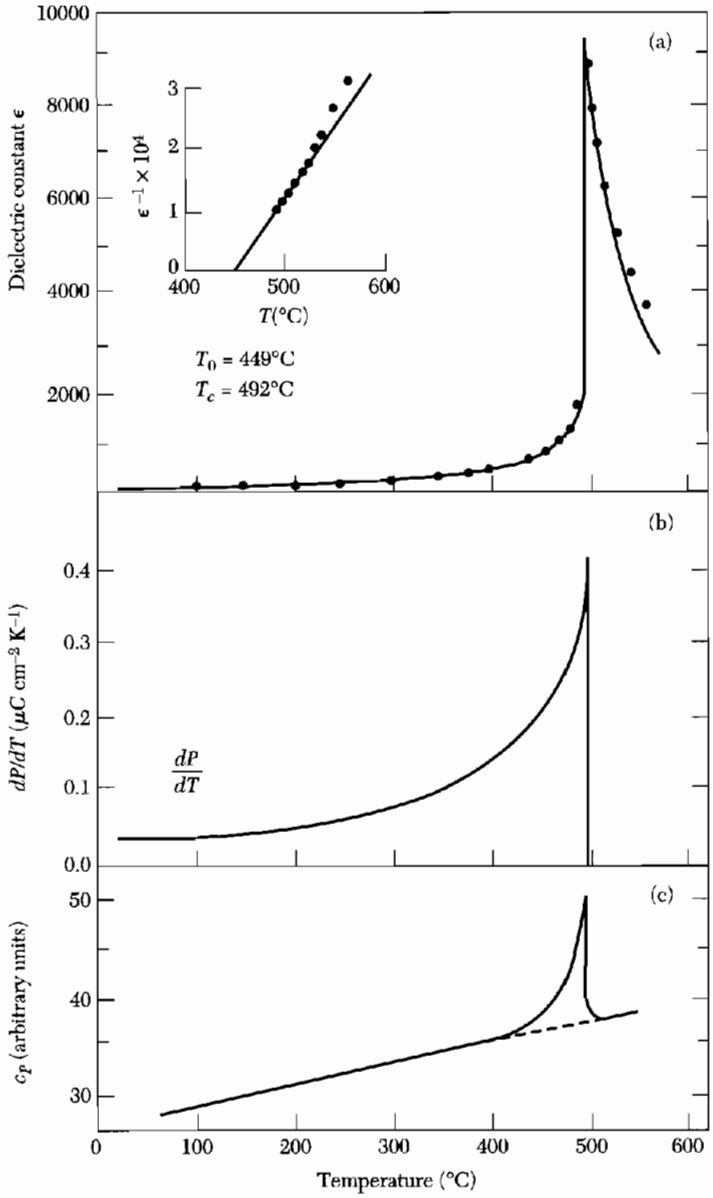


Figure 9 The temperature variation of (a) the dielectric constant ϵ, (b) the pyroelectric coefficient dP/dT, and (c) the specific heat c_p, of $PbTiO_3$. (After Remeika and Class.)

show significant hysteresis when the electric field is increased and then reversed, both slowly.

Ferroelectricity usually disappears above a certain temperature called the transition temperature. Above the transition the crystal is said to be in a **paraelectric** state. The term paraelectric suggests an analogy with paramagnetism:

there is usually a rapid drop in the dielectric constant as the temperature increases.

In some crystals the ferroelectric dipole moment is not changed by an electric field of the maximum intensity which it is possible to apply before causing electrical breakdown. In these crystals we are often able to observe a change in the spontaneous moment when the temperature is changed (Fig. 9). Such crystals are called **pyroelectric**. Lithium niobate, $LiNbO_3$, is pyroelectric at room temperature. It has a high transition temperature (T_c = 1480 K) and a high saturation polarization (50 $\mu C/cm^2$). It can be "poled," which means given a remanent polarization, by an electric field applied over 1400 K.

Classification of Ferroelectric Crystals

We list in Table 2 some of the crystals commonly considered to be ferroelectric, along with the transition temperature or Curie point T_c at which the crystal changes from the low-temperature polarized state to the high-temperature unpolarized state. Thermal motion tends to destroy the ferroelectric order. Some ferroelectric crystals have no Curie point because they melt before leaving the ferroelectric phase. The table also includes values of the spontaneous polarization P_s. Ferroelectric crystals may be classified into two main groups, order-disorder or displacive.

One may define the character of the transition in terms of the dynamics of the lowest frequency ("soft") optical phonon modes. If a soft mode can propagate in the crystal at the transition, then the transition is displacive. If the soft mode is only diffusive (non-propagating) there is really not a phonon at all,

Table 2 Ferroelectric crystals

To obtain the spontaneous polarization P_s in the CGS unit of esu cm^{-2}, multiply the value given in $\mu C\ cm^{-2}$ by 3×10^3.

		T_c, in K	P_s, in $\mu C\ cm^{-2}$, at T K	
KDP type	KH_2PO_4	123	4.75	[96]
	KD_2PO_4	213	4.83	[180]
	RbH_2PO_4	147	5.6	[90]
	KH_2AsO_4	97	5.0	[78]
	GeTe	670	—	—
TGS type	Tri-glycine sulfate	322	2.8	[29]
	Tri-glycine selenate	295	3.2	[283]
Perovskites	$BaTiO_3$	408	26.0	[296]
	$KNbO_3$	708	30.0	[523]
	$PbTiO_3$	765	>50	[296]
	$LiTaO_3$	938	50	
	$LiNbO_3$	1480	71	[296]

but is only a large amplitude hopping motion between the wells of the order-disorder system. Many ferroelectrics have soft modes that fall between these two extremes.

The order-disorder class of ferroelectrics includes crystals with hydrogen bonds in which the motion of the protons is related to the ferroelectric properties, as in potassium dihydrogen phosphate (KH_2PO_4) and isomorphous salts. The substitution of deuterons for protons nearly doubles T_c, although the fractional change in the molecular weight of the compound is less than 2 percent:

	KH_2PO_4	KD_2PO_4	KH_2AsO_4	KD_2AsO_4
Curie temperature	123 K	213 K	97 K	162 K

This extraordinarily large isotope shift is believed to be a quantum effect involving the mass-dependence of the de Broglie wavelength. Neutron diffraction data show that above the Curie temperature the proton distribution along the hydrogen bond is symmetrically elongated. Below the Curie temperature the distribution is more concentrated and asymmetric with respect to neighboring ions, so that one end of the hydrogen bond is preferred by the proton over the other end, giving a polarization.

The displacive class of ferroelectrics includes ionic crystal structures closely related to the perovskite and ilmenite structures. The simplest ferroelectric crystal is GeTe with the sodium chloride structure. We shall devote ourselves primarily to crystals with the perovskite structure, Fig. 10.

Consider the order of magnitude of the ferroelectric effects in barium titanate: the observed saturation polarization P_s at room temperature (Fig. 11)

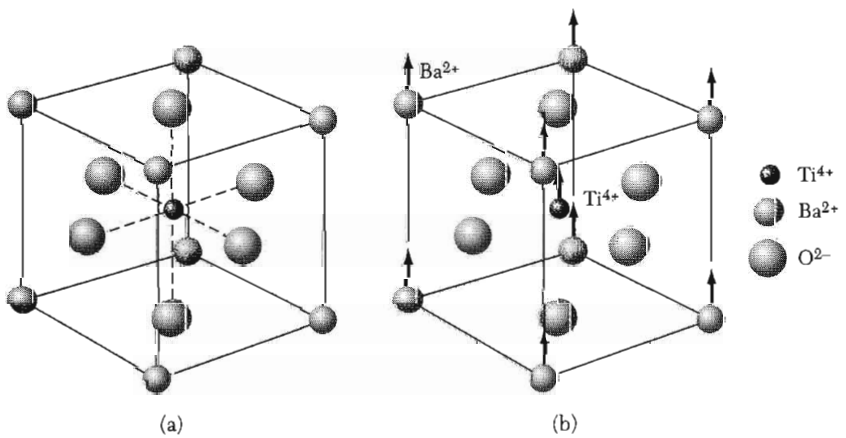


Figure 10 (a) The crystal structure of barium titanate. The prototype crystal is calcium titanate (perovskite). The structure is cubic, with Ba^{2+} ions at the cube corners, O^{2-} ions at the face centers, and a Ti^{4+} ion at the body center. (b) Below the Curie temperature the structure is slightly deformed, with Ba^{2+} and Ti^{4+} ions displaced relative to the O^{2-} ions, thereby developing a dipole moment. The upper and lower oxygen ions may move downward slightly.

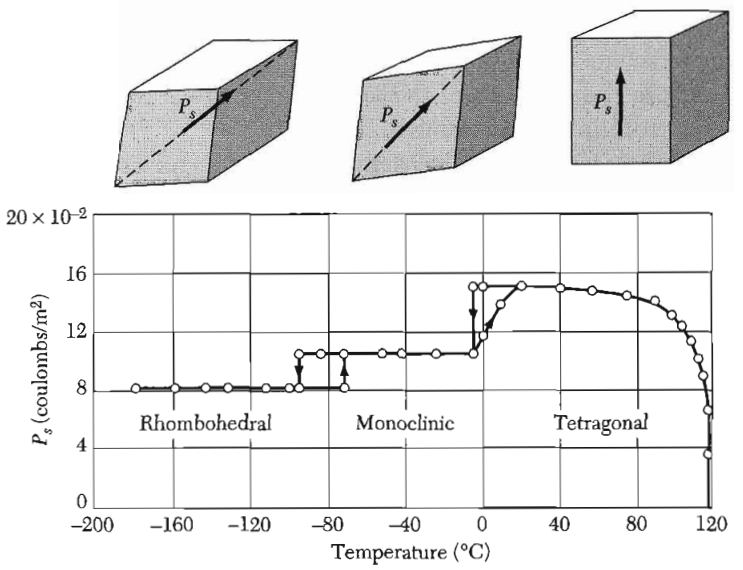


Figure 11 Spontaneous polarization projected on cube edge of barium titanate, as a function of temperature. (After W. J. Merz.)

is 8×10^4 esu cm^{-2}. The volume of a cell is $(4 \times 10^{-8})^3 = 64 \times 10^{-24}$ cm^3, so that the dipole moment of a cell is

(CGS) $p \cong (8 \times 10^4 \text{ esu cm}^{-2})(64 \times 10^{-24} \text{ cm}^{-3}) \cong 5 \times 10^{-18}$ esu cm ;

(SI) $p \cong (3 \times 10^{-1} \text{ C m}^{-2})(64 \times 10^{-30} \text{ m}^3) \cong 2 \times 10^{-29}$ C m .

If the positive ions Ba^{2+} and Ti^{4+} were moved by $\delta = 0.1$ Å with respect to the negative O^{2-} ions, the dipole moment of a cell would be $6e\delta \cong 3 \times 10^{-18}$ esu cm. In $LiNbO_3$ the displacements are considerably larger, being 0.9 Å and 0.5 Å for the lithium and niobum ions respectively, giving the larger P_s.

DISPLACIVE TRANSITIONS

Two viewpoints contribute to an understanding of a ferroelectric displacive transition and by extension to displacive transitions in general. We may speak of a polarization catastrophe in which for some critical condition the polarization or some Fourier component of the polarization becomes very large. Equally, we may speak of the condensation of a transverse optical phonon. Here the word condensation is to be understood in the Bose-Einstein sense (*TP*, p. 199) of a time-independent displacement of finite amplitude. This can occur when the corresponding TO phonon frequency vanishes at some point in the Brillouin zone. LO phonons always have higher frequencies than the TO phonons of the same wavevector, so we are not concerned with LO phonon condensation.

In a polarization catastrophe the local electric field caused by the ionic displacement is larger than the elastic restoring force, thereby giving an asymmetrical shift in the positions of the ions. Higher order restoring forces will limit the shift to a finite displacement.

The occurrence of ferroelectricity (and antiferroelectricity) in many perovskite-structure crystals suggests that this structure is favorably disposed to a displacive transition. Local field calculations make clear the reason for the favored position of this structure: the O^{2-} ions do not have cubic surroundings, and the local field factors turn out to be unusually large.

We give first the simple form of the catastrophe theory, supposing that the local field at all atoms is equal to $\mathbf{E} + 4\pi\mathbf{P}/3$ in CGS or $\mathbf{E} + \mathbf{P}/3\epsilon_0$ in SI. The theory given now leads to a second-order transition; the physical ideas can be carried over to a first-order transition. In a second-order transition there is no latent heat; the order parameter (in this instance, the polarization) is not discontinuous at the transition temperature. In a first-order transition there is a latent heat; the order parameter changes discontinuously at the transition temperature.

We rewrite (24) for the dielectric constant in the form

$$\text{(CGS)} \qquad \epsilon = \frac{1 + \frac{8\pi}{3}\Sigma N_i\alpha_i}{1 - \frac{4\pi}{3}\Sigma N_i\alpha_i} , \tag{30}$$

where α_i is the electronic plus ionic polarizability of an ion of type i and N_i is the number of ions i per unit volume. The dielectric constant becomes infinite and permits a finite polarization in zero applied field when

$$\text{(CGS)} \qquad \Sigma N_i\alpha_i = 3/4\pi . \tag{31}$$

This is the condition for a polarization catastrophe.

The value of ϵ in (30) is sensitive to small departures of $\Sigma N_i\alpha_i$ from the critical value $3/4\pi$. If we write

$$\text{(CGS)} \qquad (4\pi/3)\Sigma N_i\alpha_i = 1 - 3s , \tag{32}$$

where $s \ll 1$, the dielectric constant in (30) becomes

$$\epsilon \simeq 1/s . \tag{33}$$

Suppose near the critical temperature s varies linearly with temperature:

$$s \simeq (T - T_c)/\xi , \tag{34}$$

where ξ is a constant. Such a variation of s or $\Sigma N_i\alpha_i$ might come from normal thermal expansion of the lattice. The dielectric constant has the form

$$\epsilon \simeq \frac{\xi}{T - T_c} , \tag{35}$$

close to the observed temperature variation in the paraelectric state, Fig. 12.

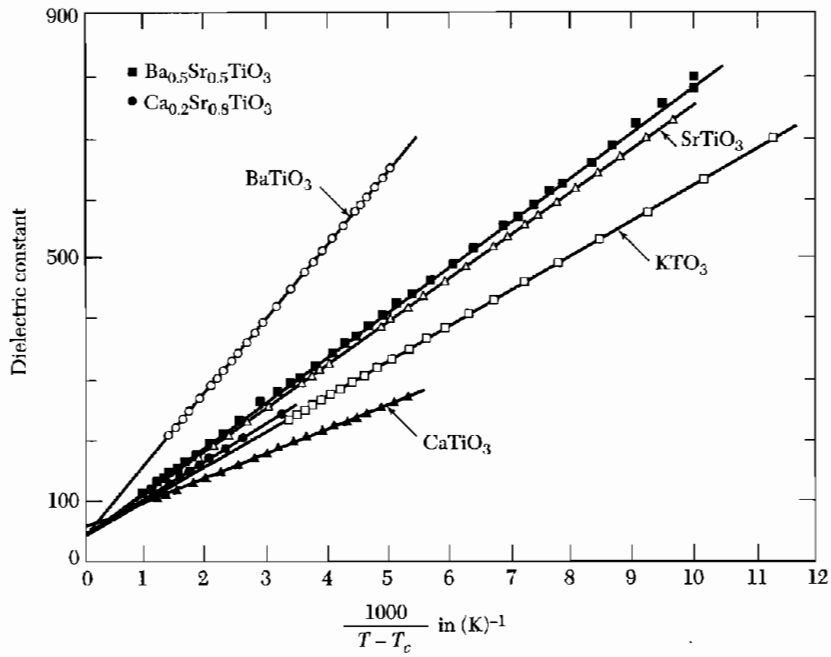


Figure 12 Dielectric constant versus $1/(T - T_c)$ in the paraelectric state ($T > T_c$) of perovskites, after G. Rupprecht and R. O. Bell.

Soft Optical Phonons

The Lyddane-Sachs-Teller relation (Chapter 14) is

$$\omega_T^2/\omega_L^2 = \epsilon(\infty)/\epsilon(0) \quad . \tag{36}$$

The static dielectric constant increases when the transverse optical phonon frequency decreases. When the static dielectric constant $\epsilon(0)$ has a high value, such as 100 to 10,000, we find that ω_T has a low value.

When $\omega_T = 0$ the crystal is unstable and $\epsilon(0)$ is infinite because there is no effective restoring force. The ferroelectric $BaTiO_3$ at 24°C has a TO mode at 12 cm^{-1}, a low frequency for an optical mode.

If the transition to a ferroelectric state is first order, we do not find $\omega_T = 0$ or $\epsilon(0) = \infty$ at the transition. The LST relation suggests only that $\epsilon(0)$ extrapolates to a singularity at a temperature T_0 below T_c.

The association of a high static dielectric constant with a low-frequency optical mode is supported by experiments on strontium titanate, $SrTiO_3$. According to the LST relation, if the reciprocal of the static dielectric constant has a temperature dependence $1/\epsilon(0) \propto (T - T_0)$, then the square of the optical mode frequency will have a similar temperature dependence: $\omega_T^2 \propto (T - T_0)$, if ω_L is independent of temperature. The result for ω_T^2 is very well confirmed by Fig. 13. Measurements of ω_T versus T for another ferroelectric crystal, SbSI, are shown in Fig. 14.

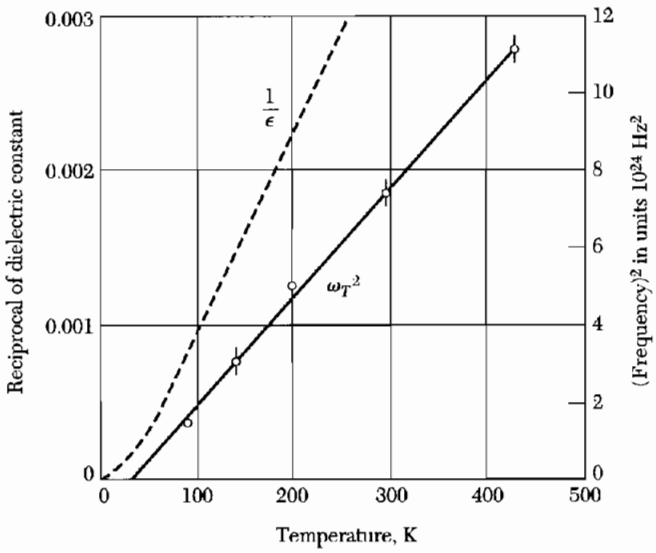


Figure 13 Plot of the square of the frequency of the zero wavevector transverse optical mode against temperature, for $SrTiO_3$, as observed in neutron diffraction experiments by Cowley. The broken line is the reciprocal of the dielectric constant from the measurements of Mitsui and Westphal.

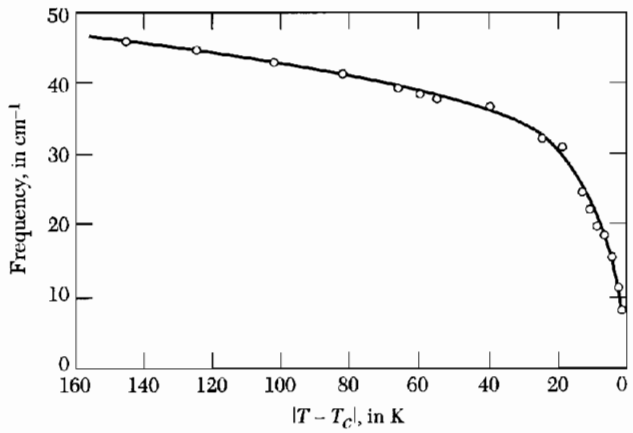


Figure 14 Decrease of a transverse phonon frequency as the Curie temperature is approached from below, in the ferroelectric crystal antimony sulphoiodide, SbSI. (After Raman scattering experiments by C. H. Perry and D. K. Agrawal.)

Landau Theory of the Phase Transition

A ferroelectric with a first-order phase transition between the ferroelectric and the paraelectric state is distinguished by a discontinuous change of the saturation polarization at the transition temperature. The transition between the normal and superconducting states is a second-order transition, as is the

transition between the ferromagnetic and paramagnetic states. In these transitions the degree of order goes to zero without a discontinuous change as the temperature is increased.

We can obtain a consistent formal thermodynamic theory of the behavior of a ferroelectric crystal by considering the form of the expansion of the energy as a function of the polarization P. We assume that the Landau[7] free energy density $\hat{F}$ in one dimension may be expanded formally as

$$\hat{F}(P;T,E) = -EP + g_0 + \tfrac{1}{2}g_2P^2 + \tfrac{1}{4}g_4P^4 + \tfrac{1}{6}g_6P^6 + \cdots , \tag{37}$$

where the coefficients g_n depend on the temperature.

The series does not contain terms in odd powers of P if the unpolarized crystal has a center of inversion symmetry, but crystals are known in which odd powers are important. Power series expansions of the free energy do not always exist, for nonanalytic terms are known to occur, especially when very near a transition. For example, the transition in KH_2PO_4 appears to have a logarithmic singularity in the heat capacity at the transition, which is not classifiable as either first or second order.

The value of P in thermal equilibrium is given by the minimum of $\hat{F}$ as a function of P; the value of $\hat{F}$ at this minimum defines the Helmholtz free energy $F(T,E)$. The equilibrium polarization in an applied electric field E satisfies the extremum condition

$$\frac{\partial \hat{F}}{\partial P} = 0 = -E + g_2P + g_4P^3 + g_6P^5 + \cdots . \tag{38}$$

In this section we assume that the specimen is a long rod with the external applied field E parallel to the long axis.

To obtain a ferroelectric state we must suppose that the coefficient of the term in P^2 in (37) passes through zero at some temperature T_0:

$$g_2 = \gamma(T - T_0) , \tag{39}$$

where γ is taken as a positive constant and T_0 may be equal to or lower than the transition temperature. A small positive value of g_2 means that the lattice is "soft" and is close to instability. A negative value of g_2 means that the unpolarized lattice is unstable. The variation of g_2 with temperature is accounted for by thermal expansion and other effects of anharmonic lattice interactions.

Second-Order Transition

If g_4 in (37) is positive, nothing new is added by the term in g_6, and this may then be neglected. The polarization for zero applied electric field is found from (38):

$$\gamma(T - T_0)P_s + g_4P_s^3 = 0 , \tag{40}$$

[7]In *TP*, see pp. 69 and 298 for a discussion of the Landau function.

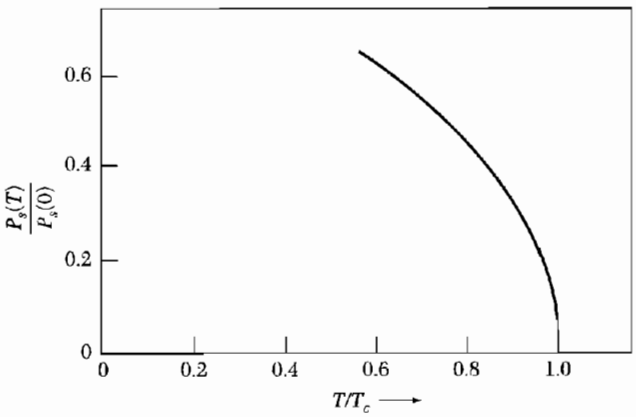


Figure 15 Spontaneous polarization versus temperature, for a second-order phase transition.

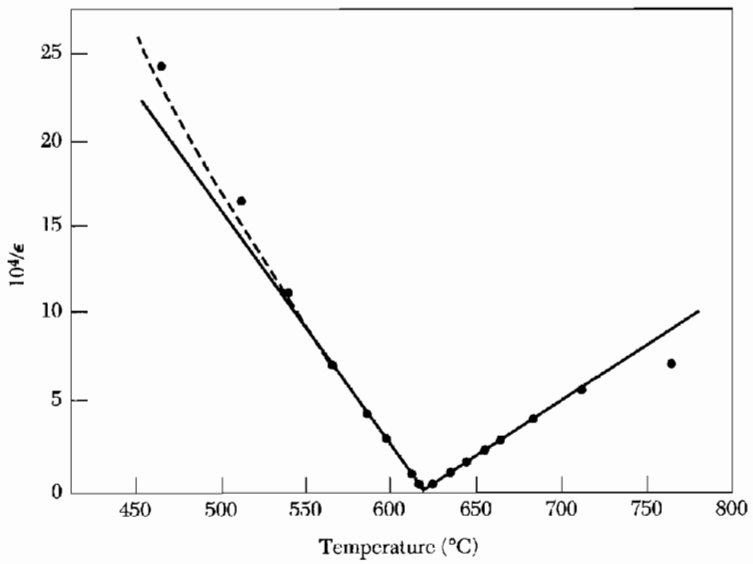


Figure 16 Temperature variation of the polar-axis static dielectric constant of $LiTaO_3$. (After Glass.)

so that either $P_s = 0$ or $P_s^2 = (\gamma/g_4)(T_0 - T)$. For $T \geq T_0$ the only real root of (40) is at $P_s = 0$, because γ and g_4 are positive. Thus T_0 is the Curie temperature. For $T < T_0$ the minimum of the Landau free energy in zero applied field is at

$$|P_s| = (\gamma/g_4)^{1/2}(T_0 - T)^{1/2} , \tag{41}$$

as plotted in Fig. 15. The phase transition is a second-order transition because the polarization goes continuously to zero at the transition temperature. The transition in $LiTaO_3$ is an example (Fig. 16) of a second-order transition.

First-Order Transition

The transition is first order if g_4 in (37) is negative. We must now retain g_6 and take it positive in order to restrain $\hat{F}$ from going to minus infinity (Fig. 17). The equilibrium condition for $E = 0$ is given by (38):

$$\gamma(T - T_0)P_s - |g_4|P_s^3 + g_6 P_s^5 = 0 \ , \tag{42}$$

so that either $P_s = 0$ or

$$\gamma(T - T_0) - |g_4|P_s^2 + g_6 P_s^4 = 0 \ . \tag{43}$$

At the transition temperature T_c the free energies of the paraelectric and ferroelectric phases will be equal. That is, the value of $\hat{F}$ for $P_s = 0$ will be equal to the value of $\hat{F}$ at the minimum given by (43). In Fig. 18 we show the characteristic variation with temperature of P_s for a first-order phase transition;

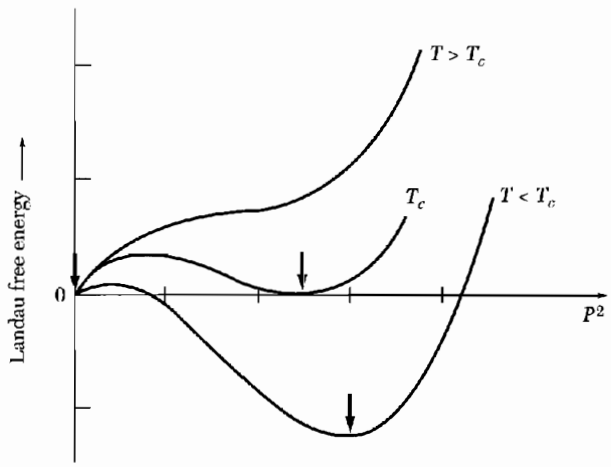


Figure 17 Landau free energy function versus (polarization)^2 in a first-order transition, at representative temperatures. At T_c the Landau function has equal minima at $P = 0$ and at a finite P as shown. For T below T_c the absolute minimum is at larger values of P; as T passes through T_c there is a discontinuous change in the position of the absolute minimum. The arrows mark the minima.

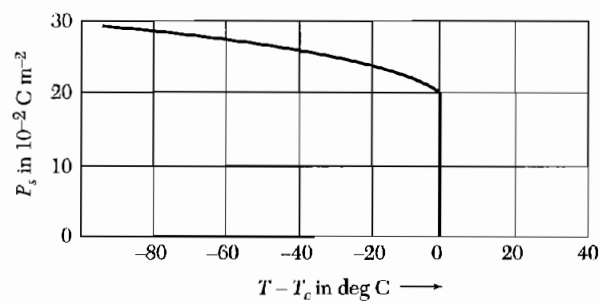


Figure 18 Calculated values of the spontaneous polarization as a function of temperature, with parameters as for barium titanate. (After W. Cochran.)

contrast this with the variation shown in Fig. 15 for a second-order phase transition. The transition in $BaTiO_3$ is first order.

The dielectric constant is calculated from the equilibrium polarization in an applied electric field E and is found from (38). In equilibrium at temperatures over the transition, the terms in P^4 and P^6 may be neglected; thus $E = \gamma(T - T_0)P$, or

$$\text{(CGS)} \qquad \epsilon(T > T_c) = 1 + 4\pi P/E = 1 + 4\pi/\gamma(T - T_0) \ , \qquad (44)$$

of the form of (36). The result applies whether the transition is of the first or second order, but if second order we have $T_0 = T_c$; if first order, then $T_0 < T_c$. Equation (39) defines T_0, but T_c is the transition temperature.

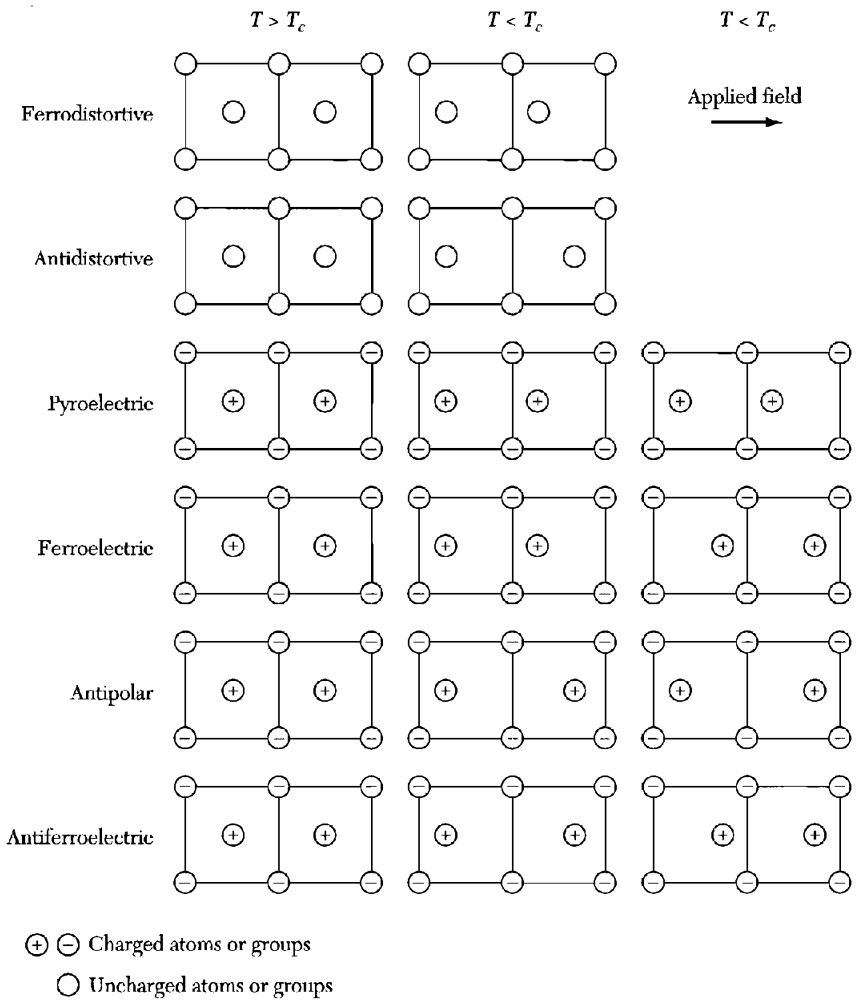


Figure 19 Schematic representation of fundamental types of structural phase transitions from a centrosymmetric prototype. (After Lines and Glass.)

Antiferroelectricity

A ferroelectric displacement is not the only type of instability that may develop in a dielectric crystal. Other deformations occur, as in Fig. 19. These deformations, even if they do not give a spontaneous polarization, may be accompanied by changes in the dielectric constant. One type of deformation is called **antiferroelectric** and has neighboring lines of ions displaced in opposite senses. The perovskite structure appears to be susceptible to many types of deformation, often with little difference in energy between them. The phase diagrams of mixed perovskite systems, such as the $PbZrO_3$–$PbTiO_3$ system, show transitions between para-, ferro-, and antiferroelectric states (Fig. 20). Several crystals believed to have an ordered nonpolar state are listed in Table 3.

Ferroelectric Domains

Consider a ferroelectric crystal (such as barium titanate in the tetragonal phase) in which the spontaneous polarization may be either up or down the c axis of the crystal. A ferroelectric crystal generally consists of regions called **domains** within each of which the polarization is in the same direction, but in adjacent domains the polarization is in different directions. In Fig. 21 the polarization is in opposite directions. The net polarization depends on the difference in the volumes of the upward- and downward-directed domains. The

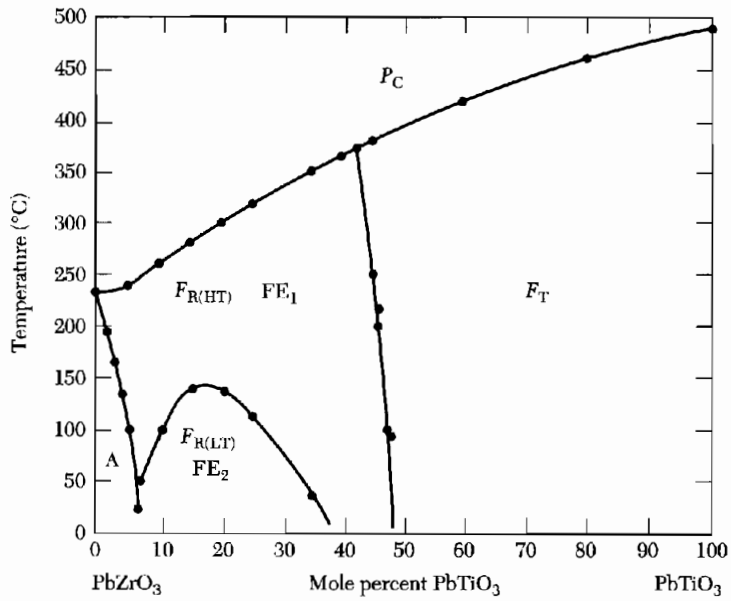


Figure 20 Ferroelectric F, antiferroelectric A, and paraelectric P phases of the lead zirconate–lead titanate solid solution system. The subscript T denotes a tetragonal phase; C a cubic phase; R a rhombohedral phase, of which there are high-temperature (HT) and low-temperature (LT) forms. Near the rhombohedral–tetragonal phase boundaries one finds very high piezoelectric coupling coefficients. (After Jaffe.)

Table 3 Antiferroelectric crystals

Crystal	Transition temperature to antiferroelectric state, in K
WO_3	1010
$NaNbO_3$	793, 911
$PbZrO_3$	506
$PbHfO_3$	488
$NH_4H_2PO_4$	148
$ND_4D_2PO_4$	242
$NH_4H_2AsO_4$	216
$ND_4D_2AsO_4$	304
$(NH_4)_2H_3IO_6$	254

From a compilation by Walter J. Merz.

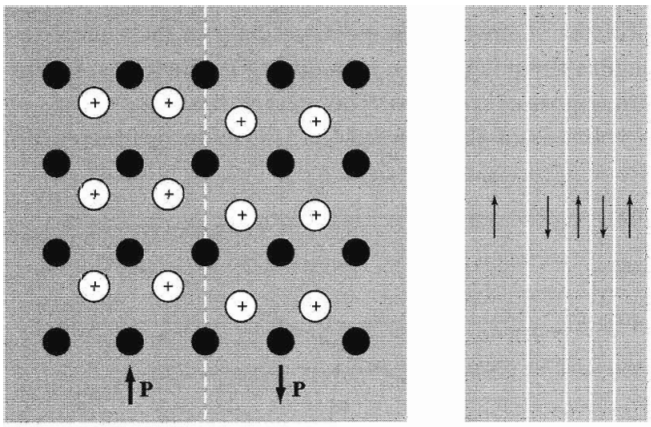


Figure 21 (a) Schematic drawing of atomic displacements on either side of a boundary between domains polarized in opposite directions in a ferroelectric crystal; (b) view of a domain structure, showing 180° boundaries between domains polarized in opposite directions.

crystal as a whole will appear to be unpolarized, as measured by the charge on electrodes covering the ends, when the volumes of domains in opposite senses are equal. The total dipole moment of the crystal may be changed by the movement of the walls between domains or by the nucleation of new domains.

Figure 22 is a series of photomicrographs of a single crystal of barium titanate in an electric field normal to the plane of the photographs and parallel to the tetragonal axis. The closed curves are boundaries between domains polarized into and out of the plane of the photographs. The domain boundaries change size and shape when the intensity of the electric field is altered.

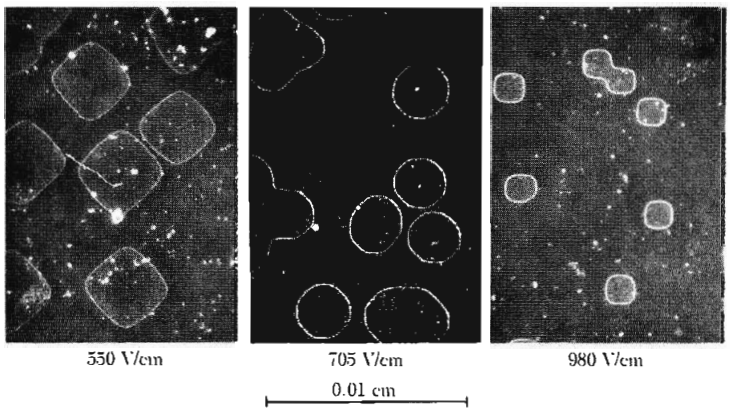


Figure 22 Ferroelectric domains on the face of a single crystal of barium titanate. The face is normal to the tetragonal or c axis. The net polarization of the crystal as judged by domain volumes is increased markedly as the electric field intensity parallel to the axis is increased from 550 volts/cm to 980 V/cm. The domain boundaries are made visible by etching the crystal in a weak acid solution. (R. C. Miller.)

Piezoelectricity

All crystals in a ferroelectric state are also piezoelectric: a stress Z applied to the crystal will change the electric polarization (Fig. 23). Similarly, an electric field E applied to the crystal will cause the crystal to become strained. In schematic one-dimensional notation, the piezoelectric equations are

$$\text{(CGS)} \qquad P = Zd + E\chi; \quad e = Zs + Ed \ , \tag{45}$$

where P is the polarization, Z the stress, d the **piezoelectric strain constant**, E the electric field, χ the dielectric susceptibility, e the elastic strain, and s the elastic compliance constant. To obtain (45) in SI, replace χ by $\epsilon_0\chi$. These relations exhibit the development of polarization by an applied stress and the development of elastic strain by an applied electric field.

A crystal may be piezoelectric without being ferroelectric: a schematic example of such a structure is given in Fig. 24. Quartz is piezoelectric, but not ferroelectric; barium titanate is both. For order of magnitude, in quartz $d \approx 10^{-7}$ cm/statvolt and in barium titanate $d \approx 10^{-5}$ cm/statvolt. The general definition of the piezoelectric strain constants is

$$d_{ik} = (\partial e_k/\partial E_i)_Z \ , \tag{46}$$

where $i \equiv x, y, z$ and $k \equiv xx, yy, zz, yz, zx, xy$. To convert to cm/stat-V from values of d_{ik} given in m/V, multiply by 3×10^4.

The lead zirconate–lead titanate system (called the PZT system), Fig. 20, is widely used in polycrystalline (ceramic) form with compositions of very high piezoelectric coupling. The synthetic polymer polyvinylidenfluoride (PVF_2) is

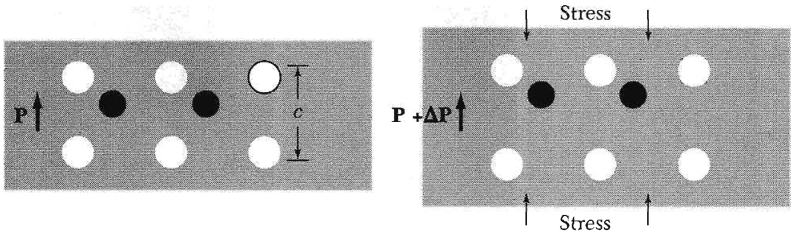


Figure 23 (a) Unstressed ferroelectric crystal and (b) stressed ferroelectric crystal. The stress changes the polarization by **ΔP**, the induced piezoelectric polarization.

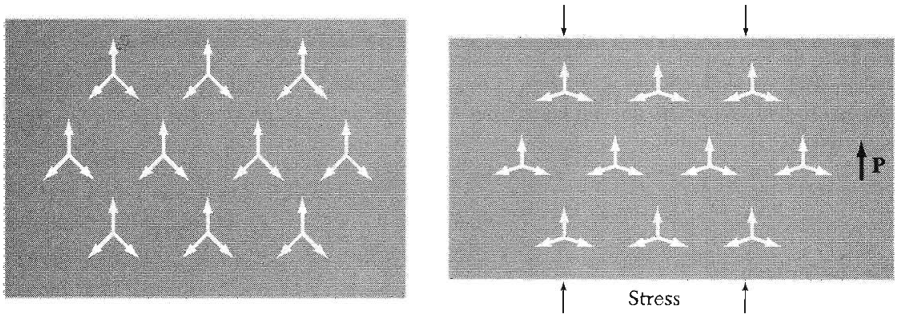


Figure 24 (a) The unstressed crystal has a threefold symmetry axis. The arrows represent dipole moments; each set of three arrows represents a planar group of ions denoted by $A_3^+B^{3-}$, with a B^{3-} ion at each vertex. The sum of the three dipole moments at each vertex is zero. (b) The crystal when stressed develops a polarization in the direction indicated. The sum of the dipole moments about each vertex is no longer zero.

five times more strongly piezoelectric than crystalline quartz. Thin stretched films of PVF_2 are flexible and as ultrasonic transducers are applied in medicine to monitor blood pressure and respiration.

SUMMARY

(In CGS Units)

- The electric field averaged over the volume of the specimen defines the macroscopic electric field **E** of the Maxwell equations.
- The electric field that acts at the site $\mathbf{r}_j$ of an atom j is the local electric field, $\mathbf{E}_{\text{loc}}$. It is a sum over all charges, grouped in terms as $\mathbf{E}_{\text{loc}}(\mathbf{r}_j) = \mathbf{E}_0 + \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3(\mathbf{r}_j)$, where only $\mathbf{E}_3$ varies rapidly within a cell. Here:
  $\mathbf{E}_0$ = external electric field;
  $\mathbf{E}_1$ = depolarization field associated with the boundary of the specimen;
  $\mathbf{E}_2$ = field from polarization outside a sphere centered about $\mathbf{r}_j$;
  $\mathbf{E}_3(\mathbf{r}_j)$ = field at $\mathbf{r}_j$ due to all atoms inside the sphere.

- The macroscopic field $\mathbf{E}$ of the Maxwell equations is equal to $\mathbf{E}_0 + \mathbf{E}_1$, which, in general, is not equal to $\mathbf{E}_{\text{loc}}(\mathbf{r}_j)$.
- The depolarization field in an ellipsoid is $E_{1\mu} = -N_{\mu\nu}P_\nu$, where $N_{\mu\nu}$ is the depolarization tensor; the polarization $\mathbf{P}$ is the dipole moment per unit volume. In a sphere $N = 4\pi/3$.
- The Lorentz field is $\mathbf{E}_2 = 4\pi\mathbf{P}/3$.
- The polarizability α of an atom is defined in terms of the local electric field as $\mathbf{p} = \alpha\mathbf{E}_{\text{loc}}$.
- The dielectric susceptibility χ and dielectric constant ϵ are defined in terms of the macroscopic electric field $\mathbf{E}$ as $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P} = \epsilon\mathbf{E} = (1 + 4\pi\chi)\mathbf{E}$, or $\chi = P/E$. In SI, we have $\chi = P/\epsilon_0 E$.
- An atom at a site with cubic symmetry has $\mathbf{E}_{\text{loc}} = \mathbf{E} + (4\pi/3)\mathbf{P}$ and satisfies the Clausius-Mossotti relation (24).

Problems

1. ***Polarizability of atomic hydrogen.*** Consider a semiclassical model of the ground state of the hydrogen atom in an electric field normal to the plane of the orbit (Fig. 25), and show that for this model $\alpha = a_H^3$, where a_H is the radius of the unperturbed orbit. *Note:* If the applied field is in the x direction, then the x component of the field of the nucleus at the displaced position of the electron orbit must be equal to the applied field. The correct quantum-mechanical result is larger than this by the factor $\frac{9}{2}$. (We are speaking of α_0 in the expansion $\alpha = \alpha_0 + \alpha_1 E + \cdots$.) We assume $x \ll a_H$. One can also calculate α_1 on this model.

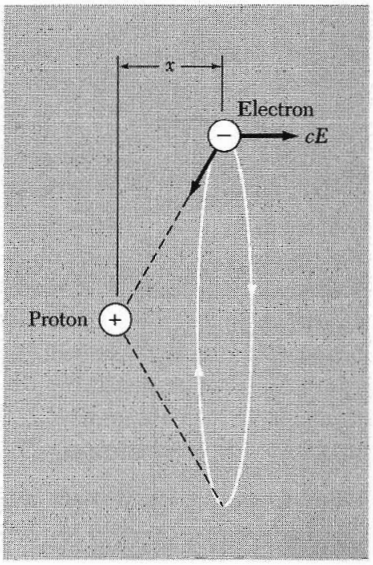


Figure 25 An electron in a circular orbit of radius a_H is displaced a distance x on application of an electric field E in the $-x$ direction. The force on the electron due to the nucleus is e^2/a_H^2 in CGS or $e^2/4\pi\epsilon_0 a_H^2$ in SI. The problem assumes $x \ll a_H$.

2. ***Polarizability of conducting sphere.*** Show that the polarizability of a conducting metallic sphere of radius a is $\alpha = a^3$. This result is most easily obtained by noting that $E = 0$ inside the sphere and then using the depolarization factor $4\pi/3$ for a sphere (Fig. 26). The result gives values of α of the order of magnitude of the observed polarizabilities of atoms. A lattice of N conducting spheres per unit volume has dielectric constant $\epsilon = 1 + 4\pi N a^3$, for $Na^3 \ll 1$. The suggested proportionality of α to the cube of the ionic radius is satisfied quite well for alkali and halogen ions. To do the problem in SI, use $\frac{1}{3}$ as the depolarization factor.

3. ***Effect of air gap.*** Discuss the effect of an air gap (Fig. 27) between capacitor plates and dielectric on the measurement of high dielectric constants. What is the highest apparent dielectric constant possible if the air gap thickness is 10^{-3} of the total thickness? The presence of air gaps can seriously distort the measurement of high dielectric constants.

4. ***Interfacial polarization.*** Show that a parallel-plate capacitor made up of two parallel layers of material—one layer with dielectric constant ϵ, zero conductivity, and thickness d, and the other layer with $\epsilon = 0$ for convenience, finite conductivity σ,

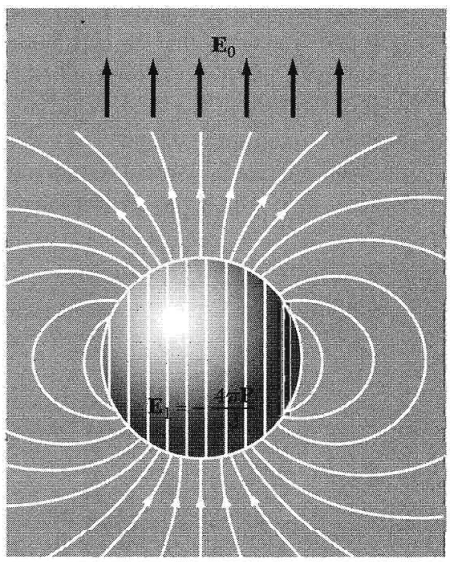


Figure 26 The total field inside a conducting sphere is zero. If a field $\mathbf{E}_0$ is applied externally, then the field $\mathbf{E}_1$ due to surface charges on the sphere must just cancel $\mathbf{E}_0$, so that $\mathbf{E}_0 + \mathbf{E}_1 = 0$ within the sphere. But $\mathbf{E}_1$ can be simulated by the depolarization field $-4\pi\mathbf{P}/3$ of a uniformly polarized sphere of polarization $\mathbf{P}$. Relate $\mathbf{P}$ to $\mathbf{E}_0$ and calculate the dipole moment $\mathbf{p}$ of the sphere. In SI the depolarization field is $-\mathbf{P}/3\epsilon_0$.

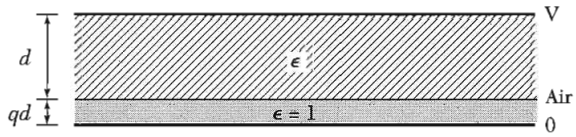


Figure 27 An air gap of thickness qd is in series in a capacitor with a dielectric slab of thickness d.

and thickness qd—behaves as if the space between the condenser plates were filled with a homogeneous dielectric with dielectric constant

$$\epsilon_{\text{eff}} = \frac{\epsilon(1+q)}{1-(i\epsilon\omega q/4\pi\sigma)} ,$$

where ω is the angular frequency. Values of ϵ_{eff} as high as 10^4 or 10^5 caused largely by this Maxwell-Wagner mechanism are sometimes found, but the high values are always accompanied by large ac losses.

5. ***Polarization of sphere.*** A sphere of dielectric constant ϵ is placed in a uniform external electric field E_0. (a) What is the volume average electric field E in the sphere? (b) Show that the polarization in the sphere is $P = \chi E_0/[1 + (4\pi\chi/3)]$, where $\chi = (\epsilon - 1)/4\pi$. Hint: You do not need to calculate E_{loc} in this problem; in fact it is confusing to do so, because ϵ and χ are defined so that $P = \chi E$. We require E_0 to be unchanged by insertion of the sphere. We can produce a fixed E_0 by placing positive charges on one thin plate of an insulator and negative charges on an opposite plate. If the plates are always far from the sphere, the field of the plates will remain unchanged when the sphere is inserted between them. The results above are in CGS.

6. ***Ferroelectric criterion for atoms.*** Consider a system of two neutral atoms separated by a fixed distance a, each atom having a polarizability α. Find the relation between a and α for such a system to be ferroelectric. Hint: The dipolar field is strongest along the axis of the dipole.

7. ***Saturation polarization at Curie point.*** In a first-order transition the equilibrium condition (43) with T set equal to T_c gives one equation for the polarization $P_s(T_c)$. A further condition at the Curie point is that $\hat{F}(P_s, T_c) = \hat{F}(0, T_c)$. (a) Combining these two conditions, show that $P_s^2(T_c) = 3|g_4|/4g_6$. (b) Using this result, show that $T_c = T_0 + 3g_4^2/16\gamma g_6$.

8. ***Dielectric constant below transition temperature.*** In terms of the parameters in the Landau free energy expansion, show that for a second-order phase transition the dielectric constant below the transition temperature is

$$\epsilon = 1 + 4\pi\Delta P/E = 1 + 2\pi/\gamma(T_c - T) .$$

This result may be compared with (44) above the transition.

9. ***Soft modes and lattice transformations.*** Sketch a monatomic linear lattice of lattice constant a. (a) Add to each of six atoms a vector to indicate the direction of the displacement at a given time caused by a longitudinal phonon with wavevector at the zone boundary. (b) Sketch the crystal structure that results if this zone boundary phonon becomes unstable ($\omega \to 0$) as the crystal is cooled through T_c. (c) Sketch on one graph the essential aspects of the longitudinal phonon dispersion relation for the monatomic lattice at T well above T_c and at $T = T_c$. Add to the graph the same information for phonons in the new structure at T well below T_c.

10. ***Ferroelectric linear array.*** Consider a line of atoms of polarizability α and separation a. Show that the array can polarize spontaneously if $\alpha \geq a^3/4\Sigma n^{-3}$, where the sum is over all positive integers and is given in tables as 1.202. . . .

17

Surface and Interface Physics

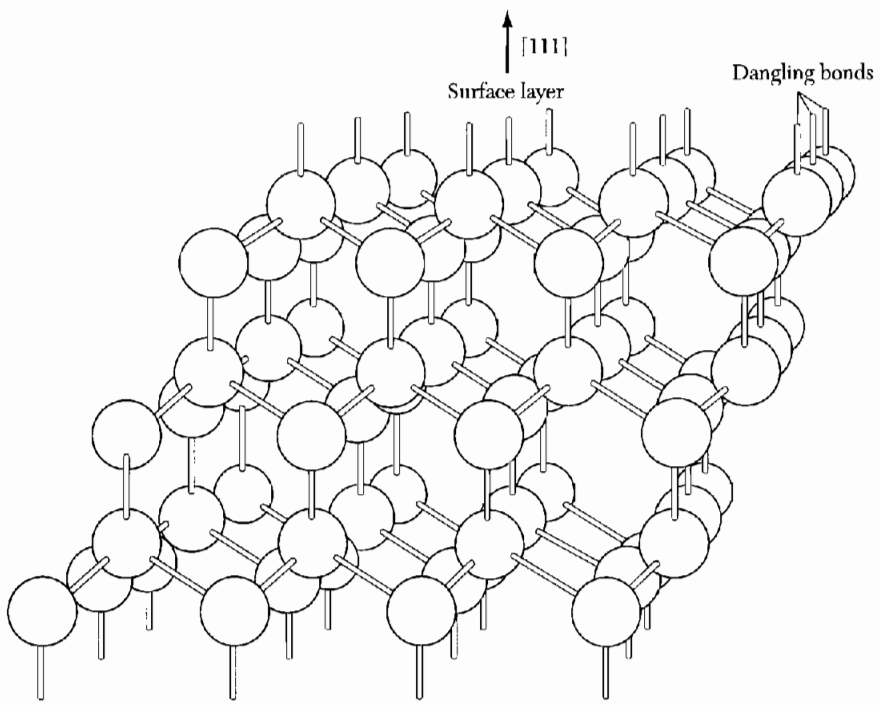


Figure 1 Dangling bonds from the (111) surface of a covalently bonded diamond cubic structure. (After M. Prutton, *Surface physics*, Clarendon, 1975.)

Reconstruction and Relaxation

The **surface** of a crystalline solid in vacuum is generally defined as the few, approximately three, outermost atomic layers of the solid that differ significantly from the bulk. The surface may be entirely clean or it may have foreign atoms deposited on it or incorporated in it. The bulk of the crystal is called the **substrate**.

If the surface is clean the top layer may be either **reconstructed** or, sometimes, unreconstructed. In unreconstructed surfaces the atomic arrangement is in registry with that of the bulk except for an interlayer spacing change (called multilayer relaxation) at the top surface.

The shrinking of the interlayer distance between the first and second layer of atoms with respect to subsequent layers in the bulk is a rather dominant phenomenon. The surface may be thought of as an intermediate between a diatomic molecule and the bulk structure. Because the interatomic distances in diatomic molecules are much smaller than in the bulk, there is a rationale for the surface relaxation. This may be contrasted with reconstruction where the relaxation of atoms yields new surface primitive cells. In relaxation the atoms maintain their structure in the surface plane as it was (according to the projection of the bulk cell on the surface); only their distance from the bulk changes.

Sometimes in metals, but most often in nonmetals, the atoms in the surface layer form superstructures in which the atoms in the layer are not in registry with the atoms in corresponding layers in the substrate. This is surface reconstruction; it can be a consequence of a rearrangement of broken covalent or ionic bonds at the surface. Under such conditions the atoms at the surface bunch into rows with alternately larger and smaller spacings than in the bulk. That is, for some crystals held together by valence bonds, creation of a surface would leave unsaturated bonds dangling into space (Fig. 1). The energy may then be lowered if neighboring atoms approach each other and form bonds with their otherwise unused valence electrons. Atomic displacements can be as large as 0.5 Å.

Reconstruction does not necessarily require formation of a superstructure. For example, on GaAs (110) surfaces a rotation of the Ga-As bond leaves the point group intact. The driving force is electron transfer from Ga to As, which fills the dangling bonds on As and depletes them on Ga.

Surfaces of planes nominally of high indices may be built up of low index planes separated by steps one (or two) atoms in height. Such terrace-step arrangements are important in evaporation and desorption because the attachment energy of atoms is often low at the steps and at kinks in the steps. The chemical activity of such sites may be high. The presence of periodic arrays of

steps may be detected by double and triple beams of diffraction in **LEED** (see below) experiments.

SURFACE CRYSTALLOGRAPHY

The surface structure is in general periodic only in two dimensions. The surface structure can be the structure of foreign material deposited on the substrate or it can be the selvage of the pure substrate. In Chapter 1 we used the term Bravais lattice for the array of equivalent points in two or in three dimensions, that is, for diperiodic or triperiodic structures. In the physics of surfaces it is common to speak of a two-dimensional **lattice**. Further, the area unit may be called a **mesh**.

We showed in Fig. 1.7 four of the five nets possible for a diperiodic structure; the fifth net is the general oblique net, with no special symmetry relation between the mesh basis vectors $\mathbf{a}_1$, $\mathbf{a}_2$. Thus the five distinct nets are the oblique, square, hexagonal, rectangular, and centered rectangular.

The substrate net parallel to the surface is used as the reference net for the description of the surface. For example, if the surface of a cubic substrate crystal is the (111) surface, the substrate net is hexagonal (Fig. 1.7b), and the surface net is referred to these axes.

The vectors $\mathbf{c}_1$, $\mathbf{c}_2$ that define the mesh of the surface structure may be expressed in terms of the reference net $\mathbf{a}_1$, $\mathbf{a}_2$ by a matrix operation $\mathbf{P}$:

$$\begin{pmatrix}\mathbf{c}_1\\ \mathbf{c}_2\end{pmatrix} = \mathbf{P}\begin{pmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{pmatrix} = \begin{pmatrix}P_{11} & P_{12}\\ P_{21} & P_{22}\end{pmatrix}\begin{pmatrix}\mathbf{a}_1\\ \mathbf{a}_2\end{pmatrix} . \tag{1}$$

Provided that the included angles of the two meshes are equal, the shorthand notation due to E. A. Wood may be used. In this notation, which is widely used, the relation of the mesh $\mathbf{c}_1$, $\mathbf{c}_2$ to the reference mesh $\mathbf{a}_1$, $\mathbf{a}_2$ is expressed as

$$\left(\frac{c_1}{a_1} \times \frac{c_2}{a_2}\right) R\alpha , \tag{2}$$

in terms of the lengths of the mesh basis vectors and the angle α of relative rotation R of the two meshes. If $\alpha = 0$, the angle is omitted. Examples of the Wood notation are given in Fig. 2.

The reciprocal net vectors of the surface mesh may be written as $\mathbf{c}_1^*$, $\mathbf{c}_2^*$, defined by

$$\mathbf{c}_1 \cdot \mathbf{c}_2^* = \mathbf{c}_2 \cdot \mathbf{c}_1^* = 0 \ ; \qquad \mathbf{c}_1 \cdot \mathbf{c}_1^* = \mathbf{c}_2 \cdot \mathbf{c}_2^* = 2\pi \text{ (or 1)} . \tag{3}$$

Here the 2π (or 1) indicates that two conventions are in use. The definitions (3) used in Fig. 3 may be compared with the definitions in Chapter 2 of the reciprocal lattice vectors of a triperiodic lattice.

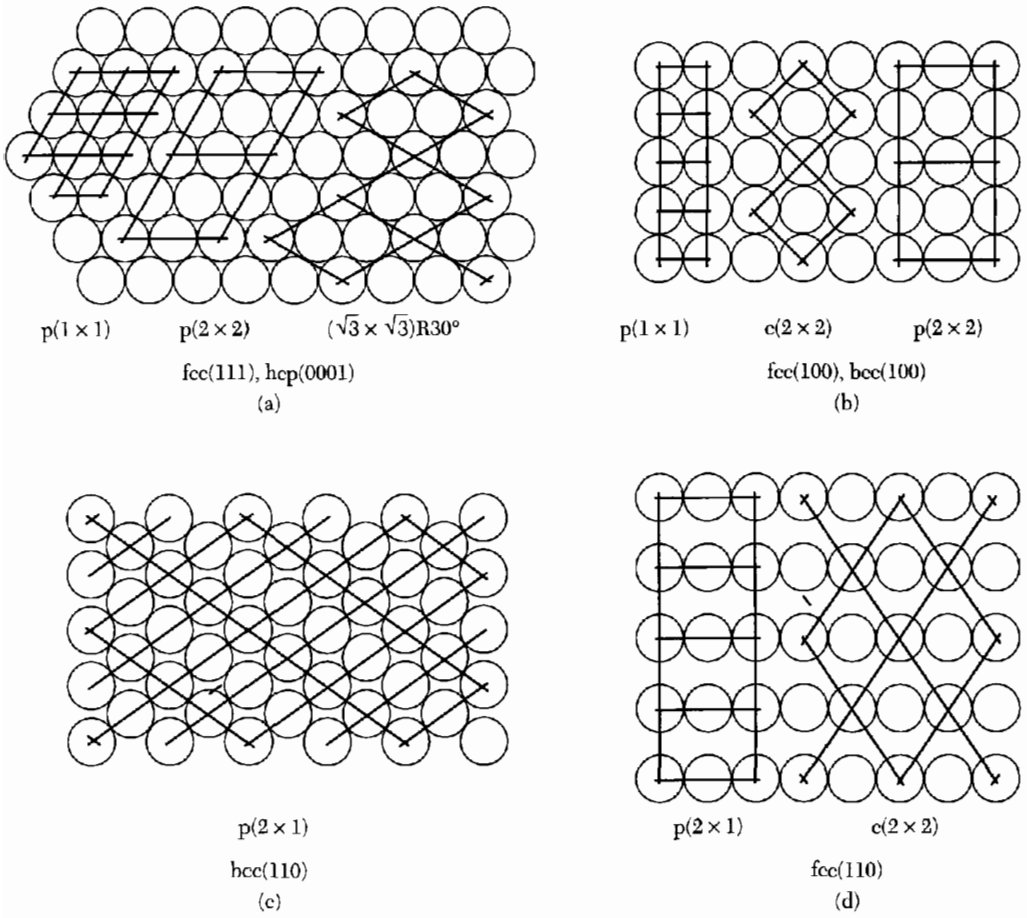


Figure 2 Surface nets of adsorbed atoms. The circles represent atoms in the top layer of the substrate. In (a) the designation fcc(111) means the (111) face of an fcc structure. This face determines a reference net. The lines represent ordered overlayers, with adatoms at the intersections of two lines. The intersection points represent diperiodic nets (lattices in two dimensions). The designation p(1 × 1) in (a) is a primitive mesh unit for which the basis is identical with the basis of the reference net. In (b) the c(2 × 2) mesh unit is a centered mesh with basis vectors twice as long as those of the reference net. Atomic adsorption on metals takes place most often into those surface sites (hollow sites) that maximize the number of nearest-neighbor atoms on the substrate. (After Van Hove.)

The reciprocal net points of a diperiodic net may be thought of—when we are in three dimensions—as rods. The rods are infinite in extent and normal to the surface plane, where they pass through the reciprocal net points. It may be helpful to think of the rods as generated by a triperiodic lattice which is expanded without limit along one of its axes. Then the reciprocal lattice points along this axis are moved closer together and in the limit form a rod.

The usefulness of the rod concept comes out with the Ewald sphere construction explained in Fig. 2.8. Diffraction occurs everywhere the Ewald

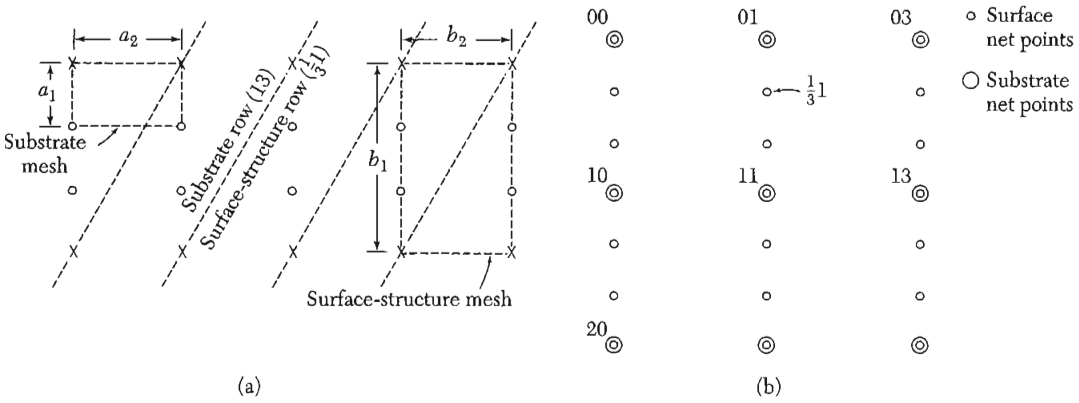


Figure 3 A (3 × 1) surface structure, (a) real-space; and (b) reciprocal-space diagrams. (After E. A. Wood.)

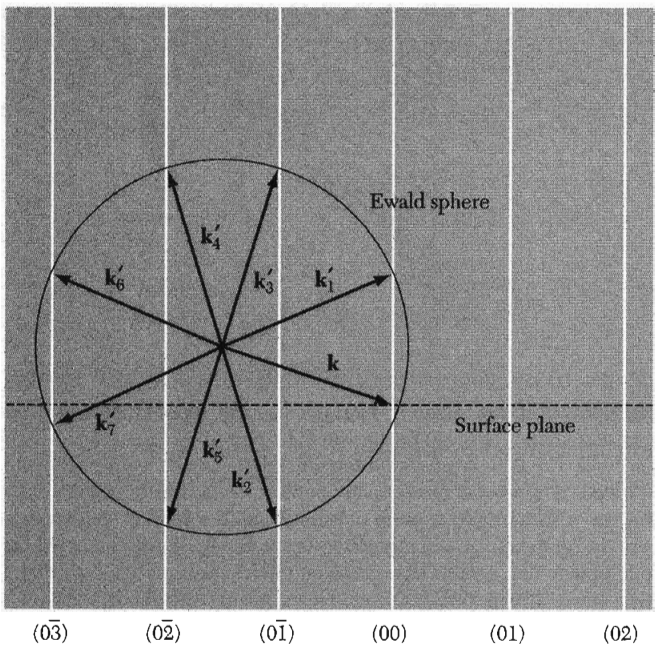


Figure 4 Ewald sphere construction for diffraction of incident wave **k** by a square net, when **k** is parallel to one axis of the mesh. The back scattered beams in the plane of the paper are $\mathbf{k}'_4$, $\mathbf{k}'_5$, $\mathbf{k}'_6$, $\mathbf{k}'_7$. Diffracted beams out of the plane of the paper will also occur. The vertical lines are the rods of the reciprocal net.

sphere intercepts a reciprocal net rod. Each diffracted beam is labelled with the indices *hk* of the reciprocal net vector

$$\mathrm{g} = h\mathbf{c}_1^* + k\mathbf{c}_2^* \tag{4}$$

forming the beam.

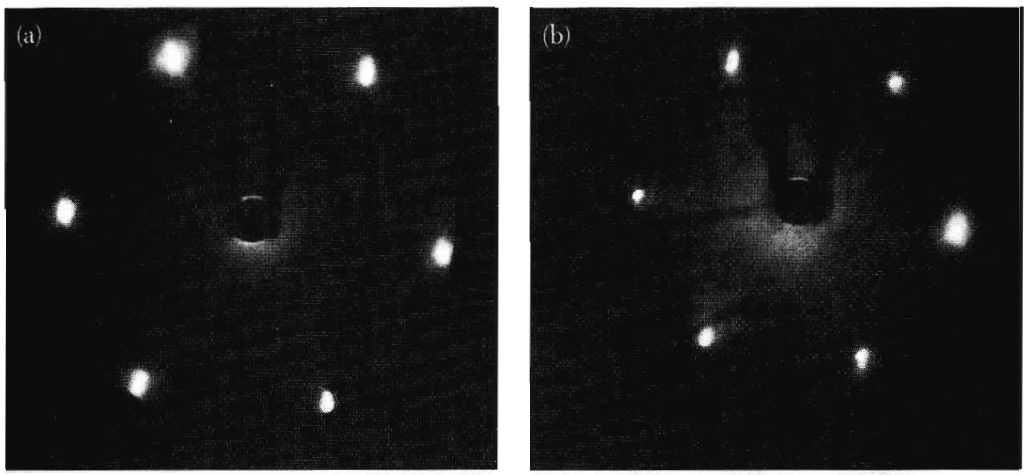


Figure 5 LEED patterns from a Pt(111) crystal surface for incident electron energies of 51 and 63.5 eV. The diffraction angle is greater at the lower energy. (After G. A. Somorjai, *Chemistry in two dimensions: surfaces*, Cornell, 1981.)

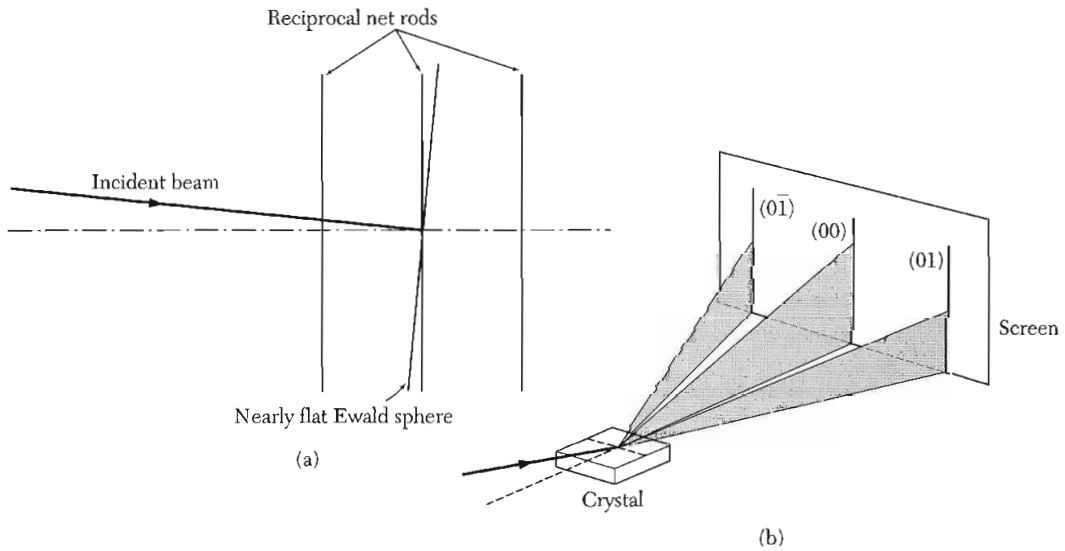


Figure 6 The RHEED method. In (a) the high-energy incident electron beam at a glancing angle to the crystal surface is associated with an Ewald sphere of large radius, so large that the surface is nearly flat in relation to the separation between adjacent rods of the reciprocal net. The formation of diffraction lines on a plane screen is shown in (b). (After Prutton.)

Low energy electron diffraction (LEED) is illustrated by Fig. 4. The electron energy is typically in the range 10–1000 eV. With this arrangement Davisson and Germer in 1927 discovered the wave nature of the electron. An experimental pattern is shown in Fig. 5.

Reflection High-Energy Electron Diffraction. In the RHEED method a beam of high-energy electrons is directed upon a crystal surface at grazing

incidence. By adjustment of the angle of incidence one can arrange the normal component of the incoming wavevector to be very small, which will minimize the penetration of the electron beam and enhance the role of the crystal surface.

The radius k of the Ewald sphere for 100 keV electrons will be $\simeq 10^3$ Å^{-1}, which is much longer than the shortest reciprocal lattice vector $2\pi/a \approx 1$ Å^{-1}. It follows that the Ewald sphere will be nearly a flat surface in the central scattering region. The intercept of the rods of the reciprocal net with the nearly flat sphere will be nearly a line when the beam is directed at grazing incidence. The experimental arrangement is shown in Fig. 6.

SURFACE ELECTRONIC STRUCTURE

Work Function

The work function W of the uniform surface of a metal is defined as the difference in potential energy of an electron between the vacuum level and the Fermi level. The vacuum level is the energy of an electron at rest at a point sufficiently far outside the surface so that the electrostatic image force on the electron may be neglected—more than 100 Å from the surface. The Fermi level is the electrochemical potential of the electrons in the metal.

Typical values of electron work functions are given in Table 1. The orientation of the exposed crystal face affects the value of the work function

Table 1 Electron work functions[a]

(Values obtained by photoemission, except tungsten obtained by field emission.)

Element	Surface plane	Work function, in eV
Ag	(100)	4.64
	(110)	4.52
	(111)	4.74
Cs	polycrystal	2.14
Cu	(100)	4.59
	(110)	4.48
	(111)	4.98
Ge	(111)	4.80
Ni	(100)	5.22
	(110)	5.04
	(111)	5.35
W	(100)	4.63
	(110)	5.25
	(111)	4.47

[a]After H. D. Hagstrum.

because the strength of the electric double layer at the surface depends on the concentration of surface positive ion cores. The double layer exists because the surface ions are in an asymmetrical environment, with vacuum (or an adsorbed foreign atom layer) on one side and the substrate on the other side.

The work function is equal to the threshold energy for photoelectric emission at absolute zero. If $\hbar\omega$ is the energy of an incident photon, then the Einstein equation is $\hbar\omega = W + T$, where T is the kinetic energy of the emitted electron and W is the work function.

Thermionic Emission

The rate of emission of thermionic electrons depends exponentially on the work function. The derivation follows.

We first find the electron concentration in vacuum in equilibrium with electrons in a metal at temperature $\tau(=k_B T)$ and chemical potential μ. We treat the electrons in the vacuum as an ideal gas, so that their chemical potential is

$$\mu = \mu_{\text{ext}} + \tau \log(n/n_Q) \ , \tag{5}$$

by *TP*, Chapter 5. Here

$$n_Q = 2(m\tau/2\pi\hbar^2)^{3/2} \ , \tag{6}$$

for particles of spin 1/2.

Now $\mu_{\text{ext}} - \mu = W$, by the definition of the work function W. Thus, from (5),

$$n = n_Q \exp(-W/\tau) \ . \tag{7}$$

The flux of electrons that leaves the metal surface when all electrons are drawn off is equal to the flux incident on the surface from outside:

$$J_n = \tfrac{1}{4} n\bar{c} = n(\tau/2\pi m)^{1/2} \tag{8}$$

by *TP*(14.95) and (14.121). Here $\bar{c}$ is the mean speed of the electrons in the vacuum. The electric charge flux is eJ_n or

$$J_e = (\tau^2 me/2\pi^2\hbar^3)\exp(-W/\tau) \ . \tag{9}$$

This is called the Richardson-Dushman equation for thermionic emission.

Surface States

At the free surface of a semiconductor there often exist surface-bound electronic states with energies in the forbidden gap between the valence and conduction bands of the bulk semiconductor. We can obtain a good impression of the nature of the surface states by considering the wave functions in the

weak binding or two-component approximation of Chapter 7, in one dimension. (The wave functions in three dimensions will have extra factors $\exp[i(k_y y + k_z z)]$ in the y, z plane of the surface.)

If the vacuum lies in the region $x > 0$, the potential energy of an electron in this region can be set equal to zero:

$$U(x) = 0 \ , \qquad x > 0 \ . \tag{10}$$

In the crystal the potential energy has the usual periodic form:

$$U(x) = \sum_G U_G \exp(iGx) \ , \qquad x < 0 \ . \tag{11}$$

In one dimension $G = n\pi/a$, where n is any integer, including zero.

In the vacuum the wave function of a bound surface state must fall off exponentially:

$$\psi_{\text{out}} = \exp(-sx) \ , \qquad x > 0 \ . \tag{12}$$

By the wave equation the energy of the state referred to the vacuum level is

$$\epsilon = -\hbar^2 s^2/2m \ . \tag{13}$$

Within the crystal the two-component wave function of a bound surface state will have the form, for $x < 0$,

$$\psi_{\text{in}} = \exp(qx + ikx)[C(k) + C(k - G)\exp(-iGx)] \ , \tag{14}$$

by analogy with (7.49), but with the addition of the factor $\exp(qx)$ which serves to bind the electron to the surface.

We now come to an important consideration that restricts the allowed values of the wavevector k. If the state is bound, there can be no current flow in the x direction, normal to the surface. This condition is assured in quantum mechanics if the wave function can be written as a real function of x, a condition already satisfied by the exterior wave function (12). But (14) can be a real function only if $k = \frac{1}{2}G$, so that

$$\psi_{\text{in}} = \exp(qx)\left[C(\tfrac{1}{2}G)\exp(iGx/2) + C(-\tfrac{1}{2}G)\exp(-iGx/2)\right] \ . \tag{15}$$

This is real provided $C^*(\frac{1}{2}G) = C(-\frac{1}{2}G)$. Thus k_x for a surface state does not have a continuum of values, but is limited to discrete states associated with Brillouin zone boundaries.

The state (15) is damped exponentially in the crystal. The constants s, q are related by the condition that ψ and $d\psi/dx$ are continuous at $x = 0$. The binding energy ϵ is determined by solving the two-component secular equation analogous to (7.46). The plot of Fig. 7.12 is helpful in this connection.

Tangential Surface Transport

We have seen that there may exist surface-bound electronic states with energies in the forbidden gap between the valence and conduction bands of the substrate crystal. These states may be occupied or vacant; their existence must affect the statistical mechanics of the problem. This means that the states modify the local equilibrium concentration of electrons and holes, as expressed as a shift of the chemical potential relative to the band edges. Because the chemical potential is independent of position in an equilibrium system, the energy bands must be displaced or bent, as in Fig. 7.

The thickness and carrier concentration in the surface layer may be changed by applying an electric field normal to the surface. The effect of an external field is utilized in the metal-oxide-semiconductor field-effect transistor (MOSFET). This has a metal electrode just outside the semiconductor surface and insulated from it by a layer of oxide. A voltage, the gate voltage V_g, is applied between the metal and semiconductor that modulates the n_s, the surface charge density per unit area:

$$\Delta n_s = C_g \Delta V_g \ ,$$

where C_g is the capacitance per unit area between the metallic gate and the semiconductor. This surface charge layer forms the conducting pathway of the MOSFET. The conductance of a surface layer of length L and width W between two electrical contacts is:

$$G = (W/L) n_s e \mu \ ,$$

where μ is the carrier mobility. The carrier density n_s, and hence the conductance, is controlled by the gate voltage. This three-terminal electronic valve is

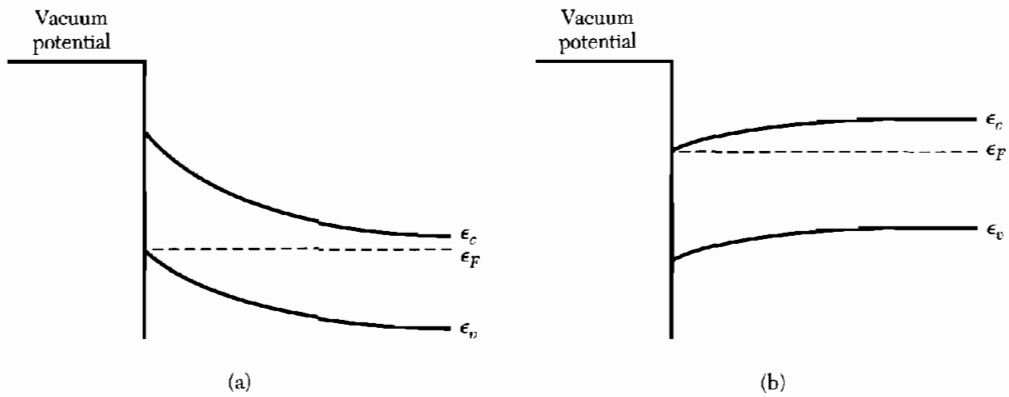


Figure 7 Band bending near a semiconductor surface that can give a highly conducting surface region. (a) Inversion layer on an n-type semiconductor. For the bending as shown, the hole concentration at the surface is far larger than the electron concentration in the interior. (b) Accumulation layer on an n-type semiconductor, with an electron concentration at the surface that is far higher than in the interior.

a principal component in microelectronic systems. The electronic states occupied by the carriers at the surface are quantized along the direction normal to the interface, as treated in Problem 2.

MAGNETORESISTANCE IN A TWO-DIMENSIONAL CHANNEL

The static magnetoconductivity tensor in 3D was found in Problem 6.9. Here we translate that result to a 2D surface conductance channel in the xy plane, with the static magnetic field in the z direction, normal to the MOS layer. We assume the surface density of electrons is $n_s = N/L^2$. The surface conductance is defined as the volume conductivity times the layer thickness. The surface current density is defined as the current crossing a line of unit length in the surface.

Thus, with (6.43) and (6.65) the surface tensor conductance components become

$$\sigma_{xx} = \frac{\sigma_0}{1 + (\omega_c\tau)^2} \; ; \qquad \sigma_{xy} = \frac{\sigma_0\omega_c\tau}{1 + (\omega_c\tau)^2} \; , \tag{16}$$

where $\sigma_0 = n_s e^2\tau/m$ and $\omega_c = eB/mc$ in CGS and eB/m in SI. The following discussion is written in CGS only, except where ohms are used.

These results apply specifically in the relaxation time approximation used in Chapter 6. When $\omega_c\tau \gg 1$, as for strong magnetic field and low temperatures, the surface conductivity components approach the limits

$$\sigma_{xx} = 0 \; ; \qquad \sigma_{xy} = n_s ec/B \; . \tag{17}$$

The limit for σ_{xy} is a general property of free electrons in crossed electric E_y and magnetic fields B_z. We establish the result that such electrons drift in the x direction with velocity $v_D = cE_y/B_z$. Consider the electrons from a Lorentz frame that moves in the x direction with this velocity. By electromagnetic theory there is in this frame an electric field $E_y' = -v_D B_z/c$ that will cancel the applied field E_y for the above choice of v_D. Viewed in the laboratory frame, all electrons drift in the x direction with velocity v_D in addition to any velocity components they had before E_y was applied.

Thus $j_x = \sigma_{xy}E_y = n_s ev_D = (n_s ec/B)E_y$, so that

$$\sigma_{xy} = n_s ec/B \tag{18}$$

as in (17). The experiments measure the voltage V in the y direction and the current I in the x direction (Fig. 8). Here $I_x = j_x L_y = (n_s ec/B)(E_y L_y) = (n_s ec/B)V_y$. The Hall resistance is

$$\rho_H = V_y/I_x = B/n_s ec \; . \tag{18a}$$

We see that j_x can flow with zero E_x, so that the effective conductance j_x/E_x can be infinite. Paradoxically, this limit occurs only when σ_{xx} and σ_{yy} are zero. Consider the tensor relations

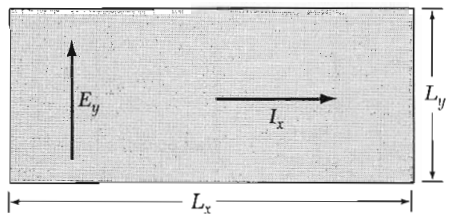


Figure 8 Applied field E_y and drift current I_x in a quantum Hall effect (IQHE) experiment.

$$j_x = \sigma_{xx}E_x + \sigma_{xy}E_y \ ; \qquad j_y = \sigma_{yx}E_x + \sigma_{yy}E_y \ . \tag{19}$$

In the Hall effect geometry $j_y = 0$, so that $E_y = (\sigma_{xy}/\sigma_{yy})E_x$, with $\sigma_{xy} = -\sigma_{yx}$. Thus

$$j_x = (\sigma_{xx} + \sigma_{xy}^2/\sigma_{yy})E_x = \sigma(\text{eff})E_x \ , \tag{20}$$

and in the limit $\sigma_{xx} = \sigma_{yy} = 0$ the effective conductance is infinite.

Integral Quantized Hall Effect (IQHE)

The results of the original measurements[1] under quantum conditions of temperature and magnetic field are shown in Fig. 9. The results are remarkable: at certain values of the gate voltage the voltage drop in the direction of current flow goes essentially to zero, as if the effective conductance were infinite. Further, there are plateaus of the Hall voltage near these same values of gate voltage, and the values of the Hall resistivity V_H/I_x at these plateaus are accurately equal to (25,813/integer) ohms, where 25,813 is the value of h/e^2 expressed in ohms.

The IQHE voltage minima V_{pp} may be explained on a model that is, however, oversimplified. Later we give a general theory. Apply a strong magnetic field such that the separation $\hbar\omega_c \gg k_BT$. It is meaningful to speak of Landau levels that are completely filled or completely empty. Let the electron surface concentration (proportional to the gate voltage) be adjusted to any of the set of values that cause the Fermi level to fall at a Landau level: from (9.33) and (9.34),

$$seB_s/hc = n_s \ , \tag{21}$$

where s is any integer and n_s is the electron surface concentration.

When the above conditions are satisfied, the electron collision time is greatly enhanced. No elastic collisions are possible from one state to another state in the same Landau level because all possible final states of equal energy are occupied. The Pauli principle prohibits an elastic collision. Inelastic collisions to a vacant Landau level are possible with the absorption of the necessary energy from a phonon, but there are very few thermal phonons of energy greater than the interlevel spacing by virtue of the assumption $\hbar\omega_c \gg k_BT$.

[1]K. von Klitzing, G. Dorda, and M. Pepper, Phys. Rev. Lett. **45**, 494 (1980).

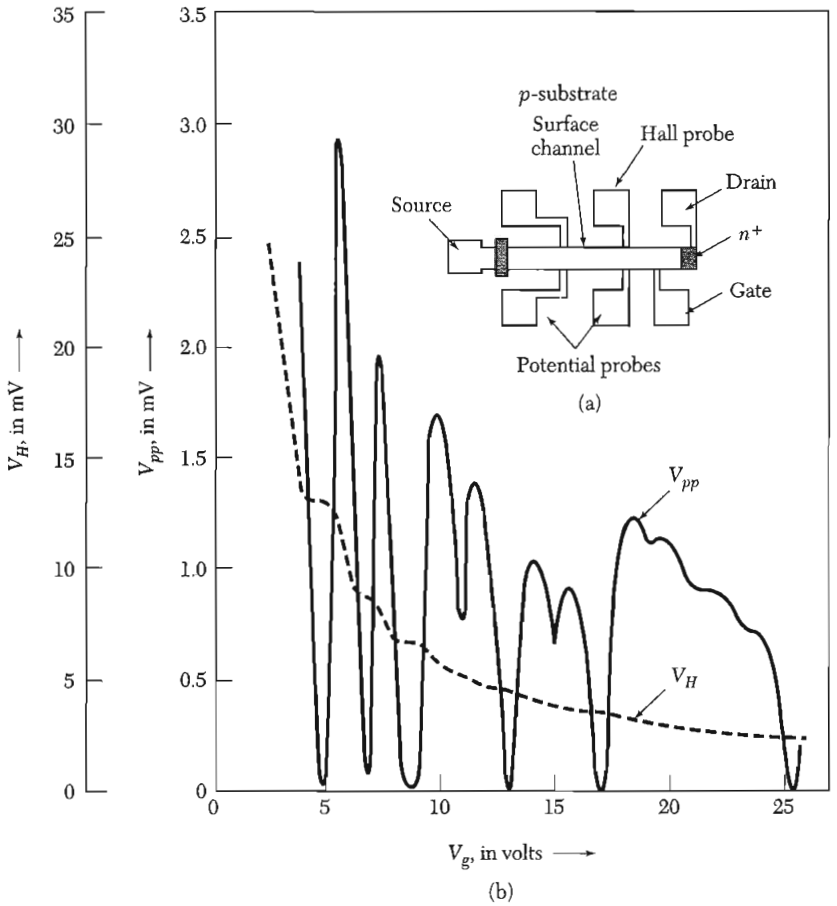


Figure 9 In the original IQHE measurements a magnetic field of 180 kG (18 T) points out of the paper. The temperature is 1.5 K. A constant current of 1 μA is made to flow between the source and the drain. Voltages V_{pp} and V_H are plotted versus the gate voltage V_g, which is proportional to the Fermi level. (After K. von Klitzing, G. Dorda, and M. Pepper.)

The quantization of the Hall resistance follows on combining (18a) and (21):

$$\rho_H = h/se^2 = 2\pi/sc\alpha \ , \tag{22}$$

where α is the fine structure constant $e^2/\hbar c \cong 1/137$, and s is an integer.

IQHE in Real Systems

The measurements (Fig. 9) suggest that the above theory of the IQHE is too good. The Hall resistivity is accurately quantized at 25,813/s ohms, whether or not the semiconductor is of very high purity and perfection. The sharp Landau levels (Fig. 10a) are broadened in the real crystal (Fig. 10b), but this does not affect the Hall resistivity. The occurrence of plateaus in the Hall

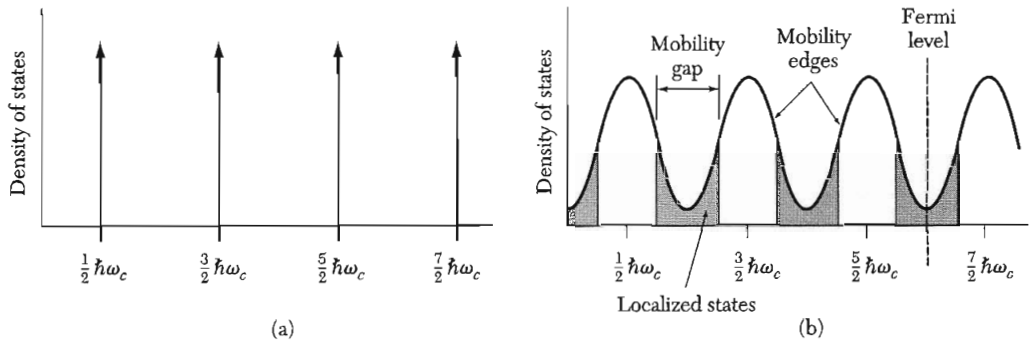


Figure 10 Density of states in a 2D electron gas in a strong magnetic field. (a) Ideal 2D crystal. (b) Real 2D crystal, with impurities and other imperfections.

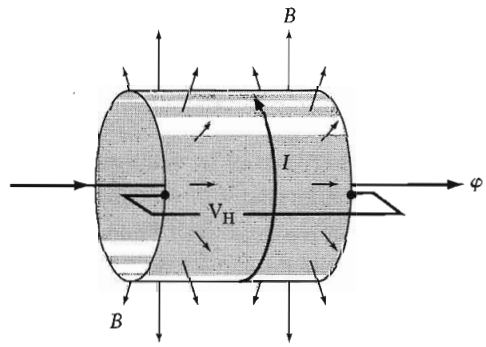


Figure 11 Geometry for Laughlin's thought-experiment. The 2D electron system is wrapped around to form a cylinder. A strong magnetic field B pierces the cylinder everywhere normal to its surface. A current I circles the loop, giving rise to the Hall voltage V_H and a small magnetic flux φ through the loop.

resistance, evident in the U_H curve of Fig. 9, is not expected in ideal systems because partially filled Landau levels will exist for all gate voltages except those for which the Fermi level exactly coincides with a Landau level. Yet the experiments show that a range of V_g values gives the exact Hall resistance.

Laughlin[2] interpreted the results for real systems as the expression of the general principle of gauge invariance. The argument is subtle and somewhat reminiscent of the flux quantization in a superconductor in Chapter 10.

In Laughlin's thought-experiment the 2D electron system is bent to form a cylinder (Fig. 11) whose surface is pierced everywhere by a strong magnetic field B normal to the surface. The current I (former I_x) circles the loop. The magnetic field B acts on the charge carriers to produce a Hall voltage V_H (former V_y) perpendicular to the current and to B; that is, V_H is developed between one edge of the cylinder and the other.

The circulating current I is accompanied by a small magnetic flux φ that threads the current loop. The aim of the thought-experiment is to find the

[2]R. B. Laughlin, Phys. Rev. B **23**, 5632 (1981); see also his article in the *McGraw-Hill year-book of science and technology*, 1984, pp. 209–214. A review is given by H. L. Stormer and D. C. Tsui, Science **220**, 1241 (1983).

relation between I and V_H. We start with the electromagnetic relation that relates I to the total energy U of a resistanceless system:

$$\frac{\partial U}{\partial t} = -V_x I_x = \frac{I}{c}\frac{\partial \varphi}{\partial t} \; ; \qquad I = c\frac{\delta U}{\delta \varphi} \; . \tag{23}$$

The value of I can now be found from the variation δU of the electronic energy that accompanies a small variation $\delta\varphi$ of the flux.

The carrier states divide into two classes:

- Localized states, which are not continuous around the loop.
- Extended states, continuous around the loop.

Localized and extended states cannot coexist at the same energy, according to our present understanding of localization.

The two classes of states respond differently to the application of the flux φ. The localized states are unaffected to first order because they do not enclose any significant part of φ. To a localized state a change in φ looks like a gauge transformation, which cannot affect the energy of the state.

The extended states enclose φ, and their energy may be changed. However, if the magnetic flux is varied by a flux quantum, $\delta\varphi = hc/e$, all extended orbits are identical to those before the flux quantum was added. The argument here is identical to that for the flux quantization in the superconducting ring treated in Chapter 10, but with the $2e$ of the Cooper pair replaced by e.

If the Fermi level falls within the localized states of Fig. 10b, all extended states (Landau levels) below the Fermi level will be filled with electrons both before and after the flux change $\delta\varphi$. However, during the change an integral number of states, generally one per Landau level, enter the cylinder at one edge and leave it at the opposite edge.

The number must be integral because the system is physically identical before and after the flux change. If the transferred state is transferred while occupied by one electron, it contributes an energy change eV_H; if N occupied states are transferred, the energy change is NeV_H.

This electron transfer is the only way the degenerate 2D electron system can change its energy. We can understand the effect by looking at a model system without disorder in the Landau gauge for the vector potential:

$$\mathbf{A} = -By\hat{\mathbf{x}} \; . \tag{24}$$

An increase δA that corresponds to the flux increase $\delta\varphi$ is equivalent to a displacement of an extended state by $\delta A/B$ in the y direction. By the Stokes theorem and the definition of the vector potential we have $\delta\varphi = L_x\delta A$. Thus $\delta\varphi$ causes a motion of the entire electron gas in the y direction.

By $\delta U = NeV_H$ and $\delta\varphi = hc/e$, we have

$$I = c(\Delta U/\Delta\varphi) = cNe^2V_H/hc = (Ne^2/h)V_H \; , \tag{25}$$

so that the Hall resistance is

$$\rho_H = V_H/I = h/Ne^2 \ , \tag{26}$$

as in (22).

Fractional Quantized Hall Effect (FQHE). A quantized Hall effect has been reported for similar systems at fractional values of the index s, by working at lower temperatures and higher magnetic fields. In the extreme quantum limit the lowest Landau level is only partially occupied, and the integral QHE treated above should not occur. It has been observed,[3] however, that the Hall resistance ρ_H is quantized in units of $3h/e^2$ when the occupation of the lowest Landau level is 1/3 and 2/3, and ρ_{xx} vanishes for these occupations. Similar breaks have been reported for occupations of 2/5, 3/5, 4/5, and 2/7.

p-n JUNCTIONS

A *p-n* junction is made from a single crystal modified in two separate regions. Acceptor impurity atoms are incorporated into one part to produce the p region in which the majority carriers are holes. Donor impurity atoms in the other part produce the n region in which the majority carriers are electrons. The interface region may be less than 10^{-4} cm thick. Away from the junction region on the p side there are $(-)$ ionized acceptor impurity atoms and an equal concentration of free holes. On the n side there are $(+)$ ionized donor atoms and an equal concentration of free electrons. Thus the majority carriers are holes on the p side and electrons on the n side, Fig. 12.

Holes concentrated on the p side would like to diffuse to fill the crystal uniformly. Electrons would like to diffuse from the n side. But diffusion will upset the local electrical neutrality of the system.

A small charge transfer by diffusion leaves behind on the p side an excess of $(-)$ ionized acceptors and on the n side an excess of $(+)$ ionized donors. This charge double layer creates an electric field directed from n to p that inhibits diffusion and thereby maintains the separation of the two carrier types. Because of this double layer the electrostatic potential in the crystal takes a jump in passing through the region of the junction.

In thermal equilibrium the chemical potential of each carrier type is everywhere constant in the crystal, even across the junction. For holes

$$k_B T \ln p(\mathbf{r}) + e\varphi(\mathbf{r}) = \text{constant} \tag{27a}$$

[3]D. C. Tsui, H. L. Stormer, and A. C. Gossard, Phys. Rev. Lett. **48**, 1562 (1982); A. M. Chang et al., Phys. Rev. Lett. **53**, 997 (1984). For a discussion of the theory see R. Laughlin in G. Bauer et al., eds., *Two-dimensional systems, heterostructures, and superlattices*, Springer, 1984.

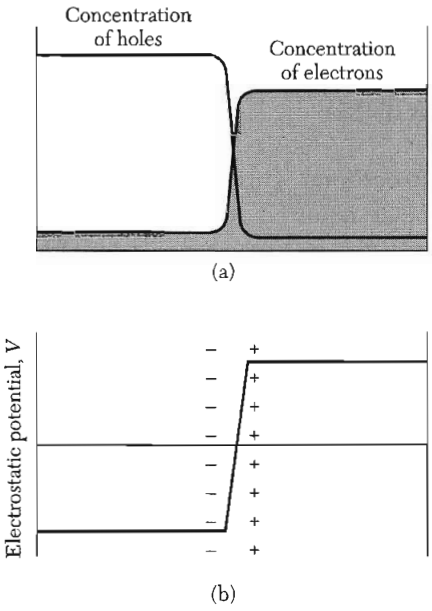


Figure 12 (a) Variation of the hole and electron concentrations across an unbiased (zero applied voltage) junction. The carriers are in thermal equilibrium with the acceptor and donor impurity atoms, so that the product pn of the hole and electron concentrations is constant throughout the crystal in conformity with the law of mass action. (b) Electrostatic potential from acceptor (−) and donor (+) ions near the junction. The potential gradient inhibits diffusion of holes from the p side to the n side, and it inhibits diffusion of electrons from the n side to the p side. The electric field in the junction region is called the built-in electric field.

across the crystal, where p is the hole concentration and φ the electrostatic potential. Thus p is low where φ is high. For electrons

$$k_B T \ln n(\mathbf{r}) - e\varphi(\mathbf{r}) = \text{constant} \ , \tag{27b}$$

and n will be low where φ is low.

The total chemical potential is constant across the crystal. The effect of the concentration gradient exactly cancels the electrostatic potential, and the net particle flow of each carrier type is zero. However, even in thermal equilibrium there is a small flow of electrons from n to p where the electrons end their lives by combination with holes. The recombination current J_{nr} is balanced by a current J_{ng} of electrons which are generated thermally in the p region and which are pushed by the built-in field to the n region. Thus in zero external applied electric field

$$J_{nr}(0) + J_{ng}(0) = 0 \ , \tag{28}$$

for otherwise electrons would accumulate indefinitely on one side of the barrier.

Rectification

A p-n junction can act as a rectifier. A large current will flow if we apply a voltage across the junction in one direction, but if the voltage is in the opposite direction only a very small current will flow. If an alternating voltage is applied across the junction the current will flow chiefly in one direction—the junction has rectified the current (Fig. 13).

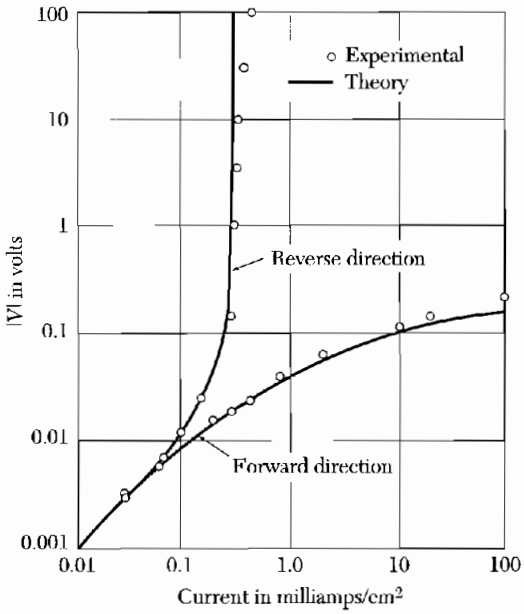


Figure 13 Rectification characteristic of a p-n junction in germanium, after Shockley.

For back voltage bias a negative voltage is applied to the p region and a positive voltage to the n region, thereby increasing the potential difference between the two regions. Now practically no electrons can climb the potential energy hill from the low side of the barrier to the high side. The recombination current is reduced by the Boltzmann factor:

$$J_{nr}(V \text{ back}) = J_{nr}(0) \exp(-e|V|/k_B T) \ . \tag{29}$$

The Boltzmann factor controls the number of electrons with enough energy to get over the barrier.

The thermal generation current of electrons is not particularly affected by the back voltage because the generation electrons flow downhill (from p to n) anyway:

$$J_{ng}(V \text{ back}) = J_{ng}(0) \ . \tag{30}$$

We saw in (28) that $J_{nr}(0) = -J_{ng}(0)$; thus the generation current dominates the recombination current for a back bias.

When a forward voltage is applied, the recombination current increases because the potential energy barrier is lowered, thereby enabling more electrons to flow from the n side to the p side:

$$J_{nr}(V \text{ forward}) = J_{nr}(0) \exp(e|V|/k_B T) \ . \tag{31}$$

Again the generation current is unchanged:

$$J_{ng}(V \text{ forward}) = J_{ng}(0) \ . \tag{32}$$

The hole current flowing across the junction behaves similarly to the electron current. The applied voltage which lowers the height of the barrier for electrons also lowers it for holes, so that large numbers of electrons flow from the n region under the same voltage conditions that produce large hole currents in the opposite direction.

The electric currents of holes and electrons are additive, so that the total forward electric current is

$$I = I_s[\exp(eV/k_BT)-1] \ , \tag{33}$$

where I_s is the sum of the two generation currents. This equation is well satisfied for p-n junctions in germanium (Fig. 13), but not quite as well in other semiconductors.

Solar Cells and Photovoltaic Detectors

Let us shine light on a p-n junction, one without an external bias voltage. Each absorbed photon creates an electron and a hole. When these carriers diffuse to the junction, the built-in electric field of the junction separates them at the energy barrier. The separation of the carriers produces a forward voltage across the barrier: forward, because the electric field of the photoexcited carriers is opposite to the built-in field of the junction.

The appearance of a forward voltage across an illuminated junction is called the **photovoltaic effect**. An illuminated junction can deliver power to an external circuit. Large area p-n junctions of silicon are used as solar panels to convert solar photons to electrical energy.

Schottky Barrier

When a semiconductor is brought into contact with a metal, there is formed in the semiconductor a barrier layer from which charge carriers are severely depleted. The barrier layer is also called a depletion layer or exhaustion layer.

In Fig. 14 an n-type semiconductor is brought into contact with a metal. The Fermi levels are coincident after the transfer of electrons to the conduction band of the metal. Positively charged donor ions are left behind in this region that is practically stripped of electrons. Here the Poisson equation is

$$\text{(CGS)} \quad \operatorname{div} \mathbf{D} = 4\pi ne \qquad \text{(SI)} \quad \operatorname{div} \mathbf{D} = ne/\epsilon_0 \ , \tag{34}$$

where n is the donor concentration. The electrostatic potential is determined by

$$\text{(CGS)} \quad d^2\varphi/dx^2 = -4\pi ne/\epsilon \qquad \text{(SI)} \quad d^2\varphi/dx^2 = ne/\epsilon\epsilon_0 \ , \tag{35}$$

which has a solution of the form

$$\text{(CGS)} \quad \varphi = -(2\pi ne/\epsilon)x^2 \qquad \text{(SI)} \quad \varphi = -(ne/2\epsilon\epsilon_0)x^2 \ . \tag{36}$$

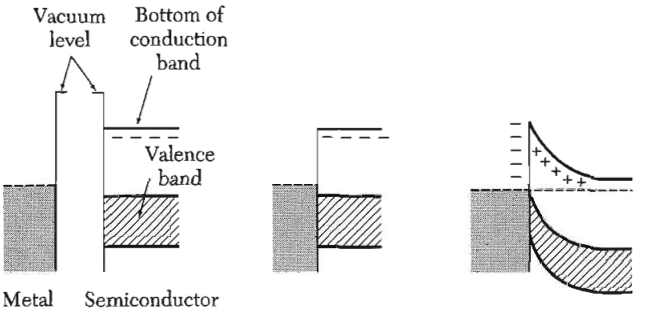


Figure 14 Rectifying barrier between a metal and an n-type semiconductor. The Fermi level is shown as a broken line.

The origin of x has been taken for convenience at the right-hand edge of the barrier. The contact is at $-x_b$, and here the potential energy relative to the right-hand side is $-e\varphi_0$, whence the thickness of the barrier is

$$\text{(CGS)} \quad x_b = (\epsilon|\varphi_0|/2\pi ne)^{1/2} \qquad \text{(SI)} \quad x_b = (2\epsilon\epsilon_0|\varphi_0|/ne)^{1/2}. \tag{37}$$

With $\epsilon = 16$; $e\varphi_0 = 0.5$ eV; $n = 10^{16}$ cm^{-3}, we find $x_b = 0.3$ μm. This is a somewhat simplified view of the metal-semiconductor contact.

HETEROSTRUCTURES

Semiconductor heterostructures are layers of two or more different semiconductors grown coherently with one common crystal structure. Heterostructures offer extra degrees of freedom in the design of semiconductor junction devices, because both the impurity doping and the conduction and valence band offsets at the junction can be controlled. Because of this freedom many devices that utilize compound semiconductors incorporate heterostructures. Examples include semiconductor lasers in CD players and high-speed devices for cell-phone systems.

A heterostructure may be viewed as a single crystal in which the occupancy of the atomic sites changes at the interface. As an example, one side of the interface can be Ge and the other side GaAs: both lattice constants are 5.65 Å. One side has the diamond structure and the other side the cubic zinc sulfide structure. Both structures are built up from tetrahedral covalent bonds and fit together coherently as if they were a single crystal. There are a few edge dislocations (Chapter 21) to relieve the strain energy near the interface.

The band gaps, however, are different, and this difference is the source of the real interest in the heterostructure, apart from the technical virtuosity in

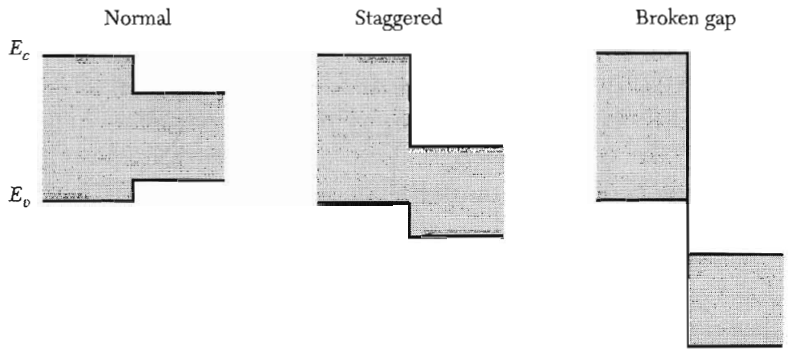


Figure 15 Three types of band edge offsets at hetero-interfaces. The forbidden gaps are shown shaded. The offset called normal occurs, for example, in GaAs/(Al,Ga)As. The "broken-gap" offset occurs in the GaSb/InAs heterojunction.

forming the structure. The band gaps are 0.67 eV for Ge and 1.43 eV for GaAs, at 300 K. The relative alignment of the conduction and valence band edges offers several possibilities, as shown in Fig. 15. Calculations suggest that the top of the valence band E_v in Ge should lie about 0.42 eV higher than in GaAs. The bottom of the conduction band E_c in Ge should lie about 0.35 eV lower than in GaAs, so that the offsets are classified as normal in the scheme of Fig. 15.

Band edge offsets act as potential barriers in opposite senses on electrons and holes. Recall that electrons lower their energy by "sinking" on an energy band diagram, whereas holes lower their energy by "floating" on the same diagram. For the normal alignment both electrons and holes are pushed by the barrier from the wide-gap to the narrow-gap side of the heterostructure.

Other important semiconductor pairs used in heterostructure are AlAs/GaAs, InAs/GaSb, GaP/Si, and ZnSe/GaAs. Good lattice matching in the range 0.1–1.0 percent is often accomplished by use of alloys of different elements, which may also adjust energy gaps to meet specific device needs.

n-N Heterojunction

As a practical example, consider two *n*-type semiconductors with a large offset of the two conduction bands, as sketched in Fig. 16a for a semiconductor pair with a normal band line-up. The *n*-type material with the higher conduction band edge is labeled with a capital letter as *N*-type, and the junction shown is called an *n-N* junction. The electron transport properties across the junction are similar to those across a Schottky barrier. Far from the interface the two semiconductors must be electrically neutral in composition. However, the two Fermi levels, each determined by the doping, must coincide if there is to be zero net electron transport in the absence of an external bias voltage.

These two considerations fix the "far-off" conduction band edge energies relative to the Fermi level, as in Fig. 16b. The combination of a specified band

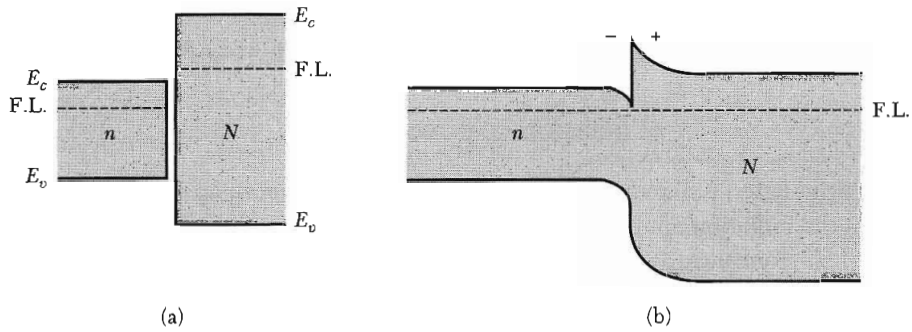


Figure 16 (a) Two semiconductors not in contact; the absolute band edge energies are labeled E_c for the conduction band edge and E_v for the valence band edge. An "absolute energy" means referred to infinite distance. The Fermi levels in the two materials are determined by the donor concentrations, as well as by the band structure. (b) The same semiconductors as a heterojunction, so that the two parts are in diffusive equilibrium. This requires that the Fermi level (F.L.) be independent of position, which is accomplished by transfer of electrons from the N-side to the n-side of the interface. A depletion layer of positively ionized donors is left behind on the N-side.

offset (determined by the host material composition) at the interface and the distant band energies (determined by the Fermi level) can be reconciled only if the bands bend near the interface, as in the figure. The necessary band bending is created by space charges consequent to the transfer of electrons from the N-side to the lower n-side. This transfer leaves behind on the N-side a positive donor space charge layer, which through the Poisson equation of electrostatics is the source of the positive second derivative (upward curvature) in the conduction band edge energy on that side.

On the n-side there is now a negative space charge because of the excess of electrons on that side. The layer of negative space charge gives a negative second derivative (downward curvature) in the conduction band edge energy. On the n-side the band as a whole bends down toward the junction. This differs from the usual p-n junction. The downward bending and the potential step form a potential well for electrons. The well is the basis for the new physical phenomena characteristic of heterostructure physics.

If the doping on the n-side (low E_c) is reduced to a negligible value, there will be very few ionized donors on that side in the electron-rich layer. The mobility of these electrons is largely limited only by lattice scattering, which falls off sharply as the temperature is lowered. Low-temperature mobilities as high as $10^7\ \mathrm{cm^2\ V^{-1}s^{-1}}$ have been observed in GaAs/(Al, Ga)As.

If now the thickness of the N-side semiconductor is reduced below the depletion layer thickness on that side, the N material will be entirely depleted of its low-mobility electrons. All of the electrical conduction parallel to the interface will be carried by the high-mobility electrons on the n-side, equal in number to the number of ionized N-side donors, but spatially separated from them by the potential step. Such high-mobility structures play a large role in solid

state studies of 2D electron gases and also in new classes of high-speed field effect transistors for computer applications at low temperatures.

SEMICONDUCTOR LASERS

Stimulated emission of radiation can occur in direct-gap semiconductors from the radiation emitted when electrons recombine with holes. The electron and hole concentrations created by illumination are larger than their equilibrium concentrations. The recombination times for the excess carriers are much longer than the times for the conduction electrons to reach thermal equilibrium with each other in the conduction band, and for the holes to reach thermal equilibrium with each other in the valence band. This steady-state condition for the electron and hole populations is described by separate Fermi levels μ_c and μ_v for the two bands, called **quasi-Fermi levels**.

With μ_c and μ_v referred to their band edges, the condition for population inversion is that

$$\mu_c > \mu_v + \epsilon_g \ . \tag{38}$$

For laser action the quasi-Fermi levels must be separated by more than the band gap.

Population inversion and laser action can be achieved by forward bias of an ordinary GaAs or InP junction, but almost all practical injection lasers employ the double heterostructure proposed by H. Kroemer (Fig. 17). Here the lasing semiconductor is embedded between two wider-gap semiconductor regions of opposite doping, creating a quantum well that confines both electrons and holes. An example is GaAs embedded in (Al,Ga)As. In such a structure there is a potential barrier that prevents the outflow of electrons to the p-type region, and an opposite potential barrier that prevents the outflow of holes to the n-type region.

The value of μ_c in the optically active layer lines up with μ_n in the n contact; similarly, μ_v lines up with μ_p in the p contact. Inversion can be achieved if we apply a bias voltage larger than the voltage equivalent of the active layer energy gap. The diode wafer provides its own electromagnetic cavity, for the reflectivity at the crystal-air interface is high. Crystals are usually polished to provide two flat parallel surfaces; the radiation is emitted in the plane of the heterojunctions.

Crystals with direct band gaps are required normally for junction lasers. Indirect gaps involve phonons as well as photons; carriers recombine less efficiently because of competing processes, and no laser action has been observed in indirect gap semiconductors.

Gallium arsenide has been widely studied as the optically active layer. It emits in the near infrared at 8383 Å or 1.48 eV; the exact wavelength depends

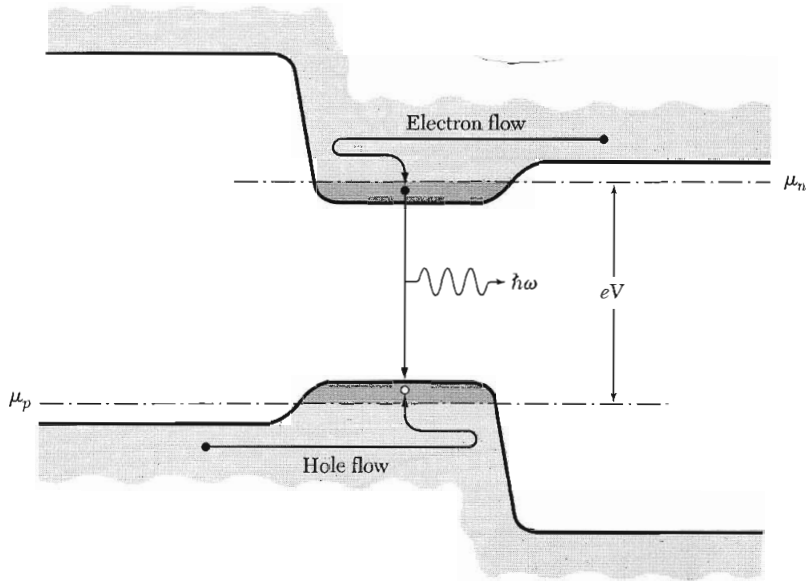


Figure 17 Double heterostructure injection laser. Electrons flow from the right into the optically-active layer, where they form a degenerate electron gas. The potential barrier provided by the wide energy gap on the p side prevents the electrons from escaping to the left. Holes flow from the left into the active layer, but cannot escape to the right.

on temperature. The gap is direct (Chapter 8). In a heterojunction the system is very efficient: the ratio of light energy output to dc electrical energy input is near 50 percent, and the differential efficiency for small changes is up to 90 percent.

The wavelength can be adjusted over a wide range in the alloy system $Ga_xIn_{1-x}P_yAs_{1-y}$, so that we can match the laser wavelength to the absorption minimum of optical fibers used as a transmission medium. The combination of double heterostructure lasers with glass fibers forms the basis of the new light-wave communication technology that is rapidly replacing transmission of signals over copper lines.

LIGHT-EMITTING DIODES

The efficiency of light-emitting diodes is now at the point of exceeding incandescent lamps. Consider a p-n junction with a voltage source V splitting the two chemical potentials μ_n and μ_p by eV, as in Figure 18. Electrons from the n side are injected into the p side, and holes from the p side are injected into the n side. These injected carriers annihilate each other across the junction, thus generating photons if the quantum efficiency is unity.

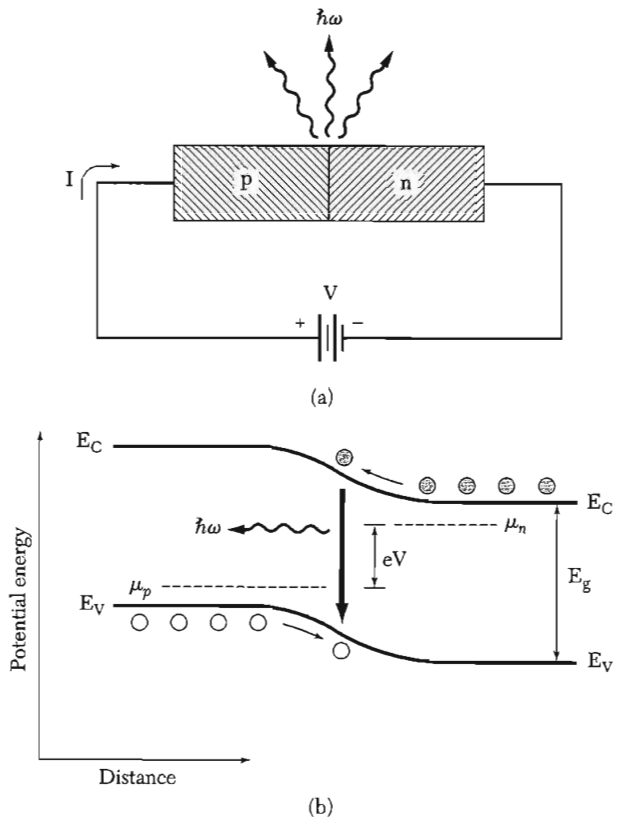


Figure 18 Electron-hole recombination into photons, across a p-n junction.

The generation or recombination process will be much stronger in a direct-gap semiconductor (Fig. 8.5a) than in an indirect gap semiconductor (Fig. 8.5b). In a direct-gap semiconductor such as GaAs, the band-to-band photons are absorbed in a distance $\approx 1\ \mu$m, which is strong absorption. The direct-gap ternary semiconductor $GaAs_{1-x}\ P_x$ gives light tuned to shorter wavelengths as the composition variable x is increased. This composition was made by Holonyak into one of the first p-n diode lasers and into the first visible-spectrum (red) LED. Blue-emitting heterostructures have now been made, such as $In_xGa_{1-x}N - Al_yGa_{1-y}N$.

The performance of LEDs has increased markedly over the years, from about 0.1 lumens/watt in 1962 to about 40 lumens/watt in 2004, compared with 15 lumens/watt for a standard white unfiltered incandescent lamp. To quote Craford and Holonyak, "We are entering an entirely new era in lighting (illumination) with an ultimate form of lamp—a direct-gap III-V alloy p-n heterostructure."

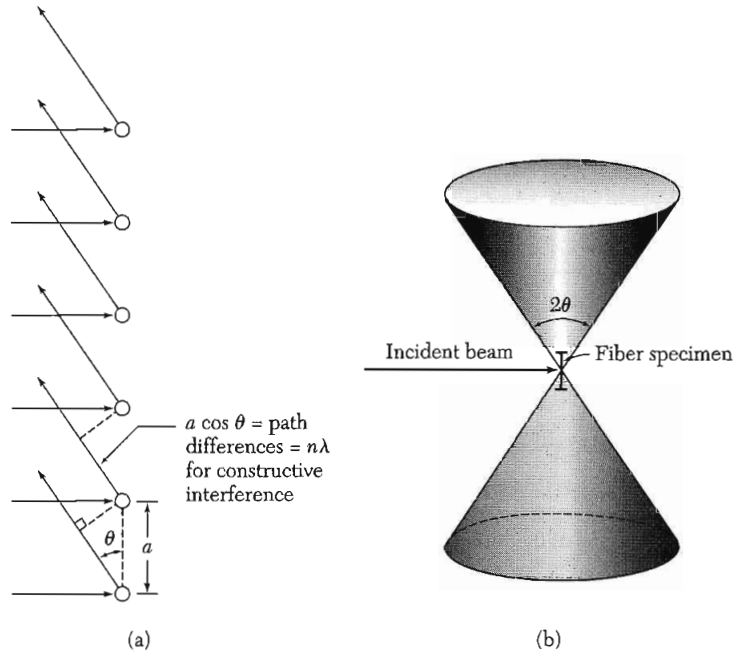
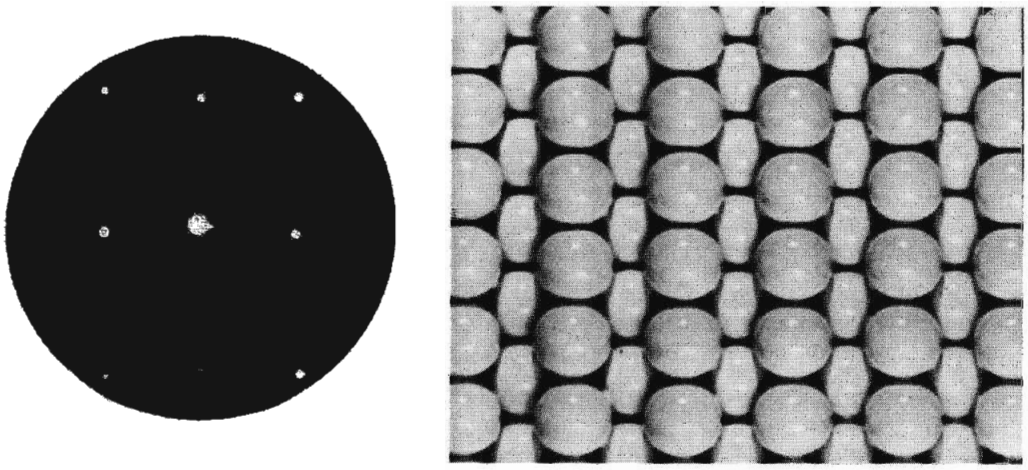


Figure 19 The diffraction pattern from a single line of lattice constant a in a monochromatic x-ray beam perpendicular to the line. (a) The condition for constructive interference is $a \cos \theta = n\lambda$, where n is an integer. (b) For given n the diffracted rays of constant λ lie on the surface of a cone.

Problems

1. ***Diffraction from a linear array and a square array.*** The diffraction pattern of a linear structure of lattice constant a is explained* in Fig. 19. Somewhat similar structures are important in molecular biology: DNA and many proteins are linear helices. (a) A cylindrical film is exposed to the diffraction pattern of Fig. 19b; the axis of the cylinder is coincident with the axis of the linear structure or fiber. Describe the appearance of the diffraction pattern on the film. (b) A flat photographic plate is placed behind the fiber and normal to the incident beam. Sketch roughly the appearance of the diffraction pattern on the plate. (c) A single plane of atoms forms a square lattice of lattice constant a. The plane is normal to the incident x-ray beam. Sketch roughly the appearance of the diffraction pattern on the photographic plate. Hint: The diffraction from a plane of atoms can be inferred from the patterns for two perpendicular lines of atoms. (d) Figure 20 shows the electron diffraction pattern in the backward direction from the nickel atoms on the (110) surface of a nickel crystal. Explain the orientation of the diffraction pattern in relation to the atomic positions of the surface atoms shown in the model. Assume that only the surface atoms are effective in the reflection of low-energy electrons.

*Another viewpoint is useful: for a linear lattice the diffraction pattern is described by the single Laue equation $\mathbf{a} \cdot \Delta\mathbf{k} = 2\pi q$, where q is an integer. The lattice sums which led to the other Laue equations do not occur for a linear lattice. Now $\mathbf{a} \cdot \Delta\mathbf{k}$ = constant is the equation of a plane; thus the reciprocal lattice becomes a set of parallel planes normal to the line of atoms.



(a) (b)

Figure 20 (a) Backward scattering pattern of 76 eV electrons incident normally on the (110) face of a nickel crystal; a model of the surface is shown in (b). (Courtesy of A. U. MacRae.)

2. ***Surface subbands in electric quantum limit.*** Consider the contact plane between an insulator and a semiconductor, as in a metal-oxide-semiconductor transistor or MOSFET. With a strong electric field applied across the SiO_2-Si interface, the potential energy of a conduction electron may be approximated by $V(x) = eEx$ for x positive and by $V(x) = \infty$ for x negative, where the origin of x is at the interface. The wavefunction is 0 for x negative and may be separated as $\psi(x,y,z) = u(x) \exp[i(k_y y + k_z z)]$, where $u(x)$ satisfies the differential equation

$$-(\hbar^2/2m)d^2u/dx^2 + V(x)u = \epsilon u \ .$$

With the model potential for $V(x)$ the exact eigenfunctions are Airy functions, but we can find a fairly good ground state energy from the variational trial function $x \exp(-ax)$. (a) Show that $\langle\epsilon\rangle = (\hbar^2/2m)a^2 + 3eE/2a$. (b) Show that the energy is a minimum when $a = (3eEm/2\hbar^2)^{1/3}$. (c) Show that $\langle\epsilon\rangle_{\min} = 1.89(\hbar^2/2m)^{1/3}(3eE/2)^{2/3}$. In the exact solution for the ground state energy the factor 1.89 is replaced by 1.78. As E is increased the extent of the wavefunction in the x direction is decreased. The function $u(x)$ defines a surface conduction channel on the semiconductor side of the interface. The various eigenvalues of $u(x)$ define what are called electric subbands. Because the eigenfunctions are real functions of x the states do not carry current in the x direction, but they do carry a surface channel current in the yz plane. The dependence of the channel on the electric field E in the x direction makes the device a field effect transistor.

3. ***Properties of the two-dimensional electron gas.*** Consider a two-dimensional electron gas (2DEG) with twofold spin degeneracy but no valley degeneracy. (a) Show that the number of orbitals per unit energy is given by: $D(\epsilon) = m/\pi\hbar^2$. (b) Show that the sheet density is related to the Fermi wavevector by: $n_s = k_F^2/2\pi$. (c) Show that, in the Drude model, the sheet resistance, i.e., the resistance of a square segment of the 2DEG, can be written as: $R_s = (h/e^2)/(k_F\ell)$ where $\ell = v_\mu\tau$ is the mean free path.

18

Nanostructures

Written by Professor Paul McEuen of Cornell University

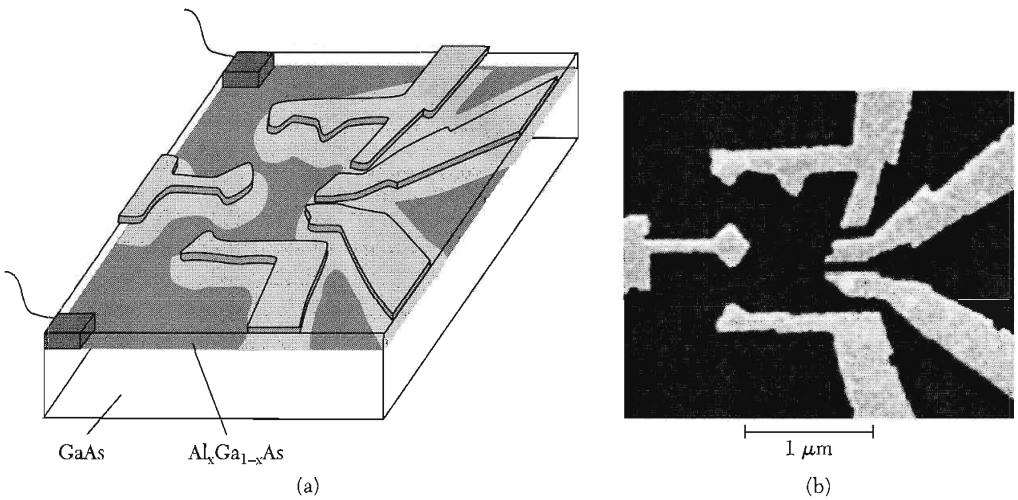


Figure 1 Schematic and scanning electron microscope (SEM) image of a gate electrode pattern on a GaAs/AlGaAs heterostructure used to create a quantum dot of complex shape in the underlying two-dimensional (2D) electron gas. (Courtesy of C. Marcus.)

The previous chapter addressed solids with spatial confinement at the nanometer scale along one direction: surfaces, interfaces, and quantum wells. These systems were effectively two-dimensional, which we define as extended in two directions but of nanometer scale in the third. Only a small number of quantized states—often only one—are occupied in the confined direction. In this chapter we discuss solids confined in either two or three orthogonal directions, creating effectively one-dimensional (1D) or zero-dimensional (0D) **nanostructures**. Important 1D examples are carbon nanotubes, quantum wires, and conducting polymers. Examples of 0D systems include semiconductor nanocrystals, metal nanoparticles, and lithographically patterned quantum dots. Some examples are shown in Figs. 1 to 3. We will almost exclusively focus on nanostructures that are created from confined periodic solids. Nonperiodic nanostructures are of great interest in other fields, such as molecular assemblies in chemistry and organic macromolecules in biology.

The techniques for the creation of nanostructures can be divided into two broad categories. Top-down approaches use lithographic patterning to structure macroscopic materials at the nanoscale, such as the metallic electrodes on top of a semiconductor heterostructure shown in Fig. 1. Bottom-up approaches utilize growth and self-assembly to build nanostructures from atomic or molecular precursors. A CdSe nanocrystal grown in solution is shown in Fig. 2. It is typically difficult to create structures smaller than 50 nm with

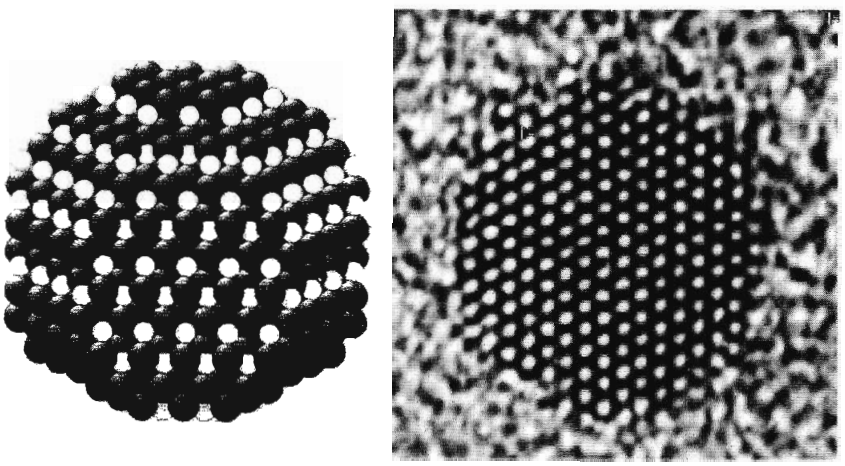


Figure 2 Model and transmission electron microscope (TEM) image of a CdSe nanocrystal. Individual rows of atoms are clearly resolved in the TEM image. (Courtesy of A. P. Alivisatos.)

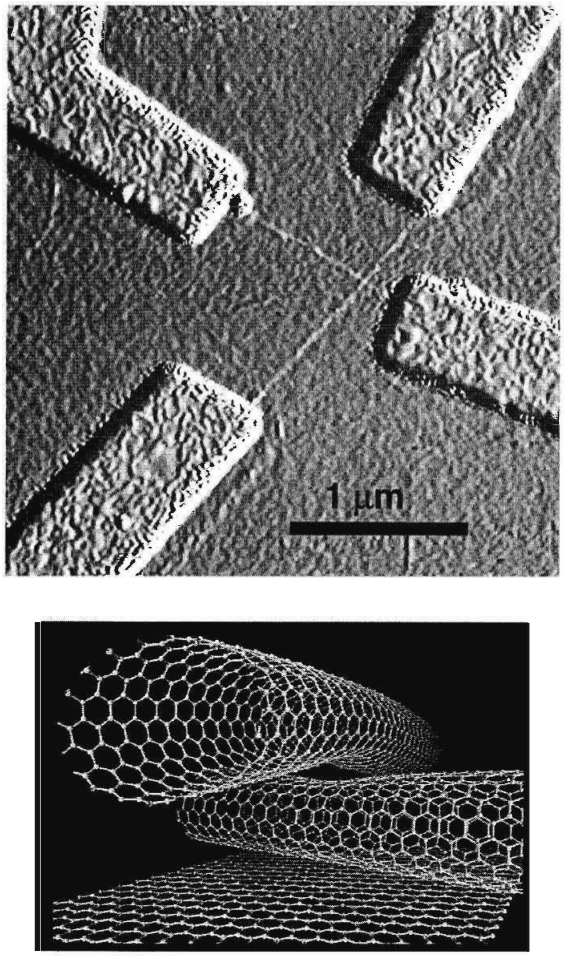


Figure 3 Atomic force microscope (AFM) image of a pair of crossed carbon nanotubes contacted by Au electrodes patterned by electron beam lithography. (Image courtesy of M. S. Fuhrer.) Also shown is a model of the nanotube cross region, showing the honeycomb lattice of the graphene sheets that form the nanotube walls. (Courtesy of P. Avouris.)

top-down techniques, while it is often difficult to create structures larger than 50 nm by bottom-up techniques. A major challenge of nanoscience and technology is to combine these approaches and develop strategies to reliably create complex systems over all length scales, from the molecular to the macroscopic. Figure 3 shows one example, where 100-nm-wide lithographic electrodes make contact to 2-nm-wide carbon nanotubes grown by chemical vapor deposition.

When the extent of a solid is reduced in one or more dimensions, the physical, magnetic, electrical, and optical properties can be dramatically altered. This makes nanostructures a subject of both fundamental and practical interest; their properties can be tailored by controlling their size and shape on

the nanometer scale. One class of effects is related to the large ratio of number of surface atoms to bulk atoms in a nanostructure. For a spherical nanoparticle of radius R composed of atoms with an average spacing a, the ratio is given by

$$N_{\text{surf}}/N \cong 3a/R \ . \tag{1}$$

For $R = 6a \sim 1$ nm, half of the atoms are on the surface. The large surface area of nanoparticles is advantageous for applications in gas storage, where molecules are adsorbed on the surfaces, or in catalysis, where reactions occur on the surface of the catalyst. It also has dramatic effects on the stability of the nanoparticle. The cohesive energy is dramatically lowered because atoms on the surface are incompletely bonded. Nanoparticles therefore melt at temperatures far below the melting temperature of the corresponding bulk solid.

The fundamental electronic and vibrational excitations of a nanostructure also become quantized, and these excitations determine many of the most important properties of the nanostructured material. These quantization phenomena will be the primary subject of this chapter. Typically, they are important in the 1–100-nanometer size range.

IMAGING TECHNIQUES FOR NANOSTRUCTURES

The development of new techniques to image and probe nanostructures has been essential to the evolution of the field. For periodic 3D structures, the diffraction of electrons or X-rays can be used to determine structure in reciprocal space, which can then be inverted to find the real-space atomic arrangements, as discussed in Chapter 2. For individual nanoscale solids, diffraction is only of limited utility for both fundamental and practical reasons. The solid's small size interrupts the periodicity of the lattice, blurring diffraction peaks, and also produces a very small scattered signal.

Real-space probes that can directly determine the properties of the nanostructure are therefore very valuable. These probes use the interaction of a particle, typically an electron or photon, with the object under study, to create an image. The techniques fall into two major classes, which we will refer to as focal and scanned probe.

In focal microscopy, the probe particle is focused by a series of lenses onto the sample. Figure 4 shows a schematic. The ultimate resolution of the system is limited by the wavelike nature of the particle through the Heisenberg Uncertainty Principle, or, equivalently, diffraction. This smallest feature spacing d that can be resolved is given by

$$d \approx \lambda/2\beta \ , \tag{2}$$

where λ is the wavelength of the probe particle and $\beta = \sin\theta$ is the numerical aperture defined in Fig. 4. Achieving nanoscale resolution requires using particles with small wavelengths and maximizing the numerical aperture.

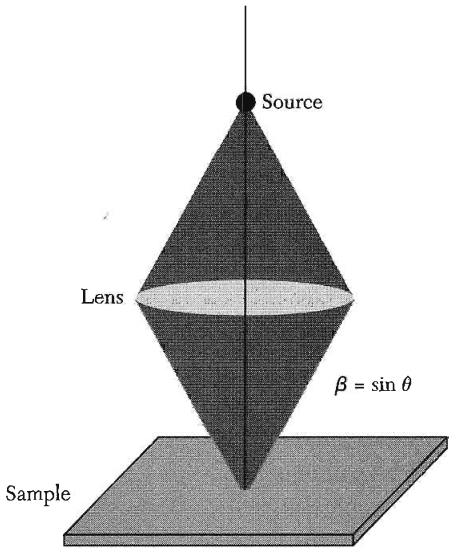


Figure 4 Schematic diagram of a focal microscope. A beam emitted from a source is focused onto the sample by a series of lenses. An equivalent focal system can be used to focus particle/waves emitted from the sample onto a detector.

In scanned probe microscopy, by contrast, a tiny probe is brought close to the sample and scanned over its surface. The resolution of the microscope is determined by the effective range of the interaction between the probe and the structure under study, rather than by the wavelength of the probe particle.

In addition to imaging, scanned and focal probes provide information about the electrical, vibrational, optical, and magnetic properties of individual nanostructures. Of particular importance is the electronic structure, expressed in the density of states. For a finite-sized system, the density of states is a series of delta functions

$$D(\varepsilon) = \sum_j \delta(\varepsilon - \varepsilon_j) \ , \tag{3}$$

where the sum is taken over all the energy eigenstates of the system. For extended solids, the density of states can be represented by a continuous function, but for a nanostructure the discrete sum form is necessary along the confined directions. This quantized density of states determines many of the most important properties of nanostructures, and it can be directly measured using the techniques described below.

Electron Microscopy

A very powerful focal tool is the electron microscope. A collimated beam of electrons is accelerated by high voltages and focused through a series of electrostatic or magnetic lenses onto the sample under study.

In **transmission electron microscopy**, or TEM, the electron beam travels through the sample and is focused on a detector plate in much the same

way as the image is focused onto the eyepiece of an optical microscope. The ultimate resolving power d is set by the wavelength of the accelerated electrons

$$d = \lambda/2\beta \cong 0.6\,\text{nm}/(\beta\sqrt{V}) \ , \tag{4}$$

where V is the accelerating voltage (measured in volts). For typical accelerating voltages (100 kV), the theoretical resolving power is therefore subatomic. Other effects, such as imperfections in the lenses, keep the TEM resolution well above this limit, but $d \sim 0.1$ nm has been achieved. Figure 2 shows a TEM image of a semiconductor nanocrystal, where rows of atoms are clearly resolved.

A major limitation of TEM is that the electron beam must penetrate the sample, making it impossible to examine structures on solid substrates. This problem is overcome in the **scanning electron microscope** (SEM). In an SEM, a high-energy (100 V to 100 kV), tightly focused electron beam is scanned over the sample. The number of backscattered electrons and/or the secondary electrons generated by the beam that emerge from the sample depends on the local composition and topography of the sample. These electrons are collected by an electron detector, and an image is formed by plotting this detector signal as a function of the beam location. This powerful technique can be used on most kinds of samples, but it typically has a lower resolution (>1 nm) than the TEM. Figure 1 is an SEM image of metallic electrodes on a GaAs/AlGaAs substrate.

In addition to imaging, the SEM beam can be used to expose an electron-sensitive material and draw small features in a technique known as **electron beam lithography.** The ultimate resolution (<10 nm) is very high, but it is a slow process because the patterns must be drawn pixel by pixel. It is therefore used primarily in research, prototyping, and optical mask fabrication.

Optical Microscopy

The optical microscope is the prototypical focal instrument. Using visible light and a high numerical aperture ($\beta \approx 1$), the highest obtainable resolution is 200–400 nm. For direct imaging, optical microscopy therefore only reaches the edge of the nanoscale realm. However, many of the optical spectroscopies discussed in Chapter 15 have been successfully adapted to study individual nanostructures. These include elastic light scattering, absorption, luminescence, and Raman scattering. Measurements of a single nanostructure, or even a single molecule, are possible if only one is in the field of view of the microscope.

Here we briefly review the emission and absorption of electromagnetic radiation by matter in a manner suitable for applications to nanostructures. Within the electric dipole approximation, Fermi's golden rule gives the transition rate between an initial state i and a higher energy state j due to absorption:

$$w_{i\to j} = (2\pi/\hbar)\left|\langle j|e\mathbf{E}\cdot\mathbf{r}|i\rangle\right|^2 \delta(\varepsilon_j - \varepsilon_i - \hbar\omega) \ . \tag{5}$$

Transitions therefore occur between states that have a nonzero dipole matrix element and whose energies differ by the absorbed photon energy $\hbar\omega$. Similarly, the emission rate from state j to i is given by

$$w_{j\to i} = (2\pi/\hbar)\left|\langle j|e\mathbf{E}\cdot\mathbf{r}|i\rangle\right|^2\delta(\varepsilon_i - \varepsilon_j + \hbar\omega) + (4\alpha\omega_{ji}^3/c^2)\left|\langle j|\mathbf{r}|i\rangle\right|^2 , \qquad (6)$$

where $\omega_{ji} = (\varepsilon_j - \varepsilon_i)/\hbar$ and α is the fine structure constant. The first and second terms represent stimulated and spontaneous emission, respectively.

By summing over all possible states, these relations can be used to calculate the total power $\sigma' E^2$ absorbed from the electromagnetic field and hence the real part of the conductivity:

$$\sigma'(\omega) = (\pi e^2\omega/V)\sum_{i,j}\left|\langle j|\hat{\mathbf{n}}\cdot\mathbf{r}|i\rangle\right|^2[f(\varepsilon_i) - f(\varepsilon_j)]\delta(\varepsilon_j - \varepsilon_i - \hbar\omega) , \qquad (7)$$

where $\hat{\mathbf{n}}$ is a unit vector pointing in the direction of the electric field. The absorption is proportional to the joint density of all initial and final states separated by an energy $\hbar\omega$, weighted by the dipole matrix element and the occupation factors of the states. The Fermi functions indicate that absorption only occurs when the initial state i is filled and the final state j is empty.

The above relations show that absorption and emission can be used to probe the electronic energy level spectra of nanostructures. Measurements can readily be performed on macroscopic collections of nominally identical nanostructures, but the effects of inhomogenous broadening due to the variation in the properties of the individual nanostructures are significant. Furthermore, sometimes only a few or even a single nanostructure is available for measurement. Optical measurements that probe single nanostructures have therefore proven to be particularly valuable.

Figure 5 shows an example of spontaneous emission, or fluorescence, from individual optically excited semiconductor quantum dots. The emission occurs from the lowest energy state in the conduction band to the highest energy state in the valence band. The linewidths of the emission lines from single nanocrystals are very narrow, but they are distributed over a range of energies due to variations in the nanocrystal size, shape, and local environment. Measurements of an ensemble therefore show a broad peak that does not accurately reflect the properties of a single nanocrystal.

In addition to their use in probing nanostructures, optical focal systems are also widely used for microfabrication. In projection photolithography, a pattern on a mask is projected onto a photosensitive resist using optical elements. Following exposure and development of the resist, the pattern is transferred into the material of interest by etching or deposition through the resist stencil. Optical lithography is the basis for the mass-fabrication of microelectronic and micromechanical systems. By using wavelengths into the

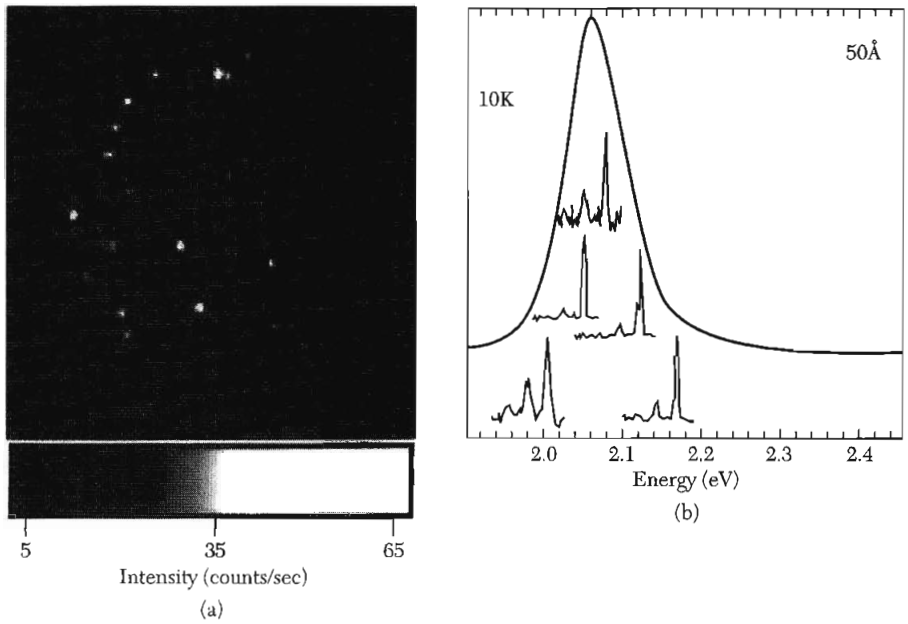
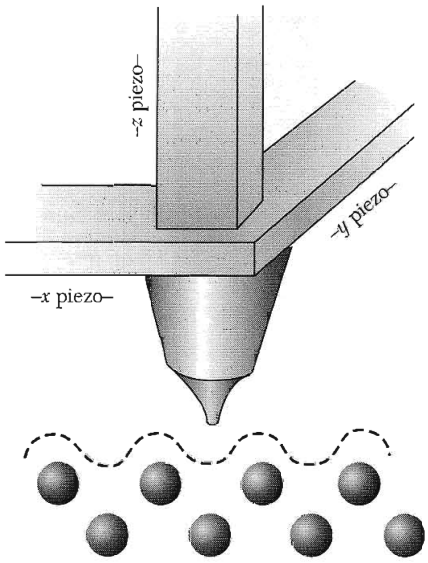


Figure 5 Left: Image of the fluorescence from individual CdSe nanocrystals dilutely distributed on a surface at $T = 10$ K. Right: Spectra of the fluorescence of a number of different individual nanocrystals. In each spectrum, the high-energy peak is the primary transition between the lowest electronic state in the conduction band and the highest energy state in the valence band. The lower energy peaks are associated transitions involving the emission of an LO phonon. Variations in the nanocrystal size and local electronic environment shift the positions of the peaks. The broad peak is the spectrum obtained for an ensemble of nominally identical nanocrystals. (After S. Empedocles et al.)

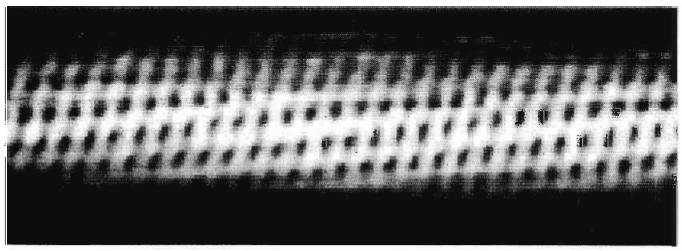
deep-UV, devices with features of 100 nm are in commercial production. Further improvements using Extreme UV light or even X-rays are possible, but the masks and focusing elements become more and more challenging to fabricate and control.

Scanning Tunneling Microscopy

The most famous scanned probe instrument is the **scanning tunneling microscope** (STM), schematically shown in Fig. 6. Its invention was a breakthrough in the field of nanoscience. In an STM, a sharp metal tip, preferably one with a single atom protruding from the end, is brought to within a nanometer of the conducting sample to be studied. The position of the tip is controlled with picometer precision using piezoelectric materials that expand or contract in response to electrical signals from a control system. A voltage bias V is applied to the sample, and a tunneling current I flowing between the tip and the sample is measured. The current is proportional to $\mathfrak{I}$, the



(a)



(b)

Figure 6 Schematic of a scanning tunneling microscope (STM). When operated in feedback mode, the piezos scan the tip over the sample and maintain a constant tunneling current between the tip and the sample. (Courtesy of D. LePage.) Lower: STM image of a carbon nanotube. (Courtesy of C. Dekker.)

tunneling probability through the gap between the tip and sample. The tunneling probability is exponentially sensitive to the tunneling distance. In the WKB approximation,

$$\mathcal{T} \propto \exp(-2\sqrt{2m\phi/\hbar^2}\, z) \ , \tag{8}$$

where z is the distance between the tip and sample and ϕ is the effective barrier height for tunneling. For typical parameters, a 0.1-nm change in the tip position leads to an order of magnitude change in $\mathcal{T}$.

When the STM operates in feedback mode, I is maintained at a constant value by changing the tip height z. The STM thus tracks the surface topography, and very small changes in the height of the surface can be detected

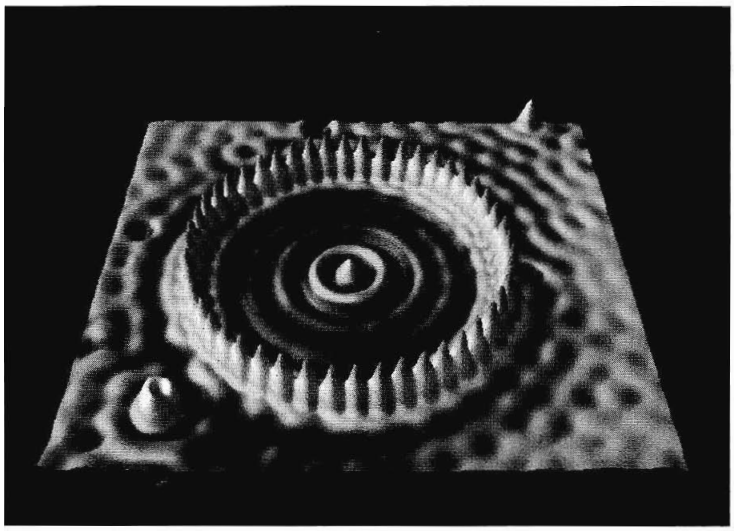


Figure 7 A "quantum corral" of mean radius 7.1 nm was formed by moving 48 Fe atoms on a Cu (111) surface. The Fe atoms scatter the surface state electrons, confining them to the interior of the corral. The rings in the corral are the density distribution of the electrons in the three quantum states of the corral that lie close to the Fermi energy. The atoms were imaged and moved into position by a low-temperature, ultra-high vacuum scanning tunneling microscope. (Image courtesy of D. M. Eigler, IBM Research Division.)

(<1 pm). This is illustrated in Fig. 6, where an STM image of a carbon nanotube is shown. The STM can also be used to manipulate individual atoms on a surface. An example is shown in Fig. 7, where the STM tip is used to construct a "quantum corral" by pushing Fe atoms on a Cu (111) surface into a ring.

The STM tunneling current I as a function of bias V can give spatial and spectroscopic information about the quantum states of a nanostructure. At zero temperature, the derivative of the current with respect to voltage is

$$dI/dV \propto \Im \sum_j \left|\psi_j(\mathbf{r}_t)\right|^2 \delta(\varepsilon_F + eV - \varepsilon_j) \ . \tag{9}$$

It is proportional to the density of states at the tunneling electron energy $\varepsilon_F + eV$, weighted by the electron probability density of those states at the STM tip position $\mathbf{r}_t$.

For the quantum corral, the electrons in the 2D surface state of Cu are reflected by the Fe atoms, creating a discrete set of states in the interior of the corral. The observed ripples in the image in Fig. 7 are due to the modulations of the probability density $|\psi_j(\mathbf{r}_t)|^2$ of these localized states near the tunneling electron energy. Images at different bias voltages yield the spatial structure of quantized states at different energies.

Atomic Force Microscopy

The **atomic force microscope** (AFM) was developed soon after the STM. It is a much more flexible technique than STM and can be used on both conducting and insulating samples. However, it typically has poorer resolution. An AFM measures the force between the tip and the sample, rather than the tunneling current. A sharp tip is mounted on the end of a millimeter-sized cantilever, as shown in Fig. 8. A force F exerted on the tip by the sample deflects the cantilever by Δz:

$$F = C\Delta z \ , \tag{10}$$

where C is the force constant of the cantilever. The displacement of the cantilever is measured as a function of tip position, often by using the back of the cantilever as a reflector for a laser beam (Fig. 8). Motion of the reflector changes the path of the laser beam, which is detected using a photodiode array; picometer-scale displacements can easily be measured. Since a typical value of the force constant is $C = 1$N/m, pN-scale forces can be transduced. Forces well below 1 fN have been measured under special circumstances.

The simplest mode of operation is contact mode, where the tip is dragged along in contact with the surface and the cantilever deflection is measured. This gives a measure of the sample topography, but it can damage the sample. Noncontact or intermittent-contact imaging modes are less invasive, and they also can give information about the long-range forces between the sample and the tip. In these techniques, the cantilever oscillates just above the sample due to an applied driving force of amplitude F_ω near the cantilever resonance

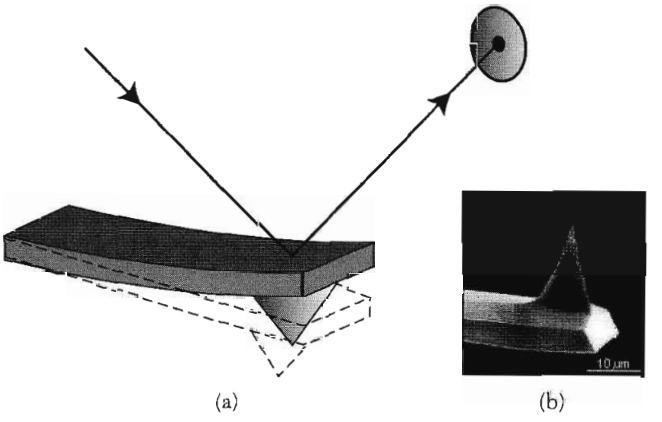


Figure 8 (a) Schematic of an atomic force microscope (AFM). Deflections of the cantilever are measured by a photodetector registering the position of a laser beam that reflects off the top of the cantilever. (Courtesy of Joost Frenken.) Inset: SEM image of an AFM tip. The effective radius of curvature of the tip can be less than 10 nm.

frequency ω_0. Modeling the cantilever as a driven simple harmonic oscillator, the magnitude of the cantilever response at a frequency ω is given by

$$|z_\omega| = \frac{F_\omega}{C} \frac{\omega_0^2}{\left[(\omega^2 - \omega_0^2)^2 + (\omega\omega_0 / Q)^2\right]^{1/2}} , \tag{11}$$

where Q, the quality factor of the oscillator, is the ratio of the energy stored in the cantilever to the energy dissipated per cycle. Note that on-resonance, $\omega = \omega_0$, the response is Q times larger than at low frequencies, making the detection of small forces possible.

The parameters characterizing the oscillating cantilever are sensitive to any forces that occur between the tip and the sample. These forces can be van der Waals, electrostatic, magnetic, or many others. The interaction shifts the resonance frequency ω_0 and/or modifies Q. This change is recorded and used to construct an image. For example, in tapping mode imaging, the tip 'taps' the surface during the closest approach of the oscillation cycle, causing both a frequency shift and additional dissipation. The nanotube device shown in Fig. 3 is imaged in tapping mode.

Another important technique is **Magnetic Force Microscopy** (MFM), briefly discussed in Chapter 12. The tip is coated with a magnetic material so that it has a magnetic moment μ normal to the surface of the sample. It then feels a force due to variations in local magnetic fields produced by the sample

$$F(z_0 + \Delta z) = F(z_0) + \partial F/\partial z\Big|_{z=z_0}\Delta z = \mu(\partial B/\partial z)\Big|_{z=z_0} + \mu(\partial^2 B/\partial z^2)\Big|_{z=z_0}\Delta z , \tag{12}$$

where z_0 is the tip's equilibrium position and Δz is the displacement during the oscillation. The term $\mu(\partial B/\partial z)$ produces a static deflection of the cantilever, but does not alter the oscillation frequency or the damping. The term $\mu(\partial^2 B/\partial z^2)\Delta z$, on the other hand, has the form of a force constant change δC, since it is linear in the displacement Δz of the cantilever. It therefore shifts the resonance frequency of the cantilever. Monitoring this frequency shift produces an image. Gradients of other local force fields can be similarly measured.

There are many other scanned probe techniques. Near-field scanning optical microscopy (NSOM) creates optical images with a resolution below the diffraction limit by using a scanned subwavelength aperture through which photons 'tunnel'. Scanning capacitance microscopy (SCM) measures capacitance variations between the tip and the sample as a function of position. This ever-growing family of techniques is increasingly used to characterize objects ranging from individual molecules to Si transistors in integrated circuits.

ELECTRONIC STRUCTURE OF 1D SYSTEMS

The quantized electronic states of nanostructures determine their electrical and optical properties, and they influence the physical and chemical properties as well. To describe these states, we take as our starting point the band structure of the bulk material. An effective mass approximation is used for the electronic dispersion of a given band, and the associated wavefunctions are treated as plane waves. These are simplifications; the bands are not always parabolic, and the true eigenstates are Bloch states, not plane waves. However, these assumptions greatly simplify the mathematics and are qualitatively (and often quantitatively) correct. We will also often neglect the Coulomb interactions between electrons. However, there are many cases in the physics of nanostructures where electron–electron interactions cannot be ignored, as discussed later in this chapter.

One-dimensional (1D) Subbands

Consider a nanoscale solid in the geometry of a wire. Its dimensions along the x and y are nanoscale, but it is continuous in z. The energies and eigenstates of such a wire are given by

$$\varepsilon = \varepsilon_{i,j} + \hbar^2 k^2 / 2m \ ; \qquad \psi(x,y,z) = \psi_{i,j}(x,y)e^{ikz} \ , \tag{13}$$

where i and j are the quantum numbers labeling the eigenstates in the x,y plane and k is the wavevector in the z direction. For the rectangular wire shown in Fig. 9, $\varepsilon_{i,j}$ and $\psi_{i,j}$ (x,y) are just particle-in-a-box energies and eigenstates discussed in Chapter 6.

The dispersion relation consists of a series of 1D subbands, each corresponding to a different transverse energy state $\varepsilon_{i,j}$. The total density of electronic states $D(\varepsilon)$ is the sum of the density of states of the individual subbands:

$$D(\varepsilon) = \sum_{i,j} D_{i,j}(\varepsilon) \ , \tag{14}$$

where $D_{i,j}(\varepsilon)$ is given by

$$\begin{aligned} D_{i,j}(\varepsilon) = \frac{dN_{i,j}}{dk}\frac{dk}{d\varepsilon} &= (2)(2)\frac{L}{2\pi}\left[\frac{m}{2\hbar^2(\varepsilon - \varepsilon_{i,j})}\right]^{1/2} = \frac{4L}{h\nu_{i,j}} \quad \text{for } \varepsilon > \varepsilon_{i,j} \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \text{for } \varepsilon < \varepsilon_{i,j} \ . \end{aligned} \tag{15}$$

The first factor of two in the middle expression is due to spin degeneracy and the second from including both positive and negative values of k. In the right expression, $\nu_{i,j}$ is the velocity of the electron in the i,j subband with kinetic energy $\varepsilon - \varepsilon_{i,j}$. Note that the density of states diverges as $(\varepsilon - \varepsilon_{i,j})^{-1/2}$ at each subband threshold. These are called **van Hove singularities**. This behavior stands in contrast to three dimensions, where $D(\varepsilon)$ goes to zero at low energies

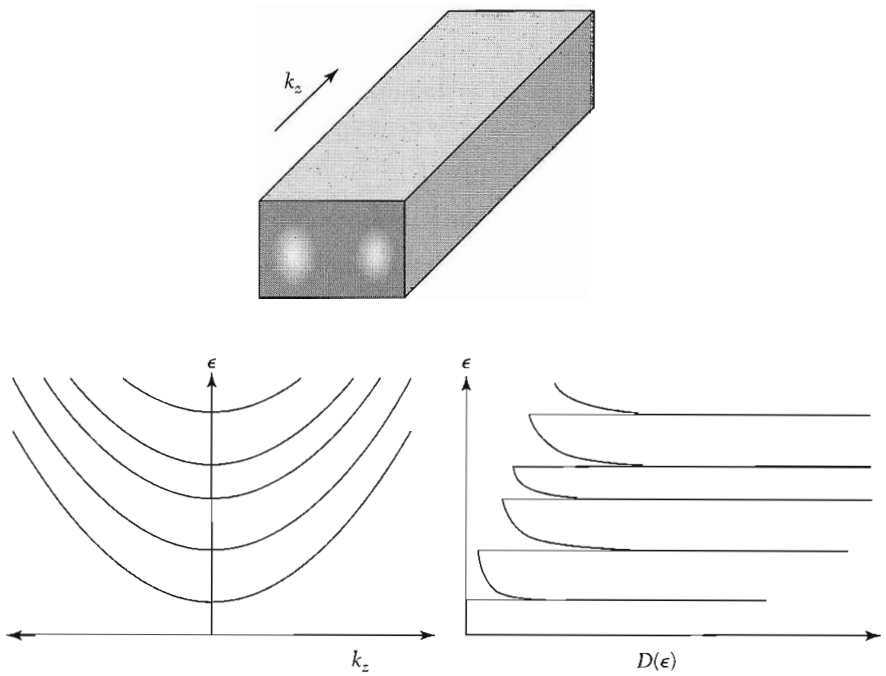


Figure 9 Schematic of a rectangular quasi-one-dimensional wire, along with the dispersion relations and the density of states of the 1D subbands. The peaks in the density of states at the subband thresholds are called Van Hove singularities. The probability density for the $i = 2, j = 1$ state is shown as a gray scale on the cross section of the wire.

(Chapter 6), and two dimensions, where $D(\varepsilon)$ steps up a constant value at the bottom of each 2D subband (Prob. 17.3)

Spectroscopy of Van Hove Singularities

The Van Hove singularities described by (15) affect the electrical and optical properties of 1D systems. Here, we discuss the case of a semiconducting carbon nanotube, whose band structure is calculated in Prob. 1 and shown in Fig. 10a. Van Hove singularities are seen in scanning tunneling spectroscopy, as shown in Fig. 10b. Peaks in the differential conductance, which is proportional to the density of states by (9), are observed at bias voltages corresponding to the energies of these singularities.

The optical absorption and emission of semiconducting nanotubes are also dominated by these singularities, since they depend on the initial and final density of states by (5)–(7). Figure 10c shows the photoluminescence intensity of a collection of carbon nanotubes as a function of the wavelength of the exciting and emitted light. The absorption of the incident light is enhanced when the energy matches $\varepsilon_{c2} - \varepsilon_{v2}$, the energy difference between the 2nd Van Hove singularities in the conductance and valence bands. The electrons and holes relax quickly to the bottom of the first subband, where they recombine,

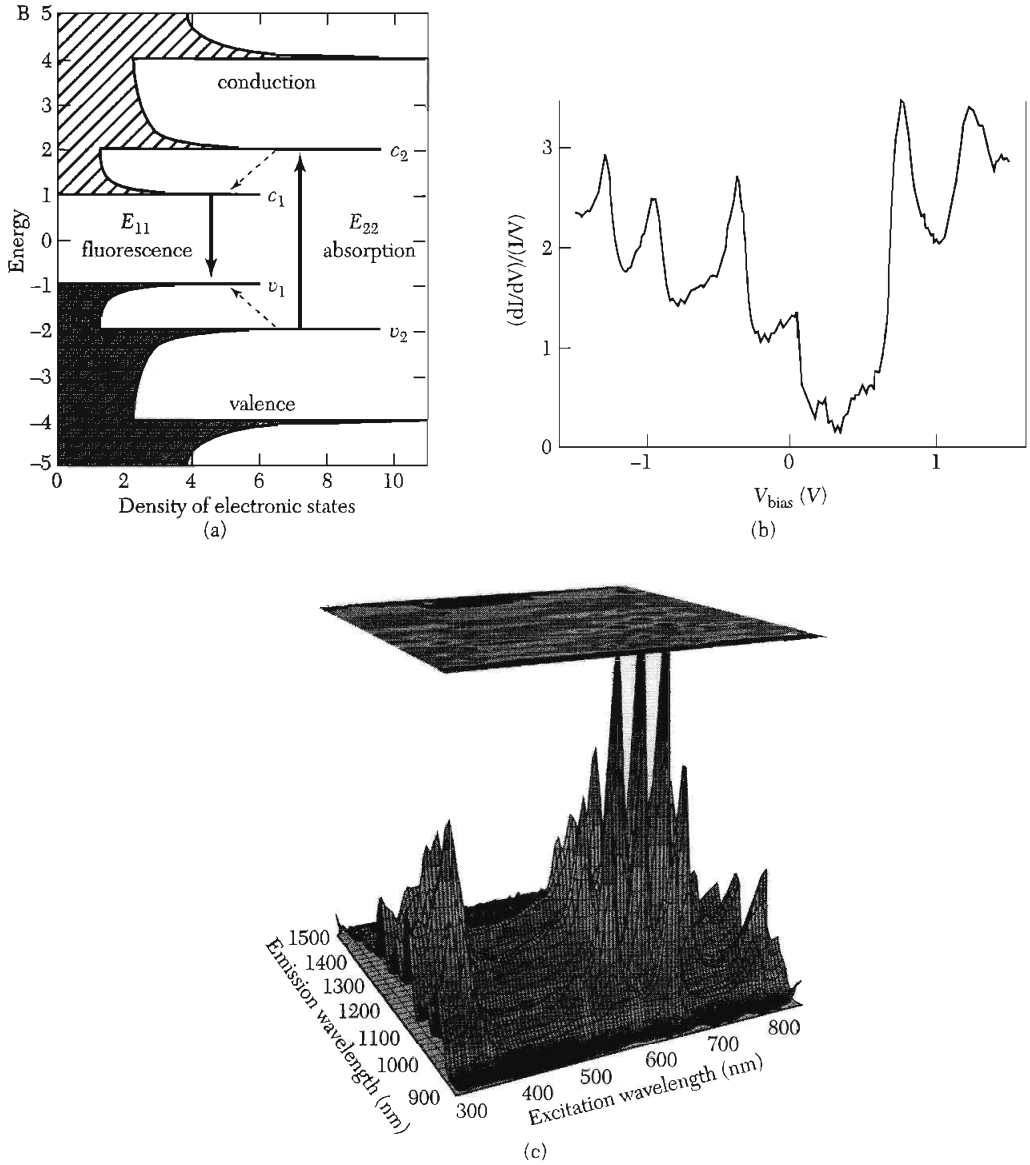


Figure 10 (a) The density of states for a semiconducting carbon nanotube as a function of energy. The Van Hove singularities are seen in the STM tunneling spectra of a nanotube shown in (b). In (c), the emission intensity is plotted as a function of the emission wavelength and the excitation wavelength. Peaks in the intensity are observed when the absorption and emission energies correspond to those shown in the diagram (a). Different peaks correspond to nanotubes with different radii and chirality. [After Bachilo et al. (a and c) and C. Dekker (b).]

producing luminescence with energy $\varepsilon_{c1} - \varepsilon_{v1}$. Peaks are therefore observed when the emitted and absorbed light simultaneously match the energies between the 1st and 2nd van Hove singularities, respectively. Different peaks in the emission intensity in the plot correspond to nanotubes of varying diameters and chiralities.

1D Metals—Coulomb Interactions and Lattice Couplings

In a quasi-one-dimensional metal, electrons fill up individual 1D subbands, with the Fermi energy and the total number of subbands occupied determined by the electron density. For a strictly 1D metal, there is only one (spin-degenerate) subband occupied. In this case, if there are n_{1D} carriers per unit length:

$$n_{1D} = 2k_F/\pi. \tag{16}$$

The Fermi surface of a 1D metal consists of just two points, at $+k_F$ and $-k_F$, as shown in Fig. 11. This is quite different from the Fermi surfaces in 3D and 2D free-electron metals, which consist of a sphere and a circle, respectively. Two consequences of this unusual Fermi surface are discussed below.

Coulomb interactions cause scattering among electrons near the Fermi energy. For 3D metals, scattering is strongly suppressed near ε_F by the restrictions of energy/momentum conservation combined with the Pauli exclusion principle. At an energy ε measured relative to ε_F, $1/\tau_{ee} \approx (1/\tau_0)\,(\varepsilon/\varepsilon_F)^2$, where $1/\tau_0$ is the

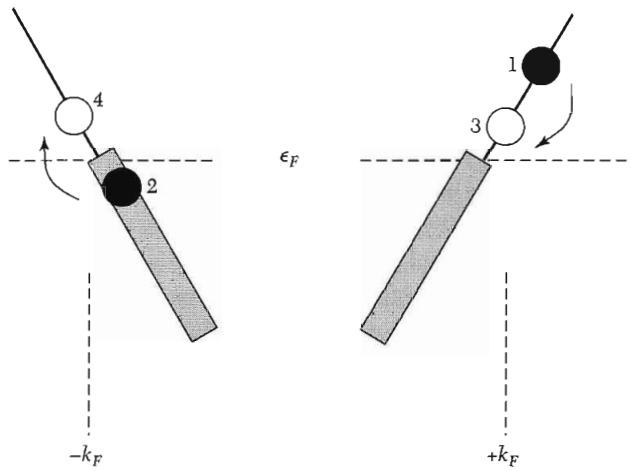


Figure 11 Electronic structure of a 1D metal near the Fermi energy. The Fermi surface consists of two points at $\pm\, k_F$. The scattering of electrons from filled states 1 and 2 to empty states 3 and 4 conserves energy as long as the energy difference is the same between 1 and 3 and between 2 and 4. Momentum is simultaneously conserved because the energy is locally linear in k.

classical scattering rate. By the uncertainty principle, this produces an uncertainty in the energy of the electron:

$$\delta\varepsilon(3D) \approx \hbar/\tau_{ee} \sim (\hbar/\tau_0)(\varepsilon/\varepsilon_F)^2 \ . \tag{17}$$

As the energy becomes small (measured relative to ε_F), the uncertainty in the energy goes to zero as the second power of ε. The uncertainty $\delta\varepsilon$ is therefore guaranteed to be small in comparison to ε sufficiently close to ε_F. This ensures that the quasiparticles near the Fermi surface are well defined.

The case of one dimension is shown in Fig. 11. Energy and momentum conservation are equivalent in this case since for small ε the energy is locally linear in the momentum change $\Delta k = k - k_F$:

$$\varepsilon \cong (\hbar^2 k_F/m)\Delta k \ . \tag{18}$$

Referring to Fig. 11, energy conservation requires that for an electron in state 1 at energy ε to scatter to state 3, a simultaneous scattering of an electron from state 2 to 4 must occur. The only restriction is that the final state energy 3 be positive and less than ε. This gives a reduction factor $1/\tau_{ee} \sim (1/\tau_0)\,(\varepsilon/\varepsilon_F)$, and, from the uncertainty principle,

$$\delta\varepsilon(1D) \simeq \hbar/\tau_{ee} \sim (\hbar/\tau_0)(\varepsilon/\varepsilon_F) \ . \tag{19}$$

Since the uncertainty is linear in ε, there is no guarantee that $\delta\varepsilon$ will be smaller than ε as $\varepsilon \to 0$. The fundamental assumption behind Fermi liquid theory, that weakly interacting quasiparticles exist as $\varepsilon \to 0$, is therefore not guaranteed in 1D. In fact, the ground state of the interacting 1D electron gas is believed *not* to be a Fermi liquid, but rather a Luttinger liquid whose low-energy excitations are collective in nature. The excitations are more analogous to phonons or plasmons—a collective motion of many objects—than isolated electrons moving independently of their neighbors. This collective nature has a number of effects. For example, tunneling into a 1D metal is suppressed at low energies because the tunneling electron must excite the collective modes. In spite of this issue, the independent electron picture remains a useful approximation for the 1D electron gas. It has been successful in describing most, but not all, experiments on real 1D systems, and will be adopted below.

A second unusual property of 1D metals is that they are unstable to perturbations at a wavevector $2k_F$. For example, a distortion of the lattice at this wavevector will open up a bandgap in the electronic spectrum, converting the metal to an insulator. This is the Peierls instability treated in detail in Chapter 14. This effect is particularly important in 1D conducting polymers such as polyacetylene (Fig. 12). It has one conduction electron per carbon atom spacing a, and therefore from (16) $k_F = \pi/2a$. Without any distortions, polyacetylene would have a half-filled band and be a metal. However, a lattice distortion

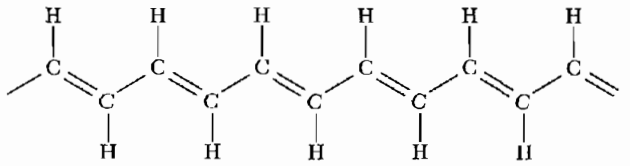


Figure 12 Structure of polyacetylene. Due to the Peierls distortion, the lattice is dimerized, with carbon atoms joined by double bonds in the diagram closer together than those linked by single bonds. The Peierls distortion opens a semiconducting gap of approximately 1.5 eV.

at $2k_F = \pi/a$, corresponding to a wavelength $2a$, opens up a gap at the Fermi energy. This corresponds to a dimerization of the lattice. This dimerization produces alternating single and double bonds along the chain and turns polyacetylene into a semiconductor with a bandgap of 1.5 eV.

Polyacetylene and related semiconducting polymers can be made into field-effect transistors, light-emitting diodes, and other semiconducting devices. They can also be doped chemically, producing metallic behavior with conductivities comparable to traditional metals. However, they retain the mechanical flexibility and ease of processing characteristic of polymers. Their discovery has led to a revolution in flexible plastic electronics.

The Peierls distortion is large in polymers because their backbone consists of a single atomic chain, which can easily distort. Other 1D systems such as nanotubes and nanowires are much stiffer, and the Peierls transition is not observed at experimentally relevant temperatures.

ELECTRICAL TRANSPORT IN 1D

Conductance Quantization and the Landauer Formula

A 1D channel has a finite current-carrying capacity for a given voltage applied across its ends. It therefore has a finite conductance even if there is no scattering in the wire. Consider a wire with one subband occupied connecting two larger reservoirs with a voltage difference V between them, as shown in Fig. 13. The right-going states will be populated up to an electrochemical potential μ_1 and left-going states will be populated up to μ_2, where $\mu_1 - \mu_2 = qV$ and $q = -e$ for electrons and $+e$ for holes. The net current flowing through the channel due to the excess right-moving carrier density Δn is then

$$I = \Delta n q v = \frac{D_R(\varepsilon) q V}{L} q v = \frac{2}{hv} v q^2 V = \frac{2e^2}{h} V \ , \tag{20}$$

where $D_R(\varepsilon)$, the density of states of right-moving carriers, is $\frac{1}{2}$ the total density of states given in (15).

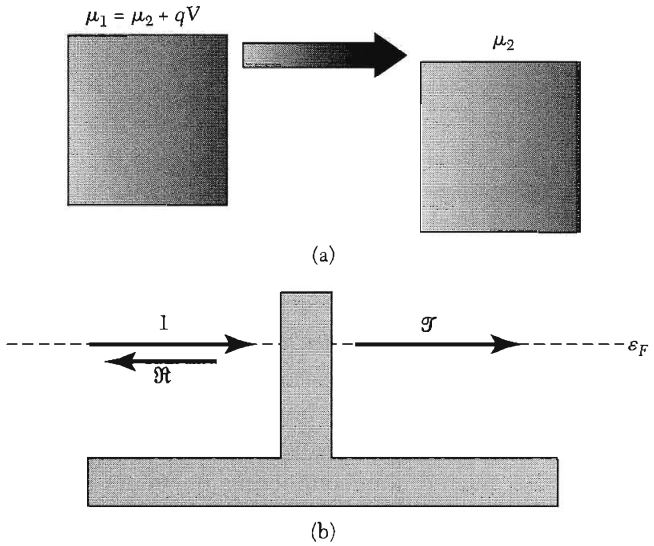


Figure 13 (a) The net current propagating between two reservoirs for an applied bias voltage difference $V_1 - V_2$. (b) Schematic representation of transmission probability $\mathcal{T}$ and reflection probability $\mathcal{R}$ from a barrier in the channel, where $\mathcal{T} + \mathcal{R} = 1$.

Remarkably, in 1D the velocity exactly cancels with the density of states to create a current that depends only on the voltages and fundamental constants. The two-terminal conductance I/V and resistance V/I are then

$$G_Q = 2e^2/h \ ; \qquad R_Q = h/2e^2 = 12.906\ \text{k}\Omega \ . \tag{21}$$

A perfectly transmitting one-dimensional channel has a finite conductance whose value is the ratio of fundamental constants. This is called the **conductance quantum** G_Q; its inverse is called the **resistance quantum** R_Q. While derived here in the effective mass approximation, it is true for a 1D band of arbitrary dispersion.

The quantization of conductance is dramatically illustrated in the data in Fig. 14. A short quasi-1D channel is formed between two regions of a 2D electron gas in a GaAs/AlGaAs heterostructure. As the carrier density of the channel is increased, the conductance increases in discrete steps of height $2e^2/h$. Each step corresponds to the occupancy of an additional 1D subband in the wire. Conductance quantization is also observed in atomic-scale bridges between macroscopic metals.

If the channel is not perfectly conducting, the overall conductance is the quantum of conductance times the probability $\mathcal{T}(\varepsilon_F)$ for electron transmission through the channel (Fig. 13):

$$G(\varepsilon_F) = (2e^2/h)\mathcal{T}(\varepsilon_F) \ . \tag{22}$$

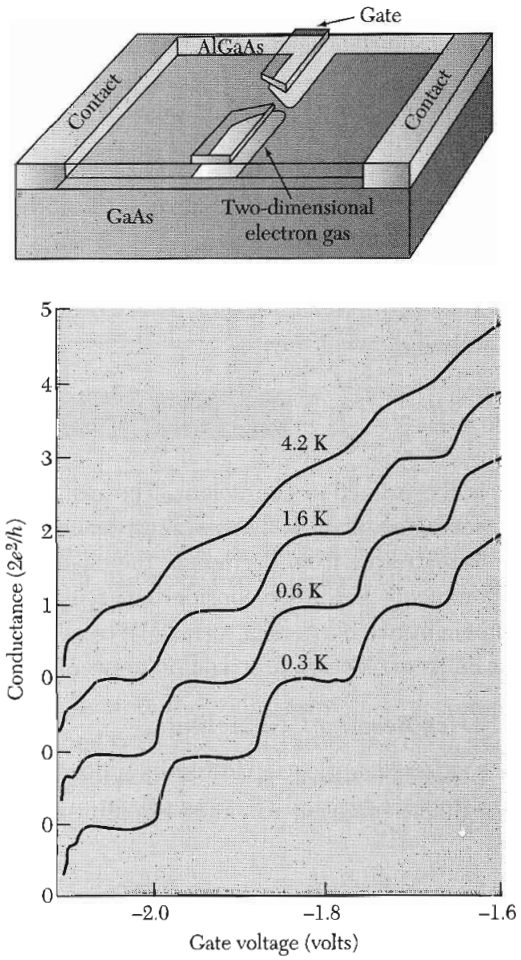


Figure 14 Conductance quantization in a short channel electrostatically defined in a GaAs/AlGaAs heterostructure at different temperatures. A negative gate voltage V_g applied to the metallic gates on the surface of the sample depletes the carriers in the underlying two-dimensional electron gas, creating a narrow channel. The channel is fully depleted of carriers at $V_g = -2.1$V. Individual 1D subbands become occupied with increasing V_g, with each new subband adding a conductance of $2e^2/h$. (Courtesy of H. Van Houten and C. Beenakker.)

This equation is often called the **Landauer Formula**. For a quasi-1D system with multiple channels, we sum over the contributions of each channel, since conductances in parallel add:

$$\mathcal{T}(\varepsilon_F) = \sum_{i,j} \mathcal{T}_{i,j}(\varepsilon_F) \ , \tag{23}$$

where i, j label the transverse eigenstates. For example, for N perfectly transmitted channels in parallel, $\mathcal{T} = N$, as for the data in Fig. 14.

At finite temperatures or biases, the Fermi-Dirac energy distributions f of the electrons in the left and right leads must be taken into account:

$$I(\varepsilon_F,V,T) = (2e/h)\int_{-\infty}^{\infty} d\varepsilon[f_L(\varepsilon - eV) - f_R(\varepsilon)]\mathfrak{I}(\varepsilon) \ . \tag{24}$$

The net current is simply the difference between the left- and right-moving currents, integrated over all energies.

The Landauer formula (22) directly relates the resistance of a system to the transmission properties of the channel. Let us rewrite the resistance for the one-channel case in the following way:

$$R = \frac{h}{2e^2}\frac{1}{\mathfrak{I}} = \frac{h}{2e^2}\frac{\mathfrak{I}+(1-\mathfrak{I})}{\mathfrak{I}} = \frac{h}{2e^2} + \frac{h}{2e^2}\frac{\mathfrak{R}}{\mathfrak{I}} \ , \tag{25}$$

where $\mathfrak{R} = 1 - \mathfrak{I}$ is the reflection coefficient. The resistance of the device is the sum of the first term, the quantized contact resistance, and the second term, the resistance due to scattering from barriers in the channel. The latter term is zero for a perfect conductor. Below we consider an application of the Landauer formula to the problem of two barriers in series. We treat this in both the coherent and incoherent limits of electron propagation between the barriers.

Two Barriers in Series-Resonant Tunneling

Consider two barriers in series separated by a distance L, with transmission/reflection amplitudes t_1, r_1 and t_2, r_2, as shown in Fig. 15. These amplitudes are complex:

$$t_j = |t_j|e^{i\varphi_{tj}} \ ; \qquad r_j = |r_j|e^{i\varphi_{rj}} \ . \tag{26}$$

To calculate the transmission probability $\mathfrak{I}$ through the entire double barrier structure, we need the corresponding transmission amplitude. For an incident wave from the left whose amplitude is 1, the amplitudes defined in Fig. 15 are given by

$$a = t_1 + r_1 b \ ; \quad b = ar_2 e^{i\varphi} \ ; \quad c = at_2 e^{i\varphi/2} \ , \tag{27}$$

where $\varphi = 2kL$ is the phase that an electron with kinetic energy $h^2k^2/2m$ accumulates propagating the distance $2L$ on a round trip between the barriers. Combining these to solve for the transmitted amplitude yields:

$$c = \frac{t_1 t_2 e^{i\varphi/2}}{1 - r_1 r_2 e^{i\varphi}} \ . \tag{28}$$

The transmission probability through the double barrier is then

$$\mathfrak{I} = |c|^2 = \frac{|t_1|^2|t_2|^2}{1 + |r_1|^2|r_2|^2 - 2|r_1||r_2|\cos(\varphi^*)} \tag{29}$$

where $\varphi^* = 2kL + \varphi_{r1} + \varphi_{r2}$.

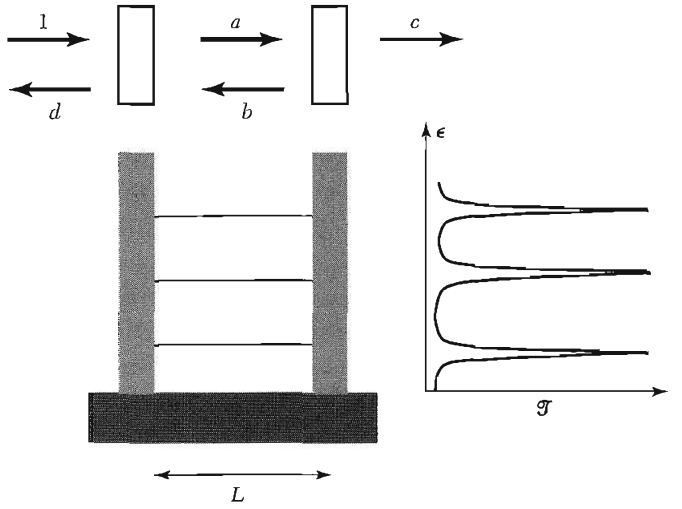


Figure 15 Resonant tunneling through two identical barriers in series separated by a length L. The upper diagram shows the transmission amplitudes between and outside the barriers for a unity amplitude incident wave. The transmission resonances at the energies of the quasibound states between the barriers are shown.

This is plotted in Fig. 15. Note the round-trip phase accumulation φ^* includes the phase shifts associated with reflections from the barriers.

The transmission probability (29) is greatly enhanced when $\cos(\varphi^*)$ approaches unity, because the denominator becomes small. This occurs for the resonance condition

$$\varphi^* = 2kL + \varphi_{r1} + \varphi_{r2} = 2\pi n \ , \tag{30}$$

where n is an integer. This is a general property of waves, and is due to the constructive interference of many pathways through the sample. This can be easily seen by rewriting (28) using the series expansion $1/(1-x) = \sum_{m=0}^{\infty} x^m$:

$$c = t_1 t_2 e^{i\varphi/2}/(1 - r_1 r_2 e^{i\varphi}) = t_1 t_2 e^{i\varphi/2}\left[1 + r_1 r_2 e^{i\varphi} + (r_1 r_2 e^{i\varphi})^2 + \ldots\ldots\right] . \tag{31}$$

The mth order in the expansion corresponds to a path with m round trips between the barriers. On resonance, these paths add in phase to yield a strongly enhanced transmission.

Consider the special case where the barriers are the same: $t_1 = t_2$. We then have

$$\mathcal{T}(\varphi^* = 2\pi n) = |t_1|^4 (1 - |r_1|^2)^{-2} = 1 \ . \tag{32}$$

The transmission on resonance through a symmetric double-barrier structure is 1, even if the transmission through each of the individual barriers is small.

This is called **resonant tunneling**. Off resonance, the denominator of (29) is of order unity for opaque barriers, and the transmission is roughly the product of the transmission coefficients of each of the two barriers in series: $\mathcal{T} \sim |t_1|^2|t_2|^2$.

The resonance condition $\varphi^* = 2\pi n$ corresponds to the energies of the quasibound electronic states confined between the two barriers. For very opaque walls, this is just the particle-in-a-box quantization condition: $kL = \pi n$. We derived the resonant tunneling condition for a one-dimensional case, but it is a general result. The transmission through a confined electron system is strongly enhanced at energies corresponding to the bound-state energy levels of the confined electrons. This is also evident from the STM tunneling expression (9); quasibound states produce peaks in the differential conductance.

For the case of opaque barriers, $|t_1|^2, |t_2|^2 \ll 1$, the cosine term in the denominator of (29) can be expanded, as shown in Prob. 3, yielding the familiar Breit-Wigner form for a resonance:

$$\mathcal{T}(\varepsilon) = \frac{4\Gamma_1\Gamma_2}{(\Gamma_1+\Gamma_2)^2 + 4(\varepsilon-\varepsilon_n)^2} \qquad \text{where} \qquad \Gamma_j = \frac{\Delta\epsilon}{2\pi}|t_j|^2 \ . \tag{33}$$

The resonances are thus Lorentzian peaks with a width in energy of $\Gamma = \Gamma_1 + \Gamma_2$ determined by the energy level spacing $\Delta\varepsilon$ and the transmission probabilities through the two barriers. This is just the uncertainty principle broadening of the level due to the finite lifetime of the double-barrier bound state.

Incoherent Addition and Ohm's Law

If we instead treat the electron classically, we add probabilities rather than amplitudes. This is valid if the electron effectively loses track of its phase between the barriers due to, for example, inelastic scattering from phonons. This corresponds to replacing (27) by

$$|a|^2 = |t_1|^2 + |r_1|^2|b|^2 \ ; \qquad |b|^2 = |a|^2|r_2|^2 \ ; \qquad |c|^2 = |a|^2|t_2|^2 \ . \tag{34}$$

This gives

$$\mathcal{T} = \frac{|t_1|^2|t_2|^2}{1-|r_1|^2|r_2|^2} \ . \tag{35}$$

Some elementary manipulations (Prob. 4) yield

$$R = (h/2e^2)(1+|r_1|^2/|t_1|^2+|r_2|^2/|t_2|^2) \ . \tag{36}$$

The resistance is just the sum of the quantized contact resistance and the intrinsic resistances of the individual barriers (see Eq. 25). This is Ohm's law—resistors in series add. It is valid if interference effects can be neglected.

Equation (36) allows us to connect to the Drude formula. Consider a process that gives a backscattering rate $1/\tau_b$. This backscattering could result from either an elastic scattering process such as impurity scattering or from an inelastic

process such as phonon scattering. For propagation over a small distance dL, the reflection probability $d\mathfrak{R}$ ($\ll 1$) is

$$d\mathfrak{R} = \frac{1}{\tau_b}\frac{dL}{v_F} = \frac{dL}{\ell_b} \ . \tag{37}$$

This gives a contribution to the resistance, yielding a resistivity:

$$\rho_{1D} = dR/dL = (h/2e^2)/\ell_b \ . \tag{38}$$

This is equal to the 1D Drude resistance $\sigma_{1D}^{-1} = (n_{1D}e^2\tau/m)^{-1}$, as shown in Prob. 4. Ignoring interference effects, the resistances of individual segments add ohmically, giving

$$R = R_Q + (h/2e^2)(L/\ell_b) \ . \tag{39}$$

Localization

Now consider when two barriers are connected in series, but coherence is *not* neglected. However, we average over all possible phases, corresponding to an average over different energies. From Eq. (29), the average resistance is

$$\langle R\rangle = \frac{h}{2e^2}\frac{1+|r_1|^2|r_2|^2-2|r_1||r_2|\langle\cos\varphi^*\rangle}{|t_1|^2|t_2|^2} = \frac{h}{2e^2}\frac{1+|r_1|^2|r_2|^2}{|t_1|^2|t_2|^2} \ . \tag{40}$$

Notably, the phase-averaged resistance (40) is *larger* than the resistance in the incoherent limit (36).

To understand the scaling with length associated with (40), consider a long conductor of length L consisting of a series of only elastic (phase-preserving) scatterers characterized by an elastic backscattering length ℓ_e. Assume that the conductor has a large resistance $\langle R\rangle$, so that $\mathfrak{R} \simeq 1$ and $\mathfrak{T} \ll 1$. For a small additional length dL, there will be an additional reflection and transmission $d\mathfrak{R} = dL/\ell_e$, as in (37), and $d\mathfrak{T} = 1 - d\mathfrak{R}$. Combining these according to the prescription of (40), and assuming that $d\mathfrak{R} \ll 1$, gives

$$\langle R + dR\rangle = \frac{h}{2e^2}\frac{1+\mathfrak{R}d\mathfrak{R}}{\mathfrak{T}(1-d\mathfrak{R})} \simeq \langle R\rangle\left(1+\frac{2dL}{\ell_e}\right) \ , \tag{41}$$

or equivalently,

$$\langle dR\rangle = \langle R\rangle(2dL/\ell_e) \ . \tag{42}$$

Separating variables and integrating both sides of the equation yields

$$\langle R\rangle = (h/2e^2)\exp(2L/\ell_e) \ . \tag{43}$$

Remarkably, the resistance grows exponentially with the length of the sample, rather than linearly as in an ohmic conductor. This behavior is a result of localization. Due to quantum interference among the states scattered by disorder, the states become localized on a size scale $\xi \sim \ell_e$, where ξ is called the localization

length. There are no extended states that traverse the entire length of the conductor, so the resistance is exponentially large. A similar result holds for quasi-1D systems, but with a localization length $\xi \sim N\ell_e$, where N is the number of 1D subbands occupied.

At very low temperatures, only coherent scattering processes occur and the resistance is exponentially large by (43). At finite temperatures, electrons retain their phase memory only over the phase coherence length ℓ_φ due to their interaction with other degrees of freedom such as phonons or electrons. This length typically is a power-law function of temperature, $\ell_\varphi = AT^{-\alpha}$, since the number of electronic and vibrational excitations present is a power law in T. The resistance of each phase coherent segment can be approximated by (43) with ℓ_φ replacing L. The resistance of each phase coherent segment decreases rapidly with increasing temperature (as the exponential of a power law in T). This dramatically decreases the overall resistance, which is the (incoherent) series combination of L/ℓ_φ such phase coherent sections. At a sufficiently high temperature where $\ell_\varphi \leq \ell_e$, all phase coherence is lost between scattering events and the ohmic expression (36) is applicable.

A related issue is the nature of the electronic states in 2D and 3D systems in the presence of disorder. In 2D, it is believed that, for noninteracting electrons, all states are also localized by disorder. In 3D, on the other hand, a critical amount of disorder is required to localize the states. The subject of localization continues to be of great fundamental interest and controversy, particularly when the effects of coulomb interactions between the electrons are included.

Voltage Probes and the Büttiker-Landauer Formalism

In many measurements, more than two probes are connected to a conductor. Some are used as voltage probes (which draw no net current from the sample) and others as current probes, as shown in Fig. 16. Büttiker extended the Landauer formalism to deal with this multiprobe case. Define $\mathfrak{T}^{(n,m)}$ as the total transmission probability for an electron leaving contact m to arrive at contact n, including the contributions from all the 1D channels. For a current probe n with N_n channels, the electrochemical potential of the contact is fixed by an applied voltage, and the net current that flows through the contact is

$$I_n = (2e^2/h)(N_n V_n - \sum_m \mathfrak{T}^{(n,m)} V_m) \quad . \tag{44}$$

This is just the current flowing out of the contact minus the currents flowing in that originated from each of the other contacts. Note that $N_n = \sum_m \mathfrak{T}^{(n,m)}$, which can be easily obtained from (44) by considering the equilibrium case where all the voltages are equal and all the currents are zero.

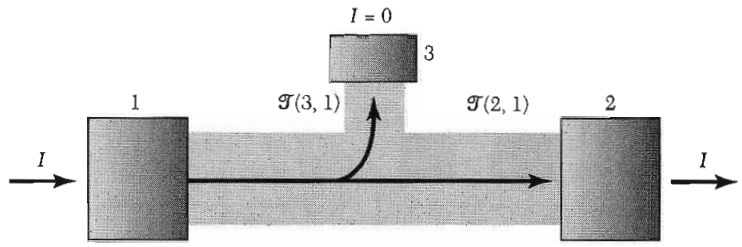


Figure 16 Schematic representation of a multiterminal conductor. Contacts 1 and 2 are current probes; contact 3 is a voltage probe. The transmission probability from contact 1 to 2 and from 1 to 3 is schematically indicated.

For a voltage probe, the potential V_n adjusts itself so that no net current flows ($I_n = 0$):

$$V_n = \frac{\sum_{m \neq n} \mathcal{T}^{(n,m)} V_m}{\sum_{m \neq n} \mathcal{T}^{(n,m)}} \ . \tag{45}$$

The electrochemical potential measured by the probe is the weighted average of the electrochemical potentials of the different contacts, where the weighting coefficients are the transmission probabilities.

Equations (44–45) have a number of surprising consequences. Since the measured currents and voltages depend on $\mathcal{T}^{(n,m)}$, the details of the path that an electron takes in traversing the sample influences the resistance. A voltage probe can disturb the paths, and the measured voltage can in turn be affected by transmission through all parts of the sample. Below we present three examples that illustrate these properties.

Consider a voltage probe connected to the center of an otherwise ballistic 1D conductor, as shown in Fig. 16. Assume that electrons leaving from probe 1 either arrive at probe 2 or 3, but none are directly backscattered. The voltage read by probe 3 is then

$$V_3 = \frac{\mathcal{T}^{(3,1)} V}{\mathcal{T}^{(3,1)} + \mathcal{T}^{(3,2)}} = \frac{V}{2} \ , \tag{46}$$

where for the last step we assumed that the voltage probe couples symmetrically to the left and right moving channels, $\mathcal{T}^{(3,1)} = \mathcal{T}^{(3,2)}$. The voltage measured in the channel is just the average of the voltage of the two contacts.

The current flowing out of contact 1 is given by:

$$I = (2e^2/h)(V - \mathcal{T}^{(1,3)} V_3) = (2e^2/h) V (1 - \tfrac{1}{2}\mathcal{T}^{(1,3)}) \ , \tag{47}$$

where (46) has been employed in the second step. Note that the presence of the voltage probe decreases the transmission below the unity value of a perfect

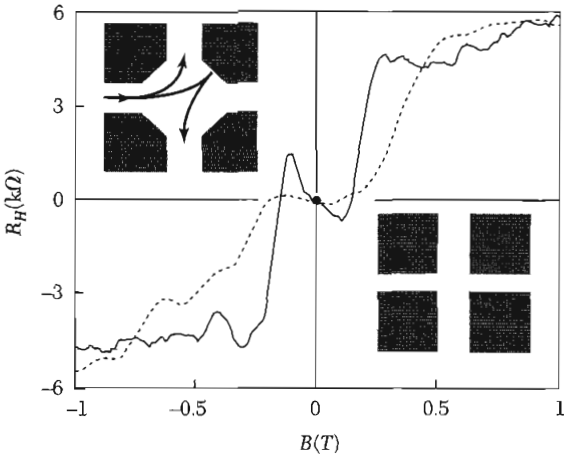


Figure 17 Four-terminal Hall resistance measurements of submicron junctions of different shapes. In the junction shown schematically in the upper left, the Hall resistance is negative at small B and positive at large B. The reason is indicated in the diagram; at small B, the electrons bounce off the wall into the "wrong" probe. (After C. Ford et al.)

channel. Some of the electrons scatter into the voltage probe, are re-emitted, and then return to contact 1. This shows that voltage probes are in general invasive; they influence what they measure unless they only couple very weakly to the system.

Figure 17 shows a measurement of the Hall resistance of two nanoscale crosses patterned in a high mobility 2D electron gas whose geometries are shown in the insets. The junction region is ballistic, meaning that there is no scattering from disorder, only from the sample walls. The measured Hall resistance is not of the form $B/n_s e$, where n_s is the sheet carrier concentration, as expected for a macroscopic 2D electron gas, but has instead a number of notable features. Most surprisingly, the Hall voltage is of the opposite sign at low B compared to high B for the sample shown in the upper left inset. This can be easily understood from the shape of the classical electron paths sketched on the figure. At high B, the Lorentz force preferentially deflects the electron into the upper electrode, giving the expected sign of the Hall voltage. At low B, however, the electron bounces off the boundary of the conductor and arrives at the lower electrode, reversing the sign of the measured Hall voltage. For a small multiprobe conductor, the resistance is a measure of the electron trajectories through the sample and not simply related to intrinsic material properties like the electron density.

Equations (44–45) can be used to treat arbitrarily complex microscopic (or even macroscopic) conductors. It has been widely used to describe measurements on small disordered metal samples at low temperatures as a function of magnetic field B. These samples have many transverse channels and contain impurities. The elastic scattering length ℓ_e is less than the sample dimensions,

but the phase coherence length ℓ_φ is greater. Electrons therefore propagate diffusively, but phase-coherently, through the sample. This is called the **mesoscopic** regime. In a semiclassical picture, the transmission amplitude between two probes n and m corresponds to the sum of many different classical paths through the sample:

(CGS) $$t^{(m,n)} \propto \sum_j a_j \exp\left((i/\hbar) \int_\ell^m (\mathbf{p} - e\mathbf{A}/c) \cdot \mathbf{dl} \right) ;$$

(SI) $$t^{(m,n)} \propto \sum_j a_j \exp\left((i/\hbar) \int_\ell^m (\mathbf{p} - e\mathbf{A}) \cdot \mathbf{dl} \right) . \tag{48}$$

Note that the phase associated with each path amplitude a_j contains a contribution from the magnetic vector potential $\mathbf{A}$, as described in Appendix G. Since $\mathcal{T}^{(n,m)} = |t^{(n,m)}|^2$, quantum interference among different conducting pathways through the sample modulates the transmission.

An interesting example is shown in Fig. 18. On the left, the four-terminal resistance of a nanoscale metallic wire is shown. Aperiodic fluctuations are seen in the conductance versus magnetic field B. These fluctuations are due to modulations of the interference between the many diffusive paths linking the contacts. Since there are many paths, the result is an essentially random variation. These modulations are referred to as conductance fluctuations.

When an additional loop is added that is outside the region between the contacts, as shown on the right of Fig. 18, the conductivity G qualitatively changes. A periodic modulation with magnetic field is seen. This is due to the **Aharonov-Bohm effect**. The vector potential modulates the quantum interference between those electron paths that encircle the ring and those that do not. For simplicity, consider the interference between just two such paths with transmission amplitudes a_1 and a_2 (Fig. 18) in the absence of a magnetic field. At finite B,

(CGS) $$|a_1 + a_2 \exp([(ie/hc) \oint_{loop} \mathbf{A} \cdot \mathbf{dl})]|^2 = |a_1|^2 + |a_2|^2 + 2|a_1||a_2| \cos[2\pi\Phi/(hc/e)] ;$$

(SI) $$|a_1 + a_2 \exp[(ie/\hbar) \oint_{loop} \mathbf{A} \cdot \mathbf{dl}]|^2 = |a_1|^2 + |a_2|^2 + 2|a_1||a_2| \cos[2\pi\Phi/(h/e)] . \tag{49}$$

In the last step we have employed Stokes's theorem, where Φ is the magnetic flux going through the loop and hc/e is the magnetic flux quantum (h/e in SI). With increasing flux, the transmission though the wire oscillates with the period of one flux quantum. This effect is closely related to the superconducting flux quantization discussed in Chapter 10, except that the charge appearing in the flux quantum here is e, not $2e$, since the carriers here are electrons, not Cooper pairs.

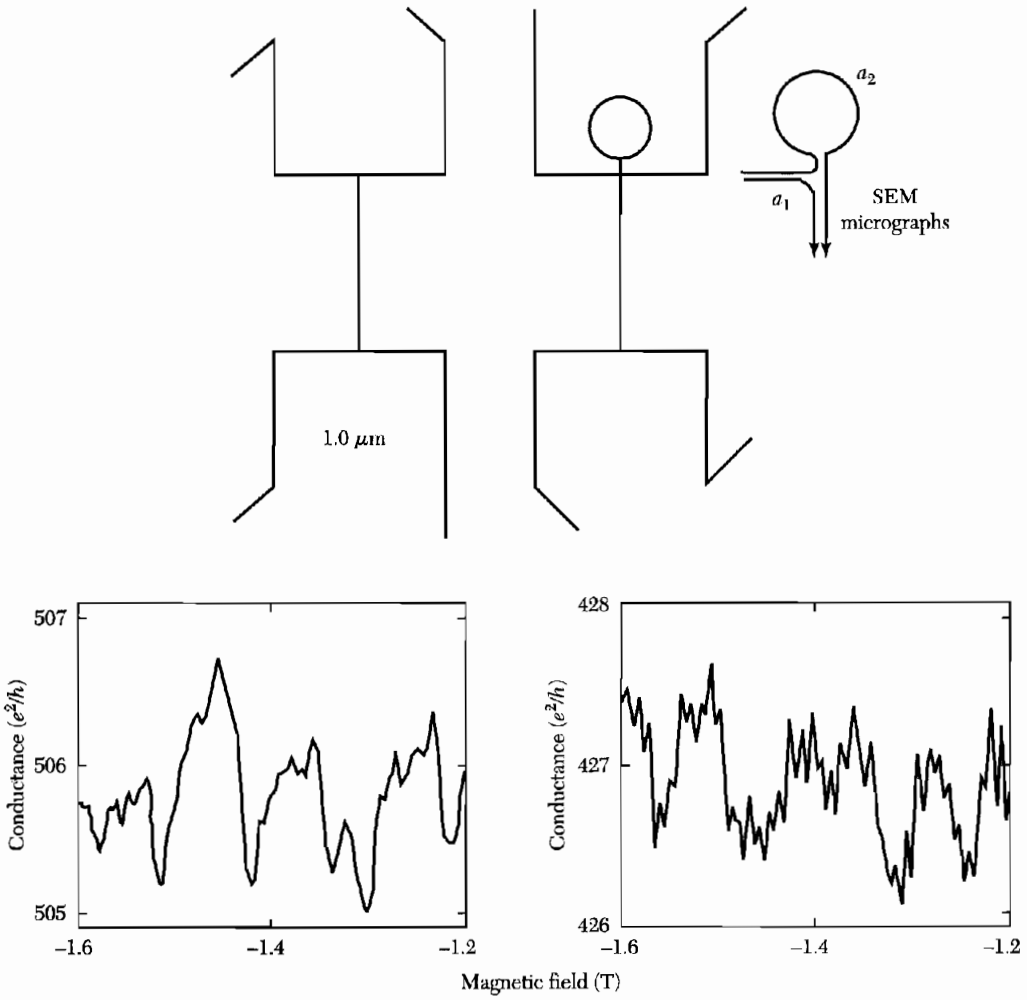


Figure 18 Upper: SEM micrographs of two vertical Au wires with current and voltage probes attached. In the device on the right, an extra loop has been added outside the region between the probes. The diagram to the right shows two paths, one that encircles the ring and one that does not. Lower left: Conductance versus magnetic field for the left sample. Aperiodic conductance fluctuations are seen due to quantum interference between the conducting paths through the sample. Lower right: Periodic oscillations are observed associated with the Aharonov-Bohm effect for paths enclosing the loop nominally outside the region between the contacts, showing the nonlocal nature of diffusive coherent transport in mesoscopic systems. (After R. Webb.)

It is remarkable that the addition of a loop outside of the region between the voltage contacts changes the measured properties. Resistances in the mesoscopic regime are nonlocal. Electrons coherently diffuse throughout the entire sample while journeying between contacts, and their phase remembers the journey.

ELECTRONIC STRUCTURE OF 0D SYSTEMS

Quantized Energy Levels

A system of electrons fully confined in all three dimensions will have discrete charge and electronic states, as do atoms and molecules. They are often called artificial atoms or **quantum dots** to reflect the importance of quantization phenomena on their properties.

As a simple example, consider an electron in a spherical potential well. Due to the spherical symmetry, the Hamitonian separates into angular and radial parts, giving eigenstates and eigenenergies:

$$\varepsilon_{n,l,m} = \varepsilon_{n,l} \; ; \qquad \psi(r,\theta,\phi) = Y_{l,m}(\theta,\phi)R_{n,l}(r) \; , \tag{50}$$

where $Y_{l,m}(\theta,\phi)$ are the spherical harmonics and $R_{n,l}(r)$ are the radial wavefunctions. The energy levels and radial wavefunctions depend on the details of the particular confining potential. For an infinite spherical well, where $V = 0$ for $r < R$ and is infinite otherwise,

$$\varepsilon_{n,l} = \hbar^2\beta_{n,l}^2/(2m^*R^2) \; ,$$
$$R_{n,l}(r) = j_l(\beta_{n,l}r/R) \; , \qquad r < R \; . \tag{51}$$

The function $j_l(x)$ is the lth spherical Bessel function and the coefficient $\beta_{n,l}$ is the nth zero of $j_l(x)$. For example, $\beta_{0,0} = \pi$ (1S), $\beta_{0,1} = 4.5$ (1P), $\beta_{0,2} = 5.8$ (1D), $\beta_{1,0} = 2\pi$ (2S), and $\beta_{1,1} = 7.7$ (2P). The labels in parentheses are the atomic notations for the states, which have the usual degeneracies associated with spin and angular momentum orientation.

Semiconductor Nanocrystals

A semiconductor nanocrystal such as the one shown in Fig. 2 can, to a good approximation, be described by the spherical model given above. Both the electron states in the conduction band and the hole states in the valence band are quantized. For a CdSe nanoparticle, the conduction band effective mass $m_e^* = 0.13\ m$, and the electron energy levels are $\varepsilon_{n,l} = (2.9\ \text{eV}/R^2)(\beta_{n,l}/\beta_{0,0})^2$, where R, the radius of the nanoparticle, is expressed in nanometers. For $R = 2$ nm, the spacing between the lowest two energy levels is $\varepsilon_{0,1} - \varepsilon_{0,0} = 0.76$ eV.

The 1S electron state increases in energy with decreasing R, while the 1S hole state decreases. The bandgap therefore grows and can be tuned over a wide range by changing R. This is shown in Fig. 19, where the absorption spectra of CdSe nanocrystals of different sizes are presented. For the smallest radii, the threshold for absorption shifts by nearly 1 eV from its bulk value. A similar shift is seen in the emission spectrum. The optical spectra of nanocrystals can be tuned continuously across the visible spectrum, making them useful in applications from fluorescent labeling to light-emitting diodes.

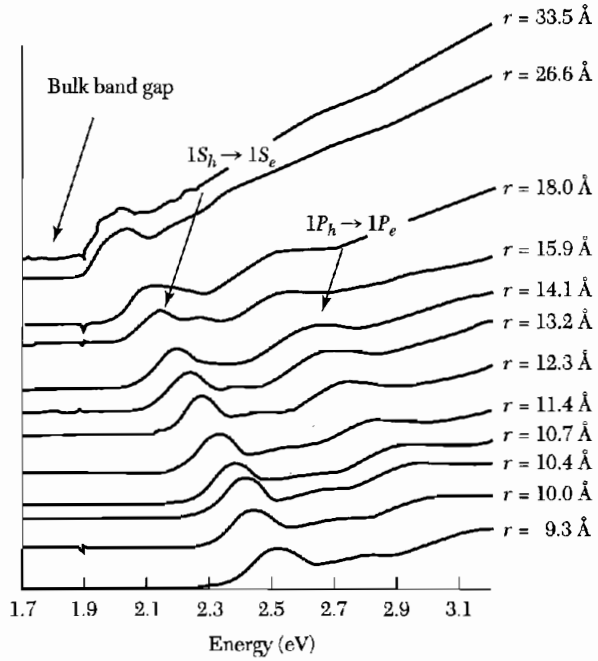


Figure 19 Optical absorption spectra for a series of CdSe nanocrystal samples of different average radii. The lowest transition energy in the smallest nanocrystal sample is shifted by nearly 1 eV from the bulk bandgap. Two dominant transitions are labeled. (Courtesy of A. P. Alivisatos.)

The absorption intensity in nanocrystals becomes concentrated at the specific frequencies corresponding to the transitions between discrete states, as described by (7). An important result for the integrated absorption can be obtained from the Kramers-Kronig relations discussed in Chapter 15. From Eq. (15.11b), we have:

$$\sigma''(\omega) = -\frac{2\omega}{\pi} P \int_0^\infty \frac{\sigma'(s)}{s^2 - \omega^2} ds \ . \tag{52}$$

At very high frequencies, $\omega \rightarrow \infty$, the electron's response is identical to that of a free electron. By (15.20),

$$\sigma''(\omega) = ne^2/m\omega \ . \tag{53}$$

In addition, as $\omega \rightarrow \infty$, the frequency s in the denominator of (52) can be neglected. Combining (52) and (53) then gives

$$\int_0^\infty \sigma'(s) ds = \pi ne^2/2m \ . \tag{54}$$

A bulk semiconductor and a nanocrystal therefore have the same overall absorption per unit volume when integrated over all frequencies. It is distributed very differently, however. The absorption spectrum of macroscopic

semiconductor is continuous, but in a nanocrystal it consists of a series of discrete transitions with very high absorption intensity at the transition frequencies. These strong transitions at particular frequencies have motivated researchers to create lasers that work on the quantized electronic transitions of quantum dots.

Metallic Dots

For small spherical metallic dots, such as alkali-metal clusters created in an atomic beam, the electrons in the conduction band fill up the quantized energy levels described by (50), as shown in Fig. 20a. These quantized levels affect the electrical and optical properties, and can even influence the stability of the dot. Small clusters can be analyzed by mass spectroscopy to determine the number of atoms in the cluster (Fig. 20b). Since there is one conduction electron per atom in an alkali metal, this is also the number of electrons in the conduction band. Large abundances are seen at certain "magic numbers" of atoms in the cluster. These result from the enhanced stability for clusters with filled electronic shells. For example, the 8-atom cluster peak corresponds to the filling of the 1S ($n = 1$, $\ell = 0$) and the 1P ($n = 1$, $\ell = 1$) shells. These filled-shell clusters are analogous to chemically stable filled-shell atoms (the noble gases).

For larger or irregularly shaped metallic dots, the shell structure is destroyed. The level spacing becomes small in comparison to the shifts in the levels due to shape imperfections, faceting of the crystals, or disorder. While the details of the level spectrum are difficult to predict, the average level spacing at the Fermi energy can be estimated using (6.21) as

$$\Delta\varepsilon = 1/D(\varepsilon_F) = 2\varepsilon_F/3N \ . \tag{55}$$

For a spherical Au nanoparticle with $R = 2$ nm, the average level spacing is $\Delta\varepsilon \sim 2$ meV. This is much smaller than the spacing between the lowest states in the CdSe nanocrystal conduction band (0.76 eV) calculated earlier. Energy-level quantization effects are much more important in semiconductor dots than they are in metallic ones. This is because the energy-level spacing is larger for low-lying states in a 3D potential well and also because the electron effective mass for semiconductors is typically smaller than for metals.

The optical properties of a small metallic dot are typically dominated by its **surface plasmon resonance**. The polarizability of a sphere is, from (16.11),

$$\text{(CGS)} \quad P = \frac{\chi E_0}{1 + 4\pi\chi/3} \ ; \qquad \text{(SI)} \quad P = \frac{\chi\varepsilon_0 E_0}{1 + \chi/3} \ , \tag{56}$$

where E_0 is the external electric field and χ is the electronic susceptibility. This relation was presented in Chap. 16 for the static case, but it applies at high frequencies as long as the dot is small enough for retardation effects to be

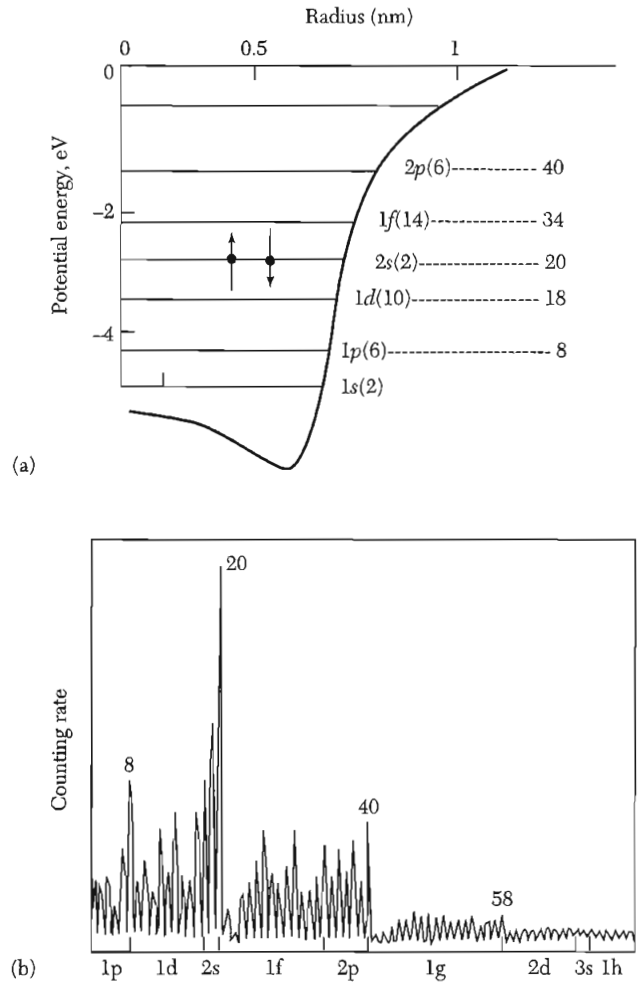


Figure 20 (a) Energy level diagram for the states in a small spherical alkali metallic cluster. The numbers at right of the diagram show the number of electrons required to fill successive electronic shells. (b) Abundance spectrum of Na clusters, showing high intensities for clusters with completely filled electronic shells. (After W. A. de Heer et al.)

absent. Modeling the carriers in the dot as a lossless free electron gas, the susceptibility is, from (14.6),

$$\text{(CGS)}\quad \chi(\omega) = -ne^2/m\omega^2\ ; \qquad \text{(SI)}\quad \chi(\omega) = -ne^2/m\varepsilon_0\omega^2\ . \tag{57}$$

Combining (56) and (57) gives

$$\text{(CGS)}\quad P(\omega) = \frac{E_0}{1-4\pi ne^2/3m\omega^2} = \frac{E_0}{1-\omega_p^2/3\omega^2}\ ;$$

$$\text{(SI)}\quad P(\omega) = \frac{\varepsilon_0 E_0}{1-ne^2/3m\varepsilon_0\omega^2} = \frac{\varepsilon_0 E_0}{1-\omega_p^2/3\omega^2}\ , \tag{58}$$

where ω_p is the plasma frequency of the bulk metal. The polarization diverges at a frequency:

$$\omega_{sp} = \omega_p/\sqrt{3} \ . \tag{59}$$

This is the surface plasmon resonance frequency for a sphere. It shifts the bulk plasma resonance for metals like Au and Ag from the UV into the visible portion of the spectrum. The result (59) is independent of particle size. In reality, however, the optical properties do depend somewhat on size due to retardation effects at larger radii and losses and intraband transitions at smaller radii.

Liquids or glasses containing metallic nanoparticles are often brilliantly colored due to absorption by the surface plasmon resonance. They have been used for hundreds of years in stained glasses. Other optical applications of metallic nanoparticles make use of the large electric field just outside the nanoparticle near resonance. In techniques like surface enhanced Raman scattering (SERS) or second harmonic generation (SHG), weak optical processes in nanostructures near the surface of the nanoparticle become measurable due to the locally high electric fields.

Discrete Charge States

If a quantum dot is relatively isolated electrically from its environment, it has a set of well-defined charge states, like an atom or molecule. Each successive charge state corresponds to the addition of one more electron to the dot. Because of the coulomb repulsion between electrons, the energy difference between successive charge states can be very large. Within the Thomas-Fermi approximation (28), the electrochemical potential for adding the $(N + 1)$th electron to a dot containing N electrons is given by:

$$\mu_{N+1} = \varepsilon_{N+1} - e\varphi = \varepsilon_{N+1} + NU - \alpha e V_g \ , \tag{60}$$

where U is the coulomb interaction energy between any two electrons on the dot, often called the **charging energy**. The dimensionless number α is the rate at which a voltage V_g applied to a nearby metal, typically referred to as the gate (see Fig. 21), shifts the electrostatic potential φ of the dot.

In general, U will vary for different electronic states in the dot, but we assume here it is a constant, as in a classical metal. In this case, we can describe the electrostatics and interactions in terms of capacitances:

$$U = e^2/C \qquad \text{and} \qquad \alpha = C_g/C \ , \tag{61}$$

where C is the total electrostatic capacitance of the dot and C_g is the capacitance between the dot and the gate. The quantity e/C is the electrostatic potential shift of the dot when one electron is added.

If the dot is in weak electrical contact with a metallic reservoir, electrons will tunnel onto the dot until the electrochemical potential for adding another electron exceeds the electrochemical potential μ of the reservoir (Fig. 21).

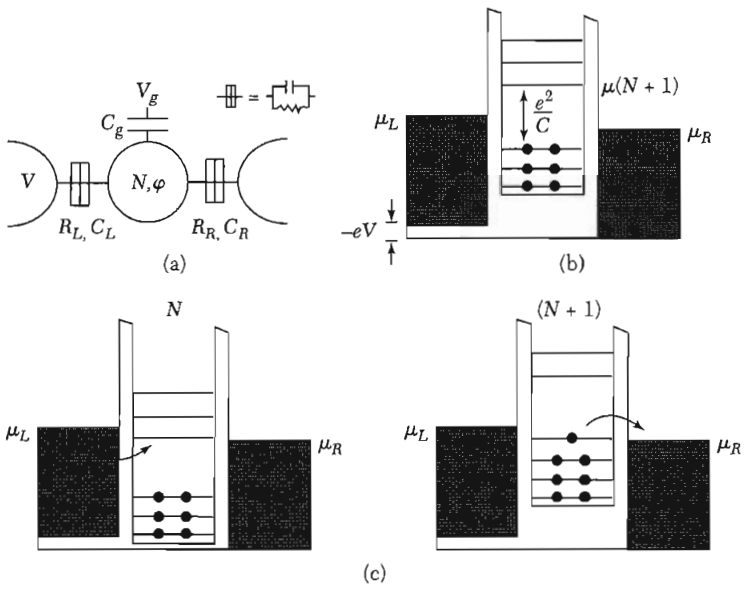


Figure 21 (a) Schematic illustration of a quantum dot in tunnel contact with two metallic reservoirs and capacitatively coupled to a gate. Main: Energy level diagrams illustrating the coulomb blockade. In (b) the gate voltage is such that the dot is stable with N electrons, so no current flows. In (c) the blockade is lifted when the electrochemical potential is lowered into the window between the potentials in the leads, allowing successive charging and discharging of the dot and a net current flow.

This sets the equilibrium occupancy N of the dot. The charge state can be changed using the gate voltage V_g. The additional gate voltage ΔV_g required to add one more electron from a reservoir of fixed μ is, from (60),

$$\Delta V_g = (1/\alpha e)(\varepsilon_{N+1} - \varepsilon_N + e^2/C) \ . \tag{62}$$

Adding an extra electron to the dot requires enough energy to fill up the next single-particle state and also enough energy to overcome the charging energy.

The charging energy U depends on both the size of the dot and the local electrostatic environment. Nearby metals or dielectrics will screen the coulomb interaction and reduce the charging energy. In general, U must be calculated for the specific geometry. As a simple model, consider a spherical dot of radius R surrounded by a spherical metal shell of radius $R + d$. This shell screens the coulomb interaction between electrons on the dot. An elementary application of Gauss's law (Problem 5) gives the capacitance and therefore the charging energy:

$$\text{(CGS)} \quad U = \frac{e^2}{\varepsilon R}\frac{d}{R+d} \ ; \qquad \text{(SI)} \quad U = \frac{e^2}{4\pi\varepsilon_0 \varepsilon R}\frac{d}{R+d} \ . \tag{63}$$

For $R = 2$ nm, $d = 1$ nm, and $\varepsilon = 1$, the charging energy is $e^2/C = 0.24$ eV. This far exceeds $k_B T \approx 0.026$ eV at room temperature, indicating that thermal

fluctuations in the charge of the dot will be strongly suppressed. It is comparable to the energy level spacing (0.76 eV) between the lowest two states in a 2-nm-radius CdSe dot. In contrast, it is much larger than the level spacing (2 meV) for a 2-nm-radius metallic dot. The addition energy of a metallic dot is therefore dominated by the charging energy, but in a semiconductor dot the charging energy and the level spacing are of comparable importance.

Charging effects are destroyed if the tunneling rate between the dot and the electrodes is too rapid. The charge resides on the dot for a time scale of order $\delta t = RC$, where R is the resistance for tunneling to the electrodes. By the uncertainty principle, the energy level will be broadened by

$$\delta\varepsilon \approx h/\delta t = h/RC = (e^2/C)(h/e^2)/R \ . \tag{64}$$

The uncertainty in the energy of the electron becomes comparable to the charging energy when $R \sim h/e^2$. For resistances below this value, quantum fluctuations due to the uncertainty principle smear out the coulomb charging effects. The conditions for well-defined charge states of a quantum dot are then

$$R \gg h/e^2 \qquad \text{and} \qquad e^2/C \gg k_B T \ . \tag{65}$$

ELECTRICAL TRANSPORT IN 0D

Coulomb Oscillations

At temperatures $T < (U + \Delta\varepsilon)/k_B$, the charging energy U and the level spacing $\Delta\varepsilon$ control the flow of electrons through a quantum dot, as shown in Fig. 21. Transport through the dot is suppressed where the Fermi levels of the leads lie between the electrochemical potential for the N and $N + 1$ charge states (Fig. 21b). This is called the **Coulomb blockade**. Current can only flow when $\mu_e(N + 1)$ is lowered to lie between the Fermi levels of the left and right leads. Then an electron can hop on the dot from the left electrode and off the dot to the right electrode, resulting in current flow (Fig. 21c). This process repeats with increasing V_g for each new charge state. This leads to so-called Coulomb oscillations in the conductance as a function of V_g, shown in Fig. 22. If $U \gg k_B T$, these peaks can be very sharp. The spacing between the Coulomb peaks is determined by (62).

Coulomb oscillations are first and foremost a result of charge quantization. They will occur if $U \gg k_B T$ even if the single particle level spacing is very small, $\Delta\varepsilon \ll k_B T$. This is often the case in metallic quantum dots. A device showing Coulomb oscillations is called a **single electron transistor** (SET), since it turns on and off periodically as the occupancy of the dot is changed by e. This effect is quite remarkable, and can be used as an ultrasensitive electrometer. It detects electric fields much as a SQUID (Chap. 10) detects magnetic fields. One is based on the quantization of charge, the other on the quantization of flux.

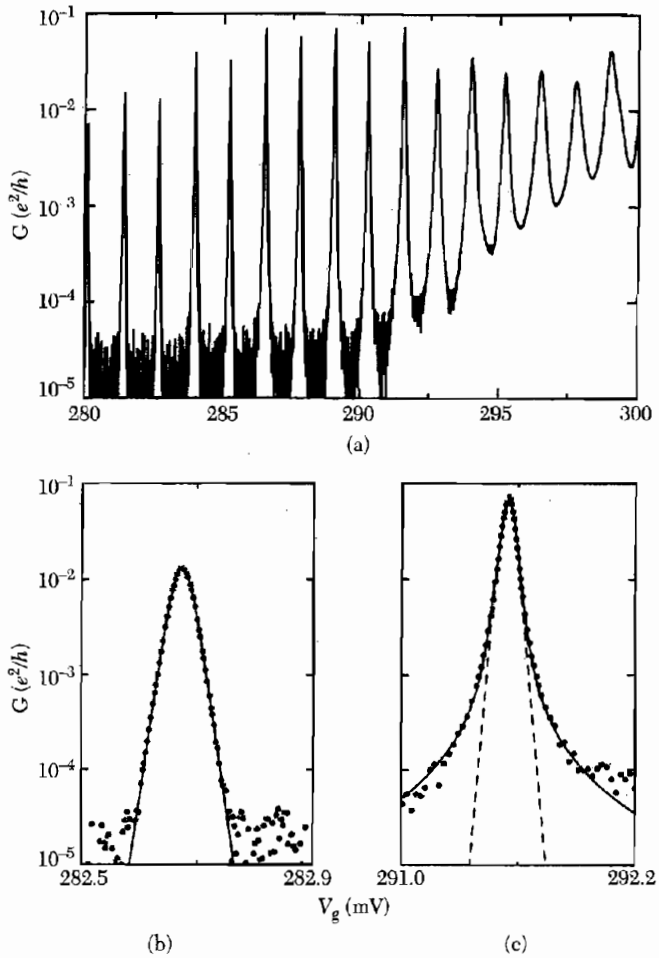


Figure 22 Conductance oscillations versus gate voltage V_g measured in a quantum dot formed in a gated GaAs/AlGaAs heterostructure at $T = 0.1$ K. The data are plotted on a log scale. As the gate voltage increases, the barriers become more transparent and the peaks get broader. The lineshape of the peak in (b) is determined by thermal broadening alone, while the one in (c) also reflects the intrinsic Breit-Wigner lineshape. (Adapted from Foxman et al.)

SETs can also be used to make single electron turnstiles and pumps. Oscillating voltages at a frequency f applied to the gates of a properly designed quantum dot system can shuttle a single electron through dot per cycle of the oscillation. This results in quantized current flowing through the dot:

$$I = ef \ . \tag{66}$$

Such devices are under investigation as current standards in metrology.

For the quantum dot in Fig. 22, the level spacing $\Delta\varepsilon \gg k_B T$. The Nth Coulomb oscillation then corresponds to resonant tunneling through a single

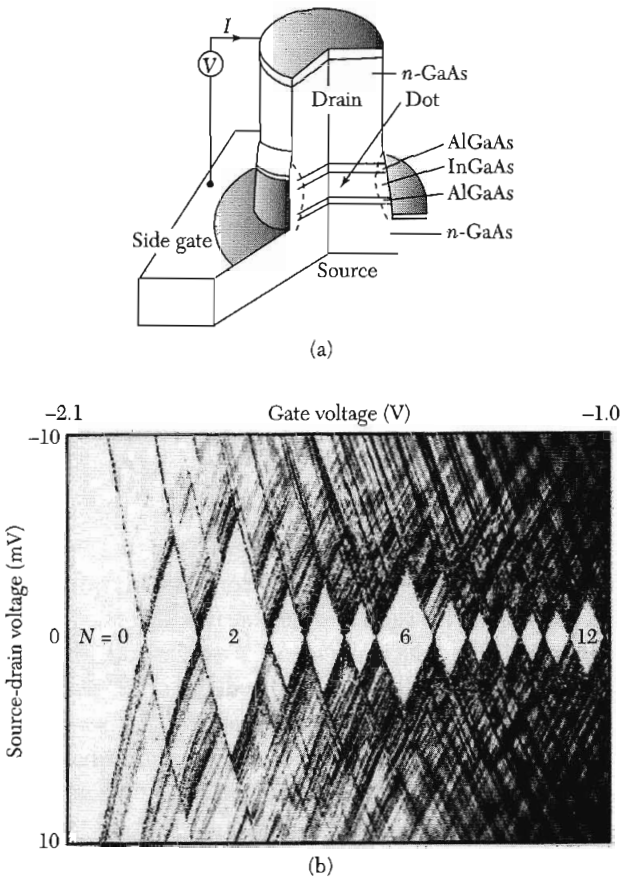


Figure 23 (a) Schematic of a 2D circular quantum dot formed in a GaAs/AlGaAs heterostructure. (b) The differential conductance dI/dV as a function of both gate voltage and source drain bias, plotted as a gray scale. The white diamond regions correspond to different charge states of the dot. A larger charging energy is observed for $N = 2$ and 6 electrons on the dot, corresponding to filled electronic shells. The additional lines on the diagram correspond to excited energy levels of the dot. (Courtesy of L. Kouwenhoven.)

quantized energy level ε_N. The coulomb oscillations in this case are analogous to the theoretical resonant tunneling peaks described by (29) and shown in Fig. 15. A crucial difference from (29) is that the positions of the coulomb peaks are determined by both the level spacing and the coulomb charging energy (62). The lower right panel of Fig. 22 shows a fit of one coulomb peak to the Breit-Wigner form for resonant tunneling (33).

The *I-V* characteristics of a quantum dot are in general complex, reflecting the interplay of the charging energy, the excited state level spacing, and the source-drain bias voltage. In Fig. 23, measurements of the first few electrons added to a small, 2D circular dot are shown. The differential conductance dI/dV is represented using a gray scale as both the gate voltage and the

source drain bias are varied. Each of the lines seen corresponds to tunneling through an individual quantum state of the dot. The white diamonds along the V_g axis, indicating $dI/dV = 0$, correspond to the Coulomb blockade. Each successive diamond corresponds to another electron on the dot. The point at which different diamonds touch along the V_g axis are the Coulomb oscillations where the charge state of the dot changes. The height of the diamond corresponds to $eV_{max} = e^2/C + \Delta\varepsilon$, the maximum voltage that can be applied without current flowing in a given charge state. The diamonds corresponding to $N = 2$ and $N = 6$ are noticeably larger than neighboring diamonds, indicating a larger addition energy for adding the third and seventh electron to the dot.

This dot can be effectively modeled by a 2D harmonic oscillator confining potential $U(x, y) = \frac{1}{2}m\omega^2(x^2 + y^2)$, with energy levels:

$$\varepsilon_{ij} = (i + j + 1)\hbar\omega \ , \tag{67}$$

where i and j are nonnegative integers. This level spectrum can be used in conjunction with (62) to find the addition energies that determine the sizes of the diamonds in the figure. The first electron fills the spin-degenerate ground state energy level ε_{00}. The second electron fills the same quantum state, but with opposite spin, at a gate voltage $\Delta V_{g2} = U/\alpha e$ after the first electron is added. The third electron fills one of the degenerate states ε_{01} or ε_{10} after a $\Delta V_{g3} = (U + \hbar\omega)/\alpha e$. The next three electrons fill the rest of these states, each spaced in gate voltage by $U/\alpha e$. The seventh electron fills one of the degenerate states ε_{11}, ε_{20}, or ε_{02}, after a gate voltage $\Delta V_{g7} = (U + \hbar\omega)/\alpha e$. This simple model correctly predicts the larger addition energies for the 3rd and 7th electrons seen in the experiment. The addition energy is larger by the level spacing when the extra electron is added to a new energy level above a filled electronic shell ($N = 2$ and $N = 6$).

Spin, Mott Insulators, and the Kondo Effect

Consider a quantum dot that is occupied by an odd number of electrons in the blockaded region, as shown in Fig. 24. The highest single particle energy level of the dot is doubly degenerate and an electron can either reside in a spin-up or spin-down state. The addition of a second electron of opposite spin is allowed by the Pauli exclusion principle but is energetically prohibited by the coulomb interaction between the electrons. This is analogous to the Mott Insulator where a half-filled band is insulating because coulomb interactions prohibit double-occupancy of the lattice sites by electrons.

The dot therefore has a spin-½ magnetic moment with two degenerate configurations, spin up and spin down, in the absence of coupling to the leads. If coupling to the leads is included, however, this degeneracy is lifted at low temperatures. The ground state is a quantum superposition of the two spin configurations, with transitions between them accomplished by a virtual

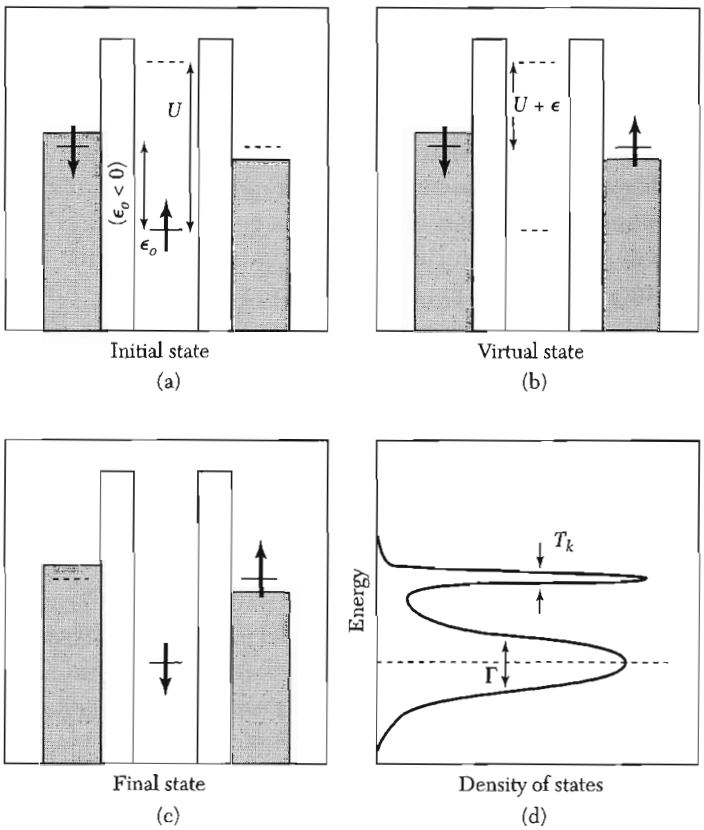


Figure 24 The Kondo effect in a quantum dot. For an unpaired spin on the dot, a virtual process (b) can occur that converts the spin up (a) to the spin down (c) state and transfers an electron from one side of the dot to the other. The ground state of the system is a coherent superposition on the initial and final states shown, creating a spin singlet between the spin on the dot and the spins in the leads. This is called the Kondo effect, and produces a narrow peak of width $\sim k_B T_K$ in the density of states at ε_F in addition to the original broadened level of width Γ, as shown in (d).

intermediate state involving an exchange of electrons with the leads, as is illustrated in Fig. 24. This is known as the **Kondo effect**. The local moment pairs with electrons in the metallic electrodes to create a spin singlet. This occurs below a temperature known as the Kondo temperature T_K:

$$T_K = \tfrac{1}{2}(\Gamma U)^{1/2} \exp\left[\pi \varepsilon_0 (\varepsilon_0 + U)/(\Gamma U)\right] \tag{68}$$

where Γ is the level width defined in (33) and ε_0 is indicated in Fig. 24 ($\varepsilon_0 < 0$). A peak in the density of states of the dot of width $k_B T_K$ appears at the Fermi energy due to the admixture of states in the electrodes at an energy ε_F. The Kondo temperature is very small unless the coupling Γ to the leads is large, since the process involves a virtual intermediate state.

Because the Kondo effect involves exchange of electrons with the leads, it causes transmission through the dot even in the blockaded region, as illustrated in Fig. 24a–c. For symmetric barriers and $T \ll T_K$, the transmission coefficient through the Kondo resonance can be unity, just as for resonant tunneling. This effect has been seen in transport through quantum dots and in STM measurements of magnetic impurities on metal surfaces. The Kondo effect was first observed in metals containing magnetic impurities. The formation of a spin singlet between the magnetic impurities and the conduction electrons enhances the scattering of the electrons. This will be discussed further in Chap. 22.

Cooper Pairing in Superconducting Dots

In a small metallic dot made of a superconductor, there is an interesting competition between single electron charging and the Cooper pairing of electrons. With an odd number of electrons residing on the dot, there is

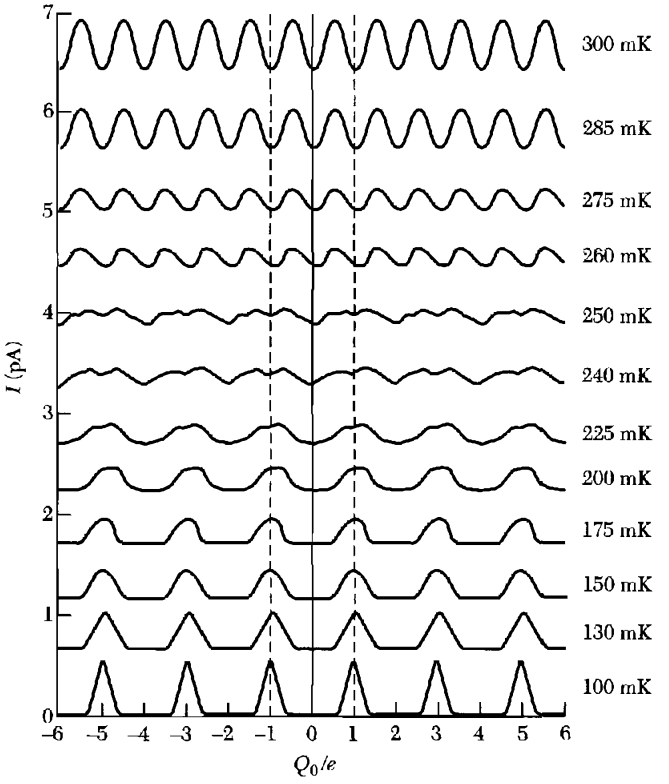


Figure 25 Measurement of coulomb oscillations in a superconducting metallic dot with decreasing temperature. A crossover from e-periodic oscillations to $2e$-periodic oscillations is seen as the temperature is lowered due to the Cooper pairing of electrons on the dot. (After M. Tinkham, J. M. Hergenrother, and J. G. Lu.)

necessarily an unpaired electron. If the Cooper pair binding energy 2Δ is larger than the charging energy U, it is energetically favorable to add an electron to the dot, paying the energy U in order to gain the pairing energy 2Δ. The odd-charge states are thus energetically unfavorable. Electrons will be added to the dot in Cooper pairs, and the Coulomb oscillations will be $2e$-periodic. This is shown in Fig. 25. This is a remarkable manifestation of Cooper pairing.

VIBRATIONAL AND THERMAL PROPERTIES

To treat the vibrational properties of nanostructures, we will begin from a continuum description of the elastic properties. This is analogous to employing the band structure as the starting point to describe the electronic properties. It is a good approximation for all but the smallest of nanostructures.

In general, the components of stress and strain in a solid are related by a matrix. A stress along one axis will produce a strain along that axis, but it will also produce strains along other axes. For example, a cube stretched along one axis will typically contract somewhat along the orthogonal axes. To simplify the discussion below, we will ignore the off-diagonal elements and treat the stress-strain matrix as diagonal and isotropic. In other words, strains will only occur along the direction of the stress and the magnitude will be independent of the axis direction. For a more complete treatment, we refer the reader to advanced texts on mechanics.

Quantized Vibrational Modes

Just as the electronic degrees of freedom are quantized, the vibrational frequencies become discrete in a 1D or 0D solid. The continuous low-frequency modes associated with the acoustic modes, $\omega = v_s K$, become instead a series of discrete frequencies ω_j. The exact frequencies and wavevectors depend in detail on the shape and boundary conditions of the solid.

An illustrative example is the vibrations around the circumference of a thin cylinder of radius R and thickness $h \ll R$, as shown in Fig. 26. In Fig. 26a,

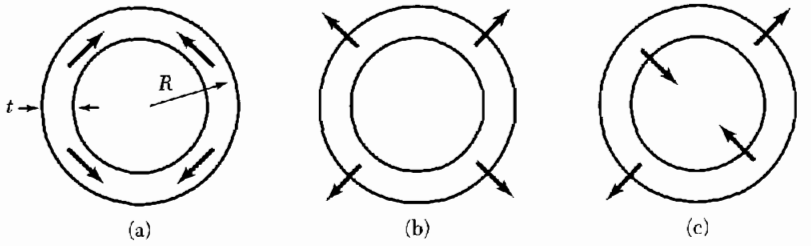


Figure 26 Fundamental vibrational modes of a thin-walled cylinder. Image (a) is a longitudinal compressional mode, (b) is the radial breathing mode (RBM), and (c) is a transverse mode.

a quantized longitudinal acoustic mode is schematically illustrated. The allowed frequencies can be found by applying periodic boundary conditions around the circumference of the cylinder:

$$K_j = j/R \ ; \qquad \omega_{Lj} = (v_L/R)j \ ; \qquad j = 1, 2 \ldots . \tag{69}$$

Another mode, called the **radial breathing mode** (RBM), is shown in Fig. 26b. The radius of the cylinder uniformly expands and contracts, producing circumferential tension and compression. From elasticity theory, the elastic energy associated with a strain e for an isotropic medium is given by

$$U_{\text{tot}} = \tfrac{1}{2}\int_V Ye^2 dV \ , \tag{70}$$

where Y is the elastic (Young's) modulus. The strain in the cylinder for a radius change dr is $e = dr/R$, yielding

$$U_{\text{tot}} = \frac{YV}{2R^2}(dr)^2 \ , \tag{71}$$

where V is the volume of the cylinder. Equation (71) has the form of a Hooke's law spring energy, where the spring constant is given by YV/R^2. The vibrational frequency is then

$$\omega_{\text{RBM}} = (YV/MR^2)^{1/2} = (Y/\rho)^{1/2}/R = v_L/R \ , \tag{72}$$

where in the last step we have defined a longitudinal sound velocity $v_L = \sqrt{Y/\rho}$.

The final class of quantized modes around the circumference, corresponding to transverse acoustic modes, are shown in Fig. 26c. Their wavevectors and frequencies are given by

$$K_j = j/R \ ; \qquad \omega_{Tj} \cong \frac{v_L h j^2}{\sqrt{12}R^2} \ ; \qquad j = 1, 2 \ldots . \tag{73}$$

Note the frequency scales like K_j^2. The origin of this behavior will be discussed further below.

Quantized vibrational modes can be measured in a variety of ways. One technique widely used to probe the vibrational structure of individual nanoscale objects is Raman spectroscopy (see Chap. 15). Raman spectroscopy of single nanotubes is shown in Fig. 27. For a nanotube, v_L = 21 km/s, and the energy of the radial breathing mode is, from (72),

$$\hbar\omega_{\text{RBM}} = 14 \text{ meV}/R[\text{nm}] \ .$$

The measured values are in good agreement with this expression. As a result, measurements of the RBM can be used as a diagnostic to infer the radius of a nanotube.

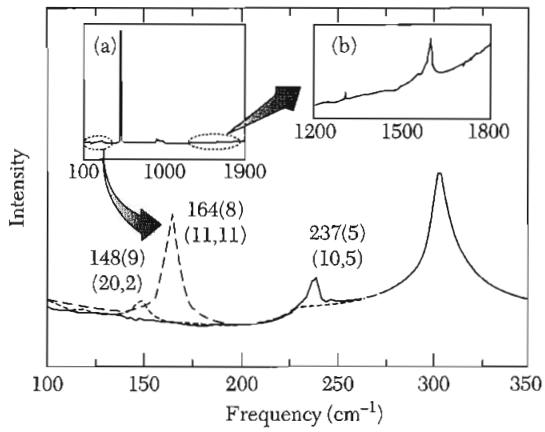


Figure 27 Raman spectra of individual carbon nanotubes. The radial breathing mode frequencies are labeled in the main panel, along with the structural assignments (n, m). Note: 160 cm^{-1} ≅ 20 meV. (After A. Jorio et al.)

Transverse Vibrations

We now address the phonons propagating in the direction of the axis of a long, thin object, such as the cylinder discussed in the previous section or a thin, solid beam (Fig. 28). The longitudinal phonons are similar to the 3D case, with a dispersion $\omega = v_L K$, where K is the continuous wavevector. However, there is a fundamental modification of the transverse phonons at wavelengths longer than the thickness h of the beam. Instead of shearing, as in a bulk transverse phonon, the solid bends, as shown in Fig. 28a. This is the classical problem of transverse flexural waves on a beam. The energy of bending comes from the longitudinal compression/stretching of the solid along the inner/outer arcs of the bend. The linear wavevector dependence $\omega_T = v_T K$ characteristic of a bulk solid is changed to dispersion quadratic in K, as we show below.

Consider a transverse standing wave on a solid rectangular beam of thickness h, width w, and length L whose displacement is given by $y(z,t) = y_0 \cos(Kz - \omega t)$. The strain at a given point inside the beam is given by the local curvature and the distance t from the center line of the beam (Fig. 28a):

$$e = -(\partial^2 y/\partial z^2)t = K^2 y t \ . \tag{74}$$

The total energy associated with this strain is again given by (70):

$$U_{\text{tot}} = (wY/2)\int_0^L \int_{-h/2}^{h/2} (K^2 y t)^2 dt\, dz = YVK^4 h^2 \langle y^2 \rangle / 24 \ , \tag{75}$$

where $\langle y^2 \rangle$ is averaged over one period of the oscillation. This again gives an effective spring constant and, from steps analogous to (70–72), an oscillation frequency:

$$\omega_T = v_L h K^2 / \sqrt{12} \ . \tag{76}$$

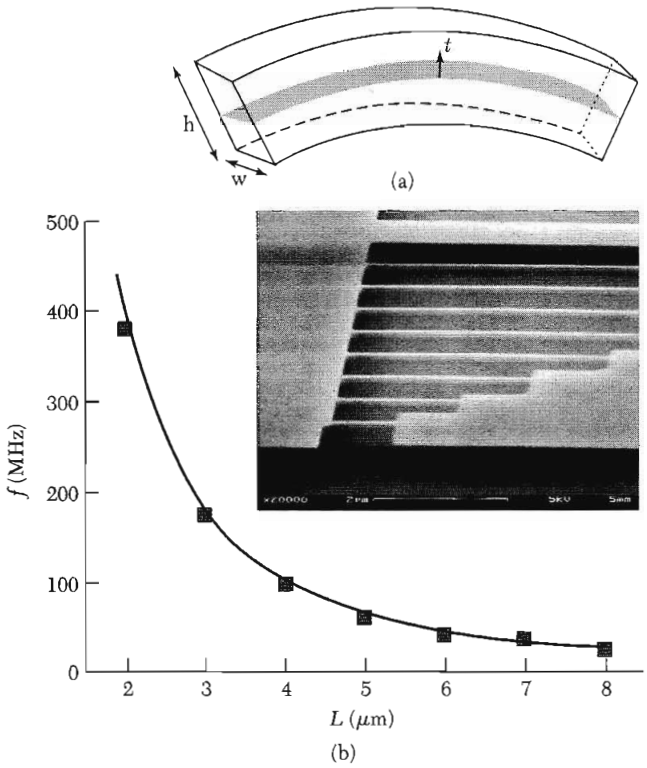


Figure 28 (a) Stresses in a bent beam, showing the inner portion under compression while the outer portion is under tension. (b) SEM micrograph of a series of suspended Si beams of varying lengths L and the measured resonance frequency as a function of L. The line is a fit to the functional form $f = B/L^2$, where B is a constant. (After D. W. Carr et al.)

Note that the frequency depends on the longitudinal sound velocity and not the transverse sound velocity, since the mode is now essentially compressional in nature. It is no longer linear in K since the effective restoring force grows stronger with increasing curvature, i.e., increasing K. In contrast, the torsional mode, corresponding to a twist of the beam along its length, retains its shear character and, $\omega_{\text{twist}} \propto K$.

Transverse vibrational modes described by (76) are frequently observed in microscale and nanoscale beams. A set of nanoscale beams constructed in Si using electron beam lithography and etching is shown in Fig. 28b. The frequencies associated with fundamental resonance $K_1 = 2\pi/L$ of these beams scale as $1/L^2$ (Fig. 28c), as expected from (76).

Note that the modification of the dispersion relation (76) for long wavelength transverse modes is not restricted to nanoscale systems. The only requirement is that the system be in the geometry of a thin beam or slab with transverse dimension h smaller than the wavelength, i.e., $Kh \ll 1$. For example, AFM cantilevers operated in the noncontact mode, as discussed above,

are well described by this relation. The dispersion relation (76) is also related to the $\omega \sim K^2$ dependence seen in (73) for the class of modes shown in Fig. 26c. Both describe the transverse flexural vibrations, one of a beam and the other of a thin shell.

A revolution is underway in the fabrication of small, complex mechanical structures using techniques adapted from microelectronic processing. The beams shown in Fig. 28b are a simple example. These structures can be integrated with electronic devices, creating microelectromechanical systems (MEMs) and nanoelectromechanical systems (NEMs). They are being explored for a variety of applications, including sensing, data storage, and signal processing.

Heat Capacity and Thermal Transport

The above relations indicate that the quantized vibrational mode energies are typically less than $k_B T$ at room temperature except in the very smallest structures. Modes along the confined directions will thus be thermally excited at room temperature. As a result, the lattice thermal properties of nanostructures will be similar to their bulk counterparts. In particular, the lattice heat capacity and thermal conductivity will be proportional to T^3, as for a 3D solid (Chap. 5).

At low temperatures, however, vibrational excitations of frequency ω in the confined directions will freeze out when $T < \hbar\omega/k_B$. In the case of a long, thin structure, the system will behave as 1D thermal system at sufficiently low temperatures, with a set of 1D phonon subbands analogous to the 1D electronic subbands shown in Fig. 9. A calculation analogous to the one performed in Chap. 5 for a 3D solid yields the heat capacity per 1D acoustic phonon subband with a dispersion $\omega = vk$ (Prob. 6):

$$C_V^{(1D)} = 2\pi^2 L k_B^2 T/3hv \ . \tag{77}$$

The thermal conductance G_{th} of the wire is defined as the ratio of the net energy flow through the wire divided by the temperature difference ΔT between its ends. The thermal conductance per 1D phonon subband $G_{th}^{(1D)}$ is

$$G_{th}^{(1D)} = (\pi^2 k_B^2 T/3h)\mathfrak{I} \ . \tag{78}$$

This result is derived in Prob. 6 using an approach analogous to the one employed to derive the Landauer formula for conductance of a 1D channel. Note $\mathfrak{I}$ is now the transmission probability for phonons through the structure. For a transverse flexural mode, where $\omega \propto K^2$, the result (77) is modified, but (78) is the same.

Both (77) and (78) are linear in temperature, in contrast to the T^3 result for 3D. The difference reflects the number of modes with energies $\hbar\omega < k_B T$, or equivalently, with wavevectors $K < k_B T/\hbar v$. The number of modes scales like K^D, where D is the dimensionality, producing T^3 in 3D and T in 1D.

Note that in the case of perfectly transmitted phonons through the channel, the thermal conductance (78) is determined only by fundamental constants and the absolute temperature. This result is analogous to the quantized electronic conductance (21) of a 1D channel, which was independent of the electron velocity in the channel. Both the ballistic thermal conductance (78) and the 1D form of the heat capacity (77) have been observed in experiments on narrow wires at very low temperatures.

SUMMARY

- Real space probes can give atomic-scale images of nanostructures.
- The density of states of a 1D subband, $D(E) = 4L/h\upsilon$, diverges at the subband thresholds. These are called van Hove singularities.
- The electrical conductance of a 1D system is given by the Landauer formula, $G = (2e^2/h)\mathfrak{I}$, where $\mathfrak{I}$ is the transmission coefficient through the sample.
- The conductance of a quasi-1D system can be strongly influenced by quantum interference among the electron paths traversing the sample, leading to resonant tunneling, localization, and the Aharonov-Bohm effect.
- The optical properties of a quantum dot can be tuned by changing its size and hence its quantized energy levels.
- Adding an extra charge e to a quantum dot requires an additional electrochemical potential given by $U + \Delta\varepsilon$, where U is the charging energy and $\Delta\varepsilon$ is the level spacing.
- The vibrational modes of a nanometer-scale object are quantized.

Problems

1. ***Carbon nanotube band structure.*** Figure 29 shows the graphene lattice with the primitive lattice translation vectors of length $a = 0.246$ nm, along with the first Brillouin zone. (a) Find the set of reciprocal lattice vectors $\mathbf{G}$ associated with the lattice. (b) Find the length of the vectors $\mathbf{K}$ and $\mathbf{K}'$ shown in Fig. 29 in terms of a.

   For energies near the Fermi energy and wavevectors near the K point, the 2D band structure can be approximated as

$$\varepsilon = \pm\hbar\upsilon_F|\Delta\mathbf{k}| \qquad \Delta\mathbf{k} = \mathbf{k} - \mathbf{K} \ ,$$

   where $\upsilon_F = 8 \times 10^5$ m/s. A similar approximation holds near the K' point. Consider a tube rolled up along the x-axis with a circumference na. By applying periodic boundary conditions along the rolled up direction, the dispersions of the 1D subbands near the K point can be found. (c) Show that, if n is divisible by 3, there exists a "massless" subband whose energy is linear in Δk_y. Sketch this subband. These nanotubes are 1D metals. (d) If n is not divisible by 3, the subband structure is that

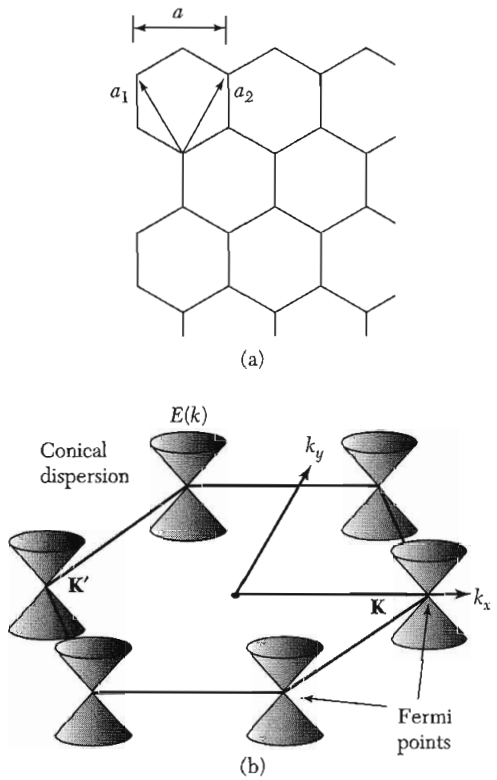


Figure 29 (a) The graphene lattice and (b) the first Brillouin zone of the graphene lattice showing the conical dispersion of the energies near the **K** and **K'** points.

as shown in Fig. 10. For the case of $n = 10$, find the magnitude of the semiconducting bandgap ε_{11} in eV and show that $\varepsilon_{22}/\varepsilon_{11} = 2$. (e) Again for the $n = 10$ case, show that the dispersion relation of the lowest electron subband is of the form of a relativistic particle, $\varepsilon^2 = (m^*c^2)^2 + (pc)^2$, where v_F plays the role of the speed of light, and find the ratio of effective mass m^* to the free electron mass m.

2. ***Filling subbands.*** For electrons in a square GaAs wire of width 20 nm, find the linear electron density at which the $n_x = 2$, $n_y = 2$ subband is first populated in equilibrium at $T = 0$. Assume an infinite confining potential at the wire boundary.

3. ***Breit-Wigner form of a transmission resonance.*** The purpose of this problem is to derive (33) from (29). (a) By expanding the cosine for small phase differences away from resonance, $\delta\varphi = \varphi^* - 2\pi n$, find a simplified form of (29) involving only $|t_1|^2$, $|t_2|^2$, and $\delta\varphi$. (b) Show that, for states in a 1D box, the following relation holds between small changes in the phases and small changes in the energy: $\delta\varepsilon/\Delta\varepsilon = \delta\varphi/2\pi$, where $\Delta\varepsilon$ is the level spacing. (c) Combine (a) and (b) to obtain (33).

4. ***Barriers in series and Ohm's Law.*** (a) Derive (36) from (35). (b) Show that the 1D Drude conductivity $\sigma_{1D} = n_{1D}e^2\tau/m$ can be written as $\sigma_{1D} = (2e^2/h)\ell_B$. (Note: The momentum relaxation rate and the backscattering rates are related as $1/\tau = 2/\tau_B$ because the former corresponds to the relaxation from **p** to 0 while the latter corresponds to relaxation from **p** to $-\mathbf{p}$.)

5. ***Energies of a spherical quantum dot.*** (a) Derive the formula (63) for the charging energy. (b) Show that, for $d \ll R$, the result is the same as that obtained using the parallel plate capacitor result, $C = \varepsilon\varepsilon_0 A/d$. (c) For the case of an isolated dot, $d \to \infty$, find the ratio of the charging energy to lowest quantized energy level. Express your answer in terms of the radius R of the dot and the effective Bohr radius a_B^*.

6. ***Thermal properties in 1D.*** (a) Derive the formula (77) for the low temperature heat capacity of a single 1D phonon mode within the Debye approximation. (b) Derive the relation for the thermal conductance (78) of a 1D phonon mode between two reservoirs by calculating the energy flow out of one reservoir at a temperature T_1 and subtracting the energy flow from the other reservoir at a temperature T_2. Use an approach analogous to that used to obtain (20) and (24) for the electrical conductance.

19

Noncrystalline Solids

CHAPTER 19: NONCRYSTALLINE SOLIDS

The terms amorphous solid, noncrystalline solid, disordered solid, glass, or liquid have no precise structural meaning beyond the description that the structure is "not crystalline on any significant scale." The principal structural order present is imposed by the approximately constant separation of nearest-neighbor atoms or molecules. We exclude from the present discussion disordered crystalline alloys (Chapter 22) where different atoms randomly occupy the sites of a regular crystal lattice.

DIFFRACTION PATTERN

The x-ray or neutron diffraction pattern of an amorphous material such as a liquid or a glass consists of one or more broad diffuse rings, when viewed on the plane normal to the incident x-ray beam. The pattern is different from the diffraction pattern of powdered crystalline material which shows a large number of fairly sharp rings as in Fig. 2.17 of Chap. 2. The result tells us that a liquid does not have a unit of structure that repeats itself identically at periodic intervals in three dimensions.

In a simple monatomic liquid the positions of the atoms show only a short range structure referred to an origin on any one atom. We never find the center of another atom closer than a distance equal to the atomic diameter, but at roughly this distance we expect to find about the number of nearest-neighbor atoms that we find in a crystalline form of the material.

Although the x-ray pattern of a typical amorphous material is distinctly different from that of a typical crystalline material, there is no sharp division between them. For crystalline powder samples of smaller and smaller particle size, the powder pattern lines broaden continuously, and for small enough crystalline particles the pattern becomes similar to the amorphous pattern of a liquid or a glass.

From a typical liquid or glass diffraction pattern, containing three or four diffuse rings, the only quantity which can be determined directly is the radial distribution function. This is obtained from a Fourier analysis of the experimental x-ray scattering curve, and gives directly the average number of atoms to be found at any distance from a given atom. The method of Fourier analysis is equally applicable to a liquid, a glass, or a powdered crystalline material.

It is convenient to begin the analysis of the diffraction pattern with Eq. (2.43). Instead of writing it for the structure factor of the basis, we write the sum for all the atoms in the specimen. Further, instead of specializing the scattering to the reciprocal lattice vectors $\mathbf{G}$ characteristic of a crystal, we consider arbitrary scattering vectors $\Delta\mathbf{k} = \mathbf{k}' - \mathbf{k}$, as in Fig. 2.6. We do this because

scattering from amorphous materials is not limited to the reciprocal lattice vectors, which in any event cannot here be defined.

Therefore the scattered amplitude from an amorphous material is described by

$$S(\Delta\mathbf{k}) = \sum_m f_m \exp(-i\Delta\mathbf{k} \cdot \mathbf{r}_m) \ , \tag{1}$$

with f_m the atomic form factor of the atom, as in Eq. (2.50). The sum runs over all atoms in the specimen.

The scattered intensity at scattering vector $\Delta\mathbf{k}$ is given by

$$I(\Delta\mathbf{k}) = S^*S = \sum_m \sum_n f_m f_n \exp[i\Delta\mathbf{k} \cdot (\mathbf{r}_m - \mathbf{r}_n)] \ , \tag{2}$$

in units referred to the scattering from a single electron. If α denotes the angle between $\Delta\mathbf{k}$ and $\mathbf{r}_m - \mathbf{r}_n$, then

$$I(K, \alpha) = \sum_m \sum_n f_m f_n \exp(iKr_{mn} \cos \alpha) \ , \tag{3}$$

where K is the magnitude of $\Delta\mathbf{k}$ and r_{mn} is the magnitude of $\mathbf{r}_m - \mathbf{r}_n$.

In an amorphous specimen the vector $\mathbf{r}_m - \mathbf{r}_n$ may take on all orientations, so we average the phase factor over a sphere:

$$\begin{aligned} \langle \exp(iKr \cos \alpha) \rangle &= \frac{1}{4\pi} 2\pi \int_{-1}^{1} d(\cos \alpha) \exp(iKr_{mn} \cos \alpha) \\ &= \frac{\sin Kr_{mn}}{Kr_{mn}} \ . \end{aligned} \tag{4}$$

Thus we have the Debye result for the scattered intensity at $\Delta\mathbf{k}$:

$$I(K) = \sum_m \sum_n (f_m f_n \sin Kr_{mn})/Kr_{mn} \ . \tag{5}$$

Monatomic Amorphous Materials

For atoms of only one type, we let $f_m = f_n = f$ and separate out from the summation (5) the terms with $n = m$. For a specimen of N atoms,

$$I(K) = Nf^2 \left[1 + \sum_m{}' (\sin Kr_{mn})/Kr_{mn} \right] \ . \tag{6}$$

The sum runs over all atoms m except the origin atom $m = n$.

If $\rho(r)$ is the concentration of atoms at distance r from a reference atom, we can write (6) as

$$I(K) = Nf^2 \left[1 + \int_0^R dr \, 4\pi r^2 \rho(r) (\sin Kr)/Kr \right] \ , \tag{7}$$

where R is the (very large) radius of the specimen. Let ρ_0 denote the average concentration; then (7) may be written as

$$I(K) = Nf^2\left\{1 + \int_0^R dr\, 4\pi r^2[\rho(r) - \rho_0](\sin Kr)/Kr + (\rho_0/K)\int_0^R dr\, 4\pi r \sin Kr\right\} . \tag{8}$$

The second integral in (8) gives the scattering from a uniform concentration and may be neglected except in the forward region of very small angles; it reduces to a delta function at $K = 0$ as $R \to \infty$.

Radial Distribution Function

It is convenient to introduce the liquid structure factor defined by

$$S(K) = I/Nf^2 . \tag{9}$$

Note that this is not at all the same as $S(\mathbf{\Delta k})$ in (1). From (8) we have, after dropping the delta function contribution,

$$S(K) = 1 + \int_0^\infty dr\, 4\pi r^2[\rho(r) - \rho_0](\sin Kr)/Kr . \tag{10}$$

We define the **radial distribution function** $g(r)$ such that

$$\rho(r) = g(r)\rho_0 . \tag{11}$$

Then (10) becomes

$$\begin{aligned} S(K) &= 1 + 4\pi\rho_0 \int_0^\infty dr[g(r) - 1]r^2(\sin Kr)/Kr \\ &= 1 + \rho_0 \int d\mathbf{r}\,[g(r) - 1]\exp(i\mathbf{K}\cdot\mathbf{r}) , \end{aligned} \tag{12}$$

because $(\sin Kr)/Kr$ is the spherically symmetric or s term in the expansion of $\exp(i\mathbf{K}\cdot\mathbf{r})$.

By the Fourier integral theorem in three dimensions,

$$\begin{aligned} g(r) - 1 &= \frac{1}{8\pi^3\rho_0}\int d\mathbf{K}\,[S(K) - 1]\exp(-i\mathbf{K}\cdot\mathbf{r}) \\ &= \frac{1}{2\pi^2\rho_0 r}\int dK\,[S(K) - 1]\,K\sin Kr . \end{aligned} \tag{13}$$

This result allows us to calculate the radial distribution function $g(r)$ (also called the two-atom correlation function) from the measured structure factor $S(K)$.

One of the simplest liquids well suited to x-ray diffraction study is liquid sodium. The plot of the radial distribution $4\pi r^2\rho(r)$ vs. r is given in Fig. 1, together with the distribution of neighbors in crystalline sodium.

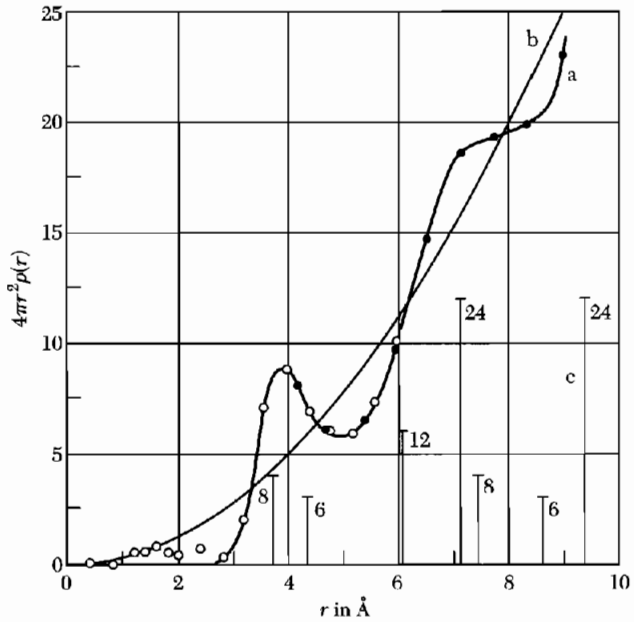


Figure 1 (a) Radial distribution curve $4\pi r^2\rho(r)$ for liquid sodium. (b) Average density curve $4\pi r^2\rho_0$. (c) Distribution of neighbors in crystalline sodium. (After Tarasov and Warren.)

Structure of Vitreous Silica, SiO_2

Vitreous silica (fused quartz) is a simple glass. The x-ray scattering curve is given in Fig. 2. The radial distribution curve $4\pi r^2\rho(r)$ vs. r is given in Fig. 3. Because there are two kinds of atoms, $\rho(r)$ is actually the superposition of two electron concentration curves, one about a silicon atom as origin and one about an oxygen atom as origin.

The first peak is at 1.62 Å, close to the average Si-O distance found in crystalline silicates. The x-ray workers conclude from the intensity of the first peak that each silicon atom is tetrahedrally surrounded by four oxygen atoms. The relative proportions of Si and O tell us that each O atom is bonded to two Si atoms. From the geometry of a tetrahedron, the O-O distance should be 2.65 Å, compatible with the distance suggested by the shoulder in Fig. 3.

The x-ray results are consistent with the standard model of an oxide glass, due to Zachariasen. Figure 4 illustrates in two dimensions the irregular structure of a glass and the regularly repeating structure of a crystal of identical chemical composition. The x-ray results are completely explained by picturing glassy silica as a random network in which each silicon is tetrahedrally surrounded by four oxygens, each oxygen bonded to two silicons, the two bonds to an oxygen being roughly diametrically opposite. The orientation of one tetrahedral group with respect to a neighboring group about the connecting

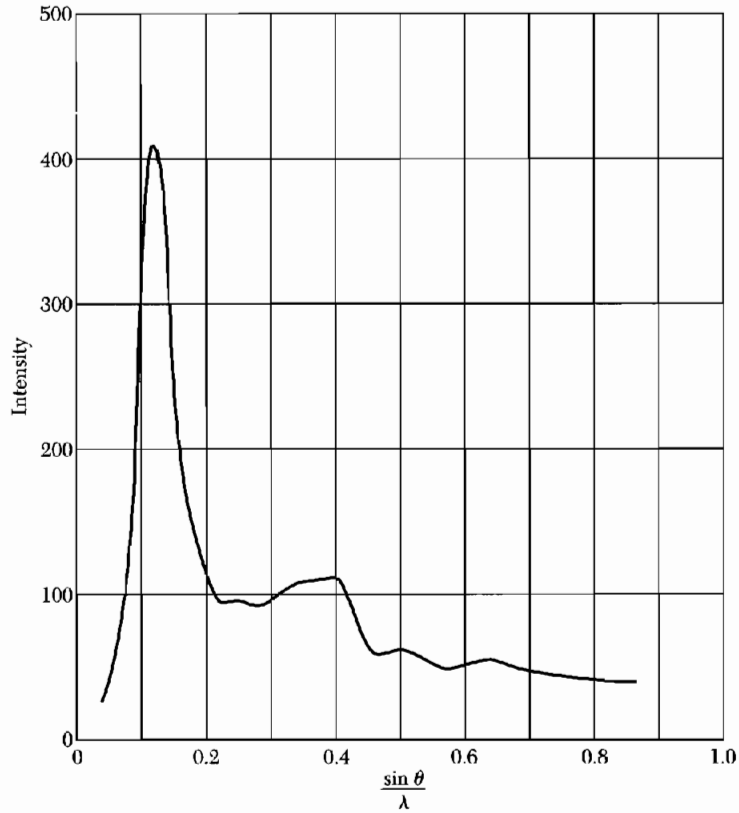


Figure 2 Scattered x-ray intensity vs. scattering angle θ, for vitreous SiO_2. (After B. E. Warren.)

Si-O-Si bond can be practically random. There is a definite structural scheme involved: each atom has a definite number of nearest neighbors at a definite distance, but no unit of structure repeats itself identically at regular intervals in three dimensions, and hence the material is not crystalline.

It is not possible to explain the x-ray results by assuming that vitreous silica consists of very small crystals of some crystalline form of quartz, such as cristoballite. Small angle x-ray scattering is not observed, but would be expected from discrete particles with breaks and voids between them. The scheme of bonding in glass must be essentially continuous, at least for the major part of the material, although the scheme of coordination about each atom is the same in vitreous silica and in crystalline cristoballite. The low thermal conductivity of glasses at room temperature, as discussed below, also is consistent with the continuous random network model.

A comparison of experimental and calculated x-ray intensity results for amorphous germanium is shown in Fig. 5. The calculations are for a random network model and for a microcrystallite model. The latter model gives a very

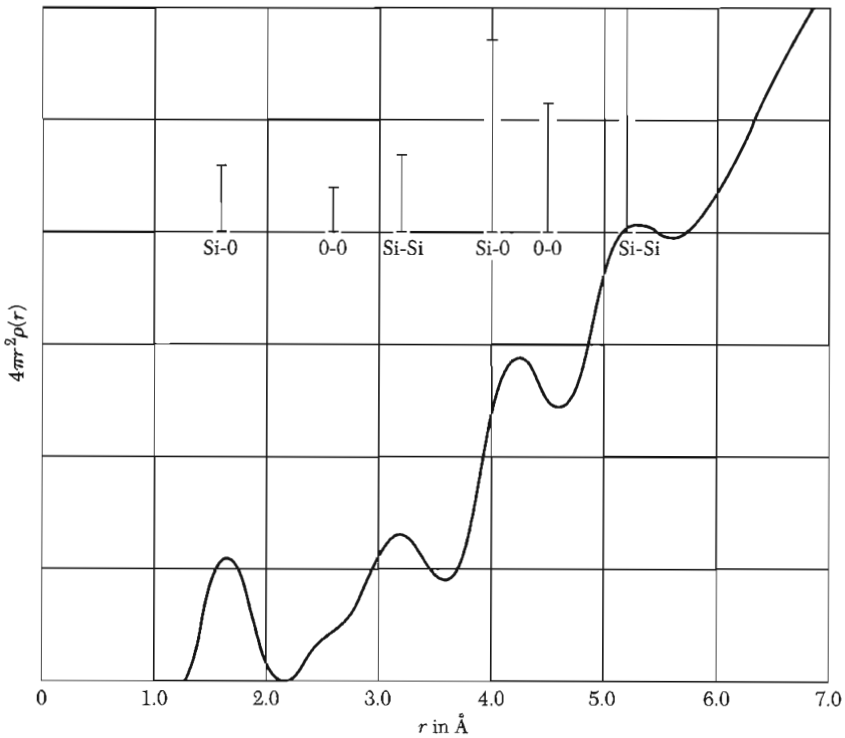
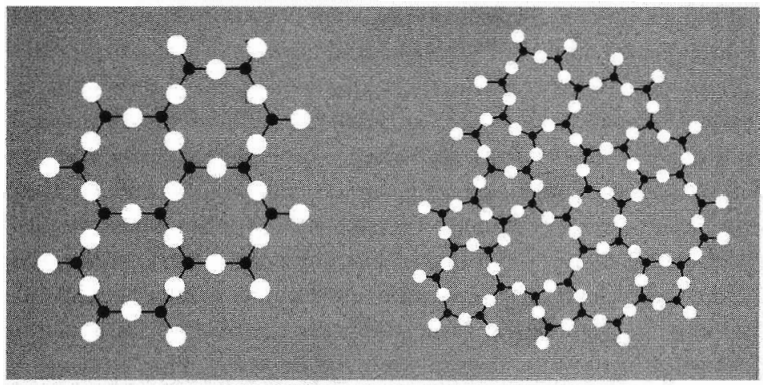


Figure 3 Radial distribution curve for vitreous SiO_2, as the Fourier transform of Fig. 2. The positions of the peaks give the distances of atoms from a silicon or an oxygen. From the areas under the peaks it is possible to calculate the number of neighbors at that distance. The vertical lines indicate the first few average interatomic distances; the heights of the lines are proportional to the peak areas. (After B. E. Warren.)



(a) (b)

Figure 4 Schematic two-dimensional analogs illustrating the differences between: (a) the regularly repeating structure of a crystal and (b) continuous random network of a glass. (After Zachariasen.)

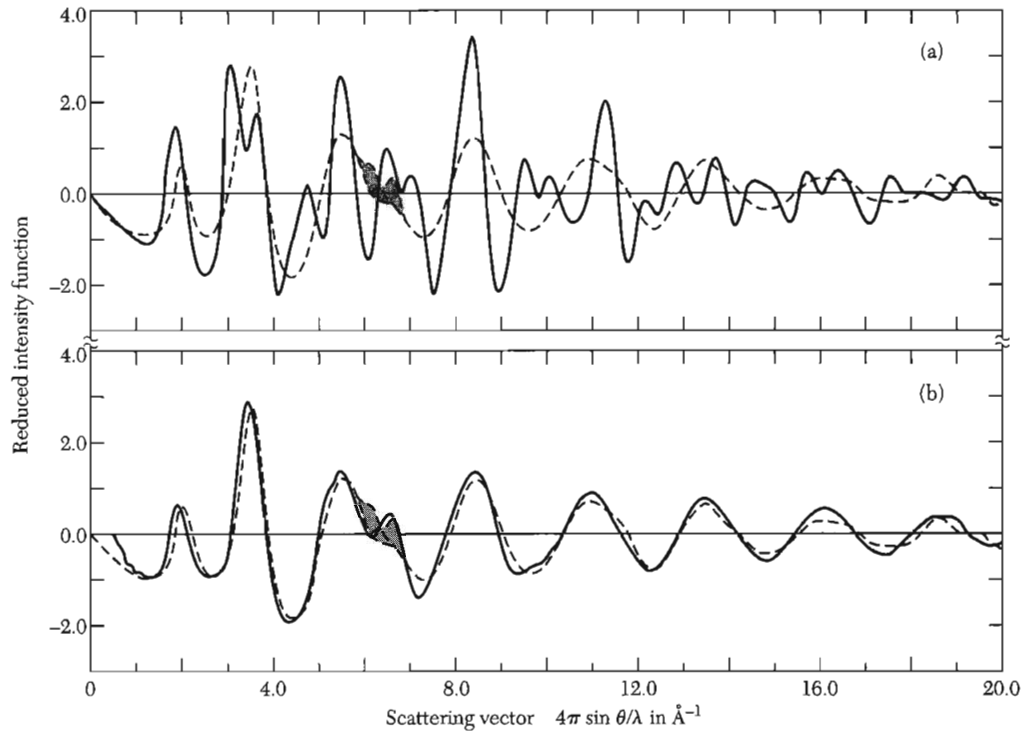


Figure 5 Comparison of experimental (dashed curve) and calculated (solid curve) reduced intensity function for amorphous germanium. (a) Amorphous germanium compared with microcrystallite model. (b) Amorphous germanium compared with random network model. (Results by J. Graczyk and P. Chaudhari.)

poor agreement. The random network model is supported for amorphous silicon by studies of the band gap and spectroscopic work on the $2p$ shell.

GLASSES

A glass has the random structure of the liquid from which it is derived by cooling below the freezing point, without crystallization. Also, a glass has the elastic properties of an isotropic solid.

By general agreement, we say that a liquid on being cooled becomes a glass when the viscosity equals 10^{13} poise, where a poise is the CGS unit of viscosity.[1] This defines the **glass transition temperature** T_g. At temperatures above T_g we have a liquid; below T_g we have a glass. The transition is not a thermodynamic phase transition, only a transition for "practical purposes."

[1]The SI unit of viscosity is 1 Nsm^{-2}, so that 1 poise = 0.1 Nsm^{-2}. It is quite common to find viscosities given in cp or centipoise, being 10^{-2} poise.

The value 10^{13} poise used to define T_g is arbitrary, but not unreasonable. If we bond a slab of glass 1 cm thick to two plane parallel vertical surfaces, the glass will flow perceptibly in one year under its own weight when the viscosity drops below 10^{13} poise. (For comparison, the viscosity of the mantle of the earth is of the order of 10^{22} poise.)

Relatively few liquids can be cooled fast enough in the bulk to form a glass before crystallization intervenes. Molecules of most substances have high enough mobility in the liquid so that on cooling a liquid-solid melting transition occurs a long time before the viscosity increases to 10^{13} poise or 10^{15} cp. Liquid water has a viscosity 1.8 cp at the freezing point; the viscosity increases enormously on freezing.

We can often make a glass by depositing a jet of atoms of a substrate cooled to a low temperature, a process which will sometimes produce an amorphous layer with glasslike properties. Amorphous ribbons of some metal alloys may be produced in this way in industrial quantities.

Viscosity and the Hopping Rate

The viscosity of a liquid is related to the rate at which molecules undergo thermal rearrangement on a local scale, as by hopping into a vacant neighbor site or by interchange of two neighbor molecules. The physics of the transport process is somewhat different from that of viscosity in the gas phase, but the gas phase result gives a qualitative lower limit to the viscosity of the liquid phase, a limit that applies to nearest-neighbor hopping of atoms.

The gas result (*TP* 14.34) is

$$\eta = \tfrac{1}{3}\rho\bar{c}\ell \ , \tag{14}$$

where η is the viscosity, ρ the density, $\bar{c}$ the mean thermal velocity, and ℓ the mean free path. In the liquid l is of the order of magnitude of the intermolecular separation a. With "typical" values $\rho \approx 2$ g cm^{-3}; $\bar{c} \approx 10^5$ cm s^{-1}; $a \approx 5 \times 10^{-8}$ cm, we have

$$\eta(\text{min}) \approx 0.3 \times 10^{-2} \text{ poise} = 0.3 \text{ cp} \tag{15}$$

as an estimate of the lower limit of the viscosity of a liquid. (Tables in chemical handbooks only rarely list values below this.)

We give now a very simple model of the viscosity of a liquid. In order to hop successfully, a molecule must surmount the potential energy barrier presented by its neighbors in the liquid. The preceding estimate of the minimum viscosity applies when this barrier may be neglected. If the barrier is of height E, the molecule will have sufficient thermal energy to pass over the barrier only a fraction

$$f \approx \exp(-E/k_B T) \tag{16}$$

of the time. Here E is an appropriate free energy and is called the activation energy for the process that determines the rate of hopping. It is related to the activation energy for self-diffusion.

The viscosity will be increased as the probability of successful hopping is decreased. Thus

$$\eta \approx \eta(\min)/f \approx \eta(\min) \exp(E/k_B T) \ . \tag{17}$$

If $\eta = 10^{13}$ poise at the glass transition, the order of magnitude of f must be

$$f \approx 0.3 \times 10^{-15} \tag{18}$$

at the transition, using (15). The corresponding activation energy is

$$E/k_B T_g = -\ln f = \ln(3 \times 10^{15}) = 35.6 \ . \tag{19}$$

If $T_g \approx 2000$ K, then $k_B T_g = 2.7 \times 10^{-13}$ erg and $E = 9.6 \times 10^{-12}$ erg ≈ 6 eV. This is a high potential energy barrier.

Glasses with lower values of T_g will have correspondingly lower values of E. (Activation energies obtained in this way are often labeled as E_{visc}.) Materials that are glass-formers are characterized by activation energies of the order of 1 eV or more; non-glass-formers may have activation energies of the order of 0.01 eV.

When being pressed into molds or drawn into tubes, glass is used in a range of temperatures at which its viscosity is 10^3 to 10^6 poises. The working range for vitreous silica begins over 2000°C, so high that the practical usefulness of the material is severely limited. In common glass, about 25 percent of Na_2O is added as a network modifier to SiO_2 in order to reduce below 1000°C the temperature needed to make the glass fluid enough for the forming operations needed to make electric lamp bulbs, window glass, and bottles.

AMORPHOUS FERROMAGNETS

Amorphous metallic alloys are formed by very rapid quenching (cooling) of a liquid alloy, commonly by directing a molten stream of the alloy onto the surface of a rapidly rotating drum. This process produces a continuous "melt-spun" ribbon of amorphous alloy in industrial quantities.

Ferromagnetic amorphous alloys were developed because amorphous materials have nearly isotropic properties, and isotropic materials should have essentially zero magnetocrystalline anisotropy energy. The absence of directions of hard and easy magnetization should result in low coercivities, low hysteresis losses, and high permeabilities. Because amorphous alloys are also random alloys, their electrical resistivity is high. All these properties have technological value for application as soft magnetic materials. The trade name Metglas is attached to several of these.

The transition metal-metalloid (TM-M) alloys are an important class of magnetic amorphous alloys. The transition metal component is usually about 80 percent of Fe, Co, or Ni, with the metalloid component B, C, Si, P, or Al. The presence of the metalloids lowers the melting point, making it possible to quench the alloy through the glass transition temperature rapidly enough to stabilize the amorphous phase. For example, the composition $Fe_{80}B_{20}$ (known as Metglas 2605) has T_g = 441°C, as compared with the melting temperature 1538°C of pure iron.

The Curie temperature of this composition in the amorphous phase is 647 K, and the value of the magnetization M_s at 300 K is 1257, compared with T_c = 1043 K and M_s = 1707 for pure iron. The coercivity is 0.04 G, and the maximum value of the permeability is 3×10^5. Coercivities as low as 0.006 G have been reported for another composition.

High coercivity materials can be produced by the same melt-spin process if the spin rate or quench rate is decreased to produce a fine-grained crystalline phase, which may be of metastable composition. If the grain size is arranged to match the optimum size for single domains, the coercivity can be

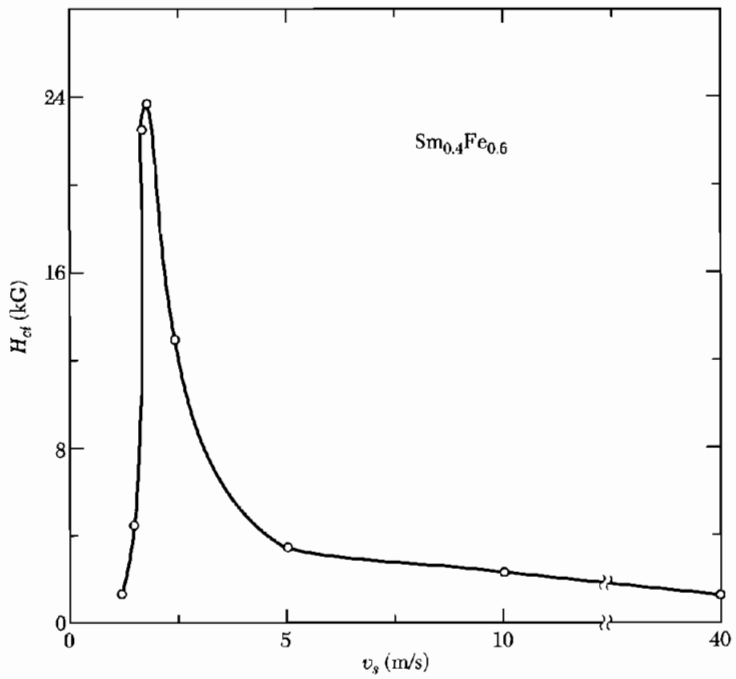


Figure 6 Coercivity at room temperature vs. melt-spin velocity v_s for $Sm_{0.4}Fe_{0.6}$. The maximum coercivity is 24 kG and occurs at 1.65 m s^{-1}, which is believed to correspond to single domain behavior in each crystallite. At higher spin rates the coercivity decreases because the deposited material becomes amorphous (more isotropic). At lower spin rates the crystallites anneal to sizes above the single domain regime; domain boundaries give a lower coercivity. (After J. L. Croat.)

quite high (Fig. 6). J. L. Croat has reported $H_{ci} = 7.5$ kG for the metastable alloy $Nd_{0.4}Fe_{0.6}$ at the optimum melt-spin velocity 5 m s^{-1}.

AMORPHOUS SEMICONDUCTORS

Amorphous semiconductors can be prepared as thin films by evaporation or sputtering, or in some materials as bulk glasses by supercooling the melt.

What happens to the electron energy band model in a solid without regular crystalline order? The Bloch theorem is not applicable when the structure is not periodic, so that the electron states cannot be described by well-defined **k** values. Thus, the momentum selection rule for optical transitions is relaxed; hence all infrared and Raman modes contribute to the absorption spectra. The optical absorption edge is rather featureless. Allowed bands and energy gaps still occur because the form of the density of states vs. energy is determined most strongly by local electron bonding configurations.

Both electrons and holes can carry current in an amorphous semiconductor. The carriers may be scattered strongly by the disordered structure, so that the mean free path may sometimes be of the order of the scale of the disorder. Anderson proposed that the states near band edges may be localized and do not extend through the solid (Fig. 7). Conduction in these states may take place by a thermally-assisted hopping process, for which the Hall effect is anomalous and cannot be used to determine the carrier concentration.

Two distinct classes of amorphous semiconductors are widely studied: tetrahedrally-bonded amorphous solids such as silicon and germanium, and the chalcogenide glasses. The latter are multicomponent solids of which one major constituent is a "chalcogen" element—sulfur, selenium, or tellurium.

The tetrahedrally-bonded materials have properties similar to those of their crystalline forms, provided the dangling-bond defects are compensated with hydrogen. They can be doped with small amounts of chemical impurities, and their conductivity can be sharply modified by injection of free carriers from a metallic contact. By contrast, the chalcogenide glasses are largely insensitive to chemical impurities and to free carrier injection.

Amorphous hydrogenated silicon is a candidate material for solar cells. Amorphous silicon is a much less expensive material than single crystal silicon.

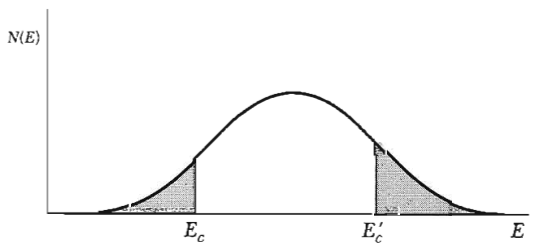


Figure 7 Density of electron states as believed to occur in amorphous solids, when states are non-localized in the center of the band. Localized states are shown shaded. The mobility band edges E_c, E_c' separate the ranges of energy where states are localized and non-localized. (After N. Mott and E. A. Davis.)

Attempts at using pure amorphous silicon, however, failed because of structural defects (dangling bonds) which were impossible to eliminate. Introduction of hydrogen into amorphous silicon appears to remove the undesirable structural defects. Relatively large proportions of hydrogen are incorporated, of the order of 10 percent or more.

LOW ENERGY EXCITATIONS IN AMORPHOUS SOLIDS

The low temperature heat capacity of pure dielectric crystalline solids is known (Chapter 5) to follow the Debye T^3 law, precisely as expected from the excitation of long wavelength phonons. The same behavior was expected in glasses and other amorphous solids. However, many insulating glasses show an unexpected linear term in the heat capacity below 1 K. Indeed, at 25 mK the observed heat capacity of vitreous silica exceeds the Debye phonon contribution by a factor of 1000. Anomalous linear terms of comparable magnitude are found in all, or nearly all, amorphous solids. Their presence is believed to be an intrinsic consequence of the amorphous states of matter, but the details of why this is so remain unclear. There is strong evidence that the anomalous properties arise from two-level systems and not from multi-level oscillator systems; in brief, the evidence is that the systems can be saturated by intense phonon fields, just as a two-level spin system can be saturated by an intense rf magnetic field.

Heat Capacity Calculation

Consider an amorphous solid with a concentration N of two-level systems at low energies; that is, with a level splitting Δ much less than the phonon Debye cutoff $k_B\theta$. The partition function of one system is, with $\tau = k_BT$,

$$Z = \exp(\Delta/2\tau) + \exp(-\Delta/2\tau) = 2\cosh(\Delta/2\tau) \ . \tag{20}$$

The thermal average energy is

$$U = -\tfrac{1}{2}\Delta \tanh(\Delta/2\tau) \ , \tag{21}$$

and the heat capacity of the single system is

$$C_V = k_B(\partial U/\partial\tau) = k_B(\Delta/2\tau)^2 \operatorname{sech}^2(\Delta/2\tau) \ . \tag{22}$$

These results are given in detail in *TP*, pp. 62–63.

Now suppose that Δ is distributed with uniform probability in the range $\Delta = 0$ to $\Delta = \Delta_0$. The average value of C_V is

$$\begin{aligned} C_V &= (k_B/4\tau^2)\int_0^{\Delta_0} d\Delta\ (\Delta^2/\Delta_0)\operatorname{sech}^2(\Delta/2\tau) \\ &= \left(\frac{2k_B\tau}{\Delta_0}\right)\int_0^{\Delta_0/2\tau} dx\, x^2 \operatorname{sech}^2 x \ . \end{aligned} \tag{23}$$

The integral cannot be evaluated in closed form.

Two limits are of special interest. For $\tau \ll \Delta_0$, the $\operatorname{sech}^2 x$ term is roughly 1 from $x = 0$ to $x = 1$, and roughly zero for $x > 1$. The value of the integral is roughly 1/3, whence

$$C_V \approx 2k_B^2 T/3\Delta_0 \ , \tag{24}$$

for $T < \Delta_0/k_B$.

For $\tau \gg \Delta_0$, the value of the integral is roughly $\frac{1}{3}(\Delta_0/2k_B T)^3$, so that in this limit

$$C_V \approx \Delta_0^2/12k_B T^2 \ , \tag{25}$$

which approaches zero as T increases.

Thus the interesting region is at low temperatures, for here by (24) the two-level system contributes to the heat capacity a term linear in the temperature. This term, originally introduced for dilute magnetic impurities in metals, has no connection with the usual conduction electron heat capacity which is also proportional to T.

The empirical result appears to be that all disordered solids have $N \sim 10^{17}$ cm^{-3} "new type" low energy excitations uniformly distributed in the energy interval from 0 to 1 K. The anomalous specific heat can now be obtained from (24). For $T = 0.1$ K and $\Delta_0/k_B = 1$ K,

$$C_V \approx \tfrac{2}{3} N k_B (0.1) \approx 1 \text{ erg cm}^{-3} \text{ K}^{-1} \ . \tag{26}$$

For comparison, the phonon contribution at 0.1 K is, from (5.35),

$$\begin{aligned} C_V &\approx 234 N k_B (T/\theta)^3 \approx (234)(2.3 \times 10^{22})(1.38 \times 10^{-16})(0.1/300)^3 \\ &\approx 2.8 \times 10^{-2} \text{ erg cm}^{-3} \text{ K}^{-1} \ , \end{aligned} \tag{27}$$

much smaller than (26).

The experimental results (Fig. 8) for vitreous SiO_2 are represented by

$$C_V = c_1 T + c_3 T^3 \ , \tag{28}$$

where $c_1 = 12$ erg g^{-1} K^{-2} and $c_3 = 18$ erg g^{-1} K^{-4}.

Thermal Conductivity

The thermal conductivity of glasses is very low. It is limited at room temperature and above by the scale of the disorder of the structure, for this scale determines the mean free path of the dominant thermal phonons. At low temperatures, below 1 K, the conductivity is carried by long wavelength phonons and is limited by phonon scattering from the mysterious two-level systems or tunneling states discussed earlier for their contribution to the heat capacity of amorphous solids.

As in Chapter 5, the expression for the thermal conductivity K has the form

$$K = \tfrac{1}{3} c v \ell \ , \tag{29}$$

where c is the heat capacity per unit volume, v is an average phonon velocity, and ℓ is the phonon mean free path. For vitreous silica at room temperature,

$$K \cong 1.4 \times 10^{-2} \text{ J cm}^{-1} \text{ s}^{-1} \text{ K}^{-1} \ ;$$

$$c \cong 1.6 \text{ J cm}^{-3} \text{ K}^{-1} \ ;$$

$$\langle v \rangle \cong 4.2 \times 10^{5} \text{ cm s}^{-1} \ .$$

Thus the mean free path $\ell \cong 6 \times 10^{-8}$ cm; by reference to Fig. 3 we see that this is of the order of magnitude of the disorder of the structure.

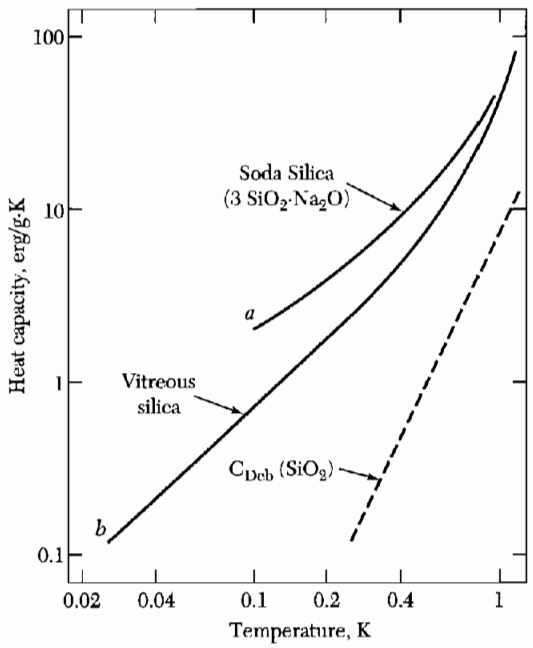


Figure 8 Heat capacity of vitreous silica and soda silica glass as a function of temperature. The heat capacity is roughly linear in T below 1 K. The dashed line represents the calculated Debye heat capacity of vitreous silica.

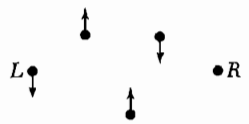


Figure 9 Short phonon mean free path in a disordered structure. A short wavelength phonon that displaces atom L, as shown, will displace atom R by a much smaller distance, because of the phase cancellation of the upper and lower paths from L to R. The displacement of R is $\uparrow + \downarrow \sim 0$, so that the wave incident from L is reflected at R.

This value of the phonon mean free path is remarkably small. At room temperature and above (that is, above the Debye temperature), most of the phonons have half-wavelengths of the order of the interatomic spacing. It is through phase cancellation processes, as in Fig. 9, that the mean free path is limited to several interatomic spacings. No other structure for fused quartz will give a 6 Å mean free path. The normal modes of vibration of the glass structure are utterly unlike plane waves. But the modes, as distorted as they are, still have quantized amplitudes and therefore may be called phonons.

FIBER OPTICS

Fibers of silica-based lightguides carry a high proportion of the data and information transmitted on the surface of the earth and under the seas. The optical fibers consist of a thin core ($\approx$10 μm) of high-refractive index glass surrounded by a cladding. The digital data are carried by light, with a minimum attenuation near 0.20 db km^{-1} at wavelengths near 1.55 μm, which is in the infrared (Fig. 10). A range of 100 km corresponds to a loss of 20 db, power readily supplied by an Eu^{3+} laser amplifier.

The optic window of high-purity glasses near this wavelength is limited on the low frequency side by phonon absorption bands and on the high frequency side by Rayleigh scattering, and, ultimately, by electronic absorption. In the

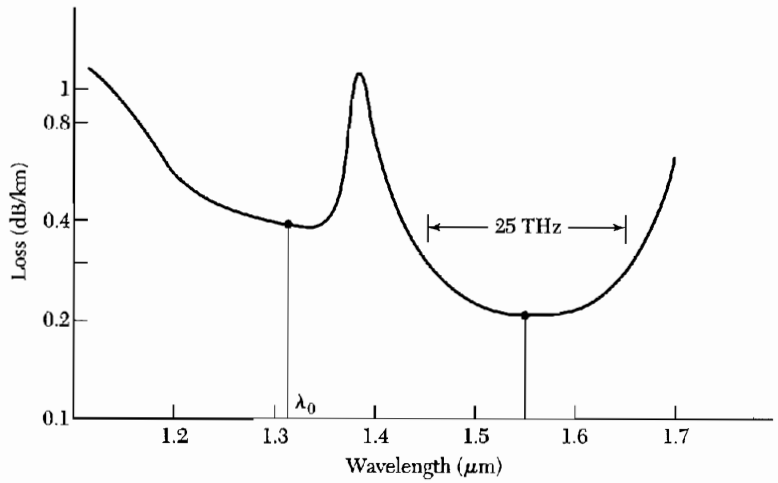


Figure 10 The transmission characteristic of communication-quality optical fibers, showing the attenuation in units of decibels per km as a function of the wavelength of light, in μm. The Rayleigh scattering regime is dominant on the left of the curve, except for a strong impurity absorption line associated with OH ions that accompany SiO_2; the line is the second harmonic of a line at 2.7 μm known as the "water line." The wavelength marked at 1.31 μm is used in 1994 transmission lines; it was replaced by the wavelength 1.55 μm available from Eu^{3+} ion amplifiers, which are used every 100 km in typical long distance applications. The power needed to pump the amplifiers is supplied by copper wires. (Courtesy of Tingye Li, AT&T Bell Laboratories.)

optic window the losses are determined by the Rayleigh scattering intrinsic to static fluctuations in the local dielectric constant of an inhomogeneous medium, and the attenuation varies as the fourth power of the frequency.

It is fortunate that an excellent source is available for radiation at 1.55 μm. As shown in Fig. 13.24, excited (pumped) erbium Er^{3+} ions can amplify in an erbium-doped section of fiber.

Rayleigh Attenuation

The attenuation of light waves in glass is dominated at wavelengths in the infrared by the same scattering process, called Rayleigh scattering, that is responsible for the blue light of the sky. The extinction coefficient, or attenuation coefficient, h, has the dimension of reciprocal length and for light scattered in a gas is given, after Rayleigh, by

$$h = (2\omega^4/3\pi c^4 N)\langle (n-1)^2\rangle \ , \tag{30}$$

where n is the local refractive index and N is the number of scattering centers per unit volume. The energy flux as a function of distance has the form $\exp(-hx)$.

Derivations of (30) are found in good texts on electrodynamics; the structure of the result may be understood by a general argument: The radiant energy scattered from a dipole element p is proportional to $(dp^2/dt^2)^2$, and this accounts for the factor ω^4. The local polarizability α enters as α^2; if there are N random scattering centers per unit volume, the scattered energy averaged over these random sources will go as $N\langle(\Delta\alpha)^2\rangle$, or $\langle(\Delta n)^2\rangle/N$. Thus we have the essential factors that appear in (30). As applied to a glass, Δn should refer to the variations in polarization around each group of Si—O bonds, and satisfactory numerical estimates of the attenuation may be made in this way.

Problem

1. ***Metallic optic fibres?*** It has been speculated that metallic wires can act as optic fibres, transmitting light with a long delay appropriate to the high refractive index characteristic of metals. Unfortunately the refractive index of a typical metal is dominated by a free-electron term in $i^{1/2}$, so that the propagation of a light wave is in fact highly damped in a metal. Show that in sodium at room temperature a wave of vacuum wavelength 10 μm will have a damping length of 0.1 μm. This may be contrasted with the 100 km damping length found for light in high-quality glass fibres.

20

Point Defects

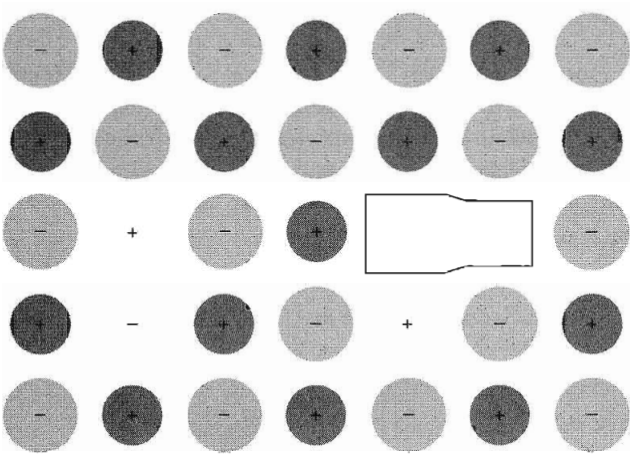


Figure 1 A plane of a pure alkali halide crystal, showing two vacant positive ion sites, and a coupled pair of vacant sites of opposite sign.

The common point imperfections in crystals are chemical impurities, vacant lattice sites, and extra atoms not in regular lattice positions. Linear imperfections are treated under dislocations, Chapter 21. The crystal surface is a planar imperfection, with electron, phonon, and magnon surface states.

Some important properties of crystals are controlled as much by imperfections as by the composition of the host crystal, which may act only as a solvent or matrix or vehicle for the imperfections. The conductivity of some semiconductors is due entirely to trace amounts of chemical impurities. The color and luminescence of many crystals arise from impurities or imperfections. Atomic diffusion may be accelerated enormously by impurities or imperfections. Mechanical and plastic properties are usually controlled by imperfections.

LATTICE VACANCIES

The simplest imperfection is a **lattice vacancy**, which is a missing atom or ion, also known as a **Schottky defect**. (Fig. 1). We create a Schottky defect in a perfect crystal by transferring an atom from a lattice site in the interior to a lattice site on the surface of the crystal. In thermal equilibrium a certain number of lattice vacancies are always present in an otherwise perfect crystal, because the entropy is increased by the presence of disorder in the structure.

In metals with close-packed structures the proportion of lattice sites vacant at temperatures just below the melting point is of the order of 10^{-3} to 10^{-4}. But in some alloys, in particular the very hard transition metal carbides such as TiC, the proportion of vacant sites of one component can be as high as 50 percent.

The probability that a given site is vacant is proportional to the Boltzmann factor for thermal equilibrium: $P = \exp(-E_V/k_BT)$, where E_V is the energy required to take an atom from a lattice site inside the crystal to a lattice site on the surface. If there are N atoms, the equilibrium number n of vacancies is given by the Boltzmann factor

$$\frac{n}{N-n} = \exp(-E_V/k_BT) \ . \tag{1}$$

If $n \ll N$, then

$$n/N \cong \exp(-E_V/k_BT) \ . \tag{2}$$

If $E_V \approx 1$ eV and $T \approx 1000$ K, then $n/N \approx e^{-12} \approx 10^{-5}$.

The equilibrium concentration of vacancies decreases as the temperature decreases. The actual concentration of vacancies will be higher than the

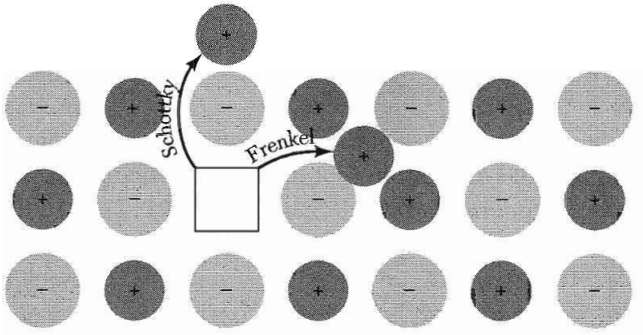


Figure 2 Schottky and Frenkel defects in an ionic crystal. The arrows indicate the displacement of the ions. In a Schottky defect the ion ends up on the surface of the crystal; in a Frenkel defect the ion is removed to an interstitial position.

equilibrium value if the crystal is grown at an elevated temperature and then cooled suddenly, thereby freezing in the vacancies (see the discussion of diffusion below).

In ionic crystals it is usually energetically favorable to form roughly equal numbers of positive and negative ion vacancies. The formation of pairs of vacancies keeps the crystal electrostatically neutral on a local scale. From a statistical calculation we obtain

$$n \cong N \exp(-E_p/2k_BT) \tag{3}$$

for the number of pairs, where E_p is the energy of formation of a pair.

Another vacancy defect is the **Frenkel defect** (Fig. 2) in which an atom is transferred from a lattice site to an **interstitial position**, a position not normally occupied by an atom. In pure alkali halides the most common lattice vacancies are Schottky defects; in pure silver halides the most common vacancies are Frenkel defects. The calculation of the equilibrium number of Frenkel defects proceeds along the lines of Problem 1. If the number n of Frenkel defects is much smaller than the number of lattice sites N and the number of interstitial sites N', the result is

$$n \cong (NN')^{1/2} \exp(-E_I/2k_BT) \ , \tag{4}$$

where E_I is the energy necessary to remove an atom from a lattice site to an interstitial position.

Lattice vacancies are present in alkali halides when these contain additions of divalent elements. If a crystal of KCl is grown with controlled amounts of $CaCl_2$, the density varies as if a K^+ lattice vacancy were formed for each Ca^{2+} ion in the crystal. The Ca^{2+} enters the lattice in a normal K^+ site and the two Cl^- ions enter two Cl^- sites in the KCl crystal (Fig. 3). Demands of charge neutrality result in a vacant metal ion site. The experimental results show that the addition of $CaCl_2$ to KCl lowers the density of the crystal. The density

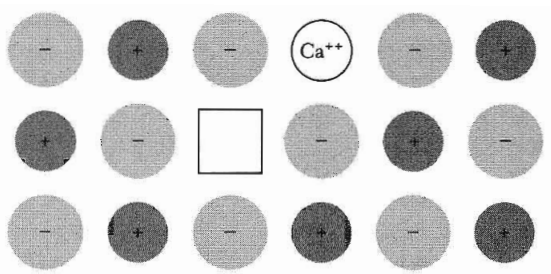


Figure 3 Production of a lattice vacancy by the solution of $CaCl_2$ in KCl: to ensure electrical neutrality, a positive ion vacancy is introduced into the lattice with each divalent cation Ca^{++}. The two Cl^- ions of $CaCl_2$ enter normal negative ion sites.

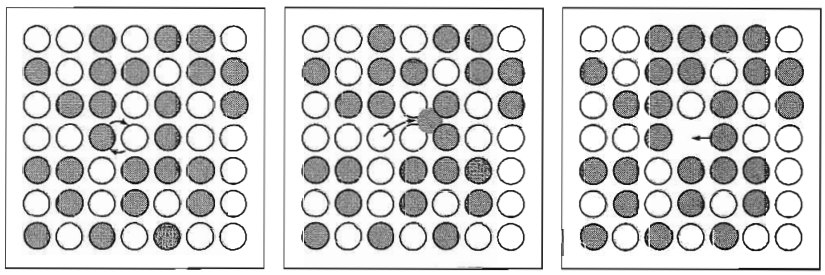


Figure 4 Three basic mechanisms of diffusion: (a) Interchange by rotation about a midpoint. More than two atoms may rotate together. (b) Migration through interstitial sites. (c) Atoms exchange position with vacant lattice sites. (After Seitz.)

would increase if no vacancies were produced, because Ca^{2+} is a heavier and smaller ion than K^+.

The mechanism of electrical conductivity in alkali and silver halide crystals is usually by the motion of ions and not by the motion of electrons. This has been established by comparing the transport of charge with the transport of mass as measured by the material plated out on electrodes in contact with the crystal.

The study of ionic conductivity is an important tool in the investigation of lattice defects. Work on alkali and silver halides containing known additions of divalent metal ions shows that at not too high temperatures the ionic conductivity is directly proportional to the amount of divalent addition. This is not because the divalent ions are intrinsically highly mobile, for it is predominantly the monovalent metal ion which deposits at the cathode. The lattice vacancies introduced with the divalent ions are responsible for the enhanced diffusion (Fig. 4c). The diffusion of a vacancy in one direction is equivalent to the diffusion of an atom in the opposite direction. When lattice defects are generated thermally, their energy of formation gives an extra contribution to the heat capacity of the crystal, as shown in Fig. 5.

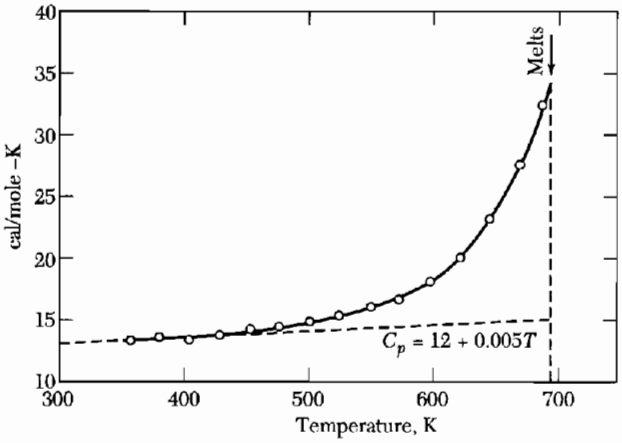


Figure 5 Heat capacity of silver bromide exhibiting the excess heat capacity from the formation of lattice defects. (After R. W. Christy and A. W. Lawson.)

An associated pair of vacancies of opposite sign exhibits an electric dipole moment, with contributions to the dielectric constant and dielectric loss due to the motion of pairs of vacancies. The dielectric relaxation time is a measure of the time required for one of the vacant sites to jump by one atomic position with respect to the other. The dipole moment can change at low frequencies, but not at high. In sodium chloride the relaxation frequency is 1000 s^{-1} at 85°C.

DIFFUSION

When there is a concentration gradient of impurity atoms or vacancies in a solid, there will be a flux of these through the solid. In equilibrium the impurities or vacancies will be distributed uniformly. The net flux J_N of atoms of one species in a solid is related to the gradient of the concentration N of this species by a phenomenological relation called **Fick's law**:

$$\boldsymbol{J}_N = -D \operatorname{grad} N \ . \tag{5}$$

Here J_N is the number of atoms crossing unit area in unit time; the constant D is the **diffusion constant** or **diffusivity** and has the units cm^2/s or m^2/s. The minus sign means that diffusion occurs away from regions of high concentration. The form (5) of the law of diffusion is often adequate, but rigorously the gradient of the chemical potential is the driving force for diffusion and not the concentration gradient alone (*TP*, p. 406).

The diffusion constant is often found to vary with temperature as

$$D = D_0 \exp(-E/k_B T) \ ; \tag{6}$$

here E is the **activation energy** for the process. Experimental results on the diffusion of carbon in alpha iron are shown in Fig. 6. The data are represented

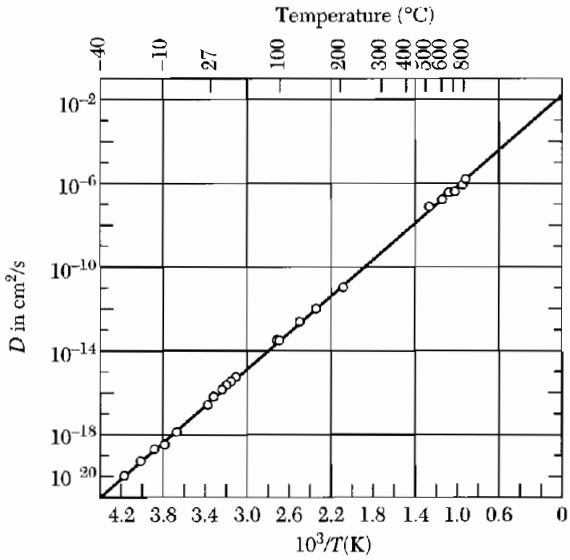


Figure 6 Diffusion coefficient of carbon in iron, after Wert. The logarithm of D is directly proportional to $1/T$.

Table 1 Diffusion constants and activation energies

Host crystal	Atom	D_0 $cm^2 s^{-1}$	E eV	Host crystal	Atom	D_0 $cm^2 s^{-1}$	E eV
Cu	Cu	0.20	2.04	Si	Al	8.0	3.47
Cu	Zn	0.34	1.98	Si	Ga	3.6	3.51
Ag	Ag	0.40	1.91	Si	In	16.0	3.90
Ag	Cu	1.2	2.00	Si	As	0.32	3.56
Ag	Au	0.26	1.98	Si	Sb	5.6	3.94
Ag	Pb	0.22	1.65	Si	Li	2×10^{-3}	0.66
Na	Na	0.24	0.45	Si	Au	1×10^{-3}	1.13
U	U	2×10^{-3}	1.20	Ge	Ge	10.0	3.1

by $E = 0.87$ eV, $D_0 = 0.020$ cm^2/s. Representative values of D_0 and E are given in Table 1.

To diffuse, an atom must overcome the potential energy barrier presented by its nearest neighbors. We treat the diffusion of impurity atoms between interstitial sites. The same argument will apply to the diffusion of vacant lattice sites. If the barrier is of height E, the atom will have sufficient thermal energy to pass over the barrier a fraction $\exp(-E/k_BT)$ of the time. Quantum tunneling through the barrier is another possible process, but is usually important only for the lightest nuclei, particularly hydrogen.

If ν is a characteristic atomic vibrational frequency, then the probability p that during unit time the atom will have enough thermal energy to pass over the barrier is

$$p \approx \nu \exp(-E/k_B T) \ . \tag{7}$$

In unit time the atom makes ν passes at the barrier, with a probability $\exp(-E/k_B T)$ of surmounting the barrier on each try. The quantity p is called the *jump frequency*.

We consider two parallel planes of impurity atoms in interstitial sites. The planes are separated by lattice constant a. There are S impurity atoms on one plane and $(S + a\, dS/dx)$ on the other. The net number of atoms crossing between the planes in unit time is $\approx -pa\, dS/dx$. If N is the total concentration of impurity atoms, then $S = aN$ per unit area of a plane.

The diffusion flux may now be written as

$$J_N \approx -pa^2(dN/dx) \ . \tag{8}$$

On comparison with (5) we have the result

$$\boxed{D = \nu a^2 \exp(-E/k_B T) \ .} \tag{9}$$

of the form (6) with $D_0 = \nu a^2$.

If the impurities are charged, we may find the ionic mobility $\bar{\mu}$ and the conductivity σ from the diffusivity by using the Einstein relation $k_B T \bar{\mu} = qD$ from *TP*, p. 406:

$$\bar{\mu} = (q\nu a^2/k_B T) \exp(-E/k_B T) \ ; \tag{10}$$

$$\sigma = Nq\bar{\mu} = (Nq^2 \nu a^2/k_B T) \exp(-E/k_B T) \ , \tag{11}$$

where N is the concentration of impurity ions of charge q.

The proportion of vacancies is independent of temperature in the range in which the number of vacancies is determined by the number of divalent metal ions. Then the slope of a plot of $\ln \sigma$ versus $1/k_B T$ gives E_+, the barrier activation energy for the jumping of positive ion vacancies (Table 2). Diffusion is very slow at low temperatures. At room temperature the jump frequency is of the order of $1\,\mathrm{s}^{-1}$, and at 100 K it is of the order of $10^{-25}\,\mathrm{s}^{-1}$.

The proportion of vacancies in the temperature range in which the concentration of defects is determined by thermal generation is given by

$$f \cong \exp(-E_f/2k_B T) \ , \tag{12}$$

where E_f is the energy of formation of a vacancy pair, according to the theory of Schottky or Frenkel defects. Here the slope of a plot of $\ln \sigma$ versus $1/k_B T$ will be $E_+ + \frac{1}{2}E_f$, according to (10) and (12). From measurements in different

Table 2 Activation energy E_+ for motion of a positive ion vacancy

Values of the energy of formation of a vacancy pair, E_f, are also given. The numbers given in parentheses for the silver salts refer to interstitial silver ions.

Crystal	E_+(eV)	E_f(eV)	Workers
NaCl	0.86	2.02	Etzel and Maurer
LiF	0.65	2.68	Haven
LiCl	0.41	2.12	Haven
LiBr	0.31	1.80	Haven
LiI	0.38	1.34	Haven
KCl	0.89	2.1–2.4	Wagner; Kelting and Witt
AgCl	0.39(0.10)	1.4[a]	Teltow
AgBr	0.25(0.11)	1.1[a]	Compton

[a]For Frenkel defect.

temperature ranges we determine the energy of formation of a vacancy pair E_f and the jump activation energy E_+.

The diffusion constant can be measured by radioactive tracer techniques. The diffusion of a known initial distribution of radioactive ions is followed as a function of time or distance. Values of the diffusion constant thus determined may be compared with values from ionic conductivities. The two sets of values do not usually agree within the experimental accuracy, suggesting the presence of a diffusion mechanism that does not involve the transport of charge. For example, the diffusion of pairs of positive and negative ion vacancies does not involve the transport of charge.

Metals

Self-diffusion in monatomic metals most commonly proceeds by lattice vacancies. **Self-diffusion** means the diffusion of atoms of the metal itself, and not of impurities. The activation energy for self-diffusion in copper is expected to be in the range 2.4 to 2.7 eV for diffusion through vacancies and 5.1 to 6.4 eV for diffusion through interstitial sites. Observed values of the activation energy are 1.7 to 2.1 eV.

Activation energies for diffusion in Li and Na can be determined from measurements of the temperature dependence of the nuclear resonance line width. As discussed under magnetic resonance, the resonance line width narrows when the jump frequency of an atom between sites becomes rapid in comparison with the frequency corresponding to the static line width. The values 0.57 eV and 0.45 eV were determined by NMR for Li and Na. Self-diffusion measurements for sodium also give 0.4 eV.

COLOR CENTERS

Pure alkali halide crystals are transparent throughout the visible region of the spectrum. A **color center** is a lattice defect that absorbs visible light. An ordinary lattice vacancy does not color alkali halide crystals, although it affects the absorption in the ultraviolet. The crystals may be colored in a number of ways:

- by the introduction of chemical impurities;
- by the introduction of an excess of the metal ion (we may heat the crystal in the vapor of the alkali metal and then cool it quickly—an NaCl crystal heated in the presence of sodium vapor becomes yellow; a KCl crystal heated in potassium vapor becomes magenta);
- by x-ray, γ-ray, neutron, and electron bombardment; and
- by electrolysis.

F Centers

The name *F* center comes from the German word for color, *Farbe*. We usually produce *F* centers by heating the crystal in excess alkali vapor or by x-irradiation. The central absorption band (*F* band) associated with *F* centers in several alkali halides are shown in Fig. 7, and the quantum energies are listed in Table 3. Experimental properties of *F* centers have been investigated in detail, originally by Pohl.

The *F* center has been identified by electron spin resonance as an electron bound at a negative ion vacancy (Fig. 8), in agreement with a model suggested by de Boer. When excess alkali atoms are added to an alkali halide crystal, a corresponding number of negative ion vacancies are created. The valence electron of the alkali atom is not bound to the atom; the electron migrates in the crystal and becomes bound to a vacant negative ion site. A negative ion vacancy in a perfect periodic lattice has the effect of an isolated positive charge: it attracts and binds an electron. We can simulate the electrostatic effect of

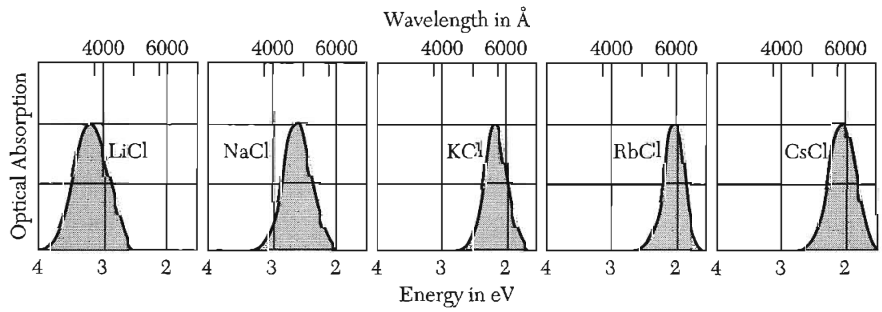


Figure 7 The *F* bands for several alkali halides: optical absorption versus wavelength for crystals that contain *F* centers.

Table 3 Experimental F center absorption energies, in eV

LiCl	3.1	NaBr	2.3
NaCl	2.7	KBr	2.0
KCl	2.2	RbBr	1.8
RbCl	2.0	LiF	5.0
CsCl	2.0	NaF	3.6
LiBr	2.7	KF	2.7

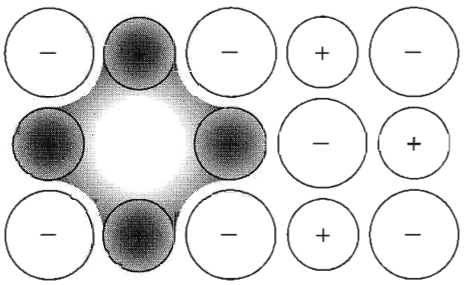


Figure 8 An F center is a negative ion vacancy with one excess electron bound at the vacancy. The distribution of the excess electron is largely on the positive metal ions adjacent to the vacant lattice site.

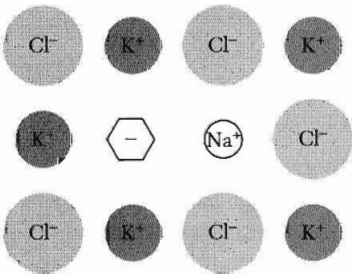


Figure 9 An F_A center in KCl: one of the six K^+ ions which bind an F center is replaced by another alkali ion, here Na^+.

a negative ion vacancy by adding a positive charge q to the normal charge $-q$ of an occupied negative ion site.

The F center is the simplest trapped-electron center in alkali halide crystals. The optical absorption of an F center arises from an electric dipole transition to a bound excited state of the center.

Other Centers in Alkali Halides

In the F_A center one of the six nearest neighbors of an F center has been replaced by a different alkali ion, Fig. 9. More complex trapped-electron centers are formed by groups of F centers, Fig. 10 and 11. Thus two adjacent F centers form an M center. Three adjacent F centers form an R center. Different centers are distinguished by their optical absorption frequencies.

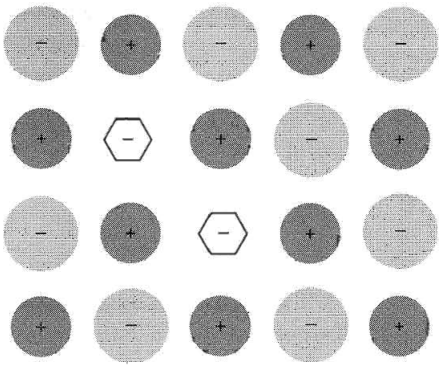


Figure 10 An M center consists of two adjacent F centers.

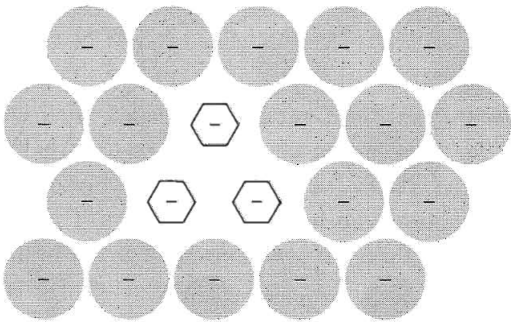


Figure 11 An R center consists of three adjacent F centers; that is, a group of three negative ion vacancies in a [111] plane of the NaCl structure, with three associated electrons.

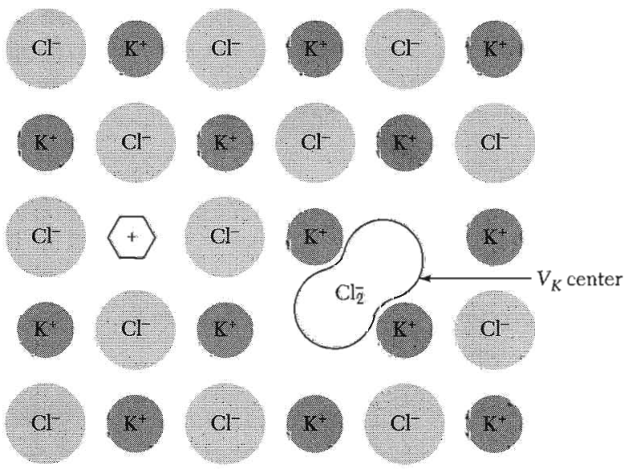


Figure 12 A V_K center formed when a hole is trapped by a pair of negative ions resembles a negative halogen molecule ion, which is Cl_2^- in KCl. No lattice vacancies or extra atoms are involved in a V_K center. The center at the left of the figure probably is not stable: the hexagon represents a hole trapped near a positive ion vacancy; such a center would be the antimorph to an F center. Holes have a lower energy trapped in a V_K center than in an anti-F center.

Holes may be trapped to form color centers, but hole centers are not usually as simple as electron centers. For example, a hole in the filled p^6 shell of a halogen ion leaves the ion in a p^5 configuration, whereas an electron added to the filled p^6 shell of an alkali ion leaves the ion in a p^6s configuration.

The chemistry of the two centers is different: p^6s acts as a spherically symmetric ion, but p^5 acts as an asymmetric ion and, by virtue of the Jahn-Teller effect, will distort its immediate surroundings in the crystal.

The antimorph to the F center is a hole trapped at a positive ion vacancy, but no such center has been identified experimentally in alkali halides; in insulating oxides the O^- (called V^-) defect is known. The best-known trapped-hole center is the V_K center, Fig. 12. The V_K center is formed when a hole is trapped by a halogen ion in an alkali halide crystal. Electron spin resonance shows that the center is like a negative halogen molecular ion, such as Cl_2^- in KCl. The Jahn-Teller trapping of free holes is the most effective form of self-trapping of charge carriers in perfect crystals.

Problems

1. ***Frenkel defects.*** Show that the number n of interstitial atoms in equilibrium with n lattice vacancies in a crystal having N lattice points and N' possible interstitial positions is given by the equation

$$E_I = k_B T \ln [(N - n)(N' - n)/n^2] ,$$

whence, for $n \ll N, N'$, we have $n \simeq (NN')^{1/2} \exp(-E_I/2k_BT)$. Here E_I is the energy necessary to remove an atom from a lattice site to an interstitial position.

2. ***Schottky vacancies.*** Suppose that the energy required to remove a sodium atom from the inside of a sodium crystal to the boundary is 1 eV. Calculate the concentration of Schottky vacancies at 300 K.

3. ***F center.*** (a) Treat an F center as a free electron of mass m moving in the field of a point charge e in a medium of dielectric constant $\epsilon = n^2$; what is the $1s$-$2p$ energy difference of F centers in NaCl? (b) Compare from Table 3 the F center excitation energy in NaCl with the $3s$-$3p$ energy difference of the free sodium atom.

21

Dislocations

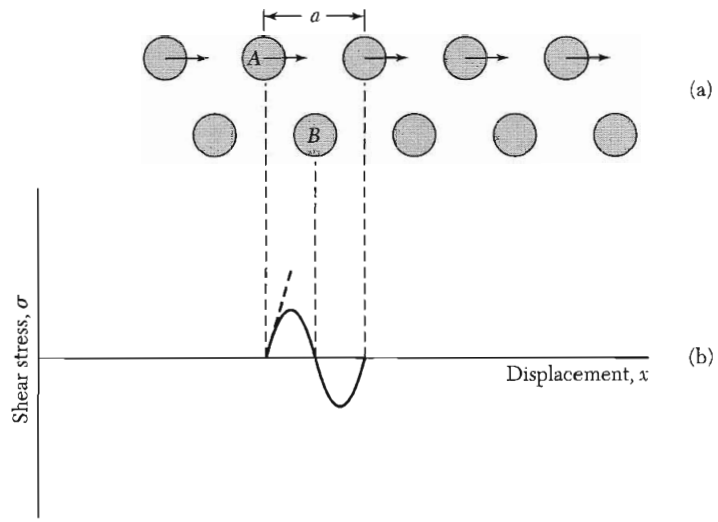


Figure 1 (a) Relative shear of two planes of atoms (shown in section) in a uniformly strained crystal; (b) shear stress as a function of the relative displacement of the planes from their equilibrium position. The heavy broken line drawn at the initial slope defines the shear modulus G.

CHAPTER 21: DISLOCATIONS

This chapter is concerned with the interpretation of the plastic mechanical properties of crystalline solids in terms of the theory of dislocations. Plastic properties are irreversible deformations; elastic properties are reversible. The ease with which pure single crystals deform plastically is striking. This intrinsic weakness of crystals is exhibited in various ways. Pure silver chloride melts at 455°C, yet at room temperature it has a cheeselike consistency and can be rolled into sheets. Pure aluminum crystals are elastic (follow Hooke's law) only to a strain of about 10^{-5}, after which they deform plastically.

Theoretical estimates of the strain at the elastic limit of perfect crystals may give values 10^3 or 10^4 higher than the lowest observed values, although a factor 10^2 is more usual. There are few exceptions to the rule that pure crystals are plastic and are not strong: crystals of germanium and silicon are not plastic at room temperature and fail or yield only by fracture. Glass at room temperature fails only by fracture, but it is not crystalline. The fracture of glass is caused by stress concentration at minute cracks.

SHEAR STRENGTH OF SINGLE CRYSTALS

Frenkel gave a simple method of estimating the theoretical shear strength of a perfect crystal. We consider in Fig. 1 the force needed to make a shear displacement of two planes of atoms past each other. For small elastic strains, the stress σ is related to the displacement x by

$$\sigma = Gx/d \ . \tag{1}$$

Here d is the interplanar spacing, and G denotes the appropriate shear modulus. When the displacement is large and has proceeded to the point that atom A is directly over atom B in the figure, the two planes of atoms are in a configuration of unstable equilibrium and the stress is zero. As a first approximation we represent the stress-displacement relation by

$$\sigma = (Ga/2\pi d) \sin (2\pi x/a) \ , \tag{2}$$

where a is the interatomic spacing in the direction of shear. This relation is constructed to reduce to (1) for small values of x/a. The critical shear stress σ_c at which the lattice becomes unstable is given by the maximum value of σ, or

$$\sigma_c = Ga/2\pi d \ . \tag{3}$$

If $a \approx d$, then $\sigma_c \approx G/2\pi$: the ideal critical shear stress is of the order of $\frac{1}{6}$ of the shear modulus.

Table 1 Comparison of shear modulus G and observed elastic limit σ_c[a]

	Shear modulus G, in dyn/cm^2	Elastic limit σ_c, in dyn/cm^2	G/σ_c
Sn, single crystal	1.9×10^{11}	1.3×10^{7}	15,000
Ag, single crystal	2.8×10^{11}	6×10^{6}	45,000
Al, single crystal	2.5×10^{11}	4×10^{6}	60,000
Al, pure, polycrystal	2.5×10^{11}	2.6×10^{8}	900
Al, commercial drawn	$\sim 2.5 \times 10^{11}$	9.9×10^{8}	250
Duralumin	$\sim 2.5 \times 10^{11}$	3.6×10^{9}	70
Fe, soft, polycrystal	7.7×10^{11}	1.5×10^{9}	500
Heat-treated carbon steel	$\sim 8 \times 10^{11}$	6.5×10^{9}	120
Nickel-chrome steel	$\sim 8 \times 10^{11}$	1.2×10^{10}	65

[a]After Mott.

The observations in Table 1 show the experimental values of the elastic limit are much smaller than (3) would suggest. The theoretical estimate may be improved by consideration of the actual form of the intermolecular forces and by consideration of other configurations of mechanical stability available to the lattice as it is sheared. Mackenzie has shown that these two effects may reduce the theoretical ideal shear strength to about $G/30$, corresponding to a critical shear strain angle of about 2 degrees. The observed low values of the shear strength can be explained only by the presence of imperfections that can act as sources of mechanical weakness in real crystals. The movement of crystal imperfections called dislocations is responsible for slip at very low applied stresses.

Slip

Plastic deformation in crystals occurs by slip, an example of which is shown in Fig. 2. In slip, one part of the crystal slides as a unit across an adjacent part. The surface on which slip takes place is known as the slip plane. The direction of motion is known as the slip direction. The great importance of lattice properties for plastic strain is indicated by the highly anisotropic nature of slip. Displacement takes place along crystallographic planes with a set of small Miller indices, such as the {111} planes in fcc metals and the {110}, {112}, and {123} planes in bcc metals.

The slip direction is in the line of closest atomic packing, ⟨110⟩ in fcc metals and ⟨111⟩ in bcc metals (Problem 1). To maintain the crystal structure after slip, the displacement or slip vector must equal a lattice translation vector. The shortest lattice translation vector expressed in terms of the lattice constant a in a fcc structure is of the form $(a/2)(\hat{\mathbf{x}} + \hat{\mathbf{y}})$; in a bcc structure it is $(a/2)(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$. But in fcc crystals one also observes partial displacements which upset the regular sequence $ABCABC \ldots$ of closest-packed planes, to

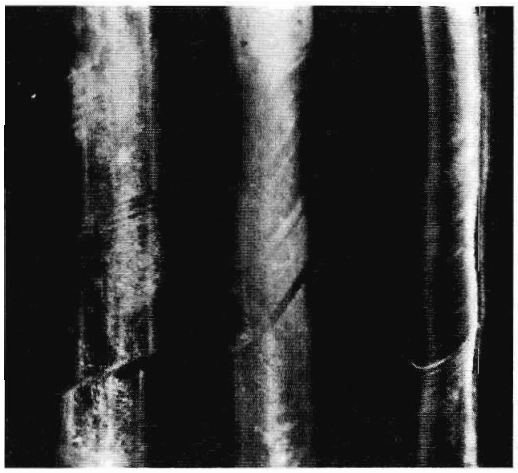


Figure 2 Translational slip in zinc single crystals. (E. R. Parker.)

produce a **stacking fault** such as *ABCABABC*. . . . The result is then a mixture of fcc and hcp stacking.

Deformation by slip is inhomogeneous: large shear displacements occur on a few widely separated slip planes, while parts of the crystal lying between slip planes remain essentially undeformed. A property of slip is the Schmid law of the critical shear stress: slip takes place along a given slip plane and direction when the corresponding *component* of shear stress reaches the critical value.

Slip is one mode of plastic deformation. Another mode, **twinning**, is observed particularly in hcp and bcc structures. During slip a considerable displacement occurs on a few widely separated slip planes. During twinning, a partial displacement occurs successively on each of many neighboring crystallographic planes. After twinning, the deformed part of the crystal is a mirror image of the undeformed part. Although both slip and twinning are caused by the motion of dislocations, we shall be concerned primarily with slip.

DISLOCATIONS

The low observed values of the critical shear stress are explained in terms of the motion through the lattice of a line imperfection known as a dislocation. The idea that slip propagates by the motion of dislocations was published in 1934 independently by Taylor, Orowan, and Polanyi; the concept of dislocations was introduced somewhat earlier by Prandtl and Dehlinger. There are several basic types of dislocations. We first describe an **edge dislocation**. Figure 3 shows a simple cubic crystal in which slip of one atom distance has occurred over the left half of the slip plane but not over the right half. The boundary between the slipped and unslipped regions is called the dislocation. Its position is marked by the termination of an extra vertical half-plane of atoms crowded into

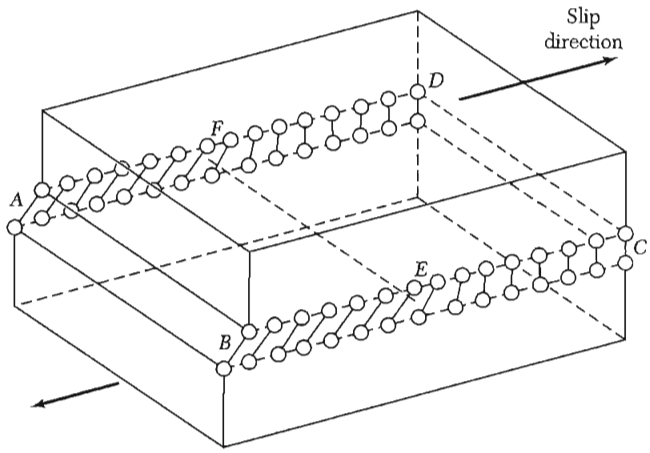


Figure 3 An edge dislocation *EF* in the glide plane *ABCD*. The figure shows the slipped region *ABEF* in which the atoms have been displaced by more than half a lattice constant and the unslipped region *FECD* with displacement less than half a lattice constant.

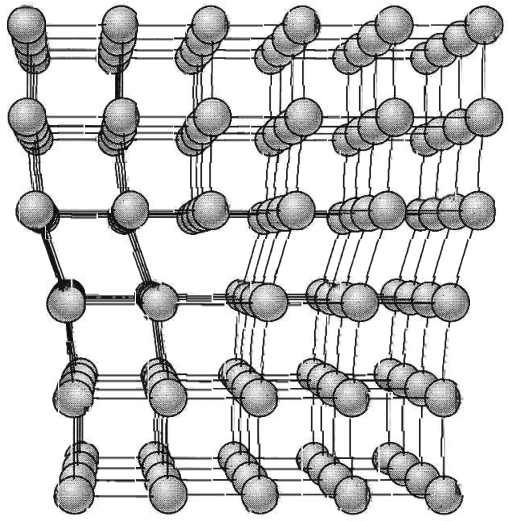


Figure 4 Structure of an edge dislocation. The deformation may be thought of as caused by inserting an extra plane of atoms on the upper half of the y axis. Atoms in the upper half-crystal are compressed by the insertion; those in the lower half are extended.

the upper half of the crystal as shown in Fig. 4. Near the dislocation the crystal is highly strained. The simple edge dislocation extends indefinitely in the slip plane in a direction normal to the slip direction. In Fig. 5 we show a photograph of a dislocation in a two-dimensional soap bubble raft obtained by the method of Bragg and Nye.

The mechanism responsible for the mobility of a dislocation is shown in Fig. 6. The motion of an edge dislocation through a crystal is analogous to the passage of a ruck or wrinkle across a rug: the ruck moves more easily than the whole rug. If atoms on one side of the slip plane are moved with respect to

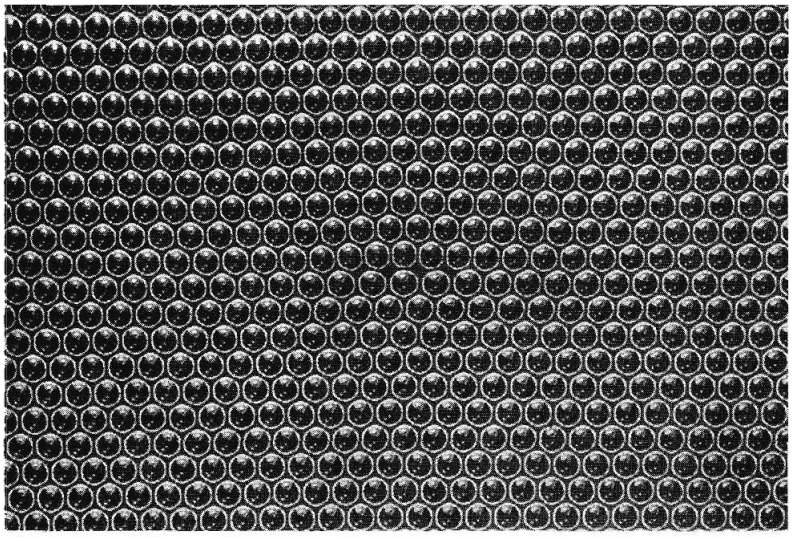


Figure 5 A dislocation in a two-dimensional bubble raft. The dislocation is most easily seen by turning the page by 30° in its plane and sighting at a low angle. (W. M. Lomer, after Bragg and Nye.)

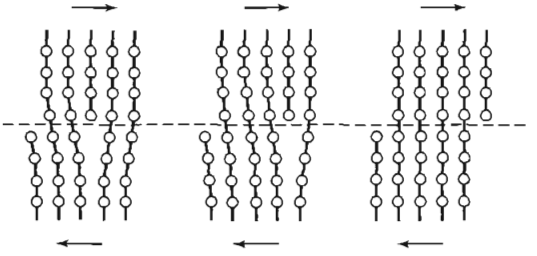


Figure 6 Motion of a dislocation under a shear tending to move the upper surface of the specimen to the right. (D. Hull.)

those on the other side, atoms at the slip plane will experience repulsive forces from some neighbors and attractive forces from others across the slip plane. These forces cancel to a first approximation. The external stress required to move a dislocation has been calculated and is quite small, below 10^5 dyn/cm^2 when the bonding forces in the crystal are not highly directional. Thus dislocations may make a crystal very plastic. Passage of a dislocation through a crystal is equivalent to a slip displacement of one part of the crystal.

The second simple type of dislocation is the **screw dislocation**, sketched in Figs. 7 and 8. A screw dislocation marks the boundary between slipped and unslipped parts of the crystal. The boundary *parallels* the slip direction, instead of lying perpendicular to it as for the edge dislocation. The screw dislocation may be thought of as produced by cutting the crystal partway through with a knife and shearing it parallel to the edge of the cut by one atom spacing. A screw dislocation transforms successive atom planes into the surface of a helix; this accounts for the name of the dislocation.

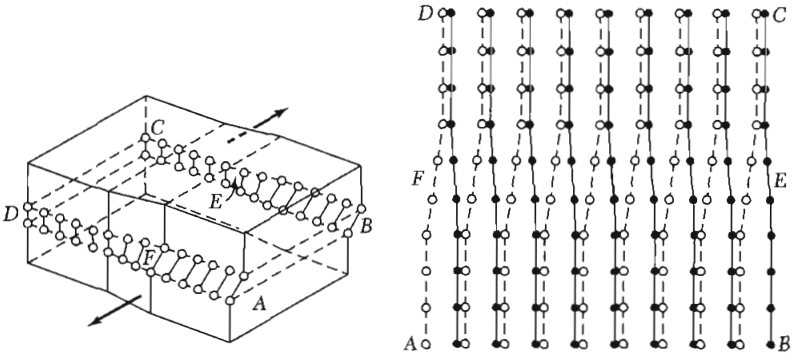


Figure 7 A screw dislocation. A part *ABEF* of the slip plane has slipped in the direction parallel to the dislocation line *EF*. A screw dislocation may be visualized as a helical arrangement of lattice planes, such that we change planes on going completely around the dislocation line. (After Cottrell.)

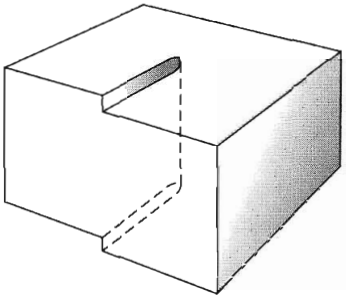


Figure 8 Another view of a screw dislocation. The broken vertical line that marks the dislocation is surrounded by strained material.

Burgers Vectors

Other dislocation forms may be constructed from segments of edge and screw dislocations. Burgers has shown that the most general form of a linear dislocation pattern in a crystal can be described as shown in Fig. 9. We consider any closed curve within a crystal, or an open curve terminating on the surface at both ends: (a) Make a cut along any simple surface bounded by the line. (b) Displace the material on one side of this surface by a vector **b** relative to the other side; here **b** is called the **Burgers vector**. (c) In regions where **b** is not parallel to the cut surface, this relative displacement will either produce a gap or cause the two halves to overlap. In these cases we imagine that we either add material to fill the gap or subtract material to prevent overlap. (d) Rejoin the material on both sides. We leave the strain displacement intact at the time of rewelding, but afterwards we allow the medium to come to internal equilibrium. The resulting strain pattern is that of the dislocation characterized jointly by the boundary curve and the Burgers vector. The Burgers vector must be equal to a lattice vector in order that the rewelding process will maintain the crystallinity of the material. The Burgers vector of a screw dislocation (Figs. 7 and 8) is parallel to the dislocation line; that of an edge dislocation (Figs. 3 and 4) is perpendicular to the dislocation line and lies in the slip plane.

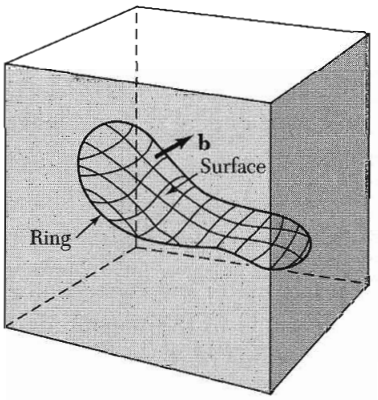


Figure 9 General method of forming a dislocation ring in a medium. The medium is represented by the rectangular block. The ring is represented by the closed curve in the interior in the block. A cut is made along the surface bounded by the curve and indicated by the contoured area. The material on one side of the cut is displaced relative to that on the other by vector distance **b**, which may be oriented arbitrarily relative to the surface. Forces will be required to effect the displacement. The medium is filled in or cut away so as to be continuous after the displacement. It is then joined in the displaced state and the applied forces are relaxed. Here **b** is the Burgers vector of the dislocation. (After Seitz.)

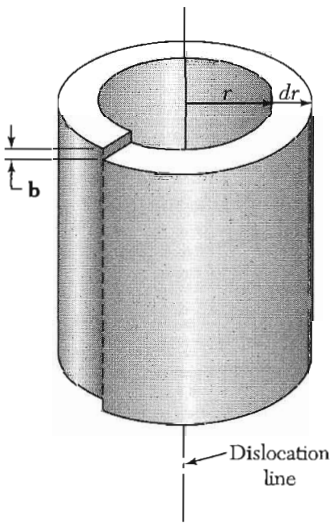


Figure 10 Shell of elastically distorted crystal surrounding screw dislocation with Burgers vector **b**; see also Fig. 16.

Stress Fields of Dislocations

The stress field of a screw dislocation is particularly simple. Figure 10 shows a shell of material surrounding an axial screw dislocation. The shell of circumference $2\pi r$ has been sheared by an amount b to give a shear strain $e = b/2\pi r$. The corresponding shear stress in an elastic continuum is

$$\sigma = Ge = Gb/2\pi r \quad . \tag{4}$$

This expression does not hold in the region immediately around the dislocation line, as the strains here are too large for continuum or linear elasticity theory to apply. The elastic energy of the shell is $dE_s = \frac{1}{2}Ge^2\,dV = (Gb^2/4\pi)\,dr/r$ per unit length. The total elastic energy per unit length of a screw dislocation is found on integration to be

$$E_S = \frac{Gb^2}{4\pi}\ln\frac{R}{r_0}\ , \tag{5}$$

where R and r_0 are appropriate upper and lower limits for the variable r. A reasonable value of r_0 is comparable to the magnitude b of the Burgers vector or to the lattice constant; the value of R cannot exceed the dimensions of the crystal. The value of the ratio R/r_0 is not very important because it enters in a logarithm term.

We now show the form of the energy of an edge dislocation. Let σ_{rr} and $\sigma_{\theta\theta}$ denote the tensile stresses in the radial and circumferential directions, and let $\sigma_{r\theta}$ denote the shear stress. In an isotropic elastic continuum, σ_{rr} and $\sigma_{\theta\theta}$ are proportional to $(\sin\theta)/r$: we need a function that falls off as $1/r$ and that changes sign when y is replaced by $-y$. The shear stress $\sigma_{r\theta}$ is proportional to $(\cos\theta)/r$; considering the plane $y = 0$, we see from Fig. 4 that the shear stress is an odd function of x. The constants of proportionality in the stress are proportional to the shear modulus G and to the Burgers vector b of the displacement. The final result is

$$\sigma_{rr} = \sigma_{\theta\theta} = -\frac{Gb}{2\pi(1-\nu)}\frac{\sin\theta}{r}\ , \qquad \sigma_{r\theta} = \frac{Gb}{2\pi(1-\nu)}\frac{\cos\theta}{r}\ , \tag{6}$$

where the Poisson ratio $\nu \approx 0.3$ for most crystals. The strain energy of a unit length of edge dislocation is

$$E_e = \frac{Gb^2}{4\pi(1-\nu)}\ln\frac{R}{r_o}\ . \tag{7}$$

We want an expression for the shear stress component σ_{xy} on planes parallel to the slip plane in Fig. 4. From the stress components σ_{rr}, $\sigma_{\theta\theta}$, and $\sigma_{r\theta}$ evaluated on the plane a distance y above the slip plane, we find

$$\sigma_{xy} = \frac{Gb}{2\pi(1-\nu)}\cdot\frac{x(x^2-y^2)}{(x^2+y^2)^2}\ . \tag{8}$$

It is shown in Problem 3 that the force caused by a resolved uniform shear stress σ is $F = b\sigma$ per unit length of dislocation. The force that an edge dislocation at the origin exerts upon a similar one at the location (y, θ) is

$$F = b\sigma_{xy} = \frac{Gb^2}{2\pi(1-\nu)}\frac{\sin 4\theta}{4y} \tag{9}$$

per unit length. Here F is the component of force in the slip direction.

Low-angle Grain Boundaries

Burgers suggested that low-angle boundaries between adjoining crystallites or crystal grains consist of arrays of dislocations. A simple example of the Burgers model of a grain boundary is shown in Fig. 11. The boundary occupies a (010) plane in a simple cubic lattice and divides two parts of the crystal that have a [001] axis in common. Such a boundary is called a pure tilt boundary: the misorientation can be described by a small rotation θ about the common [001] axis of one part of the crystal relative to the other. The tilt boundary is represented as an array of edge dislocations of spacing $D = b/\theta$, where b is the Burgers vector of the dislocations. Experiments have substantiated this model. Figure 12 shows the distribution of dislocations along small-angle grain boundaries, as observed with an electron microscope. Further, Read and Shockley derived a theory of the interfacial energy as a function of the angle of tilt, with results in excellent agreement with measurements.

Direct verification of the Burgers model is provided by the quantitative x-ray and optical studies of low-angle boundaries in germanium crystals by Vogel

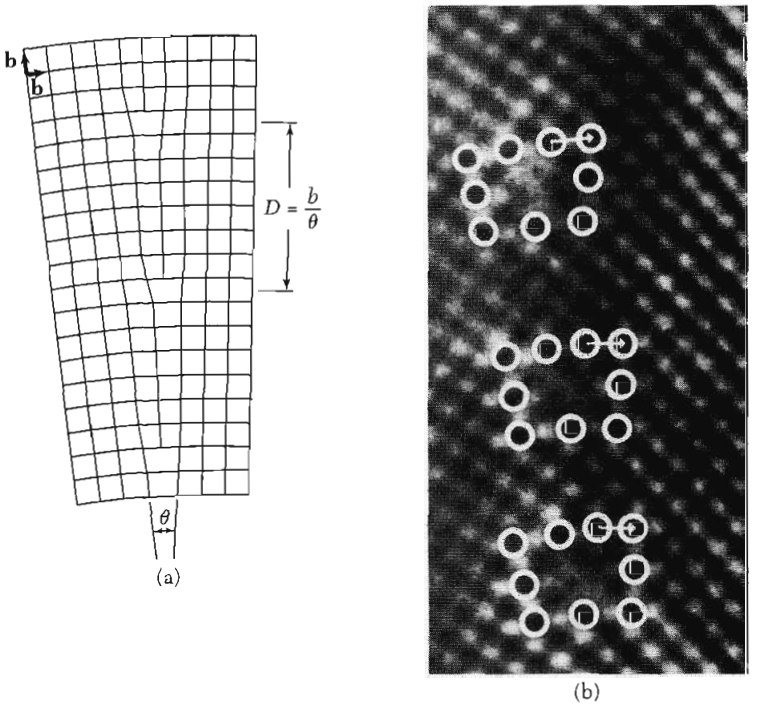


Figure 11 (*a*) Low-angle grain boundary, after Burgers. (*b*) Electron micrograph of a low-angle grain boundary in molybdenum. The three dislocations in the image each have the same Burgers vector as in the drawing in Fig. 11*a*. The white circles mark the positions of atomic columns normal to the plane of the paper. Each array of circles defines the position of a dislocation, with four circles on the top of each array and three circles below. Closure failure is indicated by the arrows which define the Burgers vectors. (Courtesy of R. Gronsky.)

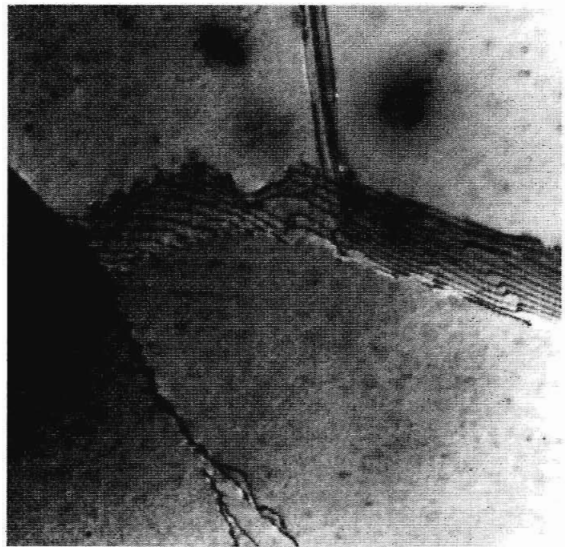


Figure 12 Electron micrograph of dislocation structures in low-angle grain boundaries in an Al-7 percent Mg solid solution. Notice the lines of small dots on the right. Mag. ×17,000. (R. Goodrich and G. Thomas.)

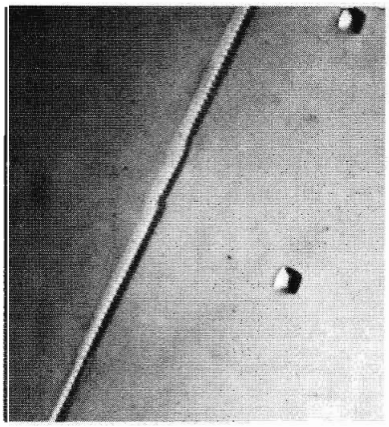


Figure 13 Dislocation etch pits in low-angle boundary on (100) face of germanium; the angle of the boundary is 27.5°. The boundary lies in a (011) plane; the line of the dislocations is [100]. The Burgers vector is the shortest lattice translation vector, or $|b| = a/\sqrt{2} = 4.0$ Å. (F. L. Vogel, Jr.)

and co-workers. By counting etch pits along the intersection of a low-angle grain boundary with an etched germanium surface (Fig. 13), they determined the dislocation spacing D. They assumed that each etch pit marked the end of a dislocation. The angle of tilt calculated from the relation $\theta = b/D$ agrees well with the angle measured directly by means of x-rays.

The interpretation of low-angle boundaries as arrays of dislocations is further supported by the fact that pure tilt boundaries move normal to themselves on application of a suitable stress. The motion has been demonstrated in

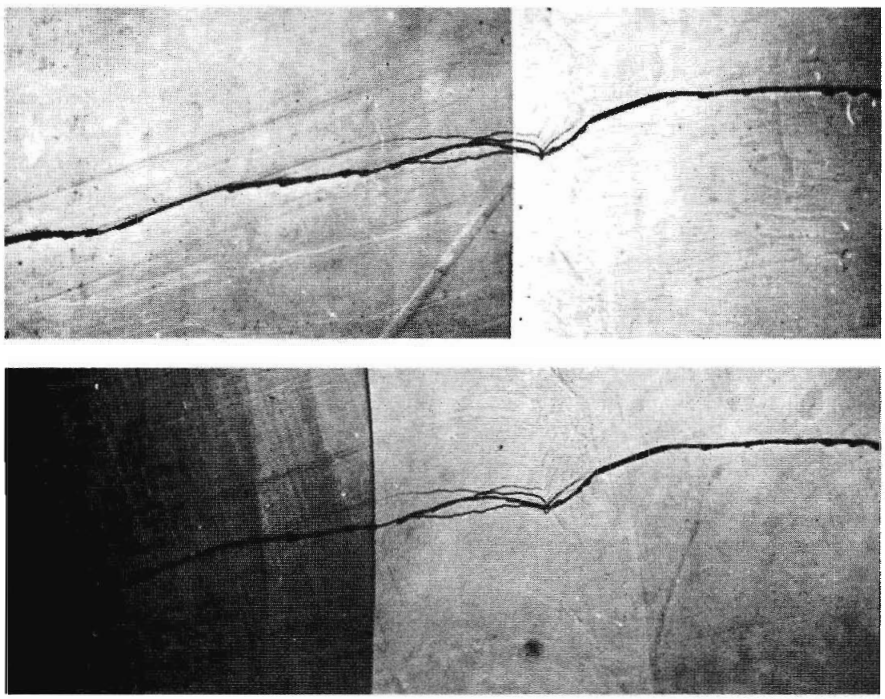


Figure 14 Motion of a low-angle grain boundary under stress. The boundary is the straight vertical line, and it is photographed under vertical illumination, thereby making evident the 2° angular change in the cleavage surface of the zinc crystal at the boundary. The irregular horizontal line is a small step in the cleavage surface which serves as a reference mark. The crystal is clamped at the left; at the right it is subject to a force normal to the plane of the page. *Top,* original position of boundary; *bottom,* moved back 0.4 mm. (J. Washburn and E. R. Parker.)

a beautiful experiment, Fig. 14. The specimen is a bicrystal of zinc containing a 2° tilt boundary with dislocations about 30 atomic planes apart. One side of the crystal was clamped, and a force was applied at a point on the opposite side of the boundary. Motion of the boundary took place by cooperative motion of the dislocations in the array, each dislocation moving an equal distance in its own slip plane. The motion was produced by stresses of the order of magnitude of the yield stress for zinc crystals, strong evidence that ordinary deformation results from the motion of dislocations.

Grain boundaries and dislocations offer relatively little resistance to diffusion of atoms in comparison with diffusion in perfect crystals. A dislocation is an open passage for diffusion. Diffusion is greater in plastically deformed material than in annealed crystals. Diffusion along grain boundaries controls the rates of some precipitation reactions in solids: the precipitation of tin from lead-tin solutions at room temperature proceeds about 10^8 times faster than expected from diffusion in an ideal lattice.

Dislocation Densities

The density of dislocations is the number of dislocation lines that intersect a unit area in the crystal. The density ranges from well below 10^2 dislocations/cm^2 in the best germanium and silicon crystals to 10^{11} or 10^{12} dislocations/cm^2 in heavily deformed metal crystals. The methods available for estimating dislocation densities are compared in Table 2. The actual dislocation configurations in cast or annealed (slowly cooled) crystals correspond either to a group of low-angle grain boundaries or to a three-dimensional network of dislocations arranged in cells, as shown in Fig. 15.

Lattice vacancies that precipitate along an existing edge dislocation will eat away a portion of the extra half-plane of atoms and cause the dislocation to climb, which means to move at right angles to the slip direction. If no dislocations are

Table 2 Methods for estimating dislocation densities[a]

Technique	Specimen thickness	Width of image	Maximum practical density, per cm^2
Electron microscopy	>1000 Å	~100 Å	$10^{11}-10^{12}$
X-ray transmission	0.1–1.0 mm	5 μm	10^4-10^5
X-ray reflection	<2 μm (min.) – 50 μm (max.)	2 μm	10^6-10^7
Decoration	~10 μm (depth of focus)	0.5 μm	2×10^7
Etch pits	no limit	0.5 μm[b]	4×10^8

[a]W. G. Johnston.
[b]Limit of resolution of etch pits.

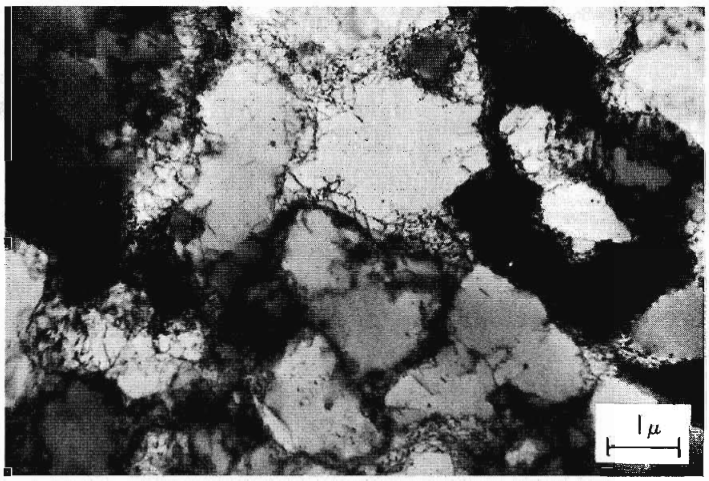


Figure 15 Cell structure of three-dimensional tangles of dislocations in deformed aluminum. (P. R. Swann.)

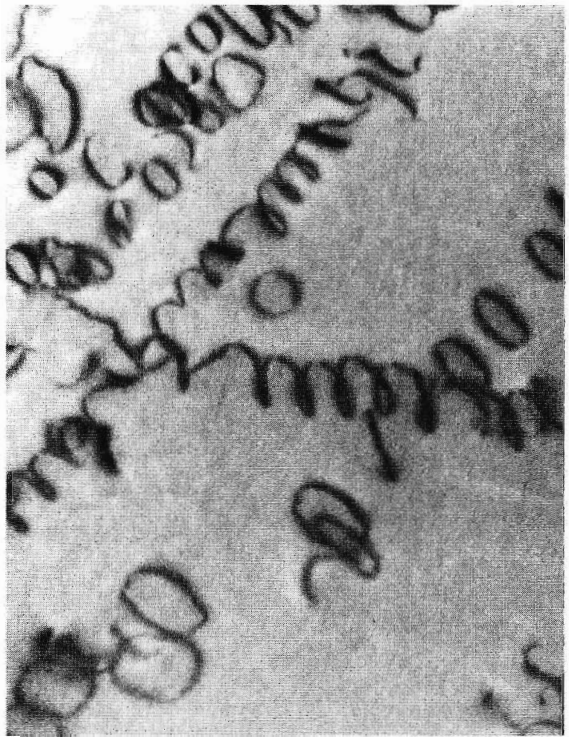


Figure 16 Electron micrograph of dislocation loops formed by aggregation and collapse of vacancies in Al–5 percent Mg quenched from 550°C. The helical dislocations are formed by vacancy condensation on a screw dislocation. Mag. ×43,000. (A. Eikum and G. Thomas.)

present, the crystal will become supersaturated with lattice vacancies; their precipitation in cylindrical vacancy plates may be followed by collapse of the plates and formation of dislocation rings that grow with further vacancy precipitation, as in Fig. 16.

Dislocation Multiplication and Slip

Plastic deformation causes a very great increase in dislocation density, typically from 10^8 to about 10^{11} dislocations/cm^2 during deformation. If a dislocation moves completely across its slip plane, an offset of one atom spacing is produced, but offsets up to 100 to 1000 atom spacings are observed. This means that dislocations multiply during deformation.

Consider a closed circular dislocation loop of radius r surrounding a slipped area having the radius of the loop. Such a loop will be partly edge, partly screw, and mostly of intermediate character. The strain energy of the loop increases as its circumference, so that the loop will tend to shrink in size. However, the loop will tend to expand if a shear stress is acting that favors slip.

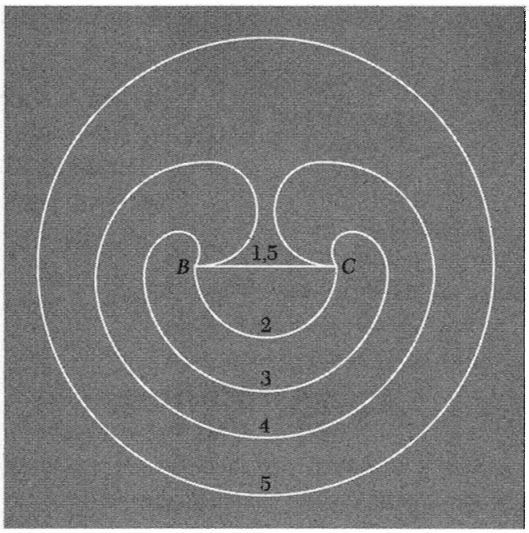


Figure 17 Frank-Read mechanism for multiplication of dislocations, showing successive stages in the generation of a dislocation loop by the segment *BC* of a dislocation line. The process can be repeated indefinitely.

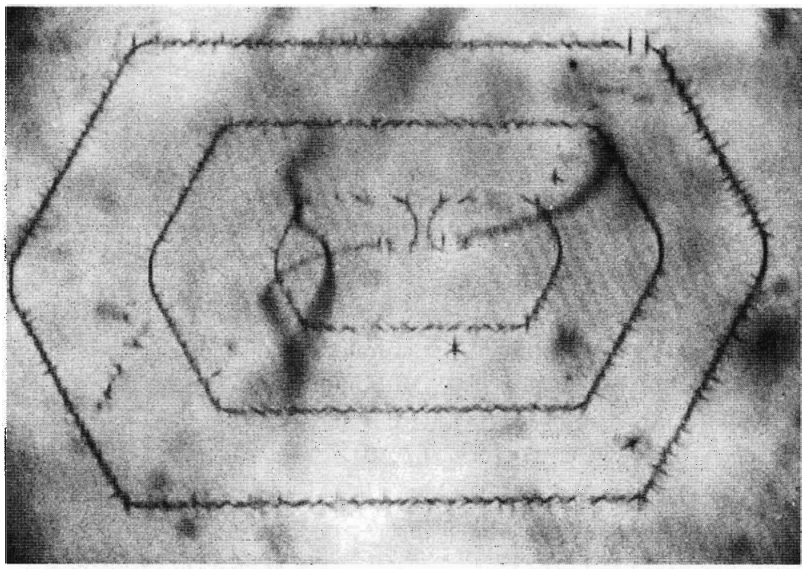


Figure 18 A Frank-Read dislocation source in silicon, decorated with copper precipitates and viewed with infrared illumination. Two complete dislocation loops are visible, and the third, innermost loop is near completion. (After W. C. Dash.)

A common feature of all dislocation sources is the bowing of dislocations. A dislocation segment pinned at each end is called a **Frank-Read source**, and it can lead (Fig. 17) to the generation of a large number of concentric dislocations on a single slip plane (Fig. 18). Related types of dislocation multiplication mechanisms account for slip and for the increased density of dislocations during plastic deformation. Double cross-slip is the most common source.

STRENGTH OF ALLOYS

Pure crystals are very plastic and yield at very low stresses. There appear to be four important ways of increasing the yield strength of an alloy so that it will withstand shear stresses as high as 10^{-2} G. They are mechanical blocking of dislocation motion, pinning of dislocations by solute atoms, impeding dislocation motion by short-range order, and increasing the dislocation density so that tangling of dislocations results. All four strengthening mechanisms depend for their success upon impeding dislocation motion. A fifth mechanism, that of removing all dislocations from the crystal, may operate for certain fine hairlike crystals (whiskers) that are discussed in the section on crystal growth.

Mechanical blocking of dislocation motion can be produced most directly by introducing tiny particles of a second phase into a crystal lattice. This process is followed in the hardening of steel, where particles of iron carbide are precipitated into iron, and in hardening aluminum, where particles of Al_2Cu are precipitated. The pinning of a dislocation by particles is shown in Fig. 19.

In strengthening by the addition of small particles there are two cases to be considered: either the particle can be deformed with the matrix, which requires that the particle can be traversed by the dislocation, or the particle cannot be traversed by the dislocation. If the particle cannot be cut, the stress

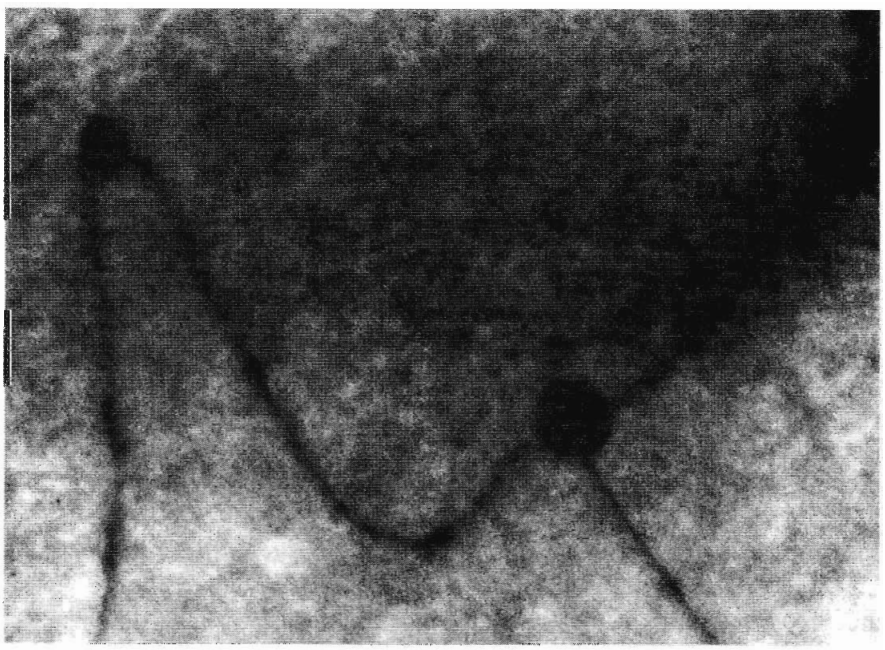


Figure 19 Dislocations pinned by particles in magnesium oxide. (Electron micrograph by G. Thomas and J. Washburn.)

necessary to force a dislocation between particles spaced L apart on a slip plane should be approximately

$$\sigma/G = b/L \ . \tag{10}$$

The smaller the spacing L, the higher is the yield stress σ. Before particles precipitate, L is large and the strength is low. Immediately after precipitation is complete and many small particles are present, L is a minimum and the strength is a maximum. If the alloy is then held at a high temperature, some particles grow at the expense of others, so that L increases and the strength drops. Hard intermetallic phases, such as refractory oxides, cannot be cut by dislocations.

The strength of dilute solid solutions is believed to result from the pinning of dislocations by solute atoms. The solubility of a foreign atom will be greater in the neighborhood of a dislocation than elsewhere in a crystal. An atom that tends to expand the crystal will dissolve preferentially in the expanded region near an edge dislocation. A small atom will tend to dissolve preferentially in the contracted region near the dislocation—a dislocation offers both expanded and contracted regions.

As a result of the affinity of solute atoms for dislocations, each dislocation will collect a cloud of associated solute atoms during cooling, at a time when the mobility of solute atoms is high. At still lower temperatures, diffusion of solute atoms effectively ceases, and the solute atom cloud becomes fixed in the crystal. When a dislocation moves, leaving its solute cloud behind, the energy of the crystal must increase. The increase in energy can only be provided by an increased stress acting on the dislocation as it pulls away from the solute atom cloud, and so the presence of the cloud strengthens the crystal.

The passage of a dislocation across a slip plane in pure crystals does not alter the binding energy across the plane after the dislocation is gone. The internal energy of the crystal remains unaffected. The same is true for random solid solutions, because the solution is equally random across a slip plane after slip. Most solid solutions, however, have short-range order. Atoms of different species are not arranged at random on the lattice sites, but tend to have an excess or a deficiency of pairs of unlike atoms. Thus in ordered alloys dislocations tend to move in pairs: the second dislocation reorders the local disorder left by the first dislocation.

The strength of a crystalline material increases with plastic deformation. The phenomenon is called **work-hardening** or **strain-hardening**. The strength is believed to increase because of the increased density of dislocations and the greater difficulty of moving a given dislocation across a slip plane that is threaded by many dislocations. Strain-hardening frequently is employed in the strengthening of materials, but its usefulness is limited to low enough temperatures so that annealing does not occur.

An important factor in strain-hardening is the total density of dislocations. In most metals dislocations tend to form cells (Fig. 15) of dislocation-free

areas of dimensions of the order of 1 μm. But unless we can get a uniform high density of dislocations we cannot strain-harden a metal to its theoretical strength, because of slip in the dislocation-free areas. A high total density is accomplished by explosive deformation or by special thermal-mechanical treatments, as of martensite in steel.

Each of the mechanisms of strengthening crystals can raise the yield strength to the range of 10^{-3} G to 10^{-2} G. All mechanisms begin to break down at temperatures where diffusion can occur at an appreciable rate. When diffusion is rapid, precipitated particles dissolve; solute clouds drift along with dislocations as they glide; short-range order repairs itself behind slowly moving dislocations; and dislocation climb and annealing tend to decrease the dislocation density. The resulting time-dependent deformation is called creep. This irreversible motion precedes the elastic limit. The search for alloys for use at very high temperatures is a search for reduced diffusion rates, so that the four strengthening mechanisms will survive to high temperatures. But the central problem of strong alloys is not strength, but ductility, for failure is often by fracture.

DISLOCATIONS AND CRYSTAL GROWTH

In some cases the presence of dislocations may be the controlling factor in crystal growth. When crystals are grown in conditions of low supersaturation, of the order of 1 percent, it has been observed that the growth rate is enormously faster than that calculated for an ideal crystal. The actual growth rate is explained in terms of the effect of dislocations on growth.

The theory of growth of ideal crystals predicts that in crystal growth from vapor, a supersaturation (pressure/equilibrium vapor pressure) of the order of 10 is required to nucleate new crystals, of the order of 5 to form liquid drops, and of 1.5 to form a two-dimensional monolayer of molecules on the face of a perfect crystal. Volmer and Schultze observed growth of iodine crystals at vapor supersaturations down to less than 1 percent, where the growth rate should have been down by the factor $\exp(-3000)$ from the rate defined as the minimum observable growth.

The large disagreement expresses the difficulty of nucleating a new monolayer on a completed surface of an ideal crystal. But if a screw dislocation is present (Fig. 20), it is never necessary to nucleate a new layer: the crystal will grow in spiral fashion at the *edge* of the discontinuity shown. An atom can be bound to a step more strongly than to a plane. The calculated growth rates for this mechanism are in good agreement with observation. We expect that nearly all crystals in nature grown at low supersaturation will contain dislocations, as otherwise they could not have grown. Spiral growth patterns have been observed on a large number of crystals. A beautiful example of the growth pattern from a single screw dislocation is given in Fig. 21.

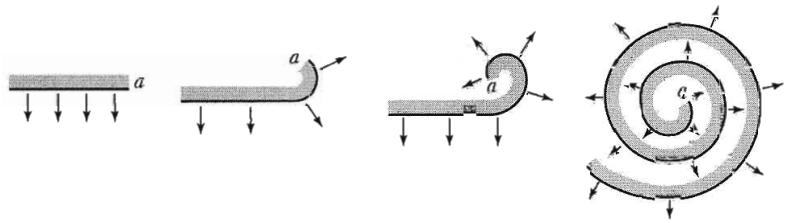


Figure 20 Development of a spiral step produced by intersection of a screw dislocation with the surface of a crystal as in Fig. 8. (F. C. Frank.)

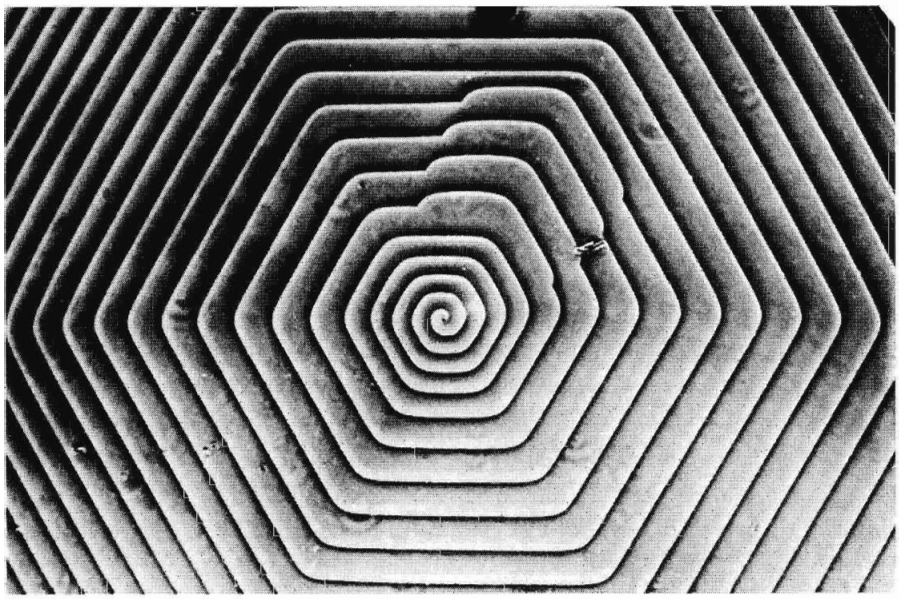


Figure 21 Phase contrast micrograph of a hexagonal spiral growth pattern on a SiC crystal. The step height is 165 Å. (A. R. Verma.)

If the growth rate is independent of direction of the edge in the plane of the surface, the growth pattern is an Archimedes spiral, $r = a\theta$, where a is a constant. The limiting minimum radius of curvature near the dislocation is determined by the supersaturation. If the radius of curvature is too small, atoms on the curved edge evaporate until the equilibrium curvature is attained. Away from the origin each part of the step acquires new atoms at a constant rate, so that $dr/dt = \text{const}$.

Whiskers

Fine hairlike crystals, or **whiskers**, have been observed to grow under conditions of high supersaturation without the necessity for more than perhaps one dislocation. It may be that these crystals contain a single axial screw dislocation

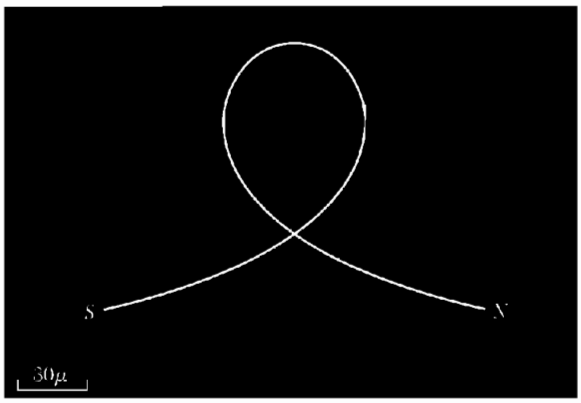


Figure 22 A nickel whisker of diameter 1000 Å bent in a loop. (R. W. De Blois.)

that aids their essentially one-dimensional growth. From the absence of dislocations we would expect these crystal whiskers to have high yield strengths, of the order of the calculated value $G/30$ discussed earlier in this chapter. A single axial screw dislocation, if present, could not cause yielding, because in bending the crystal the dislocation is not subjected to a shear stress parallel to its Burgers vector. That is, the stress is not in a direction that can cause slip. Herring and Galt observed whiskers of tin of radius $\sim 10^{-4}$ cm with elastic properties near those expected from theoretically perfect crystals. They observed yield strains of the order of 10^{-2}, which correspond to shear stresses of order $10^{-2}\ G$, about 1000 times greater than in bulk tin, confirming the early estimates of the strength of perfect crystals. Theoretical or ideal elastic properties have been observed for a number of materials as for carbon nanotubes. A single domain whisker of nickel is shown in Fig. 22.

HARDNESS OF MATERIALS

The hardness of materials is measured in several ways, the simplest test for nonmetals being the scratch test. Substance A is harder than substance B if A will scratch B but B will not scratch A. A standard scale is used for representative minerals, with diamond, the hardest, assigned the value 10 and talc, the softest, assigned the value 1:

10 diamond C	5 apatite $Ca_5(PO_4)_3F$
9 corundum Al_2O_3	4 fluorite CaF_2
8 topaz $Al_2SiO_4F_2$	3 calcite $CaCO_3$
7 quartz SiO_2	2 gypsum $CaSO_4 \cdot 2H_2O$
6 orthoclase $KAlSi_3O_8$	1 talc $3MgO \cdot 4SiO_2 \cdot H_2O$

There is great current interest in the development of materials of great hardness, for example as films for use as scratch-resistant coatings on lenses.

It is widely felt that the scale between diamond and corundum is misleading, because diamond is much, much harder than corundum. It has been suggested that one might assign diamond the hardness 15, with the gap between 9 and 15 to be filled in eventually by synthetic materials, such as compounds of C and B.

Modern scales of hardness, such as the VHN scale, are based on indenter tests in which an indenter is pressed into the surface of the material and the size of the impression is measured. The Vickers Hardness Numbers of selected materials are tabulated below, after conversion by E. R. Weber to units of GPa [GN/m^2]:

Diamond	45.3	BeO	7.01
SiC	20.0	Steel (quenched)	4.59
Si_3N_4	18.5	Cu (annealed)	0.25
Al_2O_3	14.0	Al (annealed)	0.12
B	13.5	Pb	0.032
WC	11.3		

The data are from J. C. Anderson and others.

Problems

1. ***Lines of closest packing.*** Show that the lines of closest atomic packing are ⟨110⟩ in fcc structures and ⟨111⟩ in bcc structures.

2. ***Dislocation pairs.*** (a) Find a pair of dislocations equivalent to a row of lattice vacancies; (b) find a pair of dislocations equivalent to a row of interstitial atoms.

3. ***Force on dislocation.*** Consider a crystal in the form of a cube of side L containing an edge dislocation of Burgers vector $\mathbf{b}$. If the crystal is subjected to a shear stress σ on the upper and lower faces in the direction of slip, show, by considering energy balance, that the force acting on the dislocation is $F = b\sigma$ per unit length.

22

Alloys

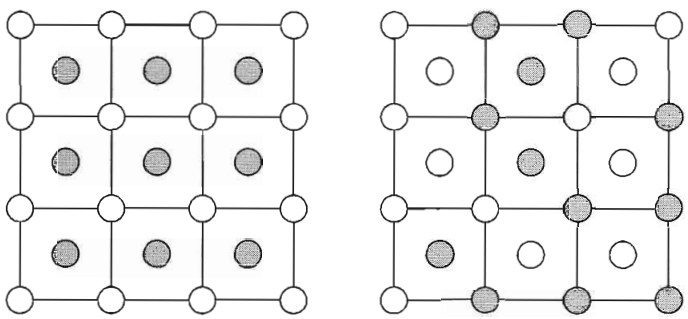


Figure 1 Ordered (a) and disordered (b) arrangements of AB ions in the alloy AB.

CHAPTER 22: ALLOYS

GENERAL CONSIDERATIONS

The theory of the band structure of solids assumes that the crystal has translational invariance. But suppose that the crystal is composed of two elements A and B that occupy at random the regular lattice sites of the structure, in proportions x and $1 - x$ for the composition A_xB_{1-x}. The translational symmetry is no longer perfect. Will we then lose the consequences of band theory, such as the existence of Fermi surfaces and of energy gaps? Will insulators become conductors because the energy gap is gone? We touched on these questions in the discussion of amorphous semiconductors in Chapter 19.

Experiment and theory agree that the consequences of the destruction of perfect translational symmetry are much less serious (nearly always) than we expect at first sight. The viewpoint of the effective screened potential of Chapter 9 is helpful in these matters, first because the effective potentials are relatively weak in comparison with a free ion potential and, second and most important, the differences between the effective potentials of the host and the additive atoms may be very weak in comparison with either alone. Alloys of Si and Ge or of Cu and Ag are classic examples of what we may call the relative ineffectiveness of alloying.

In any event, a low concentration of impurity atoms cannot have much effect on the Fourier components $U_{\mathbf{G}}$ of the effective potential $U(\mathbf{r})$ that is responsible for the band gaps and for the form of the Fermi surface. (This statement implies that the $\mathbf{G}$'s exist, which implies that a regular lattice exists. This is not an important assumption because we know that thermal phonons do not have drastic effects on the band structure, so that lattice distortions described as frozen-in phonons should not have drastic effects. If the distortions are more serious, as with amorphous solids, the electronic changes can be significant.)

It is true that an impurity atom will introduce Fourier components of $U(\mathbf{r})$ at wavevectors that are not reciprocal lattice vectors, but at low impurity concentration such components are never large in comparison with the $U_{\mathbf{G}}$, arguing from the statistics of random potentials. The Fourier components at the reciprocal lattice vectors $\mathbf{G}$ will still be large and will give the band gaps, Fermi surfaces, and sharp x-ray diffraction lines characteristic of a regular lattice.

The consequences of alloying will be particularly small when the impurity element belongs to the same column of the periodic table as the host element it replaces, because the atomic cores will make rather similar contributions to the effective potentials.

One measure of the effect of alloying is the residual electrical resistivity, defined as the low temperature limit of the resistivity. Here we must distinguish

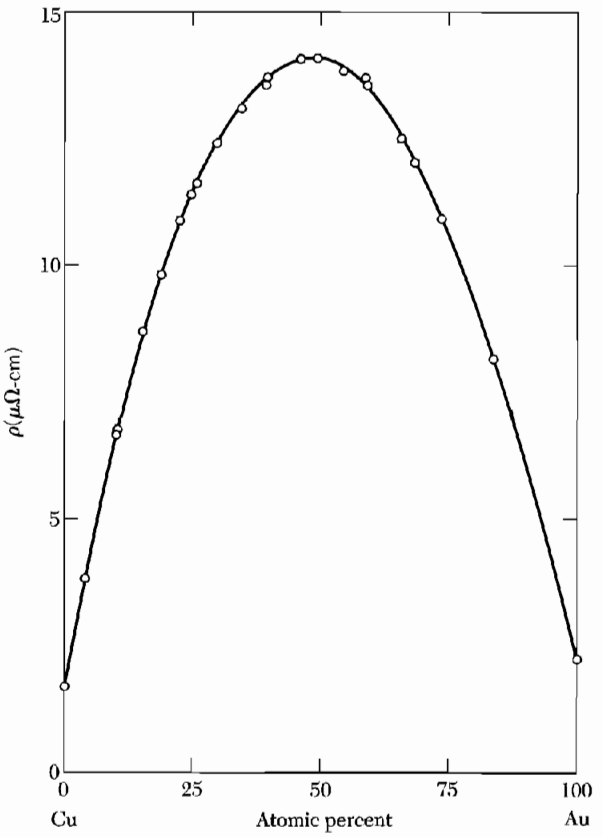


Figure 2 Resistivity of a disordered binary alloy of copper and gold. The variation of the residual resistivity depends on the composition Cu_xAu_{1-x} as $x(1-x)$, which is known as Nordheim's Rule for a disordered alloy. Here $x(1-x)$ is a measure of the degree of maximum disorder possible for a given value of x. (Johansson and Linde.)

between disordered and ordered alloys. An alloy is disordered if the A and B atoms are randomly arranged, which occurs for a general value of x in the composition A_xB_{1-x}. For special values of x, such as 1/4, 1/2, and 3/4 for a cubic structure, it is possible for ordered phases to form, phases in which the A and B atoms form an ordered array. The distinction between order and disorder is shown in Fig. 1. The effect of order on the electrical resistivity is shown in Figs. 2 and 3. The residual resistivity increases with disorder, as discussed for amorphous materials in Chapter 19. The effect is shown in Fig. 2 for the Cu-Au alloy system. When the specimen is cooled slowly from a high temperature, ordered structures are formed at Cu_3Au and CuAu; these structures have a lower residual resistivity by virtue of their order, as in Fig. 3.

Thus we can use the residual electrical resistivity to measure the effect of alloying in a disordered structure. One atomic percent of copper dissolved in silver (which lies in the same column of the periodic table) increases the

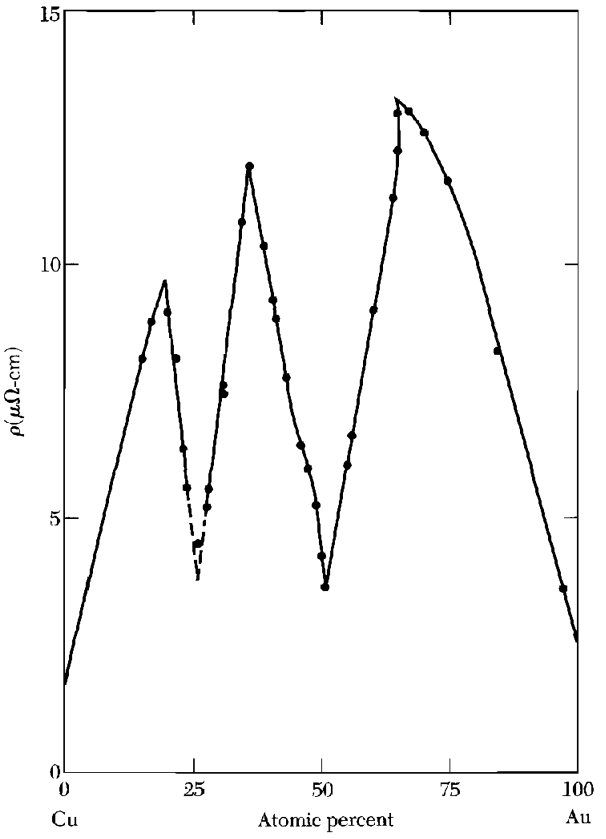


Figure 3 Effect of ordered phases on the resistivity of a binary alloy Cu_xAu_{1-x}. The alloys here have been annealed, whereas those in Fig. 2 have been quenched (cooled rapidly). The compositions of low residual resistivity correspond to the ordered compositions Cu_3Au and CuAu. (Johansson and Linde.)

residual resistivity by 0.077 μohm-cm. This corresponds to a geometrical scattering cross section which is only 3 percent of the naive "projected area" of the impurity atom, so that the scattering effect is very small.

In insulators there is no experimental evidence for a significant reduction of band gap caused by the random potential components. For example, silicon and germanium form homogeneous solid solutions, known as substitutional alloys, over the entire composition range, but the band edge energies vary continuously with composition from the pure Si gap to the pure Ge gap.

It is widely believed, however, that the density of states near the band edges in amorphous materials is smeared by the gross absence of translational symmetry. Some of the new states thus formed just inside the gap may not necessarily be current-carrying states because they may not extend throughout the crystal.

SUBSTITUTIONAL SOLID SOLUTIONS—HUME-ROTHERY RULES

We now discuss substitutional solid solutions of one metal A in another metal B of different valence, where A and B occupy, at random, equivalent sites in the structure. Hume-Rothery treated the empirical requirements for the stability of a solid solution of A and B as a single phase system.

One requirement is that the atomic diameters be compatible, which means that they should not differ by more than 15 percent. For example, the diameters are favorable in the Cu (2.55 Å) − Zn (2.65 Å) alloy system: zinc dissolves in copper as an fcc solid solution up to 38 atomic percent zinc. The diameters are less favorable in the Cu (2.55 Å) − Cd (2.97 Å) system, where only 1.7 atomic percent cadmium is soluble in copper. The atomic diameters referred to copper are 1.04 for zinc and 1.165 for cadmium.

Although the atomic diameters may be favorable, solid solutions will not form when there is a strong chemical tendency for A and B to form "intermetallic compounds," which are compounds of definite chemical proportions. If A is strongly electronegative and B strongly electropositive, compounds such as AB and A_2B may precipitate from the solid solution. (This is different from the formation of an ordered alloy phase only by the greater chemical bonding strength of the intermetallic compounds.) Although the atomic diameter ratio is favorable for As in Cu (1.02), only 6 atomic percent As is soluble. The diameter ratio is also favorable for Sb in Mg (1.06), yet the solubility of Sb in Mg is very small.

The electronic structure of alloys can often be described by the average number of conduction electrons (or valence electrons) per atom, denoted by n. In the alloy CuZn the value of n is 1.50; in CuAl, $n = 2.00$. Changes in electron concentration determine structural changes in many alloy systems.

The phase diagram of the copper-zinc system[1] is shown in Fig. 4. The fcc structure of pure copper ($n = 1$) persists on the addition of zinc ($n = 2$) until the electron concentration reaches 1.38. A bcc structure occurs at a minimum electron concentration of about 1.48. The γ phase exists for the approximate range of n between 1.58 and 1.66, and the hcp phase ϵ occurs near 1.75.

The term **electron compound** denotes an intermediate phase (such as the β phase of CuZn) whose crystal structure is determined by a fairly well defined electron to atom ratio. For many alloys the ratio is close to the **Hume-Rothery rules**: 1.50 for the β phase, 1.62 for the γ phase, and 1.75 for the ϵ phase. Representative experimental values are collected in Table 1, based on the usual chemical valence of 1 for Cu and Ag; 2 for Zn and Cd; 3 for Al and Ga; 4 for Si, Ge, and Sn.

The Hume-Rothery rules find a simple expression in terms of the band theory of nearly free electrons. The observed limit of the fcc phase occurs

[1]The phases of interest are usually denoted by metallurgists by Greek characters: in the Cu-Zn system we have α (fcc), β (bcc), γ (complex cubic cell of 52 atoms), ϵ (hcp) and η (hcp); ϵ and η differ considerably in c/a ratio. The meaning of the characters depends on the alloy system.

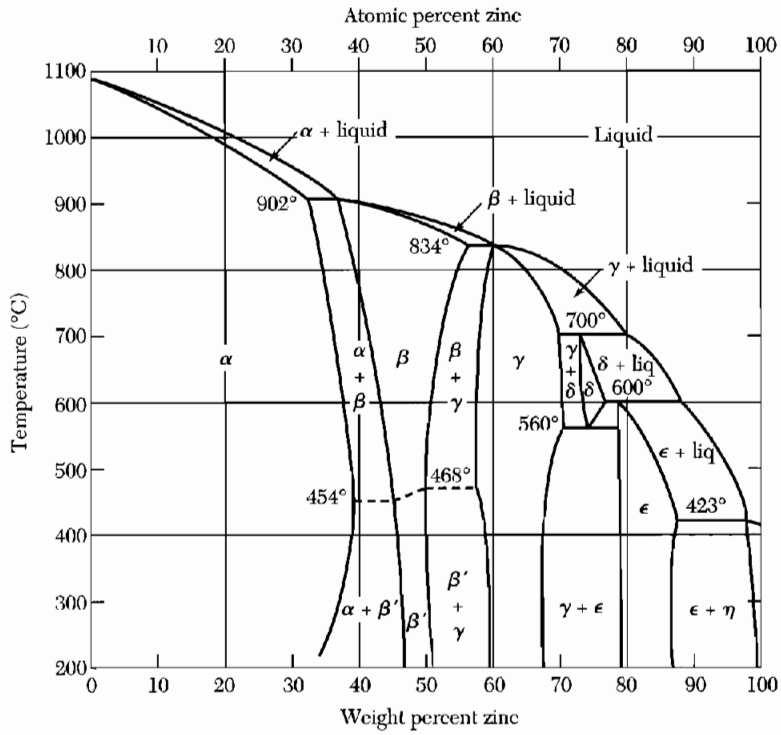


Figure 4 Equilibrium diagram of phases in the copper-zinc alloy system. The α phase is fcc; β and β' are bcc; γ is a complex structure; ϵ and η are both hcp, but ϵ has a c/a ratio near 1.56 and η (for pure Zn) has $c/a = 1.86$. The β' phase is ordered bcc, by which we mean that most of the Cu atoms occupy sites on one sc sublattice and most of the Zn atoms occupy sites on a second sc sublattice that interpenetrates the first sublattice. The β phase is disordered bcc: any site is equally likely to be occupied by a Cu or Zn atom, almost irrespective of what atoms are in the neighboring sites.

Table 1 Electron/atom ratios of electron compounds

Alloy	fcc phase boundary	Minimum bcc phase boundary	γ-phase boundaries	hcp phase boundaries
Cu-Zn	1.38	1.48	1.58–1.66	1.78–1.87
Cu-Al	1.41	1.48	1.63–1.77	
Cu-Ga	1.41			
Cu-Si	1.42	1.49		
Cu-Ge	1.36			
Cu-Sn	1.27	1.49	1.60–1.63	1.73–1.75
Ag-Zn	1.38		1.58–1.63	1.67–1.90
Ag-Cd	1.42	1.50	1.59–1.63	1.65–1.82
Ag-Al	1.41			1.55–1.80

close to the electron concentration of 1.36 at which an inscribed Fermi sphere makes contact with the Brillouin zone boundary for the fcc lattice. The observed electron concentration of the bcc phase is close to the concentration 1.48 at which an inscribed Fermi sphere makes contact with the zone boundary for the bcc lattice. Contact of the Fermi sphere with the zone boundary for the γ phase is at the concentration 1.54. Contact for the hcp phase is at the concentration 1.69 for the ideal c/a ratio.

Why is there a connection between the electron concentrations at which a new phase appears and at which the Fermi surface makes contact with the boundary of the Brillouin zone? We recall that the energy bands split into two at the region of contact on the zone boundary (Chapter 9). If we add more electrons to the alloy at this stage, they will have to be accommodated in the upper band or in states of high energy near the zone corners of the lower band. Both options are possible, and both involve an increase of energy. It may also be energetically favorable for the crystal structure to change to one which can contain a Fermi surface of larger volume (more electrons) before contact is made with the zone boundary. In this way H. Jones made plausible the sequence of structures fcc, bcc, γ, hcp with increasing electron concentration.

Measurements of the lattice parameter of Li-Mg alloys are shown in Fig. 5. In the range shown the structure is bcc. The lattice contracts during the initial stages of the addition of Mg to Li. When the lithium content drops below 50 atomic percent, corresponding to an average electron concentration increasing above 1.5 per atom, the lattice starts to expand. We have seen that

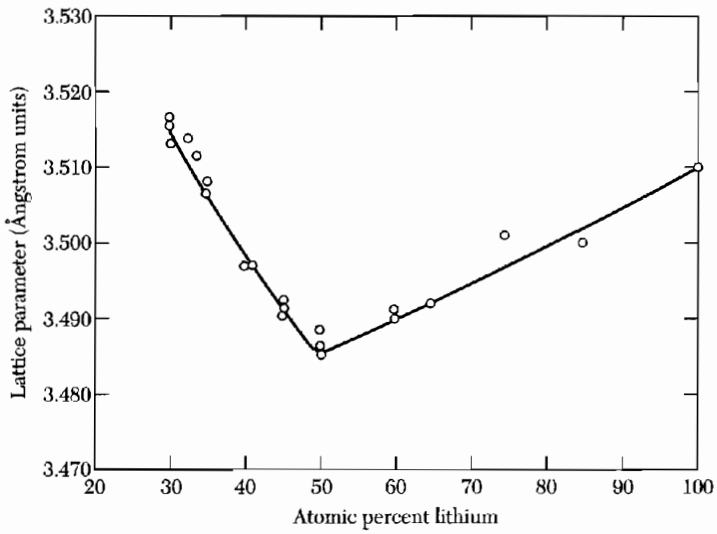


Figure 5 Lattice parameter of body-centered cubic magnesium-lithium alloys. (After D. W. Levinson.)

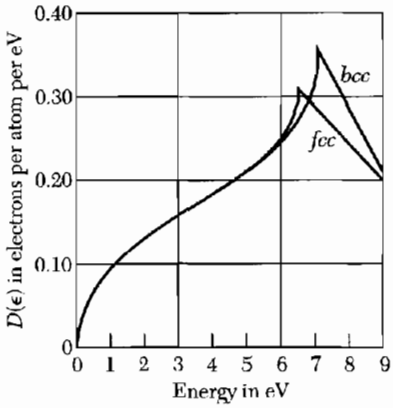


Figure 6 Number of orbitals per unit energy range for the first Brillouin zone of the fcc and bcc lattices, as a function of energy.

for a spherical Fermi surface, contact with the zone boundary is established at $n = 1.48$ electrons per atom, in a bcc lattice. It appears that the expansion of the lattice arises from the onset of overlap across the zone boundary.

The transformation from fcc to bcc is illustrated by Fig. 6; this shows the number of orbitals per unit energy range as a function of energy, for the fcc and bcc structures. As the number of electrons is increased, a point is reached where it is easier to accommodate additional electrons in the Brillouin zone of the bcc lattice rather than in the Brillouin zone of the fcc lattice. The figure is drawn for copper.

ORDER-DISORDER TRANSFORMATION

The dashed horizontal line in the beta-phase (bcc) region of the phase diagram (Fig. 4) of the Cu-Zn system represents the transition temperature between the ordered (low temperature) and disordered (high temperature) states of the alloy. In the common ordered arrangement of an AB alloy with a bcc structure, all the nearest-neighbor atoms of a B atom are A atoms, and vice versa. This arrangement results when the dominant interaction among the atoms is an attraction between A and B atoms. (If the AB interaction is weakly attractive or repulsive, a two-phase system is formed in which some crystallites are largely A and other crystallites are largely B.)

The alloy is completely ordered in equilibrium at absolute zero. It becomes less ordered as the temperature is increased, until a transition temperature is reached above which the structure is disordered. The transition temperature marks the disappearance of **long-range order**, which is order over many interatomic distances, but some **short-range order** or correlation among near neighbors may persist above the transition. The long-range order in an AB alloy is shown in Fig. 7a. Long- and short-range order for an alloy of composition AB_3 is given in Fig. 7b. The degree of order is defined below.

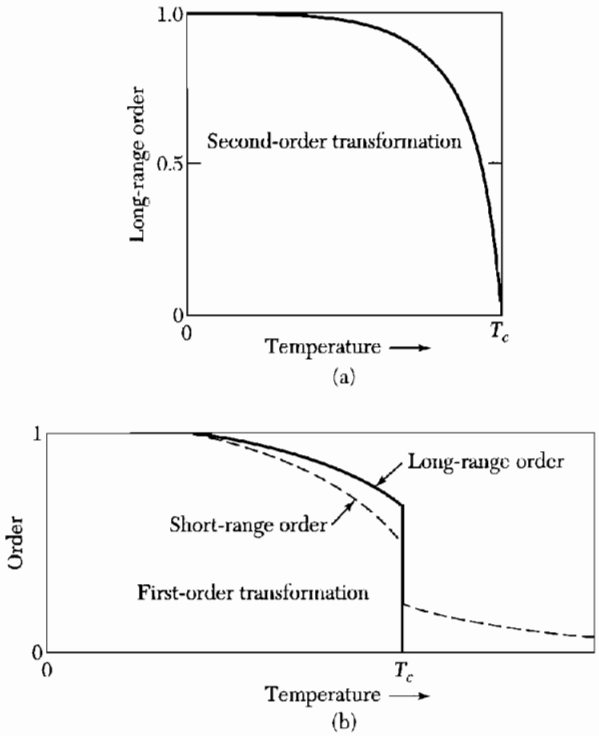
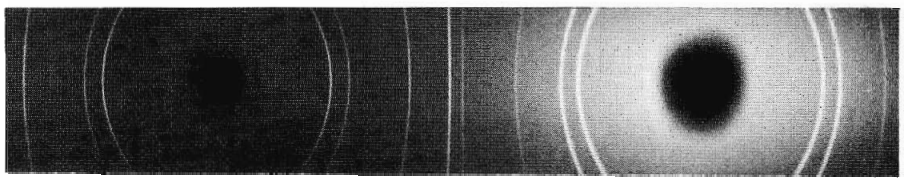


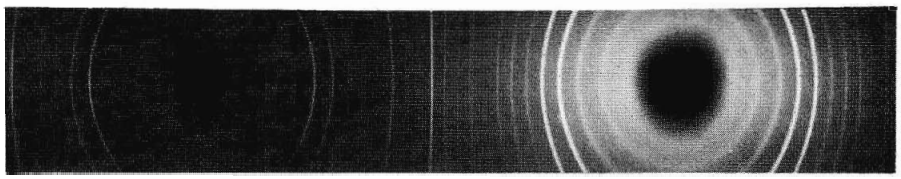
Figure 7 (a) Long-range order versus temperature for an AB alloy. The transformation is second order. (b) Long-range and short-range order for an AB_3 alloy. The transformation for this composition is first order.

If an alloy is cooled rapidly from high temperatures to a temperature below the transition, a metastable condition may be produced in which a nonequilibrium disorder is frozen in the structure. The reverse effect occurs when an ordered specimen is disordered at constant temperature by heavy irradiation with nuclear particles. The degree of order may be investigated experimentally by x-ray diffraction. The disordered structure in Fig. 8 has diffraction lines at the same positions as if the lattice points were all occupied by only one type of atom, because the effective scattering power of each plane is equal to the average of the A and B scattering powers. The ordered structure has extra diffraction lines not possessed by the disordered structure. The extra lines are called **superstructure lines**.

The use of the terms order and disorder in this chapter always refers to regular lattice sites; it is the occupancy that is randomly A or B. Do not confuse this usage with that of Chapter 19 on noncrystalline solids where there are no regular lattice sites and the structure itself is random. Both possibilities occur in nature.



(a)



(b)

Figure 8 X-ray powder photographs in $AuCu_3$ alloy. (a) Disordered by quenching from $T > T_c$; (b) ordered by annealing at $T < T_c$. (Courtesy of G. M. Gordon.)

The structure of the ordered CuZn alloy is the cesium chloride structure of Chapter 1. The space lattice is simple cubic, and the basis has one Cu atom at 000 and one Zn atom at $\frac{1}{2}\frac{1}{2}\frac{1}{2}$. The diffraction structure factor

$$S(hkl) = f_{Cu} + f_{Zn}\, e^{-i\pi(h+k+l)} \ . \tag{1}$$

This cannot vanish because $f_{Cu} \neq f_{Zn}$; therefore all reflections of the simple cubic space lattice will occur. In the disordered structure the situation is different: the basis is equally likely to have either Zn or Cu at 000 and either Zn or Cu at $\frac{1}{2}\frac{1}{2}\frac{1}{2}$. Then the average structure factor is

$$\langle S(hkl)\rangle = \langle f\rangle + \langle f\rangle\, e^{-i\pi(h+k+l)} \ , \tag{2}$$

where $\langle f\rangle = \frac{1}{2}(f_{Cu} + f_{Zn})$. Equation (2) is exactly the form of the result for the bcc lattice; the reflections vanish when $h + k + l$ is odd. We see that the ordered lattice has reflections (the superstructure lines) not present in the disordered lattice (Fig. 8).

Elementary Theory of Order

We give a simple statistical treatment of the dependence of order on temperature for an AB alloy with a bcc structure. The case A_3B differs from AB, the former having a first-order transition marked by a latent heat and the latter having a second-order transition marked by a discontinuity in the heat capacity (Fig. 9). We introduce a measure of the long-range order. We call one simple cubic lattice a and the other b: the bcc structure is composed of the two interpenetrating sc lattices, and the nearest neighbors of an atom on one lattice lie on the other lattice. If there are N atoms A and N atoms B in the alloy, the

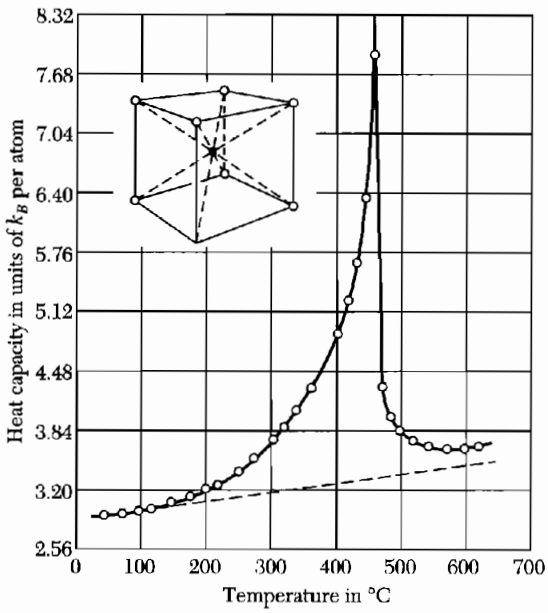


Figure 9 Heat capacity versus temperature of CuZn alloy (β-brass).

long-range order parameter P is defined so that the number of A's on the lattice a is equal to $\frac{1}{2}(1 + P)N$. The number of A's on lattice b is equal to $\frac{1}{2}(1-P)N$. When $P = \pm 1$, the order is perfect and each lattice contains only one type of atom. When $P = 0$, each lattice contains equal numbers of A and B atoms and there is no long-range order.

We consider that part of the internal energy associated with the bond energies of AA, AB, and BB nearest-neighbor pairs. The total bond energy is

$$E = N_{AA}U_{AA} + N_{BB}U_{BB} + N_{AB}U_{AB} \ , \tag{3}$$

where N_{ij} is the number of nearest-neighbor ij bonds and U_{ij} is the energy of an ij bond.

The probability that an atom A on lattice a will have an AA bond is equal to the probability that an A occupies a particular nearest-neighbor site on b, times the number of nearest-neighbor sites, which is 8 for the bcc structure. We assume that the probabilities are independent. Thus, by the preceding expressions for the number of A's on a and b,

$$\begin{aligned} N_{AA} &= 8[\tfrac{1}{2}(1 + P)N][\tfrac{1}{2}(1 - P)] = 2(1 - P^2)N \ ; \\ N_{BB} &= 8[\tfrac{1}{2}(1 + P)N][\tfrac{1}{2}(1 - P)] = 2(1 - P^2)N \ ; \\ N_{AB} &= 8N[\tfrac{1}{2}(1 + P)]^2 + 8N[\tfrac{1}{2}(1 - P)]^2 = 4(1 + P^2)N \ . \end{aligned} \tag{4}$$

The energy (3) becomes

$$E = E_0 + 2NP^2U \ , \tag{5}$$

where

$$E_0 = 2N(U_{AA} + U_{BB} + 2U_{AB}) \ ; \qquad U = 2U_{AB} - U_{AA} - U_{BB} \ . \tag{6}$$

We now calculate the entropy of this distribution of atoms. There are $\frac{1}{2}(1 + P)N$ atoms A and $\frac{1}{2}(1 - P)N$ atoms B on lattice a; there are $\frac{1}{2}(1 - P)N$ atoms A and $\frac{1}{2}(1 + P)N$ atoms B on lattice b. The number of arrangements G of these atoms is

$$G = \left[\frac{N!}{[\frac{1}{2}(1+P)N]![\frac{1}{2}(1-P)N]!}\right]^2 \ . \tag{7}$$

From the definition of the entropy as $S = k_B \ln G$, we have, using Stirling's approximation,

$$S = 2Nk_B \ln 2 - Nk_B[(1 + P)\ln(1 + P) + (1 - P) \ln (1 - P)] \ . \tag{8}$$

This defines the **entropy of mixing**. For $P = \pm 1$, $S = 0$; for $P = 0$, $S = 2Nk_B \ln 2$.

The equilibrium order is determined by the requirement that the free energy $F = E - TS$ be a minimum with respect to the order parameter P. On differentiating F with respect to P, we have as the condition for the minimum

$$4NPU + Nk_BT \ln \frac{1+P}{1-P} = 0 \ . \tag{9}$$

The transcendental equation for P may be solved graphically; we find the smoothly decreasing curve shown in Fig. 7a. Near the transition we may expand (9) to find $4NPU + 2Nk_BTP = 0$. At the transition temperature $P = 0$, so that

$$T_c = -2U/k_B \ . \tag{10}$$

For a transition to occur, the effective interaction U must be negative.

The **short-range order parameter** r is a measure of the fraction of the average number q of nearest-neighbor bonds that are AB bonds. When completely disordered, an AB alloy has an average of four AB bonds about each atom A. The total possible is eight. We may define

$$r = \tfrac{1}{4}(q - 4) \ , \tag{11}$$

so that $r = 1$ in complete order and $r = 0$ in complete disorder. Observe that r is a measure only of the local order about an atom, whereas the long-range order parameter P refers to the purity of the entire population on a given sublattice. Above the transition temperature T_c the long-range order is rigorously zero, but the short-range order is not.

PHASE DIAGRAMS

There is a large amount of information in a phase diagram even for a binary system, as in Fig. 4. The areas enclosed by curves relate to the equilibrium state in that region of composition and temperature. The curves mark the course of phase transitions as plotted in the T-x plane, where x is the composition parameter.

The equilibrium state is the state of minimum free energy of the binary system at given T, x. Thus the analysis of a phase diagram is the subject of thermodynamics. Several extraordinary results come out of this analysis, in particular the existence of low-melting-point eutectic compositions. Because the analysis has been treated in Chapter 11 of *TP*, we only outline the principal results here.

Two substances will dissolve in each other and form a homogeneous mixture if that is the configuration of lowest free energy accessible to the components. The substances will form a heterogeneous mixture if the combined free energy of the two separate phases side by side is lower than the free energy of the homogeneous mixture. Now we say that the mixture exhibits a **solubility gap**. In Fig. 4 we see that compositions near $Cu_{0.60}Zn_{0.40}$ are in a solubility gap and are mixtures of fcc and bcc phases of different structures and compositions. The phase diagram represents the temperature dependence of the solubility gaps.

When a small fraction of a homogeneous liquid freezes, the composition of the solid that forms is almost always different from that of the liquid. Consider a horizontal section near the composition $Cu_{0.80}Zn_{0.20}$ in Fig. 4. Let x denote the weight percent of zinc. At a given temperature, there are three regions:

$x > x_L$, the equilibrium system is a homogeneous liquid.
$x_S < x < x_L$, there is a solid phase of composition x_S and a liquid phase of composition x_L.
$x < x_S$, equilibrium system is a homogeneous solid.

The point x_L traces a curve called the **liquidus** curve, and the point x_S traces the **solidus** curve.

Eutectics. Mixtures with two liquidus branches in their phase diagram are called eutectics, as in Fig. 10 for the Au-Si system. The minimum solidification temperature is called the eutectic temperature; here the composition is the eutectic composition. The solid at this composition consists of two separate phases, as in the microphotograph of Fig. 11.

There are many binary systems in which the liquid phase persists to temperatures below the lower melting temperature of the constituents. Thus $Au_{0.69}Si_{0.31}$ solidifies at 370°C as a two-phase heterogeneous mixture, although Au and Si solidify at 1063°C and 1404°C, respectively. One phase of the eutectic is nearly pure gold; the other nearly pure silicon.

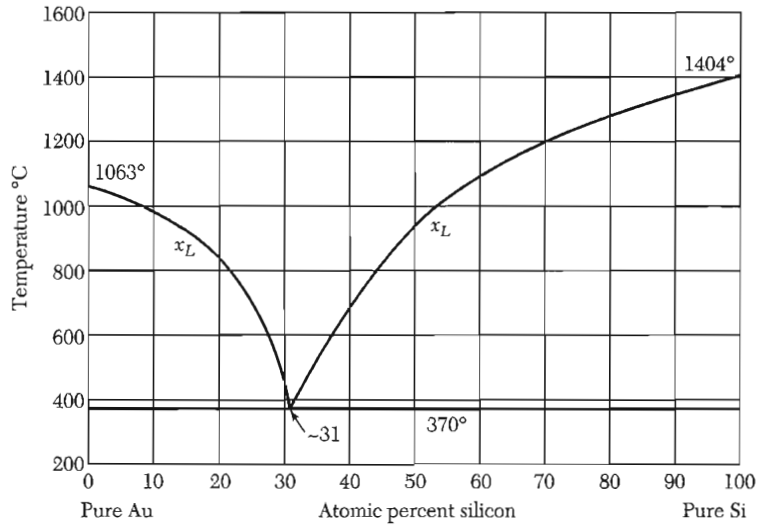


Figure 10 Eutectic phase diagram of gold-silicon alloys. The eutectic consists of two branches that come together at T_e = 370 °C and x_g = 0.31 atomic percent Si. (After Kittel and Kroemer, *TP*.)

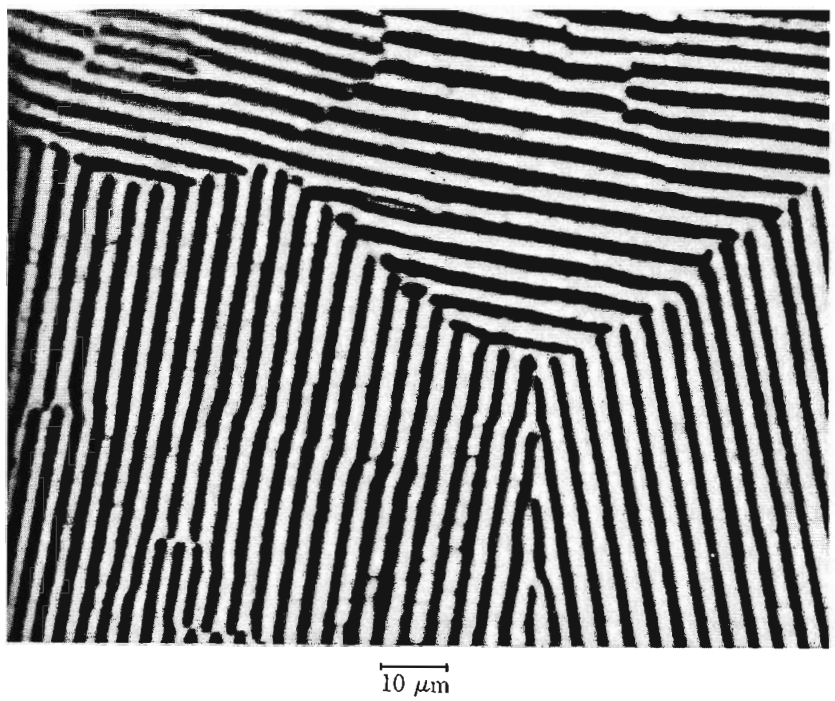


Figure 11 Microphotograph of the Pb-Sn eutectic. (Courtesy of J. D. Hunt and K. A. Jackson.)

The Au-Si eutectic is important in semiconductor technology because the eutectic permits low temperature welding of gold contact wires to silicon devices. Lead-tin alloys have a similar eutectic of $Pb_{0.26}Sn_{0.74}$ at 183°C. This or nearby compositions are used in solder: nearby if a range of melting temperatures is desired for ease in handling.

TRANSITION METAL ALLOYS

When we add copper to nickel, the effective magneton number per atom decreases linearly and goes through zero near $Cu_{0.60}Ni_{0.40}$, as shown in Fig. 12. At this composition the extra electron from the copper has filled the 3*d* band, or the spin-up and spin-down 3*d* sub-bands that were shown in Fig. 12.7b. The situation is shown schematically in Fig. 13.

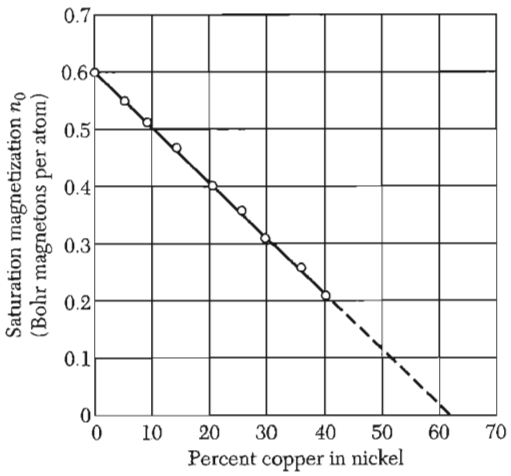


Figure 12 Bohr magneton numbers of nickel-copper alloys.

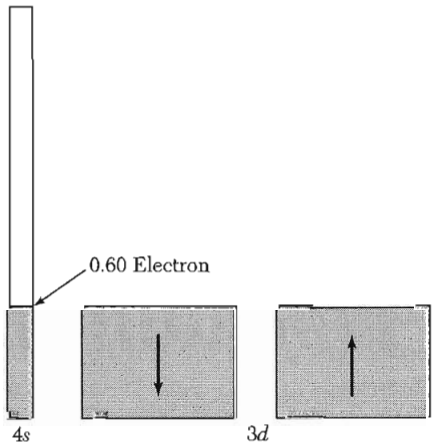


Figure 13 Distribution of electrons in the alloy 60Cu40Ni. The extra 0.6 electron provided by the copper has filled the *d* band entirely and increased slightly the number of electrons in the *s* band with respect to Fig. 12.7b.

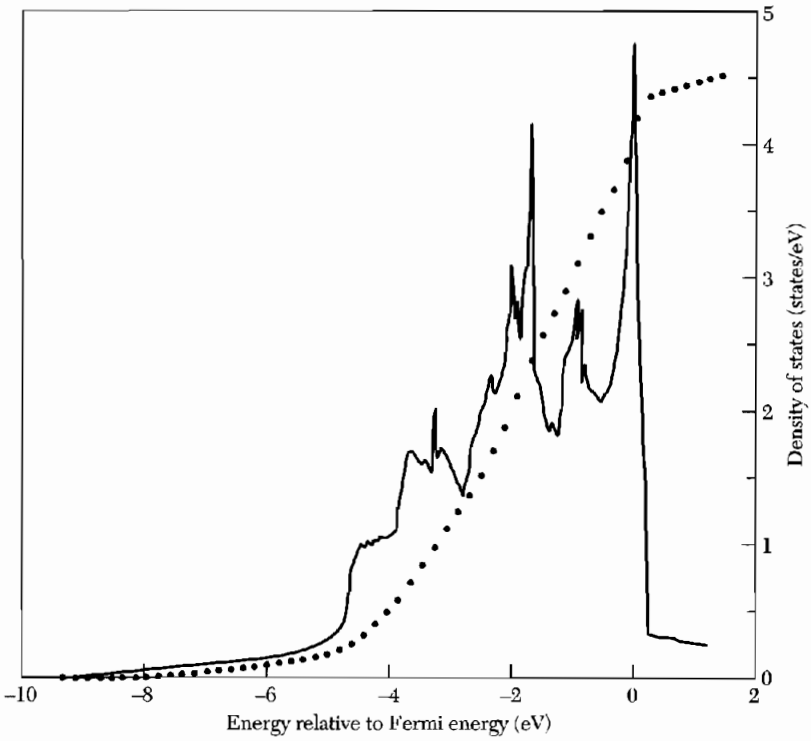


Figure 14 Density of states in nickel. (V. L. Moruzzi, J. F. Janak, and A. R. Williams.)

For simplicity the block drawings represent the density of states as uniform in energy. The actual density is known to be far from uniform; the result of a modern calculation is shown in Fig. 14 for nickel. The width of the 3*d* band is about 5 eV. At the top, where the magnetic effects are determined, the density of states is particularly high. The average density of states is an order of magnitude higher in the 3*d* band than in the 4*s* band. This enhanced density of states ratio gives a rough indication of the expected enhancement of the electronic heat capacity and of the paramagnetic susceptibility in the nonferromagnetic transition metals as compared with the simple monovalent metals.

Figure 15 shows the effect of the addition of small amounts of other elements to nickel. On the band model an alloying metal with z valence electrons outside a filled d shell is expected to decrease the magnetization of nickel by approximately z Bohr magnetons per solute atom. This simple relation holds well for Sn, Al, Zn, and Cu, with $z = 4$, 3, 2, and 1, respectively. For Co, Fe, and Mn the localized moment model of Friedel accounts for effective z values of -1, -2, and -3, respectively.

The average atomic magnetic moments of binary alloys of the elements in the iron group are plotted in Fig. 16 as a function of the concentration of

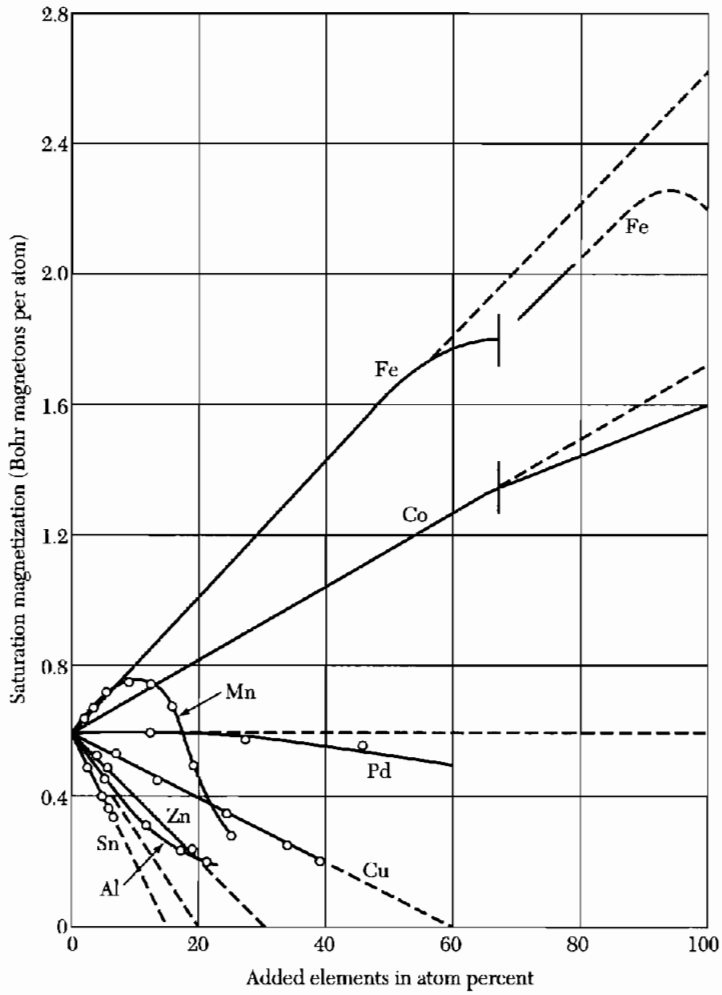


Figure 15 Saturation magnetization of nickel alloys in Bohr magnetons per atom as a function of the atomic percent of solute element.

electrons outside the 3*p* shell. This is called a Slater-Pauling plot. The main sequence of alloys on the right-hand branch follows the rules discussed in connection with Fig. 15. As the electron concentration is decreased, a point is reached at which neither of the 3*d* sub-bands is entirely filled, and the magnetic moment then decreases toward the left-hand side of the plot.

Electrical Conductivity. It might be thought that in the transition metals the availability of the 3*d* band as a path for conduction in parallel with the 4*s* band would increase the conductivity, but this is not the way it works out. The resistivity of the *s* electron path is increased by collisions with the *d* electrons; this is a powerful extra scattering mechanism not available when the *d* band is filled.

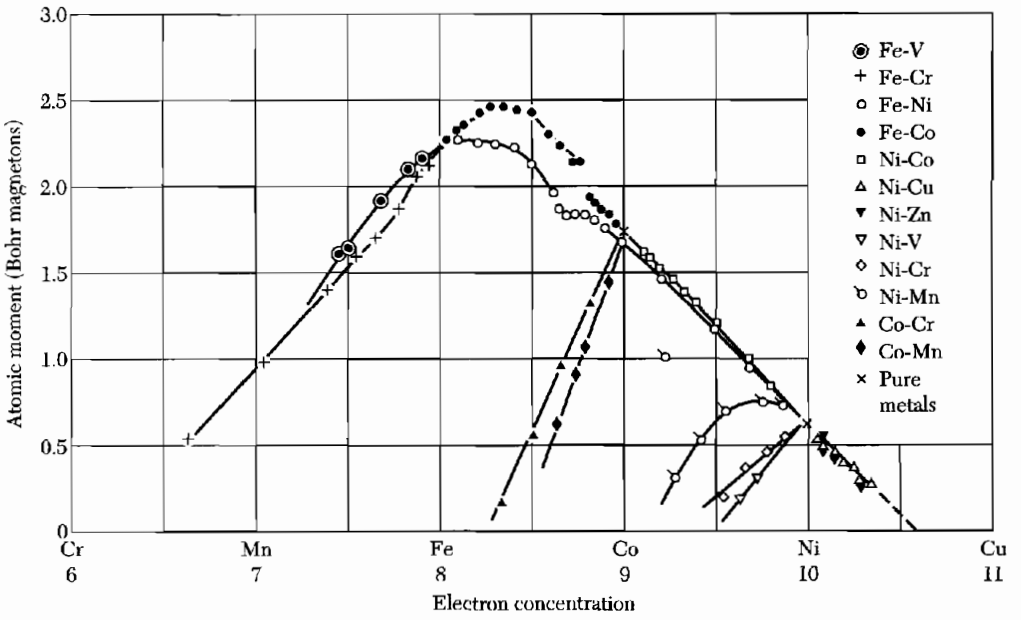


Figure 16 Average atomic moments of binary alloys of the elements in the iron group. (Bozorth.)

We compare the values of the electrical resistivities of Ni, Pd, and Pt in *microhm-cm* at 18°C with that of the noble metals Cu, Ag, and Au immediately following them in the periodic table:

Ni	Pd	Pt
7.4	10.8	10.5

Cu	Ag	Au
1.7	1.6	2.2

The resistivities of the noble metals are lower than those of the transition metals by a factor of the order of 5. This shows the effectiveness of the *s-d* scattering mechanism.

KONDO EFFECT

In dilute solid solutions of a magnetic ion in a nonmagnetic metal crystal (such as Mn in Cu), the exchange coupling between the ion and the conduction electrons has important consequences. The conduction electron gas is

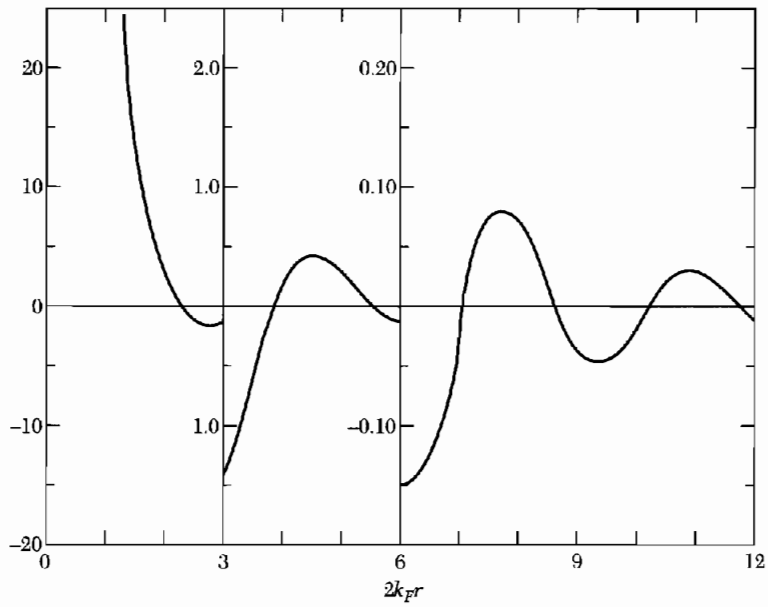


Figure 17 Magnetization of a free electron Fermi gas at $T = 0$ in neighborhood of a point magnetic moment at the origin $r = 0$, according to the RKKY theory. The horizontal axis is $2k_Fr$, where k_F is the wavevector at the Fermi surface. (de Gennes.)

magnetized in the vicinity of the magnetic ion, with the spatial dependence shown in Fig. 17. This magnetization causes an indirect exchange interaction[2] between two magnetic ions, because a second ion perceives the magnetization induced by the first ion. The interaction, known as the Friedel or RKKY interaction, also plays a role in the magnetic spin order of the rare-earth metals, where the spins of the 4*f* ion cores are coupled together by the magnetization induced in the conduction electron gas.

A consequence of the magnetic ion-conduction electron interaction is the **Kondo effect**, discussed in a different context in Chapter 18. A minimum in the electrical resistivity-temperature curve of dilute magnetic alloys at low temperatures has been observed in alloys of Cu, Ag, Au, Mg, Zn with Cr, Mn, and Fe as impurities, among others.

The occurrence of a resistance minimum is connected with the existence of localized magnetic moments on the impurity atoms. Where a resistance minimum is found, there is inevitably a local moment. Kondo showed that the anomalously high scattering probability of magnetic ions at low temperatures

[2]A review of indirect exchange interactions in metals is given by C. Kittel, Solid state physics **22**, 1 (1968); a review of the Kondo effect is given by J. Kondo, "Theory of dilute magnetic alloys," Solid state physics **23**, 184 (1969) and A. J. Heeger, "Localized moments and nonmoments in metals: the Kondo effect," Solid state physics **23**, 248 (1969). The notation RKKY stands for Ruderman, Kittel, Kasuya, and Yosida.

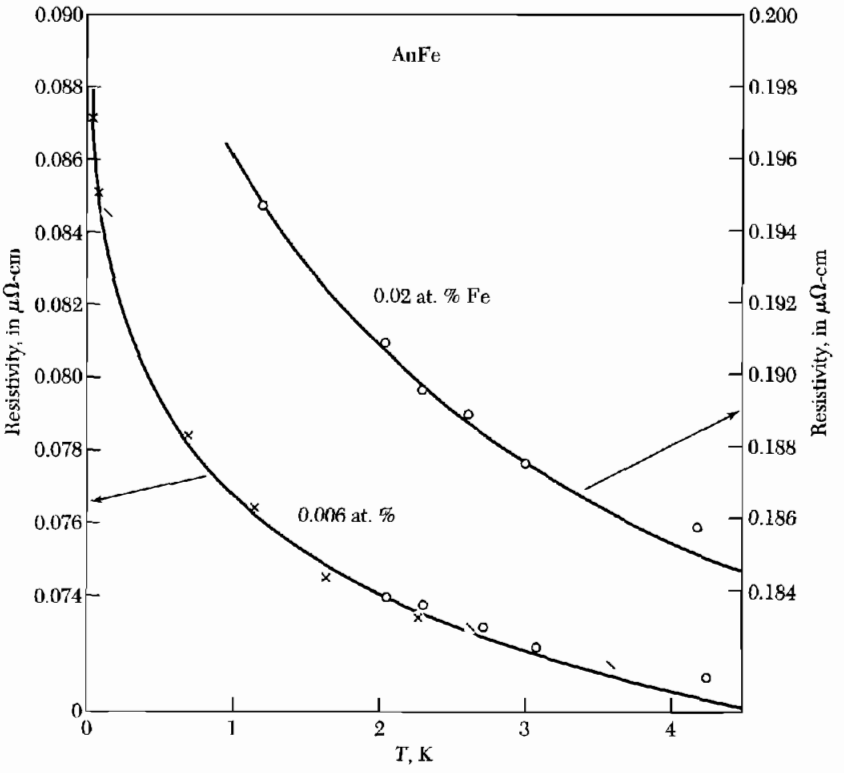


Figure 18 A comparison of experimental and theoretical results for the increase of electrical resistivity at low temperatures in dilute alloys of iron in gold. The resistance minimum lies to the right of the figure, for the resistivity increases at high temperatures because of scattering of electrons by thermal phonons. The experiments are due to D. K. C. MacDonald, W. B. Pearson, and I. M. Templeton; the theory is by J. Kondo. An exact solution was given by K. Wilson.

is a consequence of the dynamic nature of the scattering by the exchange coupling and of the sharpness of the Fermi surface at low temperatures. The temperature region in which the Kondo effect is important is shown in Fig. 18.

The central result is that the spin-dependent contribution to the resistivity is

$$\rho_{\text{spin}} = c\rho_M \left[1 + \frac{3zJ}{\epsilon_F} \ln T \right] = c\rho_0 - c\rho_1 \ln T \ , \tag{12}$$

where J is the exchange energy; z the number of nearest neighbors; c the concentration; and ρ_M is a measure of the strength of the exchange scattering. We see that the spin resistivity increases toward low temperatures if J is negative. If the phonon contribution to the electrical resistivity goes as T^5 in the region of interest and if the resistivities are additive, then the total resistivity has the form

$$\rho = \alpha T^5 + c\rho_0 - c\rho_1 \ln T \ , \tag{13}$$

with a minimum at

$$d\rho/dT = 5aT^4 - c\rho_1/T = 0 \ , \tag{14}$$

whence

$$T_{\min} = (c\rho_1/5a)^{1/5} \ . \tag{15}$$

The temperature at which the resistivity is a minimum varies as the one-fifth power of the concentration of the magnetic impurity, in agreement with experiment, at least for Fe in Cu.

Problems

1. ***Superlattice lines in Cu_3Au.*** Cu_3Au alloy (75% Cu, 25% Au) has an ordered state below 400°C, in which the gold atoms occupy the 000 positions and the copper atoms the $\frac{1}{2}\frac{1}{2}0$, $\frac{1}{2}0\frac{1}{2}$, and $0\frac{1}{2}\frac{1}{2}$ positions in a face-centered cubic lattice. Give the indices of the new x-ray reflections that appear when the alloy goes from the disordered to the ordered state. List all new reflections with indices ≤ 2.

2. ***Configurational heat capacity.*** Derive an expression in terms of $P(T)$ for the heat capacity associated with order/disorder effects in an AB alloy. [The entropy (8) is called the configurational entropy or entropy of mixing.]

APPENDIX A: TEMPERATURE DEPENDENCE OF THE REFLECTION LINES

. . . I came to the conclusion that the sharpness of the interference lines would not suffer but that their intensity should diminish with increasing angle of scattering, the more so the higher the temperature.

P. Debye

As the temperature of the crystal is increased, the intensity of the Bragg-reflected beams decreases, but the angular width of the reflected line does not change. Experimental intensities for aluminum are shown in Fig. 1. It is surprising that we can get a sharp x-ray reflection from atoms undergoing large amplitude random thermal motion, with instantaneous nearest-neighbor spacings differing by 10 percent at room temperature. Before the Laue experiment

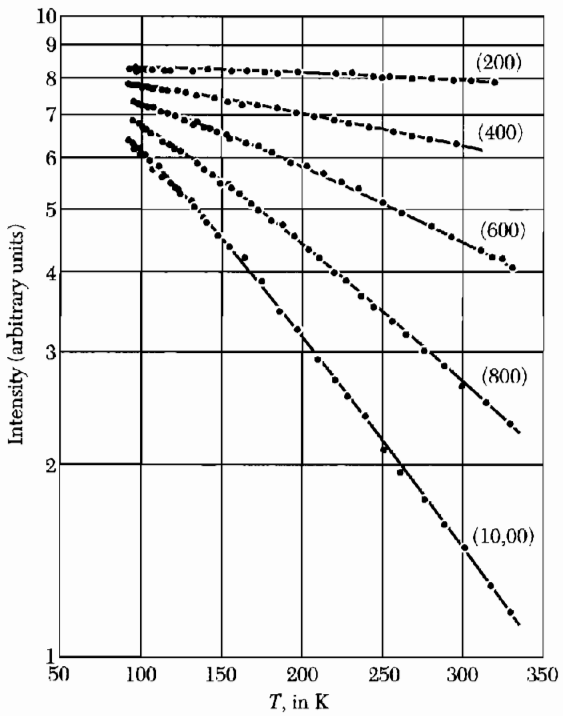


Figure 1 The dependence of intensity on temperature for the (h00) x-ray reflections of aluminum. Reflections (h00) with h odd are forbidden for an fcc structure. (After R. M. Nicklow and R. A. Young.)

was done, but when the proposal was discussed[1] in a coffee house in Munich, the objection was made that the instantaneous positions of the atoms in a crystal at room temperature are far from a regular periodic array, because of the large thermal fluctuation. Therefore, the argument went, one should not expect a well-defined diffracted beam.

But such a beam is found. The reason was given by Debye. Consider the radiation amplitude scattered by a crystal: let the position of the atom nominally at $\mathbf{r}_j$ contain a term $\mathbf{u}(t)$ fluctuating in time: $\mathbf{r}(t) = \mathbf{r}_j + \mathbf{u}(t)$. We suppose each atom fluctuates independently about its own equilibrium position.[2] Then the thermal average of the structure factor (2.43) contains terms

$$f_j \exp(-i\mathbf{G}\cdot\mathbf{r}_j)\langle\exp(-i\mathbf{G}\cdot\mathbf{u})\rangle \ , \tag{1}$$

where $\langle\cdots\rangle$ denotes thermal average. The series expansion of the exponential is

$$\langle\exp(-i\mathbf{G}\cdot\mathbf{u})\rangle = 1 - i\langle\mathbf{G}\cdot\mathbf{u}\rangle - \tfrac{1}{2}\langle(\mathbf{G}\cdot\mathbf{u})^2\rangle + \cdots \ . \tag{2}$$

But $\langle\mathbf{G}\cdot\mathbf{u}\rangle = 0$, because $\mathbf{u}$ is a random thermal displacement uncorrelated with the direction of $\mathbf{G}$. Further,

$$\langle(\mathbf{G}\cdot\mathbf{u})^2\rangle = G^2\langle u^2\rangle\langle\cos^2\theta\rangle = \tfrac{1}{3}\langle u^2\rangle G^2 \ .$$

The factor $\frac{1}{3}$ arises as the geometrical average of $\cos^2\theta$ over a sphere.

The function

$$\exp(-\tfrac{1}{6}\langle u^2\rangle G^2) = 1 - \tfrac{1}{6}\langle u^2\rangle G^2 + \cdots \tag{3}$$

has the same series expansion as (2) for the first two terms shown here. For a harmonic oscillator all terms in the series (2) and (3) can be shown to be identical. Then the scattered intensity, which is the square of the amplitude, is

$$I = I_0 \exp(-\tfrac{1}{3}\langle u^2\rangle G^2) \ , \tag{4}$$

where I_0 is the scattered intensity from the rigid lattice. The exponential factor is the **Debye-Waller factor**.

Here $\langle u^2\rangle$ is the mean square displacement of an atom. The thermal average potential energy $\langle U\rangle$ of a classical harmonic oscillator in three dimensions is $\frac{3}{2}k_BT$, whence

$$\langle U\rangle = \tfrac{1}{2}C\langle u^2\rangle = \tfrac{1}{2}M\omega^2\langle u^2\rangle = \tfrac{3}{2}k_BT \ , \tag{5}$$

[1] P. P. Ewald, private communication.

[2] This is the Einstein model of a solid; it is not a very good model at low temperatures, but it works well at high temperatures.

where C is the force constant, M is the mass of an atom, and ω is the frequency of the oscillator. We have used the result $\omega^2 = C/M$. Thus the scattered intensity is

$$I(hkl) = I_0 \exp(-k_B T G^2/M\omega^2) \ , \tag{6}$$

where hkl are the indices of the reciprocal lattice vector $\mathbf{G}$. This classical result is a good approximation at high temperatures.

For quantum oscillators $\langle u^2 \rangle$ does not vanish even at $T = 0$; there is zero-point motion. On the independent harmonic oscillator model the zero-point energy is $\frac{3}{2}\hbar\omega$; this is the energy of a three-dimensional quantum harmonic oscillator in its ground state referred to the classical energy of the same oscillator at rest. Half of the oscillator energy is potential energy, so that in the ground state

$$\langle U \rangle = \tfrac{1}{2} M\omega^2 \langle u^2 \rangle = \tfrac{3}{4}\hbar\omega \ ; \qquad \langle u^2 \rangle = 3\hbar/2M\omega \ , \tag{7}$$

whence, by (4),

$$I(hkl) = I_0 \exp(-\hbar G^2/2M\omega) \tag{8}$$

at absolute zero. If $G = 10^9$ cm^{-1}, $\omega = 10^{14}$ s^{-1}, and $M = 10^{-22}$ g, the argument of the exponential is approximately 0.1, so that $I/I_0 \simeq 0.9$. At absolute zero, 90 percent of the beam is elastically scattered and 10 percent is inelastically scattered.

We see from (6) and from Fig. 1 that the intensity of the diffracted line decreases, but not catastrophically, as the temperature is increased. Reflections of low G are affected less than reflections of high G. The intensity we have calculated is that of the coherent diffraction (or the elastic scattering) in the well-defined Bragg directions. The intensity lost from these directions is the inelastic scattering and appears as a diffuse background. In inelastic scattering the x-ray photon causes the excitation or de-excitation of a lattice vibration, and the photon changes direction and energy.

At a given temperature the Debye-Waller factor of a diffraction line decreases with an increase in the magnitude of the reciprocal lattice vector $\mathbf{G}$ associated with the reflection. The larger $|\mathbf{G}|$ is, the weaker the reflection at high temperatures. The theory we have worked out here for x-ray reflection applies equally well to neutron diffraction and to the **Mössbauer effect**, the recoilless emission of gamma rays by nuclei bound in crystals.

X-rays can be absorbed in a crystal also by the inelastic processes of photoionization of electrons and Compton scattering. In the photoeffect the x-ray photon is absorbed and an electron is ejected from an atom. In the Compton effect the photon is scattered inelastically by an electron: the photon loses energy and the electron is ejected from an atom. The depth of penetration of the x-ray beam depends on the solid and on the photon energy, but 1 cm is

typical. A diffracted beam in Bragg reflection may remove the energy in a much shorter distance, perhaps 10^{-3} cm in an ideal crystal.

APPENDIX B: EWALD CALCULATION OF LATTICE SUMS

The problem is to calculate the electrostatic potential experienced by one ion in the presence of all the other ions in the crystal. We consider a lattice made up of ions with positive or negative charges and shall assume that the ions are spherical.

We compute the total potential $\varphi = \varphi_1 + \varphi_2$ at an ion as the sum of two distinct but related potentials. The potential φ_1 is that of a structure with a Gaussian distribution of charge situated at each ion site, with signs the same as those of the real ions. According to the definition of the Madelung constant, the charge distribution on the reference point is not considered to contribute to the potential φ_1 or φ_2 (Fig. 1a). We therefore calculate the potential φ_1 as the difference

$$\varphi_1 = \varphi_a - \varphi_b$$

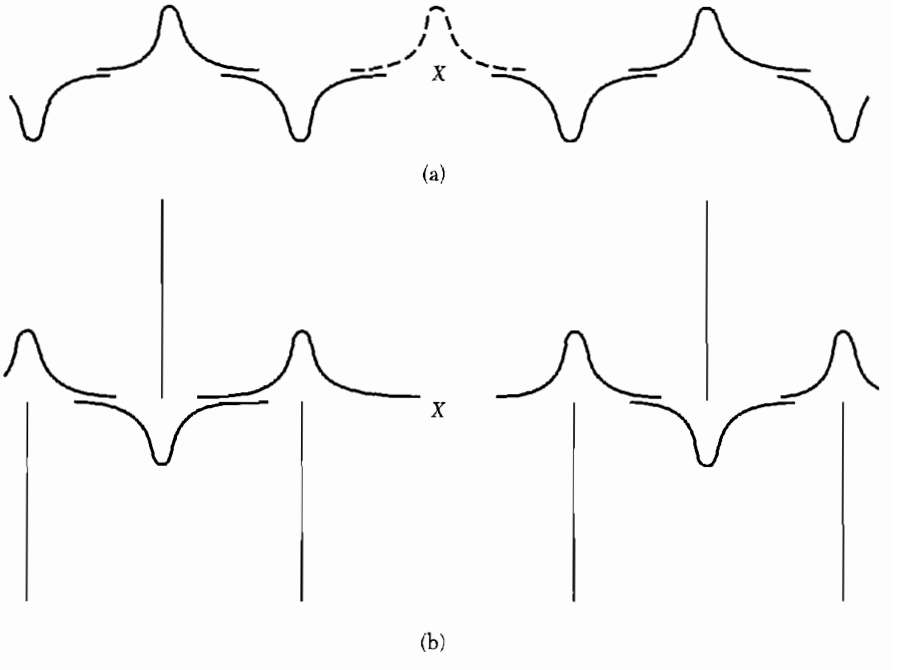


Figure 1 (a) Charge distribution used for computing potential φ_1; the potential φ_a is computed (it includes the dashed curve at the reference point), while φ_b is the potential of the dashed curve alone. (b) Charge distribution for potential φ_2. The reference point is denoted by an *X*.

of two potentials, φ_a being the potential of a continuous series of Gaussian distributions and φ_b being the potential of the single Gaussian distribution on the reference point.

The potential φ_2 is that of a lattice of point charges with an additional Gaussian distribution of opposite sign superposed upon the point charges (Fig. 1b).

The point of splitting the problem into the two parts φ_1 and φ_2 is that by a suitable choice of the parameter determining the width of each Gaussian peak we can get very good convergence of both parts at the same time. The Gaussian distributions drop out completely on taking the sum of the separate charge distributions giving rise to φ_1 and φ_2, so that the value of the total potential φ is independent of the width parameter, but the rapidity of convergence depends on the value chosen for that parameter.

We calculate first the potential φ_a of a continuous Gaussian distribution. We expand φ_a and the charge density ρ in Fourier series:

$$\varphi_a = \sum_{\mathbf{G}} c_{\mathbf{G}} \exp(i\mathbf{G} \cdot \mathbf{r}) \ ; \tag{1}$$

$$\rho = \sum_{\mathbf{G}} \rho_{\mathbf{G}} \exp(i\mathbf{G} \cdot \mathbf{r}) \ , \tag{2}$$

where $\mathbf{G}$ is 2π times a vector in the reciprocal lattice. The Poisson equation is

$$\nabla^2 \varphi_a = -4\pi\rho \ ,$$

or

$$\Sigma\, G^2 c_{\mathbf{G}} \exp(i\mathbf{G} \cdot \mathbf{r}) = 4\pi\, \Sigma\, \rho_{\mathbf{G}} \exp(i\mathbf{G} \cdot \mathbf{r}) \ ,$$

so that

$$c_{\mathbf{G}} = 4\pi\rho_{\mathbf{G}}/G^2 \ . \tag{3}$$

We suppose in finding $\rho_{\mathbf{G}}$ that there is associated with each lattice point of the Bravais lattice a basis containing ions of charge q_t at positions $\mathbf{r}_t$ relative to the lattice point. Each ion point is therefore the center of a Gaussian charge distribution of density

$$\rho(\mathbf{r}) = q_t(\eta/\pi)^{3/2} \exp(-\eta r^2) \ ,$$

where the factor in front of the exponential ensures that the total charge associated with the ion is q_t; the range parameter η is to be chosen judiciously to ensure rapid convergence of the final result (6), which is in value independent of η.

We would normally evaluate $\rho_{\mathbf{G}}$ by multiplying both sides of (2) by $\exp(-i\mathbf{G} \cdot \mathbf{r})$ and integrating over the volume Δ of one cell, so that the charge distribution to be considered is that originating on the ion points within the cell and also that of the tails of the distributions originating in all other cells. It

is easy to see, however, that the integral of the total charge density times $\exp[-(i\mathbf{G}\cdot\mathbf{r})]$ over a single cell is equal to the integral of the charge density originating in a single cell times $\exp[-(i\mathbf{G}\cdot\mathbf{r})]$ over all space.

We have therefore

$$\rho_G \int_{\substack{\text{one}\\\text{cell}}} \exp(i\mathbf{G}\cdot\mathbf{r})\exp(-i\mathbf{G}\cdot\mathbf{r})\,d\mathbf{r} = \rho_G\Delta$$

$$= \int_{\substack{\text{all}\\\text{space}}} \sum_t q_t(\eta/\pi)^{3/2}\exp[-\eta(r-r_t)^2]\exp(-i\mathbf{G}\cdot\mathbf{r})\,d\mathbf{r} \ .$$

This expression is readily evaluated:

$$\rho_G\Delta = \sum_t q_t\exp(-i\mathbf{G}\cdot\mathbf{r}_t)\,(\eta/\pi)^{3/2} \int_{\substack{\text{all}\\\text{space}}} \exp[-(i\mathbf{G}\cdot\boldsymbol{\xi}+\eta\xi^2)]\,d\xi$$

$$= \left(\sum_t q_t\exp(-i\mathbf{G}\cdot\mathbf{r}_t)\right)\exp(-G^2/4\eta) = S(\mathbf{G})\exp(-G^2/4\eta) \ ,$$

where $S(\mathbf{G}) = \sum_t q_t\exp(-i\mathbf{G}\cdot\mathbf{r}_t)$ is just the structure factor (Chapter 2) in appropriate units. Using (1) and (3),

$$\varphi_a = \frac{4\pi}{\Delta}\sum_{\mathbf{G}} S(\mathbf{G})G^{-2}\exp(i\mathbf{G}\cdot\mathbf{r} - G^2/4\eta) \ . \tag{4}$$

At the origin $\mathbf{r} = 0$ we have

$$\varphi_a = \frac{4\pi}{\Delta}\sum_{\mathbf{G}} S(\mathbf{G})G^{-2}\exp(-G^2/4\eta) \ .$$

The potential φ_b at the reference ion point i due to the central Gaussian distribution is

$$\varphi_b = \int_0^\infty (4\pi r^2\,dr)(\rho/r) = 2q_i(\eta/\pi)^{1/2} \ ,$$

and so

$$\varphi_1(i) = \frac{4\pi}{\Delta}\sum_{\mathbf{G}} S(\mathbf{G})G^{-2}\exp(-G^2/4\eta) - 2q_i(\eta/\pi)^{1/2} \ .$$

The potential φ_2 is to be evaluated at the reference point, and it differs from zero because other ions have the tails of their Gaussian distributions overlapping the reference point. The potential is due to three contributions from each ion point:

$$q_l\left[\frac{1}{r_l} - \frac{1}{r_l}\int_0^{r_l}\rho(\mathbf{r})\,d\mathbf{r} - \int_{r_l}^\infty \frac{\rho(\mathbf{r})}{r}\,d\mathbf{r}\right] \ ,$$

where the terms are from the point charge, from the part of the Gaussian distribution lying inside a sphere of radius r_l about the lth ion point, and from that part lying outside the sphere, respectively. On substituting for $\rho(\mathbf{r})$ and carrying out elementary manipulations, we have

$$\varphi_2 = \sum_l \frac{q_l}{r_l} F(\eta^{1/2} r_l) \ , \tag{5}$$

where

$$F(x) = (2/\pi^{1/2}) \int_x^\infty \exp(-s^2)\, ds \ .$$

Finally,

$$\varphi(i) = \frac{4\pi}{\Delta} \sum_{\mathbf{G}} S(\mathbf{G}) G^{-2} \exp(-G^2/4\eta) - 2q_l(\eta/\pi)^{1/2} + \sum_l \frac{q_l}{r_l} F(\eta^{1/2} r_l) \tag{6}$$

is the desired total potential of the reference ion i in the field of all the other ions in the crystal. In the application of the Ewald method the trick is to choose η such that both sums in (6) converge rapidly.

Ewald-Kornfeld Method for Lattice Sums for Dipole Arrays

Kornfeld extended the Ewald method to dipolar and quadrupolar arrays. We discuss here the field of a dipole array at a point which is not a lattice point. According to (4) and (5) the potential at a point $\mathbf{r}$ in a lattice of positive unit point charges is

$$\varphi = (4\pi/\Delta) \sum_{\mathbf{G}} S(\mathbf{G}) G^{-2} \exp[i\mathbf{G} \cdot \mathbf{r} - G^2/4\eta] + \sum_l F(\sqrt{\eta}\, r_l)/r_l \ , \tag{7}$$

where r_l is the distance from $\mathbf{r}$ to the lattice point l.

The first term on the right gives the potential of the charge distribution $\rho = (\eta/\pi)^{3/2} \exp(-\eta r^2)$ about each lattice point. By a well-known relation in electrostatics we obtain the potential of an array of unit dipoles pointing in the z direction by taking $-d/dz$ of the above potential. The term under discussion contributes

$$-(4\pi i/\Delta) \sum_{\mathbf{G}} S(\mathbf{G})(G_z/G^2) \exp[i\mathbf{G} \cdot \mathbf{r} - G^2/4\eta] \ ,$$

and the z component of the electric field from this term is $E_z = \partial^2\varphi/\partial z^2$, or

$$-(4\pi/\Delta) \sum_{\mathbf{G}} S(\mathbf{G})(G_z^2/G^2) \exp[i\mathbf{G} \cdot \mathbf{r} - G^2/4\eta] \ . \tag{8}$$

The second term on the right of (7) after one differentiation gives

$$-\sum_l z_l[(F(\sqrt{\eta} r_l)/r_l^3) + (2/r_l^2)(\eta/\pi)^{1/2} \exp(-\eta r_l^2)] \ ,$$

and the z component of this part of the field is

$$\sum_l [z_l^2[(3F(\sqrt{\eta}r_l)/r_l^5) + (6/r_l^4)(\eta/\pi)^{1/2}\exp(-\eta r_l^2) + (4/r_l^2)(\eta^3/\pi)^{1/2}\exp(-\eta r_l^2)] - [(F(\sqrt{\eta}r_l)/r_l^3) + (2/r_l^2)(\eta/\pi)^{1/2}\exp(-\eta r_l^2)]] \quad (9)$$

The total E_z is given by the sum of (8) and (9). The effects of any number of lattices may be added.

APPENDIX C: QUANTIZATION OF ELASTIC WAVES: PHONONS

Phonons were introduced in Chapter 4 as quantized elastic waves. How do we quantize an elastic wave? As a simple model of phonons in a crystal, consider the vibrations of a linear lattice of particles connected by springs. We can quantize the particle motion exactly as for a harmonic oscillator or set of coupled harmonic oscillators. To do this we make a transformation from particle coordinates to phonon coordinates, also called wave coordinates because they represent a traveling wave.

Let N particles of mass M be connected by springs of force constant C and length a. To fix the boundary conditions, let the particles form a circular ring. We consider the transverse displacements of the particles out of the plane of the ring. The displacement of particle s is q_s and its momentum is p_s. The Hamiltonian of the system is

$$H = \sum_{s=1}^{n} \left\{ \frac{1}{2M} p_s^2 + \frac{1}{2} C(q_{s+1} - q_s)^2 \right\} . \quad (1)$$

The Hamiltonian of a harmonic oscillator is

$$H = \frac{1}{2M} p^2 + \frac{1}{2} Cx^2 , \quad (2)$$

and the energy eigenvalues are, where $n = 0, 1, 2, 3, \ldots$,

$$\epsilon_n = \left(n + \frac{1}{2}\right)\hbar\omega . \quad (3)$$

The eigenvalue problem is also exactly solvable for a chain with the different Hamiltonian (1).

To solve (1) we make a Fourier transformation from the coordinates p_s, q_s to the coordinates P_k, Q_k, which are known as phonon coordinates.

Phonon Coordinates

The transformation from the particle coordinates q_s to the phonon coordinates Q_k is used in all periodic lattice problems. We let

$$q_s = N^{-1/2} \sum_k Q_k \exp(iksa) \ , \tag{4}$$

consistent with the inverse transformation

$$Q_k = N^{-1/2} \sum_s q_s \exp(-iksa) \ . \tag{5}$$

Here the N values of the wavevector k allowed by the periodic boundary condition $q_s = q_{s+N}$ are given by:

$$k = 2\pi n/Na \ ; \ n = 0, \pm 1, \pm 2, \ldots, \pm\left(\frac{1}{2}N - 1\right), \frac{1}{2}N \ . \tag{6}$$

We need the transformation from the particle momentum p_s to the momentum P_k that is canonically conjugate to the coordinate Q_k. The transformation is

$$p_s = N^{-1/2} \sum_k P_k \exp(-iksa); \quad P_k = N^{-1/2} \sum_s p_s \exp(iksa) \ . \tag{7}$$

This is not quite what one would obtain by the naive substitution of p for q and P for Q in (4) and (5), because k and $-k$ have been interchanged between (4) and (7).

We verify that our choice of P_k and Q_k satisfies the quantum commutation relation for canonical variables. We form the commutator

$$\begin{aligned} [Q_k,P_{k'}] &= N^{-1}\left[\sum_r q_r \exp(-ikra), \sum_s p_s \exp(ik'sa)\right] \\ &= N^{-1} \sum_r \sum_s [q_r,p_s] \exp[-i(kr - k's)a] \ . \end{aligned} \tag{8}$$

Because the operators q, p are conjugate, they satisfy the commutation relation

$$[q_r,p_s] = i\hbar\delta(r,s) \ , \tag{9}$$

where $\delta(r,s)$ is the Kronecker delta symbol.

Thus (8) becomes

$$[Q_k,P_{k'}] = N^{-1} i\hbar \sum_r \exp[-i(k - k')ra] = i\hbar\delta(k,k') \ , \tag{10}$$

so that Q_k, P_k also are conjugate variables. Here we have evaluated the summation as

$$\begin{aligned} \sum_r \exp[-i(k - k')ra] &= \sum_r \exp[-i2\pi(n - n')r/N] \\ &= N\delta(n,n') = N\delta(k,k') \ , \end{aligned} \tag{11}$$

where we have used (6) and a standard result for the finite series in (11).

We carry out the transformations (7) and (4) on the hamiltonian (1), and make use of the summation (11):

$$\begin{aligned}\sum_s p_s^2 &= N^{-1}\sum_s\sum_k\sum_{k'} P_kP_{k'}\exp[-i(k+k')sa] \\ &= \sum_k\sum_{k'} P_kP_k\delta(-k,k') = \sum_k P_kP_{-k}\ ;\end{aligned} \tag{12}$$

$$\begin{aligned}\sum_s (q_{s+1}-q_s)^2 &= N^{-1}\sum_s\sum_k\sum_{k'} Q_kQ_{k'}\exp(iksa)[\exp(ika)-1] \\ &\quad\times\exp(ik'sa)[\exp(ik'a)-1] = 2\sum_k Q_kQ_{-k}(1-\cos ka)\ .\end{aligned} \tag{13}$$

Thus the hamiltonian (1) becomes, in phonon coordinates,

$$H = \sum_k\left\{\frac{1}{2M}P_kP_{-k} + CQ_kQ_{-k}(1-\cos ka)\right\}\ . \tag{14}$$

If we introduce the symbol ω_k defined by

$$\omega_k \equiv (2C/M)^{1/2}(1-\cos ka)^{1/2}\ , \tag{15}$$

we have the phonon hamiltonian in the form

$$H = \sum_k\left\{\frac{1}{2M}P_kP_{-k} + \frac{1}{2}M\omega_k^2\,Q_kQ_{-k}\right\}\ . \tag{16}$$

The equation of motion of the phonon coordinate operator Q_k is found by the standard prescription of quantum mechanics:

$$i\hbar\dot{Q}_k = [Q_k, H] = i\hbar P_{-k}/M\ , \tag{17}$$

with H given by (14). Further, using the commutator (17),

$$i\hbar\ddot{Q}_k = [\dot{Q}_k, H] = M^{-1}[P_{-k},H] = i\hbar\omega_k^2 Q_k\ , \tag{18}$$

so that

$$\ddot{Q}_k + \omega_k^2 Q_k = 0\ . \tag{19}$$

This is the equation of motion of a harmonic oscillator with the frequency ω_k.

The energy eigenvalues of a quantum harmonic oscillator are

$$\epsilon_k = \left(n_k + \frac{1}{2}\right)\hbar\omega_k\ , \tag{20}$$

where the quantum number $n_k = 0, 1, 2, \ldots$. The energy of the entire system of all phonons is

$$U = \sum_k\left(n_k + \frac{1}{2}\right)\hbar\omega_k\ . \tag{21}$$

This result demonstrates the quantization of the energy of elastic waves on a line.

Creation and Annihilation Operators

It is helpful in advanced work to transform the phonon hamiltonian (16) into the form of a set of harmonic oscillators:

$$H = \sum_k \hbar\omega_k\left(a_k^+ a_k + \frac{1}{2}\right) . \tag{22}$$

Here a_k^+, a_k are harmonic oscillator operators, also called creation and destruction operators or boson operators. The transformation is derived below.

The boson creation operator a^+ which "creates a phonon" is defined by the property

$$a^+|n\rangle = (n+1)^{1/2}|n+1\rangle , \tag{23}$$

when acting on a harmonic oscillator state of quantum number n, and the boson annihilation operator a which "destroys a phonon" is defined by the property

$$a|n\rangle = n^{1/2}|n-1\rangle . \tag{24}$$

It follows that

$$a^+a|n\rangle = a^+n^{1/2}|n-1\rangle = n|n\rangle , \tag{25}$$

so that $|n\rangle$ is an eigenstate of the operator a^+a with the integral eigenvalue n, called the quantum number or occupancy of the oscillator. When the phonon mode k is in the eigenstate labeled by n_k, we may say that there are n_k phonons in the mode. The eigenvalues of (22) are $U = \Sigma\,(n_k + \frac{1}{2})\hbar\omega_k$, in agreement with (21).

Because

$$aa^+|n\rangle = a(n+1)^{1/2}|n+1\rangle = (n+1)|n\rangle , \tag{26}$$

the commutator of the boson wave operators a_k^+ and a_k satisfies the relation

$$[a, a^+] \equiv aa^+ - a^+a = 1 . \tag{27}$$

We still have to prove that the hamiltonian (16) can be expressed as (19) in terms of the phonon operators a_k^+, a_k. This can be done by the transformation

$$a_k^+ = (2\hbar)^{-1/2}[(M\omega_k)^{1/2}Q_{-k} - i(M\omega_k)^{-1/2}P_k] ; \tag{28}$$

$$a_k = (2\hbar)^{-1/2}[(M\omega_k)^{1/2}Q_k + i(M\omega_k)^{-1/2}P_{-k}] . \tag{29}$$

The inverse relations are

$$Q_k = (\hbar/2M\omega_k)^{1/2}(a_k + a_{-k}^+) ; \tag{30}$$

$$P_k = i(\hbar M\omega_k/2)^{1/2}(a_k^+ - a_{-k}) . \tag{31}$$

By (4), (5), and (29) the particle position operator becomes

$$q_s = \sum_k (\hbar/2NM\omega_k)^{1/2}[a_k \exp(iks) + a_k^+ \exp(-iks)] . \tag{32}$$

This equation relates the particle displacement operator to the phonon creation and annihilation operators.

To obtain (29) from (28), we use the properties

$$Q_{-k}^{+} = Q_k \ ; \qquad P_k^{+} = P_{-k} \tag{33}$$

which follow from (5) and (7) by use of the quantum mechanical requirement that q_s and p_s be hermitian operators:

$$q_s = q_s^{+} \ ; \qquad p_s = p_s^{+} \ . \tag{34}$$

Then (28) follows from the transformations (4), (5), and (7). We verify that the commutation relation (33) is satisfied by the operators defined by (28) and (29):

$$[a_k, a_k^{+}] = (2\hbar)^{-1}(M\omega_k[Q_k, Q_{-k}] - i[Q_k,P_k] + i[P_{-k},Q_{-k}] + [P_{-k},P_k]/M\omega_k) \ . \tag{35}$$

By use of $[Q_k,P_{k'}] = i\hbar\delta(k,k')$ from (10) we have

$$[a_k, a_{k'}^{+}] = \delta(k, k') \ . \tag{36}$$

It remains to show that the versions of (16) and (22) of the phonon hamiltonian are identical. We note that $\omega_k = \omega_{-k}$ from (15), and we form

$$\hbar\omega_k(a_k^{+}a_k + a_{-k}^{+}a_{-k}) = \frac{1}{2M}(P_kP_{-k} + P_{-k}P_k) + \frac{1}{2}M\omega_k^2(Q_kQ_{-k} + Q_{-k}Q_k) \ .$$

This exhibits the equivalence of the two expressions (14) and (22) for H. We identify $\omega_k = (2C/M)^{1/2}(1 - \cos ka)^{1/2}$ in (15) with the classical frequency of the oscillator mode of wavevector k.

APPENDIX D: FERMI-DIRAC DISTRIBUTION FUNCTION[1]

The Fermi-Dirac distribution function[1] may be derived in several steps by use of a modern approach to statistical mechanics. We outline the argument here. The notation is such that conventional entropy S is related to the fundamental entropy σ by $S = k_B\sigma$, and the Kelvin temperature T is related to the fundamental temperature τ by $\tau = k_BT$, where k_B is the Boltzmann constant with the value 1.38066×10^{-23} J K.

The leading quantities are the entropy, the temperature, the Boltzmann factor, the chemical potential, the Gibbs factor, and the distribution functions. The

[1]This appendix follows closely the introduction to C. Kittel and H. Kroemer, *Thermal Physics*, 2nd ed., Freeman, 1980.

entropy measures the number of quantum states accessible to a system. A closed system might be in any of these quantum states and (we assume) with equal probability. The fundamental assumption is that quantum states are either accessible or inaccessible to the system, and the system is equally likely to be in any one accessible state as in any other accessible state. Given g accessible states, the entropy is defined as $\sigma = \log g$. The entropy thus defined will be a function of the energy U, the number of particles N, and the volume V of the system.

When two systems, each of specified energy, are brought into thermal contact, they may transfer energy; their total energy remains constant, but the constraints on their individual energies are lifted. A transfer of energy in one direction, or perhaps in the other, may increase the product g_1g_2 that measures the number of accessible states of the combined systems. What we call the fundamental assumption biases the outcome in favor of that allocation of the total energy that maximizes the number of accessible states: more is better, and more likely. This statement is the kernel of the law of increase of entropy, which is the general expression of the second law of thermodynamics.

We have brought two systems into thermal contact so that they may transfer energy. What is the most probable outcome of the encounter? One system will gain energy at the expense of the other, and meanwhile the total entropy of the two systems will increase. Eventually the entropy will reach a maximum for the given total energy. It is not difficult to show that the maximum is attained when the value of $(\partial\sigma/\partial U)_{N,V}$ for one system is equal to the value of the same quantity for the second system. This equality property for two systems in thermal contact is the property we expect of the temperature. Accordingly, the fundamental temperature τ is defined by the relation

$$\frac{1}{\tau} \equiv \left(\frac{\partial \sigma}{\partial U}\right)_{N,V} . \tag{1}$$

The use of $1/\tau$ assures that energy will flow from high τ to low τ; no more complicated relation is needed.

Now consider a very simple example of the Boltzmann factor. Let a small system with only two states, one at energy 0 and one at energy ϵ, be placed in thermal contact with a large system that we call the reservoir. The total energy of the combined systems is U_0; when the small system is in the state of energy 0, the reservoir has energy U_0 and will have $g(U_0)$ states accessible to it. When the small system is in the state of energy ϵ, the reservoir will have energy $U_0 - \epsilon$ and will have $g(U_0 - \epsilon)$ states accessible to it. By the fundamental assumption, the ratio of the probability of finding the small system with energy ϵ to the probability of finding it with energy 0 is

$$\frac{P(\epsilon)}{P(0)} = \frac{g(U_0 - \epsilon)}{g(U_0)} = \frac{\exp[\sigma(U_0 - \epsilon)]}{\exp[\sigma(U_0)]} . \tag{2}$$

The reservoir entropy σ may be expanded in a Taylor series:

$$\sigma(U_0 - \epsilon) \simeq \sigma(U_0) - \epsilon(\partial\sigma/\partial U_0) = \sigma(U_0) - \epsilon/\tau \ , \tag{3}$$

by the definition (1) of the temperature. Higher order terms in the expansion may be dropped. Cancellation of the term $\exp[\sigma(U_0)]$, which occurs in the numerator and denominator of (2) after the substitution of (3), leaves us with

$$P(\epsilon)/P(0) = \exp(-\epsilon/\tau) \ . \tag{4}$$

This is Boltzmann's result. To show its use, we calculate the thermal average energy $\langle\epsilon\rangle$ of the two-state system in thermal contact with a reservoir at temperature τ:

$$\langle\epsilon\rangle = \sum_i \epsilon_i P(\epsilon_i) = 0 \cdot P(0) + \epsilon P(\epsilon) = \frac{\epsilon \exp(-\epsilon/\tau)}{1 + \exp(-\epsilon/\tau)} \ , \tag{5}$$

where we have imposed the normalization condition on the sum of the probabilities:

$$P(0) + P(\epsilon) = 1 \ . \tag{6}$$

The argument can be generalized immediately to find the average energy of a harmonic oscillator at temperature τ, as in the Planck law.

The most important extension of the theory is to systems that can transfer particles as well as energy with the reservoir. For two systems in diffusive and thermal contact, the entropy will be a maximum with respect to the transfer of particles as well as to the transfer of energy. Not only must $(\partial\sigma/\partial U)_{N,V}$ be equal for the two systems, but $(\partial\sigma/\partial N)_{U,V}$ must also be equal, where N refers to the number of particles of a given species. The new equality condition is the occasion for the introduction[2] of the chemical potential μ:

$$-\frac{\mu}{\tau} = \left(\frac{\partial\sigma}{\partial N}\right)_{U,V} \ . \tag{7}$$

For two systems in thermal and diffusive contact, $\tau_1 = \tau_2$ and $\mu_1 = \mu_2$. The sign in (7) is chosen to ensure that the direction of particle flow is from high chemical potential to low chemical potential as equilibrium is approached.

The Gibbs factor is an extension of the Boltzmann factor (4) and allows us to treat systems that can transfer particles. The simplest example is a system with two states, one with 0 particles and 0 energy, and one with 1 particle and energy ϵ. The system is in contact with a reservoir at temperature τ and chemical potential μ. We extend (3) for the reservoir entropy:

$$\begin{aligned} \sigma(U_0 - \epsilon; N_0 - 1) &= \sigma(U_0; N_0) - \epsilon(\partial\sigma/\partial U_0) - 1 \cdot (\partial\sigma/\partial N_0) \\ &= \sigma(U_0; N_0) - \epsilon/\tau + \mu/\tau \ . \end{aligned} \tag{8}$$

[2] *TP* Chapter 5 has a careful treatment of the chemical potential.

By analogy with (4), we have the Gibbs factor

$$P(1,\epsilon)/P(0,0) = \exp[(\mu - \epsilon)/\tau] \ , \tag{9}$$

for the ratio of the probability that the system is occupied by 1 particle at energy ϵ to the probability that the system is unoccupied, with energy 0. The result (9) after normalization is readily expressed as

$$P(1,\epsilon) = \frac{1}{\exp[(\epsilon - \mu)/\tau] + 1} \ . \tag{10}$$

This is the Fermi-Dirac distribution function.

APPENDIX E: DERIVATION OF THE *dk/dt* EQUATION

The simple and rigorous derivation that follows is due to Kroemer. In quantum mechanics, for any operator A we have

$$d\langle A\rangle/dt = (i/\hbar)\langle [H, A]\rangle \ , \tag{1}$$

where H is the hamiltonian. See also C. L. Cook, American J. Phys. **55**, 953 (1987).

We let A be the lattice translation operator T defined by

$$Tf(x) = f(x + a) \ , \tag{2}$$

where a is a basis vector, here in one dimension. For a Bloch function

$$T\psi_k(x) = \exp(ika)\psi_k(x) \ . \tag{3}$$

This result is usually written for one band, but it holds even if ψ_k is a linear combination of Bloch states from any number of bands, but having the identical wavevector k in the reduced zone scheme.

The crystal hamiltonian H_0 commutes with the lattice translation operator T, so that $[H_0,T] = 0$. If we add a uniform external force F, then

$$H = H_0 - Fx \ , \tag{4}$$

and

$$[H, T] = FaT \ . \tag{5}$$

From (1) and (5),

$$d\langle T\rangle/dt = (i/\hbar)(Fa)\langle T\rangle \ . \tag{6}$$

From (6) we form

$$\langle T\rangle^* d\langle T\rangle/dt = (iFa/\hbar)|\langle T\rangle|^2 \ ;$$

$$\langle T\rangle d\langle T^*\rangle/dt = -(iFa/\hbar)|\langle T\rangle|^2 \ .$$

On addition,

$$d|\langle T\rangle|^2/dt = 0 \ . \tag{7}$$

This is the equation of a circle in the complex plane. The coordinate axes in the plane are the real and imaginary parts of the eigenvalue $\exp(ika)$. If $\langle T\rangle$ is initially on the unit circle, it will remain on the unit circle.

For ψ's that satisfy periodic boundary conditions, $\langle T\rangle$ can lie on the unit circle only if ψ_k is a single Bloch function or a superposition of Bloch functions from different bands, but with the same reduced k.

As $\langle T\rangle$ moves around the unit circle, the wavevector k changes exactly at the same rate for the components of ψ_k in all bands. With $\langle T\rangle = \exp(ika)$, we have from (6) that

$$ia\, dk/dt = iFa/\hbar \ , \tag{8}$$

or

$$dk/dt = F/\hbar \ , \tag{9}$$

an exact result.

This does not mean that interband mixing (such as Zener tunneling) does not occur under the influence of applied electric fields. It just means that k evolves at a constant rate for every component of a wave packet. The result is easily extended to three dimensions.

APPENDIX F: BOLTZMANN TRANSPORT EQUATION

The classical theory of transport processes is based on the Boltzmann transport equation. We work in the six-dimensional space of Cartesian coordinates $\mathbf{r}$ and velocity $\mathbf{v}$. The classical distribution function $f(\mathbf{r},\mathbf{v})$ is defined by the relation

$$f(\mathbf{r},\mathbf{v})d\mathbf{r}d\mathbf{v} = \text{number of particles in } d\mathbf{r}d\mathbf{v} \ . \tag{1}$$

The Boltzmann equation is derived by the following argument. We consider the effect of a time displacement dt on the distribution function. The Liouville theorem of classical mechanics tells us that if we follow a volume element along a flowline the distribution is conserved:

$$f(t + dt,\mathbf{r} + d\mathbf{r},\mathbf{v} + d\mathbf{v}) = f(t,\mathbf{r},\mathbf{v}) \ , \tag{2}$$

in the absence of collisions. With collisions

$$f(t + dt, \mathbf{r} + d\mathbf{r}, \mathbf{v} + d\mathbf{v}) - f(t, \mathbf{r}, \mathbf{v}) = dt(\partial f/\partial t)_{\text{coll}} \ . \tag{3}$$

Thus

$$dt(\partial f/\partial t) + d\mathbf{r} \cdot \text{grad}_{\mathbf{r}} f + d\mathbf{v} \cdot \text{grad}_{\mathbf{v}} f = dt(\partial f/\partial t)_{\text{coll}} \ . \tag{4}$$

Let α denote the acceleration $d\mathbf{v}/dt$; then

$$\boxed{\partial f/\partial t + \mathbf{v} \cdot \text{grad}_{\mathbf{r}} f + \alpha \cdot \text{grad}_{\mathbf{v}} f = (\partial f/\partial t)_{\text{coll}}} \tag{5}$$

This is the **Boltzmann transport equation**.

In many problems the collision term $(\partial f/\partial t)_{\text{coll}}$ may be treated by the introduction of a relaxation time $\tau_c(\mathbf{r},\mathbf{v})$, defined by the equation

$$(\partial f/\partial t)_{\text{coll}} = -(f - f_0)/\tau_c \ . \tag{6}$$

Here f_0 is the distribution function in thermal equilibrium. Do not confuse τ_c for relaxation time with τ for temperature. Suppose that a nonequilibrium distribution of velocities is set up by external forces which are suddenly removed. The decay of the distribution towards equilibrium is then obtained from (6) as

$$\frac{\partial(f - f_0)}{\partial t} = -\frac{f - f_0}{\tau_c} \ , \tag{7}$$

if we note that $\partial f_0/\partial t = 0$ by definition of the equilibrium distribution. This equation has the solution

$$(f - f_0)_t = (f - f_0)_{t=0} \exp(-t/\tau_c) \ . \tag{8}$$

It is not excluded that τ_c may be a function of $\mathbf{r}$ and $\mathbf{v}$.

We combine (1), (5), and (6) to obtain the Boltzmann transport equation in the relaxation time approximation:

$$\boxed{\frac{\partial f}{\partial t} + \alpha \cdot \text{grad}_{\mathbf{v}} f + \mathbf{v} \cdot \text{grad}_{\mathbf{r}} f = -\frac{f - f_0}{\tau_c} \ .} \tag{9}$$

In the steady state $\partial f/\partial t = 0$ by definition.

Particle Diffusion

Consider an isothermal system with a gradient of the particle concentration. The steady-state Boltzmann transport equation in the relaxation time approximation becomes

$$v_x df/dx = -(f - f_0)/\tau_c \ , \tag{10}$$

where the nonequilibrium distribution function f varies along the x direction. We may write (10) to first order as

$$f_1 = f_0 - v_x \tau_c df_0/dx \ , \tag{11}$$

where we have replaced $\partial f/\partial x$ by df_0/dx. We can iterate to obtain higher order solutions when desired. Thus the second order solution is

$$f_2 = f_0 - v_x \tau_c df_1/dx = f_0 - v_x \tau_c df_0/dx + v_x^2 \tau_c^2 d^2 f_0/dx^2 \ . \tag{12}$$

The iteration may be used in the treatment of nonlinear effects.

Classical Distribution

Let f_0 be the distribution function in the classical limit:

$$f_0 = \exp[(\mu - \epsilon)/\tau] \ . \tag{13}$$

We are at liberty to take whatever normalization for the distribution function is most convenient because the transport equation is linear in f and f_0. We can take the normalization as in (13) rather than as in (1). Then

$$df_0/dx = (df_0/d\mu)(d\mu/dx) = (f_0/\tau)(d\mu/dx) \ , \tag{14}$$

and the first order solution (11) for the nonequilibrium distribution becomes

$$f = f_0 - (v_x \tau_c f_0/\tau)(d\mu/dx) \ . \tag{15}$$

The particle flux density in the x direction is

$$J_n^x = \int v_x f D(\epsilon) d\epsilon \ , \tag{16}$$

where $D(\epsilon)$ is the density of electron states per unit volume per unit energy range:

$$D(\epsilon) = \frac{1}{2\pi^2}\left(\frac{2M}{\hbar^2}\right)^{3/2} \epsilon^{1/2} \ . \tag{17}$$

Thus

$$J_n^x = \int v_x f_0 D(\epsilon) d\epsilon - (d\mu/dx) \int (v_x^2 \tau_c f_0/\tau) D(\epsilon) d\epsilon \ . \tag{18}$$

The first integral vanishes because v_x is an odd function and f_0 is an even function of v_x. This confirms that the net particle flux vanishes for the equilibrium distribution f_0. The second integral will not vanish.

Before evaluating the second integral, we have an opportunity to make use of what we may know about the velocity dependence of the relaxation time τ_c. Only for the sake of example we assume that τ_c is constant, independent of velocity; τ_c may then be taken out of the integral:

$$J_n^x = -(d\mu/dx)(\tau_c/\tau) \int v_x^2 f_0 D(\epsilon) d\epsilon \ . \tag{19}$$

The integral may be written as

$$\tfrac{1}{3}\int v^2 f_0 D(\epsilon)d\epsilon = \frac{2}{3M}\int (\tfrac{1}{2}Mv^2) f_0 D(\epsilon)d\epsilon = n\tau/M \ , \tag{20}$$

because the integral is just the kinetic energy density $\frac{3}{2}n\tau$ of the particles. Here $\int f_0 D(\epsilon)d\epsilon = n$ is the concentration. The particle flux density is

$$J_n^x = -(n\tau_c/M)(d\mu/dx) = -(\tau_c\tau/M)(dn/dx) \ , \tag{21}$$

because

$$\mu = \tau \log n + \text{constant} \ . \tag{22}$$

The result (21) is of the form of the diffusion equation with the diffusivity

$$D_n = \tau_c\tau/M = \tfrac{1}{3}\langle v^2\rangle\tau_c \ . \tag{23}$$

Another possible assumption about the relaxation time is that it is inversely proportional to the velocity, as in $\tau_c = l/v$, where the mean free path l is constant. Instead of (19) we have

$$J_n^x = -(d\mu/dx)(l/\tau)\int (v_x^2/v) f_0 D(\epsilon)d\epsilon \ , \tag{24}$$

and now the integral may be written as

$$\tfrac{1}{3}\int v f_0 D(\epsilon)d\epsilon = \tfrac{1}{3}n\bar{c} \ , \tag{25}$$

where $\bar{c}$ is the average speed. Thus

$$J_n^x = -\tfrac{1}{3}(l\bar{c}n/\tau)(d\mu/dx) = -\tfrac{1}{3}l\bar{c}(dn/dx) \ , \tag{26}$$

and the diffusivity is

$$D_n = \tfrac{1}{3}l\bar{c} \ . \tag{27}$$

Fermi-Dirac Distribution

The distribution function is

$$f_0 = \frac{1}{\exp[(\epsilon - \mu)/\tau] + 1} \ . \tag{28}$$

To form df_0/dx as in (14) we need the derivative $df_0/d\mu$. We argue below that

$$df_0/d\mu = \delta(\epsilon - \mu) \ , \tag{29}$$

at low temperatures $\tau \ll \mu$. Here δ is the Dirac delta function, which has the property for a general function $F(\epsilon)$ that

$$\int_{-\infty}^{\infty} F(\epsilon)\delta(\epsilon - \mu)d\epsilon = F(\mu) \ . \tag{30}$$

Now consider the integral $\int_0^\infty F(\epsilon)(df_0/d\mu)d\epsilon$. At low temperatures $df_0/d\mu$ is very large for $\epsilon \simeq \mu$ and is small elsewhere. Unless the function $F(\epsilon)$ is very rapidly varying near μ, we may take $F(\epsilon)$ outside the integral, with the value $F(\mu)$:

$$\begin{aligned}\int_0^\infty F(\epsilon)(df_0/d\mu)d\epsilon &\simeq F(\mu)\int_0^\infty (df_0/d\mu)d\epsilon = -F(\mu)\int_0^\infty (df_0/d\epsilon)d\epsilon \\ &= -F(\mu)[f_0(\epsilon)]_0^\infty = F(\mu)f_0(0) \ ,\end{aligned} \tag{31}$$

where we have used $df_0/d\mu = -df_0/d\epsilon$. We have also used $f_0 = 0$ for $\epsilon = \infty$. At low temperatures $f(0) \simeq 1$; thus the right-hand side of (31) is just $F(\mu)$, consistent with the delta function approximation. Thus

$$df_0/dx = \delta(\epsilon - \mu)d\mu/dx \ . \tag{32}$$

The particle flux density is, from (16),

$$J_n^x = -(d\mu/dx)\tau_c \int v_x^2 \delta(\epsilon - \mu)D(\epsilon)d\epsilon \ , \tag{33}$$

where τ_c is the relaxation time at the surface $\epsilon = \mu$ of the Fermi sphere. The integral has the value

$$\tfrac{1}{3}v_F^2(3n/2\epsilon_F) = n/m \ , \tag{34}$$

by use of $D(\mu) = 3n/2\epsilon_F$ at absolute zero, where $\epsilon_F \equiv \frac{1}{2}mv_F^2$ defines the velocity v_F on the Fermi surface. Thus

$$J_n^x = -(n\tau_c/m)d\mu/dx \ . \tag{35}$$

At absolute zero $\mu(0) = (\hbar^2/2m)(3\pi^2 n)^{2/3}$, whence

$$\begin{aligned}d\mu/dx &= [\tfrac{2}{3}(\hbar^2/2m)(3\pi^2)^{2/3}/n^{1/3}]dn/dx \\ &= \tfrac{2}{3}(\epsilon_F/n)dn/dx \ ,\end{aligned} \tag{36}$$

so that (33) becomes

$$J_n^x = -(2\tau_c/3m)\epsilon_F\, dn/dx = -\tfrac{1}{3}v_F^{\,2}\tau_c\, dn/dx \ . \tag{37}$$

The diffusivity is the coefficient of dn/dx:

$$D_n = \tfrac{1}{3}v_F^2\tau_c \ , \tag{38}$$

closely similar in form to the result (23) for the classical distribution of velocities. In (38) the relaxation time is to be taken at the Fermi energy.

We see we can solve transport problems where the Fermi-Dirac distribution applies, as in metals, as easily as where the classical approximation applies.

Electrical Conductivity

The isothermal electrical conductivity σ follows from the result for the particle diffusivity when we multiply the particle flux density by the particle charge q and replace the gradient $d\mu/dx$ of the chemical potential by the gradient $qd\varphi/dx = -qE_x$ of the external potential, where E_x is the x component of the electric field intensity. The electric current density follows from (21):

$$\mathbf{J}_q = (nq^2\tau_c/m)\mathbf{E} \; ; \qquad \sigma = nq^2\tau_c/m \; , \tag{39}$$

for a classical gas with relaxation time τ_c. For the Fermi-Dirac distribution, from (35),

$$\mathbf{J}_q = (nq^2\tau_c/m)\mathbf{E} \; ; \qquad \sigma = nq^2\tau_c/m \; . \tag{40}$$

APPENDIX G: VECTOR POTENTIAL, FIELD MOMENTUM, AND GAUGE TRANSFORMATIONS

This section is included because it is hard to find the magnetic vector potential **A** discussed thoroughly in one place, and we need the vector potential in superconductivity. It may seem mysterious that the hamiltonian of a particle in a magnetic field has the form derived in (18) below:

$$H = \frac{1}{2M}\left(\mathbf{p} - \frac{Q}{c}\mathbf{A}\right)^2 + Q\varphi \; , \tag{1}$$

where Q is the charge; M is the mass; **A** is the vector potential; and φ is the electrostatic potential. This expression is valid in classical mechanics and in quantum mechanics. Because the kinetic energy of a particle is not changed by a static magnetic field, it is perhaps unexpected that the vector potential of the magnetic field enters the hamiltonian. As we shall see, the key is the observation that the momentum **p** is the sum of two parts, the kinetic momentum

$$\mathbf{p}_{\text{kin}} = M\mathbf{v} \tag{2}$$

which is familiar to us, and the potential momentum or field momentum

$$\mathbf{p}_{\text{field}} = \frac{Q}{c}\mathbf{A} \; . \tag{3}$$

The total momentum is

$$\boxed{\mathbf{p} = \mathbf{p}_{\text{kin}} + \mathbf{p}_{\text{field}} = M\mathbf{v} + \frac{Q}{c}\mathbf{A}\ ,} \tag{4}$$

and the kinetic energy is

$$\frac{1}{2}Mv^2 = \frac{1}{2M}(Mv)^2 = \frac{1}{2M}\left(\mathbf{p} - \frac{Q}{c}\mathbf{A}\right)^2\ . \tag{5}$$

The vector potential[1] is related to the magnetic field by

$$\mathbf{B} = \text{curl}\,\mathbf{A}\ . \tag{6}$$

We assume that we work in nonmagnetic material so that **H** and **B** are treated as identical.

Lagrangian Equations of Motion

To find the Hamiltonian, the prescription of classical mechanics is clear: we must first find the Lagrangian. The Lagrangian in generalized coordinates is

$$L = \frac{1}{2}M\dot{q}^2 - Q\varphi(\mathbf{q}) + \frac{Q}{c}\dot{\mathbf{q}}\cdot\mathbf{A}(\dot{\mathbf{q}})\ . \tag{7}$$

This is correct because it leads to the correct equation of motion of a charge in combined electric and magnetic fields, as we now show.

In Cartesian coordinates the Lagrange equation of motion is

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = 0\ , \tag{8}$$

and similarly for y and z. From (7) we form

$$\frac{\partial L}{\partial x} = -Q\frac{\partial \varphi}{\partial x} + \frac{Q}{c}\left(\dot{x}\frac{\partial A_x}{\partial x} + \dot{y}\frac{\partial A_y}{\partial x} + \dot{z}\frac{\partial A_z}{\partial x}\right)\ ; \tag{9}$$

$$\frac{\partial L}{\partial \dot{x}} = M\dot{x} + \frac{Q}{c}A_x\ ; \tag{10}$$

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = M\ddot{x} + \frac{Q}{c}\frac{dA_x}{dt} = M\ddot{x} + \frac{Q}{c}\left(\frac{\partial A_x}{\partial t} + \dot{x}\frac{\partial A_x}{\partial x} + \dot{y}\frac{\partial A_x}{\partial y} + \dot{z}\frac{\partial A_x}{\partial z}\right)\ . \tag{11}$$

Thus (8) becomes

$$M\ddot{x} + Q\frac{\partial \varphi}{\partial x} + \frac{Q}{c}\left[\frac{\partial A_x}{\partial t} + \dot{y}\left(\frac{\partial A_x}{\partial y} - \frac{\partial A_y}{\partial x}\right) + \dot{z}\left(\frac{\partial A_z}{\partial z} - \frac{\partial A_z}{\partial x}\right)\right] = 0\ , \tag{12}$$

[1]For an elementary treatment of the vector potential see E. M. Purcell, *Electricity and magnetism*, 2nd ed., McGraw-Hill, 1984.

or

$$M\frac{d^2x}{dt^2} = QE_x + \frac{Q}{c}[\mathbf{v}\times\mathbf{B}]_x \ , \tag{13}$$

with

$$E_x = -\frac{\partial\varphi}{\partial x} - \frac{1}{c}\frac{\partial A_x}{\partial t} \ ; \tag{14}$$

$$\mathbf{B} = \text{curl}\,\mathbf{A} \ . \tag{15}$$

Equation (13) is the Lorentz force equation. This confirms that (7) is correct. We note in (14) that **E** has one contribution from the electrostatic potential φ and another from the time derivative of the magnetic vector potential **A**.

Derivation of the Hamiltonian

The momentum **p** is defined in terms of the Lagrangian as

$$\mathbf{p} \equiv \frac{\partial L}{\partial\dot{\mathbf{q}}} = M\dot{\mathbf{q}} + \frac{Q}{c}\mathbf{A} \ , \tag{16}$$

in agreement with (4). The hamiltonian $H(\mathbf{p},\mathbf{q})$ is defined by

$$H(\mathbf{p},\mathbf{q}) \equiv \mathbf{p}\cdot\dot{\mathbf{q}} - L \ , \tag{17}$$

or

$$H = M\dot{q}^2 + \frac{Q}{c}\dot{\mathbf{q}}\cdot\mathbf{A} - \frac{1}{2}M\dot{q}^2 + Q\varphi - \frac{Q}{c}\dot{\mathbf{q}}\cdot\mathbf{A} = \frac{1}{2M}\left(\mathbf{p} - \frac{Q}{c}\mathbf{A}\right)^2 + Q\varphi \ , \tag{18}$$

as in (1).

Field Momentum

The momentum in the electromagnetic field that accompanies a particle moving in a magnetic field is given by the volume integral of the Poynting vector, so that

$$\mathbf{p}_{\text{field}} = \frac{1}{4\pi c}\int dV\,\mathbf{E}\times\mathbf{B} \ . \tag{19}$$

We work in the nonrelativistic approximation with $v \ll c$, where v is the velocity of the particle. At low values of v/c we consider **B** to arise from an external source alone, but **E** arises from the charge on the particle. For a charge Q at $\mathbf{r}'$,

$$\mathbf{E} = -\nabla\varphi \ ; \qquad \nabla^2\varphi = -4\pi Q\,\delta(\mathbf{r}-\mathbf{r}') \ . \tag{20}$$

Thus

$$\mathbf{p}_f = -\frac{1}{4\pi c}\int dV\,\nabla\varphi\times\text{curl}\,\mathbf{A} \ . \tag{21}$$

By a vector relation we have

$$\int dV\, \nabla\varphi \times \operatorname{curl} \mathbf{A} = -\int dV\, [\mathbf{A} \times \operatorname{curl} (\nabla\varphi) - \mathbf{A} \operatorname{div} \nabla\varphi - (\nabla\varphi) \operatorname{div} \mathbf{A}] \ . \quad (22)$$

But curl $(\Delta\varphi) = 0$, and we can always choose the gauge such that div $\mathbf{A} = 0$. This is the transverse gauge.

Thus, we have

$$\mathbf{p}_f = -\frac{1}{4\pi c} \int dV\, \mathbf{A}\, \nabla^2\varphi = \frac{1}{c} \int dV\, \mathbf{A} Q\, \delta(\mathbf{r} - \mathbf{r}') = \frac{Q}{c} \mathbf{A} \ . \quad (23)$$

This is the interpretation of the field contribution to the total momentum $\mathbf{p} = M\mathbf{v} + Q\mathbf{A}/c$.

GAUGE TRANSFORMATION

Suppose $H\psi = \epsilon\psi$, where

$$H = \frac{1}{2M} \left(\mathbf{p} - \frac{Q}{c} \mathbf{A} \right)^2 \ . \quad (24)$$

Let us make a gauge transformation to $\mathbf{A}'$, where

$$\mathbf{A}' = \mathbf{A} + \nabla\chi \ , \quad (25)$$

where χ is a scalar. Now $\mathbf{B} = \operatorname{curl} \mathbf{A} = \operatorname{curl} \mathbf{A}'$, because curl $(\nabla\chi) \equiv 0$. The Schrödinger equation becomes

$$\frac{1}{2M} \left(\mathbf{p} - \frac{Q}{c} \mathbf{A}' + \frac{Q}{c} \nabla\chi \right)^2 \psi = \epsilon\psi \ . \quad (26)$$

What ψ' satisfies

$$\frac{1}{2M} \left(\mathbf{p} - \frac{Q}{c} \mathbf{A}' \right)^2 \psi' = \epsilon\psi' \ , \quad (27)$$

with the same ϵ as for ψ? Equation (27) is equivalent to

$$\frac{1}{2M} \left(\mathbf{p} - \frac{Q}{c} \mathbf{A} - \frac{Q}{c} \nabla\chi \right)^2 \psi' = \epsilon\psi' \ . \quad (28)$$

We try

$$\psi' = \exp(iQ\chi/\hbar c)\psi \ . \quad (29)$$

Now

$$p\psi' = \exp(iQ\chi/\hbar c)\mathbf{p}\psi + \frac{Q}{c} (\nabla\chi)\exp(iQ\chi/\hbar c)\psi \ ,$$

so that

$$\left(\mathbf{p} - \frac{Q}{c} \nabla\chi \right)\psi' = \exp(iQ\chi/\hbar c)\mathbf{p}\psi$$

and

$$\frac{1}{2M}\left(\mathbf{p} - \frac{Q}{c}\mathbf{A} - \frac{Q}{c}\nabla\chi\right)^2 \psi' = \exp(iQ\chi/\hbar c)\,\frac{1}{2M}\left(\mathbf{p} - \frac{Q}{c}\mathbf{A}\right)^2 \psi$$
$$= \exp(iQ\chi/\hbar c)\epsilon\psi \ . \tag{30}$$

Thus $\psi' = \exp(iQ\chi/\hbar c)\psi$ satisfies the Schrödinger equation after the gauge transformation (25). The energy ϵ is invariant under the transformation.

The gauge transformation on $\mathbf{A}$ merely changes the local phase of the wavefunction. We see that

$$\psi'^*\psi' = \psi^*\psi \ , \tag{31}$$

so that the charge density is invariant under a gauge transformation.

Gauge in the London Equation

Because of the equation of continuity in the flow of electric charge we require that in a superconductor

$$\operatorname{div}\mathbf{j} = 0 \ ,$$

so that the vector potential in the London equation $\mathbf{j} = -c\mathbf{A}/4\pi\lambda_L^2$ must satisfy

$$\operatorname{div}\mathbf{A} = 0 \ . \tag{32}$$

Further, there is no current flow through a vacuum/superconductor interface. The normal component of the current across the interface must vanish: $j_n = 0$, so that the vector potential in the London equation must satisfy

$$A_n = 0 \ . \tag{33}$$

The gauge of the vector potential in the London equation of superconductivity is to be chosen so that (32) and (33) are satisfied.

APPENDIX H: COOPER PAIRS

For a complete set of states of a two-electron system that satisfy periodic boundary conditions in a cube of unit volume, we take plane wave product functions

$$\varphi(\mathbf{k}_1,\mathbf{k}_2;\mathbf{r}_1,\mathbf{r}_2) = \exp[i(\mathbf{k}_1\cdot\mathbf{r}_1 + \mathbf{k}_2\cdot\mathbf{r}_2)] \ . \tag{1}$$

We assume that the electrons are of opposite spin.

We introduce center-of-mass and relative coordinates:

$$\mathbf{R} = \tfrac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2) \; ; \qquad \mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2 \; ; \tag{2}$$

$$\mathbf{K} = \mathbf{k}_1 + \mathbf{k}_2 \; ; \qquad \mathbf{k} = \tfrac{1}{2}(\mathbf{k}_1 - \mathbf{k}_2) \; , \tag{3}$$

so that

$$\mathbf{k}_1 \cdot \mathbf{r}_1 + \mathbf{k}_2 \cdot \mathbf{r}_2 = \mathbf{K} \cdot \mathbf{R} + \mathbf{k} \cdot \mathbf{r} \; . \tag{4}$$

Thus (1) becomes

$$\varphi(\mathbf{K},\mathbf{k};\mathbf{R},\mathbf{r}) = \exp(i\mathbf{K} \cdot \mathbf{R}) \exp(i\mathbf{k} \cdot \mathbf{r}) \; , \tag{5}$$

and the kinetic energy of the two-electron system is

$$\epsilon_{\mathbf{K}} + E_{\mathbf{k}} = (\hbar^2/m)(\tfrac{1}{4}K^2 + k^2) \; . \tag{6}$$

We give special attention to the product functions for which the center-of-mass wavevector $\mathbf{K} = 0$, so that $\mathbf{k}_1 = -\mathbf{k}_2$. With an interaction H_1 between the two electrons, we set up the eigenvalue problem in terms of the expansion

$$\chi(\mathbf{r}) = \Sigma g_{\mathbf{k}} \exp(i\mathbf{k} \cdot \mathbf{r}) \; . \tag{7}$$

The Schrödinger equation is

$$(H_0 + H_1 - \epsilon)\chi(\mathbf{r}) = 0 = \Sigma_{\mathbf{k}'} [(E_{\mathbf{k}'} - \epsilon)g_{\mathbf{k}'} + H_1 g_{\mathbf{k}'}] \exp(i\mathbf{k}' \cdot \mathbf{r}) \; , \tag{8}$$

where H_1 is the interaction energy of the two electrons. Here ϵ is the eigenvalue.

We take the scalar product with $\exp(i\mathbf{k} \cdot \mathbf{r})$ to obtain

$$(E_{\mathbf{k}} - \epsilon)g_{\mathbf{k}} + \Sigma_{\mathbf{k}} g_{\mathbf{k}'} (\mathbf{k}|H_1|\mathbf{k}') = 0 \; , \tag{9}$$

the secular equation of the problem.

Now transform the sum to an integral:

$$(E - \epsilon)g(E) + \int dE' \, g(E')H_1(E,E')N(E') = 0 \; , \tag{10}$$

where $N(E')$ is the number of two electron states with total momentum $\mathbf{K} = 0$ and with kinetic energy in dE' at E'.

Now consider the matrix elements $H_1(E,E') = (\mathbf{k}|H_1|\mathbf{k}')$. Studies of these by Bardeen suggest that they are important when the two electrons are confined to a thin energy shell near the Fermi surface—within a shell of thickness $\hbar\omega_D$ above E_F, where ω_D is the Debye phonon cutoff frequency. We assume that

$$H_1(E,E') = -V \tag{11}$$

for E,E' within the shell and zero otherwise. Here V is assumed to be positive.

Thus (10) becomes

$$(E - \epsilon)g(E) = V \int_{2\epsilon_F}^{2\epsilon_m} dE' \, g(E')N(E') = C \; , \tag{12}$$

with $\epsilon_m = \epsilon_F + \hbar\omega_D$. Here C is a constant, independent of E.

From (12) we have

$$g(E) = \frac{C}{E - \epsilon} \tag{13}$$

and

$$1 = V \int_{2\epsilon_F}^{2\epsilon_m} dE' \frac{N(E')}{E' - \epsilon} \ . \tag{14}$$

With $N(E')$ approximately constant and equal to N_F over the small energy range between $2\epsilon_m$ and $2\epsilon_F$, we take it out of the integral to obtain

$$1 = N_F V \int_{2\epsilon_F}^{2\epsilon_m} dE' \frac{1}{E' - \epsilon} = N_F V \log \frac{2\epsilon_m - \epsilon}{2\epsilon_F - \epsilon} \ . \tag{15}$$

Let the eigenvalue ϵ of (15) be written as

$$\epsilon = 2\epsilon_F - \Delta \ , \tag{16}$$

which defines the binding energy Δ of the electron pair, relative to two free electrons at the Fermi surface. Then (15) becomes

$$1 = N_F V \log \frac{2\epsilon_m - 2\epsilon_F + \Delta}{\Delta} = N_F V \log \frac{2\hbar\omega_D + \Delta}{\Delta} \ , \tag{17}$$

or

$$1/N_F V = \log(1 + 2\hbar\omega_D/\Delta) \ . \tag{18}$$

This result for the binding energy of a Cooper pair may be written as

$$\Delta = \frac{2\hbar\omega_D}{\exp(1/N_F V) - 1} \ . \tag{19}$$

For V positive (attractive interaction) the energy of the system is lowered by excitation of a pair of electrons above the Fermi level. Therefore the Fermi gas is unstable in an important way. The binding energy (19) is closely related to the superconducting energy gap E_g. The BCS calculations show that a high density of Cooper pairs may form in a metal.

APPENDIX I: GINZBURG-LANDAU EQUATION

We owe to Ginzburg and Landau an elegant theory of the phenomenology of the superconducting state and of the spatial variation of the order parameter in that state. An extension of the theory by Abrikosov describes the structure of the vortex state which is exploited technologically in superconducting magnets. The attractions of the GL theory are the natural introduction of the

coherence length and of the wavefunction used in the theory of the Josephson effects in Chapter 12.

We introduce the **order parameter** $\psi(\mathbf{r})$ with the property that

$$\psi^*(\mathbf{r})\psi(\mathbf{r}) = n_S(\mathbf{r}) \ , \tag{1}$$

the local concentration of superconducting electrons. The mathematical formulation of the definition of the function $\psi(\mathbf{r})$ will come out of the BCS theory. We first set up a form for the free energy density $F_S(\mathbf{r})$ in a superconductor as a function of the order parameter. We assume that in the general vicinity of the transition temperature

$$F_S(\mathbf{r}) = F_N - \alpha|\psi|^2 + \tfrac{1}{2}\beta|\psi|^4 + (1/2m)|(-i\hbar\nabla - qA/c)\psi|^2 - \int_0^{B_a} \mathbf{M}\cdot d\mathbf{B}_a \ , \tag{2}$$

with the phenomenological positive constants α, β, and m, of which more will be said. Here:

1. F_N is the free energy density of the normal state.

2. $-\alpha|\psi|^2 + \frac{1}{2}\beta|\psi|^4$ is a typical Landau form for the expansion of the free energy in terms of an order parameter that vanishes at a second-order phase transition. This term may be viewed as $-\alpha n_S + \frac{1}{2}\beta n_S^2$ and by itself is a minimum with respect to n_S when $n_S(T) = \alpha/\beta$.

3. The term in $|\text{grad}\ \psi|^2$ represents an increase in energy caused by a spatial variation of the order parameter. It has the form of the kinetic energy in quantum mechanics.[1] The kinetic momentum $-i\hbar\nabla$ is accompanied by the field momentum $-q\mathbf{A}/c$ to ensure the gauge invariance of the free energy, as in Appendix G. Here $q = -2e$ for an electron pair.

4. The term $-\int\mathbf{M}\cdot d\mathbf{B}_a$, with the fictitious magnetization $\mathbf{M} = (\mathbf{B} - \mathbf{B}_a)/4\pi$, represents the increase in the superconducting free energy caused by the expulsion of magnetic flux from the superconductor.

The separate terms in (2) will be illustrated by examples as we progress further. First let us derive the GL equation (6). We minimize the total free energy $\int dV\, F_S(\mathbf{r})$ with respect to variations in the function $\psi(\mathbf{r})$. We have

$$\delta F_S(\mathbf{r}) = [-\alpha\psi + \beta|\psi|^2\psi + (1/2m)(-i\hbar\nabla - q\mathbf{A}/c)\psi\cdot(i\hbar\nabla - q\mathbf{A}/c)\delta\psi^* + \text{c.c.}] \tag{3}$$

We integrate by parts to obtain

$$\int dV\,(\nabla\psi)(\nabla\delta\psi^*) = -\int dV\,(\nabla^2\psi)\delta\psi^* \ , \tag{4}$$

if $\delta\psi^*$ vanishes on the boundaries. It follows that

$$\delta\int dV\, F_S = \int dV\,\delta\psi^*[-\alpha\psi + \beta|\psi|^2\psi + (1/2m)(-i\hbar\nabla - q\mathbf{A}/c)^2\psi] + \text{c.c.} \tag{5}$$

[1] A contribution of the form $|\nabla\mathbf{M}|^2$, where $\mathbf{M}$ is the magnetization, was introduced by Landau and Lifshitz to represent the exchange energy density in a ferromagnet; see *QTS*, p. 65.

This integral is zero if the term in brackets is zero:

$$\boxed{[(1/2m)(-i\hbar\nabla - q\mathbf{A}/c)^2 - \alpha + \beta|\psi|^2]\psi = 0 \ .} \tag{6}$$

This is the Ginzburg-Landau equation; it resembles a Schrödinger equation for ψ.

By minimizing (2) with respect to $\delta\mathbf{A}$ we obtain a gauge-invariant expression for the supercurrent flux:

$$\mathbf{j}_S(\mathbf{r}) = -(iq\hbar/2m)(\psi^*\nabla\psi - \psi\nabla\psi^*) - (q^2/mc)\psi^*\psi\mathbf{A} \ . \tag{7}$$

At a free surface of the specimen we must choose the gauge to satisfy the boundary condition that no current flows out of the superconductor into the vacuum: $\hat{\mathbf{n}} \cdot \mathbf{j}_S = 0$, where $\hat{\mathbf{n}}$ is the surface normal.

Coherence Length. The intrinsic coherence length ξ may be defined from (6). Let $\mathbf{A} = 0$ and suppose that $\beta|\psi|^2$ may be neglected in comparison with α. In one dimension the GL equation (6) reduces to

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} = \alpha\psi \ . \tag{8}$$

This has a wavelike solution of the form $\exp(ix/\xi)$, where ξ is defined by

$$\xi \equiv (\hbar^2/2m\alpha)^{1/2} \ . \tag{9}$$

A more interesting special solution is obtained if we retain the nonlinear term $\beta|\psi|^2$ in (6). Let us look for a solution with $\psi = 0$ at $x = 0$ and with $\psi \to \psi_0$ as $x \to \infty$. This situation represents a boundary between normal and superconducting states. Such states can coexist if there is a magnetic field H_c in the normal region. For the moment we neglect the penetration of the field into the superconducting region: we take the field penetration depth $\lambda \ll \xi$, which defines an extreme type I superconductor.

The solution of

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} - \alpha\psi + \beta|\psi|^2\psi = 0 \ , \tag{10}$$

subject to our boundary conditions, is

$$\psi(x) = (\alpha/\beta)^{1/2}\tanh(x/\sqrt{2}\xi) \ . \tag{11}$$

This may be verified by direct substitution. Deep inside the superconductor we have $\psi_0 = (\alpha/\beta)^{1/2}$, as follows from the minimization of the terms $-\alpha|\psi|^2 + \frac{1}{2}\beta|\psi|^4$ in the free energy. We see from (11) that ξ marks the extent of the coherence of the superconducting wavefunction into the normal region.

We have seen that deep inside the superconductor the free energy is a minimum when $|\psi_0|^2 = \alpha/\beta$, so that

$$F_S = F_N - \alpha^2/2\beta = F_N - H_c^2/8\pi \ , \tag{12}$$

by definition of the thermodynamic critical field H_c as the stabilization free energy density of the superconducting state. It follows that the critical field is related to α and β by

$$H_c = (4\pi\alpha^2/\beta)^{1/2} \ . \tag{13}$$

Consider the penetration depth of a weak magnetic field ($B \ll H_c$) into a superconductor. We assume that $|\psi|^2$ in the superconductor is equal to $|\psi_0|^2$, the value in the absence of a field. Then the equation for the supercurrent flux reduces to

$$\mathbf{j}_S(\mathbf{r}) = -(q^2/mc)|\psi_0|^2\mathbf{A} \ , \tag{14}$$

which is just the London equation $\mathbf{j}_S(\mathbf{r}) = -(c/4\pi\lambda^2)\mathbf{A}$, with the penetration depth

$$\lambda = \left(\frac{mc^2}{4\pi q^2|\psi_0|^2}\right)^{1/2} = \left(\frac{mc^2\beta}{4\pi q^2\alpha}\right)^{1/2} \ . \tag{15}$$

The dimensionless ratio $\kappa \equiv \lambda/\xi$ of the two characteristic lengths is an important parameter in the theory of superconductivity. From (9) and (15) we find

$$\kappa = \frac{mc}{q\hbar}\left(\frac{\beta}{2\pi}\right)^{1/2} \ . \tag{16}$$

We now show that the value $\kappa = 1/\sqrt{2}$ divides type I superconductors ($\kappa < 1/\sqrt{2}$) from type II superconductors ($\kappa > 1/\sqrt{2}$).

Calculation of the Upper Critical Field. Superconducting regions nucleate spontaneously within a normal conductor when the applied magnetic field is decreased below a value denoted by H_{c2}. At the onset of superconductivity $|\psi|$ is small and we linearize the GL equation (6) to obtain

$$\frac{1}{2m}(-i\hbar\nabla - q\mathbf{A}/c)^2\psi = \alpha \ . \tag{17}$$

The magnetic field in a superconducting region at the onset of superconductivity is just the applied field, so that $\mathbf{A} = B(0,x,0)$ and (17) becomes

$$-\frac{\hbar}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2}\right)\psi + \frac{1}{2m}\left(i\hbar\frac{\partial}{\partial y} + \frac{qB}{c}x\right)^2\psi = \alpha\psi \ . \tag{18}$$

This is of the same form as the Schrödinger equation of a free particle in a magnetic field.

We look for a solution in the form $\exp[i(k_y y + k_z z)]\varphi(x)$ and find

$$(1/2m)[-\hbar^2 d^2/dx^2 + \hbar^2 k_z^2 + (\hbar k_y - qBx/c)^2]\varphi = \alpha\varphi \ , \tag{19}$$

this is the equation for an harmonic oscillator, if we set $E = \alpha - (\hbar^2/2m)(k_y^2 + k_z^2)$ as the eigenvalue of

$$(1/2m)[-\hbar^2 d^2/dx^2 + (q^2B^2/c^2)x^2 - (2\hbar k_y qB/c)x]\varphi = E\varphi \ . \tag{20}$$

The term linear in x can be transformed away by a shift of the origin from 0 to $x_0 = \hbar k_y qB/2mc$, so that (20) becomes, with $X = x - x_0$,

$$-\left[\frac{\hbar^2}{2m}\frac{d^2}{dX^2} + \tfrac{1}{2}m(qB/mc)^2X^2\right]\varphi = (E + \hbar^2k_y^2/2m)\varphi \ . \tag{21}$$

The largest value of the magnetic field B for which solutions of (21) exist is given by the lowest eigenvalue, which is

$$\tfrac{1}{2}\hbar\omega = \hbar qB_{\max}/2mc = \alpha - \hbar^2k_z^2/2m \ , \tag{22}$$

where ω is the oscillator frequency qB/mc. With k_z set equal to zero,

$$B_{\max} \equiv H_{c2} = 2\alpha mc/q\hbar \ . \tag{23}$$

This result may be expressed by (13) and (16) in terms of the thermodynamic critical field H_c and the GL parameter $\kappa = \lambda/\xi$:

$$H_{c2} = \frac{2\alpha mc}{q\hbar} \cdot \frac{H_c}{(4\pi\alpha^2/\beta)^{1/2}} = \sqrt{2}\,\frac{mc}{\hbar q}\sqrt{\frac{\beta}{2\pi}}\,H_c = \sqrt{2}\kappa H_c \ . \tag{24}$$

When $\lambda/\xi > 1/\sqrt{2}$, a superconductor has $H_{c2} > H_c$ and is said to be of type II.

It is helpful to write H_{c2} in terms of the flux quantum $\Phi_0 = 2\pi\hbar c/q$ and $\xi^2 = \hbar^2/2m\alpha$:

$$H_{c2} = \frac{2mc\alpha}{q\hbar} \cdot \frac{q\Phi_0}{2\pi\hbar c} \cdot \frac{\hbar^2}{2m\alpha\xi^2} = \frac{\Phi_0}{2\pi\xi^2} \ . \tag{25}$$

This tells us that at the upper critical field the flux density H_{c2} in the material is equal to one flux quantum per area $2\pi\xi^2$, consistent with a fluxoid lattice spacing of the order of ξ.

APPENDIX J: ELECTRON-PHONON COLLISIONS

Phonons distort the local crystal structure and hence distort the local band structure. This distortion is sensed by the conduction electrons. The important effects of the coupling of electrons with phonons are

- Electrons are scattered from one state $\mathbf{k}$ to another state $\mathbf{k}'$, leading to electrical resistivity.

- Phonons can be absorbed in the scattering event, leading to the attenuation of ultrasonic waves.
- An electron will carry with it a crystal distortion, and the effective mass of the electron is thereby increased.
- A crystal distortion associated with one electron can be sensed by a second electron, thereby causing the electron-electron interaction that enters the theory of superconductivity.

The deformation potential approximation is that the electron energy $\epsilon(\mathbf{k})$ is coupled to the crystal dilation $\Delta(\mathbf{r})$ or fractional volume change by

$$\epsilon(\mathbf{k},\mathbf{r}) = \epsilon_0(\mathbf{k}) + C\Delta(\mathbf{r}) \ , \tag{1}$$

where C is a constant. The approximation is useful for spherical band edges $\epsilon_0(\mathbf{k})$ at long phonon wavelengths and low electron concentrations. The dilation may be expressed in terms of the phonon operators $a_\mathbf{q}$, $a_\mathbf{q}^+$ of Appendix C by

$$\Delta(\mathbf{r}) = i\sum_q (\hbar/2M\omega_\mathbf{q})^{1/2}|\mathbf{q}|[a_\mathbf{q}\exp(i\mathbf{q}\cdot\mathbf{r}) - a_\mathbf{q}^+\exp(-i\mathbf{q}\cdot\mathbf{r})] \ . \tag{2}$$

as in *QTS*, p. 23. Here M is the mass of the crystal. The result (2) also follows from (C.32) on forming $q_s - q_{s-1}$ in the limit $k \ll 1$.

In the Born approximation for the scattering we are concerned with the matrix elements of $C\Delta(\mathbf{r})$ between the one-electron Bloch states $|\mathbf{k}\rangle$ and $|\mathbf{k}'\rangle$, with $|\mathbf{k}\rangle = \exp(i\mathbf{k}\cdot\mathbf{r})u_\mathbf{k}(\mathbf{r})$. In the wave field representation the matrix element is

$$\begin{aligned} H' &= \int d^3x\, \psi^+(\mathbf{r})C\Delta(\mathbf{r})\psi(\mathbf{r}) = \sum_{\mathbf{k}'\mathbf{k}} c_\mathbf{k}^+ c_\mathbf{k}\langle\mathbf{k}'|C\Delta|\mathbf{k}\rangle \\ &= iC\sum_{\mathbf{k}'\mathbf{k}} c_{\mathbf{k}'}^+ c_\mathbf{k} \sum_\mathbf{q} (\hbar/2M\omega_\mathbf{q})^{1/2}|\mathbf{q}|(a_\mathbf{q}\int d^3x\, u_{\mathbf{k}'}^* u_\mathbf{k} e^{i(\mathbf{k}-\mathbf{k}'+\mathbf{q})\cdot\mathbf{r}} \\ &\quad - a_\mathbf{q}^+ \int d^3x\, u_{\mathbf{k}'}^* u_\mathbf{k} e^{i(\mathbf{k}-\mathbf{k}'-\mathbf{q})\cdot\mathbf{r}}) \ , \end{aligned} \tag{3}$$

where

$$\psi(\mathbf{r}) = \sum_\mathbf{k} c_\mathbf{k}\varphi_\mathbf{k}(\mathbf{r}) = \sum_\mathbf{k} c_\mathbf{k}\exp(i\mathbf{k}\cdot\mathbf{r})u_\mathbf{k}(\mathbf{r}) \ , \tag{4}$$

where $c_\mathbf{k}^+$, $c_\mathbf{k}$ are the fermion creation and annihilation operators. The product $u_{\mathbf{k}'}^*(\mathbf{r})u_\mathbf{k}(\mathbf{r})$ involves the periodic parts of the Bloch functions and is itself periodic in the lattice; thus the integral in (3) vanishes unless

$$\mathbf{k} - \mathbf{k}' \pm \mathbf{q} = \begin{cases} 0 \\ \text{vector in the reciprocal lattice.} \end{cases}$$

In semiconductors at low temperatures only the possibility zero (N processes) may be allowed energetically.

Let us limit ourselves to N processes, and for convenience we approximate $\int d^3x\, u_{\mathbf{k}'} u_{\mathbf{k}}$ by unity. Then the deformation potential perturbation is

$$H' = iC \sum_{\mathbf{kq}} (\hbar/2M\omega_{\mathbf{q}})^{1/2} |\mathbf{q}| (a_{\mathbf{q}} c^+_{\mathbf{k}+\mathbf{q}} c_{\mathbf{k}} - a^+_{\mathbf{q}} c^+_{\mathbf{k}-\mathbf{q}} c_{\mathbf{k}}) \ . \tag{5}$$

Relaxation Time. In the presence of the electron-phonon interaction the wavevector $\mathbf{k}$ is not a constant of the motion for the electron alone, but the sum of the wavevectors of the electron and virtual phonon is conserved. Suppose an electron is initially in the state $|\mathbf{k}\rangle$; how long will it stay in that state?

We calculate first the probability ω per unit time that the electron in $\mathbf{k}$ will emit a phonon $\mathbf{q}$. If $n_{\mathbf{q}}$ is the initial population of the phonon state,

$$w(\mathbf{k}-\mathbf{q};n_{\mathbf{q}}+1|\mathbf{k};n_{\mathbf{q}}) = (2\pi/\hbar)|\langle \mathbf{k}-\mathbf{q};n_{\mathbf{q}}+1|H'|\mathbf{k};n_{\mathbf{q}}\rangle|^2 \delta(\epsilon_{\mathbf{k}} - \hbar\omega_{\mathbf{q}} - \epsilon_{\mathbf{k}-\mathbf{q}}) \tag{6}$$

by time-dependent perturbation theory. Here

$$|\langle \mathbf{k}-\mathbf{q};n_{\mathbf{q}}+1|H'|\mathbf{k};n_{\mathbf{q}}\rangle|^2 = [C^2\hbar q/2Mc_S(n_{\mathbf{q}}+1)] \ . \tag{7}$$

The total collision rate W of an electron in the state $|\mathbf{k}\rangle$ with a phonon system at absolute zero is, with $n_{\mathbf{q}} = 0$,

$$W = \frac{C^2}{4\pi\rho c_s} \int_{-1}^{1} d(\cos\theta_{\mathbf{q}}) \int_0^{q_m} dq\, q^3 \delta(\epsilon_{\mathbf{k}} - \epsilon_{\mathbf{k}-\mathbf{q}} - \hbar\omega_{\mathbf{q}}) \ , \tag{8}$$

where ρ is the mass density.

The argument of the delta function is

$$\frac{\hbar^2}{2m^*}(2\mathbf{k}\cdot\mathbf{q} - q^2) - \hbar c_s q = \frac{\hbar^2}{2m^*}(2\mathbf{k}\cdot\mathbf{q} - q^2 - qq_c) \ , \tag{9}$$

where $q_c = 2\hbar m^* c_s$, with c_s the velocity of sound. The minimum value of k for which the argument can be zero is $k_{\min} = \frac{1}{2}(q + q_c)$, which for $q = 0$ reduces to $k_{\min} = \frac{1}{2}q_c = m^* c_s/\hbar$. For this value of k the electron group velocity $v_g = k_{\min}/m^*$ is equal to the velocity of sound. Thus the threshold for the emission of phonons by electrons in a crystal is that the electron group velocity should exceed the acoustic velocity. This requirement resembles the Cerenkov threshold for the emission of photons in crystals by fast electrons. The electron energy at the threshold is $\frac{1}{2}m^* c_s^2 \sim 10^{-27} \cdot 10^{11} \sim 10^{-16}$ erg ~ 1 K. An electron of energy below this threshold will not be slowed down in a perfect crystal at absolute zero, even by higher order electron-phonon interactions, at least in the harmonic approximation for the phonons.

For $k \gg q_c$ we may neglect the qq_c term in (9). The integrals in (8) become

$$\int_{-1}^{1} d\mu \int dq\, q^3 (2m^*/\hbar^2 q)\delta(2k\mu - q) = (8m^*/\hbar^2) \int_0^1 d\mu\, k^2\mu^2 = 8m^* k^2/3\hbar^2 \ , \tag{10}$$

and the phonon emission rate is

$$W(\text{emission}) = \frac{2C^2 m^* k^2}{3\pi\rho c_s \hbar^2} , \tag{11}$$

directly proportional to the electron energy $\epsilon_{\mathbf{k}}$. The loss of the component of wavevector parallel to the original direction of the electron when a phonon is emitted at an angle θ to $\mathbf{k}$ is given by $q \cos\theta$. The fractional rate of loss of k_z is given by the transition rate integral with the extra factor $(q/k) \cos\theta$ in the integrand. Instead of (10), we have

$$(2m^*/\hbar^2 k) \int_0^1 d\mu \, 8k^3\mu^4 = 16m^* k^2/5\hbar^2 , \tag{12}$$

so that the fractional rate of decrease of k_z is

$$W(k_z) = 4C^2 m^* k^2/5\pi\rho c_s \hbar^2 . \tag{13}$$

This quantity enters into the electrical resistivity.

The above results apply to absolute zero. At a temperature $k_B T \gg \hbar c_s k$ the integrated phonon emission rate is

$$W(\text{emission}) = \frac{C^2 m^* k k_B T}{\pi c_s^2 \rho \hbar^3} . \tag{14}$$

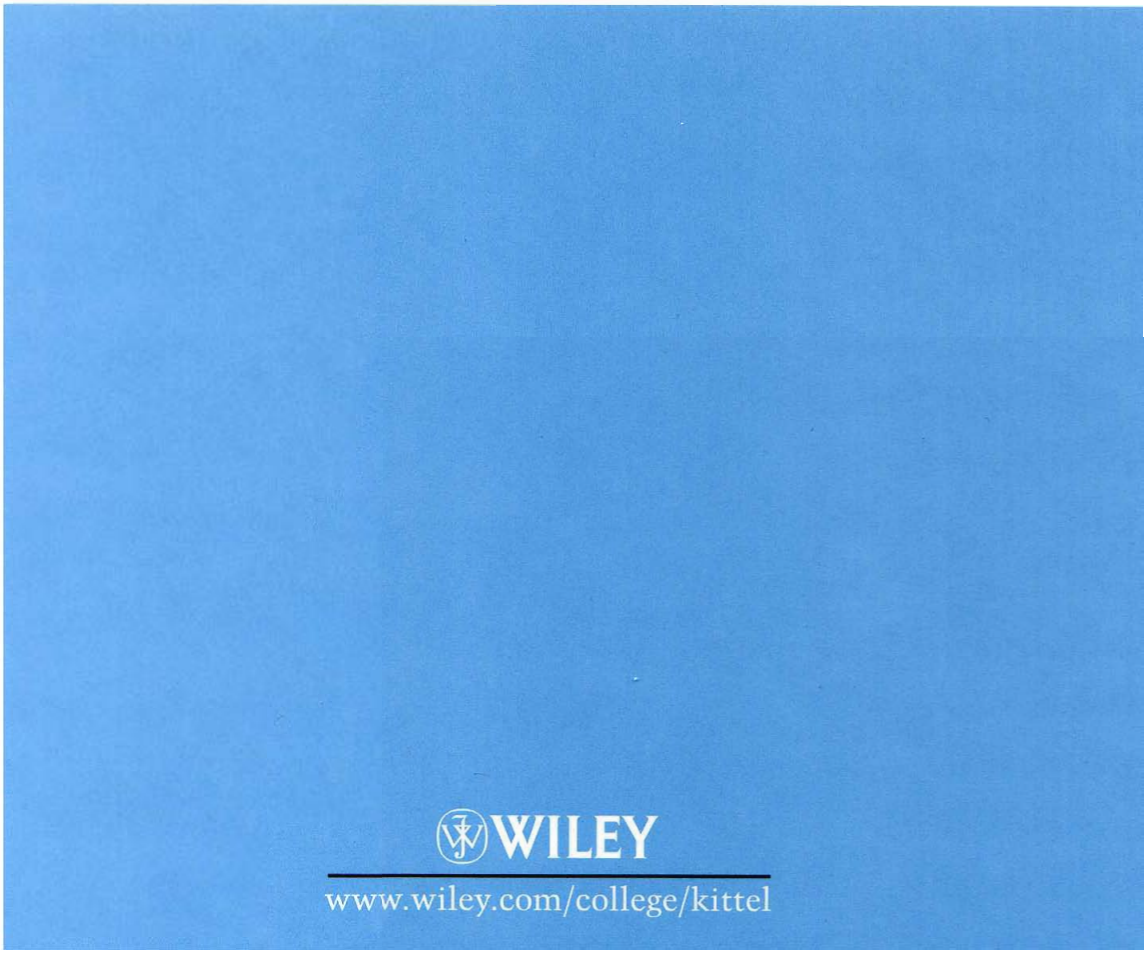
For electrons in thermal equilibrium at not too low temperatures the required inequality is easily satisfied for the rms value of k. If we take $C = 10^{-12}$ erg; $m^* = 10^{-27}$ g; $k = 10^7$ cm^{-1}; $c_s = 3 \times 10^5$ cm s^{-1}; $\rho = 5$ g cm^{-3}; then $W \simeq 10^{12}$ s^{-1}. At absolute zero (13) gives $W \simeq 5 \times 10^{10}$ s^{-1} with these same parameters.

Index

Table of Values

Quantity	Symbol	Value	CGS	SI
Velocity of light	c	2.997925	10^{10} cm s ^{1}	10^{8} m s^{-1}
Proton charge	e	1.60219	–	10^{-19} C
		4.80325	10^{-10} esu	–
Planck's constant	h	6.62620	10^{-27} erg s	10^{-34} J s
	$\hbar = h/2\pi$	1.05459	10^{-27} erg s	10^{-34} J s
Avogadro's number	N	6.02217×10^{23} mol^{-1}	–	–
Atomic mass unit	amu	1.66053	10^{-24} g	10^{-27} kg
Electron rest mass	m	9.10956	10^{-28} g	10^{-31} kg
Proton rest mass	M_p	1.67261	10^{-24} g	10^{-27} kg
Proton mass/electron mass	M_p/m	1836.1	–	–
Reciprocal fine structure constant $\hbar c/e^2$	$1/\alpha$	137.036	–	–
Electron radius e^2/mc^2	r_e	2.81794	10^{-13} cm	10^{-15} m
Electron Compton wavelength $\hbar/mc$	$\lambda\!\!\!^{-}_e$	3.86159	10^{-11} cm	10^{-13} m
Bohr radius $\hbar^2/me^2$	r_0	5.29177	10^{-9} cm	10^{-11} m
Bohr magneton $e\hbar/2mc$	μ_B	9.27410	10^{-21} erg G^{-1}	10^{-24} J T^{-1}
Rydberg constant $me^4/2\hbar^2$	R_∞ or Ry	2.17991	10^{-11} erg	10^{-18} J
		13.6058 eV		
1 electron volt	eV	1.60219	10^{-12} erg	10^{-19} J
	eV/h	2.41797×10^{14} Hz	–	–
	eV/hc	8.06546	10^{3} cm^{-1}	10^{5} m^{-1}
	eV/k_B	1.16048×10^{4} K	–	–
Boltzmann constant	k_B	1.38062	10^{-16} erg K^{-1}	10^{-23} J K^{-1}
Permittivity of free space	ϵ_0	–	1	$10^{7}/4\pi c^2$
Permeability of free space	μ_0	–	1	$4\pi \times 10^{-7}$

Source. B. N. Taylor, W. H. Parker, and D. N. Langenberg, Rev. Mod. Phys. 41, 375 (1969). See also E. R. Cohen and B. N. Taylor, Journal of Physical and Chemical Reference Data 2(4), 663 (1973).


WILEY
www.wiley.com/college/kittel