ELEMENTS OF SOLID STATE PHYSICS

SECOND EDITION

J.P. SRIVASTAVA
Former Professor of Physics
Aligarh Muslim University
Aligarh

PHI Learning Private Limited
New Delhi-110001
2009

Contents

Chapter 1

The Crystalline State

The solidification of elements and their chemical compounds from liquid and gaseous states is generally understood in terms of the combined effect of attractive and repulsive forces among the constituent particles (atoms/molecules). The details of this process are far from simple to picturize as this requires us to know mainly the nature of response of valence electrons in atoms to the presence of other atoms. The solidification or freezing of matter may lead to the formation of an ordered or disordered state. The ordered state is commonly called the *crystalline* state whereas the disordered one is termed the *amorphous* state.

The crystalline state refers to an infinite array of atoms or a group of atoms. The regularity in the arrangement or the periodicity extends over the whole volume of a crystalline matter or crystal. The whole volume of a crystal can be constructed by moving a building block of the smallest acceptable size along its edges. This block consisting of atoms or a group of atoms is called a *unit cell.*

Even with the complete available knowledge of the laws of atomic packing and geometrical considerations for a certain material, it is generally not possible to predict the material's crystal structure. Predictions turn out to be wrong as often as right. Various crystal structures existing in nature can be understood by systematically exploring all the possible ways of arranging atoms in periodic arrays. This process involves the visualization of motional operations performed on atoms within a unit cell such that the operations leave the unit cell unchanged. These operations, called 'Symmetry Operations', form the most vital component of Geometrical Crystallography. Every crystal structure satisfies the requirements of a specific group (combination) of certain symmetry operations. Different combinations of symmetry operations result in different crystal structures. Thus the knowledge of the symmetry operations helps us in making a complete list of possible crystal structures and in classifying crystals.

We take up below the relationship between the geometrical symmetry and the constituent particles of a crystal. This helps us in appreciating the meaning of unit cell, defined earlier. Further, the fact that the unit cell design is not always unique is very crucial in the theory of crystalline solids. The illustration of these ideas needs to be carried out in the immediate continuation of our discussion as it will equip us better to give an adequate account of symmetry operations.

1.1 BASIS OF CRYSTAL STRUCTURE

An infinite periodic array of points in space is called a *lattice.* In more general terms it is known as the *space lattice.* The arrangement of points defines the lattice symmetry. When an atom or an identical group of atoms is attached to every lattice point, we obtain a crystal structure. The attached atom or the group of atoms, called *basis,* is identical for every lattice point in terms of composition, relative orientation and separation. These features are explained in Fig. 1.1. The number of atoms in the basis of an inorganic crystal may approach 100 whereas it is known to be as large as a few thousands in organic crystals like protein.

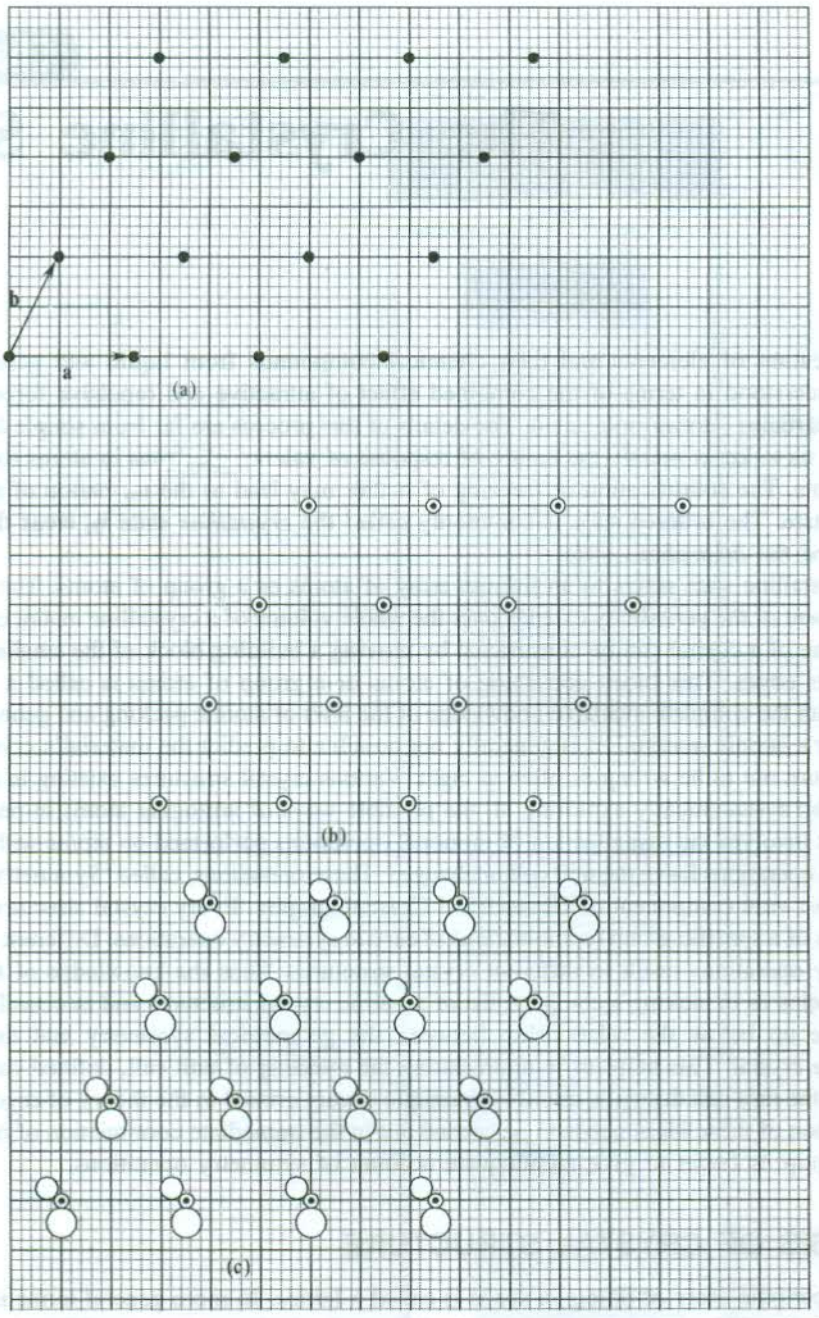


FIG. 1.1 (a) A regular arrangement of points on a plane forming a two-dimensional space lattice. The points are called *lattice points*. Vectors **a** and **b** define the symmetry and size of the lattice. (b) A two-dimensional crystal formed when an atom is placed at each lattice point of the space lattice shown in (a). All atoms are identical (of the same element). The single atom forms the *basis* of the crystal structure. The basis of the crystal structure is one. (c) A crystal structure with the space lattice (a) and the basis of three atoms, all of different elements. The relative positions and orientations of atoms in the basis are identical for all lattice points which may now be taken as imaginary points.

In Fig. 1.1(a), **a** and **b** represent vectors that determine the size of a unit cell. Since a unit cell is another name for the smallest structural unit that may be used as the building block of a crystal, a few words about its acceptable size are very much in order. Examine the structure of the following two-dimensional crystal.

The unit cell shown in Fig. 1.2 is represented by $\mathbf{a}_2$, $\mathbf{b}_2$ and not by $\mathbf{a}_1$, $\mathbf{b}_1$. This takes us closer to the introduction of symmetry and symmetry operations in crystals. Figure 1.2 can be used to describe the simplest of the symmetry operations—the translation. In this structure the translation of the unit cell represented by $\mathbf{a}_2$ and $\mathbf{b}_2$ vectors is able to map the infinite pattern of the crystal whereas it is not possible with $\mathbf{a}_1$ and $\mathbf{b}_1$ as the translation by these vectors takes the smaller atom to the position of the bigger atom and vice versa. Thus in the present case the size of the smallest structural unit representing the unit cell is given by $\mathbf{a}_2$, $\mathbf{b}_2$.

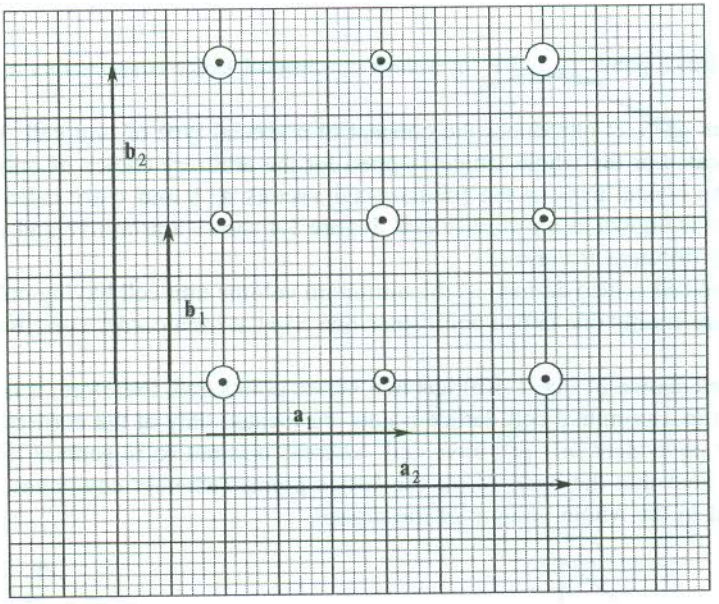


FIG. 1.2 Identifying the periodicity of the crystal structure to determine the primitive translation vectors ($\mathbf{a}_2$ and $\mathbf{b}_2$).

1.2 UNIT CELL—PRIMITIVE CELL STRUCTURES

The translations along the crystal axes **a**, **b** and **c** are the simplest of many translations that may map the crystal lattice completely. These translations are actually the various combinations of **a**, **b** and **c**, naturally, with certain restrictions. These restrictions are closely linked with the structural aspects of a unit cell which are discussed below.

A proper description of the subject may be carried out by referring to the three-dimensional space lattices. Draw a vector **t** connecting two lattice points P_1 and P_2 (Fig. 1.3) represented by the position vectors $\mathbf{r}_1$ and $\mathbf{r}_2$, respectively. Then, the vector **t** is given by the relation

$$\mathbf{r}_2 = \mathbf{r}_1 + \mathbf{t} \tag{1.1}$$

such that

$$\mathbf{t} = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c} \tag{1.2}$$

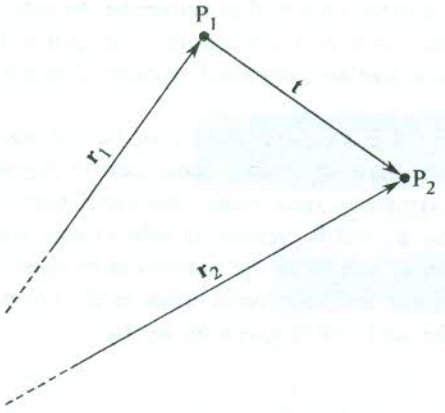


FIG. 1.3 Two lattice points P_1 and P_2 of a three-dimensional lattice connected by a translation vector **t**.

The interpretation of (1.2) introduces the concept of a special type of unit cell. When all the lattice points can be located or the whole lattice mapped for the arbitrary choice of only integral values of n_1, n_2 and n_3, the crystal axes or the fundamental vectors **a**, **b** and **c** are called *primitive* and the resulting unit cell is known as a *primitive cell*. Figure 1.4 shows a primitive and a non-primitive cell belonging to two different materials.

The primitive cell is also defined as the minimum volume unit cell having only one lattice point. In Fig. 1.4(a) there is only one lattice point since one point at each corner contributes 1/8th of a point to the cell.

The other cell shown in Fig. 1.4(b) has two lattice points and, therefore, it is non-primitive. We can also say that in a primitive cell, lattice points are located only at the corners of the cell. All the lattice points are equivalent in a primitive cell, whereas it is not so in general [Fig. 1.4(b)]. But we must not forget here that the basis of all the lattice points (whether equivalent or non-equivalent) remains identical in the whole volume of the crystal.

FIG. 1.4 (a) A primitive cell. There is one lattice point in the cell. A point at the corner of the cube contributes the fraction (1/8). (b) A non-primitive cell with two lattice points per cell.

It is desirable to mention here that there can be more than one choice for the primitive cell of a given crystal. All of these primitive cells are equally acceptable as each of them has the same volume. This property can be easily demonstrated in a two-dimensional space lattice.

Figure 1.5 shows that all acceptable primitive cells for a square lattice have a common value of the area enclosed.

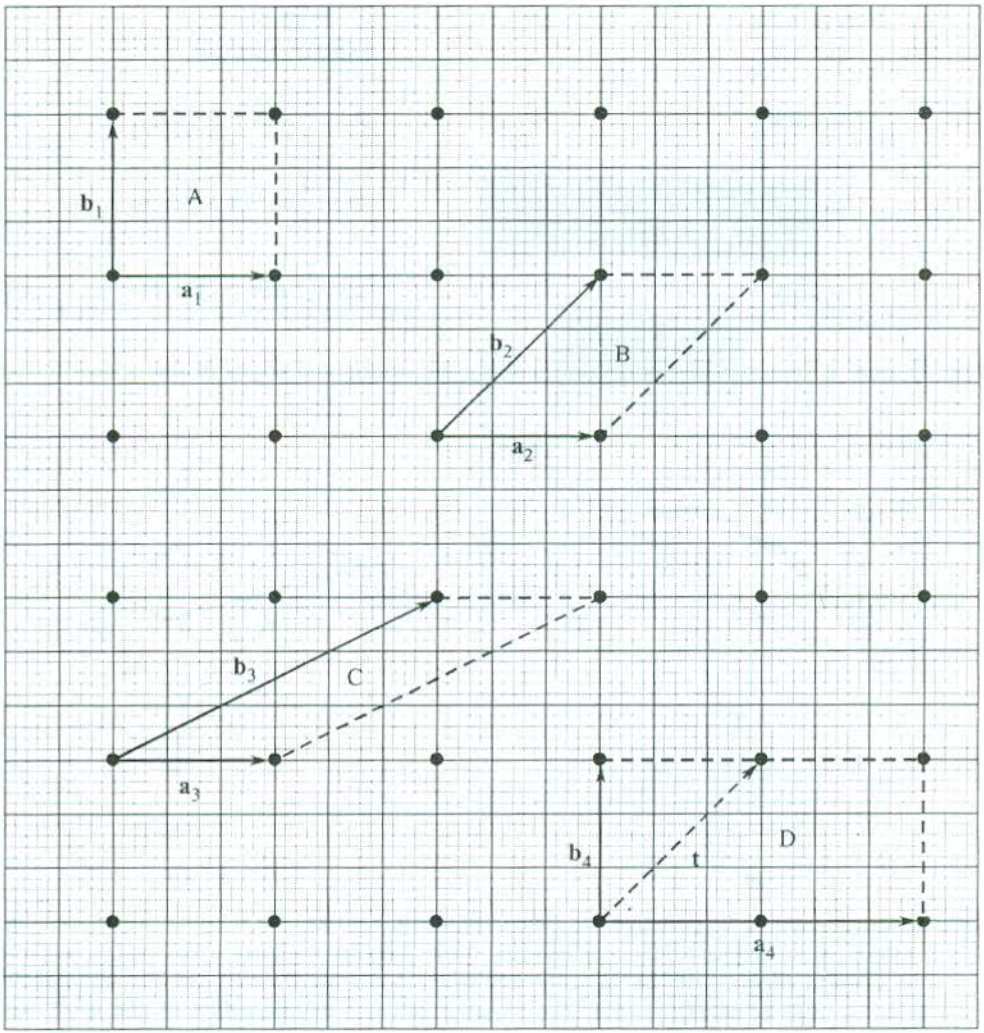


FIG. 1.5 Choices for the primitive cell of a square lattice. The choice D is not acceptable since all lattice points cannot be mapped with integral values of n_1 and n_2 in the vector, $\mathbf{t} = n_1\mathbf{a}_4 + n_2\mathbf{b}_4$.

The use of a primitive cell or a non-primitive unit cell for the study of a crystalline material is purely a matter of convenience. There is no hard and fast rule in this respect. Figure 1.6(a) shows the non-primitive and primitive cells for the three-dimensional space lattice pictured in Fig. 1.4(b).

The space lattice of the unit cell shown in Fig. 1.6(a) is the body centred cubic (BCC). But the same structure has a rhombohedral symmetry for the design of its primitive cell [Fig. 1.6(a)]. The edges of the unit cell in the two cases are as follows:

Non-primitive (BCC)		Primitive (Rhombohedral)	
$\mathbf{a} = a\hat{\mathbf{x}}$		$\mathbf{a} = \frac{1}{2}a(\hat{\mathbf{x}} + \hat{\mathbf{y}} - \hat{\mathbf{z}})$	
$\mathbf{b} = a\hat{\mathbf{y}}$	(1.3)	$\mathbf{b} = \frac{1}{2}a(-\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$	(1.4a)
$\mathbf{c} = a\hat{\mathbf{z}}$		$\mathbf{c} = \frac{1}{2}a(\hat{\mathbf{x}} - \hat{\mathbf{y}} + \hat{\mathbf{z}})$	

where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ represent the unit vectors along the orthogonal edges of the non-primitive cell.

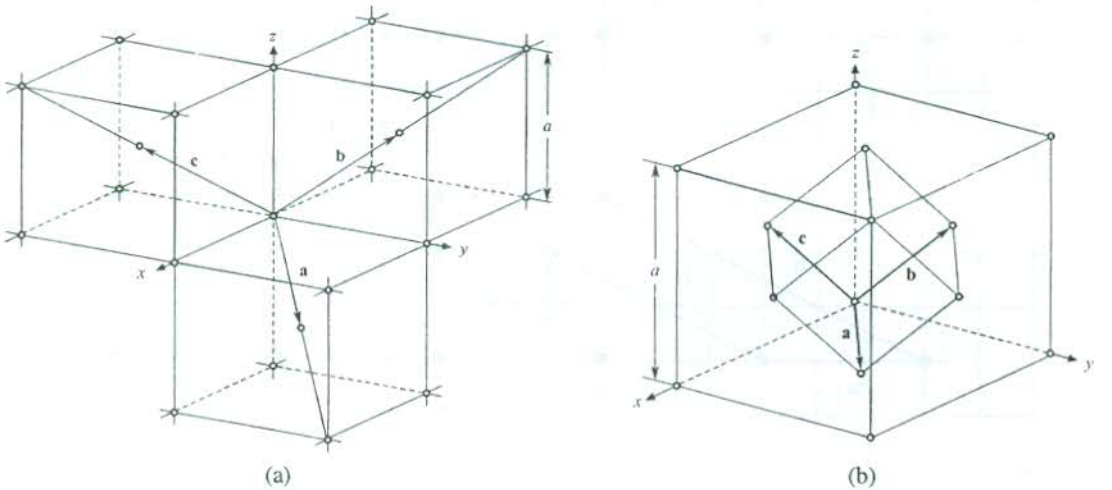


FIG. 1.6 (a) Orientations of the primitive translation vectors (**a, b, c**) for a BCC lattice. These form the edges of the primitive cell having the rhombohedral shape. The edge measures $\sqrt{3}\,a/2$ and the angle between the adjacent edges is 109°28′ (a is the size of the non-primitive cell). (b) The non-primitive and primitive cells of an FCC lattice. The primitive cell is a rhombohedron of size $a/\sqrt{2}$ with axes inclined at 60° to each other.

An alternative primitive cell, known as Wigner–Seitz cell, is of historic importance in the theory of Solid State Physics. The first proper electron energy bands calculation was made by Wigner and Seitz using this design of the primitive cell. The method of construction is as follows.

A lattice point is joined to all the nearby lattice points with the help of lattice vectors. Then, a plane perpendicular to each of these vectors, connecting the central lattice point, is drawn at the mid-point of the vector. The planes form a completely closed polyhedron which contains only one lattice point at its centre. The polyhedron serves as an ideal primitive cell and is named after Wigner and Seitz who are credited with having given the philosophy of its construction. Figure 1.7 shows the pictures of Wigner–Seitz cell of a two-dimensional lattice. It should be noticed that the Wigner–Seitz cell of BCC lattice can in effect be constructed by chopping off all the corners of its lattice cube halfway along a diagonal from the centre to a corner point. In this construction the lattice vectors connecting the lattice point at the centre and the next nearest lattice points are also involved. The resulting polyhedron turns out to be a truncated octahedron.

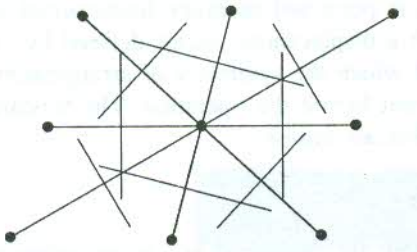


FIG. 1.7 The construction of the Wigner–Seitz cell for a two-dimensional space lattice.

Another common space lattice to which structures of many solids correspond is the FCC lattice. Its conventional unit cell has four lattice points and the three edge vectors are represented as for the BCC lattice (1.3). The primitive cell is again a rhombohedron with primitive translation vectors as depicted in Fig. 1.6(b). It is straightforward to check that they can be expressed as

$$\begin{aligned} \mathbf{a} &= \frac{1}{2}a(\hat{\mathbf{x}} + \hat{\mathbf{y}}) \\ \mathbf{b} &= \frac{1}{2}a(\hat{\mathbf{y}} + \hat{\mathbf{z}}) \\ \mathbf{c} &= \frac{1}{2}a(\hat{\mathbf{z}} + \hat{\mathbf{x}}) \end{aligned} \tag{1.4b}$$

It thus becomes obvious that in order to make a list of various types of crystal structures, we must know how many types of unit cells are possible. The task involves the compilation of knowledge of the symmetries of unit cells or symmetry operations under which a unit cell remains unchanged. The importance of symmetry operations, as already emphasized in the beginning of this chapter, has now become amply evident. An attempt to give a simple description of the symmetry operations is made below.

1.3 SYMMETRY OPERATIONS

An operation that takes the crystal into itself is called a symmetry operation. Imagine an observer placed inside a crystal. He is asked to close his eyes and then an operation is performed on the crystal. If the observer opens eyes after the operation is completed and does not find any change in his surroundings, the performed operation will be called a symmetry operation.

Symmetry operations may be grouped into three classes:

1. Translation operations
2. Point operations
3. Hybrid operations

1.3.1 Translation Operations

As already stated, these are the simplest of the symmetry operations. A translation operation is defined as the displacement of a crystal parallel to itself by a crystal translation vector.

The translation operation is permitted in every fundamental space lattice of crystals when a lattice point is moved through a displacement vector defined by (1.2). The final point of journey is another lattice point around which the symmetry of arrangement of lattice points is identical to that around the first lattice point before the operation. The fundamental space lattice of a crystal is commonly known as the *Bravais lattice*.

1.3.2 Point Operations

Point operations described below are performed at a point within a Bravais lattice or a crystalline matter. A simple description of these operations is given below.

The mirror reflection

In this operation, the reflection of a structure at a mirror plane m passing through a lattice point leaves the crystal unchanged. The mirror plane is composed of the atoms lying on the concerned imaginary plane. Two different mirror planes in a structure are demonstrated in Fig. 1.8.

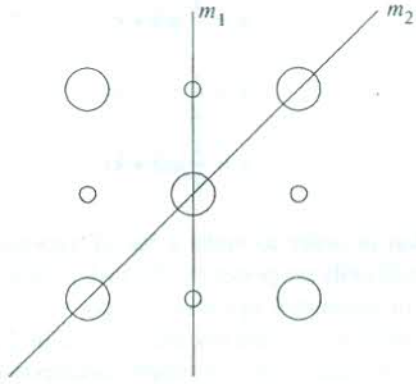


FIG. 1.8 Examples of two mirror planes in a two-dimensional crystal.

The inversion

Let the position vector of a lattice point be represented by $\mathbf{r}$, the origin being another lattice point (Fig. 1.9). If the crystal lattice has a lattice point with the position vector $-\mathbf{r}$, we say that the crystal has the inversion symmetry and the origin taken in the example is called the *centre of inversion* represented by symbol i. The inversion operation shows that the inversion of the lattice point A at the centre of inversion O (another lattice point) keeps the lattice unchanged. In this operation the point A goes to the point A' which is also a lattice point. After studying Section 1.4, the readers must verify that every Bravais lattice has inversion symmetry.

The rotation

It is a well-known fact that objects can be repeated by rotation. An object can be rotated about an axis through a certain angle so as to give the unchanged picture of the object to a stationary observer. The axis is called the *rotation axis*. If the angle θ, for which this condition is met, be given by

$$\theta = \frac{2\pi}{n} \qquad \text{with } n = 1, 2, 3, \ldots \tag{1.5}$$

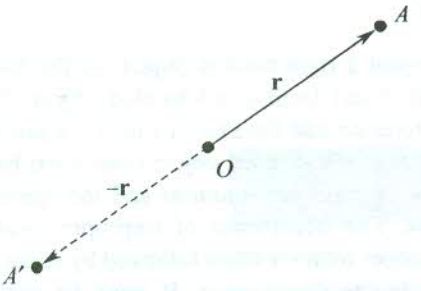


FIG. 1.9 The illustration of the inversion operation.

then the axis of rotation is termed n-fold. This implies that such n rotations of an object about its n-fold axis bring the object back to its identical configuration. This shows that every object has 1-fold axis of rotation.

These axes of rotation, also known as *symmetry axes*, are very common in the crystalline state of matter. The location of a 4-fold and a 6-fold axis is shown in Fig. 1.10 in two different crystals.

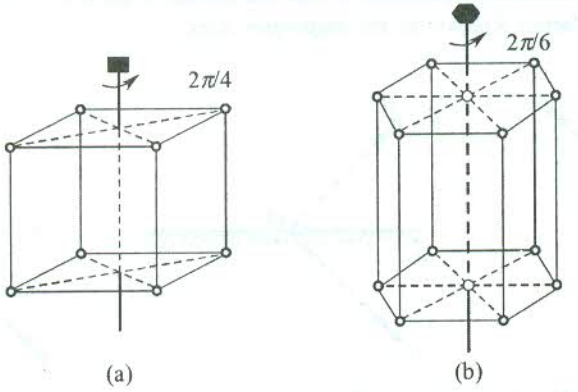


FIG. 1.10 (a) A 4-fold axis of rotation. (b) A 6-fold axis of rotation.

The symmetry axes most commonly observed in crystals are 1-fold, 2-fold, 3-fold, 4-fold, and 6-fold. The evidence for a 5-fold rotational symmetry is encountered in the structural studies of some materials, classified as quasicrystals (see Section 1.8). The lack of conviction in getting a periodic structure by arranging units with pentagonal faces is strongly supported by the absence of 5-fold symmetry in all generally talked about crystals.

The rotations discussed above always produce a congruent set and are known as proper rotations. But our discussion on rotational symmetry remains incomplete without making a reference to a very special type of rotation. The rotations that repeat a right-handed object from a left-handed one are quite frequent in crystals. They are known as *improper rotations.* A combination of such a rotation with the reflection operation takes a crystal into itself (the congruent set). It can be identified as a *rotoreflection* operation. It is one of the several hybrid symmetry operations which are required to explain or account for the structure of many crystals.

1.3.3 Hybrid Operations

The rotations discussed above repeat a right-handed object (or the basis of structure) from a right-handed object and a left-handed object from a left-handed object. Such a rotation resulting in a congruent set is called a *proper rotation* and the axis of rotation a *proper rotation axis*. On the other hand, there are rotations that repeat a right-handed object from a left-handed object (incongruent set) and vice versa. These are known as *improper rotations* and the operation is referred to as rotation about an improper rotation axis. The occurrence of improper rotation axes in crystals is well documented in literature. An improper rotation when followed by some other operation like reflection or translation sometimes leads to the congruence. It must be noticed that none of these two operations is independently a symmetry operation of the crystal whereas their combination is. Similarly, a reflection may combine with a translation to produce congruence in some crystals. The combined symmetry operation in these cases is named as a hybrid symmetry operation. Some of the hybrid symmetry operations that frequently occur in crystals are described below.

The rotoreflection

It is the combination of an n-fold rotation and a reflection at the plane, perpendicular to the rotation axis. Its general representation is n/m with $n = 1, 2, 3, 4, 6$. It can be shown that the rotoreflection $2/m$ is equivalent to the inversion operation, described earlier. Figure 1.11 explains this. The axis of rotation in this case is better known as the *improper axis*.

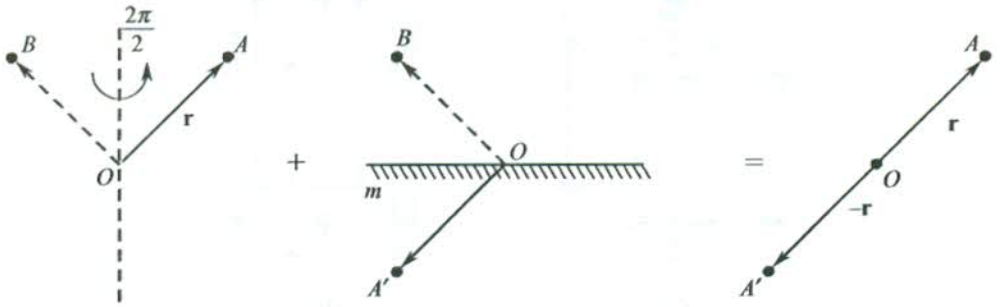


FIG. 1.11 The illustration of a rotoreflection $2/m$; it is equivalent to an inversion operation.

The rotoinversion

This operation includes an n-fold rotation followed by an inversion. Its common representation is $\bar{n}$, where $n = 1, 2, 3, 4, 6$. Thus a rotoinversion composed of a 3-fold rotation and an inversion is depicted as $\bar{3}$. The axis of rotation is again an improper axis.

The screw translation

Similar to the two operations explained above, an operation combining an n-fold rotation with a translation parallel to the rotation axis is found to be permitted by the symmetry of several crystals. Such an operation with a screw diad axis, denoted by 2_1 is demonstrated in Fig. 1.12.

The glide reflection

We can also combine a reflection with a translation parallel to the reflection plane to achieve congruence in some crystals. This combination is called a glide reflection. The translation component

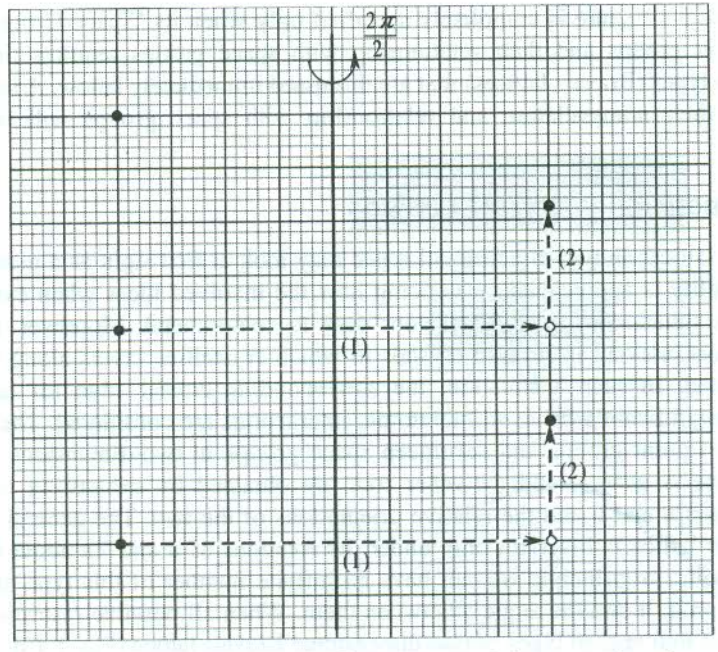


FIG. 1.12 The illustration of a screw translation operation, 2_1.

of a glide reflection is expressed as the combination of translations by fractions of repeat distances in a crystal. For example, the translation in a diagonal glide may be written as

$$\mathbf{t} = \frac{1}{2}\mathbf{a} + \frac{1}{2}\mathbf{b}, \quad \frac{1}{2}\mathbf{b} + \frac{1}{2}\mathbf{c} \quad \text{or} \quad \frac{1}{2}\mathbf{c} + \frac{1}{2}\mathbf{a}$$

The list of symmetry operations is not without a very special symmetry operation, known as the *identity operation*. This operation means doing no operation on the crystal or keeping the crystal undisturbed. Thus every crystal has the identity operation. It is generally denoted by the symbol *I*. In the context of symmetry operations the identity operation appears irrelevent at the first instance. But this has to be included in the list to meet the mathematical requirements of Group Theory which arms us with the quickest methods of making a quantitative calculation of many physical properties of crystals.

1.4 CRYSTAL TYPES

In order to have a complete list of different types of crystals, we basically require to know the number of fundamental space lattices, referred to as Bravais lattices. The number of these lattices is obtained by imposing restrictions on **a**, **b** and **c** and getting different arrangements of lattice points each of which is represented by the respective unit cell. The point symmetry operations of a Bravais lattice may be combined in different ways such that each of the combinations represents a different structure in terms of symmetry. Each combination which is a collection of certain symmetry operations is called a 'Point Group'. A symmetry operation in this reference is given another name 'Symmetry Element'. For example, we say that the centre of inversion i is a symmetry element of the octahedral point group O_h [see Table 1.2(b)]. It may, however, be remarked that the

symmetry of a crystal structure is completely specified only when its 'Space Group' is known. The description of point groups and their relationship with space groups requires the knowledge of fundamental lattice types (Bravais lattices) that are discussed immediately below. A suitable discussion on point groups, their nomenclature and relationship with space groups will be undertaken in Section 1.6.

1.4.1 Two-dimensional Crystal Lattices

The discussion on lattices in two-dimensional space may look at first sight as a matter of theoretical concern. But its utility is appreciated on noticing how the involved ideas when extended make the relatively complex situation in three-dimensions easy to comprehend. Moreover, two-dimensional lattices become most relevant in layered crystals in which crystalline layers along one crystal edge are largely separated. In such a situation, the corresponding crystal edge (or lattice parameter) becomes almost insignificant whenever interatomic interactions are considered. There are in all five Bravais lattices, ten point groups and seventeen space groups in two-dimensions. Out of the five Bravais lattices, one is general and the other four are obtained by exhausting the feasible axial relationships between **a** and **b** and the relative orientations of the two. The general lattice is termed *oblique lattice.* The lattice is invariant only under the rotations of π and 2π about the axes passing through a lattice point. In principle we can have an infinite number of this type of lattice as there are no restrictions on **a**, **b** and the angle γ between them. When all possible restrictions are imposed on **a**, **b** and γ, we get four special types of two-dimensional Bravais lattice [Fig. 1.13]. All these lattices are primitive except the second of the two derived from the same centred rectangular arrangement of lattice points.

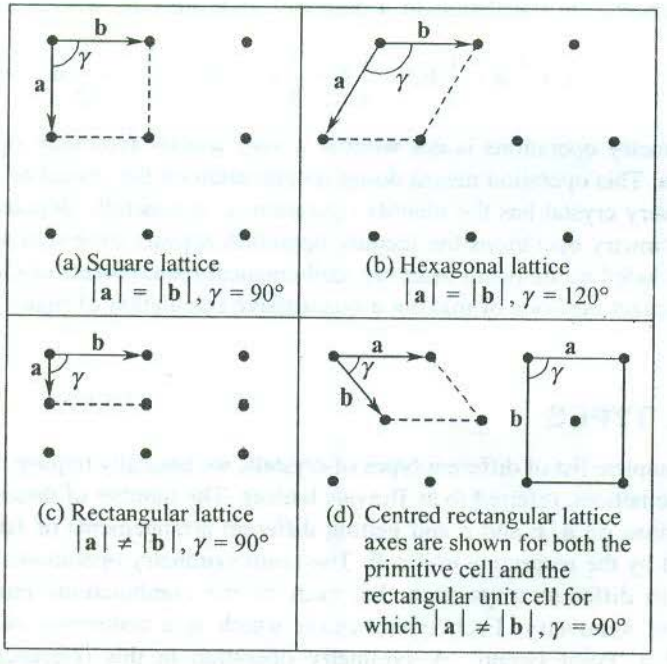


FIG. 1.13 The types (Bravais lattices) of two-dimensional lattices.

1.4.2 Three-dimensional Crystal Lattices

A three-dimensional unit cell is defined by vectors **a**, **b** and **c** representing its edges or crystal axes and the angles α, β and γ as defined in Fig. 1.14.

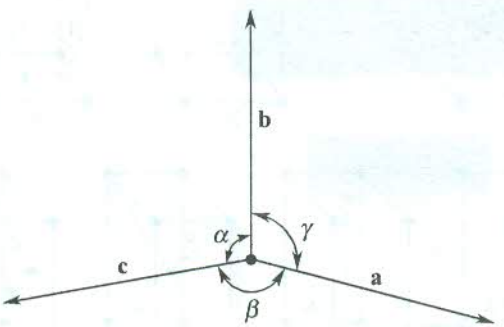


FIG. 1.14 Notation for angles between the crystal axes.

The number of Bravais lattices is 14 with 32 point groups and 230 space groups. Based on relationships between **a**, **b** and **c** in terms of magnitude and relative orientation (α, β and γ), the 14 types of unit cells are grouped in seven different classes of crystal lattices: triclinic, monoclinic, orthorhombic, tetragonal, cubic, trigonal and hexagonal. A description of the 14 Bravais lattices of three-dimensions along with the axial relationship for the class of crystal lattices to which each belongs is entered in Table 1.1. The shapes are shown in Fig. 1.15.

Table 1.1 The Bravais lattices in three dimensions

Class	*Lattice type*	*No. of lattices*	*Axial relationship*
Triclinic	P	1	$a \neq b \neq c$ $\alpha \neq \beta \neq \gamma$
Monoclinic	P, B	2	$a \neq b \neq c$ $\alpha = \gamma = 90° \neq \beta$
Orthorhombic	P, B, F, C	4	$a \neq b \neq c$ $\alpha = \beta = \gamma = 90°$
Tetragonal	P, C	2	$a = b \neq c$ $\alpha = \beta = \gamma = 90°$
Cubic	P, F, C	3	$a = b = c$ $\alpha = \beta = \gamma = 90°$
Trigonal	P or R	1	$a = b = c$ $\alpha = \beta = \gamma < 120°, \neq 90°$
Hexagonal	P	1	$a = b \neq c$ $\alpha = \beta = 90°$ $\gamma = 120°$

Representation of symbols: P—primitive, B—base centred, C—body centred, F—face centred.

The number of lattice points in unit cell can be calculated by appreciating the following:

Contribution of the lattice point at the corner = 1/8th of the point
Contribution of the lattice point at the face = 1/2 of the point
Contribution of the lattice point at the centre = 1 of the point

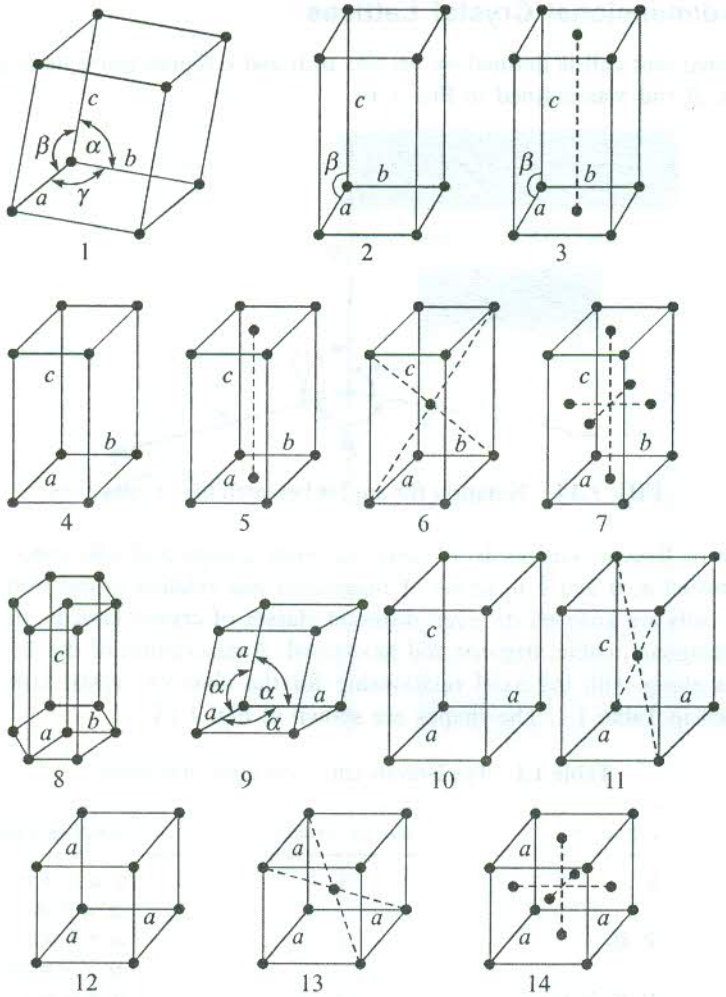


FIG. 1.15 The three-dimensional lattices; 1: Triclinic, 2–3: Monoclinic, 4–7: Orthorhombic, 8: Hexagonal, 9: Trigonal, 10–11: Tetragonal, 12–14 Cubic.

Every type of unit cell is characterized by the number of lattice points (not the atoms) in it. For example, the number of lattice points per unit cell for simple cubic (SC), body centred cubic (BCC) and face centred cubic (FCC) lattices are one, two and four, respectively. Our knowledge about unit cell may not be complete without having a quantitative estimate of its volume. It can be calculated using the relation

$$V_c = \mathbf{a} \cdot [\mathbf{b} \times \mathbf{c}] \tag{1.6}$$

where V_c stands for the volume of the cell. The **a**, **b** and **c**, defined so far as the measure of unit cell edges, are more popularly known as *lattice parameters.*

1.5 INDICES OF A LATTICE DIRECTION AND A LATTICE PLANE

The exercise of describing a crystal structure essentially amounts to locating the positions of different atoms in the unit cell. But the study of many physical properties of crystals becomes more meaningful and complete with the knowledge of directions and various atomic planes visualized within their crystalline medium. These are identified by using the index systems discussed below.

Indices of a lattice direction

A vector is drawn along (or parallel to) the desired direction from the origin fixed for the axes system. The components of this vector along the crystal axes are determined and reduced to smallest integers. These integers when written within square brackets represent the indices of the direction in question. Some of the prominent directions in a cubic unit cell are shown in Fig. 1.16.

A bar on an index means that the respective component of the vector is negative.

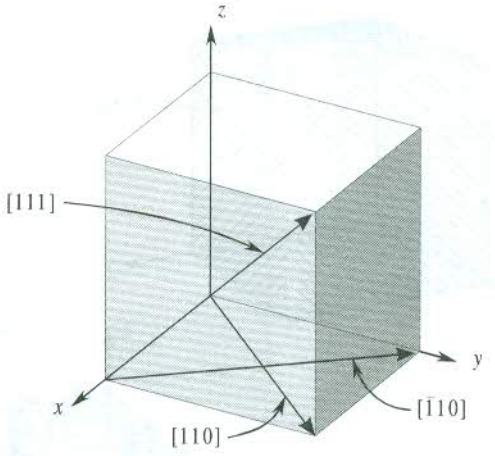


FIG. 1.16 Some prominent directions in a cubic crystal.

Indices of a lattice plane

Every crystal plane is identified by its orientation which is characterized by plane's intercepts with the crystal axes. The equation of a plane having intercepts p, q and r with x, y, z axes is given by

$$\frac{x}{p}+\frac{y}{q}+\frac{z}{r}=1 \tag{1.7}$$

The form of this equation has been exploited by Miller to develop an index system, for whom the indices are named—Miller indices. The Miller indices of a plane are determined by the following procedure.

Let the intercepts of a plane with **a**, **b** and **c** measure as pa, qb and rc, respectively. Reduce the set $\frac{1}{p},\frac{1}{q},\frac{1}{r}$ to the three integers h, k, l having the same ratio as $\frac{1}{p},\frac{1}{q},\frac{1}{r}$. This set of indices, when written in parentheses as (h k l) gives the Miller indices of the plane. For example, take a plane

having the intercepts of lengths $2a$, b and $2c$ with the three crystal axes. In order to find its Miller indices, first take the reciprocals of 2, 1 and 2 getting $\frac{1}{2}, 1, \frac{1}{2}$ and then find the integers that are in the same ratio as $\frac{1}{2}, 1, \frac{1}{2}$. This gives the Miller indices (121).

Some prominent planes of a cubic unit cell can be seen in Fig. 1.17. If a plane intersects a certain crystal axis on its negative side, a bar is put over the corresponding Miller index. For example, (123) and $(1\bar{2}3)$ are two parallel planes, their intercepts with the **b** axis being on the two sides of the origin. Examine planes (110), (220) and (330) in Fig. 1.17. They are parallel to each other and lie on the same side of the origin. In rare situations we need to distinguish them and generally they are represented by the smallest common indices obtained by dividing through with the largest common factor. Thus all these three planes may be represented by (110).

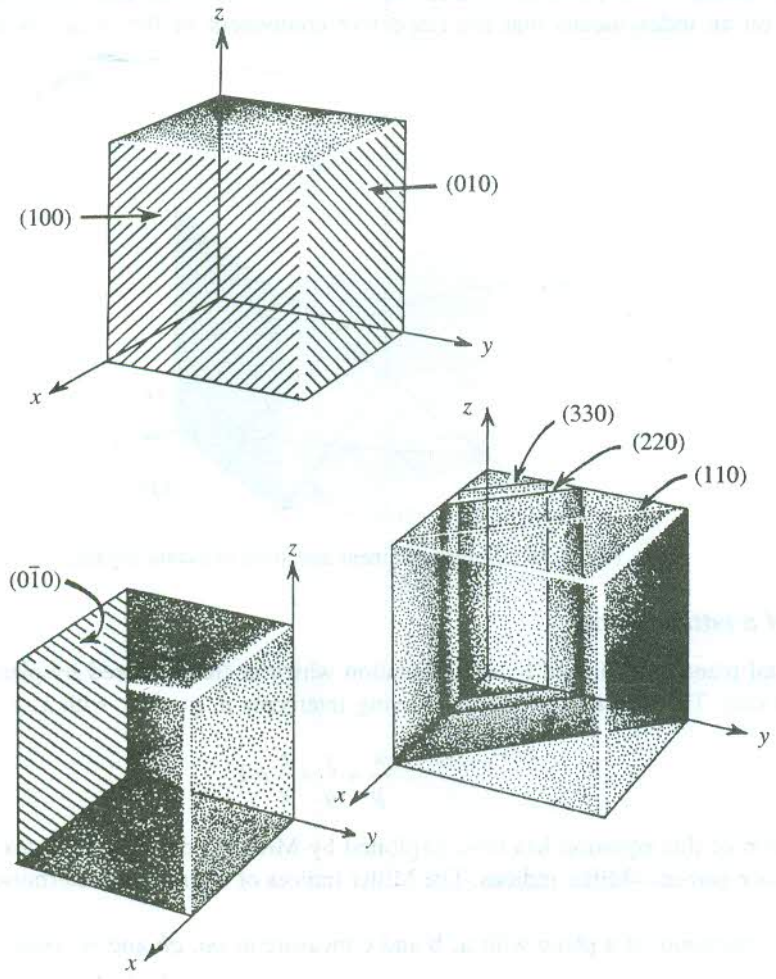


FIG. 1.17 Some prominent planes of a cubic crystal.

1.6 CRYSTAL POINT GROUPS AND SPACE GROUPS

1.6.1 Point Groups

In continuation with the discussion on crystal types initiated in Section 1.4, we would like to repeat that the symmetry elements may be combined with one another in different ways. In the converse sense, this means that every crystal structure may be described by a combination of symmetry elements (symmetry operations). But the symmetry elements in each such combination are required to satisfy certain conditions. For example, when one symmetry operation A is followed by another operation B, the resultant effect is capable of being produced by a third symmetry operation C, that is $A * B = C$, where A, B and C are members of a combination of certain symmetry elements, permitted by the symmetry of the concerned crystal lattice. Furthermore, the symmetry elements in a combination should also satisfy the following mathematical conditions:

$$\begin{aligned} (A \times B) \times C &= A \times (B \times C) \qquad \text{(the associative law)} \\ A \times I &= A \end{aligned} \tag{1.8}$$

where I is the identity operation.

$$A^{-1} \times A = I$$

where A^{-1} is an operation which is the reverse of operation A.

A set of symmetry elements that satisfies the above conditions is said to form a group, commonly known as the 'point group', since all the operations contained in the group are performed at a point in the crystal lattice. There are in total 32 distinct combinations of symmetry operations performed at a point in various crystals. This number exhausts all the possible ways of combining the symmetry elements such that the symmetry elements in each combination meet the requirements imposed by the mathematical conditions (1.8). This gives 32 point groups in three-dimensional lattices. Twenty-seven of these are non-cubic and five are cubic point groups. There exist two standard nomenclature systems for point group—the Schönflies and international. The Schönflies symbols are most common in group theoretical and spectroscopic illustrations. The composition of these groups together with their symbols in both the nomenclature systems is described in Table 1.2.

It is an established practice in crystallography to represent the point groups by the so-called stereographic projections developed essentially to obtain a systematic classification of the exposed surfaces of naturally grown crystals. The points of intersection of surface normals with a sphere are marked. These points are projected on to the plane that is perpendicular to the highest order symmetry axis. Points referring to the points of intersection on the upper-half of the sphere are indicated by dark full circles and those corresponding to intersections below the plane are represented with cross marks or open circles. Thus the highest order symmetry axis passes through the centre in a systematic representation of point groups. As an example, the stereographic projection of point group $D_{3d} \equiv \bar{3}\frac{2}{m}$ is shown in Fig. 1.18.

1.6.2 Space Groups

The information about a crystal structure is complete only when its point group and space group both are known. The precise knowledge about the composition and the symmetry of the basis of atoms attached to a lattice point is a prerequisite to the determination of the space group of a crystal

Table 1.2(a) Twenty-seven (27) non-cubic crystal point groups

Schönflies	*International*	*Meaning in terms of symmetry elements*	*Number*
C_n	n	These groups have only an n-fold axis of rotation (n = 1, 2, 3, 4, 6).	5
C_{nv}	$n\ m\ m$ (n even) $n\ m$ (n odd)	The groups have an n-fold rotation axis, a mirror plane containing the rotation axis, and as many additional mirror planes as may be required by the presence of the n-fold axis (n = 2, 3, 4, 6).	4
C_{nh}		In addition to the n-fold rotation axis, there is a single mirror plane perpendicular to the rotation axis.	
	$\left(\frac{n}{m}\right)$	n = 2, 4, 6 (even)	
	$\bar{n}$	n = 1, 3 (odd)	5
S_n	$\bar{n}$	These groups have only an n-fold rotoreflection axis (n = 2, 4, 6).	3
D_n	$n\ 2\ 2$ (n even) $n\ 2$ (n odd)	These groups contain an n-fold axis of rotation, a 2-fold rotation axis perpendicular to the n-fold axis, and as many additional 2-fold axes as the presence of the n-fold axis demands (n = 2, 3, 4, 6).	4
D_{nh}	$\frac{n}{m}\frac{2}{m}\frac{2}{m}$	In addition to all the elements of D_n, there is a mirror plane perpendicular to the n-fold rotation axis.	4
	$\left(\frac{n}{mmm}\right)$	n = 2, 4, 6 (even)	
	$\bar{n}\ 2\ m$	n = 3 (odd)	
D_{nd}		In addition to all the elements of D_n, there are mirror planes containing the n-fold rotation axis such that the planes bisect the angles between the 2-fold axes of rotation.	2
	$\bar{n}\ 2\ m$	n = 2 (even)	
	$\bar{n}\ \frac{2}{m}$	n = 3 (odd)	

structure whose point group is known. For a given crystal system there can be a certain number of point groups, i.e. the number of different combinations of the symmetry elements of the system. The group of atoms comprising the basis can have any of these point group symmetries. One point group symmetry for the basis gives a space group for each of the possible Bravais lattices of the crystal system. For example, there are five point groups and three Bravais lattices (P, F and C) for a cubic system [see Tables 1.1 and 1.2(b)]. So, five different symmetries for the basis are possible in this case. With these possibilities of symmetry for the basis in each of the three Bravais lattices, 15 different crystal structures can occur, represented by 15 different space groups.

Following the procedure described above, the total number of space groups for real crystals have been determined. In considering these combinations, one must be reminded that symmetry elements

Table 1.2(b) The five (5) cubic crystal point groups

Schönflies	*International*	*Meaning in terms of symmetry elements*
T	2 3	The tetrahedral point group (in a crystal the alternate corners of the cube are vacant). It contains three mutually perpendicular 2-fold axes, plus four 3-fold axes, angles between which are bisected by the 2-fold axes.
T_d	$\bar{4}$ 3 m	In addition to all the elements of T, there is a mirror plane through each pair of 3-fold axes (i.e. two mutually perpendicular mirror planes through each 2-fold axis. In all, there are thus six mirror planes.
T_h	$\frac{2}{m}$ $\bar{3}$ (m 3)	The group has all symmetry elements of T, plus a centre of inversion i.
O	4 3 2	The octahedral point group. It contains three mutually perpendicular 4-fold axes and four 3-fold axes which have the same orientation with respect to one another as the 2-fold and 3-fold axes of the tetrahedral point group.
O_h	$\frac{4}{m}\,\bar{3}\,\frac{2}{m}$ (m 3 m)	The group has all symmetry elements of O, plus a centre of inversion i.

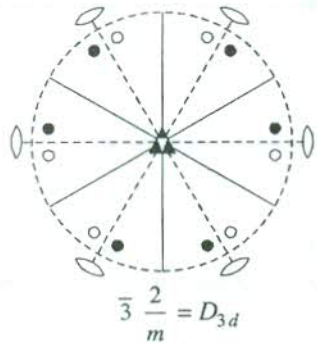


FIG. 1.18 Stereographic projection of the symmetry elements of point group $D_{3d} \equiv \bar{3}\,\frac{2}{m}$. The symbols ⬯ and ▲ denote 2- and 3-fold axes of rotation, respectively. The full lines represent mirror planes. If the plane of the figure were to be a mirror plane, the outer circle would be shown as a continuous circle.

permitted for a space lattice can have translational components as well. This demands that in addition to thirty-two point groups, the symmetry groups with identical angular relationships but having screw axes in place of proper rotation axes and glide planes in place of pure reflection planes should also be taken into account. With the inclusion of these groups, the total number of ways in which the symmetry groups may be combined with space lattices in three dimensions are found to be 230. This gives the total number of space groups for real crystals. It should surprise none that this number would be mere 14 for structures with a completely (spherically) symmetric basis (e.g. a monatomic basis). The space group of a crystal is usually represented on a plane by a repetitive pattern of the allowed point symmetry elements, by using standard symbols for different symmetry elements. Both the nomenclature systems, Schönflies and International are used. Readers interested in more details may consult Buerger's *Elementary Crystallography* (Wiley, 1963).

1.7 COMMON CRYSTAL STRUCTURES

1.7.1 The Simplest Crystal

The simplest crystal structure that we can think of is that of simple cubic symmetry with a basis of one atom. Polonium is the lone known example of this class in nature. The atoms are positioned at the corners of a cube. Thus the crystal has just one atom per primitive cell.

1.7.2 The CsCl Structure

The structure of a CsCl crystal is shown in Fig. 1.19. There is an atom at the centre of the cube surrounded by eight atoms of the other type at the corners. The space lattice is simple cubic (SC) and the primitive cell has one formula unit. The basis is composed of two atoms situated at 000 and $\frac{1}{2}\frac{1}{2}\frac{1}{2}$, positions being measured in units of the lattice parameter. Notice that this structure cannot be interpreted as body centred cubic (BCC). If the cube centre is treated as the second lattice point,

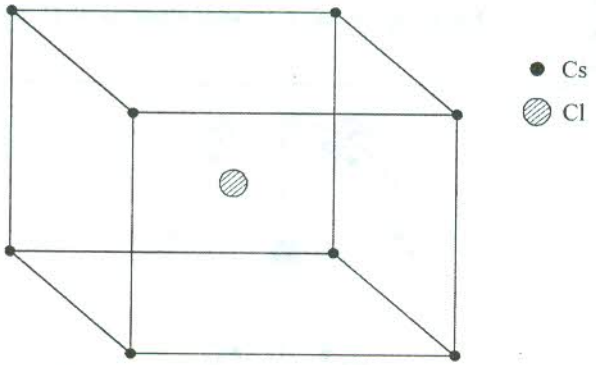


FIG. 1.19 The unit cell structure of CsCl crystal.

its basis is found to be different from the one at the corner which is not permitted, since every lattice point is required to have an identical basis in a crystal. The nearest neighbours are the atoms of the other element and their number, also known as the *coordination number,* is eight. The lattice parameter of some ionic crystals having this structure is given below:

Crystal	*Lattice parameter* (Å) (Å = 10^{-10} m)
CsCl	4.11
CsBr	4.29
CsI	4.56
TlCl	3.84
TlBr	3.97
TlI	4.18

1.7.3 Crystals of Alkali Metals

Crystals of alkali metals (Li, Na, K, Rb, Cs) are typical representatives of the body centred cubic (BCC) structure. The unit cell is non-primitive with two lattice points and the basis of one atom. The positions of atoms at the two non-equivalent lattice points are 000 and $\frac{1}{2}\frac{1}{2}\frac{1}{2}$. The coordination number is eight. As stated earlier, its primitive cell is rhombohedral [see Fig. 1.6(a)] and its edges in vector form are expressed by (1.4a).

Crystal	*Lattice parameter* (Å)	
Li	3.50	at room temperature
Na	4.28	
K	5.33	
Rb	5.62	at 92 K
Cs	6.05	

1.7.4 Crystals of Noble Metals

The crystalline noble metals (Cu, Ag, Au) represent another important structure, the face centred cubic (FCC). The unit cell has four lattice points with the basis of one atom. The positions of the atoms in the unit cell are:

$$000, \frac{1}{2}\frac{1}{2}0, \frac{1}{2}0\frac{1}{2}, 0\frac{1}{2}\frac{1}{2}$$

The coordination number is 12. It will become clear later when close-packed structures are discussed. One of the two close-packed structures has FCC symmetry. The primitive cell of an FCC crystal is a rhombohedron with primitive translation vectors as already shown in Fig. 1.6(b).

Crystal	*Lattice parameter* (Å)
Cu	3.61
Ag	4.08
Au	4.07

1.7.5 The NaCl Structure

The space lattice of the NaCl crystal is FCC. The arrangement of Na and Cl atoms on this lattice is shown in Fig. 1.20(a). Notice that none of the sodium atoms is located at any lattice point. One of them is at the centre and the others are at the mid-points of the 12 edges of the cube. An atom on the edge contributes 1/4th of the atom. Thus the unit cell gets three sodium atoms from those on the edges (12/4 = 3) and one from that at the centre. This gives four sodium atoms in the cell. Similarly, the contribution of chlorine atoms to the unit cell comes to four [(8/8) + (6/2) = 4]. A single unit cell accommodates four formula unit cells of NaCl. The positions of atoms in the unit cell are:

$$\text{Na: } \frac{1}{2}\frac{1}{2}\frac{1}{2} \quad 00\frac{1}{2} \quad 0\frac{1}{2}0 \quad \frac{1}{2}00$$

$$\text{Cl: } 000 \quad \frac{1}{2}\frac{1}{2}0 \quad \frac{1}{2}0\frac{1}{2} \quad 0\frac{1}{2}\frac{1}{2}$$

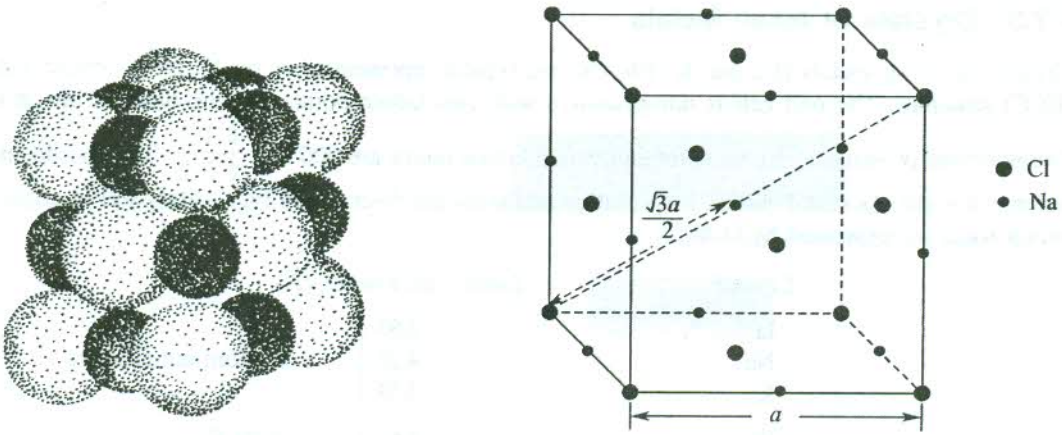


FIG. 1.20 (a) The arrangement of Na and Cl (shown bigger) atoms on the unit cell of NaCl crystal. (b) Locations of Na and Cl (shown bigger) atoms in the cube representing the unit cell. The separation between the basis partners is half the length of the body diagonal $\sqrt{3}a/2$.

The basis is composed of two atoms (Na and Cl) which are separated by half the length of the body diagonal, $\sqrt{3}a/2$, where a represents the length of the cube edge. The basis partner of each of the chlorine atoms at the corners is the sodium atom located at the centre whereas each of those at the face centres has partners in sodium atoms lying on two such edges as the distance between Na and Cl atoms in the basis remains $\sqrt{3}a/2$ [Fig. 1.20(b)]. This shows that 1/8th of a formula unit is associated with the lattice point at the corner and 1/2 that at the face centre, explaining the distribution of four formula units among four lattice points of the FCC lattice. In this structure the nearest neighbours are the atoms of different type and their number is six.

Some of the crystals, representative of NaCl arrangement, along with their lattice parameter (cube edge) are included in the following table.

Crystal	*Lattice parameter* (Å)
NaCl	5.63
LiH	4.08
KBr	6.50
RbI	7.33
NH_4I	4.37
NiO	4.17
UO	4.92
PbS	5.92

1.7.6 Close-packed Structures

The term *close-packing* needs a careful discussion. We may talk of the close-packed structure of a crystal of specific symmetry by referring to the structure with the smallest possible unit cell. Thus for every type of unit cell, there can be a maximum degree of closeness to which the atoms may be packed. The degree of closeness is represented by 'Packing Fraction', expressed as

$$\text{Packing fraction} = \frac{\text{Volume of the unit cell actually occupied by atoms}}{\text{Total volume of the unit cell}}$$

Considering unit cells having the basis of one atom, one can easily calculate the packing fraction of the closest-packed cells of different symmetries. Figure 1.21 shows the picture of a closest-packed BCC cell. Calculations show that the packing fraction is the highest in hexagonal and FCC unit cell and it has the same value (0.74) for both of these structures. This reveals that the packing in hexagonal and FCC structures can be made ideally the closest. But conventionally these two are referred to as the close-packed structures in literature. This point must be clearly borne in mind while using this nomenclature.

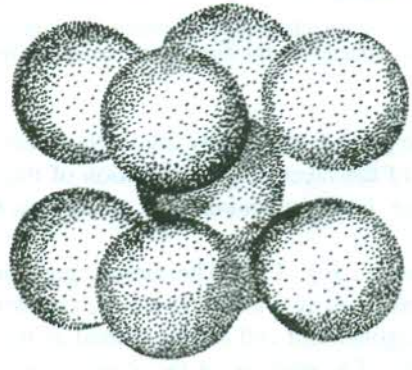


FIG. 1.21 The arrangement of atoms in the closest-packed BCC crystal.

It is desirable to appreciate that the ideas mentioned just above can be visualized if one tries to arrange identical balls in all possible regular arrays. We observe that the closest-packed hexagonal and FCC patterns maximize the packing fraction.

While giving a description of the hexagonal close-packed structure, the way of arranging identical balls to get a close-packed FCC structure is explained. The main feature of the close-packed FCC structure is that the two atoms on the corners along the face diagonal touch the face centred atom. With this information its packing fraction can be easily calculated and this is left as an exercise for the reader.

1.7.7 Hexagonal Close-packed Structure (HCP)

When we try to make the closest arrangement of identical balls by placing them on a plane surface, we find that the largest number of balls that can touch a certain ball is six. This forms the layer *A* in Fig. 1.22. In arranging the closest packing of balls above this layer, we observe that a ball of layer *A* is in contact with three balls of the new layer, marked as *B*. Similarly, we can visualize in a layered structure like this that a ball of layer *A* will be in contact with three balls of the layer just below the layer *A*. It shows that in the closest-packed arrangement of identical atoms, an atom touches 12 atoms. This means that the maximum number of the nearest neighbours that is possible in the resulting crystalline structure is 12. If the building of layers is continued, we have two ways of arranging balls in the third layer *C* above the layer *B*. One way is to place balls above holes left in the layer *A* and not occupied in the layer *B*. The repetition of this arrangement of layers (*ABC ABC ABC* ...) produces an FCC structure in which the layer *A* forms its (111) plane. The other way

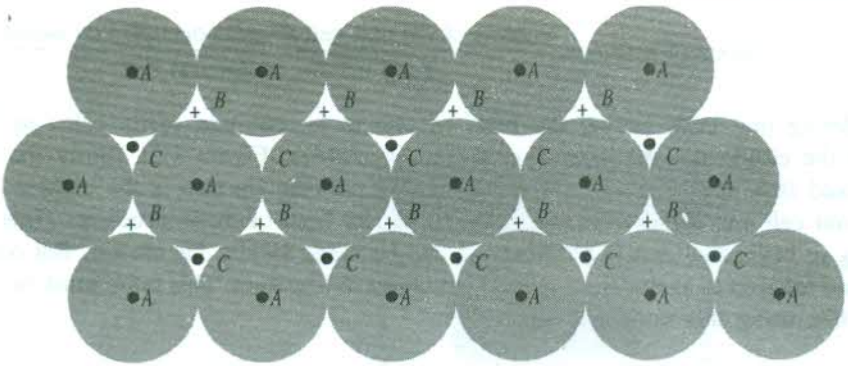


FIG. 1.22 Three successive layers of the close-packed structure of identical spheres. The points *A*, *B*, *C* represent the centres of spheres in the first layer, second layer (in the middle) and third layer, respectively.

of arranging the third layer is to place balls exactly above the centres of balls in layer *A*. Thus the third layer becomes the replica of the layer *A*. The repetition of this arrangement of layers (*AB AB AB* ...) results in an HCP structure. In this structure, layer *A* serves as the basal plane of the hexagonal unit cell.

The HCP structure is shown in Fig. 1.23. The atomic positions do not represent the lattice points. The structure's basis consists of two identical atoms, i.e. two identical atoms are associated with each lattice point. The hexagonal unit cell can be treated as the combination of three rhombus based right prism shape unit cells. The position of the atom lying within a rhombohedron in layer *B* can be determined on the basis of simple geometrical considerations of the closest packing. Its height above the base is *c*/2. The picture of the arrangement of the atoms *P*, *Q*, *R*, *S* and *M* of Fig. 1.23(b) looks as shown in Fig. 1.24(a).

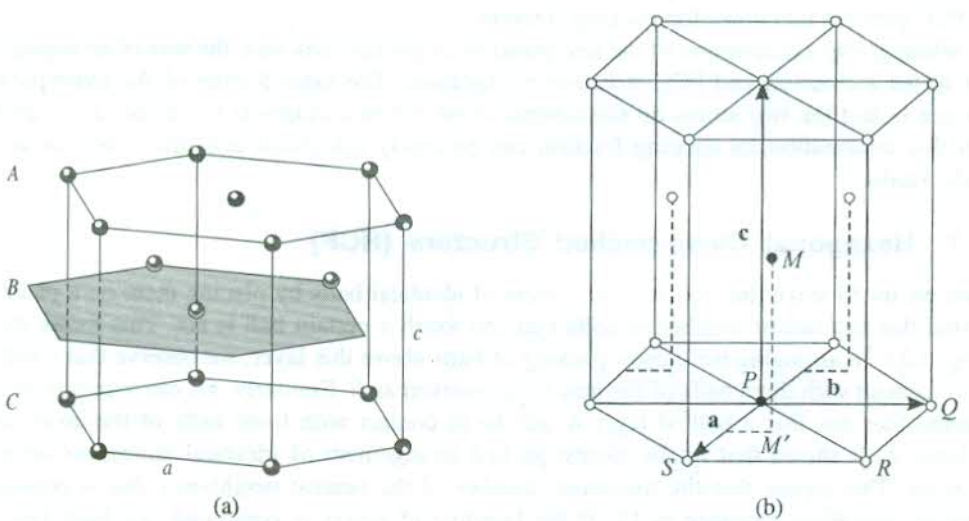


FIG. 1.23 (a) The hexagonal close-packed structure. (b) The primitive axes of the HCP crystal. The **c** is normal to the plane of **a** and **b**. The two atoms of one basis are shown as solid circles.

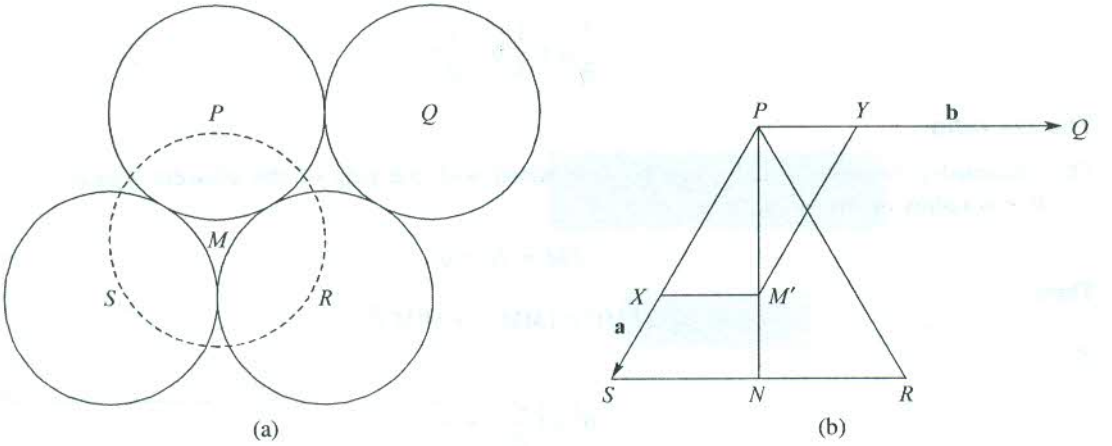


FIG. 1.24 The atoms in the basal plane (centred at *P*, *Q*, *R*, *S*) and the atom in the mid-plane (centred at *M*) of the primitive HCP structure.

The point M' is the foot of the perpendicular dropped from M on the plane PRS and coincides with the centroid of the equilateral triangle PRS [Fig. 1.24(b)]. Thus

$$PM' = \frac{2}{3} PN = \frac{2}{3} \times \frac{\sqrt{3}\, a}{2} = \frac{a}{\sqrt{3}}$$

Therefore,

$$PX = \frac{2}{3} PS \quad (\because XM' \parallel^l \text{ to } SN)$$

$$= \frac{2}{3} a$$

From $\Delta PM'X$,

$$(XM')^2 = (PX)^2 - (PM')^2$$

$$= \left(\frac{2}{3} a\right)^2 - \left(\frac{a}{\sqrt{3}}\right)^2$$

$$= \frac{a^2}{9}$$

or

$$XM' = \frac{a}{3}$$

This gives the position of M, the other atom in the basis as

$$\left(\frac{2}{3} a, \frac{1}{3} a, \frac{c}{2}\right)$$

with respect to the origin at P, the first atom in the basis. The position vector of M can be expressed as

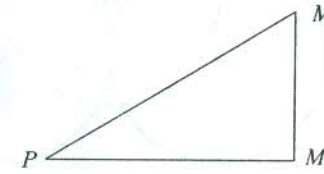
$$\frac{2}{3}\mathbf{a}+\frac{1}{3}\mathbf{b}+\frac{1}{2}\mathbf{c}$$

The c/a ratio

The relationship between a and c can be determined with the help of the adjacent figure.

If r is radius of the atom, then

$$PM = 2r = a$$



Then

$$(PM)^2 = (MM')^2 + (PM')^2$$

or

$$a^2 = \left(\frac{c}{2}\right)^2 + \frac{a^2}{3}$$

or

$$\frac{c}{a} = \left(\frac{8}{3}\right)^{1/2} = 1.633$$

This value is the ideal value for the HCP structure composed of identical atoms whose shape is assumed as exactly spherical. The ratio departs from this theoretical value in crystals as can be seen from the table given below:

HCP crystal	a (Å)	c (Å)	c/a
Be	2.28	3.59	1.574
Mg	3.20	5.20	1.625
α–Co	2.51	4.11	1.637
Zn	2.65	4.93	1.860

Packing fraction

The packing fraction can be easily calculated by using the primitive cell of the HCP structure. The primitive cell is a rhombus based right prism. Three such primitive cells combine to produce the non-primitive hexagonal unit cell. There are two atoms in a primitive cell.

$$\text{Packing fraction} = \frac{\text{Volume of two atoms}}{\text{Volume of the primitive cell}}$$

$$= \frac{2 \times \frac{4}{3}\pi r^3}{[\mathbf{a} \times \mathbf{b}] \cdot \mathbf{c}}$$

$$= \frac{\pi}{3\sqrt{2}} \quad (\text{use } a = 2r)$$

$$= 0.743$$

It can be shown that the closest-packed FCC structure has the same packing fraction as well. This value for some other standard crystal types is as under:

Crystal type	*Packing fraction*
Simple cubic	0.52
BCC	0.68
Diamond	0.34

The above data indicates that the HCP and the close-packed FCC structures have the closest-packing. Approximately, two-third of the metals crystallize either in HCP or in FCC structure. Most of the remaining one-third that are not close-packed are alkali metals and transition metals. They tend to solidify in a BCC symmetry. In common terminology the two closest-packed structures are referred to only as close-packed structures. The diamond structure, to be described in the following section, appears as the most empty lattice. The basis of these unique phenomena will be discussed in Chapter 2.

1.7.8 Diamond Structure

The space lattice of the diamond structure has FCC symmetry. Its unit cell has in all eight atoms with a basis of two atoms located at 000, $\frac{1}{4}\frac{1}{4}\frac{1}{4}$. From this structure it is impossible to derive a primitive cell whose basis may be composed of only one atom. The structure may be visualized as a combination of two FCC lattices with one being displaced from the other along the body diagonal by one-fourth of its length. The arrangement of atoms as projected on a face of the unit cell is shown in Fig. 1.25. The dark shaded atoms are positioned at $\frac{1}{4}\frac{1}{4}\frac{1}{4}, \frac{3}{4}\frac{3}{4}\frac{1}{4}, \frac{3}{4}\frac{1}{4}\frac{3}{4}$ and $\frac{1}{4}\frac{3}{4}\frac{3}{4}$ and each of them makes a tetrahedral bond with one atom at the nearest neighbour and three centred on the nearest faces (Fig. 1.26). Thus the coordination number is four and not 12 as in the closest-packed structure with a maximum of only 34 per cent of the total volume being actually occupied by atoms.

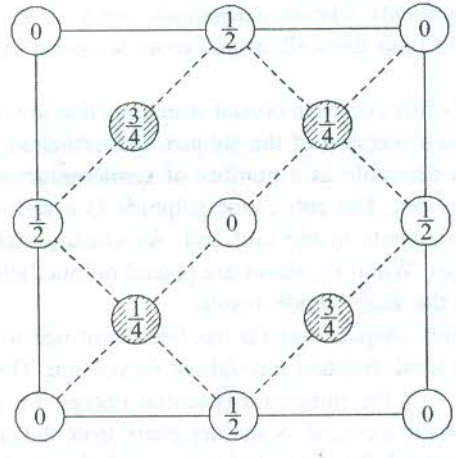


FIG. 1.25 Projection of atomic arrangement on a cube face in a diamond unit cell. The unit cell consists of two similar FCC lattices, displaced along the body diagonal of the cell by one-fourth of its length. The positions 0 and 1/2 refer to atoms on one lattice while 1/4 and 3/4 give positions of atoms on the second lattice.

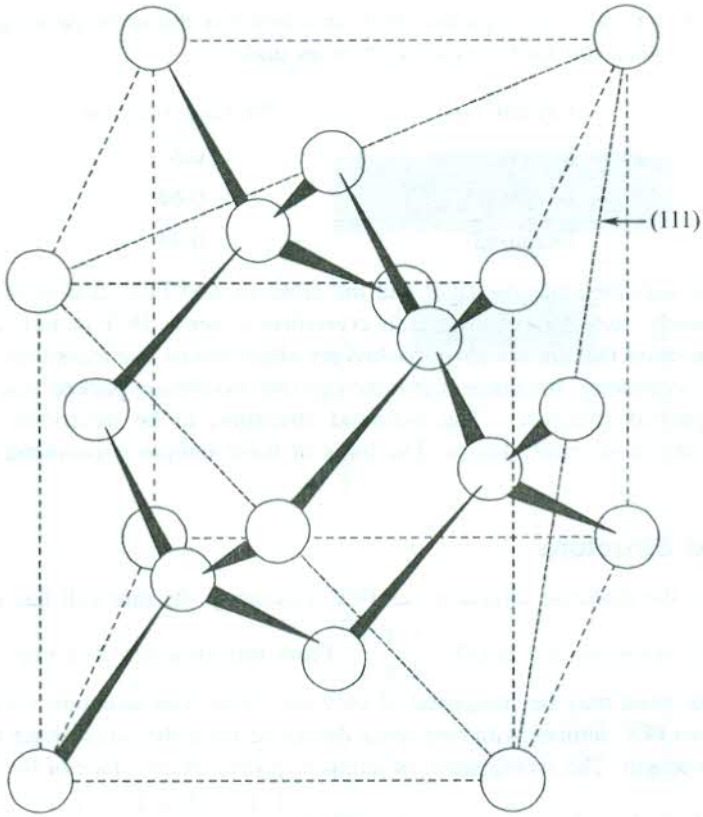


FIG. 1.26 The bonding model of the diamond structure.

Figure 1.26 shows that the diamond structure can be described as a stacking of sheets made up of continuously linked pluckered hexagonal rings of atoms parallel to (111) planes of the cubic crystal. Crystals of carbon (diamond), silicon, germanium and grey tin are well known examples of diamond structures. The edge of their unit cell around room temperature measures as 3.56, 5.43, 5.65 and 6.46 Å, respectively.

The description of several other common crystal structures that are important crystallographically cannot find a place in the present version of the subject. Nevertheless, a brief reference to the zinc blende structure is very much desirable as a number of semiconductors having this structure form the basis of certain current devices. The cubic zinc sulphide is a well-known representative of this structure with atoms of two elements in the unit cell. As already stated, the diamond lattice is a combination of two FCC lattices. When Zn atoms are placed on one lattice and S atoms on the other, the conventional unit cell of the zinc blende results.

The whole discussion in this chapter thus far has been confined to ideal crystals. But in reality no crystal of finite size has the ideal structure throughout its volume. Thermodynamic considerations, such as the requirement of having the minimum potential energy for a configuration to be stable, control the arrangement of atoms in a crystal. Some departure from the crystallographically prescribed structure is observed in every crystal. These deviations are called imperfections. In no way we should

infer that the presence of imperfections always affects the usefulness of a crystal adversely. In many applications some of the imperfections prove as a boon and make the material in question extremely useful for applications. The study of various imperfections and their effect on the properties of crystals will be taken up in Chapter 12.

1.8 QUASICRYSTALS

Some materials diffract electrons like single crystals but have point group symmetry that is inconsistent with lattice translations. This structure was discovered by Shechtman et al.[1] in 1984 in grains of size up to 2 μm of rapidly cooled alloys of Al with 10–14 atomic% Mn, Fe or Cr. It has long-range orientational order. The translational order is, however, not periodic and the structure does not have a rotational point symmetry. Materials showing this new structure have many of the properties of crystals and hence are classified as quasi periodic crystals or simply quasicrystals, in short. Unlike in crystals, the basis of structure (a group of atoms/molecules) does not repeat periodically in a quasicrystal. The repeating basis in this case is called the *unit cell*. An ideal quasicrystal is constructed by the infinite repetition in space of two or more unit cells.

The structure of aluminium alloys that forms the signature of quasicrystals is identified with the icosahedral[1,2] phase with the point group symmetry m $\bar{3}\bar{5}$. The icosahedra* are the packing unit with Mn, Fe or Cr atom at the centre surrounded by twelve Al atoms located like the corners of an icosahedron. Although the icosahedral symmetry is an approximate representation of the site symmetry in these structures, the aperiodic lattice translations do not permit it to survive. Quasicrystals cannot and do not exhibit the point group symmetry. For example, the icosahedral phase under discussion shows five-fold axes of symmetry together with others only over a selected area when the specimen is rotated through angles of the icosahedral group. The observation must be read with the indications of elementary crystallography[3] that five-fold axes are inconsistent with translational orders.

The crystal structure analyses confirm that the symmetry of the fundamental space lattice of a quasicrystal belongs to none of the Bravais lattices (five in two dimensions and fourteen in three dimensions) that are used to describe the symmetry of crystals. The symmetries are instead intermediate to those for a crystal and for a liquid. The quasicrystals, discovered so far, are intermetallic alloys and show nearly insulating behaviour. They are also of great academic interest, as they enlarge the concept of crystal lattice.

SUMMARY

1. Crystal = Space lattice + Basis
   Space lattice: An infinite periodic array of points in space
   Basis: An atom or an identical group of atoms attached to every lattice point. It is identical for every lattice point in terms of composition, relative orientation and separation.

---

1. D. Shechtman, L. Blech, D. Gratias and J.W. Cahn, *Phys. Rev. Lett.* **53**, 1951 (1984).
2. D. Levine and Paul J. Steinhardt, *Phys. Rev. Lett.* **53**, 2477 (1984).

* Icosahedron is a polyhedron with twenty faces.

3. M.J. Burger, *Elementary Crystallography* (MIT Press, Cambridge, 1978).

2. Lattice points are connected to each other by the translation vector $\mathbf{t} = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}$, where n_1, n_2, n_3 are integers and **a**, **b**, **c** are primitive translation vectors (the axes of a primitive cell).
3. Primitive cell: A minimum volume unit cell having only one lattice point.
4.

Crystal types	*Bravais lattices*	*Point groups*	*Space groups*
Two-dimensional crystals	5	10	17
Three-dimensional crystals	14	32	230

5. Identical atoms may crystallize into the closest-packed structure either with HCP or with FCC symmetry, having the pattern of layers as:
   HCP: AB AB AB AB . . .
   FCC: ABC ABC ABC . . .
   For both the structures:
   (i) The number of nearest neighbours (coordination number) = 12
   (ii) The packing fraction = 0.743

PROBLEMS

1.1 Determine the lattice constant of NaCl crystal. The molecular weight of NaCl is 58.44 and the density is 2.167 g cm^{-3}.

1.2 It is given that the primitive basis vectors of a lattice are:

$$\mathbf{a} = 3\hat{\mathbf{x}}, \quad \mathbf{b} = 3\hat{\mathbf{y}} \quad \text{and} \quad \mathbf{c} = \frac{3}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$$

What is the Bravais lattice?

1.3 In a simple hexagonal lattice, $c/a = \sqrt{8/3}$. Determine the volume of its direct primitive cell in terms of a.

1.4 (a) Show that every edge of the Wigner–Seitz cell for the BCC lattice has the length $\sqrt{2}a/4$, where a is the lattice constant of the BCC structure.
(b) Show that the two diagonals on each face of the Wigner–Seitz cell for the FCC lattice are in the ratio $\sqrt{2}$:1.

1.5 Show that, for the closest packing of atoms, the densities of the FCC, BCC, SC and diamond monatonic lattices are approximately in the ratio 1.4 : 1.3 : 1.0 : 0.65.

1.6 The angles between the tetrahedral bonds of diamond are the same as the angles between the body diagonals of a cube. Determine the value of the angle using elementary vector analysis.

1.7 (a) What will be the indices of (100) and (001) planes of a conventional FCC lattice when referred to the primitive axes?
(b) Prove that in a crystal of cubic symmetry, a direction [h k l] is perpendicular to the plane (h k l) with the same indices. Assume that the primitive basis vectors of the lattice are given.

1.8 Show geometrically that the face centred tetragonal structure is equivalent to a body centred tetragonal structure.

1.9 (a) Show that the Wigner–Seitz cell for any two-dimensional Bravais lattice is either a hexagon or a rectangle.
(b) Prove that the hexagonal faces of the BCC Wigner–Seitz cell are all regular hexagons.

1.10 Sodium metal transforms from BCC to HCP structure at about 23 K. Calculate the lattice constant in the hexagonal phase. Take the value of the lattice constant as 4.23 Å in the cubic phase and assume that the density remains unchanged during the transition and the ratio *c/a* is close to the ideal value.

SUGGESTED FURTHER READING

Buerger, M.J., *Introduction to Crystal Geometry* (McGraw-Hill, 1971).

Fujiwara, T. and T. Ogawa (Eds.), "Quasicrystals", *Springer Ser. Solid State Sci.*, Vol. **93** (Springer, 1990).

Megaw, H.D., *Crystal Structures: A Working Approach* (Saunders, 1973).

Phillips, F.C., *An Introduction to Crystallography,* 4th ed. (Wiley, 1971).

Chapter 2

Atomic Cohesion and Crystal Binding

As discussed in Chapter 1, the symmetry and the structure of a crystal depend on the nature of the binding forces acting between the constituent atoms or particles. The analyses of crystal structure revealed the important fact that there is no essential difference between the forces which bind a chemical molecule and those which hold the atoms in a crystal together. This has proved to be of tremendous help to physicists in studying the nature of the binding forces in crystals by suitably extending the ideas of chemists. The present chapter is devoted to the description of various kinds of binding forces that result in different types of crystals. Other related properties will also be discussed.

2.1 COHESION OF ATOMS

Consider two atoms kept at a finite separation. The nature of the forces between atoms or the binding energy depends primarily on the distribution of positive charge over the inner ion core and of negative charge over the outer space within the atoms. Their potential energy is zero when the atoms are infinite distance apart. When they are brought closer to a finite separation, there will be a potential energy for the system whose sign depends on the relative order of magnitude of the repulsive and attractive forces. The attractive part of the potential energy is conventionally treated as negative since the atoms themselves do the work of attraction. The positive sign for the repulsive potential energy comes from the concept that in binding the atoms closer, the work is done against the repulsive force by an external agency. The total potential energy may be represented by a general potential of the form

$$U(r) = -\frac{A}{r^n} + \frac{B}{r^m} \tag{2.1}$$

where r is the interatomic distance, and A and B are constants. The net force is expressed as

$$\mathbf{F}(r) = -\frac{dU(r)}{d\mathbf{r}} = -\frac{nA}{\mathbf{r}^{n+1}} + \frac{mB}{\mathbf{r}^{m+1}} \tag{2.2}$$

The first term represents the attractive force and the second the repulsive. The variation of potential energy and force as a function of the interatomic separation is traced in Fig. 2.1. The net force at the separation r_0, when the atoms are in stable equilibrium forming a diatomic molecule, is zero [Fig. 2.1(b)]. Obviously, the potential energy at this position is minimum with the negative sign

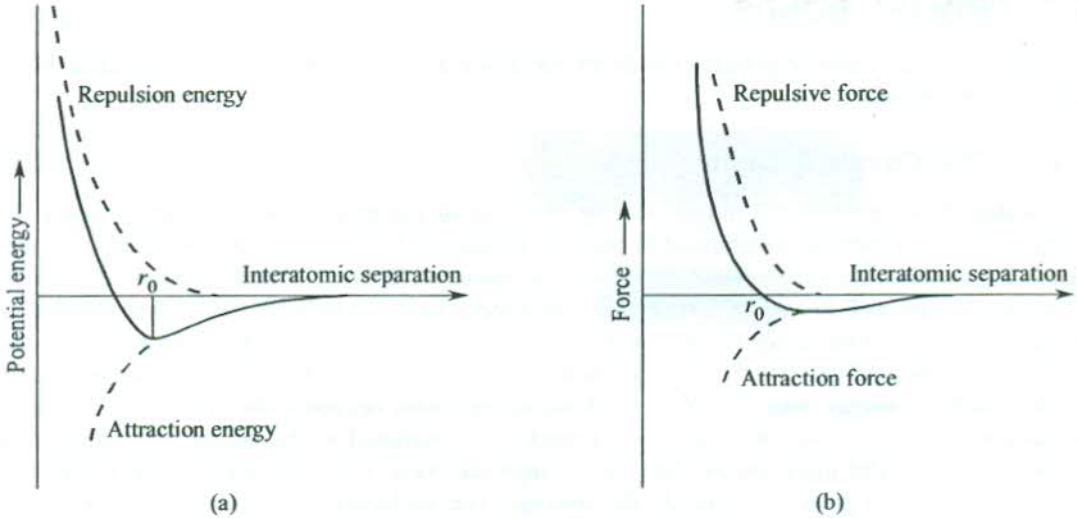


FIG. 2.1 (a) Potential energy of a system of two atoms as a function of their interatomic separation. (b) Force between two atoms as a function of their interatomic separation. r_0 denotes the separation at which the bond formation occurs.

[Fig. 2.1(a)]. It is also evident that as the atoms are brought closer, the repulsive force increases faster than the attractive force at short distances. When the two atoms stay in stable equilibrium at a certain separation they are said to have formed a chemical bond between them. The atoms spend a part of their energy in bond formation or we can say that a bond stores a part of the energy of the atoms. Therefore, the electron energy of atoms decreases when a large number of them collect together to form a solid. This suggests that when a solid is heated to break it finally into isolated atoms the energy stored in bonds must be carried away by the individual isolated atoms. How strongly the atoms are bound together in a solid is represented by its cohesive energy (or binding energy). A proper discussion on cohesive energy will be taken up later in this chapter after dealing with the types of binding in solids. When on solidification a pronounced lowering of electron energies takes place, the bonds formed are strong and called *primary bonds.* On the other hand if the lowering of electron energies is small, weak bonds are formed and referred to as *secondary bonds.*

The quantum mechanical interpretation of the probability function comes handy in understanding why some bonds are strong and directional while the others are weak and non-directional. It can be done qualitatively by acquiring information about the energies and the locations of the bonding electrons with respect to the positively charged ion cores. The probability that an electron, confined in the small volume dV, may lie in the orbital ψ is given by $|\psi|^2 dV$. For the large value of the probability amplitude $|\psi|^2$ in a certain direction, the bond formed is strong and concentrated along this direction and it behaves as directional. The small values of $|\psi|^2$ correspond to the weaker non-directional bonds. Furthermore, if $|\psi|^2$ is spherically symmetric non-directional bonds result. The spherical symmetry means that the chance of finding a bonding electron in all directions is equal and an atom can approach the other atom from any direction over the solid angle 4π for making the bond. Such a bonding gives rise to close packing exhibited by metals.

Now, we set out to outline the description of various forms of binding mechanisms in solids. Based on their comparative strength, the bonds formed in various binding processes may broadly be classified as (a) primary bonds and (b) secondary bonds.

2.2 PRIMARY BONDS

The three limiting cases of primary bonds are identified as (a) covalent bonds, (b) metallic bonds, and (c) ionic bonds.

2.2.1 The Covalent Bond

A covalent bond involves the mutual sharing of a pair of electrons between a pair of atoms. The spins of the two electrons are oriented in opposite directions. In solid state, the most stable covalent bonds are formed between non-metallic atoms like those of N, O, C, F and Cl. Some of the other elements that are well known to form covalent-bonded crystals include Si, Ge, As and Se. But, the nature of bonds in these solids is only partly covalent. Strong covalent bonds are formed when each atom has at least one-half filled orbital. In such a situation only will there be a substantial lowering in the electrons energy when each of the bonding electrons occupies the orbitals of two atoms simultaneously. The lowering of electrons energy is proportional to the degree of overlap of the bonding orbitals. The more the overlap, the stronger the bond. Either the electrostatic repulsion or the Pauli exclusion principle controls the overlap. The molecular hydrogen offers itself as the simplest candidate for the explanation of the covalent bonding. Hydrogen has a single electron which occupies the 1s orbital in the ground state. The orbital is half-filled and the Pauli principle allows it to accommodate one more electron with the opposite spin. Thus when two hydrogen atoms are brought close, their electron charge distributions overlap and the covalent bond is formed by having the two electrons with opposite spins in the 1s orbital belonging to both the atoms. This configuration, being lower in energy than the one in which the second electron went to the 2s orbital with the possibility of the same spin orientation, belongs to the hydrogen molecule in the ground state.

The effect of repulsive interaction as a compulsion of Pauli principle is best demonstrated in rare gas solids whose atoms have completely filled orbitals with little chance of overlap. This results in a large interatomic distance (3.76 Å in Ar) compared to the bond length in a covalent-bonded molecule such as Cl_2 (2 Å). There is a deficiency of one electron in the outermost shell of chlorine ($3p^5$) to saturate it. Because of this the outermost orbital of a chlorine atom has the tendency to overlap with that of the neighbouring chlorine atom in search of the deficit electron. Since this forms the basis of the covalent bond formation, the covalent bond is also known as a *saturable bond.*

The discussion on the covalent binding remains incomplete without a few remarks on elements in group IV of the periodic table (C, Si, Ge and grey Sn). These are extremely covalent elements and crystallize in the tetrahedrally coordinated diamond structure (see Fig. 1.26). Each atom participates in four covalent bonds with its four neighbours. These bonds are extremely directional and difficult to tilt. This provides the material with unusual strength and hardness. The wide use of diamond for making cutting tools is a well-acknowledged fact. Also, the strong directional nature of the covalent bonds rules out close packing and increases emptiness in covalent bonded crystals. For example, the fraction of volume of a diamond crystal actually occupied by atoms is a mere 0.34 which is 46 per cent of the value for the close-packed FCC and HCP structures. In view of this fact, the diamond lattice is called the *empty lattice.*

2.2.2 The Metallic Bond

In the modern theory of metals the valence electrons, being loosely bound to the respective parent atoms, are treated as common to the whole assembly of atoms comprising the metal. The valence electrons, generally one or two per atom, are considered free and, therefore, allowed to move freely

over the whole volume of the metal. The metal is pictured as an assembly of positive metal ions embedded in a sea of the free electron gas. The attraction between the metal ions and the electron gas gives rise to a strong cohesive force. The free electron gas serves as the glue. The concept of free electrons contributes significantly to the success of this model in explaining most of the properties of metals such as high electrical and thermal conductivities, high reflectivity and opacity. In reality the motions of valence electrons, also called *conduction electrons* with reference to metals, are affected, no matter how slightly, because of the presence of other particles including the fellow electrons. A complete theory does take care of this effect which is of crucial advantage in explaining some of the dramatic properties (e.g. superconductivity) of solids in general.

Since the metallic bonds involve loosely-bound electrons, there can be a relatively large number of probable orientations for a metal bond. In other words, the bonds are weak and non-directional. This characterizes metals as having the tendency to crystallize in relatively close-packed structures with a large number of nearest neighbours. This fact is in agreement with the observed structures of metals—HCP, FCC and BCC. In all these structures, the metallic bond is not saturated. For example, in the BCC structure of lithium, each bond has only 1/4th of an electron since the only outermost electron ($3s^1$) is involved in eight bonds with atoms at the corners of the cube. Still the modern picture assumes that the metallic bond is more closely related to the electron–pair or covalent bond than to an ionic-type bond, which is discussed next.

2.2.3 The Ionic Bond

The formation of an ionic bond is based on an electrostatic attraction between the positive and negative ions that are derived from the free atoms by the loss or gain of electrons. The ionic bond is responsible for the binding of salts that comprise the combinations of the elements located on the right- and left-hand sides of the periodic table. Alkali halides are the typical representatives of such salts. The electronic configurations of the alkali and halogen atoms are nearly closed-shell configurations. The bond formation is facilitated by the ease with which the alkali atom loses one electron and the halogen atom accepts it to acquire closed-shell configurations. In another way, we say that the origin of binding lies in the low ionization energy of the alkali atom and the high electron affinity of the halogen atom. Take the example of NaCl in which the sodium atom has one electron more ($1s^2\ 2s^2\ 2p^6\ 3s^1$) than the neon ($1s^2\ 2s^2\ 2p^6$) and the chlorine atom is one electron short ($1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^5$) of the argon configuration ($1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6$). This shows how Na and Cl atoms get bonded when brought close because of their easy ionization to Na^+ and Cl^- ions. Each ion has the tendency of being surrounded by as many ions of the opposite type as possible. The coordination number and the nearest neighbour distance are governed jointly by the geometrical factors and the repulsive interaction between the like ions. But the tendency of one type of ions to surround an ion of the opposite type binds a continuous network of ions to form the crystal instead of forming small discrete molecules.

Ionic bonds are neither saturable nor directional. Even then they are strong enough as is confirmed by the hardness, high melting point and low coefficient of expansion of ionic crystals.

2.2.4 Mixed Bonding

The discussion on primary bonds remains inconclusive without commenting on the purity of bonds in crystals. The mixture of bonding types has been discussed by Pauling* in great detail. In particular it is a matter of general observation that the bonding in most of the covalent-bonded crystals has

* L. Pauling, *The Nature of the Chemical Bond,* Cornell Univ. Press (1960).

a small component of ionic bonding. Ashcroft and Mermin* elucidate it for GaAs by picturizing the electron density of Ga and As ions in the crystal. Phillips** developed a semiempirical theory of fractional covalent or ionic character of bonds in a dielectric crystal. His calculations show that bonds in Ge and Si are purely covalent. But the bonding in most of the binary crystals turns out to be of mixed nature. LiF, NaCl and RbF are found to be almost purely ionic as the fraction of ionic character in them turns out be 0.92, 0.94 and 0.96, respectively.

The ionic character of a solid may be determined by analyzing its x-ray scattering data. This is an effective tool as the scattering power depends on the number of electrons possessed by the constituent ions. For example, in ionic-bonded KCl each of the ions, K^+ and Cl^-, has 18 electrons. Hence the scattering power for K^+ and Cl^- is found to be equal. The electron density distribution in the basal plane of NaCl, as derived by G. Schoknecht† is shown in Fig. 2.2.

FIG. 2.2 Electron density distribution in the basal plane of NaCl. The numbers of contours express the relative electron density. [From Kittle, C., *An Introduction to Solid State Physics,* 6th ed., p. 65, John Wiley (1986).]

2.3 SECONDARY BONDS

The limiting cases of secondary bonds are not easily separable. van der Waals and hydrogen bonds will be discussed in this category. The electric dipole–dipole interaction forms the basis of these bonds.

* N.W. Ashcroft and N.D. Mermin, *Solid State Physics,* Saunders College (1988).
** J.C. Phillips, *Bonds and Bands in Semiconductors,* Academic Press (1973).
† In C. Kittle, *Introduction to Solid State Physics,* 6th ed., p. 65, John Wiley (1986).

2.3.1 The van der Waals Bond

We may be curious to know why neutral molecules or noble gases should undergo liquefaction and crystallization. Consider noble gas atoms which have close-shell structure and are represented as spherical rigid charge distributions. Such electron distributions are reluctant to overlap when any two atoms are brought close to each other. But there must be some attractive forces between atoms bringing about the cohesion and finally the solidification. Of these forces, one is the van der Waals force we are interested in. These are weak forces arising out of the attractive interaction between fluctuating electric dipoles. By chance, for a fraction of the second, there could be more number of electrons on one side of the nucleus than on the other. This destroys the spherical symmetry of the electron charge distribution and momentarily displaces the centre of the negative charge (the electrons) from the centre of the positive charge (the nucleus). Thus an atom becomes a tiny electric dipole which is capable of inducing an electric dipole moment in the neighbouring atom. The two dipoles attract each other, though weakly, resulting in the van der Waals binding. As the electron charge distribution keeps fluctuating, the electric dipoles in question are called *fluctuating dipoles.*

Suppose we have two atoms 1 and 2 of a noble gas separated by distance **r**. As described above, when atom 1 acquires an instantaneous electric dipole moment $\mathbf{p}_1$, an instantaneous electric dipole moment $\mathbf{p}_2$ may be induced in atom 2 because of the polarization caused by the electric field **E** of the atomic dipole 1 at atom 2. It is now a trivial exercise in electrostatics (refer to Problem 2.2) to show that the interaction energy of two dipoles $\mathbf{p}_1$ and $\mathbf{p}_2$ $(= -\mathbf{p}_2 \cdot \mathbf{E})$ varies with separation as $1/r^6$. This interaction energy is essentially the van der Waals binding energy of the two atoms.

An alternative explanation of van der Waals interaction is the quantum mechanical basis in which the creation of the electric dipole is attributed to the zero point electronic motion. In fact, the van der Waals interaction more completely identified as the van der Waals–London interaction turns out to be a quantum effect. The quantum mechanical calculation confirms $1/r^6$ dependence of the interaction. The interaction is expressed as A/r^6 (for derivation see Appendix A), where A is a constant involving the Planck constant. This implies that the interaction vanishes when the Planck constant is taken as zero, proving that the van der Waals interaction has its proper explanation in quantum physics.

It must be noticed that the van der Waals bonds are neither saturable nor directional. Barring helium, all the noble gases crystallize in the close-packed FCC structure. These weakly-bound crystals are transparent insulators and characterized by low melting points.

2.3.2 The Hydrogen Bond

A number of covalently-bonded molecules behave as permanent electric dipoles. HF and H_2O are the two most talked about examples. The attraction between the positive end of one dipole and the negative end of the neighbouring dipole forms clusters or large aggregates of molecules on cooling when the crystallization may take place. The mechanism of bond formation involves the attraction of a hydrogen atom to two strongly electronegative atoms. These bonds are called *hydrogen bonds.* These bonds have the directional property and are stronger than the van der Waals bonds. This is endorsed by the observation that supermolecules like H_2F_2, H_3F_3, . . . are formed even in the gaseous phase.

In an effort to understand the mechanism of hydrogen bonding clearly, we take up ice crystals which exhibit some unusual behaviour. The outer electron configuration of the oxygen atom in H_2O is represented by four sp^3 hybrid orbitals. Figure 2.3 shows that two of these are involved in covalent bonds to the hydrogen atoms and the rest two are doubly occupied by oxygen electrons. Out of the

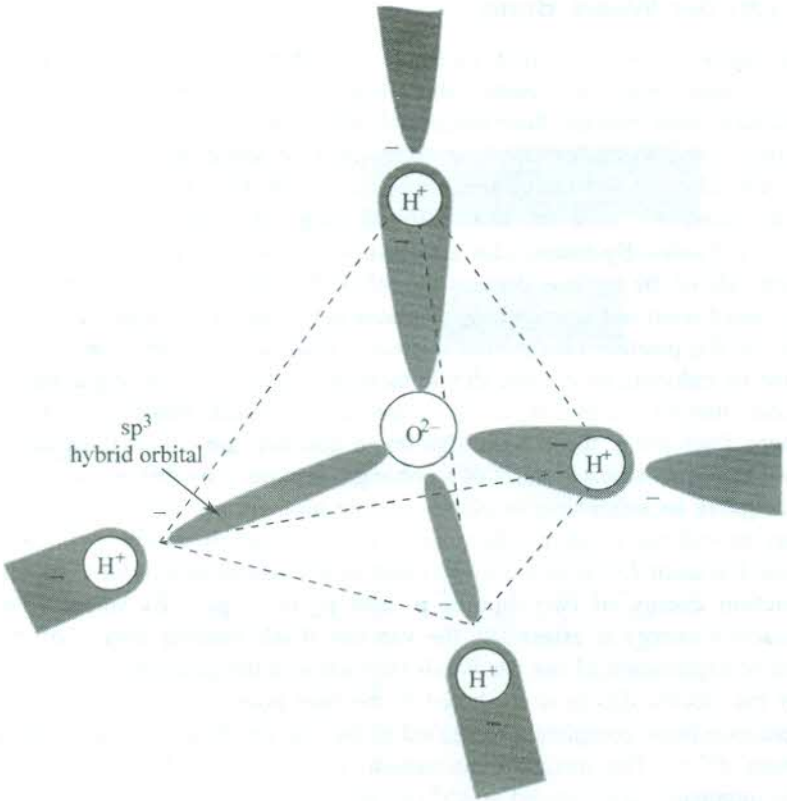


FIG. 2.3 The tetrahedral covalent bonding in ice involving the sp^3 hybrid orbitals of H_2O.

six outer oxygen electrons, two are engaged in bonding and the other four find a place in spare orbitals. This bonding pattern gives a tetrahedral shape to bonded molecules in ice. Two vertices of a tetrahedron are occupied by hydrogen atoms which are essentially the localized protons. The other two vertices exhibit relatively diffused negative charges.

The negligible small size of the proton is responsible for the peculiar structure of ice crystals as it practically sits on the oxygen ions, something impossible for any other positive ion to do. One of the many phases of ice crystals as shown in Pauling's book* is pictured in Fig. 2.4. Two oxygen atoms are bound by a proton that is localized close to one oxygen atom along the line joining it to one of its neighbouring oxygen atoms. There are two protons close to each oxygen atom, giving a large number of ways to attach a proton to either end of the bond. This is reflected in the irregular positions of protons which are very well accounted by the observed large residual entropy of ice crystals at low temperatures. Lastly, a word about the floating property of ice. The molecular clusters of water are smaller and less stable in liquid state than they are in ice because of which they are on the average more closely packed in this state. This increases the density of water and ice floats when immersed in it.

* L. Pauling, *The Nature of the Chemical Bond*, Cornell Univ. Press (1960).

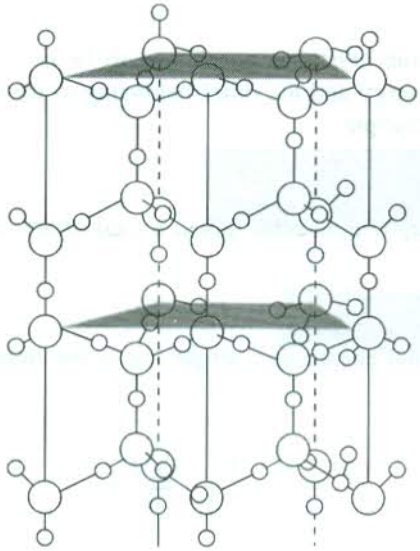


FIG. 2.4 The hydrogen bonding in ice. The large circles denote oxygen ions and the small circles protons.

2.4 THE COHESIVE ENERGY

As stated earlier, the cohesive energy of a crystal is the measure of how strongly its constituent atoms or particles are bound. It may be defined as the energy required to break the crystal into its isolated neutral constituent parts. By constituent parts we precisely mean atoms or molecules. For example, it sounds more reasonable to define the cohesive energy of solid nitrogen as the energy required to disassemble it into isolated nitrogen molecules instead of atoms. The two definitions are, however, interconvertible. In the early development of the solids classification, the theory based itself heavily on the nature of cohesion. The role of spatial electronic arrangement was almost sidelined. But with the increasing need to study the non-equilibrium properties for the development of devices, the concern for cohesive energy has ceased to be of any significant note.

Now, we present here a brief sketch of the classical theory to discuss the cohesive energy of crystals at 0 K. Calculations are made in the static lattice approximation which treats the atoms to be at rest at equilibrium sites. It neglects the zero point motion which relates to one of the basic tenets (the uncertainty principle) of quantum theory. The resulting error for most of the crystals turns out to be 1 per cent or less. But there may be a question mark against this approximation while treating the lighter noble gases as they have the origin of their cohesion in the zero point motion. This is not to suggest that the noble gases cannot be handled with a simplified theory, like the one under discussion. Surprisingly, the theory which is reasonably simple in this case enjoys a high degree of success.

The oversimplified theory discussed above is applied to ionic crystals with maximum ease as the dominating long-range attractive interaction between the oppositely charged ions renders other interactions of little concern. On the other hand, it is difficult to have a simple theory for calculating the cohesive energy of covalent and metallic crystals. The electronic configurations in crystals have far more distorted forms than what they are in isolated atoms or ions. This makes it imperative to calculate the energy by treating the valence electrons in the field of the periodic crystal potential. It leads to the problem of band structure calculation complicating the whole procedure. The level of the present book cannot accommodate the theory of covalent and metallic crystals. The description of simple methods of calculating the cohesive energy of ionic and noble gas crystals would suffice.

2.4.1 Ionic Crystals

The potential energy of an ionic crystal is considered to be composed of two components, one representing the electrostatic energy and the other belonging to the repulsive overlap which has its origin in the Pauli exclusion principle.

The electrostatic energy

Consider an ionic crystal having N molecules, given a total of $2N$ ions. The electrostatic energy of the crystal can be written as

$$U_e = NU_i \tag{2.3}$$

where U_i is the average potential energy of a single ion in the field of the other remaining ions. Also

$$U_i = \sum_{i=1}^{2N-1} U_{ij} \tag{2.4}$$

with

$$U_{ij} = \pm \frac{1}{4\pi \epsilon_0} \frac{q^2}{|\mathbf{r}_{ij}|}; \qquad \mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j \tag{2.5}$$

Here U_{ij} represents the electrostatic interaction energy in SI units between two ions bearing an equal charge q and their positions being given by the vectors $\mathbf{r}_i$ and $\mathbf{r}_j$. Taking the origin of the coordinate system at the position of one of the positive ions, we have

$$\begin{aligned} \mathbf{r}_{ij} &= (\hat{i}n_1 + \hat{j}n_2 + \hat{k}n_3)r \\ |\mathbf{r}_{ij}| &= (n_1^2 + n_2^2 + n_3^2)^{\frac{1}{2}} r \end{aligned} \tag{2.6}$$

where n_1, n_2, n_3 represent the number of units of the nearest neighbour distance r along the x, y, z axes of the crystal. Having a positive ion at the origin, we observe that $(n_1, n_2, n_3)\, r$ represents the location of

a negative ion, if $(n_1 + n_2 + n_3)$ is an odd integer

a positive ion, if $(n_1 + n_2 + n_3)$ is an even integer.

The Coulomb energy between the ion at the origin and any other ion located at $\mathbf{r}_j$ will be

$$U_{ij} = \frac{(-1)^{n_1 + n_2 + n_3}\, q^2}{4\pi \epsilon_0 (n_1^2 + n_2^2 + n_3^2)^{1/2}\, r} \tag{2.7}$$

or

$$U_i = \frac{\alpha q^2}{4\pi \epsilon_0 r} \tag{2.8}$$

where α is a constant known as the *Madelung constant*, expressed by the following relation:

$$\alpha = \sum_{n_1=0}^{2N-1} \sum_{n_2=0}^{2N-1} \sum_{n_3=0}^{2N-1} (-1)^{n_1+n_2+n_3} (n_1^2 + n_2^2 + n_3^2)^{-1/2} \qquad (2.9)$$
$$(n_1, n_2, n_3) \neq (0, 0, 0)$$

For a NaCl crystal, the first term is $-6/\sqrt{1}$ that is contributed by the six nearest neighbours of the opposite type located at $(\pm 1, 0, 0)$, $(0, \pm 1, 0)$ and $(0, 0, \pm 1)$ in units of the nearest neighbour distance. The second term $12/\sqrt{2}$ comes from the 12 nearest neighbours of the same type, located at $(\pm 1, \pm 1, 0)$, $(0, \pm 1, \pm 1)$ and $(\pm 1, 0, \pm 1)$ and at a distance of $\sqrt{2}r$ from the origin. The sum in (2.9) can be made to converge rapidly using some tricky mathematical method. The calculations yield the value of Madelung constant as -1.7476 for a NaCl crystal. For some other important ionic crystals such as CsCl, zinc blende (ZnS) and würtzite (ZnS), the values are -1.7627, -1.6381 and -1.641, respectively. A detailed account of the method of calculation has been provided by Born and Huang.*

The repulsive overlap energy

An approximate analytical form of the potential energy for a pair of ions as originally introduced by Born and Mayer** is given by

$$U_{ij}^{(r)} = \lambda_{ij} \exp\left(\frac{-|\mathbf{r}_{ij}|}{\rho}\right) \qquad (2.10)$$

where λ_{ij} and ρ are the empirical parameters which depend on the nature of the ions i and j but are independent of the distance between them.

The repulsive term is representative of the fact that the overlap between the electron configurations of neighbouring ions is resisted. The constants λ and ρ stand for the strength and the range of interaction, respectively, and can be determined with the knowledge of the experimental values of the lattice constant and compressibility. The range ρ is defined as the value of $|\mathbf{r}_{ij}|$ for which the interaction is reduced to $1/e$ of the interaction λ_{ij} when the two ions are in contact, treating them as ideal point charges.

Thus the total potential energy of the ions at $\mathbf{r}_i$ and $\mathbf{r}_j$ can be written as the sum of the attractive potential energy $U_{ij}^{(a)}$ and the repulsive potential energy $U_{ij}^{(r)}$ such that

$$U_{ij}^{(t)} = U_{ij}^{(a)} + U_{ij}^{(r)}$$
$$= -\frac{1}{4\pi\epsilon_0} \frac{q^2}{|\mathbf{r}_{ij}|} + \lambda_{ij} \exp\left(\frac{-|\mathbf{r}_{ij}|}{\rho}\right) \qquad (2.11)$$

The total potential energy of a crystal of NaCl structure will then be written as

$$U = N\left(-\frac{\alpha q^2}{4\pi\epsilon_0 r} + 6\lambda_{+-} \exp\left(\frac{-\sqrt{1}\, r}{\rho}\right) + 12\lambda_{++} \exp\left(\frac{-\sqrt{2}\, r}{\rho}\right) + \ldots\right) \qquad (2.12)$$

---

* M. Born and K. Huang, *Dynamical Theory of Crystal Lattice,* Oxford Univ. Press (1954).
** M. Born and J.E. Mayer, *Z. Physik,* **75**, 1 (1932).

The subscripts on λ in (2.12) indicate that its value is different with reference to the interaction between the like ions and that between the unlike ions.

If we consider the repulsion only between the nearest neighbours, the relation (2.12) assumes the general form

$$U = N\left(-\frac{\alpha q^2}{4\pi\epsilon_0 r} + z\lambda e^{-r/\rho}\right) \tag{2.13}$$

where z stands for the number of nearest neighbours (six in the NaCl structure). To ensure that (2.13) applies to a stable crystal, α is assigned the positive sign in the relation.

Since the cohesive energy is referred to as the minimum value of the potential energy, we can achieve the objective by expressing (2.13) in terms of the nearest neighbour distance for which the minimum of potential energy occurs. For U to be minimum,

$$\frac{dU}{dr} = 0$$

From (2.13), we get

$$\frac{dU}{dr} = N\left(\frac{\alpha q^2}{4\pi\epsilon_0 r^2} - \frac{z\lambda}{\rho} e^{-r/\rho}\right)_{r=r_0} = 0$$

where r_0 is the nearest neighbour distance at equilibrium. This condition gives

$$z\lambda e^{-r_0/\rho} = \frac{\alpha q^2 \rho}{4\pi\epsilon_0 r_0^2} \tag{2.14}$$

From (2.13) and (2.14), we get

$$U_0 = \frac{-N\alpha q^2}{4\pi\epsilon_0 r_0}\left(1 - \frac{\rho}{r_0}\right) \tag{2.15}$$

As an exercise, we apply (2.15) to a NaCl crystal. The potential energy of a single ion in the crystal is given by

$$U_{ion} = -\frac{\alpha q^2}{4\pi\epsilon_0 r_0}\left(1 - \frac{\rho}{r_0}\right) \tag{2.16}$$

For NaCl, $\alpha = 1.748$, $r_0 = 2.81$ Å and ρ is generally found to be $r_0/9$. (A small value of ρ shows the short range of repulsive interaction.)

Substituting these values in (2.16), we get

$$U_{ion} = -7.97 \text{ eV}$$

This is not the correct energy of a single ion as every ion is counted twice while taking the pair

interaction. Therefore, the true potential energy of an ion is $-7.97/2$ eV $= -3.99$ eV, which is essentially the contribution of a single ion to the potential energy. With reference to ionic crystals the quantity, 7.97 eV, is referred to as the lattice energy which is only the cohesive energy per ion pair. This is in excellent agreement with the measured value of 7.96 eV.

The cohesive energy per atom may as well be calculated easily. The ionization potential of Na is 5.14 eV and the electron affinity of Cl is – 3.61 eV. In the formation of Na^+–Cl^- ion pair, the energy spent in the electron transfer is equal to (5.14 – 3.61) eV, i.e. 1.53 eV, meaning thereby that each atom contributes 0.77 eV to the cohesive energy. Thus the cohesive energy/atom of NaCl

$$= (-3.99 + 0.77) \text{ eV}$$

$$= -3.22 \text{ eV}$$

The success of the underlined theory is further emphasized by the closeness of the measured value (3.28 eV/atom).

2.4.2 Noble Gas Crystals

The discussion under Section 2.3.1 needs extension for obtaining a complete picture of interactions in noble gas crystals. The repulsive potential energy is conventionally expressed in the form of an empirical power law as B/r^{12}. The total potential energy of a pair of atoms at the separation r is generally represented by

$$U(r) = -\frac{A}{r^6} + \frac{B}{r^{12}} \tag{2.17a}$$

and this is known as the Lennard–Jones potential. The constants A and B are empirical parameters.

On the requirement that the repulsive force increases faster than the attractive force at short distances, the exponent in the repulsive term has to be more than 6. In addition, if the analytic simplicity is taken into account the choice falls on 12. Further, with this choice the theoretical estimates of several physical properties of nobel gases, exclusive of He^3 and He^4, are found to be in excellent accord with the experiment. A satisfactory explanation for the disagreement in the case of helium is not possible at this stage to avoid digression. However, it should suffice to remark that helium is identified as a unique quantum matter in the solid state theory which has developed provisions for dealing with it especially. Finally, it must be emphasized that the van der Waals interaction is present in all the three states of every material. Being weak in nature it is not the main cause of cohesion in many crystals where other strong interactions are present and the crystal bindings are named after them.

Equation [2.17(a)] is usually expressed as

$$U(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2.17b}$$

where

$$\sigma = \left(\frac{B}{A}\right)^{1/6} \quad \text{and} \quad \varepsilon = \frac{A^2}{4B}$$

The Lennard–Jones potential in the form of relation (2.17b) is shown in Fig. 2.5. Suitable values of the parameters and σ have been obtained by fitting the theoretically calculated values of certain

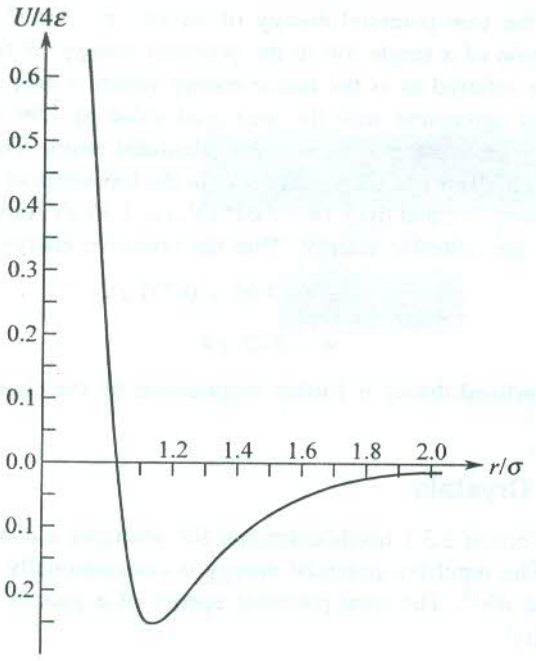


FIG. 2.5 The Lennard–Jones potential as given by equation (2.17b).

physical properties involving these parameters with those experimentally observed in the gaseous phase. In principle, these values cannot be used for solids where the interaction is not a sum of pair potentials on account of high densities. The values of parameters serve as a measure of the strength of attraction and the radius of the repulsive core.

The hydrogen crystals are among the weakest bonded crystals with a cohesive energy of 0.01 eV per molecule which is one-tenth of that of methane's (0.1 eV per molecule). The cohesive energies of a number of crystals of different types are given in Table 2.1 for the purpose of a comparative study. For a thorough study the reader may consult the book by Ashcroft and Mermin* which gives a more comprehensive view of the subject and treasures large data on the related physical quantities of molecular and ionic crystals in particular.

It must now be abundantly clear that the calculations we have discussed in this section are closely linked with lattice constants (or lattice parameters) of crystals. Let us briefly comment on the involvement of lattice constants in a few other important physical properties. The pressure required to maintain a certain volume can be determined by calculating the rate of variation of cohesive energy with lattice constants. In effect, we succeed in reproducing the experimental value of the equilibrium lattice constant maintainable at zero pressure. Similarly, it is possible to study the change in volume caused by a change in pressure and calculate the compressibility. The significance of the lattice constant and compressibility must be emphasized in view of their utility in estimating the empirical parameters λ and ρ of (2.13). As the lattice constants have a close relationship with the bond length and the atomic radii, we propose to give a short account of the same in the following section.

* N.W. Ashcroft and N.D. Mermin, *Solid State Physics*, Saunders College (1988).

Table 2.1 Cohesive energy of crystals

Crystal	*Binding type*	*Cohesive energy*	*Melting point* (K)
LiF	Ionic	10.31 eV/ion	1143
NaCl	Ionic	7.96 eV/ion pair 3.28 eV/atom	1074
RbI	Ionic	6.17 eV/ion pair	915
Diamond	Covalent	7.37 eV/atom	> 3773
Si	Covalent	4.63 eV/atom	1687
SiC	Covalent	12.04 eV/molecule	2873 (subl.)
Na	Metallic	1.11 eV/atom	371
Cu	Metallic	3.49 eV/atom	1358
Au	Metallic	3.81 eV/atom	1338
Ne	van der Waals	0.02 eV/atom	24.6
H_2	van der Waals	0.01 eV/molecule	14
CH_4	van der Waals	0.10 eV/molecule	90
Ice (H_2O)	Hydrogen	0.51 eV/molecule	273
HF	Hydrogen	0.30 eV/molecule	180.7

2.5 ATOMIC RADII vs LATTICE CONSTANTS

The atomic and ionic radii have been compiled by Pearson.* An examination of this data shows that on ion formation the size of a positive ion is smaller and that of a negative ion larger than the size of the neutral atom. The best examples illustrating this fact are the radii of C (0.72 Å), C^{4+} (0.15 Å) and C^{4-} (2.60 Å). This can be understood by taking simple examples of Na and Cl. The Na atom, whose radius is 1.86 Å, loses one electron in becoming the Na^+ ion of radius of 0.96 Å. On losing this outermost electron ($3s^1$) which is loosely bound, the atom assumes a more tightly bound configuration of Na^+ ion with the reduced size. This is one of the best examples to state, as the Na^+ ion turns out to be smaller than even the isoelectronic Ne atom of the preceding period. Similarly, the bigger size of the Cl^- ion makes sense on the ground that the Cl atom is readily willing to turn into a Cl^- ion whose electronic configuration is identical with the rigid spherical charge distribution of the nearest (the next in the same period) noble gas atom (Ar).

The extent of charge clouds of a constituent ion in any crystal is generally referred to as the *crystal ionic radius*. It needs to be distinguished from the free ion radius. According to the quantum mechanical interpretation, the free ion radius may be defined as the radius at which the probability amplitude $|\psi|^2$ for the outermost electrons to be in the orbital represented by the wavefunction ψ is maximum. The wavefunction ψ is then a solution of the Schrödinger wave equation whose hamiltonian includes the ordinary Coulomb potential. The boundary condition requires that the magnitude and the derivative of the wavefunction ψ vanish at infinitely large distances. In contrast, for an ion in the crystal the magnitude and derivative of the wavefunction must be zero at the

* W.B. Pearson, *Crystal Chemistry and Physics of Metals and Alloys,* Wiley-Interscience (1972).

boundary of the ion on the demand of the boundary condition. The potential in this case is not Coulomb but of the type shown in Fig. 2.1(a).

The above discussion indicates that the ionic radii may be calculated by solving one-electron Schrödinger wave equation. This is usually done by the Hartree–Fock self-consistent field method, treated in great detail by Slater.* Some semi-empirical methods of calculating the ionic radii in crystals are also in practice. For example, the distance between two atoms shown as r_0 (the nearest neighbour distance in a crystal) in Fig. 2.1 is approximately equal to the sum of the radii of the two neighbouring atoms. This property is known as the additive rule. The diffraction method is the standard way at present to measure lattice constants. From the diffraction pattern it is possible to determine the radius ratio of any two types of atoms from which the radii of other atoms comprising the crystal can be determined. Goldschmidt** was the first to initiate calculations based on the experimental data. Later Pauling† used his own theory to calculate the ionic radii in crystals. His results are very well accounted in literature because of their amazing success in explaining most of the related physical properties of solids.

Using the additive rule of atomic radii it is possible to predict the bond lengths or the interatomic separations in a crystal. This is successfully done for crystalline phases even before they are crystallized. But we must be aware that the charge distribution is not rigid and spherically symmetric in every atom. This may introduce an appreciable error in our estimate. Therefore, it is a *must* to know accurately the average of charge clouds of the constituent atoms in a crystal for making any estimates.

At first instance the knowledge of atomic radii can be used to determine the coordination of an atom and suggest the probable arrangement of atoms in the crystal. Though this point has more relevance to Chapter 1, it is worth digressing to emphasize the role of atomic radii being discussed here. Beiser†† has excellently explained the threefold coordination for a binary ionic crystal. It is shown that the coordination is stable (bonds formed) only when the positive ion A^+ touches all the surrounding negative ions B^-. The mutual contact among the three B^- ions is not the necessary requirement for the coordination to crystallize. A simple geometrical calculation shows that the minimum ionic radius ratio r_A/r_B for this coordination to materialize is 0.155. We calculate below this ratio for the fourfold coordination in a similar ionic crystal.

In the fourfold coordination, the A^+ ion positioned at the centre of a cube is surrounded by four B^- ions located at the alternate corners of the cube (Fig. 2.6) whose edge is a. For stable coordination, all the four B^- ions should touch the A^+ ion. Two B^- ions on every face are also in contact with each other, the distance between their centres being equal to the length of the face diagonal ($\sqrt{2}a$). Let θ be the angle between the line joining a B^- ion to an A^+ ion and the line joining the two B^- ions.

From Fig. 2.6, we have

$$\cos\theta = \frac{\sqrt{2}a/2}{\sqrt{3}a/2} \qquad \text{(length of the body diagonal} = \sqrt{3}a\text{)}$$

$$= \sqrt{\frac{2}{3}}$$

* J.C. Slater, *Quantum Theory of Atomic Structure,* Vols. I and II, McGraw-Hill Book Company Inc. (1960).
** V.M. Goldschmidt, *Skerifter Norske Videnskaps Akad,* No. **2** (1926); No. **8** (1927).
† L. Pauling, *The Nature of Chemical Bond,* Cornell Univ. Press (1960).
†† A. Beiser, *Perspective of Modern Physics,* pp. 426–427, McGraw-Hill, Kogakush Ltd. (1969).

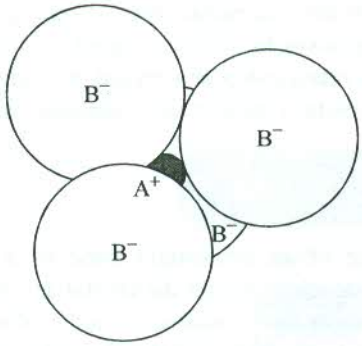


FIG. 2.6 The fourfold coordination: The smaller ion A^+ at the centre of a cube is surrounded by four B^- ions located at the alternate corners of the cube. The minimum value of r_A/r_B for the coordination is 0.225.

Also

$$\cos\theta = \frac{r_B}{r_A + r_B} = \sqrt{\frac{2}{3}} = 0.816$$

This gives $r_A/r_B = 0.225$, which is the minimum value for the fourfold coordination to be stable. These calculations show that for the formation of a stable threefold coordination, the ion radius ratio must lie between 0.155 and 0.225.

Finally, we demonstrate how successfully the bond lengths and the lattice constants are reproduced with the knowledge of the self-consistent atomic/ionic sizes. The degree of success achieved can be perceived by the data in Table 2.2.

Table 2.2 Data on the nearest neighbour separation and lattice constant, predicted from the radii of the constituent ions/atoms, for some crystals

Crystal	*Radius of the constituent ions/atoms* (Å)	*Predicted nearest neighbour separation, d* (Å)	*Relation between d and a (lattice constant)*	*Lattice constant a* (Å)	
				Predicted	*Observed*
NaF	Na^+ (0.97) F^- (1.63)	2.33	$a = 2d$	4.66	4.62
NaCl	Na^+ (0.97) Cl^- (1.81)	2.78	$a = 2d$	5.56	5.63
Diamond	C (0.72)	1.54	$a = \frac{4d}{\sqrt{3}}$	3.56	3.56
Ne	Ne (1.58)	3.16	$a = \sqrt{2}d$	4.47	4.46

2.6 ELASTIC CONSTANTS OF CRYSTALS

The crystal binding discussed thus far in this chapter strongly controls the elastic behaviour of solids. The nature of binding forces in a solid is often reflected in its elastic response. The study

of elastic properties is essential for the interpretation of several properties of solids. For example, certain elastic constants relate themselves to thermal properties, like the Debye temperature. In this section we discuss the stress–strain relationship in a crystal, treating it as a homogeneous continuous elastic medium. Effectively, the crystal's picture as a periodic array of atoms is replaced with a homogeneous elastic continuum.

2.6.1 Elastic Stress

Consider the uniform deformation of an elementary cube of a crystal. Under the action of a deforming force, an internal force develops within the crystal as a reaction to the applied force. The internal force acting on the unit area of the crystal is defined as the stress. For the present treatment it is assumed that the applied force is not large and the Hooke's law remains valid (stress $\propto$ strain). The stress acting on the six faces of the cube is expressed by nine components: σ_{xx}, σ_{xy}, σ_{xz}, σ_{yx}, σ_{yy}, σ_{yz}, σ_{zx}, σ_{zy}, σ_{zz}. The first subscript denotes the direction of the applied force and the second subscript gives the direction of the normal to the face on which the force is applied. The stress is a tensor of second rank and denoted by a (3×3) matrix

$$[\sigma_{\alpha\beta}] = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix} \tag{2.18}$$

with $\alpha, \beta = x, y, z$.

The components σ_{xx}, σ_{yy}, σ_{zz} denote the normal stress components acting on the yz, zx, and xy faces, respectively. The remaining six components represent the tangential stress (two components on each of the three pairs of the faces). If the cube is in the state of static equilibrium and it does not rotate under the influence of tangential stress components, $\sigma_{\alpha\beta}$ and $\sigma_{\beta\alpha}$ would produce equal and opposite rotations. Hence $\sigma_{\alpha\beta} = \sigma_{\beta\alpha}$ and the nine stress components reduce to six independent components. Figure 2.7 shows a normal component σ_{yy} and a tangential component σ_{yz} acting on the respective faces of a cube. The stress components have the dimensions of force per unit area or energy per unit volume.

2.6.2 Elastic Strain

Let three orthogonal unit vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ be safely embedded in an unstrained crystal. Suppose on straining the crystal by a small deforming force these vectors are transformed to non-orthogonal vectors $\mathbf{l}$, $\mathbf{m}$, $\mathbf{n}$, respectively, with individual magnitudes differing from unity. We treat the two sets of vectors as the new and old coordinate axes. The new coordinate axes may be expressed as

$$\begin{aligned} \mathbf{l} &= (1 + \varepsilon_{xx})\ \hat{\mathbf{i}} + \varepsilon_{xy}\ \hat{\mathbf{j}} + \varepsilon_{xz}\ \hat{\mathbf{k}} \\ \mathbf{m} &= \varepsilon_{yx}\ \hat{\mathbf{i}} + (1 + \varepsilon_{yy})\ \hat{\mathbf{j}} + \varepsilon_{yz}\ \hat{\mathbf{k}} \\ \mathbf{n} &= \varepsilon_{zx}\ \hat{\mathbf{i}} + \varepsilon_{zy}\ \hat{\mathbf{j}} + (1 + \varepsilon_{zz})\ \hat{\mathbf{k}} \end{aligned} \tag{2.19}$$

where $\varepsilon_{\alpha\beta}$ define the deformation. These coefficients are dimensionless and very small ($<< 1$) for a small strain.

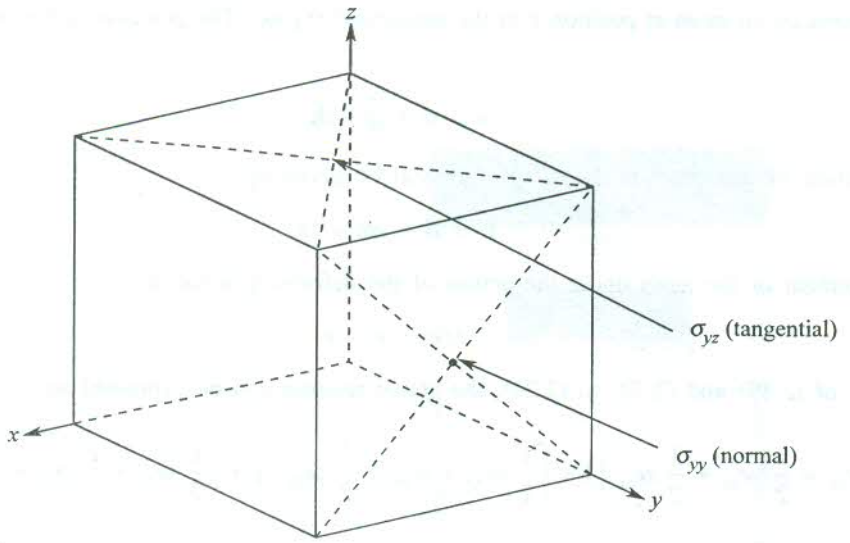


FIG. 2.7 The normal stress component, σ_{yy}, produced by a force applied in the $+y$ direction on the xz face (with normal along the $+y$ direction) of a cube. The tangential stress component, σ_{yz}, produced by a force applied in the $+y$ direction on the xy face (with normal along the $+z$ direction) of a cube.

By taking the dot products $\mathbf{l} \cdot \mathbf{l}$, $\mathbf{m} \cdot \mathbf{m}$, $\mathbf{n} \cdot \mathbf{n}$, we can easily show that the magnitude of each of the three new vectors is different from unity. Also, the dot products $\mathbf{l} \cdot \mathbf{m}$, $\mathbf{m} \cdot \mathbf{n}$, $\mathbf{n} \cdot \mathbf{l}$ do not vanish indicating that $\mathbf{l}$, $\mathbf{m}$ and $\mathbf{n}$ are not orthogonal vectors.

The strain components $e_{\alpha\beta}$ are defined in terms of $\varepsilon_{\alpha\beta}$ as

$$e_{xx} = \varepsilon_{xx}, \quad e_{yy} = \varepsilon_{yy}, \quad e_{zz} = \varepsilon_{zz} \tag{2.20a}$$

and

$$\begin{aligned} e_{xy} &= \mathbf{l} \cdot \mathbf{m} \simeq \varepsilon_{yx} + \varepsilon_{xy} \\ e_{yz} &= \mathbf{m} \cdot \mathbf{n} \simeq \varepsilon_{zy} + \varepsilon_{yz} \\ e_{zx} &= \mathbf{n} \cdot \mathbf{l} \simeq \varepsilon_{xz} + \varepsilon_{zx} \end{aligned} \tag{2.20b}$$

In relations (2.20b) the sign $\simeq$ may be replaced by the sign $=$, if the terms of order ε^2 are neglected ($\varepsilon_{\alpha\beta} << 1$). It is to be noticed that the strain components in (2.20b) are defined in terms of the changes in angle between the axes. Therefore, for a rigid rotation in which angles do not change, $e_{xy} = e_{yz} = e_{zx} = 0$ and $\varepsilon_{xy} = -\varepsilon_{yx}$; $\varepsilon_{yz} = -\varepsilon_{zy}$; $\varepsilon_{zx} = -\varepsilon_{xz}$.

If we do not consider pure rotations since they are not deformation, we may always take

$$\varepsilon_{xy} = \varepsilon_{yx}, \qquad \varepsilon_{yz} = \varepsilon_{zy} \qquad \text{and} \qquad \varepsilon_{zx} = \varepsilon_{xz} \tag{2.21}$$

Using (2.21) in (2.20b), we get

$$\begin{aligned} \varepsilon_{xy} = \varepsilon_{yx} &= \frac{1}{2} e_{xy} \\ \varepsilon_{yz} = \varepsilon_{zy} &= \frac{1}{2} e_{yz} \\ \varepsilon_{zx} = \varepsilon_{xz} &= \frac{1}{2} e_{zx} \end{aligned} \tag{2.22}$$

Now, consider an atom at position **r** in the unstrained crystal. The position vector is expressed as

$$\mathbf{r} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}} \tag{2.23}$$

Let the position of this atom in the strained crystal be given by

$$\mathbf{r}' = x\mathbf{l} + y\mathbf{m} + z\mathbf{n} \tag{2.24}$$

The displacement of the atom under the action of the deforming force is

$$\Delta(\mathbf{r}) = \mathbf{r}' - \mathbf{r} \tag{2.25}$$

Making use of (2.19) and (2.22) to (2.24), the above relation can be expressed as

$$\Delta(\mathbf{r}) = \left(xe_{xx} + \frac{1}{2} ye_{yx} + \frac{1}{2} ze_{zx} \right)\hat{\mathbf{i}} + \left(\frac{1}{2} xe_{xy} + ye_{yy} + \frac{1}{2} ze_{zy} \right)\hat{\mathbf{j}} + \left(\frac{1}{2} xe_{xz} + \frac{1}{2} ye_{yz} + ze_{zz} \right)\hat{\mathbf{k}} \tag{2.26}$$

On rewriting it as

$$\Delta(\mathbf{r}) = u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}} \tag{2.27}$$

we get

$$\begin{aligned} u_1 &= xe_{xx} + \frac{1}{2} ye_{yx} + \frac{1}{2} ze_{zx} \\ u_2 &= \frac{1}{2} xe_{xy} + ye_{yy} + \frac{1}{2} ze_{zy} \\ u_3 &= \frac{1}{2} xe_{xz} + \frac{1}{2} ye_{yz} + ze_{zz} \end{aligned} \tag{2.28}$$

where u_1, u_2, u_3 are the displacement components along the coordinate axes of the unstrained crystal.

All the six dimensionless strain components can now be defined as follows by taking the partial derivatives of u_1, u_2 and u_3.

$$\begin{aligned} e_{xx} &= \frac{\partial u_1}{\partial x}; \quad e_{xy} = \frac{\partial u_2}{\partial x} + \frac{\partial u_1}{\partial y} \\ e_{yy} &= \frac{\partial u_2}{\partial y}; \quad e_{yz} = \frac{\partial u_3}{\partial y} + \frac{\partial u_2}{\partial z} \\ e_{zz} &= \frac{\partial u_3}{\partial z}; \quad e_{zx} = \frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial x} \end{aligned} \tag{2.29}$$

Relations (2.29) give alternative definitions of strain components. Relations (2.19) and (2.20a) indicate that e_{xx}, e_{yy} and e_{zz} represent linear strains, i.e. changes in length per unit length along the three axes.

Each of the other type of three components interprets a combination of two simple shears. Take for example,

$$e_{yz} = \frac{\partial u_3}{\partial y} + \frac{\partial u_2}{\partial z}$$

It describes two shears: One in which the planes normal to y-axis slide in the z-direction and the other in which the planes normal to z-axis slide along the y-direction. Like stress, strain is also a tensor of second rank. In general, it is described by nine components with the matrix representation as

$$[e_{\alpha\beta}] = \begin{bmatrix} e_{xx} & e_{xy} & e_{xz} \\ e_{yx} & e_{yy} & e_{yz} \\ e_{zx} & e_{zy} & e_{zz} \end{bmatrix} \tag{2.30}$$

with $\alpha, \beta = x, y, z$.

2.6.3 Dilation

The fractional increase in volume created by deformation is called *dilation.* It is useful in determining some elastic constants such as the bulk modulus.

Volume of a unit cube after the deformation is

$$V' = \mathbf{l} \cdot \mathbf{m} \times \mathbf{n} \tag{2.31}$$

Substituting for **l**, **m** and **n** from (2.19) into (2.31) and neglecting the product of two strain components ($e_{\alpha\beta} << 1$), we get

$$V' = 1 + e_{xx} + e_{yy} + e_{zz} \tag{2.32}$$

Therefore, the dilation δ can be expressed as

$$\delta = \frac{V' - V}{V} = e_{xx} + e_{yy} + e_{zz} \tag{2.33}$$

using $V = 1$ for a unit cube.

2.7 ELASTIC COMPLIANCE AND STIFFNESS CONSTANTS

According to Hooke's law, the strain is directly proportional to the stress for sufficiently small deformations. Therefore, for appreciable small elastic deformations of a crystal, the stress tensor components and the strain tensor components are linearly related as

$$\begin{bmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{yz} \\ \sigma_{zx} \\ \sigma_{xy} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} \\ C_{21} & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} \\ C_{31} & C_{32} & C_{33} & C_{34} & C_{35} & C_{36} \\ C_{41} & C_{42} & C_{43} & C_{44} & C_{45} & C_{46} \\ C_{51} & C_{52} & C_{53} & C_{54} & C_{55} & C_{56} \\ C_{61} & C_{62} & C_{63} & C_{64} & C_{65} & C_{66} \end{bmatrix} \begin{bmatrix} e_{xx} \\ e_{yy} \\ e_{zz} \\ e_{yz} \\ e_{zx} \\ e_{xy} \end{bmatrix} \tag{2.34}$$

Conversely, the strain components can be expressed as the linear functions of the stress components.

$$\begin{bmatrix} e_{xx} \\ e_{yy} \\ e_{zz} \\ e_{yz} \\ e_{zx} \\ e_{xy} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} & S_{15} & S_{16} \\ S_{21} & S_{22} & S_{23} & S_{24} & S_{25} & S_{26} \\ S_{31} & S_{32} & S_{33} & S_{34} & S_{35} & S_{36} \\ S_{41} & S_{42} & S_{43} & S_{44} & S_{45} & S_{46} \\ S_{51} & S_{52} & S_{53} & S_{54} & S_{55} & S_{56} \\ S_{61} & S_{62} & S_{63} & S_{64} & S_{65} & S_{66} \end{bmatrix} \begin{bmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{zz} \\ \sigma_{yz} \\ \sigma_{zx} \\ \sigma_{xy} \end{bmatrix} \tag{2.35}$$

The coefficients C_{11}, C_{12}, etc. are called *elastic stiffness constants* and represent moduli of elasticity with dimensions of $\frac{\text{force}}{\text{area}}$ or $\frac{\text{energy}}{\text{volume}}$. The other coefficients S_{11}, S_{12}, etc. are called *elastic compliance constants* and have dimensions of $\frac{\text{area}}{\text{force}}$ or $\frac{\text{volume}}{\text{energy}}$.

2.7.1 Elastic Energy Density

By analogy with the expression for the energy of a stretched spring (to be discussed in Chapter 4), the elastic energy density ϕ is a quadratic function of strains in the approximation of Hooke's law

$$\phi = \frac{1}{2} \sum_{\mu=1}^{6} \sum_{\nu=1}^{6} \bar{C}_{\mu\nu} e_\mu e_\nu \tag{2.36}$$

where the indices 1 to 6 should read as

$$1 \equiv xx; \quad 2 \equiv yy; \quad 3 \equiv zz; \quad 4 \equiv yz; \quad 5 \equiv zx; \quad 6 \equiv xy \tag{2.37}$$

The coefficients $\bar{C}$s are found to be related to Cs of (2.34), as we will see below. We will exploit the definition (2.36) to show that the 36 coefficients in (2.34) or (2.35) can be reduced in number. The very definition of potential energy allows us to obtain the stress components from the derivative of ϕ with respect to the associated strain component. For example, when the stress σ_{xx} acts on one face of a unit cube and the opposite face is held at rest, we have

$$\sigma_{xx} = \frac{\partial \phi}{\partial e_{xx}} = \frac{\partial \phi}{\partial e_1} = \bar{C}_{11} e_1 + \frac{1}{2} \sum_{\beta=2}^{6} (\bar{C}_{1\beta} + \bar{C}_{\beta 1}) e_\beta \tag{2.38}$$

An inspection of relation (2.38) reveals that only the combination $\frac{1}{2}(\bar{C}_{\alpha\beta} + \bar{C}_{\beta\alpha})$ enters the stress–strain relations, implying that the elastic stiffness constants are symmetrical. Thus we have

$$C_{\alpha\beta} = \frac{1}{2}(\bar{C}_{\alpha\beta} + \bar{C}_{\beta\alpha}) = C_{\beta\alpha} \tag{2.39}$$

The above symmetrical property reduces the number of constants from 36 to 21.

2.7.2 Application to Cubic Crystals

In accordance with the Neumann's principle, the number of independent elastic stiffness constants decreases as the symmetry of a crystal increases. This number is 21, 13, 5 and 3, respectively, for triclinic, monoclinic, hexagonal and cubic systems. The cubic crystals, being the most symmetric, have the least number of independent elastic stiffness constants. We now indulge in the exercise of deriving this result.

We pronounce that the relation for the elastic energy density of a cubic crystal has the form

$$\phi = \frac{1}{2}C_{11}(e_{xx}^2 + e_{yy}^2 + e_{zz}^2) + \frac{1}{2}C_{44}(e_{yz}^2 + e_{zx}^2 + e_{xy}^2) + C_{12}(e_{yy}e_{zz} + e_{zz}e_{xx} + e_{xx}e_{yy}) \tag{2.40}$$

The above relation does not have the quadratic terms

$$(e_{xx}e_{xy} + \ldots); \qquad (e_{yz}e_{zx} + \ldots); \qquad (e_{xx}e_{yz} + \ldots); \tag{2.41}$$

The correctness of (2.40) can be confirmed by showing that ϕ is invariant under all symmetry operations permitted in a cubic crystal. The minimum symmetry requirement for cubic symmetry is the presence of four threefold rotation axes along the directions of the four body diagonals of the cube (see Fig. 2.8). Counterclockwise rotations by $2\pi/3$ about the [111] direction and three other equivalent directions interchange the x, y, z axes according to the following four schemes:

$$\begin{aligned} &x \to y \to z \to x; \qquad -x \to z \to -y \to -x \\ &x \to z \to -y \to x; \qquad -x \to y \to z \to -x \end{aligned} \tag{2.42}$$

It is straightforward to check that the relation for ϕ remains unchanged when x, y, z are interchanged in (2.40) according to any one of the four schemes (2.42). But every term appearing in (2.41) is odd in one or more indices. One of the schemes in (2.42) is surely such that its application to (2.40) would change the sign of the term in (2.41) ($e_{xy} = -e_{x(-y)}$, for example). This only confirms that the terms included in (2.41) have rightly been excluded from the relation (2.40) for ϕ.

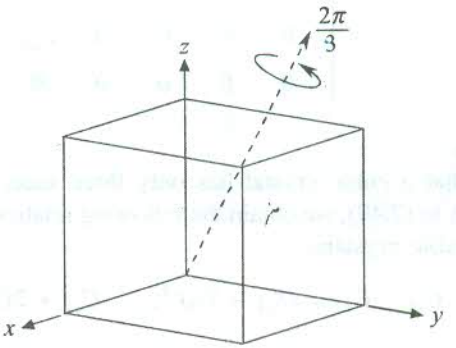


FIG. 2.8 Rotation of $2\pi/3$ about the [111] direction (the direction of a body diagonal) in a cubic crystal leaves the crystal unchanged in space by the way of interchange of the orthogonal axes: $x \to y \to z \to x$. Thus the [111] direction represents one of the four threefold rotation axes. The rest three are in other three directions that are equivalent to the [111] direction (i.e. the directions of the other three body diagonals).

We may now easily derive the stress components from (2.40). Thus

$$\sigma_{xx} = \frac{\partial \phi}{\partial e_{xx}} = C_{11}e_{xx} + C_{12}(e_{zz} + e_{yy}) \tag{2.43}$$

The corresponding relation given by (2.34) is

$$\sigma_{xx} = C_{11}e_{xx} + C_{12}e_{yy} + C_{13}e_{zz} + C_{14}e_{yz} + C_{15}e_{zx} + C_{16}e_{xy} \tag{2.44}$$

On comparing (2.43) with (2.44), we get

$$C_{12} = C_{13}; \qquad C_{14} = C_{15} = C_{16} = 0 \tag{2.45}$$

Further, from (2.40)

$$\sigma_{xy} = \frac{\partial \phi}{\partial e_{xy}} = C_{44}e_{xy} \tag{2.46}$$

The corresponding relation from (2.34) is

$$\sigma_{xy} = C_{61}e_{xx} + C_{62}e_{yy} + C_{63}e_{zz} + C_{64}e_{yz} + C_{65}e_{zx} + C_{66}e_{xy} \tag{2.47}$$

Comparison of (2.46) with (2.47) gives

$$C_{66} = C_{44}; \qquad C_{61} = C_{62} = C_{63} = C_{64} = C_{65} = 0 \tag{2.48}$$

Proceeding this way for other stress components, we find that the array of values of the elastic stiffness constants of a cubic crystal may be expressed in the following matrix representation.

$$[C_{\alpha\beta}] = \begin{bmatrix} C_{11} & C_{12} & C_{12} & 0 & 0 & 0 \\ C_{12} & C_{11} & C_{12} & 0 & 0 & 0 \\ C_{12} & C_{12} & C_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & C_{44} & 0 \\ 0 & 0 & 0 & 0 & 0 & C_{44} \end{bmatrix} \tag{2.49}$$

with $\alpha, \beta = 1, 2, 3, 4, 5, 6$.

Matrix (2.49) shows that a cubic crystal has only three independent stiffness constants. By evaluating the inverse matrix to (2.49), we obtain the following relationships between the stiffness and compliance constants for cubic crystals.

$$C_{44} = 1/S_{44}; \qquad C_{11} - C_{12} = (S_{11} - S_{12})^{-1}; \qquad C_{11} + 2C_{12} = (S_{11} + 2S_{12})^{-1} \tag{2.50}$$

2.7.3 Bulk Modulus and Compressibility

Let us consider a strained crystal which is uniformly dilated. This refers to the mathematical condition

$$e_{xx} = e_{yy} = e_{zz} = \frac{1}{3}\delta \tag{2.51}$$

From (2.40) we obtain the following relation for the elastic energy density of a cubic crystal.

$$\phi = \frac{1}{6}(C_{11} + 2C_{12})\delta^2 \tag{2.52}$$

The bulk modulus B is usually defined by the relation

$$\phi = \frac{1}{2} B\delta^2 \tag{2.53}$$

Comparing (2.52) with (2.53), we express the bulk modulus of a cubic crystal as

$$B = \frac{1}{3}(C_{11} + 2C_{12}) \tag{2.54}$$

The inverse of B has been interpreted as another useful elastic property called the *compressibility* K. The compressibility of a cubic crystal is accordingly given by

$$K = \frac{3}{C_{11} + 2C_{12}} \tag{2.55}$$

2.8 ELASTIC WAVES IN CUBIC CRYSTAL

Consider an elementary cube of edges $\Delta x = \Delta y = \Delta z$ within the volume of a cubic crystal. The cube edges are in the directions of the x-, y- and z-axes. When the cube is strained, let the stress $\sigma_{xx}(x)$ act on the face at x. Assuming that the variation of σ_{xx} is uniform along the x-direction, the stress on the face parallel to that at x can be expressed as $[\sigma_{xx}(x) + \{\partial\sigma_{xx}/(\partial/\partial_x)\}\Delta x]$ (see Fig. 2.9). The net force on the cube due to σ_{xx} component is equal to $[\{\partial\sigma_{xx}/(\partial/\partial_x)\}\Delta x]\,\Delta y \cdot \Delta z$. The other forces in the x-direction arise from the variation of σ_{xy} and σ_{xz} across the cube. Therefore, the net force on the cube along the x-direction is

$$F_x = \left(\frac{\partial\sigma_{xx}}{\partial x} + \frac{\partial\sigma_{xy}}{\partial y} + \frac{\partial\sigma_{xz}}{\partial z}\right)\Delta x \cdot \Delta y \cdot \Delta z \tag{2.56}$$

The above force is actually the restoring force that tends to bring the cube to its unstrained state. As a result, the particles in the crystal are thrown to a motion described by the relevant equation of motion. If ρ is the density of the crystal, the force per unit volume on the crystal along the x-direction is $\rho(\partial^2 u_1/\partial t^2)$ and the equation of motion in the x-direction becomes

$$\rho\frac{\partial^2 u_1}{\partial t^2} = \frac{\partial\sigma_{xx}}{\partial x} + \frac{\partial\sigma_{xz}}{\partial z} + \frac{\partial\sigma_{xy}}{\partial y} \tag{2.57}$$

Using (2.34) and (2.49), the above equation is reduced to the following form for a cubic crystal.

$$\rho\frac{\partial^2 u_1}{\partial t^2} = C_{11}\frac{\partial e_{xx}}{\partial x} + C_{12}\left(\frac{\partial e_{yy}}{\partial x} + \frac{\partial e_{zz}}{\partial x}\right) + C_{44}\left(\frac{\partial e_{xy}}{\partial y} + \frac{\partial e_{zx}}{\partial z}\right) \tag{2.58}$$

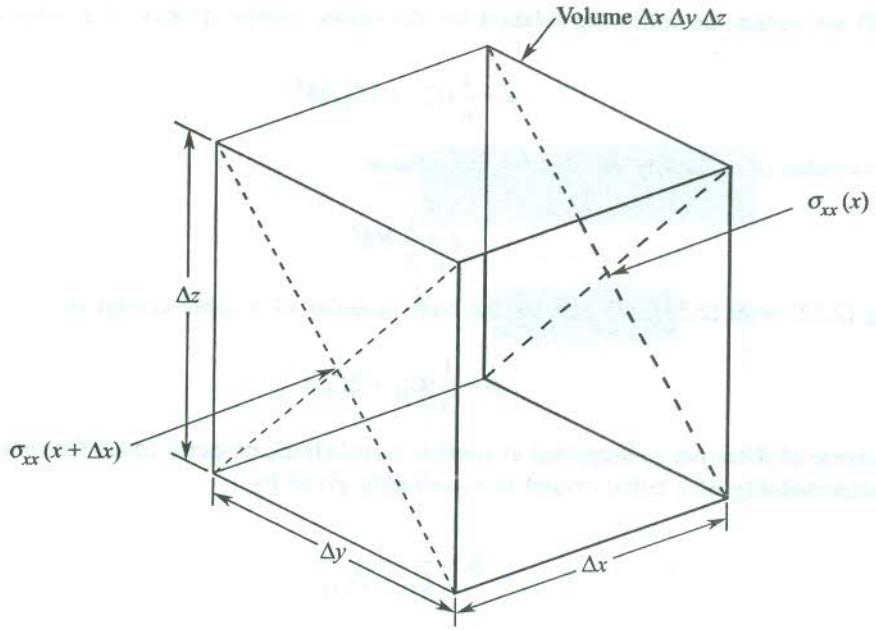


FIG. 2.9 Variation of a normal stress component (σ_{xx}) within a cube of volume $\Delta x \cdot \Delta y \cdot \Delta z$ (with $\Delta x = \Delta y = \Delta z$) along the x-edge.

On eliminating the strain components with the help of (2.29), the above equation assumes the form

$$\rho \frac{\partial^2 u_1}{\partial t^2} = C_{11} \frac{\partial^2 u_1}{\partial x^2} + C_{44} \left(\frac{\partial^2 u_1}{\partial y^2} + \frac{\partial^2 u_1}{\partial z^2} \right) + (C_{12} + C_{44}) \left(\frac{\partial^2 u_2}{\partial x \partial y} + \frac{\partial^2 u_3}{\partial x \partial z} \right) \tag{2.59}$$

Here u_1, u_2, u_3 are the components of the atomic displacement $\Delta(\mathbf{r})$ along the axes of the unstrained crystal [see equation (2.27)]. The solution to (2.59) turns out to have a waveform, indicating that waves propagate within the crystal when it is strained in such a way as there exists a non-zero stress on the crystal. These waves are called the *elastic waves* because they are produced here in an elastic continuum by elastic deformation.

Similar to (2.59), there are equations of motion in the y- and z-directions. We write them by symmetry on the basis of (2.59):

$$\rho \frac{\partial^2 u_2}{\partial t^2} = C_{11} \frac{\partial^2 u_2}{\partial y^2} + C_{44} \left(\frac{\partial^2 u_2}{\partial x^2} + \frac{\partial^2 u_2}{\partial z^2} \right) + (C_{12} + C_{44}) \left(\frac{\partial^2 u_1}{\partial x \partial y} + \frac{\partial^2 u_3}{\partial y \partial z} \right) \tag{2.60}$$

$$\rho \frac{\partial^2 u_3}{\partial t^2} = C_{11} \frac{\partial^2 u_3}{\partial z^2} + C_{44} \left(\frac{\partial^2 u_3}{\partial x^2} + \frac{\partial^2 u_3}{\partial y^2} \right) + (C_{12} + C_{44}) \left(\frac{\partial^2 u_1}{\partial x \partial z} + \frac{\partial^2 u_2}{\partial z \partial x} \right) \tag{2.61}$$

Now, we proceed to study the solutions of these equations in some common directions in cubic crystals.

2.8.1 Propagation of Waves in the [100] Direction

First, consider a longitudinal or compressional wave that propagates along the x-cube edge. Its propagation constant or wavevector k is parallel to the particle displacement u_1, given by

$$u_1 = (u_1)_0 \exp\,[i(kx - \omega t)] \tag{2.62}$$

where $k = 2\pi/\lambda$, and ω is the angular frequency. When (2.62) is used as a trial solution of (2.59) and placed in it, we obtain the dispersion relation

$$\omega^2\rho = C_{11}k^2 \tag{2.63}$$

This gives the velocity of a longitudinal wave in the [100] direction in a cubic crystal as

$$v_l = \frac{\omega}{k} = \left(\frac{C_{11}}{\rho}\right)^{1/2} \tag{2.64}$$

Next, we consider a transverse or shear wave with its wavevector k along the x-cube edge and the particle displacement u_2 in the y-direction. Thus,

$$u_2 = (u_2)_0 \exp\,[i(kx - \omega t)] \tag{2.65}$$

The substitution of (2.65) into (2.60) gives

$$\omega^2\rho = C_{44}k^2 \tag{2.66}$$

Therefore, the velocity of a transverse wave in the [100] direction in a cubic crystal is

$$v_s = \frac{\omega}{k} = \left(\frac{C_{44}}{\rho}\right)^{1/2} \tag{2.67}$$

It can be shown that a transverse wave with wavevector along the x-cube edge and the particle displacement u_3 in the z-direction moves with the identical velocity. This result asserts that two independent shear waves whose wavevectors point along the [100] direction propagate in a cubic crystal with equal velocities. For a general direction of the wavevector this result is not applicable. The geometry of the longitudinal and transverse waves propagation in the [100] direction in a cubic crystal is shown in Fig. 2.10(a).

2.8.2 Propagation of Waves in the [110] Direction

The study of elastic waves propagating in the [110] direction (the direction of a face diagonal) is especially gainful because the three elastic constants can be obtained simply from the three measured propagation velocities.

Let us first consider a shear wave with wavevector $\mathbf{k} = k_x\hat{\mathbf{i}} + k_y\hat{\mathbf{j}}$ propagating in the xy plane and causing a particle displacement u_3 in the z-direction. Thus

$$u_3 = (u_3)_0 \exp\,[i(k_x x + k_y y - \omega t)] \tag{2.68}$$

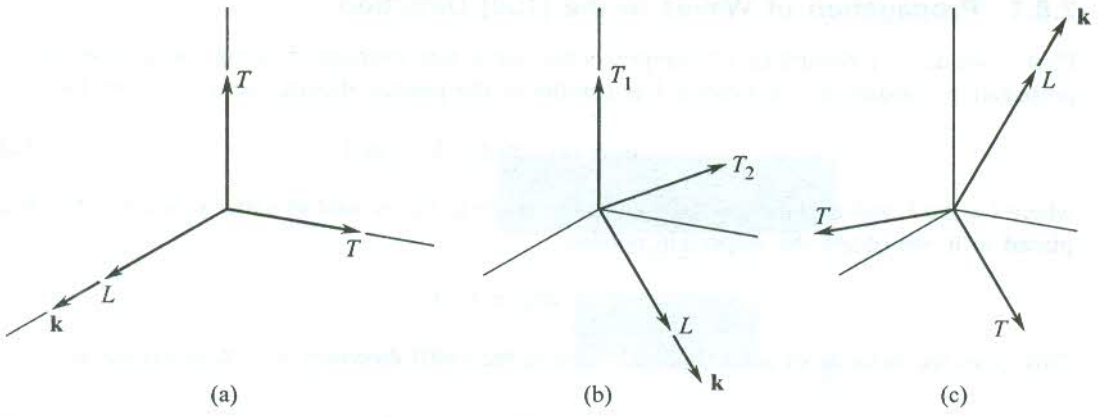


FIG. 2.10 Propagation of elastic waves in a cubic crystal—L denotes a longitudinal wave and T denotes a shear wave. (a) Waves in the [100] direction ($\mathbf{k}$ parallel to $\hat{\mathbf{i}}$)—one longitudinal wave with velocity characterized by C_{11}, two degenerate shear waves with velocity characterized by C_{44}. (b) Waves in the [110] direction [$\mathbf{k}$ parallel to $(\hat{\mathbf{i}} + \hat{\mathbf{j}})$]: one longitudinal wave with velocity characterized by $1/2(C_{11} + C_{12} + 2C_{44})$, two shear waves with velocities characterized respectively by C_{44} and $1/2(C_{11} - C_{12})$. (c) Waves in the [111] direction [$\mathbf{k}$ parallel to $(\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}})$]: one longitudinal wave with velocity characterized by $1/3(C_{11} + 2C_{12} + C_{44})$, two degenerate shear waves with velocity characterized by $1/3(C_{11} - C_{12} + C_{44})$. $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, $\hat{\mathbf{k}}$ denote unit vectors along the cube edges defining the x-, y- and z-coordinate axes.

Substituting (2.68) into (2.61), we obtain

$$\omega^2 = \left(\frac{C_{44}}{\rho}\right) k^2 \tag{2.69}$$

Next, consider the other two waves propagating in the xy plane with particle velocity in the xy-plane. Let these be represented as

$$\begin{aligned} u_1 &= (u_1)_0 \exp i(k_x x + k_y y - \omega t) \\ u_2 &= (u_2)_0 \exp i(k_x x + k_y y - \omega t) \end{aligned} \tag{2.70}$$

Placing these solutions in (2.59) and (2.60), we get the following pair of equations:

$$\begin{aligned} \omega^2 \rho u_1 &= (C_{11} k_x^2 + C_{44} k_y^2) u_1 + (C_{12} + C_{44}) k_x k_y u_2 \\ \omega^2 \rho u_2 &= (C_{11} k_y^2 + C_{44} k_x^2) u_2 + (C_{12} + C_{44}) k_x k_y u_1 \end{aligned} \tag{2.71}$$

For a wave in the [110] direction for which $k_x = k_y = k/\sqrt{2}$, equations (2.71) have a characteristically simple solution. The solution exists only if the determinant of the coefficients of u_1 and u_2 in (2.71) vanishes. Thus

$$\begin{vmatrix} \frac{1}{2}(C_{11} + C_{44}) k^2 - \omega^2 \rho & \frac{1}{2}(C_{12} + C_{44}) k^2 \\ \frac{1}{2}(C_{12} + C_{44}) k^2 & \frac{1}{2}(C_{11} + C_{44}) k^2 - \omega^2 \rho \end{vmatrix} = 0 \tag{2.72}$$

On solving (2.72), we get two dispersion relations

$$\omega^2 = \left(\frac{C_{11} + C_{12} + 2C_{44}}{2\rho}\right)k^2; \quad \omega^2 = \left(\frac{C_{11} - C_{12}}{2\rho}\right)k^2 \tag{2.73}$$

These two roots refer to two different types of waves. We now determine the nature of the waves by finding the direction of the particle displacement caused by the respective waves. When we substitute the first root in the first equation of (2.71) we get $u_1 = u_2$. Since the particle displacement occurs in the xy plane, $\Delta(\mathbf{r}) = u_1(\hat{\mathbf{i}} + \hat{\mathbf{j}})$. This shows that the displacement takes place in the [110] direction that represents the direction of $(\hat{\mathbf{i}} + \hat{\mathbf{j}})$ and happens to be the direction of propagation of the wave [see Fig. 2.10(b)]. Thus we infer that the first root in (2.73) belongs to a longitudinal wave.

Similarly, on substituting the second root in the first equation of (2.71), we obtain $u_1 = -u_2$, implying that $\Delta(\mathbf{r}) = u_1(\hat{\mathbf{i}} - \hat{\mathbf{j}})$. The direction of the vector $(\hat{\mathbf{i}} - \hat{\mathbf{j}})$ is indicated as $[1\,\bar{1}\,0]$ which is perpendicular to the [110] direction, the direction of propagation of the wave. Hence the second root in (2.73) must refer to a shear wave [Fig. 2.10(b)]. The treatment of waves in the [111] direction is relatively lengthy though not so complicated as for other general directions. The main features of propagation may, however, be found in Fig. 2.10(c).

The subject matter of the last section is closely linked to the analysis of normal modes of vibration of crystals. Generally, the direction of particle displacement or the polarization of normal modes may not be exactly parallel or perpendicular to the direction of the wavevector $\mathbf{k}$. The analysis becomes easier when $\mathbf{k}$ is along any symmetry axis of the crystal. A suitable discussion dealing with these aspects is presented in Section 4.6.

SUMMARY

1. (a) Bonds are broadly classified on the basis of strength as (i) primary bonds and (ii) secondary bonds
   (b) The three limiting cases of primary bonds are identified as (a) covalent, (b) metallic, and (c) ionic
   (c) Some primary bonds are of mixed type.
   (d) van der Waals and hydrogen bonds fall in the category of secondary bonds.
2. van der Waals interaction (induced dipole–dipole interaction) varies with interatomic separation as $1/r^6$. It is a quantum effect. The interaction potential vanishes when the Planck constant $h = 0$. Example, inert gas solids.
3. The cause of repulsive interaction between atoms lies generally in the electrostatic repulsion of overlapping charge distributions and the Pauli exclusion principle that forces overlapping electrons of parallel spin occupy higher energy states.
4. The potential energy per ion in an ionic crystal is given by

$$U_{\text{ion}} = -\frac{\alpha q^2}{4\pi \epsilon_0 r_0}\left(1 - \frac{\rho}{r_0}\right)$$

where

α is the Madelung constant
q is the charge on an ion
ρ is an empirical parameter
r_0 is the separation in a pair of ions of opposite type at equilibrium.

5. The overlap of charge distributions of antiparallel electron spin gives rise to covalent bonding.

6. There are only three independent elastic stiffness constants for a cubic crystal: C_{11}, C_{12}, C_{44}. This number increases with decrease in the crystalline symmetry (21 for triclinic symmetry).

PROBLEMS

2.1 Briefly answer the following questions.

(a) If positive and negative ions attract each other, why does the structure not collapse?
(b) Why is it that the cohesive energy of an ionic crystal is almost equal to the attractive Coulomb energy?
(c) What is the origin of cohesion in metals? Can it be described by interatomic potentials or covalent or any other bonds?

2.2 Consider a system of two argon atoms with position of the second atom with respect to the location of the first atom being denoted by vector **r** at certain instant of time. If for a fraction of second the centre of positive charges is shifted from the centre of negative charges (electrons) in the first atom, show that this would lead to an attractive potential of the system expressible as

$$U = -\frac{\alpha_e p^2}{(4\pi \epsilon_0)^2 r^6} (1 + 3\cos^2\theta)$$

where

α_e is the electronic polarizability of the argon atom
p is the instantaneous electric dipole moment of the first atom
θ is the angle between **p** and **r**.

2.3 Show that the Madelung constant for one-dimensional array of ions of alternating sign with a distance a between successive ions is given by 2 ln 2.

2.4 Calculate the cohesive energy per ion-pair of LiCl and KCl crystals using the following data:

	LiCl	KCl
Madelung constant	1.748	1.748
$Li^+ - Cl^-/K^+ - Cl^-$ spacing	2.57 Å	3.14 Å
Ionization energy of Li/K	5.4 eV	4.34 eV

Assume that the repulsive potential energy is negligibly small.

2.5 Assume that the repulsive potential energy of an ion-pair in an ionic crystal arises because of the action of the Pauli exclusion principle and that it can be expressed as

$$U_{\text{rep.}} = \frac{B}{r^n}$$

where r is the separation between the ions of the opposite sign, n is a large number and B is a constant of the crystal.

Show that the total potential energy per ion-pair for equilibrium separation r_0 is

$$U = -\frac{\alpha e^2}{4\pi \epsilon_0 r_0}\left(1 - \frac{1}{n}\right)$$

The observed cohesive energy of LiCl crystal is 6.8 eV per ion-pair. Using the data given in Problem 2.4, evaluate the value of n in the repulsive potential energy term (refer to the above problem).

2.6 Calculate the cohesive energy of hydrogen in kJ per mole using the Lennard–Jones parameters

$$\varepsilon = 5 \times 10^{-22} \text{ J} \quad \text{and} \quad \sigma = 2.96 \text{ Å}$$

Assume that hydrogen molecules are spherical and crystallize into an FCC structure. Compare your result with the measured value (0.751 kJ per mole) and comment.

2.7 In a crude model of alkali metals, the charge of each electron is treated as uniformly distributed over the volume of a sphere of radius r_s centred at each ion. Prove that the electrostatic energy per electron is then given by

$$U_{\text{coul}} = -\frac{9\,a_0}{5\,r_s} \text{ rydberg}$$

$$= \frac{24.49}{r_s/a_0} \text{ eV}$$

where a_0 is the Bohr radius.

2.8 Symbols $r^>$ and $r^<$ represent the radius of the bigger and smaller atom, respectively, in a crystal of diatomic basis. Show that the critical ratio,

$$\frac{r^>}{r^<} = \frac{\sqrt{3}+1}{2} \text{ for the CsCl structure}$$

$$= 2 + \sqrt{6} \text{ for the zinc blende structure}$$

2.9 Show that the longitudinal and shear wave velocities in the [111] direction in a cubic crystal are, respectively, given by

$$v_l = \left[\frac{C_{11} + 2C_{12} + 4C_{44}}{3\rho}\right]^{1/2}$$

$$v_s = \left[\frac{C_{11} - C_{12} + C_{44}}{3\rho}\right]^{1/2}$$

2.10 Show that in a cubic crystal, the effective elastic constant for a shear across the (110) plane in the $[1\bar{1}0]$ direction is equal to $\frac{(C_{11} - C_{12})}{2}$.

SUGGESTED FURTHER READING

Evans, R.C., *An Introduction to Crystal Chemistry* (Cambridge, 1964).

Horton, G.K., "Ideal Rare Gas Crystals", *Amer. J. Phys.*, **36**, 93(1968).

Maitland, G., et al., *Intermolecular Forces: Their origin and determination* (Oxford, 1981).

Pauling, L., *Nature of the Chemical Bond,* 3rd ed. (Cornell, 1960).

Phillips, J.C., *Covalent Bonding in Crystals, Molecules and Polymers* (The Univ. of Chicago Press, 1969).

Chapter 3

Reciprocal Lattice and Determination of Crystal Structure

This chapter is devoted to the description of principles and experimental methods that form the basis for determining crystal structures. We learnt about crystal structures in Chapter 1 in terms of the lattice symmetry, basis composition, and various intersecting families of planes as visualized in the space defined by the three crystal edges. As we already know, the most common experimental technique employed to investigate crystal structures is the x-ray diffraction method. The crystal photograph as recorded on the x-ray film represents the diffraction pattern of the crystal in which each of the spots corresponds to the diffraction maximum of a different family of parallel planes, with a certain orientation and common spacing. The derivation of the actual crystal structure from this photographic pattern in terms of atomic positions and orientation of planes in the usual three-dimensional space is not straightforward. This is done by having various geometrical constructions in the recorded photograph the details of which can be found in any book on crystallography.

Nevertheless, it is worthwhile to find some correlation between the structure of a crystal and its planar x-ray photograph. A family of parallel planes may as well be represented by their common normal. It is far more easy to visualize the motion of a normal (equivalent to a rod) than that of a plane which is a two-dimensional figure. If a length in some proportion of the common spacing be chosen and signified by a point along the normal, the point acts as a perfect representative of the family of planes under consideration. The theory discussed in this chapter will show that the x-rays diffracted from the family of planes produce the maximum intensity at this point. Further, the analysis shows that the distance of the point from an origin fixed at an arbitrary lattice point on the normal is proportional to the reciprocal of the interplanar spacing. The dimension of the representative position vector is thus the reciprocal of length. It is due to this reason that each of such points is called a *reciprocal lattice point.* These points happen to be arranged in a regular pattern forming a space lattice which is referred to as *the reciprocal lattice.* In this spirit, the regular pattern of spots on the film (Fig. 3.1) may be treated as the two-dimensional map of a three-dimensional reciprocal lattice for a certain incident direction of x-rays in the crystal. An appropriate treatment of the reciprocal lattice is taken up in the following section with the aim to establish the relationship between the direct lattice and the reciprocal lattice of a crystal.

3.1 RECIPROCAL LATTICE

The simple interpretation of reciprocal lattice as made above needs a suitable analytical extension for its complete treatment. Let us consider some functions whose values are periodic in the crystal lattice. These could be functions representing physical quantities such as electron charge density, or

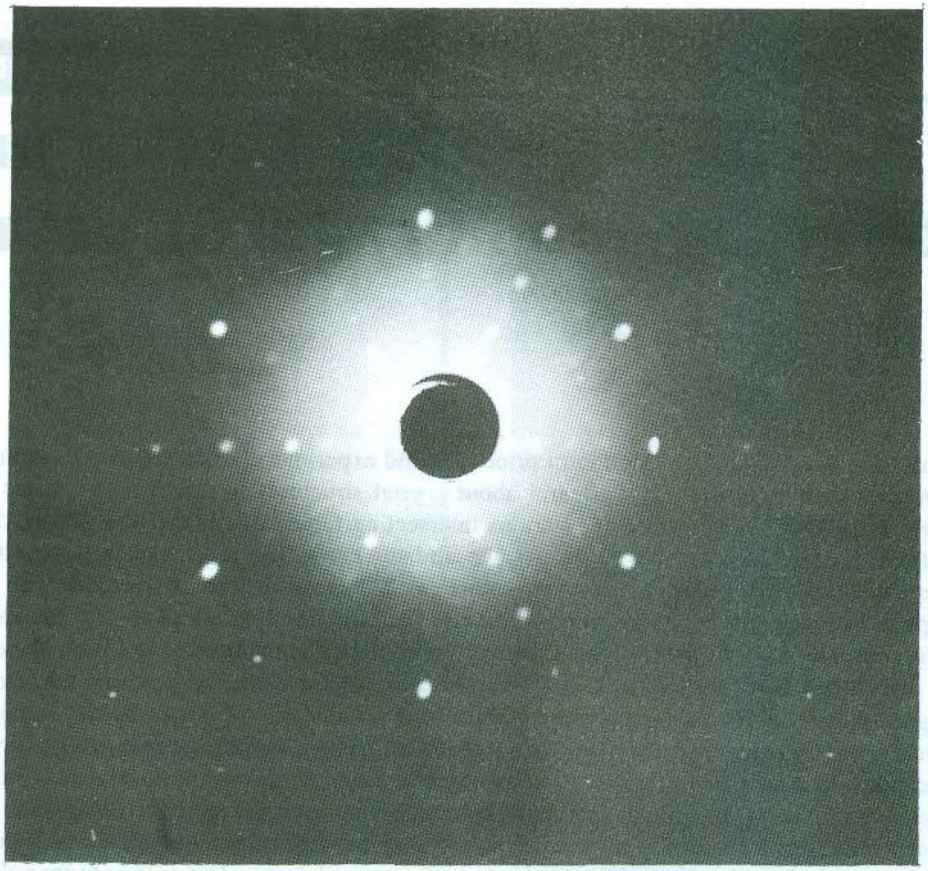


FIG. 3.1 Laue photograph of CdTe crystal (a = 6.48 Å) in [111] orientation. x-rays are incident normally on a (111) plane in [111] direction of the crystal; true for cubic crystals only. (Courtesy B.B. Sharma)

electrostatic crystal potential. We take the example of the crystal potential which is periodic or whose value is the same at all such points that are connected by translation vectors of the lattice. For a monatomic one-dimensional crystal (a linear chain of identical atoms), this potential $U(x)$ is shown in Fig. 3.2. The periodic character of $U(x)$ demands

$$U(\mathbf{x} + \mathbf{t}) = U(\mathbf{x}) \tag{3.1}$$

with $\mathbf{t} = m\mathbf{a}$

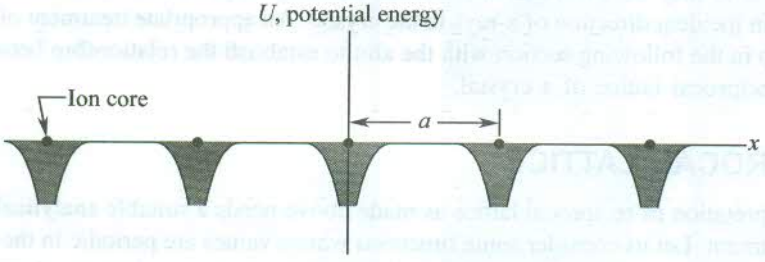


FIG. 3.2 The periodic potential energy of an electron inside a linear crystal of identical atoms.

where m is an integer and $|\mathbf{a}|$ the lattice constant. Since $U(\mathbf{x})$ is a periodic function, it can be expanded in a Fourier series

$$U(\mathbf{x}) = \sum_n U_n \exp \frac{i2\pi n\mathbf{x}}{\mathbf{a}} \tag{3.2}$$

n being an integer.

This form of $U(\mathbf{x})$ ensures that its periodic nature expressed by (3.1) remains valid. The series (3.2) is generally expressed in the form

$$U(\mathbf{x}) = \sum_{\mathbf{g}} U_{\mathbf{g}} \exp\ (i\mathbf{g} \cdot \mathbf{x}) \tag{3.3}$$

where

$$\mathbf{g} = \frac{2\pi n}{\mathbf{a}} \tag{3.4}$$

with $n = 1, 2, 3, \ldots$.

The vector $\mathbf{g}$ has the dimension of the inverse of length vector and it is accordingly identified as the reciprocal vector. The coefficients $U_{\mathbf{g}}$ are the usual Fourier coefficients. The set of reciprocal vectors given by (3.4) may be viewed as translation vectors of a one-dimensional lattice with the repeat interval of $2\pi/\mathbf{a}$. In other words the vectors $2\pi n/\mathbf{a}$ represent position vectors of the lattice points of a one-dimensional reciprocal lattice. The set of vectors $2\pi n/\mathbf{a}$ is known as the *reciprocal lattice vectors*. Further, the periodic character of $U(\mathbf{x})$ requires,

$$\exp\ (i\mathbf{g} \cdot \mathbf{t}) = 1 \tag{3.5}$$

or

$$\mathbf{g} \cdot \mathbf{t} = \text{an even multiple of } \pi$$

This condition is readily satisfied by the expressions for $\mathbf{g}$ and $\mathbf{t}$. The above treatment may be easily extended to the three dimensions where

$$U(\mathbf{r}) = \sum_{\mathbf{g}} U_{\mathbf{g}} \exp\ (i\mathbf{g} \cdot \mathbf{r}) \tag{3.6}$$

with the expression for $\mathbf{g}$ being given by

$$\mathbf{g} = m_1\mathbf{a}^* + m_2\mathbf{b}^* + m_3\mathbf{c}^* \tag{3.7}$$

where

$$\mathbf{a}^* = \frac{2\pi}{\mathbf{a}}, \quad \mathbf{b}^* = \frac{2\pi}{\mathbf{b}} \quad \text{and} \quad \mathbf{c}^* = \frac{2\pi}{\mathbf{c}} \tag{3.8}$$

Here $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$ represent the primitive translation vectors of the reciprocal lattice as against $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ for the direct lattice. In other words, $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$ form the three edges of the primitive cell of the reciprocal lattice under discussion. But it is customary to use a Wigner–Seitz type primitive cell for the purpose of calculations related to a reciprocal lattice. The method of construction of this cell is the same as in the direct space described in Chapter 1. The Wigner–Seitz cell of a reciprocal lattice is known by a special name—First Brillouin Zone. The first Brillouin zone for a two-dimensional reciprocal lattice is drawn in Fig. 3.3. These zones each of which is a closed figure fill the entire space of the reciprocal lattice. From the nomenclature we may question if there exist second, third or higher

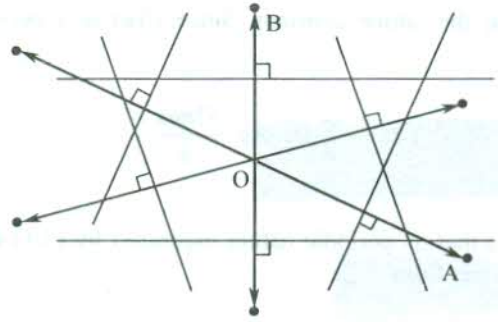


FIG. 3.3 Construction of the first Brillouin zone for a two-dimensional reciprocal lattice.

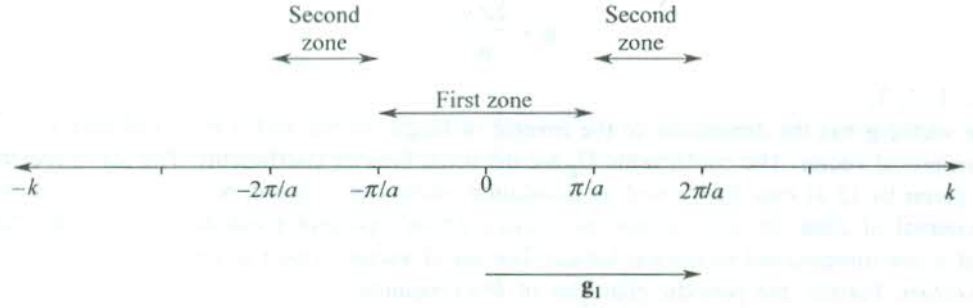


FIG. 3.4 Brillouin zones of a one-dimensional reciprocal lattice. Note that only the first zone appears as a single piece. $\mathbf{g}_1$ denotes the shortest reciprocal lattice vector $2\pi/\mathbf{a}$.

order Brillouin zones as well. The answer is yes. The Brillouin zones for a reciprocal lattice in one dimension are shown in Fig. 3.4.

The concept of reciprocal space may easily gain ground in our mind if we can identify some physical quantities that are measured in this space. One such simple physical quantity is the propagation constant which is more popularly known as wavevector (**k**) in solid state theory. This vector, having the dimension of the inverse of length ($|\mathbf{k}| = 2\pi/\lambda$; λ is the wavelength), is a reciprocal vector and, therefore, rightly lies in the reciprocal space. On account of this, the reciprocal space is often referred to as the *k*-space.

A complete description of the reciprocal lattice cannot be achieved without coming to grips with the basic principles of the theory of x-ray diffraction. There are two parallel approaches to this theory. One is based on the Bragg law and the other is Max von Laue's formulation. But it will be shown that the two in effect are equivalent to one another. The theory succeeds in giving the direction of reciprocal lattice vectors that connect the reciprocal lattice points to the origin in the reciprocal space. A reciprocal lattice point will be found to be along the direction of x-rays reflected by a family of parallel planes with a certain orientation and common interplanar spacing. We discuss below first the Bragg law and then follow it with the Laue's version of x-ray diffraction.

3.2 BRAGG LAW

Strong x-ray diffraction produced by crystalline solids is based on Bragg law. Crystals are composed of various intersecting planes each of which itself contains a number of atoms. Since the interatomic

separations in crystals are of the same order of magnitude as the wavelength of x-rays (~Å), crystals act as diffraction gratings and produce diffraction on being irradiated by x-rays. W.H. Bragg and his son W.L. Bragg observed in 1913 such characteristic effects in the x-ray radiation reflected from crystals. In this phenomenon, the crystal planes are believed to act like plane mirrors. Radiations reflected from two successive parallel planes under certain conditions may interfere constructively to produce a diffraction maximum. The Bragg diffraction shown in Fig. 3.5 occurs for specular reflections (angle of incidence = angle of reflection). The constructive interference occurs when the path difference ($2d \sin \theta$) between the interfering rays equals an integral multiple of the x-ray wavelength λ. That is,

$$2d \sin \theta = n\lambda \tag{3.9}$$

where

d is the interplanar spacing

θ is the angle of the incident radiation with the plane

$n = 1, 2, 3, \ldots$ (order of diffraction)

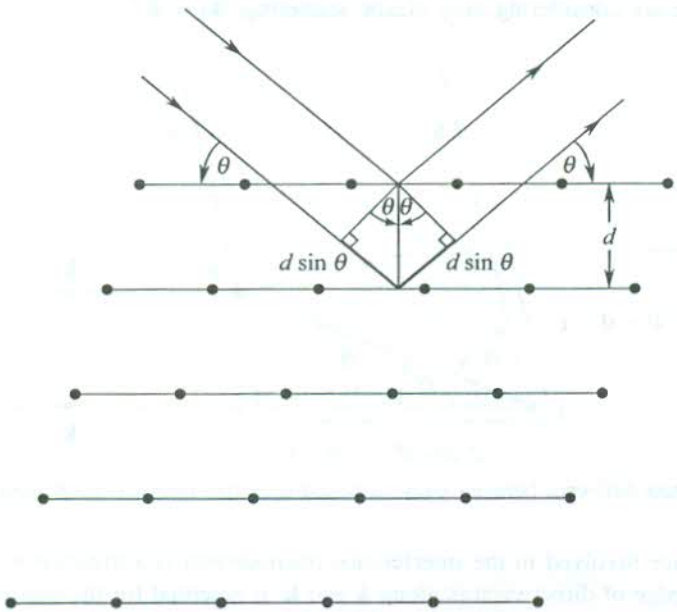


FIG. 3.5 Bragg reflection from a family of planes with interplanar spacing d. Note that the incident beam is deflected by twice the Bragg angle θ.

The relation (3.9), which is a mathematical statement of the Bragg law, shows that the diffraction effects cannot be observed from a family of planes for any arbitrary angle of incidence. Even if monochromatic x-rays are used, the match among d, θ and λ has to be sought to satisfy (3.9) and get the Bragg diffraction. So, it is clear that Bragg diffraction is very much different from ordinary diffraction which generally puts no restriction on the incident angle.

Further, since $\sin \theta \leq 1$,

$$\lambda \leq 2d \tag{3.10}$$

The condition expressed by (3.10) clearly explains why Bragg diffraction cannot occur for the visible

radiation. The Bragg law as expressed by (3.9) is essentially the consequence of the periodicity with only the elastic scattering of radiation taken into consideration.

3.3 LAUE'S INTERPRETATION OF X-RAY DIFFRACTION BY CRYSTALS

Max von Laue is credited for giving the idea that x-rays can produce constructive interference after getting diffracted from crystals. In 1912 two of Laue's colleagues Friedrich and Knipping at Münich did experiments which proved his theory. The Bragg law was established at Cambridge in 1913 with the awareness of the research activity of the Laue's group. In Laue's theory, the crystal is considered to be made up of identical microscopic set of ions or atoms. The x-rays scattered from these individual objects within each plane are believed to interfere constructively.

Let the x-rays of the wavevector **k** be incident on two identical atoms at a separation of d (Fig. 3.6). The two atoms being identical would scatter x-rays in the direction with a common wavevector **k′**. As we are considering only elastic scattering, $|\mathbf{k}| = |\mathbf{k}'|$.

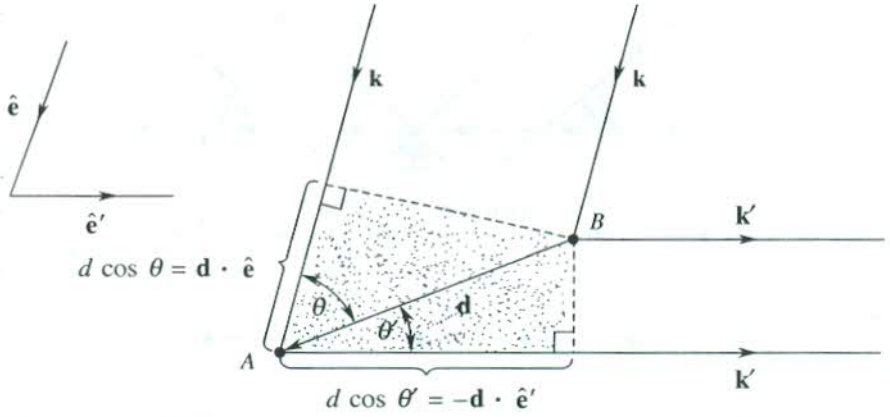


FIG. 3.6 Finding the path difference between x-rays scattered from two points A and B separated by distance d.

The path difference involved in the interference phenomenon is a distance in the direct space. Therefore, the knowledge of direct vectors along **k** and **k′** is essential for the calculation of the path difference between the x-rays scattered by atom A and those scattered by atom B. On the assumption of the direct unit vectors $\hat{\mathbf{e}}$ and $\hat{\mathbf{e}}'$ along **k** and **k′**, respectively, we solve the problem.

$$\text{The path difference} = d\cos\theta + d\cos\theta'$$

$$= \mathbf{d}\cdot\hat{\mathbf{e}} - \mathbf{d}\cdot\hat{\mathbf{e}}'$$

$$= \mathbf{d}\cdot(\hat{\mathbf{e}} - \hat{\mathbf{e}}')$$

The path difference at a point of constructive interference should meet the condition,

$$\mathbf{d}\cdot(\hat{\mathbf{e}} - \hat{\mathbf{e}}') = m\lambda \tag{3.11}$$

where m is an integer.

Multiplying both sides of (3.11) by $2\pi/\lambda$, we get

$$\mathbf{d} \cdot (\mathbf{k} - \mathbf{k}') = 2\pi m \tag{3.12}$$

When we consider the full crystal as an array of scatterers, the vector **d** in (3.12) should be replaced by a general lattice vector **t**, and thus

$$\mathbf{t} \cdot (\mathbf{k} - \mathbf{k}') = 2\pi m \tag{3.13}$$

which can be further transformed to

$$\exp [i(\mathbf{k} - \mathbf{k}')] \cdot \mathbf{t} = 1 \tag{3.14}$$

Comparison of (3.14) with (3.5) suggests that

$$\mathbf{k} - \mathbf{k}' = \mathbf{g} \tag{3.15}$$

where **g** is a reciprocal lattice vector. The relation (3.15) states the Laue condition of diffraction. From (3.15), we have

$$\mathbf{k}' = \mathbf{k} - \mathbf{g}$$

which leads the Laue condition to the following form

$$2\mathbf{k} \cdot \mathbf{g} = \mathbf{g}^2 \text{ (using } |\mathbf{k}| = |\mathbf{k}'|) \tag{3.16}$$

The Laue condition for a wavevector **k** to produce constructive interference in the diffracted radiation is contained in (3.16). The geometrical relationship between **k**, **k′** and **g** is demonstrated by Fig. 3.7. The plane perpendicular to the plane of the figure and passing through the horizontal line can be taken as the Bragg plane in Laue's formulation, which is essentially the reflecting plane in the Bragg picture. In this picture, the figure clearly represents the property that the reciprocal lattice vector (**g**) is perpendicular to the reflecting plane (and hence to the respective family of the reflecting planes).

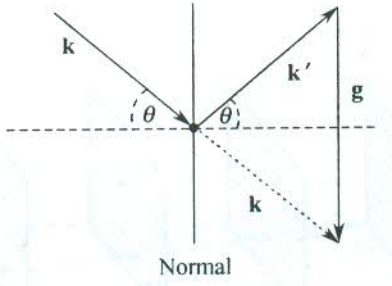


FIG. 3.7 The geometrical relationship between the wavevector of incident (**k**) and scattered (**k′**) x-rays. For elastic scattering $|\mathbf{k}| = |\mathbf{k}'|$. The scatterers are two identical atoms.

By analogy with (3.4), the general relation for **g** is

$$\mathbf{g} = \frac{2\pi n}{\mathbf{d}} \tag{3.17}$$

since the repeat distance in the present case is d.

The shortest $\mathbf{g}(=\mathbf{g}_1)$ of the set of vectors expressed by (3.17) measures $(2\pi/\mathbf{d})$. It is now very simple to derive the Bragg law from the Laue's theory.

From Fig. 3.7, we get

$$\mathbf{k} \cdot \mathbf{g} = kg \cos(90 - \theta)$$

$$= kg \sin \theta \tag{3.18}$$

Using (3.17) and (3.18), Laue condition (3.16) transforms to

$$2d \sin \theta = n\lambda \tag{3.19}$$

The parameters appearing in (3.19) and used in the two formulations independently refer to the same peak in the diffraction pattern of the crystal under consideration. The analysis of geometrical scattering as per Laue's formulation thus proves beyond doubt that the approaches of Bragg and Laue are equivalent. One important point emerging from the theory discussed above is that the reciprocal lattice point, representing a family of parallel planes with a common interplanar spacing, lies along the common normal of planes drawn from the origin in the reciprocal space. This concept is utilized in constructing the reciprocal lattice of a crystal lattice.

3.4 CONSTRUCTION OF RECIPROCAL LATTICE

In this section we describe in brief how the reciprocal lattice of a crystal lattice (or direct lattice) can be constructed. First of all, fix the origin at any arbitrary lattice point O in the direct lattice and then identify the three crystal axes (**a**, **b**, **c**) as shown in Fig. 3.8. Let us now draw a family of d_{100} planes where d stands for the interplanar separation and (100) are the Miller indices. Drop a normal **OP** to these set of planes from the origin O. The vector **OP** in this case represents the direction of **a***. The complete set of reciprocal lattice vectors along this direction is given by

$$\mathbf{g} = \frac{2\pi n^{\dagger}}{\mathbf{d}_{100}}, \qquad (n = 1, 2, 3, \ldots) \tag{3.20}$$

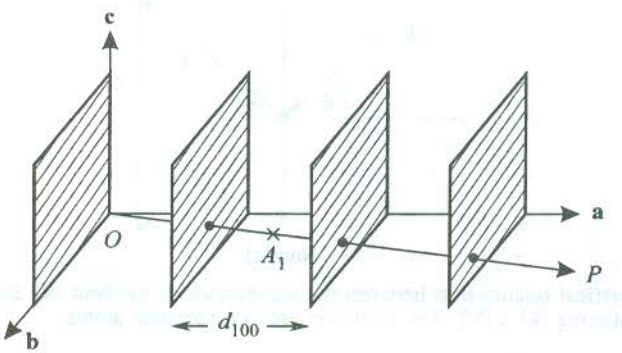


FIG. 3.8 Construction of the reciprocal lattice. **a**, **b**, **c** represent the primitive axes of the direct lattice. A_1 is the reciprocal lattice point referred to d_{100} planes and lies on the normal to d_{100} planes. The normal **OP** passing through the origin O gives the direction of the **a*** axis of the reciprocal lattice ($|\mathbf{OA}_1| = |\mathbf{a}^*|$).

[†] Here 2π is called the scale factor or the magnification factor for OA_1 that defines the reciprocal point. The value of this factor is a matter of convenience. Other values commonly used are 1 and λ (the wavelength of radiation).

The tip of these vectors will be located at the respective reciprocal lattice points A_1, A_2, A_3, ... The smallest of these $\mathbf{OA}_1$, is denoted by $\mathbf{a}^*$ and expressed as

$$\mathbf{OA}_1 = \mathbf{a}^* = \frac{2\pi}{\mathbf{d}_{100}} = \frac{2\pi\hat{\mathbf{e}}}{d_{100}} \tag{3.21}$$

where $\hat{\mathbf{e}}$ stands for the unit vector along **OP** or $\mathbf{a}^*$.

Thus A_1 represents the reciprocal lattice point corresponding to d_{100} planes and A_2 will correspond to d_{200} planes which are parallel to d_{100} planes with half common spacing.

Therefore, it is reasonable to write

$$\mathbf{OA}_2 = \frac{2\pi}{\mathbf{d}_{200}} = \frac{2\pi\hat{\mathbf{e}}}{d_{200}}$$

It is also very clear that a reciprocal lattice point truly symbolizes a family of parallel planes because its position is defined implicitly by the orientation and explicitly by the common spacing of planes. And the general expression for reciprocal lattice vectors along $\mathbf{a}^*$, identified as one of the three primitive translation vectors, is

$$\mathbf{g}_{h00} = h\mathbf{a}^*, \qquad (h = 1, 2, 3, \ldots) \tag{3.22a}$$

We can similarly write vectors along the other two primitive vectors $\mathbf{b}^*$ and $\mathbf{c}^*$ as

$$\mathbf{g}_{0k0} = k\mathbf{b}^*, \qquad (k = 1, 2, 3, \ldots) \tag{3.22b}$$

and

$$\mathbf{g}_{00l} = l\mathbf{c}^*, \qquad (l = 1, 2, 3, \ldots) \tag{3.22c}$$

Following this procedure for all the planes of a crystal lattice, its reciprocal lattice may be mapped. The general reciprocal lattice vector should read as

$$\mathbf{g} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* \tag{3.23}$$

3.5 RELATIONSHIPS BETWEEN a, b, c AND a*, b*, c*

Consider a direct crystal lattice as shown in Fig. 3.9. Draw a normal *OA* from *O* on the face opposite to *bc* face.

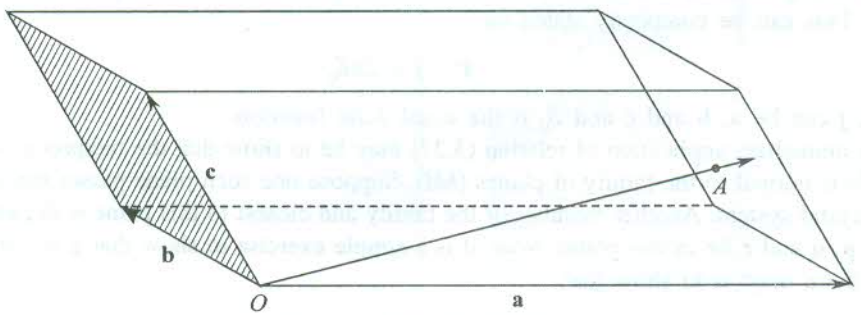


FIG. 3.9 A general direct crystal lattice.

Volume of the unit cell, $V = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$

$$= OA \times (\text{area of the face bc})$$

$$= OA \times |\mathbf{S}| \tag{3.24}$$

Therefore,

$$\frac{1}{OA} = \frac{|\mathbf{S}|}{V}$$

or

$$\frac{1}{d_{100}} = \frac{|\mathbf{S}|}{V} \qquad (\because OA = d_{100})$$

Thus

$$\frac{\hat{\mathbf{e}}}{d_{100}} = \frac{\mathbf{S}}{V}$$

where $\hat{\mathbf{e}}$ is the unit vector in the direction of the normal OA. Therefore, we have

$$\frac{2\pi\hat{\mathbf{e}}}{d_{100}} = \frac{2\pi\mathbf{S}}{V} = 2\pi \frac{\mathbf{b} \times \mathbf{c}}{\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})} \tag{3.25}$$

Comparing (3.25) with (3.21), we get

$$\mathbf{a}^* = 2\pi \frac{\mathbf{b} \times \mathbf{c}}{\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})} \tag{3.26a}$$

Similarly, we can show that

$$\mathbf{b}^* = 2\pi \frac{\mathbf{c} \times \mathbf{a}}{\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})} \tag{3.26b}$$

and

$$\mathbf{c}^* = 2\pi \frac{\mathbf{a} \times \mathbf{b}}{\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})} \tag{3.26c}$$

It is interesting to observe from (3.26) that each of $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$ is orthogonal to two crystal axis vectors. This can be compactly stated as

$$\mathbf{i}^* \cdot \mathbf{j} = 2\pi\delta_{ij} \tag{3.27}$$

where $\mathbf{i}$, $\mathbf{j}$ can be $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ and δ_{ij} is the usual delta function.

The immediate application of relation (3.27) may be to show that the reciprocal vector $\mathbf{g}$ given by (3.23) is normal to the family of planes (*hkl*). Suppose one such plane passes through the origin O in a crystal system. Another member of the family and closest to this plane is drawn in Fig. 3.10. Vectors $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{r}$ lie in this plane. Now, it is a simple exercise to show that $\mathbf{g}$ is normal to planes (*hkl*). All we need is to show that

$$\mathbf{g} \cdot \mathbf{t} = 0 \quad \text{with } \mathbf{t} \equiv \mathbf{p}, \mathbf{q}, \mathbf{r}.$$

From the definition of the Miller indices, it follows that the intercepts of the plane, shown in Fig. 3.10, with **a**, **b** and **c** axis vectors of the direct crystal are vectorially written as **a**/h, **b**/k, and **c**/l, respectively. This enables us to write

$$\mathbf{p} = \left(\frac{\mathbf{a}}{h} - \frac{\mathbf{b}}{k}\right); \quad \mathbf{q} = \left(\frac{\mathbf{c}}{l} - \frac{\mathbf{b}}{k}\right) \quad \text{and} \quad \mathbf{r} = \left(\frac{\mathbf{c}}{l} - \frac{\mathbf{a}}{h}\right)$$

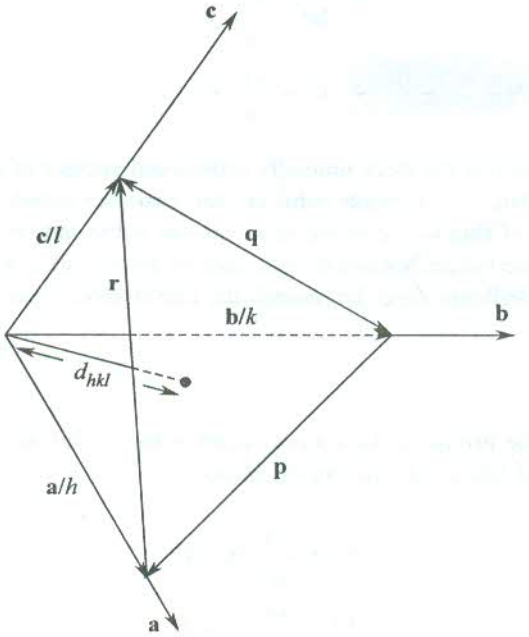


FIG. 3.10 The figure shows intersections of a plane (hkl) with crystal axes **a**, **b**, **c** at general orientations. The three intercepts measure **a**/h, **b**/k and **c**/l, respectively. Vectors **p**, **q** and **r** lie in the plane (hkl).

It can be checked that the scalar product of any of these vectors with **g** vanishes, proving thereby that every reciprocal vector is normal to a family of parallel planes with common interplanar spacing.

One point of caution: h, k and l in (3.23) need not always represent the Miller indices and in principle they can have a common factor. In view of this fact it is safer to use the form given by (3.7), i.e.

$$\mathbf{g} = m_1\mathbf{a}^* + m_2\mathbf{b}^* + m_3\mathbf{c}^*$$

where m_1, m_2 and m_3 are any integers.

3.6 APPLICATION TO SOME CRYSTAL LATTICES

We will now apply the relationships obtained above to determine the primitive translation vectors of lattices, reciprocal to certain crystal lattices.

Simple cubic (SC) lattice

For this lattice, $|\mathbf{a}| = |\mathbf{b}| = |\mathbf{c}|$ and **a**, **b** and **c** are mutually orthogonal vectors given by $a\hat{\mathbf{x}}$, $a\hat{\mathbf{y}}$ and $a\hat{\mathbf{z}}$,

respectively, where $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$ are unit vectors along the three orthogonal coordinate axes x, y, and z directed along the three edges of the lattice.

Therefore, using (3.26)

$$\begin{aligned} \mathbf{a}^* &= \frac{2\pi}{a}\,\hat{\mathbf{x}} \\ \mathbf{b}^* &= \frac{2\pi}{a}\,\hat{\mathbf{y}} \\ \mathbf{c}^* &= \frac{2\pi}{a}\,\hat{\mathbf{z}} \end{aligned} \tag{3.28}$$

The set (3.28) again represents the three mutually orthogonal vectors of equal magnitude $2\pi/a$. This shows that the reciprocal lattice of a simple cubic crystal is another simple cubic lattice. We can easily show that the reciprocal of this lattice is the same simple cubic lattice of the direct space.

The six planes, that are the perpendicular bisectors of the vectors $\pm\,\mathbf{a}^*$, $\pm\,\mathbf{b}^*$ and $\pm\,\mathbf{c}^*$, constitute the boundary of the first Brillouin zone. To remind, the first Brillouin zone serves as the primitive cell of the reciprocal lattice.

BCC lattice

Using (1.4a) (that give the primitive translation vectors) and (3.26) we get the following primitive vectors for the reciprocal lattice of this crystal lattice.

$$\begin{aligned} \mathbf{a}^* &= \frac{2\pi}{a}\,(\hat{\mathbf{x}} + \hat{\mathbf{y}}) \\ \mathbf{b}^* &= \frac{2\pi}{a}\,(\hat{\mathbf{y}} + \hat{\mathbf{z}}) \\ \mathbf{c}^* &= \frac{2\pi}{a}\,(\hat{\mathbf{z}} + \hat{\mathbf{x}}) \end{aligned} \tag{3.29}$$

These vectors represent an FCC lattice of lattice constant $4\pi/a$ [see (1.4b)], showing that an FCC lattice is the reciprocal lattice of the BCC lattice. The rhombohedron formed by $\mathbf{a}^*$, $\mathbf{b}^*$, $\mathbf{c}^*$ represents the primitive cell of volume V expressed as

$$\begin{aligned} V &= \mathbf{a}^* \cdot (\mathbf{b}^* \times \mathbf{c}^*) \\ &= \frac{16\pi^3}{a^3} \end{aligned}$$

Similarly, it can be shown that the reciprocal lattice of the FCC lattice is a BCC lattice. There are in all 12 shortest vectors for the lattice described by (3.29):

$$\frac{2\pi}{a}\,(\pm\hat{\mathbf{y}} \pm \hat{\mathbf{z}})\,; \quad \frac{2\pi}{a}\,(\pm\hat{\mathbf{z}} \pm \hat{\mathbf{x}})\,; \quad \frac{2\pi}{a}\,(\pm\hat{\mathbf{x}} \pm \hat{\mathbf{y}})$$

with the choices of signs being independent.

Planes perpendicular to these vectors at their mid-points enclose the volume of the first Brillouin zone which turns out to be a rhombododecahedron (Fig. 3.11).

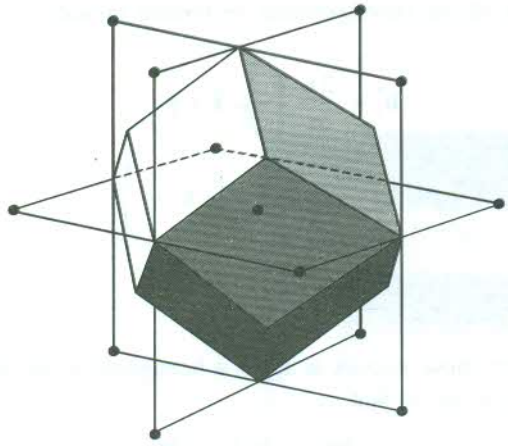


FIG. 3.11 The first Brillouin zone of a BCC crystal. It is rhombododecahedral in shape.

Hexagonal space lattice

The primitive translation vectors $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3$ are drawn in Fig. 3.12. In relation to the directions of x, y, z axes as shown in the figure, the primitive vectors of the direct lattice are given by

$$\mathbf{b}_1 = \frac{\sqrt{3}\,a}{2}\,\hat{\mathbf{x}} + \frac{a}{2}\,\hat{\mathbf{y}}$$

$$\mathbf{b}_2 = -\frac{\sqrt{3}\,a}{2}\,\hat{\mathbf{x}} + \frac{a}{2}\,\hat{\mathbf{y}} \tag{3.30}$$

$$\mathbf{b}_3 = c\hat{\mathbf{z}}$$

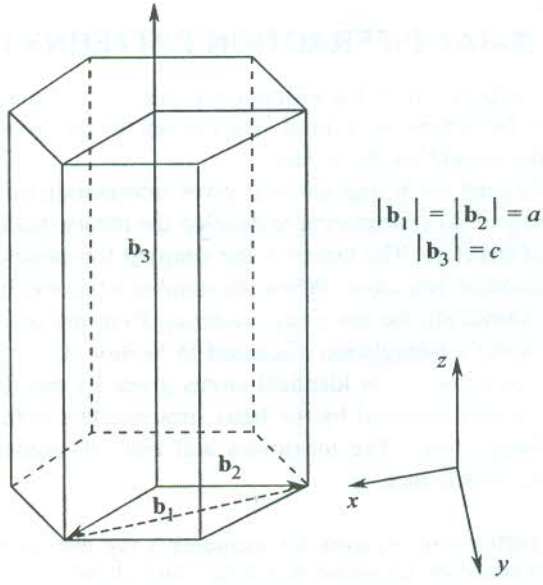


FIG. 3.12 Hexagonal space lattice—$\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3$ are the primitive vectors and the angle between $\mathbf{b}_1$ and $\mathbf{b}_2$ (on the hexogonal face) is 120°.

Then the primitive vectors of the corresponding reciprocal lattice are

$$\mathbf{b}_1^* = \frac{2\pi}{a}\left(\frac{1}{\sqrt{3}}\hat{\mathbf{x}} + \hat{\mathbf{y}}\right)$$
$$\mathbf{b}_2^* = \frac{2\pi}{a}\left(-\frac{1}{\sqrt{3}}\hat{\mathbf{x}} + \hat{\mathbf{y}}\right) \tag{3.31}$$
$$\mathbf{b}_3^* = \frac{2\pi}{c}\hat{\mathbf{z}}$$

The unit cell formed by these vectors is again a hexagonal prism as shown by Fig. 3.13. The significant point about this prism is that

$$|\mathbf{b}_3^*| < |\mathbf{b}_1^*| \text{ or } |\mathbf{b}_2^*|$$

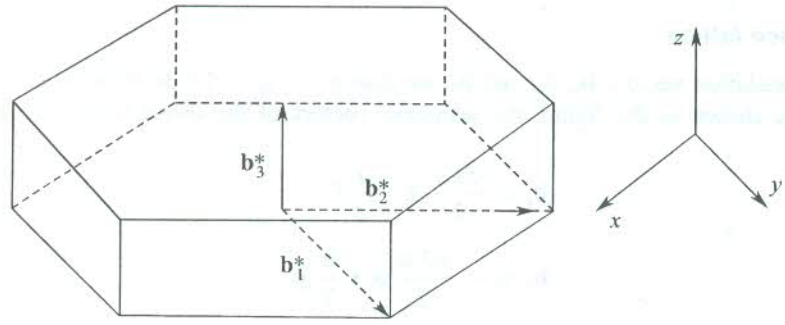


FIG. 3.13 The reciprocal lattice of the hexagonal space lattice shown in Fig. 3.12.

3.7 ANALYSIS OF X-RAY DIFFRACTION PATTERNS FROM CRYSTALS

The determination of crystal structure from the diffraction pattern of a crystal is a lengthy geometrical and mathematical exercise. Nevertheless, a brief description of the involved basic principles is provided in this book for the benefit of the reader.

We have to look well beyond the Bragg law that gives information on interplanar spacing. We capitalize on the Laue's theory of x-ray scattering to develop the theory that is capable of giving most of the information on crystal structure. The theory is the simplest for monatomic crystals (composed of identical atoms) with a basis of one atom. When the number of atoms in a primitive cell (basis) is more than one (all being identical), the net x-ray scattering from the cell can be treated easily by extending only slightly the Laue's formulation discussed in Section 3.3.

Let a primitive cell be composed of n identical atoms given by position vectors $\mathbf{r}_1$, $\mathbf{r}_2$, ..., $\mathbf{r}_n$. The extent to which the radiations scattered by the basis sites interfere with one another determines the intensity of a certain Bragg line.† The maximum and zero of intensity occurs for complete constructive and destructive interference.

† In a modern instrument (Diffractometer) used for recording x-ray diffraction pattern of crystals the photographic recording is replaced by electronic recording which shows lines (peaks) when constructive interference takes place. See Fig. 3.15.

If the change of wavevector attributed to certain peak is

$$\mathbf{k}' - \mathbf{k} = \mathbf{g}$$

we know that the corresponding path difference between the x-rays scattered by atoms at $\mathbf{r}_i$ and $\mathbf{r}_j$ will be $\mathbf{g} \cdot (\mathbf{r}_i - \mathbf{r}_j)$ and the phase factor representing the phase difference between two rays equals $\exp [i\mathbf{g} \cdot (\mathbf{r}_i - \mathbf{r}_j)]$.

As such, the phases of rays scattered by atoms at $\mathbf{r}_1$, $\mathbf{r}_2$, ..., $\mathbf{r}_n$ are in the ratios

$$e^{i\mathbf{g}\cdot\mathbf{r}_1}, e^{i\mathbf{g}\cdot\mathbf{r}_2}, \ldots, e^{i\mathbf{g}\cdot\mathbf{r}_n}$$

This suggests that the amplitude of the x-ray beam scattered by the whole primitive cell should have a factor expressible as

$$\sum_{j=1}^{n} \exp(i\mathbf{g} \cdot \mathbf{r}_j)$$

The amplitude of x-rays scattered by a single primitive cell is conventionally expressed as

$$S_{\mathbf{g}} = \sum_{j=1}^{n} f_j \exp(i\mathbf{g} \cdot \mathbf{r}_j) \tag{3.32}$$

where $S_{\mathbf{g}}$ is called the structure factor and f_j stands for the atomic form factor of the atom positioned at $\mathbf{r}_j$. The atomic form factor of an atom depends entirely on the internal structure of the atom. As in the present case, the primitive cell is composed of identical atoms, we can rewrite (3.32) as

$$S_{\mathbf{g}} = f \sum_{j=1}^{n} \exp(i\mathbf{g} \cdot \mathbf{r}_j) \tag{3.33}$$

If there are N primitive cells in a crystal, the total scattering amplitude of a diffraction peak may be expressed as

$$S = NS_{\mathbf{g}} \tag{3.34}$$

This Bragg peak appears because of the constructive interference of the scattered x-rays whose wavevector differs from that of the incident x-rays by reciprocal vector $\mathbf{g}$. Since intensity is known to be proportional to the square of amplitude, the intensity of a Bragg line is proportional to $|S_{\mathbf{g}}|^2$. Thus, the relative intensity of Bragg lines can be obtained by calculating their $S_{\mathbf{g}}$.

Structure factor of a BCC crystal

Its conventional unit cell is non-primitive. But it can be pictured as a simple cubic lattice with a basis of two atoms at 000 and $\frac{1}{2}\frac{1}{2}\frac{1}{2}$. Such a lattice is primitive and we can use (3.33) for calculating the structure factor, $S_{\mathbf{g}}$. Thus $\mathbf{r}_1 = 0$ and $\mathbf{r}_2 = \frac{1}{2} a(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})$.

From (3.33), we have

$$S_{\mathbf{g}} = f + f \exp\left[i\mathbf{g} \cdot \frac{a}{2} (\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})\right] \tag{3.35a}$$

$$= f\{1 + \exp[i\pi(m_1 + m_2 + m_3)]\}$$

$$= f[1 + (-1)^{m_1 + m_2 + m_3}] \tag{3.35b}$$

because for a simple cubic lattice, we have

$$\mathbf{g} = \frac{2\pi}{a}(m_1\hat{\mathbf{x}} + m_2\hat{\mathbf{y}} + m_3\hat{\mathbf{z}}) \tag{3.36}$$

Equation (3.35b) shows that

$$S_{\mathbf{g}} = 0, \quad \text{if } m_1 + m_2 + m_3 = \text{an odd integer} \tag{3.37a}$$

$$= 2f, \quad \text{if } m_1 + m_2 + m_3 = \text{an even integer} \tag{3.37b}$$

It is important to note that $\mathbf{g} \neq 0$ though the respective $S_{\mathbf{g}}$ vanishes for an odd integral value of $(m_1 + m_2 + m_3)$. There cannot occur Bragg peaks for planes such as (100), (111), (120), etc. but peaks caused by reflections from planes, for example, (110), (112), (222) will appear in the diffraction pattern.

For example, in sodium metal, which has a BCC structure, the reflections from (100) plane are missing in the diffraction pattern as required by (3.37a). In a cubic crystal the phases $(\mathbf{g} \cdot \mathbf{r}_j)$ for atoms on the bounding (100) planes are 0 and 2π. For this phase difference, normally the (100) diffraction peak appears. But in a BCC crystal there is a plane midway between (100) planes. The phase difference between rays scattered from an atom on this plane and another atom on any of the two nearest (100) planes thus becomes π. As the density of atoms on the (100) and midway planes is equal, a destructive interference takes place between the rays reflected from successive (100) and midway planes. This explains why (100) reflections are not observed from sodium crystals. In fact, the destructive interference in this case is complete because of a common atomic form factor for all atoms present in the crystal.

Finally, we must recall that we have treated a BCC lattice as a simple cubic (SC) with a basis of two identical atoms. The reciprocal lattice of an SC lattice is again an SC lattice with lattice constant $2\pi/a$. But in accordance with (3.37a), a lattice point the sum of whose coordinates in the reciprocal lattice is an odd number should be missing in the reciprocal lattice. Every lattice point corresponds to the reflection from a certain family of parallel planes. This means that the alternate corners of a cube of side $2\pi/a$ are non-existent. Therefore, this cube cannot be used to map the infinite reciprocal lattice and therefore, it is not the required unit cell. Interestingly, on choosing the cube of side $4\pi/a$, the condition (3.37a) allows all of its corners to be present along with a point at the centre of each bounding face. This arrangement of lattice points belongs to an FCC lattice, which effectively shows that the reciprocal lattice of a BCC lattice is an FCC lattice. This result has been proved earlier in Section 3.6. The conventional unit cell of an FCC lattice in reciprocal space is shown in Fig. 3.14.

Structure factor of a monatomic FCC crystal

The unit cell in this case has four identical atoms located at its corners and at the centres of its faces. This FCC cell can as well be interpreted as SC structure whose basis is composed of four atoms. The positions of basis atoms in the SC primitive cell are: (0 0 0), (1/2 1/2 0), (1/2 0 1/2) and (0 1/2 1/2). Their position vectors as measured from a corner of the cell will be written as:

$$\mathbf{r}_1 = 0, \mathbf{r}_2 = \frac{1}{2}a(\hat{\mathbf{x}} + \hat{\mathbf{y}}); \mathbf{r}_3 = \frac{1}{2}a(\hat{\mathbf{z}} + \hat{\mathbf{x}}) \quad \text{and} \quad \mathbf{r}_4 = \frac{1}{2}a(\hat{\mathbf{y}} + \hat{\mathbf{z}})$$

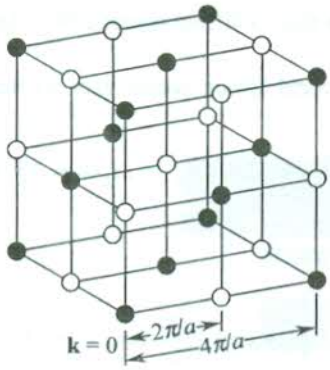


FIG. 3.14 A cubic reciprocal lattice of side $2\pi/a$. The white circles represent the lattice points for which the structure factor vanishes (3.37a). When these circles are removed from the lattice, a conventional FCC lattice of side $4\pi/a$ is obtained.

Substituting these vectors and the general reciprocal lattice vector **g** (3.36) in (3.33), one gets the following expression for the structure factor.

$$S_g = f\left[1 + e^{i\pi(m_1 + m_2)} + e^{i\pi(m_3 + m_1)} + e^{i\pi(m_2 + m_3)}\right] \tag{3.38}$$

which gives

$$S_g = 4f \text{ , only if } m_1, m_2 \text{ and } m_3 \text{ are all odd or all even} \tag{3.39a}$$

$$= 0, \text{ for all odd-even combinations of } m_1, m_2 \text{ and } m_3. \tag{3.39b}$$

The result (3.39) reveals that the peaks having origin in the diffraction from planes such as (100), (110), (121) will not be observed, whereas those from the likes of (111), (200), (222) will appear in the diffraction pattern of an FCC crystal.

Structure factor of monatomic diamond lattice

This lattice represents the crystals of diamond, germanium, silicon, grey tin, etc. Each of these is composed of identical atoms and their unit cell is interpreted as two FCC lattices displaced from one another along a body diagonal by one-fourth of the body diagonal. But it can still be treated as an FCC lattice with a basis of two identical atoms whose positions $\mathbf{r}_j$ may be represented by

$$\mathbf{r}_1 = 0 \quad \text{and} \quad \mathbf{r}_2 = \frac{1}{4}\, a(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}}) \qquad (j = 1, 2)$$

Putting direct primitive vectors from (1.4b) in set (3.26), we get the following primitive vectors of the reciprocal lattice:

$$\begin{aligned} \mathbf{a}^* &= \frac{2\pi}{a}\,(\hat{\mathbf{x}} + \hat{\mathbf{y}} - \hat{\mathbf{z}}) \\ \mathbf{b}^* &= \frac{2\pi}{a}\,(-\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}}) \\ \mathbf{c}^* &= \frac{2\pi}{a}\,(\hat{\mathbf{x}} - \hat{\mathbf{y}} + \hat{\mathbf{z}}) \end{aligned} \tag{3.40}$$

A comparison with (1.4a) shows that (3.40) represent a BCC lattice the side of whose conventional cell is $4\pi/a$.

Using (3.40), the general reciprocal lattice vector may written as

$$\mathbf{g} = \frac{2\pi}{a}\left[(-m_1 + m_2 + m_3)\,\hat{\mathbf{x}} + (m_1 - m_2 + m_3)\,\hat{\mathbf{y}} + (m_1 + m_2 - m_3)\,\hat{\mathbf{z}}\right] \tag{3.41}$$

Substitution of $\mathbf{g}$ and $\mathbf{r}_j$ in (3.33) yields

$$S_{\mathbf{g}} = f\left\{1 + \exp\left[\frac{i\pi}{2}(m_1 + m_2 + m_3)\right]\right\} \tag{3.42}$$

which gives

$$S_{\mathbf{g}} = f(1 \pm i), \text{ if } (m_1 + m_2 + m_3) = \text{an odd number} \tag{3.43a}$$

$$= 0, \text{ if } (m_1 + m_2 + m_3) = \text{twice an odd number} \tag{3.43b}$$

$$= 2f, \text{ if } (m_1 + m_2 + m_3) = \text{twice an even number} \tag{3.43c}$$

On lines similar to those followed for a BCC lattice, it can be shown that the reciprocal lattice of a diamond lattice is a BCC lattice with cubic cell of side $8\pi/a$. There will be three kinds of reciprocal lattice points in this case corresponding to the three values of the structure factor, $f(1 \pm i)$, $2f$ and 0 defined by (3.43). Details can be found elsewhere.*

Structure factor of a polyatomic crystal

Let us take the general case of a polyatomic crystal in which atoms in the basis are not identical, i.e. the crystal is composed of atoms belonging to more than one element. The structure factor of any diffraction line from this crystal is calculated by using the general relation given by (3.32).

As an example, we take the simplest case of an alloy. The binary CuZn alloy, when ordered, has CsCl structure. By ordered we mean that all nearest neighbours of a Cu atom are Zn atoms and vice versa. The lattice is simple cubic and the basis has two atoms; Cu atom at 000 and Zn atom at $\frac{1}{2}\frac{1}{2}\frac{1}{2}$. The structure factor is

$$S_{\mathbf{g}} = f_{\text{Cu}} + f_{\text{Zn}} \exp\left[i\frac{2\pi}{a}(m_1\hat{\mathbf{x}} + m_2\hat{\mathbf{y}} + m_3\hat{\mathbf{z}}) \cdot \frac{a}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}} + \hat{\mathbf{z}})\right]$$

$$= f_{\text{Cu}} + f_{\text{Zn}} \exp\left[i\pi(m_1 + m_2 + m_3)\right] \tag{3.44}$$

In principle, $S_{\mathbf{g}}$ can never vanish since $f_{\text{Cu}} \neq f_{\text{Zn}}$ and, therefore, all diffraction lines of the simple cubic lattice should be observed. But in practice every alloy is disordered to some extent. In the disordered structure there is an equal probability for a Cu atom and a Zn atom to be located at 000 or $\frac{1}{2}\frac{1}{2}\frac{1}{2}$. The relation (3.44) is found to be of tremendous value in such a situation as it can help in distinguishing between an ordered and a disordered structure. Therefore, it is prudential to talk in terms of average structure factor while discussing a disordered structure. In this case,

---

* N.W. Ashcroft and N.D. Mermin, *Solid State Physics,* Chapter 6, pp. 106–107, Saunders College (1988).

$$\langle S_g \rangle = \langle f \rangle + \langle f \rangle \exp\left[i\pi(m_1 + m_2 + m_3)\right] \tag{3.45}$$

where

$$\langle f \rangle = \frac{1}{2}\left[f_{Cu} + f_{Zn}\right]$$

From (3.45) we find that the reflections, for which $(m_1 + m_2 + m_3)$ is odd, vanish. It means that the disordered alloy will show a smaller number of diffraction lines and this fact can be used to make a distinction between an ordered and a disordered crystal in general.

Another interesting example is that of the KCl crystal. The KCl is an ionic crystal and its stable constituents are K^+ and Cl^- ions both of which possess an equal number of electrons. This prompts us to expect that

$$f_{K^+} \simeq f_{Cl^-}$$

The diffraction pattern really confirms our conjecture as it suggests that KCl is a monatomic simple cubic lattice of lattice constant $a/2$. Based on a cubic lattice of lattice constant a, we find that lines of only those planes are observed for which the sum of the plane indices $(m_1 + m_2 + m_3)$ is an even number (Fig. 3.15). The result is of special significance as the basis is not composed of identical atoms in this case. In addition, the dependence of the atomic form factor on the internal structure of an atom/ion is established.

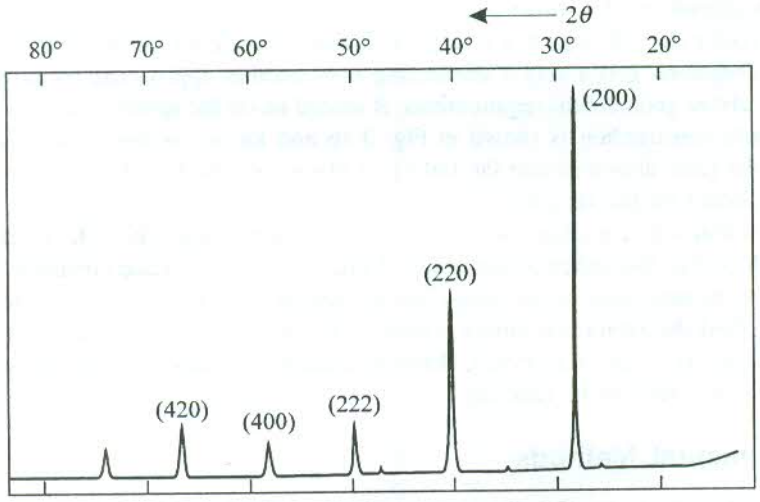


FIG. 3.15 *x*-ray diffraction from KCl crystal. The ions K^+ and Cl^- have an equal number of electrons characterizing them with a common form factor. This results in the vanishing of reflections from planes for which the sum $(m_1 + m_2 + m_3)$ is an odd number. Reflections from only those planes for which this sum is an even number are present in the diffractogram.

3.8 MEASUREMENT OF DIFFRACTION PATTERN OF CRYSTALS

Geometries of various experimental arrangements employed for this purpose are linked with the Laue condition of diffraction. A diffraction line appears only if the tip of the incident wavevector (**k**) lies on a Bragg plane in *k*-space. The Bragg plane is only the family of parallel planes which meets the

requirements of constructive interference for reflections of the incident radiation **k**. For this family of planes the direction of **k** is also specific in order to observe the diffraction peak. One diffraction line or one lattice point in the reciprocal space corresponds to one family of planes. Obviously, there are many other families of planes composing the crystal. The whole reciprocal lattice or the diffraction pattern can be completed by making the constructive interference possible for reflections from all of these planes. In practice, it is done by having provisions for:

(i) Varying the direction of **k** which can be done by rotating the crystal.

(ii) A continuous x-ray beam whose varying set of wavelengths may find appropriate allies among various crystal planes in order to satisfy the Laue condition for their direction of incidence.

The provisions mentioned above give rise to a number of experimental geometries which we will describe later in this section. But, before doing so, we will discuss one of the most beautiful geometrical constructions known as the Ewald construction.

3.8.1 The Ewald Construction

Ewald gave the following simple geometrical construction that guided experimentalists to develop various methods for deducing the crystal structure from the observed diffraction pattern.

Draw a sphere of radius k in the reciprocal space with centre at A (Fig. 3.16). Let **k** represent the wavevector of the incident x-ray beam terminating at the origin of the reciprocal lattice O. The direction and magnitude of the wavevector of reflected x-rays (say given by **k′**) will be such that change in wavevector (**k′** – **k**) equals a reciprocal lattice vector **g** on the demand of Laue condition for diffraction. Therefore, **g** is a vector connecting O to another appropriate reciprocal lattice point B. To satisfy the above geometrical requirements, B should lie on the sphere's surface and **k′** be given by **AB**. The whole construction is shown in Fig. 3.16 and known as the Ewald construction. The equidistant parallel lines drawn denote the family of planes involved in the reflection shown. These planes are represented by the reciprocal lattice point B.

The Ewald construction implies the elastic x-ray scattering, since $|\mathbf{k}| = |\mathbf{k}'|$ (= the radius of the Ewald sphere). With this, the vector relationships of Fig. 3.16 give the Laue condition $2\mathbf{k} \cdot \mathbf{g} = \mathbf{g}^2$ that has been shown to be equivalent to the Bragg law in Section 3.3. The construction clearly establishes how rare it is to find the reciprocal lattice points on the surface of the Ewald sphere to get Bragg reflections from various planes of a crystal. Different methods are used to create situations that render the recording of these reflections feasible.

3.8.2 Experimental Methods

In this section we describe in brief the three main techniques employed for the determination of crystal structure.

The Laue method

This method produced the first ever recorded x-ray diffraction photograph of a crystal. The Laue photograph is historic in this sense. The Laue condition given by (3.15) can be written as

$$\Delta\mathbf{k} = \mathbf{g} \quad . \tag{3.46}$$

where **g** is defined by (3.7).

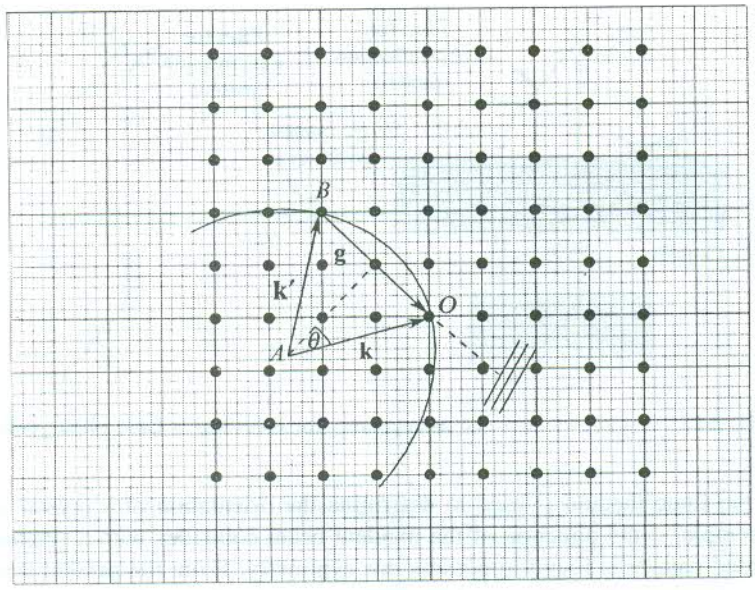


FIG. 3.16 Ewald construction for Bragg reflections. The change in wavevector on reflection is equal to a reciprocal lattice vector **g**. The **g** terminates at a reciprocal lattice point falling on the Ewald sphere.

Taking the scalar product of direct lattice primitive vectors **a, b, c** with both sides of (3.46), we have

$$\begin{aligned} \mathbf{a} \cdot \Delta\mathbf{k} &= 2\pi m_1 \\ \mathbf{b} \cdot \Delta\mathbf{k} &= 2\pi m_2 \\ \mathbf{c} \cdot \Delta\mathbf{k} &= 2\pi m_3 \end{aligned} \tag{3.47}$$

Relations (3.47), called the Laue equations, indicate that $\Delta\mathbf{k}$ lies on the surface of each of the three separate cones formed about **a**, **b** and **c**. This implies that the three cones intersect each other along a common line along $\Delta\mathbf{k}$. It is in fact a modified form of the Laue condition which can in practice be satisfied by using the x-ray beam with a wide range of wavelength and sweeping through crystal orientations. The condition getting fulfilled by sheer accident, however, cannot form the basis for any experiment. Laue's technique does make use of a continuous source of x-rays which simplifies the problem to a large extent. If the wavelength ranges from λ_1 to λ_2, the reciprocal points that may be searched will be in the region between the spheres of radii $\mathbf{k}_1 \left(= \frac{2\pi}{\lambda_1}\hat{\mathbf{n}}\right)$ and $\mathbf{k}_2 \left(= \frac{2\pi}{\lambda_2}\hat{\mathbf{n}}\right)$, respectively. The larger the difference, $(\lambda_1 - \lambda_2)$, the larger the region into which the Ewald sphere expands.

A schematic illustration of the Laue's set-up or Laue camera is made in Fig. 3.17. A well-collimated beam of x-rays is obtained with the help of a pinhole arrangement. A typical range of the source wavelength is from 0.2 Å to 2 Å. There is no critical limitation on crystal dimensions. Two flat films are used; one to record the diffraction of reflected x-rays and the other to record the pattern produced by transmitted x-rays. The diffraction pattern is an array of spots as shown in Fig. 3.1 for a CdTe crystal.

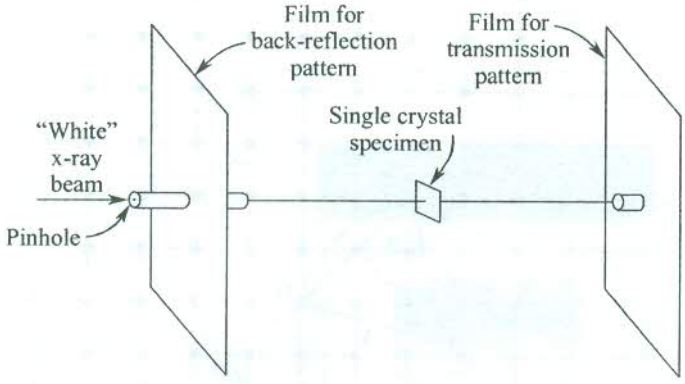


FIG. 3.17 A Laue camera: the arrangement for recording a Laue photograph.

In practice, the Laue method is used to determine the orientation of a crystal whose structure is known. For example, if the direction of the incident beam coincides with a symmetry axis of the crystal, the recorded pattern is found to possess this symmetry. From the present status of the Laue technique we must not infer that the crystal structure cannot be deduced from the Laue photograph. This method is not used simply because it is tedious involving the complicated construction of several geometrical projections.

The oscillation (rotation) method

In this method monochromatic x-rays (λ kept at a constant value) are incident on a crystal rotating about an axis. The rotation of the crystal enables the variation of the angle of incidence with the Bragg plane. Whenever a set of λ, θ and d satisfies the Bragg condition (3.9), the corresponding Bragg peak appears. In a complete rotation some of the planes are repeated and thus overlapping of spots occurs. Therefore, the crystal is actually oscillated through a limited angular range to reduce overlapping. A crystal-rotating camera is shown in Fig. 3.18.

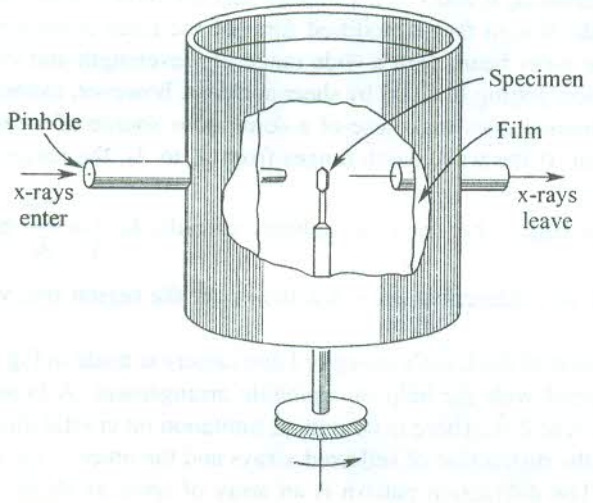


FIG. 3.18 A camera used in the oscillation method.

The basis of diffraction spots on the film is excellently explained in the Ewald picture. On fixing the direction of the monochromatic x-ray (i.e. **k**), the location of the Ewald sphere in *k*-space gets fixed. The rotation of a crystal is equivalent to the rotation of the corresponding reciprocal lattice. So while the reciprocal lattice is under rotation, every reciprocal lattice point generates a circular path about the rotation axis. As soon as this circle intersects the fixed Ewald sphere, a diffraction peak occurs.

This method is in extensive use as it is relatively fast. But a noteworthy drawback in this method is that reflections corresponding to a three-dimensional lattice net are recorded on a two-dimensional photographic film, recording thereby a two-dimensional network of the reciprocal lattice. The problem is handled well in moving film methods. One such method, known as Weissenberg method, is almost the most used technique today. For details the reader is advised to consult Buerger's book.*

The powder method

This technique in principle provides for the variation of the axis of rotation over all possible orientations of a rotating crystal. This is practically achieved by using a powdered sample instead of a single piece of crystal. The powdered sample is filled in a glass capillary which is rotated about its length. Particles in the powder, though small, are yet substantially large on atomic scale and as such each of them can be treated as a tiny crystal. The capillary filled with the powdered sample serves as a collective unit of a large number of tiny crystals in all possible orientations. The rotation of this unit is equivalent to the rotations described in the beginning of this paragraph. The design of the Debye–Scherrer camera which is used in this method is explained in Fig. 3.19.

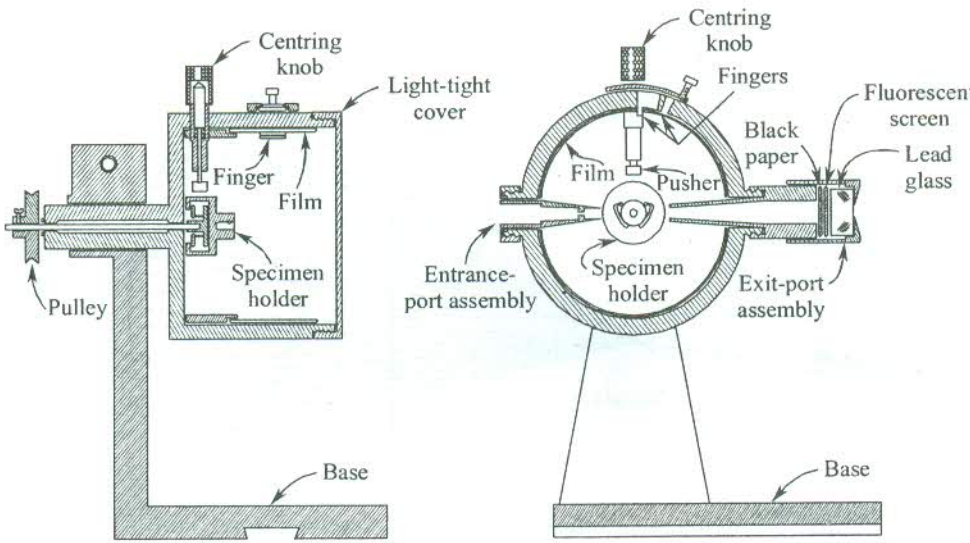


FIG. 3.19 Sectional views of a Debye–Scherrer camera used in the powder method.

This method is known to be the most fast and accurate for the determination of lattice constants. Its another important use is in the study of phase diagrams of alloy systems. Since lattice constants

* M.J. Buerger, *X-Ray Crystallography*, John Wiley & Sons (1942).

are unique for a material, this method is widely used for the identification of materials. Further, it is the only method which works for all crystalline materials. In view of this fact we have chosen this technique for a complete description of its use for determining the lattice constants.

The nature of the diffraction pattern produced by the present technique can be visualized more easily in the Bragg formulation. For the interpretation involving Ewald construction, the reader is advised to consult the book by Ashcroft and Mermin.*

As is evident from Fig. 3.20, the diffracted rays leave the sample along the generators of cones concentric with the incident beam. An angle of 2θ is formed between the generators and the direction of the incident beam where θ denotes the usual Bragg angle given by

$$\sin\theta = \left(\frac{n\lambda}{2d_{hkl}}\right) \tag{3.48}$$

There exists one cone for each solution of the Bragg equation (3.48). In a perfect experimental set-up, the number of diffracted rays is enormously large and they outline the curved surface of the cone very densely. Figure 3.20 shows that these rays cut the photographic film in a continuous circle provided they are sufficiently dense. All such circles each of which corresponds to a certain Bragg plane are truncated by the edges of the film because of its limited width (see Fig. 3.21). The figure shows two sets of concentric rings with centres on the film coinciding with the entrance and exit

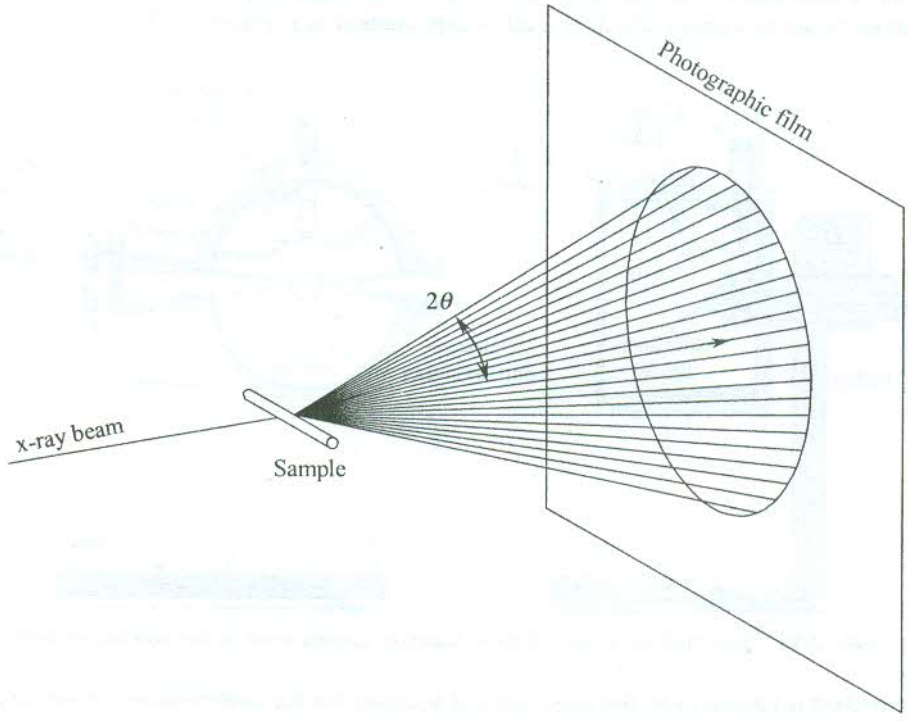


FIG. 3.20 The diffracted rays lie on the curved surface of a cone. The photographic film intercepts the cone such that the axis of the cone is perpendicular to the plane of the film.

* Ibid, pp. 102–104.

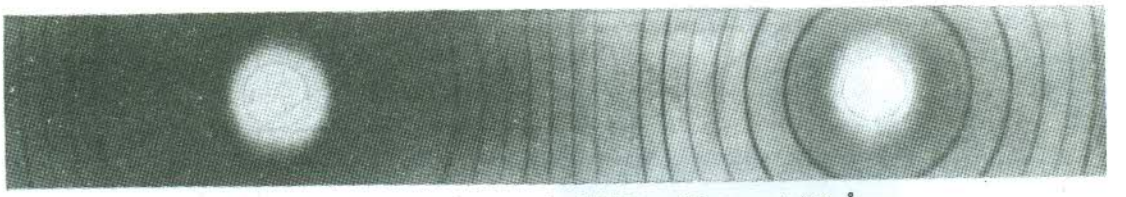


FIG. 3.21 Powder photograph of PbTe, cubic, $a = 6.439$ Å.

pinholes of the camera. The fainter set of rings represents back-reflections which occur for large values of θ. If S is the diameter of a certain circle (Fig. 3.21), then using Fig. 3.20 the θ for the respective diffraction line is expressed as

$$\theta(\text{radians}) = \frac{S}{4R} \tag{3.49}$$

where R is the radius of the camera. Using (3.48) and (3.49), the interplanar spacing d_{hkl} is calculated.

It is a matter of great significance to note that diffraction lines produced by back-reflections, though weak in intensity, yield more accurate results. On differentiating the Bragg equation (3.48) at constant wavelength (since monochromatic x-rays are used), we get

$$\frac{\Delta\theta}{\Delta d} = -\frac{\tan\theta}{d} \tag{3.50}$$

The relation (3.50) indicates that the change in θ for a given change in d is the largest as θ approaches 90°. Since reflections take place at 2θ from the direction of the incident beam, the highest sensitivity as defined by (3.50) refers to the back-reflections.

3.9 DETERMINATION OF LATTICE CONSTANTS

With knowledge of interplanar spacings (d_{hkl}), lattice constants of a crystal can be determined using the mathematical relation valid for the symmetry of the crystal. The derivation of the relationship between the unit cell dimensions and the interplanar spacing when the edges of cells make general angles with one another is a difficult exercise. Relations for crystals with well-known symmetries can be found in standard books on crystallography. Given below is the treatment for crystal systems with orthogonal axes. A unit cell with orthogonal edges, a, b, c, is shown in Fig. 3.22. A family of parallel planes (hkl) with a common interplanar spacing d_{hkl} intersects a, b, c axes into h, k, l parts, respectively. Supposing that the first plane of this family passes through the origin O and the next plane intersects the a, b, c axes at P, Q, R, respectively, and

$$OP = n_1 a; \qquad OQ = n_2 b; \qquad OR = n_3 c \tag{3.51}$$

Then,

$$h = \frac{1}{n_1}; \qquad k = \frac{1}{n_2}; \qquad l = \frac{1}{n_3} \tag{3.52}$$

Therefore,

$$OP = \frac{a}{h}; \qquad OQ = \frac{b}{k}; \qquad OR = \frac{c}{l} \tag{3.53}$$

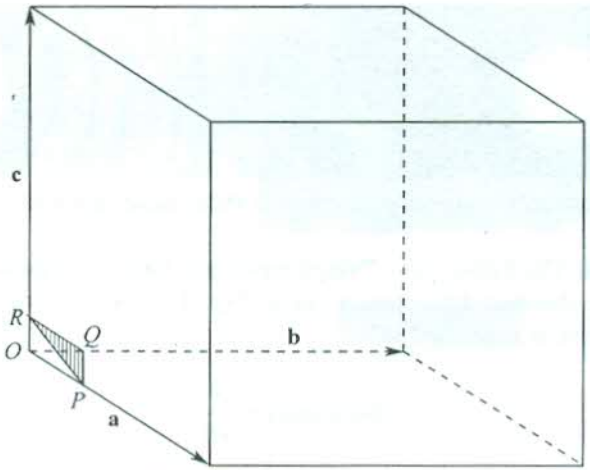


FIG. 3.22 A plane PQR of the family of planes (hkl) in a crystal with orthogonal axes.

Figure 3.23 gives an enlarged geometry of the region of the first two planes of the family (hkl). Let us perform a geometrical construction of drawing a plane perpendicular to the planes (hkl) such that it contains the c-axis. This plane intersects the planes (hkl) in the line RS, and the plane POQ in the line OS. The perpendicular OM dropped from O on the plane PQR obviously measures the interplanar spacing (i.e. $OM = d_{hkl}$). The plane POQ is drawn separately in Fig. 3.24. It should be noted that OS is perpendicular to PQ and triangles POQ and PSO are similar. Therefore,

$$\frac{OP}{OS} = \frac{PQ}{OQ} \tag{3.54}$$

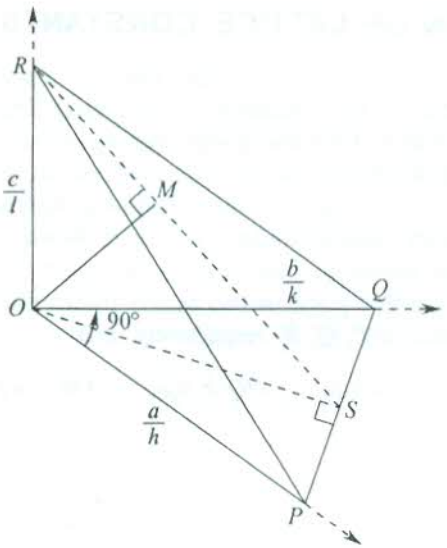


FIG. 3.23 An enlarged view of the region of the plane PQR. It is assumed that the first member of the family of planes (hkl) passes through the origin (O) and PQR is the second member of the family. The ROS represents an imaginary plane that is perpendicular to PQR and contains the c-axis. The perpendicular OM is equal to the interplanar spacing (d_{hkl}).

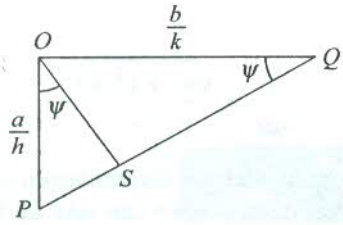


FIG. 3.24 The ΔPOQ is redrawn to show that Δs POQ and PSO are similar.

Placing the values of these lengths from (3.53), we have

$$\frac{1}{p^2} = \frac{h^2}{a^2} + \frac{k^2}{b^2} \tag{3.55}$$

where we have used $OS = p$.

Similarly, for the plane ROS shown in Fig. 3.25, OM is perpendicular to SR, and triangles SOR and SMO are similar. Hence,

$$\frac{OS}{OM} = \frac{SR}{OR} \tag{3.56}$$

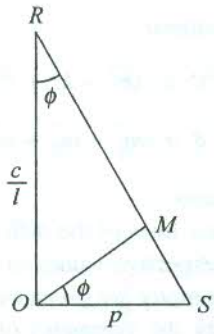


FIG. 3.25 The ΔROS is redrawn to show that Δs SOR and SMO are similar.

giving

$$\frac{1}{d_{hkl}} = \sqrt{\frac{1}{p^2} + \frac{l^2}{c^2}} \tag{3.57}$$

or

$$\frac{1}{d_{hkl}} = \sqrt{\frac{h^2}{a^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2}} \qquad \text{[using (3.55)]} \tag{3.58}$$

Equation (3.58) represents the relationship for an orthorhombic crystal.

For a tetragonal system ($a = b \neq c$):

$$\frac{1}{d_{hkl}} = \sqrt{\frac{h^2 + k^2}{a^2} + \frac{l^2}{c^2}} \tag{3.59}$$

For a cubic system ($a = b = c$):

$$\frac{1}{d_{hkl}} = \frac{\sqrt{h^2 + k^2 + l^2}}{a} \tag{3.60}$$

The simple form of (3.60) suggests that the determination of lattice constant of a cubic metal like copper can be one of the easiest exercises we can take up for learning the ideas discussed in this section.

Identification of cubic crystal structures

We have now developed sufficient background to enable us to identify a cubic crystal structure from the analysis of its diffraction pattern.

Combining (3.60) with (3.9) for the first order of diffraction,

$$\sin\theta = \frac{\lambda}{2a}\sqrt{h^2 + k^2 + l^2}$$

or

$$\sin^2\theta = \frac{\lambda^2}{4a^2}(h^2 + k^2 + l^2) \tag{3.61}$$

Therefore, for a certain diffraction pattern

$$\sin^2\theta \propto (h^2 + k^2 + l^2)$$

or

$$\sin^2\theta \propto (m_1^2 + m_2^2 + m_3^2) \tag{3.62}$$

where m_1, m_2 and m_3 have a common factor.

One can measure $\sin^2\theta$ for the first few lines of the diffraction pattern. According to (3.62), their values must be in the same ratio as the respective values of $(m_1^2 + m_2^2 + m_3^2)$. The values of m_1, m_2 and m_3 that are permitted for a crystal symmetry are given by the conditions for its non-zero structure factor derived in Section 3.7. In this way the symmetry of the crystal is determined.

SC crystal

All possible values of m_1, m_2 and m_3 allowed.

Ratio of $\sin^2\theta$ 1 : 2 : 3 : 4 : 5 : 6 : 7 ...

BCC crystal

$(m_1 + m_2 + m_3)$ = an odd integer, not allowed

$(m_1 + m_2 + m_3)$	*Planes*	$(m_1^2 + m_2^2 + m_3^2)$	$(\sin^2\theta/constant)$
2	(110)	2	1
	(200)	4	2
4	(112)	6	3
	(220)	8	4
	(130)	10	5
6	(222)	12	6
	(123)	14	7

FCC crystal

Values of m_1, m_2, m_3 : all odd or all even integers, allowed

$(m_1 + m_2 + m_3)$	*Planes*	$(m_1^2 + m_2^2 + m_3^2)$	$(sin^2\theta/constant)$
3	(111)	3	3
2	(200)	4	4
4	(220)	8	8
5	(131)	11	11
6	(222)	12	12
7	(133)	19	19

Diamond structure

$(m_1 + m_2 + m_3)$ = an odd integer or twice an even integer, allowed

$(m_1 + m_2 + m_3)$ = twice an odd integer, not allowed

$(m_1 + m_2 + m_3)$	*Planes*	$(m_1^2 + m_2^2 + m_3^2)$	$(sin^2\theta/constant)$
3	(111)	3	3
4	(220)	8	8
5	(113)	11	11
4	(400)	16	16
7	(133)	19	19

3.10 SELECTION OF INCIDENT BEAM

The diffraction phenomenon exhibited by crystals was presumed all through our discussion to have been produced using the incident beam of x-rays. But in practice it is not always so. The main requirement on incident beam (according to the Bragg law) is that the wavelength of radiation should be of the order of interatomic separation in the crystal or less than that. The x-rays are electromagnetic waves. In de Broglie picture, every moving particle is treated as a wave with a finite wavelength. The wavelength of a particle can be manipulated by varying its energy or momentum ($\lambda = h/p$), where p is the momentum. When this wavelength is of the order of a few Å, a beam of corresponding particles can be used to produce diffraction from crystals. Neutrons and electrons are two such particles whose beams have been widely used for the study of crystal structure. The purpose of this discussion is to emphasize that the use of one specific source for the study of a certain class of materials has an added advantage. We discuss below this aspect of compatibility with reference to x-rays, neutrons and electrons, separately.

X-rays

These are electromagnetic waves and their wavelength is expressed as $\lambda = hc/\varepsilon$, where ε is the photon energy. For λ to be of the order of a few angstrom units, the energy ε should be in the range from 10 to 50 keV which can be manipulated in an x-ray generator. The adjustment of wavelength of the outcoming beam is made by an appropriate choice of the target in the x-ray tube. For example, $K_{\alpha 1}$

line from copper target is at 1.541 Å and $K_{\alpha 1}$ from molybdenum is at 0.709 Å. The x-rays are the most commonly used source of radiation for all kinds of crystals. But all materials cannot be studied using a single target. For an accurate measurement, an appropriate target is selected on the basis of interatomic separation (guessed from rough data or some other measurement) in the crystal. Crystal monochromators are used to get discrete wavelengths.

Neutrons

A neutron beam has a distinct advantage over x-rays in studying the structure of magnetic materials. Neutron itself has a magnetic moment and, therefore, its scattering pattern carries information about the magnetic moment (in both magnitude and direction) of constituent magnetic atoms of the crystal under study. This happens because of the interaction between neutron and atomic magnetic moments, the details of which form the subject matter of the chapters on magnetism. The size of the unit cell of magnetically ordered materials turns out to be very different when determined by using x-rays and neutrons, separately. The reason will be discussed while dealing with ordered magnetic materials.

In the light of the de Broglie hypothesis on wave nature of particles duly confirmed by Davisson and Germer experiments, we express the wavelength of a neutron as

$$\frac{h}{p} = \frac{h}{\sqrt{2M_n\varepsilon}}$$

where M_n is the neutron mass (1.675×10^{-27} kg) and ε its energy.

It is simple to see that for $\lambda = 1$ Å,

$$\varepsilon \cong 0.08 \text{ eV}$$

This value is of the order of energy of slow neutrons produced in a nuclear reactor. It indicates that the extra neutrons of a reactor may be used for doing the diffraction experiment. This is really the case; the neutron diffraction experiment is arranged around a nuclear reactor where all precautions are taken in the most strict sense on account of neutrons being a potent source of health hazard. Neutrons can be used for all materials as in non-magnetic materials they interact with the nuclei of constituent atoms.

Electrons

The wave nature of particles was in fact established through experiments with electron beams. Thompson observed diffraction effects when he passed a narrow electron beam through crystalline films. Elastically scattered electrons have proved more suitable than x-rays for the study of surfaces because the penetrating power of electrons is limited by repulsion from electrons of the constituent atoms. Therefore, an electron beam is very frequently used to investigate the structure of thin crystals, layers, films, surfaces and gases in particular.

Substituting the electron mass (9.11×10^{-31} kg) in the relation for de Broglie wavelength, we find that an electron beam of a few hundred electron volt energy has the appropriate wavelength for obtaining the desired diffraction pattern.

For $\lambda = 1$ Å, $\quad \varepsilon = 144$ eV

This energy is very low compared to that of an x-ray photon. It is for this reason that this method is called 'low-energy electron diffraction (LEED) technique'.

3.11 FIELD ION MICROSCOPY (FIM)

It is an excellent technique for studying the surface of a sharply-pointed specimen. The shape of the surface should be ideally close to hemispherical on the atomic scale, with radius of the order of a few thousand angstroms. The elegance of FIM lies in its ability to reveal the structure of the surface charge density in real space. The LEED, like any other diffraction method, provides this information in k-space.

Figure 3.26 gives the schematic representation of a field ion microscope. A large voltage is applied between the specimen and a plate with hemispherical surface. The specimen is given the positive polarity and the plate made to serve as the negative electrode. The whole assembly is evacuated and then neutral helium atoms are allowed into the chamber. The electric field polarizes the atoms which consequently acquire induced electric dipole moment. On account of the sustained interaction with the electric field, the helium atoms are swept to the region of the strongest field. When any of these atoms comes within a few atomic spacings above the tip of the specimen, it is ionized by the electric field which is strongest in this region. The positively charged helium ions feel the repulsion from the pointed surface and move along the directions of the radial electric field to be finally collected at the plate.

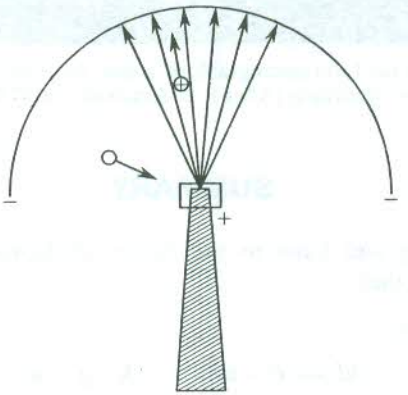


FIG. 3.26 Schematic geometry of a field ion microscope. The radial arrows indicate the direction of the electric field. The specimen (hatched cone) is given a high positive potential and a plate (represented by the semicircle) serves as the negative electrode. The empty circle denotes a neutral helium atom and the circle marked with a plus sign represents an ionized helium atom.

The field strength is maintained at such a high value as to cause the ionization of helium atoms in the close vicinity of the specimen's surface. In this condition the angular distribution of helium ions collected at the plate can be correlated with the microscopic structure of the surface whose picture on the micrograph is magnified by the ratio of the radius of curvature of the plate to the radius of curvature of the tip of the specimen. The field ion micrograph shows the actual positions of individual atoms, in addition to reflecting the complete symmetry of a crystalline sample. One such micrograph of a gold tip is shown by Fig. 3.27.

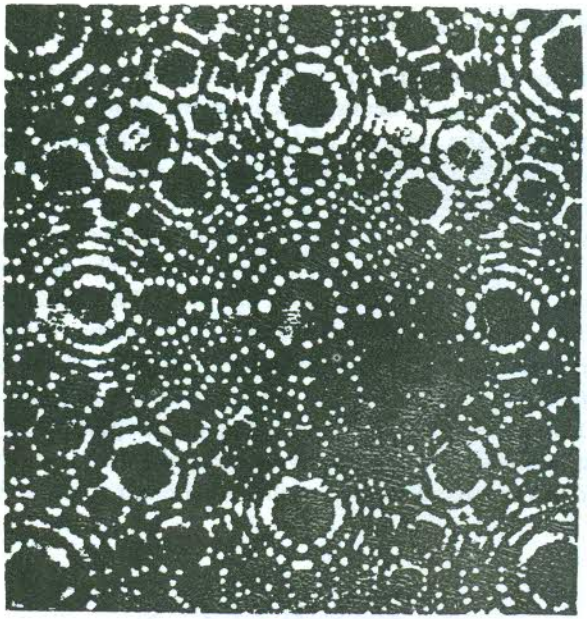


FIG. 3.27 The figure shows the field ion micrograph of a gold tip. Instead of helium atoms, Au atoms were used for imaging in this recording [After R.S. Averbach and D.N. Seidman, *Surface Science,* **40**, 249 (1973)].

SUMMARY

1. Approaches of Bragg and Laue to the theory of diffraction produced by crystals are equivalent to one another.
2. Forms of Bragg law are:

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

3. The reciprocal lattice point, representing a family of parallel planes with a common interplanar spacing, lies along the common normal of planes drawn from the origin in the reciprocal space.
4. Primitive translation vectors **a**, **b**, **c** of a crystal lattice are related to primitive translation vectors **a***, **b***, **c*** of its reciprocal lattice through

$$\mathbf{a}^* = 2\pi \frac{\mathbf{b} \times \mathbf{c}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}}; \quad \mathbf{b}^* = 2\pi \frac{\mathbf{c} \times \mathbf{a}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}}; \quad \mathbf{c}^* = 2\pi \frac{\mathbf{a} \times \mathbf{b}}{\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}}$$

5. Each of the primitive translation vectors **a***, **b***, **c*** is orthogonal to two crystal axis vectors, stated as

$$\mathbf{i}^* \cdot \mathbf{j} = 2\pi\delta_{ij}$$

where **i**, **j** = **a**, **b**, **c** and δ_{ij} is the usual delta function.

6. The general form of a reciprocal lattice vector is

$$\mathbf{g} = m_1\mathbf{a}^* + m_2\mathbf{b}^* + m_3\mathbf{c}^*$$

If m_1, m_2 and m_3 have a common factor say n, then the Miller indices h, k, l of the family of planes represented by **g** are

$$h = \frac{m_1}{n}; \quad k = \frac{m_2}{n}; \quad l = \frac{m_3}{n}$$

7.

Direct lattice	*Corresp. reciprocal lattice*
SC	SC
BCC	FCC
FCC	BCC
Hexagonal	Hexagonal

8. The Laue condition of diffraction can be written as

$$\Delta\mathbf{k} = \mathbf{g}$$

The Laue equations can be expressed as

$$\mathbf{a} \cdot \Delta\mathbf{k} = 2\pi m_1$$
$$\mathbf{b} \cdot \Delta\mathbf{k} = 2\pi m_2$$
$$\mathbf{c} \cdot \Delta\mathbf{k} = 2\pi m_3$$

with $\mathbf{g} = m_1\mathbf{a}^* + m_2\mathbf{b}^* + m_3\mathbf{c}^*$.

9. The structure factor for a crystal is

$$S_{\mathbf{g}} = \sum_{j=1}^{n} f_j \exp(i\mathbf{g} \cdot \mathbf{r}_j)$$

with j running over all atoms in the primitive cell, and where f_j denotes the form factor of an atom whose position vector is $\mathbf{r}_j$.

10. A general relation for crystals with orthogonal axes is

$$\frac{1}{d_{hkl}} = \sqrt{\frac{h^2}{a^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2}}$$

11.

Symmetry	*Ratio of* $\sin^2\theta$
SC	1 : 2 : 3 : 4 : 5 : 6 ...
BCC	1 : 2 : 3 : 4 : 5 : 6 ...
FCC	3 : 4 : 8 : 11 : 12 : 19 ...
Diamond	3 : 8 : 11 : 16 : 19 ...

PROBLEMS

3.1 Calculate the energy carried by an e.m. wave of wavelength 1 Å (i) in eV and (ii) in kelvin. Explain why no Bragg diffraction is observed for visible light.

3.2 The Bragg angle for reflection from (111) planes in Al is 19.2° for an x-ray beam of wavelength 1.54 Å. Calculate (a) the lattice constant of Al and (b) the interplanar spacing for these planes.

3.3 A two-dimensional direct lattice is formed by repeating a parallelogram of size 4 cm × 3 cm. If one of the angles of the parallelogram be $\pi/3$, determine the primitive vectors of the reciprocal lattice.

3.4 Prove that the reciprocal lattice vector

$$\mathbf{g} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^* \text{ is perpendicular to the plane } (hkl).$$

3.5 Calculate the angle between the reciprocal lattice vectors $\mathbf{g}_{100}$ and $\mathbf{g}_{111}$ of a simple cubic crystal. Find out the plane in the direct lattice to which $(\mathbf{g}_{100} \times \mathbf{g}_{111})$ is perpendicular.

3.6 Show that for BCC structure with the basis of one atom, a scheme of assigning 1/8 of each corner atom at its respective cell coordinate location to the unit cell plus the body centre atom in the central position leads to the same result for the geometrical structure factor for this lattice as the relation (3.35b).

3.7 Derive a relation for the geometrical structure factor of a crystal with CsCl structure. Would you expect (100) reflections to be present in the diffraction pattern?

3.8 Find the geometrical structure factor for an FCC crystal of monatomic basis. Which of the following reflections, (100), (110), (111), (200), (220), (222), (211), (221) and (123) would be missing in the diffraction pattern?

3.9 Show that the structure factor for a monatomic HCP structure can have any of the six values $f\left[1+\exp\left(\frac{in\pi}{3}\right)\right]$ with $n = 1, \ldots, 6$ as the reciprocal lattice vector ranges through the points of the simple hexagonal reciprocal lattice.

3.10 Cuprous oxide has a cubic unit cell with oxygen atoms at the centre (0, 0, 0) and at the corners $a(\pm1, \pm1, \pm1)$. The copper atoms are arranged in a tetrahedron around the central oxygen, at $1/2a$ (1,1,1), $1/2a$ (1, –1, –1), $1/2\ a$ (–1, 1, – 1), and $1/2a$ (–1, –1, 1). Calculate the structure factor and show that some reflections are determined only by the copper, others only by the oxygen atoms.

3.11 Powder diffraction patterns of three monatomic cubic crystals with BCC, FCC and diamond structures are recorded using a Debye–Scherrer camera. The angles 2θ in degrees of the first four diffraction lines for the three samples marked as *A, B, C* are as under:

A	*B*	*C*
42.2	28.8	42.8
49.2	41.0	73.2
72.0	50.8	89.0
87.3	59.6	115.0

(a) Determine the structure type of each sample.

(b) What is the size of the cubic cell in each case?

Take the wavelength of incident x-rays as 1.5 Å.

3.12 For a cubic crystal the diffraction line from the planes with $(h^2 + k^2 + l^2) = 8$ is observed at the angle of diffraction 10.23°. If only one line is observed at an angle lower than this, what is the crystal structure? Assuming the wavelength of x-rays used as 0.71 Å, calculate the lattice parameters.

3.13 The S values for the first three lines in the powder pattern of a cubic crystal are 34.88, 40.36 and 54.40 mm respectively. Given that the wavelength of x-rays used is 0.71 Å and the camera radius as 57.30 mm, find out the crystal structure and the lattice parameter.

3.14 Rock salt (sodium chloride) has lattice parameter of 5.63 Å. When the $K_{\alpha 1}$ line from copper target ($\lambda = 1.54$ Å) is used with a powder camera, what would be the first three S values? Take the camera diameter as 57.30 mm.

SUGGESTED FURTHER READING

Barrett, C.S. and T.B. Massalski, *Structure of Metals: Crystallographic Methods, Principles, Data,* 3rd ed. (McGraw-Hill, 1966).

Buerger, M.J., *Contemporary Crystallography* (McGraw-Hill, 1970).

Vainshtein, B.K., Modern Crystallography I, *Springer Ser.* in *Solid State Sci.,* Vol. **15** (Springer, 1981).

Chapter 4

Lattice Vibrations

The basis of crystal structures is often described in terms of ions for the interpretation of the properties of solids. The valence electrons are considered to have been placed in the force field of the lattice of ions. The word *ions* when used in this general way stands for ions in ionic crystals, ion cores in metals and covalent crystals, and atoms in a rare gas solid. The roles of ionic and electronic motions are crucial to the determination of the properties of solids. While some properties depend heavily on the electronic motion, several others are closely linked to the ionic dynamics. Lattice heat capacity, thermal expansion and hardness are some examples of properties that belong to the latter class. In the present chapter, a classical theory will be developed to describe small vibrations of atoms in crystalline solids in terms of normal modes (independent motions of characteristic frequency) of motion of constituent ions. In a normal mode all the ions move with well-defined amplitude and phase. A normal mode has the same amplitude in each cell, but varies from one unit cell to the other across the crystal like a wave with a certain wavevector. Such a wave is called *lattice wave* and the vibration* with which it is associated is commonly known as *lattice vibration*.

The analysis of lattice vibrations is faced with the major difficulty of finding a way to treat the motion of ions that are heavy, separately from that of the light electrons. This is accomplished by working in the so-called 'adiabatic approximation' which in the context of molecules is famous as the 'Born–Oppenheimer approximation'. We will realize in Chapter 6 that the significant electron velocity is the one at the Fermi level, $v_F \approx 10^6$ m s^{-1}. On the other hand, the typical ionic velocities as derived from the analysis of vibration frequencies are found at the most of the order of 10^3 m s^{-1}. Therefore, no meaningful error can be introduced by assuming that the electrons remain in their ground state at any moment for the respective ionic configuration, because the ions move so slowly on the scale of velocities of relevance to electrons. This simplifies the calculation of the potential energy of ions, enabling us to write the equations of motion of ions that are crucial to the determination of the normal modes of vibration.

Our immediate concern is to develop the mathematical structure of a simple classical theory, that is, that of lattice vibrations. We approach the task on the basis of a mechanical model of crystals, which is described below.

4.1 THE 'BALLS AND SPRINGS' MODEL OF A HARMONIC CRYSTAL

We discussed all types of binding in Chapter 2 and observed that the equilibrium interionic separation in a crystal is determined by the balance between the attractive force (at large separations) and the repulsive force (at small separations). When thermal agitations displace ions from their equilibrium

* Localized vibrations such as the internal oscillations of a SO_4^{--} ion in a crystal do not produce a lattice wave.

positions, each of them experiences a net force in the form of a restoring force. This force tends to bring the ions back to their equilibrium positions and accounts for the elastic property of solids. In a simple mechanical model we visualize a crystal as a three-dimensional periodic array of balls each of which is connected to its neighbours by massless springs. The balls and the springs are, respectively, the representations for the ions and the coupling bonds of the real crystal. The force on an ion in the crystal is calculated by drawing an analogy between the crystal and its mechanical model. For small contractions or extensions of springs, the Hooke's law is applicable to the mechanical model. Accordingly, the force on an ion is supposed to be linearly proportional to the small contractions or extensions of its nearest-neighbour distances and not at all on the more remote neighbours. At first sight this model appears to do little justice to the quantum mechanical picture of coupling between ions. But experiments confirm that small displacements in most of the solids do conform to linear Hooke's law force dependence.

Each ion inside the crystal moves in a potential well of the force field of its neighbouring ions. For small displacements or small deviations from the equilibrium value of separation of ions, the potential energy curve is parabolic, as is evident from Fig. 2.1(a) drawn for a pair of ions. Since this form of potential energy is a well-known attribute of simple harmonic motion, the ionic motion involving small displacements may be treated as simple harmonic in nature. In this book, we shall conduct discussions on lattice vibrations only within the domain of the harmonic approximation. In fact, the harmonic approximation is the starting point for all theories of lattice dynamics. For completeness it must be remarked that the harmonic approximation is known to break down in some materials of which solid helium is the best example. In solid helium the displacement of an atom from its equilibrium position in a vibration may be as large as 40 per cent of the interatomic separation. Solid helium has established its identity as a quantum solid. Its vibrational properties are described by a sophisticated theory with quantum mechanical basis.

Some of the rough estimates made below demonstrate that the harmonic approximation is generally acceptable in solids.

When the separation of a pair of ions changes by δa from the equilibrium value a, the change in the potential energy of the ions is written as

$$\delta U = \frac{1}{2} f(\delta a)^2$$

where f is called the force constant or the spring constant associated with the massless spring joining the two ions viewed as the harmonic oscillators.

For $\delta a = 0.1$ Å, Fig. 2.1(a) gives the value of δU as 0.01 eV, yielding $f = 32$ N m^{-1}. Using this value of f for a simple cubic crystal, we can readily deduce the macroscopic elastic modulus E for extension along one edge of the crystal. Thus we have

$$E = \frac{f(\delta a)/a^2}{(\delta a)/a} = \frac{f}{a} \simeq 10^{11} \text{ N m}^{-2} \qquad (\text{for } a = 3\text{Å})$$

This value of E is typically of the order of Young's modulus, identified alternatively as a stiffness constant. By ignoring the fact that an ion is coupled to several other ions, let us calculate the frequency of vibration of an ion attached to a spring whose other end is rigidly fixed. We find that the order of magnitude of frequency for $f = 30$ N m^{-1} refers to the infrared region which indeed represents the region of vibration frequencies of solids.

Another important aspect is the propagation of long waves in solids. To these waves such as sound waves, a crystal behaves as an elastic continuum. It enables us to estimate the velocity v_s of these acoustic waves as

$$v_s = \sqrt{\frac{E}{\rho}} = a\left(\frac{f}{M}\right)^{1/2}$$

where ρ is the density, M/a^3, of a simple cubic crystal with M as the ionic mass.

The above relation gives $v_s \simeq 10^4$ m s^{-1}, which is a correct order of magnitude; most of the solids have v_s in the range 1–10 km s^{-1}. This calculation, though rough, also underlines the importance of stiffness constants in the continuum approach to wave propagation in solids. In a commonly employed technique* for measuring stiffness constants, an ultrasonic pulse is sent through the crystal. The ultrasonic waves whose wavelength ($\sim \mu$m) measures several thousand times the atomic spacing treat the crystal as a continuum.

4.2 NORMAL MODES OF A ONE-DIMENSIONAL MONATOMIC CHAIN

Let us consider the simplest crystal which can be a linear chain of N identical atoms (Fig. 4.1). This is equivalent to a linear chain of N primitive cells, with one atom in each of them. If we wish to describe the vibrations of the chain we are confronted with the problem of accounting appropriately the motion of ions in the middle and the motion of two ions at the ends. The broad feature of the motion may still be obtained by considering only the nearest neighbour interactions and ignoring the ends. The results of such a calculation are most acceptable when the number of atoms is large. If we denote the displacement of an ion at any moment from its site l in the static lattice by $\mathbf{s}_l$, the effective potential energy of a chain of interionic separation, a, in the harmonic approximation is

$$U^{\text{eff}} = \sum_l \frac{1}{2} f(\mathbf{s}_l - \mathbf{s}_{l+a})^2 \tag{4.1}$$

where only the nearest neighbour interactions are included.

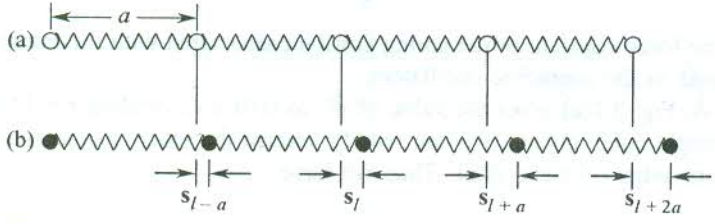


FIG. 4.1 The displacement of ions in a linear chain of identical ions connected by springs.

The restoring force on the ion situated in the static lattice at the site l can be expressed as

$$M\ddot{\mathbf{s}}_l = -f[(\mathbf{s}_l - \mathbf{s}_{l+a}) + (\mathbf{s}_l - \mathbf{s}_{l-a})]$$

$$= -f(2\mathbf{s}_l - \mathbf{s}_{l+a} - \mathbf{s}_{l-a}) \tag{4.2}$$

* R.F.S. Hearmon, *Rev. Mod. Phys.*, **18**, 409 (1946).

The preceding equation represents the equation of motion of a single ion in the chain. We expect from symmetry a normal mode solution in the form

$$\mathbf{s}_l = \mathbf{s}_0 \exp\,[i(\mathbf{k} \cdot \mathbf{l} - \omega t)] \tag{4.3}$$

which is a wave with wavevector $\mathbf{k}$ varying as $\exp\,(i\mathbf{k} \cdot \mathbf{x})$ along the line of ions (the $\mathbf{x}$ direction) with amplitude s_0.

Substituting $\mathbf{s}_l$ from (4.3) in (4.2), we get

$$M\omega^2 = f[2 - \exp\,(ika) - \exp\,(-ika)]$$

or

$$\omega^2 = \frac{2f}{M}\,(1 - \cos ka) = \frac{f}{M}\,4\sin^2\left(\frac{1}{2}\,ka\right)$$

which gives the angular frequency of oscillations as

$$\omega(k) = 2\sqrt{\frac{f}{M}}\left|\sin\left(\frac{1}{2}\,ka\right)\right|$$

$$= \omega_m\left|\sin\left(\frac{1}{2}\,ka\right)\right| \tag{4.4}$$

if the positive root is chosen.

ω_m is the frequency maximum $\left(2\sqrt{\frac{f}{M}}\right)$ observed at $k = \pm\,\pi/a$. It must be noted that l does not figure in (4.4), indicating that the equation of motion of every ion gives the same algebraic relation between ω and k. This shows that the trial function is indeed a solution of (4.2). Relation (4.4) is the required dispersion relation and the ω versus k plot derived from this is known as the *dispersion curve*.

Another important observation is that we started with equations of N coupled harmonic oscillators (4.2), implying that if one ion starts vibrating it does not continue with constant amplitude, but transfers energy to others in a complicated way. Thus the vibrations of individual ions are not simple harmonic on account of this energy exchange. Our solutions, on the other hand, are uncoupled oscillations called normal modes. They represent independent non-interacting modes in harmonic approximation. Any pattern of displacement of the atoms can be written as a superposition of the normal modes just as any vector in three dimensions is a superposition of three independent (i.e., orthogonal) unit vectors along x, y and z axes.

4.2.1 The Periodic Boundary Condition

The dispersion relation (4.4) shows that $\omega = 0$ at $k = 0$ and $\omega = \omega_m$ at $k = \pm\,\pi/a$. This indicates that all the normal mode frequencies lie in the range of k-values from $-\,\pi/a$ to $+\,\pi/a$, defining the extent of the first Brillouin zone of a linear lattice as described in Chapter 3. The complete spectrum of normal modes is usually determined by invoking the periodic boundary condition, originally given by Born and von Karman for the electron gas. The condition requires the two ends of the crystal to be joined.

It is appreciated most easily in the case of the electronic conduction. The electrical conductivity depends on the details of electronic motion within the crystal. It implies as if an electron enters the crystal at one end as soon as it leaves the other end. The fact is equivalent to stating that the two ends of the crystal behave as being in contact. In the present case the ions at the ends of the linear chain are imagined to have been joined by an additional spring which is identical with those considered to couple the successive ions in the chain. The linear chain is thus transformed into a ring as shown in Fig. 4.2.

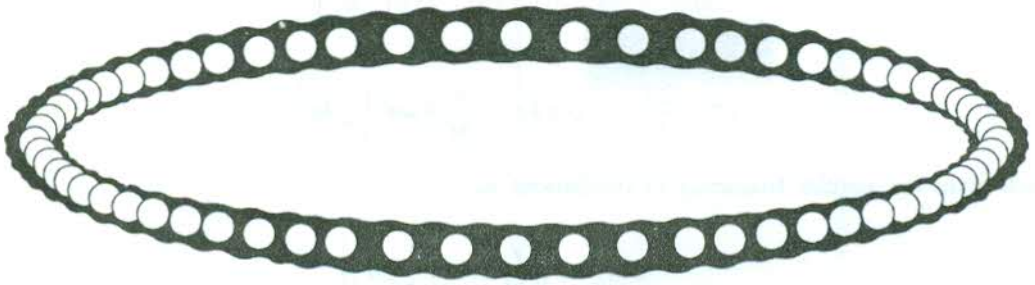


FIG. 4.2 The Born–von Karman periodic boundary condition for a linear chain of N identical ions. If s_l denotes the magnitude of displacement of an ion at site l, $s_l = s_{l+Na}$; a is the interatomic spacing when the ions are at rest.

When we move away from a certain ion along the ring and take N steps, each of length a (the interatomic spacing), we are back to the same ion. Then, we require that

$$s_l = s_{l+Na} \tag{4.5}$$

where

$l = pa$, p being an integer

Na = the length of the chain L.

Using the solution (4.3) in the condition given by (4.5), we have

$$\exp(ik \cdot Na) = 1 = \exp(i2\pi n) \tag{4.6}$$

or

$$k = \frac{2\pi}{a} \cdot \frac{n}{N}$$

$$= n \cdot \frac{2\pi}{L} \tag{4.7}$$

where n is an integer.

So, the allowed values of k are given by

$$k = 0, \pm \frac{2\pi}{L}, \pm 2\frac{2\pi}{L}, \ldots \tag{4.8}$$

From (4.3) it is evident that if k changes by $2\pi/a$ and or by its multiples, the solution remains unchanged. But such changes ($2\pi n/a$) in k denote the reciprocal lattice vectors of the one-dimensional lattice under discussion. As such a root of frequency ω corresponding to a certain k-value repeats

whenever the k-value is changed by a reciprocal lattice vector. This result forms one of the founding principles of the theory of solid state. The range of the first Brillouin zone is from $-\pi/a$ to π/a, totalling $2\pi/a$ in width. Therefore, for every k-value in the second, third or any other zone, there is a k-value in the first zone such that the vibration frequency for the two k-values is common. It happens so because the points of k-values in the set are connected by the reciprocal lattice vectors (see Fig. 4.3). This leads to a conclusion of great significance that all unique values of k that satisfy the solutions (4.3) lie within the first zone and we need not search for the k-values in other zones as they yield no new roots of ω.

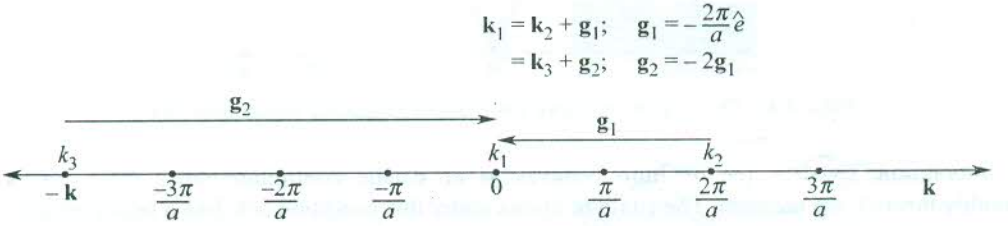


FIG. 4.3 Brillouin zones of a one-dimensional crystal of lattice constant a. Wavevectors $\mathbf{k}_2$ (extending to the second zone) and $\mathbf{k}_3$ (extending to the third zone) are connected to wavevector $\mathbf{k}_1$ (within the first zone) by reciprocal lattice vectors $\mathbf{g}_1$ and $\mathbf{g}_2$, respectively. The form of the solution to the equation of motion is such as to yield the same value of vibration frequency for $\mathbf{k}_1$, $\mathbf{k}_2$ and $\mathbf{k}_3$. Thus all important $\mathbf{k}$-values (known as unique values) lie within the first zone.

Now, we are in a position to fix the last term in the set of the allowed k-values given by (4.8) and rewrite the set as

$$k = 0, \pm \frac{2\pi}{L}, \pm 2\frac{2\pi}{L}, \ldots, \frac{N}{2}\frac{2\pi}{L}\left(= \frac{\pi}{a}\right) \tag{4.9}$$

The last term coincides with the boundary π/a of the first Brioullin zone. The negative sign is dropped since $-\pi/a$ is not independent on account of being connected to π/a by the shortest reciprocal lattice vector of magnitude $2\pi/a$. The total number of unique k-values in the allowed set is N and so will be the total number of normal modes.

4.2.2 Salient Features of the Dispersion Curve

The dispersion relation (4.4) and the dispersion curve displayed in Fig. 4.4 show only the nearest neighbour interactions. In this spirit a monatomic linear chain may be considered to act as a low pass filter. The features of the dispersion curve are further revealing in the extreme limit of the wavevector $\mathbf{k}$. In the limit of small k-values, more often referred to as the long wavelength limit, sin $(ka/2)$ in relation (4.4) may be replaced with $ka/2$ in the first approximation. The dispersion relation would then read as

$$\omega(k) = \sqrt{\frac{f}{M}}\, ka \tag{4.10}$$

From (4.10) we infer that the frequency varies linearly with the wavevector for small wavevectors. This behaviour is clearly evident in Fig. 4.4. The behaviour of elastic waves in continuum is of exactly similar nature. In the long wavelength limit of the waves where the wavelength is much larger than

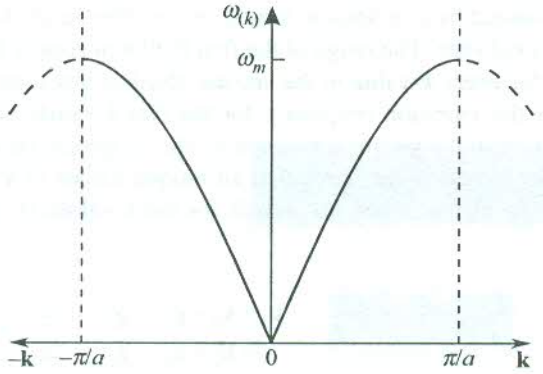


FIG. 4.4 The dispersion curve for a one-dimensional monatomic chain.

the interatomic spacing, the medium behaves as an elastic continuum since these waves pass smoothly through the medium. The chain of atoms under this condition acts like a heavy elastic string.

In this limit the group velocity $d\omega/dk$ and the phase velocity ω/k of the elastic waves (or sound waves) are equal and both become independent of frequency. But as k changes to larger values, the discreteness of the medium begins to show up and at the zone boundary ($k = \pm \pi/a$), the tangent to the dispersion curve is horizontal showing thereby that the group velocity is zero here. This refers to the total dispersion and no waves propagate through the crystal indicating that it acts as a discrete medium in this situation. This sounds perfectly logical as at $k = \pm \pi/a$, the wavelength is twice the interatomic spacing ($\lambda = 2a$) for which the crystal cannot be treated as a continuous medium. The zero value of the group velocity also shows that the motions of the neighbouring atoms are out of phase and the elastic waves suffer Bragg reflection at this point in the k-space. This behaviour is consistent with the condition for Bragg reflection (discussed in Chapter 3) in a one-dimensional crystal. The elastic waves are no more the travelling waves defined by (4.3) and get transformed into standing waves at the zone boundary.

4.3 NORMAL MODES OF ONE-DIMENSIONAL DIATOMIC CHAIN

This chain is different from the earlier one in the sense that there are two different types of atoms whose positions alternate along its length (Fig. 4.5). The chain may be viewed as a one-dimensional crystal whose primitive cell contains two atoms of different masses M and m. The relevance of treating one-dimensional atomic chains as against the reality of crystals being three-dimensional may be questioned. But these calculations bring out a few broad features of the vibrations of real crystals,

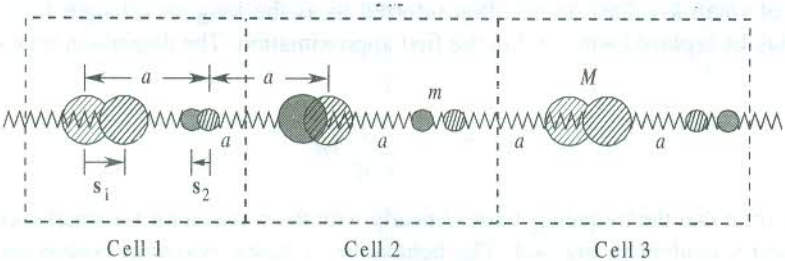


FIG. 4.5 The displacement of ions in three consecutive unit cells of a one-dimensional diatomic crystal.

as seen in the earlier section. In other words, it suffices to acknowledge that these exercises introduce us to the basics of lattice dynamics using simple mathematics.

From Fig. 4.5 we see that the springs are identical. Therefore, if we consider only the nearest neighbour interactions, a single force constant (say, f) will be involved in the equations of motion. Supposing that the heavy ion (M) occupies the site 1 and the light one (m) is at the site 2 in each primitive cell, we obtain the following equations of motion for these ions in that order:

$$M\ddot{\mathbf{s}}_{1n} = -f(2\,\mathbf{s}_{1n} - \mathbf{s}_{2n} - \mathbf{s}_{2,n-1}) \tag{4.11}$$

$$m\ddot{\mathbf{s}}_{2n} = -f(2\,\mathbf{s}_{2n} - \mathbf{s}_{1n} - \mathbf{s}_{1,n+1}) \tag{4.12}$$

where $\mathbf{s}_{1n}$ stands for the displacement of the ion at site 1 in the primitive cell n.

We seek the following solutions to the above equations:

$$\begin{aligned} \mathbf{s}_{1n} &= \mathbf{u} \exp\,[i(kna - \omega t)] \\ \text{and} \qquad \mathbf{s}_{2n} &= \mathbf{v} \exp\,[i(kna - \omega t)] \end{aligned} \tag{4.13}$$

Substituting these solutions in (4.11) and (4.12), we have

$$\begin{aligned} (2f - M\omega^2)u - f[1 + \exp(-ika)]v = 0 \\ -f[1 + \exp(ika)]u + (2f - m\omega^2)v = 0 \end{aligned} \tag{4.14}$$

where u and v (the amplitudes) are unknowns in the above homogeneous equations. The equations have solutions only if the determinant of the coefficients of u and v in them vanishes. That is,

$$\begin{vmatrix} (2f - M\omega^2) & -f[1 + \exp(-ika)] \\ -f\,[1 + \exp(ika)] & (2f - m\omega^2) \end{vmatrix} = 0 \tag{4.15}$$

or

$$Mm\omega^4 - 2f(M + m)\omega^2 + 2f^2(1 - \cos ka) = 0 \tag{4.16}$$

Based on our experience with the monatomic lattice it is advisable to solve (4.16) for small k (i.e. long wavelength limit) and for the largest k, i.e. at the first zone boundary. These describe the distinct features of the dispersion curves.

For small values of k, we have

$$\cos ka = 1 - \frac{1}{2!}(ka)^2 + \cdots$$

Retaining the first two terms of the series and putting this value of cos ka in (4.16), we get the following two roots:

$$\omega^2 \cong 2f\left(\frac{1}{M} + \frac{1}{m}\right) \tag{4.17}$$

and

$$\omega^2 = \frac{\frac{1}{2}f}{M + m}k^2a^2 \tag{4.18}$$

The dispersion curve obtained from (4.17) is called *optical branch* while the one from (4.18) is known as *acoustical branch*.

For the maximum value of k, i.e. at $k = \pm \pi/a$, the roots are

$$\omega^2 = \frac{2f}{m} \quad \text{(optical branch)} \tag{4.19}$$

and

$$\omega^2 = \frac{2f}{M} \quad \text{(acoustical branch)} \tag{4.20}$$

4.3.1 Salient Features of the Dispersion Curves

The most prominent feature of the dispersion curves shown in Fig. 4.6 is the manifestation of a frequency gap between the acoustical and optical branches. This brings out the fact that a diatomic linear chain acts as a band pass filter. It contrasts the behaviour of a monatomic linear chain which was shown to act as a low pass filter (see Section 4.2). Other features of the observed two branches representing two different types of normal modes are discussed below.

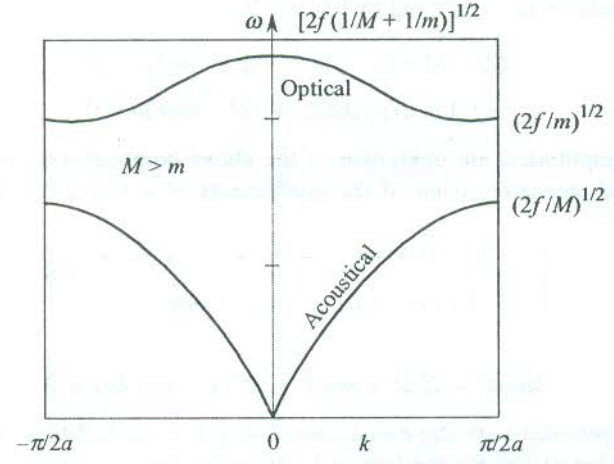


FIG. 4.6 Optical and acoustical branches of dispersion curves for a one-dimensional diatomic crystal of lattice constant a. The frequency extrema for two branches are given.

The frequency of the optical branch is nearly constant in the limit $k \to 0$ as made out by the approximate nature of (4.17). But it decreases slowly as k increases, dropping to the value $2f/m$ at the zone boundary. The acoustic branch corresponds to the single branch, obtained for the linear chain of monatomic atoms of Section 4.2. The linear behaviour of ω with k in the limit of small k (or long wavelength) is in the limit of sound waves which are longitudinal and treat the crystal as elastic continuum.

We may, further, exploit the above treatment to derive the state of ionic motions in the two branches, again for the same two limiting cases.

Substituting ω^2 from (4.17) in (4.14), at $k = 0$, we have for the optical branch,

$$\frac{u}{v} = -\frac{m}{M}$$

or

$$Mu = -mv \tag{4.21}$$

Relation (4.21) shows that the movements of the heavy and light ions are out of phase, i.e. they move towards each other or away from each other such that their centre of mass remains at rest. This refers to the situation in ionic crystals where such a motion of positive and negative ions may displace the centre of positive charges with the centre of negative charges creating an electric dipole. The oscillating electric dipole on interaction with the electromagnetic radiation may absorb the radiation. From a rough calculation we can find that the frequency of vibration of ionic crystals (e.g. NaCl, KBr, LiF etc.) lies in the infrared region. That is why ionic crystals are known to absorb infrared light. This forms the basis for giving the name *optical branch* to the branch under discussion. The relevance of treating a diatomic linear chain is thus amply justified. The results can be only instructive as we have allowed the linear chain to produce only longitudinal waves. In real crystals, there can be two transverse waves for each longitudinal wave. Generally, the frequencies of all the modes are different with the exception that the two transverse modes along directions of high symmetry in the crystal are degenerate (see Fig. 4.10). The longitudinal and transverse vibrational modes of a crystal can be clearly separated only in certain symmetry directions of crystals. The modes for any arbitrary direction are mixed in character.

For acoustical branch at $k = 0$, we obtain

$$\frac{u}{v} = \frac{m}{M} \tag{4.22}$$

which shows that the two ions move in phase.

The state of motions of the ions in the optical and acoustical branches is illustrated for a transverse wave in Fig. 4.7. The example of a transverse wave is chosen since the difference between the motions of ions in the two branches is more striking in appearance for the transverse motion. Though this motion is not allowed in the linear chain, it is present in all real crystals.

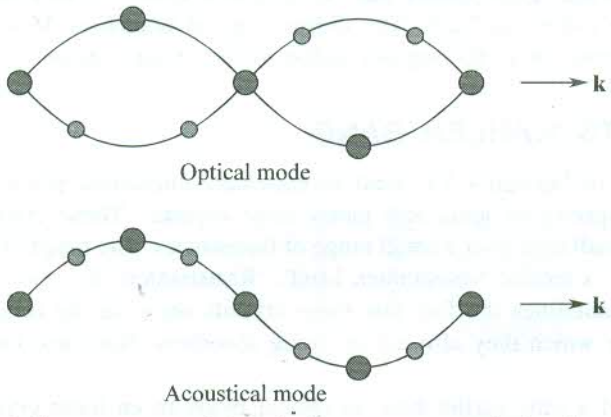


FIG. 4.7 Illustration of the movement of ions in the transverse optical and acoustical modes at equal wavelength in a diatomic linear crystal.

The variation of the amplitude ratio u/v with the wavevector can be seen in Fig. 4.8. The state of motion at the zone end ($k = \pm \pi/a$, the maximum k) can be interpreted with the help of this figure. The light ions are at rest in the acoustical mode while the heavy ions are at rest in the optical mode.

Another significant feature of the dispersion curves is that the gap at the zone end decreases with the decrease in the mass of heavy ion and approaches zero as $M \to m$. But it would surprise us to know that we have still two different branches though $M = m$ which should give only the

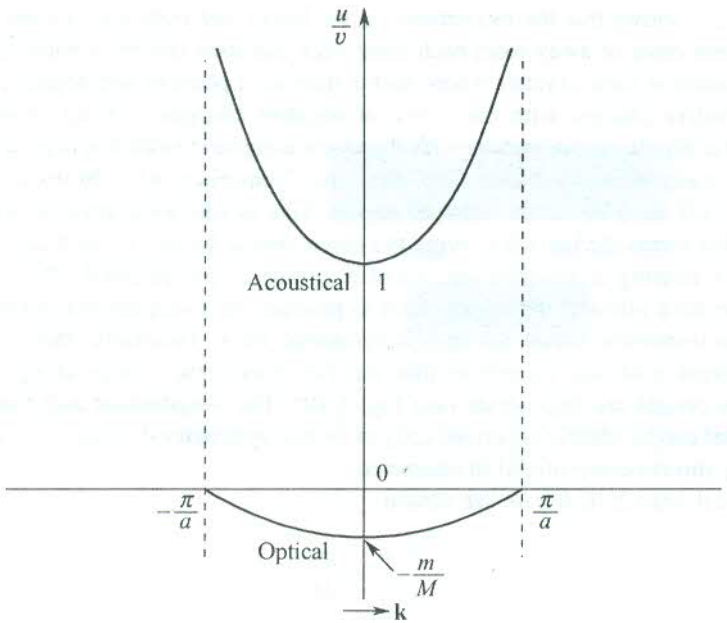


FIG. 4.8 The amplitude ratio u/v as a function of the wavevector **k** for the acoustical branch (upper curve) and the optical branch (lower curve); u belongs to M and v to m ($M > m$).

acoustical branch. For the present chain, if we have monatomic basis, we need to correct the size of the first zone which will now be double in size owing to the interionic spacing or the lattice constant being halved $a/2$. This ensures that the acoustical branch is continuous over whole region of the modified zone ($-2\pi/a$ to $2\pi/a$). The illusory optical branch of $M = m$ in the original zone gets effectively reflected on to the regions added to the original zone.

4.4 THE RESTSTRAHLEN BAND

We apply the theory of Section 4.3 to treat an extremely interesting phenomenon observed in the infrared absorption spectra of ionic and partly ionic crystals. These crystals show very intense reflection of infrared radiation over a small range of frequencies. The range of frequencies over which this reflection occurs is termed 'reststrahlen band'. 'Reststrahlen' is a German word which means residual rays. This underlines the fact that these crystals show strong reflection to a radiation of certain frequency, for which they also act as strong absorbers. NaCl and GaAs are good examples of such crystals.

It was mentioned a little earlier how an optical mode in an ionic crystal is excited when an electromagnetic radiation is incident on its surface. The transverse electric field of the radiation does the trick by exerting force on cations and anions in opposite directions. The strong absorption (resonance) takes place when the frequency of radiation matches the frequency of a transverse optical mode of the crystal. A proper explanation to the reststrahlen phenomenon follows from the theory of optical constants to be discussed in Chapter 10. According to this theory, when

$$\omega_T < \omega < \left(\frac{\epsilon_s}{\epsilon_\infty}\right)^{1/2} \omega_T \tag{4.23}$$

no wave can propagate through the crystals.

Here ω_T is the frequency of a transverse optical mode, ω the frequency of the incident radiation and ϵ_s and ϵ_∞ are the static dielectric constant and the dielectric constant at high frequencies respectively. In the above condition we have an evanescent wave that decays exponentially with the increasing distance in the crystal. Therefore, in the specified range of frequencies the radiation incident on the crystal from outside suffers total external reflection. This is what we know as the reststrahlen phenomenon. The range of frequencies over which this occurs is called the *reststrahlen band.*

Now we make an estimate of ω_0, the frequency at which the strong absorption occurs using equations of motion set up in Section 4.3 for a diatomic lattice. This also gives the measure of the reststrahlen frequency which is the same as ω_0. For NaCl crystal, the lattice constant a is equal to 5.63 Å. With $\lambda = 2a$, we get the maximum cut-off wavevector $k_m = 2\pi/\lambda \approx 10^{10}$ m^{-1}. The wavevector associated with a typical infrared radiation of wavelength 10,000 Å is about 10^6 m^{-1}. Therefore, the vibration with these relatively small wavevectors in the optical branch can be determined in the limit $k \to 0$. In this limit, relations (4.14) reduce to

$$\begin{aligned} -M\omega^2 u &= 2f(v - u) \\ -m\omega^2 v &= -2f(v - u) \end{aligned} \tag{4.24}$$

Relations (4.24) essentially give the force on masses M and m, respectively. When an electromagnetic radiation with the transverse electric field $E = E_0 \exp(i\omega t)$ is incident, these force terms get corrected by $\mp eE_0$. Then relations (4.24) reduce to

$$\begin{aligned} -M\omega^2 u &= 2f(v - u) - eE_0 \\ -m\omega^2 v &= -2f(v - u) + eE_0 \end{aligned} \tag{4.25}$$

Relations (4.25) give the two amplitudes of vibration as

$$u = \frac{-\dfrac{eE_0}{M}}{\omega_0^2 - \omega^2}$$

$$v = \frac{\dfrac{eE_0}{m}}{\omega_0^2 - \omega^2} \tag{4.26}$$

with

$$\omega_0^2 = 2f\left(\frac{1}{M} + \frac{1}{m}\right) \tag{4.27}$$

Though relations (4.26) indicate a single frequency at which the maximum absorption should occur, experiments show that there is a small range of frequencies of the electromagnetic radiation showing the reststrahlen phenomenon. This small range of frequencies forms the reststrahlen band. The frequency ω_0 is referred to as the *reststrahlen frequency.*

4.5 GENERAL THEORY OF HARMONIC APPROXIMATION

Consider a primitive cell as shown in Fig. 4.9. Let us use n to index the primitive cells with $\mathbf{n} = (n_1, n_2, n_3)$, where n_1, n_2 and n_3 form a triplet of integers that mark the cells along the primitive

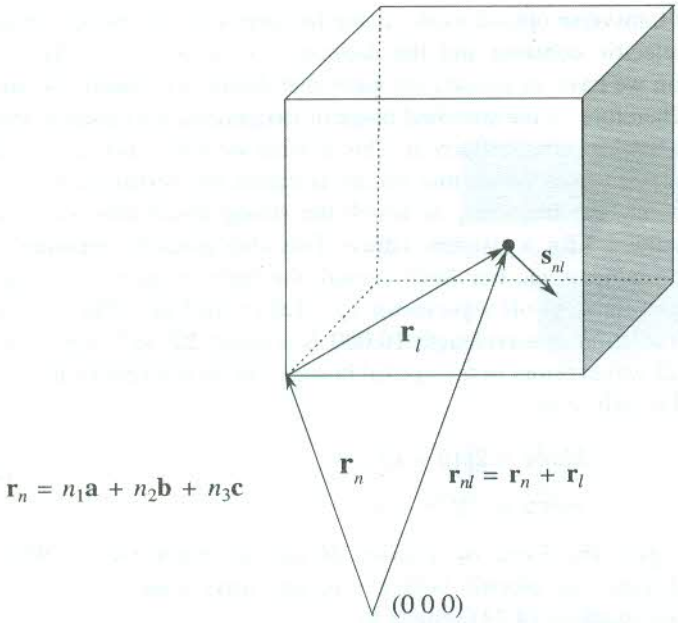


FIG. 4.9 The picture explains the relationship between the position vectors ($\mathbf{r}_{nl}$ and $\mathbf{r}_l$) of an atom at site **l** within the nth unit cell as measured from the origin (0, 0, 0) fixed at a lattice point and from a corner of the unit cell $\mathbf{n} = (n_1, n_2, n_3)$ respectively. $\mathbf{s}_{nl}$ denotes the displacement vector of the atom from its equilibrium position.

axes. We measure the position vector of an ion within a cell from one of its corners, identified as the origin of the cell. Thus the position vector of an ion at site **l** within the cell **n** may be given by

$$\mathbf{r}_{nl} = \mathbf{r}_n + \mathbf{r}_l$$

where
$$\mathbf{r}_n = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c} \tag{4.28}$$

Assume that the average position of the ion at site l is still given by $\mathbf{r}_l$ when the ion vibrates. Thus the symmetry of the Bravais lattice is supposed to remain unchanged in a vibrating crystal. Further, it is assumed that the displacement or deviation ($\mathbf{s}_{ln}$) of the ion from its equilibrium position is much small in comparison with the interionic distance.

Small values of the atomic displacements allow us to expand the total potential energy about its equilibrium value in a Taylor's series. This makes the exercise of obtaining the equation of motion quite straightforward, which is otherwise extremely tedious.

The kinetic energy T of the crystal can be written as

$$T = \sum_{ln} \frac{1}{2} M_l \, |\dot{\mathbf{s}}_{ln}|^2 \tag{4.29}$$

where M_l is the mass of the ion at site l and n stands for the unit cell running through all the unit cells.

And the potential energy U is defined by its Taylor's expansion as

$$U = U_0 + \sum_{lnj} \mathbf{s}_{lnj} \left(\frac{\partial U}{\partial s_{lnj}} \right)_0 + \frac{1}{2} \sum_{ll', nn', jj'} \mathbf{s}_{lnj}\, \mathbf{s}_{l'n'j'} \left[\frac{\partial^2 U}{\partial \mathbf{s}_{lnj} \cdot \partial \mathbf{s}_{l'n'j'}} \right]_0 + \ldots \tag{4.30}$$

where index j denotes the three Cartesian x, y, z components.

The first term U_0 refers to the potential energy at equilibrium which is a constant. It can be dropped from the expression as it is the changes in potential energy from U_0 that determine the vibration spectrum of a crystal. The coefficient of $\mathbf{s}_{lnj}$ in the linear term is equal to the magnitude of the net force on the ion at site l exerted by all the other ions when each of the ions is at its equilibrium position. At equilibrium, this force vanishes and so should the linear term. The terms of higher orders beyond the third term are not considered in the harmonic approximation. Therefore, the lone non-vanishing third term in the expansion expresses the effective potential energy of the crystal in the harmonic approximation. That is,

$$U^{\text{eff}} = \frac{1}{2} \sum_{ll', nn', jj'} \mathbf{s}_{lnj}\, \mathbf{s}_{l'n'j'} \left(\frac{\partial^2 U}{\partial \mathbf{s}_{lnj} \cdot \partial \mathbf{s}_{l'n'j'}} \right)_0 \tag{4.31}$$

We combine (4.29) and (4.31) to write the usual Lagrange's equation from the equations of motion for the Cartesian components obtained in the form

$$M_l\, \ddot{\mathbf{s}}_{lnj} = - \sum_{l', n', j'} \left(\frac{\partial^2 U}{\partial \mathbf{s}_{lnj} \cdot \partial \mathbf{s}_{l'n'j'}} \right) \mathbf{s}_{l'n'j'} \tag{4.32}$$

Relation (4.32) represents the jth component (which could be the x-, y- or z-component) of the net force on the ion at site l in the unit cell n owing to the rest of the atoms in the crystal. This equation marks the beginning of all theories of lattice dynamics.

In view of the elastic nature of crystals, the coupling between any two atoms is customarily represented in the material form of a spring. Each spring in a crystal is assigned a constant, whose value is characteristic of the nature of the ions being coupled. In classical mechanics, this constant is called by the name *spring constant* or *force constant.*

A comparison of (4.32) with the equation of motion of a particle executing simple harmonic motion shows that the derivatives in (4.32) serve as force constants in these equations of motion. For convenience, we will use from now onwards a shorter notation for these derivatives given by

$$\left(\frac{\partial^2 U}{\partial \mathbf{s}_{lnj} \cdot \partial \mathbf{s}_{l'n'j'}} \right) = U_{lnj}^{l'n'j'} \tag{4.33}$$

The equation of motion (4.32) then reads as

$$M_l\, \ddot{\mathbf{s}}_{lnj} = - \sum_{l'n'j'} U_{lnj}^{l'n'j'} \mathbf{s}_{l'n'j'} \tag{4.34}$$

Equation (4.34) is the general equation of motion for ions in crystals. It is to be observed that the equations of motion for the linear chains of monatomic basis (4.2) and diatomic basis [(4.11) and

(4.12)] are only the special cases of this equation with $U_{lnj}^{l'n'j'}$ having the dimensions of force constants. They denote the generalized force constants of a system with many degrees of freedom. The isotropy of space, the translational invariance and the point group symmetry require the coupling constants to satisfy certain conditions.

It must be noted that each term within the summation in (4.32) denotes a force that depends on the relative position of unit cells n and n' and not on their absolute position. It is the consequence of the translation invariance which effectively requires that

$$U_{lnj}^{l'n'j'} = U_{loj}^{l'(n'-n)j'} \tag{4.35}$$

We have, now, come to grips with the basic workable technique of solving the equations of motion in crystals.

4.6 NORMAL MODES OF REAL CRYSTALS

This exercise concerns real crystals which are three-dimensional and in general may have a polyatomic basis having atoms of different elements. The problem is most general and, therefore, the task of finding its solution is bound to be tedious. But, in principle, the solution is tractable. We give below the procedure for calculating the frequencies of normal modes in the framework of a theory based on the harmonic approximation.

The equation of motion of an ion, whose site in the unit cell n is denoted by the position vector l_n, is given by (4.34). If the crystal is composed of N unit cells with p atoms in a unit cell, we get in total $3pN$ equations analogous to (4.34). It is proper to write the solutions or the displacement $\mathbf{s}_{lnj}$ in terms of a plane wave with respect to the cell coordinates. That is,

$$\mathbf{s}_{lnj} = \frac{1}{\sqrt{M_l}} \mathbf{u}_{lj}(\mathbf{k}) \exp[i(\mathbf{k}\cdot\mathbf{r}_n - \omega t)] \tag{4.36}$$

This plane wave is defined only at the lattice sites, unlike a normal plane wave.

Using this solution for (4.34), we have

$$-\omega^2 \mathbf{u}_{lj}(\mathbf{k}) + \sum_{l'j'} \sum_{n'} \frac{1}{\sqrt{M_l M_{l'}}} U_{lnj}^{l'n'j'} \exp[i\mathbf{k}\cdot(\mathbf{r}_{n'} - \mathbf{r}_n)] \cdot \mathbf{u}_{l'j'}(\mathbf{k}) = 0 \tag{4.37}$$

Let us write (4.37) as

$$-\omega^2 \mathbf{u}_{lj}(\mathbf{k}) + \sum_{l'j'} D_{lj}^{l'j'}(\mathbf{k})\, \mathbf{u}_{l'j'}(\mathbf{k}) = 0 \tag{4.38}$$

where

$$D_{lj}^{l'j'}(\mathbf{k}) = \sum_{n'} \frac{1}{\sqrt{M_l M_{l'}}} U_{lnj}^{l'n'j'} \exp[i\mathbf{k}\cdot(\mathbf{r}_{n'} - \mathbf{r}_n)] \tag{4.39}$$

According to (4.35), the terms in the sum of the above equations depend only on the difference $(n' - n)$ and not on the absolute values of n and n'. The quantity $D_{lj}^{l'j'}(\mathbf{k})$, which is obtained by summing over n', is independent of n. It is instructive to note that it couples amplitudes with each

other without having to depend on n. This also explains why the amplitudes in (4.13) appear without index n. A very common term in lattice dynamics known as the *dynamical matrix* is formed by the quantities defined by (4.39). The set of equations (4.38) belongs to a set of linear homogeneous set of order $3p$. The set of linear homogeneous equations has a solution only if the

$$\text{determinant: } \{ D_{ij}^{l'j'}(\mathbf{k}) - \omega^2 \} \text{ vanishes.} \tag{4.40}$$

The above equation gives $3p$ different solutions, i.e. for one value of $\mathbf{k}$ there are $3p$ values of ω, each of which lies on a separate branch. A branch is characterized by its dispersion relation. Thus, there are in total $3p$ branches out of which three are acoustical—one longitudinal (LA) and two transverse (TA). The number of acoustical branches is three in all crystals as it does not depend on the number of atoms. The rest $(3p - 3)$ belong to the optical branches; $(p - 1)$ longitudinal (LO) and $2(p - 1)$ transverse (TO). Since there are N (the number of unit cells) unique values of $\mathbf{k}$ for each branch, the total number of vibrational modes comes to $3pN$.

A complete description of the normal modes of a crystal requires the knowledge of directions along which the ions move for a certain value of $\mathbf{k}$. These directions are known as the directions of *polarization*. In an isotropic crystal one mode will have ions moving in the direction of $\mathbf{k}$. This determines the longitudinal branch. The two transverse branches are degenerate. We can choose any two orthogonal directions in a plane perpendicular to $\mathbf{k}$ that represent the polarization of the two independent transverse modes. The picture in anisotropic crystals is not so straightforward. We can describe the dispersion curves of these crystals in terms of longitudinal and transverse branches, only if $\mathbf{k}$ lies along any n-fold axis of rotation permitted by the crystal symmetry. In such a case the longitudinal mode is polarized along $\mathbf{k}$, and transverse modes, that are degenerate, along directions perpendicular to $\mathbf{k}$.

It can now well be imagined that the picture of dispersion curves for crystals with polyatomic basis will be fairly complex. We take a relatively simple example of diamond whose primitive cell has two identical atoms. The dispersion curves plotted from the experimental data* are shown in Fig. 4.10 and refer to the propagation of waves along two important directions [100] and [111]. The wavevector is expressed as a dimensionless quantity, $q = k(2\pi/a)$, called the reduced wavevector as measured from the centre of the Brillouin zone. In principle, there should be six branches in total. But we observe only four, as the transverse modes along [100] and [111] directions are degenerate in the acoustical and the optical branches, separately. The other important point exhibited by the curves is that the LA and LO modes at the zone boundary are degenerate. This confirms the zero gap between the acoustical and the optical branches at this point as derived in Section 4.3 for the linear chain of a diatomic lattice composed of identical atoms.

4.7 QUANTIZATION OF LATTICE VIBRATIONS

The energy of lattice waves is quantized the same way as that of the electromagnetic waves. In a harmonic crystal, the atomic oscillators are treated analogous to the Planck's radiation oscillators. The energy of a vibrational mode of angular frequency ω with wavevector $\mathbf{k}$ in the branch s is expressed as $\left(n_{\mathbf{k}s} + \frac{1}{2}\right)\hbar\omega_s(\mathbf{k})$, where $n_{\mathbf{k}s}$ is an integer denoting the excitation number or the order of the

* J.L. Warren, R.G. Wenzel and J.L. Yarnell, *Inelastic Scattering of Neutrons* (Vienna, International Atomic Energy Agency, 1965).

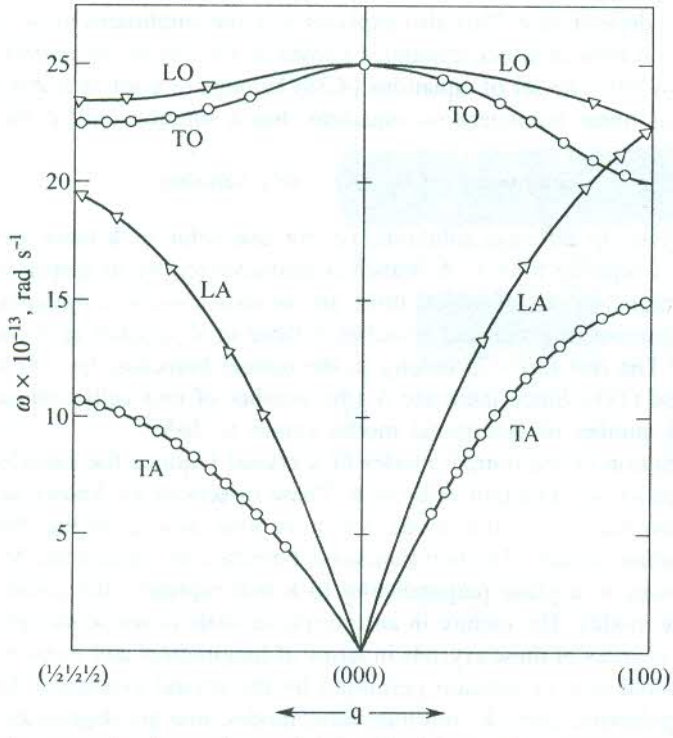


FIG. 4.10 Dispersion curve for diamond crystal (after J.L. Warren, R.G. Wenzel and J.L. Yarnell, *Inelastic Scattering of Neutrons* (Vienna, International Atomic Energy Agency, 1965).

excitation state of the classical normal mode. The fact of the normal mode being in its n_{ks} excited state is expressed in the language of quantum theory by saying that there are n_{ks} phonons of wavevector $\mathbf{k}$ in branch s. The usage of the term *phonon* is analogous to the term *photon* for the electromagnetic radiation. Phonon is the corpuscular representation for a quantum of vibration of energy $\hbar\omega_s(\mathbf{k})$ carried by a sound wave in the same way as photon represents a quantum of radiation in an electromagnetic wave. The n_{ks} is also defined as the phonon occupancy expressed by the Planck's distribution function which is a function of $\omega_s(\mathbf{k})$ and $\mathbf{k}$.

The thermal energy of a harmonic crystal is given by

$$E = \sum_{\mathbf{k}\,s} \left(n_{\mathbf{k}s} + \frac{1}{2} \right) \hbar\omega_s(\mathbf{k}) \tag{4.41}$$

Relation (4.41) shows that the energy of an oscillator is not zero even in the lowest vibration state ($n = 0$) and has a value $\frac{1}{2}\hbar\omega_s(\mathbf{k})$. It indicates that even the lowest state is not vibrationless. This is known as zero point motion and finds its interpretation only with the use of quantum mechanics. The quantity $\frac{1}{2}\hbar\omega_s(\mathbf{k})$ is called the *zero point energy*.

We must appreciate that phonon is the particle name of a quantized lattice vibration. In this nomenclature, the vibration spectra of crystals are frequently addressed as the phonon spectra. In

the quantum picture, its momentum is expressed as $\hbar\mathbf{k}$. The phonon momentum is often required for the description of several physical properties of solids.

4.8 MEASUREMENT OF PHONON DISPERSION BY INELASTIC NEUTRON SCATTERING

The experiments performed to obtain the phonon dispersion curves are based on the exchange of energy between lattice vibrations and a probe. In principle, x-rays or thermal neutrons can be used as the probe in these experiments. The energy of thermal neutrons is ~ 0.025 eV, which is of the order of their average thermal energy ($k_B T$) at room temperature. Because of small energy, neutrons show an appreciable change in their energy on being scattered while exchanging the energy of a vibrational quantum. This change in energy is easily measurable on the respective energy scale. On the other hand, the energy of x-rays is much high (~ keV). The energy change for such an energetic photon is only 1 part in 10^6 which is difficult to measure on the scale of the photon energy. On account of this reason, thermal neutrons, also known as slow neutrons, are preferred for the measurement of phonon spectra. The assembly of apparatus used for this purpose is called a triple-axis spectrometer. Its schematic diagram is shown in Fig. 4.11. The experiment is generally performed around a nuclear reactor which acts as the source of thermal neutrons. The term *slow neutrons* used for these neutrons is only relative on the energy scale as the typical order of their velocity is 10^5 cm s^{-1}, which is large compared to the motion of material objects in our daily life.

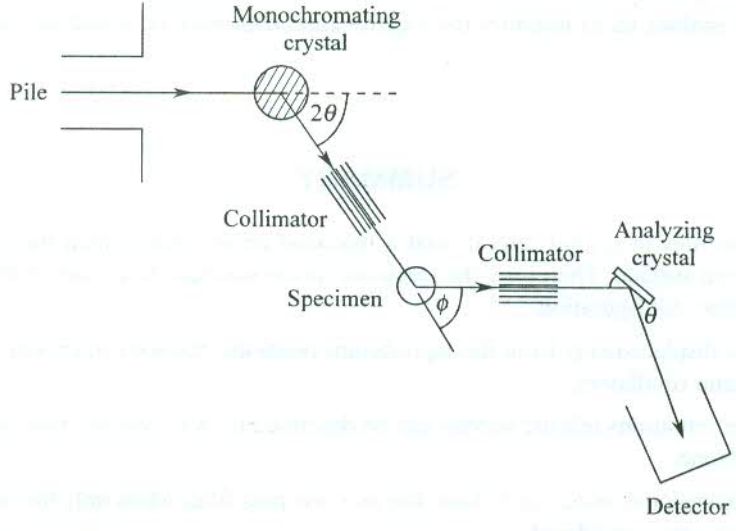


FIG. 4.11 Outlines of the experimental set-up for measuring phonon spectra using a triple-axis neutron spectrometer.

The information from the inelastically scattered neutrons is extracted mainly by exploiting the consequences of the momentum and the energy conservation laws. If a neutron of a wavevector $\mathbf{K}$ is scattered to a state of $\mathbf{K}'$ and a phonon of wavevector $\mathbf{k}_s$ is created or destroyed in the process, the momentum conservation requires that

$$\mathbf{K} - \mathbf{K}' = \pm\, \mathbf{k}_s + \mathbf{g} \qquad (4.42)$$

where (+) and (–) signs refer respectively to the creation and destruction of a phonon of energy $\hbar\omega_s(\mathbf{k})$. The symbol s stands for the branch.

The reciprocal lattice vector $\mathbf{g}$ appears for the reason that phonons are plane waves that are modified for a crystalline medium on the ground of its periodic character. The value of $\mathbf{k}_s$ in (4.42) lies within the first Brillouin zone.

According to the principle of energy conservation,

$$\frac{\hbar^2\mathbf{K}^2}{2m_n} = \frac{\hbar^2\mathbf{K}'^2}{2m_n} + \hbar\omega_s(\mathbf{k}) \tag{4.43}$$

where m_n is the neutron mass.

The monochromator crystal which rotates about the first axis of the triple axis spectrometer fixes the energy and direction of the incident neutron beam and hence the value of $\mathbf{K}$. The second axis is used to rotate the crystal. This rotates the reciprocal lattice relative to $\mathbf{K}$. For each position of the crystal, the energy of the scattered neutron is obtained by rotating the analyzer crystal about the third axis. The analyzer crystal then gives the direction and magnitude of $\mathbf{K}'$. If we take the origin of the reciprocal lattice and draw two vectors $\mathbf{K}$ and $\mathbf{K}'$, the difference $(\mathbf{K} - \mathbf{K}')$ gives the wavevector of the phonon responsible for the scattering. But this vector may lie outside the first Brillouin zone. By convention we take the phonon wavevector to be within the first Brillouin zone. Hence, we add a reciprocal lattice vector $\mathbf{g}$ to the phonon wavevector $\mathbf{k}_s$ to get it in the first Brillouin zone (4.42). The difference in the incoming and scattered neutron energies gives $\hbar\omega$, where ω is the frequency of phonon of wavevector $\mathbf{k}_s$. This leads to the determination of the dispersion. An elegant treatment of the theory that enables us to interpret the experimental data may be found in the book authored by David Pines*.

SUMMARY

1. Ionic velocities in crystals are at least a thousand times smaller than the electron velocity at the Fermi surface. Therefore, the electrons can be assumed to remain in their ground state for an ionic configuration.
2. For small displacements from the equilibrium positions, the ions in crystals may be treated as harmonic oscillators.
3. All lattice vibrations (elastic waves) can be described by wavevectors that lie within the first Brillouin zone.
4. A one-dimensional monatomic chain acts as a low-pass filter when only the nearest neighbour interactions are considered.
5. A one-dimensional diatomic chain acts as a band-pass filter by virtue of a frequency gap occurring between the acoustical and optical branches.

* David Pines, *Elementary Excitations in Solids* (Benjamin, 1963).

6. The reststrahlen frequency for an ionic crystal containing ions of masses M and m is

$$\omega_0^2 = 2f\left(\frac{1}{M} + \frac{1}{m}\right)$$

where f denotes the force constant.

7. For a crystal with p atoms in the primitive cell:

Number of acoustical branches = 3(independent of p)
= 1(LA) + 2(TA)

Number of optical branches = $3(p - 1)$
= $(p - 1)$ . . . (LO) + $2(p - 1)$ . . . (TO)

8. A quantized crystal vibration or the quantum unit of a crystal vibration is called a *phonon.* Its energy is given by $\hbar\omega$, where ω is the angular frequency of a crystal vibration.

9. The thermal energy of a harmonic crystal is given by

$$E = \sum_{\mathbf{k}s}\left(n_{\mathbf{k}s} + \frac{1}{2}\right)\hbar\omega_s\ (\mathbf{k})$$

where $n_{\mathbf{k}s}$ denotes the number of phonons with frequency ω and wavevector $\mathbf{k}$ in the branch s; and $\frac{1}{2}\hbar\omega_s(\mathbf{k})$ is the zero point energy.

10. When in a crystal a neutron of wavevector $\mathbf{K}$ is inelastically scattered to a state of wavevector $\mathbf{K}'$ and a phonon of wavevector $\mathbf{k}_s$ is created or destroyed in the process, the momentum conservation requires that

$$\mathbf{K} - \mathbf{K}' = \pm\,\mathbf{k}_s + \mathbf{g}$$

where (+) and (–) signs refer respectively to the creation and destruction of a phonon of energy $\hbar\omega_s(\mathbf{k})$; and $\mathbf{g}$ is a reciprocal lattice vector.

PROBLEMS

4.1 Localized vibrations in a crystal can be represented by a superposition of phonon modes with different wavevectors. Show that the centre of gravity of such a wave packet moves with the group velocity

$$\mathbf{v}_g = \frac{d\omega}{d\,\mathbf{k}}$$

4.2 In a linear monatomic chain of four atoms, the end atoms are fixed. Considering only the nearest neighbour interaction and assuming that the force between any two atoms is proportional to their relative displacement, set up the equations of motion for longitudinal vibrations of the free atoms. Show that the frequencies of the two normal modes are related as

$$\omega_1^2 = 3\,\omega_2^2$$

4.3 Calculate the minimum wavelength of a wave travelling along the (i) [100] direction, (ii) [111] direction in an FCC crystal. Take the lattice constant as 4.08 Å.

4.4 Show that the total momentum of the vibrating linear chain considered in Section 4.2 is given by

$$\mathbf{p}(\mathbf{k}) = -i\omega M\, \mathbf{s}_0 \exp(-i\omega t) \sum_{n=1}^{N} \exp(ikna)$$

and hence that with periodic boundary conditions $\mathbf{s}_{N+n} = \mathbf{s}_n$, $\mathbf{p}(\mathbf{k}) \equiv 0$ for $\mathbf{k} \neq 0$, i.e. a phonon carries no momentum. For $\mathbf{k} = 0$, $\mathbf{p}(\mathbf{k}) \neq 0$ if $\lim_{\omega \to 0} (\omega s_0) \neq 0$, describe the motion of the linear chain.

4.5 In a linear chain of alternate atoms having mass m_1 and m_2, suppose that only nearest neighbours interact.

(a) Show that the dispersion relation for the normal mode is

$$\omega^2 = \frac{f}{m_1 m_2}\left(m_1 + m_2 \pm \sqrt{m_1^2 + m_2^2 + 2m_1 m_2 \cos ka}\right)$$

(b) Discuss the nature of the normal modes and the form of the dispersion relation if $m_1 >> m_2$.

(c) How does the dispersion relation compare with that of the monatomic chain when $m_1 \simeq m_2$?

4.6 In a linear chain all the atoms have equal mass but are connected alternatively by springs of force constants, f_1 and f_2. Derive the frequency–wavevector relation for this chain. Are there still two branches? Explain.

4.7 In Problem 4.6, if ω_1, ω_2 and ω_3 are, respectively, the maximum frequency of the acoustic branch and minimum and maximum frequencies of the optical branch, show that for $f_1 > f_2$,

$$\omega_1 = \sqrt{\frac{2f_2}{M}}, \quad \omega_2 = \sqrt{\frac{2f_1}{M}} \quad \text{and} \quad \omega_3 = \sqrt{\frac{2(f_1 + f_2)}{M}}$$

4.8 The lattice constant of NaCl is 5.6 Å and its Young's modulus in the [100] direction is 5×10^{10} N m^{-2}. Calculate the wavelength at which the e.m. radiation is strongly reflected from a NaCl crystal. State the assumptions under which your calculation is valid. Take the atomic weights of Na and Cl as 23 and 37 respectively.
(*Hint:* Young's modulus = fla).

4.9 Consider a longitudinal wave

$$\mathbf{s}_n = \mathbf{s}_0 \cos(\omega t - nka)$$

which propagates in a monatomic linear lattice of atoms of mass M, spacing a and the force constant f.

(a) Show that the total energy of the wave is

$$\varepsilon = \frac{1}{2} M \sum_n \left(\frac{d\mathbf{s}_n}{dt} \right)^2 + \frac{1}{2} f \sum_n (\mathbf{s}_n - \mathbf{s}_{n+1})^2$$

where n runs over all atoms.

(b) By substituting $\mathbf{s}_n$ in this expression and using the dispersion relation (4.4), show that the time-average total energy per atom is

$$\frac{1}{4} M\omega^2 s_0^2 + \frac{1}{2} f(1 - \cos ka)\, s_0^2 = \frac{1}{2} M\omega^2 s_0^2$$

SUGGESTED FURTHER READING

Ashcroft, N.W. and N.D. Mermin, *Solid State Physics* (Saunders College, 1988).

Bottger, H., *Principles of the Theory of Lattice Dynamics* (Physik Verlag, 1983).

Ghatak, A.K. and L.S. Kothari, *An Introduction to Lattice Dynamics* (Addison-Wesley, 1972).

Maradudin, A.A., E.W. Montroll, G.H. Weiss, and I.P. Ipatova, *Theory of Lattice Dynamics in the Harmonic Approximation,* 2nd ed. (Academic Press, 1971).

Ziman, J.M., *Principles of the Theory of Solids* (Cambridge, 1972).

Chapter 5

Thermal Properties of Solids

In this chapter we discuss the contribution of lattice vibrations to the thermal properties of solids. For our convenience, we classify the properties into two groups:

1. Properties whose broad features can be explained in the harmonic approximation.
2. Properties whose broad features cannot be explained without considering the anharmonic effects.

From group 1 we choose 'heat capacity' and from group 2, 'thermal conductivity' and 'thermal expansion' for a proper treatment of the thermal properties of solids.

5.1 CLASSICIAL LATTICE HEAT CAPACITY

The major contribution to the heat capacity of solids comes from lattice vibrations. In non-magnetic insulators, lattice vibrations are the only contributors. Other contributions come from conduction electrons in metals and the ordering in magnetic materials. In this chapter, as already stated, we deal with lattice vibrations only.

In classical theory, the energy of a solid can be expressed as

$$\varepsilon = \varepsilon_s + \varepsilon_v \tag{5.1}$$

where

ε_s is the energy of the static lattice

ε_v is the vibration energy of the solid.

The classical equipartition energy of vibration of an atom is $k_B T$ at temperature T, k_B being the Boltzmann constant. This gives the total vibrational energy $3Nk_B T$ if there are N atoms in the solid. Substituting this in (5.1), we get

$$\varepsilon = \varepsilon_s + 3Nk_B T \tag{5.2}$$

At $T = 0$, $\varepsilon = \varepsilon_s$ which does not include zero point energy, as in classical physics there is no concept of zero point motion. The heat capacity is defined as (the value at constant volume)

$$C_V = \left(\frac{\partial \varepsilon}{\partial T}\right)_V \tag{5.3}$$

Using (5.2), we obtain

$$C_V = 3Nk_B \tag{5.4}$$

The experimentally measured value of heat capacity is referred to as the heat capacity at constant pressure C_P. But for a harmonic solid $C_V = C_P$, since the difference between the two is known to depend on the square of the temperature coefficient of linear expansion which is zero for a harmonic crystal. It will be shown later that the expansion of solids occurs because of the anharmonic motion of atoms.

The equation for heat capacity as given by (5.4) was originally derived by Dulong and Petit in the year 1869. This is now known as the Dulong–Petit law. According to this law, heat capacity is constant for a solid and independent of temperature. The measured value of heat capacity approaches this value at high temperatures and remains almost constant thereafter in the solid state (Fig. 5.1). This is how Dulong and Petit explained the behaviour of solids at high temperatures. The Dulong–Petit law, however, fails to explain the variation of heat capacity with temperature. Most significantly, the heat capacity drops to zero as the temperature approaches 0 K. But by no means the limited success of the Dulong–Petit law can be considered a mean achievement because the behaviour it fails to account for is explained only with the use of quantum theory which came much later.

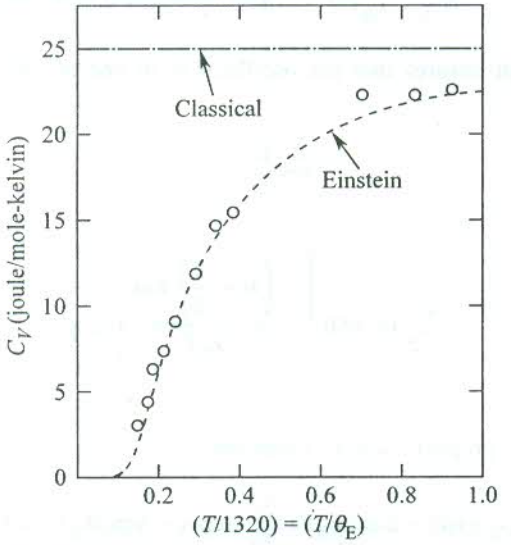


FIG. 5.1 A comparison of the experimental curve (represented by the circles) with the Einstein's theoretical curve (dotted) for molar heat capacity of diamond with θ_E = 1320 K. The horizontal dotted line gives the classical Dulong–Petit value. [After A. Einstein, *Ann. Physik,* **22**, 180 (1907).]

5.2 QUANTUM THEORY OF LATTICE HEAT CAPACITY

In this section we build up the structure for calculating lattice heat capacity capitalizing on the principles of quantum theory. Since lattice vibrations are pictured as quantized in this theory, the term phonon heat capacity sounds more appropriate than the conventional term. The calculation of phonon heat capacity is concerned mainly with the evaluation of the average thermal energy, which in turn requires the knowledge of the vibration spectrum of the solid under study. Since we have decided to work within the harmonic approximation, our first attempt would be to calculate the average thermal energy of a harmonic oscillator and then see how to proceed further.

5.2.1 Average Thermal Energy of a Harmonic Oscillator

The energy levels of a quantum harmonic oscillator as already defined in Chapter 4 are given by

$$\varepsilon_n = \left(n + \frac{1}{2}\right)\hbar\omega \tag{5.5}$$

where ω is the angular frequency of the oscillator and other indices are ignored. The average thermal energy of the oscillator ε in thermal equilibrium at temperature T is written as

$$\varepsilon = \sum_n \mathrm{p}_n\, \varepsilon_n \tag{5.6}$$

where p_n represents the probability of finding the oscillator in the energy level ε_n. It is given by

$$\mathrm{p}_n \propto \exp\left(-\frac{\varepsilon_n}{k_\mathrm{B}T}\right) \qquad \text{(the Boltzmann distribution)} \tag{5.7}$$

The proportionality constant ensures that the oscillator is in one of allowed levels. Therefore,

$$\sum_{n=0}^{\infty} \mathrm{p}_n = 1 \tag{5.8}$$

or

$$\sum_{n=0}^{\infty} \mathrm{p}_0 \exp\left[-\frac{\left(n + \frac{1}{2}\right)\hbar\omega}{k_\mathrm{B}T}\right] = 1$$

where p_0 is the constant of proportionality. Therefore,

$$\mathrm{p}_0 \exp\left(-\hbar\omega/2k_\mathrm{B}T\right) \sum_{n=0}^{\infty} \exp\left(-\hbar\omega n/k_\mathrm{B}T\right) = 1$$

or

$$\mathrm{p}_0 = \exp\left(\frac{\hbar\omega}{2k_\mathrm{B}T}\right) \cdot \left(1 - \exp\left(-\frac{\hbar\omega}{k_\mathrm{B}T}\right)\right) \tag{5.9}$$

Therefore,

$$\mathrm{p}_n = \exp\left(-n\hbar\omega/k_\mathrm{B}T\right)\left(1 - \exp\left(-\hbar\omega/k_\mathrm{B}T\right)\right) \tag{5.10}$$

Substituting p_n from (5.10) in (5.6), we get

$$\varepsilon(\omega\, T) = \left[1 - \exp\left(-\hbar\omega/k_\mathrm{B}T\right)\right]\hbar\omega \sum_{n\,=\,0}^{\infty} \left(n + \frac{1}{2}\right) \exp\left(-\hbar\omega n/k_\mathrm{B}T\right)$$

Putting $\exp(-\hbar\omega/k_BT) = x$, we obtain

$$\varepsilon(\omega, T) = (1-x)\,\hbar\omega\left[\sum_{n=0}^{\infty}(nx^n) + \frac{1}{2(1-x)}\right] \tag{5.11}$$

$$= \frac{1}{2}\hbar\omega + (1-x)\hbar\omega\sum_{n=0}^{\infty} nx^n$$

Since

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \tag{5.12}$$

on differentiating (5.12) and then multiplying both sides by x, we get

$$\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2} \tag{5.13}$$

From (5.13) and (5.11), we have

$$\varepsilon(\omega, T) = \left(\frac{1}{2} + \frac{1}{\exp(\hbar\omega/k_BT) - 1}\right)\hbar\omega \tag{5.14}$$

or

$$\varepsilon(\omega, T) = \left(\langle n\rangle + \frac{1}{2}\right)\hbar\omega \tag{5.15}$$

where

$$\langle n\rangle = \frac{1}{\exp(\hbar\omega/k_BT) - 1} \tag{5.16}$$

The comparison of (5.15) with (4.41) shows that $\langle n\rangle$ must denote the phonon occupancy, also known as the Planck's distribution function obtained by Planck for the radiation oscillators. The above exercise is a simple demonstration of the fact that lattice vibrations are quantized the same way as the radiation oscillators. $\langle n\rangle$ stands for the occupancy of phonons in the energy level ε_n which is also interpreted as the expected number of phonons in the energy state ε_n of an oscillator in thermal equilibrium at temperature T.

The calculation of total thermal energy of a solid is, however, not straightforward. Practically, it is done by making use of dispersion curves which provide the valuable information on the number of modes per unit frequency interval. The term giving this information is known as the density of states. It is a concept of central importance in solid state physics as it is concerned with the theory of several forms of excitations. We deliberately postpone for a while the proper treatment of the density of states to discuss a model in which the total thermal energy is obtained by an alternative method based on some extreme assumption. The oversimplification of the theory surprisingly does not prove to be of any serious consequence as the results of the calculation explain most of the qualitative features of phonon heat capacity. This model, proposed by Einstein, serves as a landmark in the progress of theory of phonon heat capacity.

5.2.2 Einstein Model

Einstein assumed that all the atoms in a solid (say N) vibrate with the same frequency independently. In other words, the frequency of vibration for all the $3N$ normal modes is common and the atoms oscillate independent of each other. Einstein treated the atomic harmonic oscillators as similar to the radiation oscillators of the Planck's theory of blackbody radiation and carried out the quantization of energy levels on parallel lines:

$$\varepsilon_n = n\hbar\omega \qquad (n = 0, 1, 2, 3, \ldots) \tag{5.17}$$

where ε_n represents the vibrational energy of the oscillator in its nth state of excitation. According to (5.15), the average thermal energy of the solid for a single polarization of the normal mode is equal to $N\hbar\omega\langle n\rangle$. The factor $\frac{1}{2}\hbar\omega$ for the zero point energy was not known to Einstein as this factor arises on the requirement of quantum mechanics which was formulated about twenty years later. Moreover, its omission does not affect the value of heat capacity since $C_V = (\partial\varepsilon/\partial T)_V$ and the omitted factor is not a function of temperature.

Therefore, in the Einstein model the thermal energy (or internal energy) of a solid containing N atoms is expressed as

$$= 3N\hbar\omega\langle n\rangle \tag{5.18}$$

which includes the vibrational energy in all the three polarizations of the normal mode.

The phonon heat capacity is given by

$$C_V = \left(\frac{\partial\varepsilon}{\partial T}\right)_V = 3\,Nk_B\left(\frac{\hbar\omega}{k_BT}\right)^2 \frac{\exp\left(\dfrac{\hbar\omega}{k_BT}\right)}{\left[\exp\left(\dfrac{\hbar\omega}{k_BT}\right) - 1\right]^2} \tag{5.19}$$

The plot of the above equation for diamond is shown in Fig. 5.1 which also shows the measured values at different temperatures. It shows a good agreement between the theoretical and experimental curves at high temperatures, and the value of heat capacity approaches its value in the classical limit. The two curves tend to be separated at low temperatures. An overall view indicates that there is an agreement between the theory and the experiment for most of the temperature range excepting the low temperatures. The value of C_V in the limit of high temperatures as evaluated from (5.19) is $3Nk_B$, the classical value given by the Dulong and Petit law. In the low temperatures limit it is found to decrease as $\exp(-\hbar\omega/k_BT)$ and drops to zero at $T = 0$ K. The experimental value also approaches zero as $T \to 0$ K. Therefore, the biggest achievement of the Einstein model lies in its ability to demonstrate quickly that the heat capacity drops to zero at 0 K. However, the experimental curve (see Fig. 5.1) shows that the drop is faster than that of the exponential. The curve fits with a T^3 dependence in the low temperature region. For every solid there is a characteristic temperature above which the model works and well below this temperature the model fails. This temperature is known as Einstein characteristic temperature θ_E and defined by

$$\theta_E = \frac{\hbar\omega_E}{k_B} \tag{5.20}$$

where ω_E is called the Einstein characteristic frequency.

The failure of the Einstein model at low temperatures exposes its inadequacy. The model is physically unrealistic because all vibrational modes cannot have a common frequency unless all the atoms in the solid vibrate in a manner independent of each other. The movements of adjacent atoms always tend to be correlated because of the coupling forces within the solid. Therefore, the drawback lies with the basic assumption on which the model is based. In an effort to explain the low temperature behaviour of phonon heat capacity, Debye put forward another approach in which he used an interpolation scheme to calculate the effective contribution from the frequency spectrum of a solid. The Debye model enjoys an overall success with its special achievement in the form of the Debye T^3–law that fits the low temperature heat capacity data reasonably well.

5.2.3 Phonon Density of States

As emphasized at the end of Section 5.2.1, the knowledge of density of states of normal modes is a prerequisite to the determination of thermal energy of a solid. Consider a cube of a periodic solid of side L. Let there be N^3 primitive cells within the volume of the cube. On applying the periodic boundary condition [given by (4.5)] over these cells for a three-dimensional crystal, we have

$$\exp i(k_x x + k_y y + k_z z) = \exp i[k_x(x + L) + k_y(y + L) + k_z (z + L)] \tag{5.21}$$

where (x, y, z) are the Cartesian coordinates of any lattice site. Equation (5.21) is valid for the following restricted set

$$k_x, k_y, k_z = 0; \pm \frac{2\pi}{L}; \pm \frac{4\pi}{L}; \cdots; \frac{N\pi}{L} \tag{5.22}$$

analogous to the one expressed by (4.9) for a one-dimensional crystal.

The set (5.22) shows that there lies one allowed value of k within a volume of $(2\pi/L)^3$ in k-space. Thus the number of allowed k-values per unit volume of k-space is

$$\left(\frac{L}{2\pi}\right)^3 = \frac{V}{(2\pi)^3} \tag{5.23}$$

where V denotes the volume of the crystal. If we represent the density of the allowed k-values or the density of phonon modes by $D(\omega)$, the number of modes lying within the interval of frequency ω and $\omega + d\omega$ is given by

$$D(\omega)d\omega = \frac{V}{(2\pi)^3} \cdot (\text{volume enclosed between } k\text{-surfaces, } \omega(\mathbf{k}) = \text{constant and } \omega(\mathbf{k}) + d\omega = \text{constant})$$

$$= \frac{V}{(2\pi)^3} \int_{\omega}^{\omega + d\omega} d^3 k \tag{5.24}$$

Let ds_ω denote a small surface element of $\omega(\mathbf{k})$ = constant and $d\mathbf{k}_\perp$ be a vector normal to it (Fig. 5.2) with its tip touching the surface, $\omega(\mathbf{k}) + d\omega$ = constant. Then

$$D(\omega)\, d\omega = \frac{V}{(2\pi)^3} \int_{\omega(\mathbf{k}) = \text{const.}} ds_\omega \cdot d\mathbf{k}_\perp \tag{5.25}$$

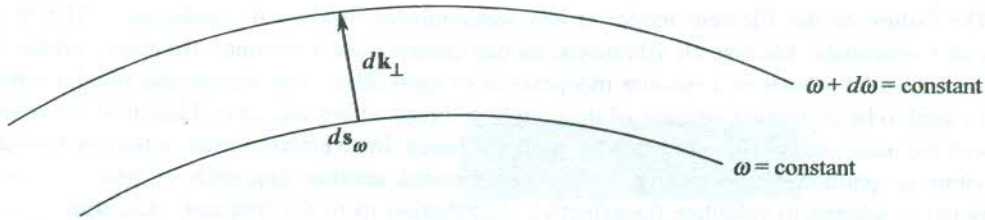


FIG. 5.2 Neighbouring constant energy surfaces $\omega(\mathbf{k})$ = constant and $\overline{\omega + d\omega}\ (\mathbf{k})$ = constant being connected by a vector $d\mathbf{k}_\perp$ normal to the surface element ds_ω of the lower energy surface.

With $d\omega = |\text{grad}_\mathbf{k}\ \omega| d\mathbf{k}_\perp$, we have

$$D(\omega)\, d\omega = \frac{V}{(2\pi)^3}\, d\omega \int_{\omega(\mathbf{k})\ =\ \text{const.}} \frac{ds_\omega}{|\text{grad}_\mathbf{k}\,\omega|} \tag{5.26}$$

and

$$D(\omega) = \frac{V}{(2\pi)^3} \int_{\omega(\mathbf{k}) = \text{const.}} \frac{ds_\omega}{|\text{grad}_\mathbf{k}\,\omega|} \tag{5.27}$$

The frequency derivative of the density of states has singularities (known as van Hove singularities) at points where the group velocity is zero or the tangents on the dispersion curve are horizontal. The one-dimensional crystal is a special case for which the density of states $D(\omega)$ itself is singular. The variation of phonon density of states with frequency for silicon is shown in Fig. 5.3.

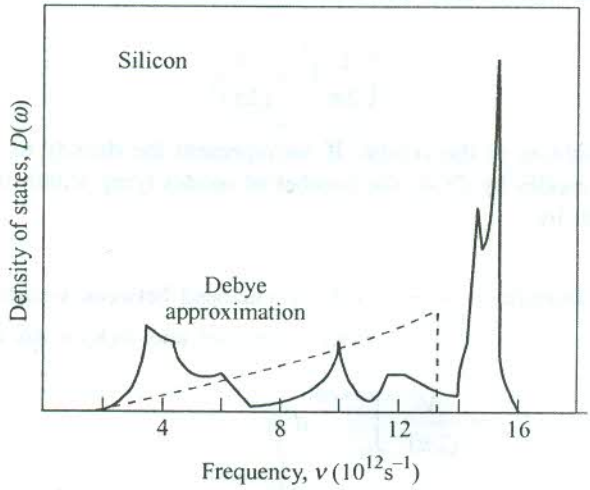


FIG. 5.3 Behaviour of the phonon density of states in silicon. The dotted curve belongs to an isotropic medium of elastic continuum for θ_D = 640 K. [After G. Dolling and R.A. Cowley, *Proc. R. Soc.*, London, **88**, 463 (1966).]

5.2.4 Debye Continuum Model

In the Debye model a crystal is treated as an isotropic elastic continuum in which the velocity of sound is constant ($c = \omega/k$). It takes into consideration only the three acoustical branches with different sound velocities for longitudinal and transverse waves. In this case the surface $\omega(\mathbf{k}) = \text{const.}$ is spherical and the velocity $|\text{grad}_{\mathbf{k}}\ \omega|$ is equal to the sound velocity. Let c_i denote the sound velocity in the branch i. Being independent of $\mathbf{k}$, it can be taken outside the integral of the relation (5.27) and the surface integral becomes

$$\int_{\omega(\mathbf{k}) = \text{const.}} ds_\omega = 4\pi k^2 \tag{5.28}$$

Using (5.28), the relation for the density of states (5.27) is put down as

$$D_i(\omega) = \frac{V}{2\pi^2} \cdot \frac{k^2}{c_i}$$

or

$$D_i(\omega) = \frac{V}{2\pi^2} \cdot \frac{\omega^2}{c_i^3} \tag{5.29}$$

The total phonon density of states is then given by

$$D(\omega) = \frac{V}{2\pi^2} \left(\frac{1}{c_L^3} + \frac{2}{c_T^3} \right) \omega^2 \tag{5.30}$$

where c_L and c_T are the velocities of longitudinal and transverse waves respectively.

Debye put a cut-off at the upper limit of frequency to ensure that the waves are long enough to have a smooth passage through the crystal which could then be justifiably treated as an elastic continuum. The cut-off frequency is denoted by ω_D. The density of states function of this model is plotted as a function of frequency for silicon in Fig. 5.3. At ω_D it drops from the maximum to the zero value giving a sharp cut-off. The total thermal energy of the crystal in thermal equilibrium, i.e. the internal energy $U(T)$ may be written as

$$U(T) = \int_0^{\omega_D} D(\omega)\, \varepsilon(\omega, T)\, d\omega \tag{5.31}$$

where $\varepsilon(\omega, T)$ is the average thermal energy stored in the mode of frequency ω at temperature T.

Substituting for $D(\omega)$ from (5.30) in (5.31), we have

$$U(T) = \frac{V}{2\pi^2} \left(\frac{1}{c_L^3} + \frac{2}{c_T^3} \right) \int_0^{\omega_D} \omega^2 \varepsilon(\omega, T)\, d\omega \tag{5.32}$$

Fixing the cut-off at a common frequency for longitudinal and transverse waves is not supported

by reason. But, to our surprise it gives a better agreement with the measured value of heat capacity than when separate cut-off values are chosen for calculations.

The phonon heat capacity may be expressed as

$$C_V(T) = \left(\frac{\partial U}{\partial T}\right)_V = \frac{V}{2\pi^2}\left(\frac{1}{c_L^3} + \frac{2}{c_T^3}\right)\int_0^{\omega_D} \omega^2 \frac{\partial}{\partial T}\,\varepsilon(\omega, T)\,d\omega \tag{5.33}$$

We determine the Debye cut-off frequency by applying the condition that the total number of modes or states be equal to $3N$, where N is the toal number of atoms in the crystal. Applying this condition, we have

$$\int_0^{\omega_D} D(\omega)\,d\omega = 3N \tag{5.34}$$

or

$$\frac{V}{2\pi^2}\left(\frac{1}{c_L^3} + \frac{2}{c_T^3}\right)\int_0^{\omega_D} \omega^2 d\omega = 3N$$

or

$$\frac{V}{2\pi^2}\left(\frac{1}{c_L^3} + \frac{2}{c_T^3}\right) = \frac{9N}{\omega_D^3} \tag{5.35}$$

Using (5.35) and taking the value $\varepsilon(\omega, T)$ as $\langle n\rangle\hbar\omega$, we have from (5.33)

$$C_V = \frac{9N}{\omega_D^3}\frac{\partial}{\partial T}\int_0^{\omega_D} \frac{\hbar\omega^3}{\exp(\hbar\omega/k_B T) - 1}\,d\omega \tag{5.36}$$

The Debye continuum model is most relevant at low temperatures where only low frequencies can be excited and, therefore, waves of only long wavelength will be present. Thus a Debye characteristic temperature θ_D is defined below, for which the model is valid. It is given by

$$\theta_D = \frac{\hbar\omega_D}{k_B} \tag{5.37}$$

Using (5.37) and the substitution, $\hbar\omega/k_B T = x$ in the relation (5.36), we get

$$C_V = 9Nk_B\left(\frac{T}{\theta_D}\right)^3 \int_0^{\theta_D/T} \frac{x^4 e^x}{(e^x - 1)^2}\,dx \tag{5.38}$$

The behaviour of phonon heat capacity in the extreme limits of temperature can be deduced from (5.38) by evaluating the integral between these limits.

At high temperatures, $T >> \theta_D$

In this limit ($k_B T >> \hbar\omega_D$), C_V as expressed by (5.38) approaches the classical value $3Nk_B$.

At very low temperatures, $T << \theta_D$

The upper limit of the integral in (5.38) approaches infinity and we obtain the value of the integral as $4\pi^4/15$. This gives

$$C_V = \frac{12\,\pi^4}{5} Nk_B \left(\frac{T}{\theta_D}\right)^3 \tag{5.39}$$

The relation (5.39) expresses the Debye T^3–law that explains the temperature dependence of the experimental data at low temperatures. The variation of phonon heat capacity in the Debye model together with the experimental curve for yttrium is shown in Fig. 5.4. A comparative study in low temperature region demonstrates that the drop of heat capacity in the Debye curve is faster than that given by the Einstein curve (Fig. 5.5) and matches the experimental curve characterized by T^3 behaviour.

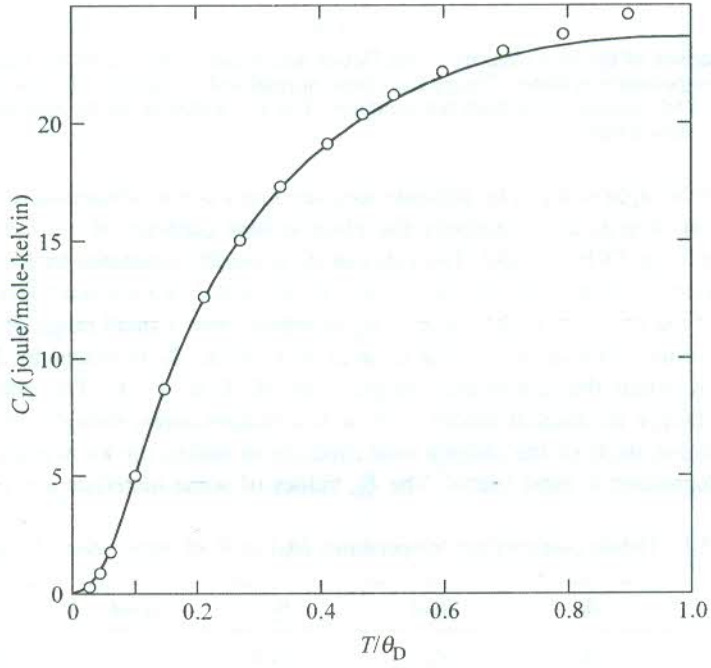


FIG. 5.4 Debye curve of heat capacity for yttrium for $\theta_D = 200$ K. The experimental points (circles) are above the Debye curve at high temperatures. For fitting points at lower temperatures, a slightly higher θ_D is required. [After L.D. Jennings, R.E. Miller, and F.H. Spedding, *J. Chem. Phys.*, **33**, 1849 (1960).]

According to the Debye model, the T^3–law is expected to be valid in the temperature range $T \leq 0.1\theta_D$. But the experimental reports suggest a much lower temperature range, $T \leq \frac{\theta_D}{50}$. Though the success of the Debye model is largely accepted, the departure from the theoretically predicted

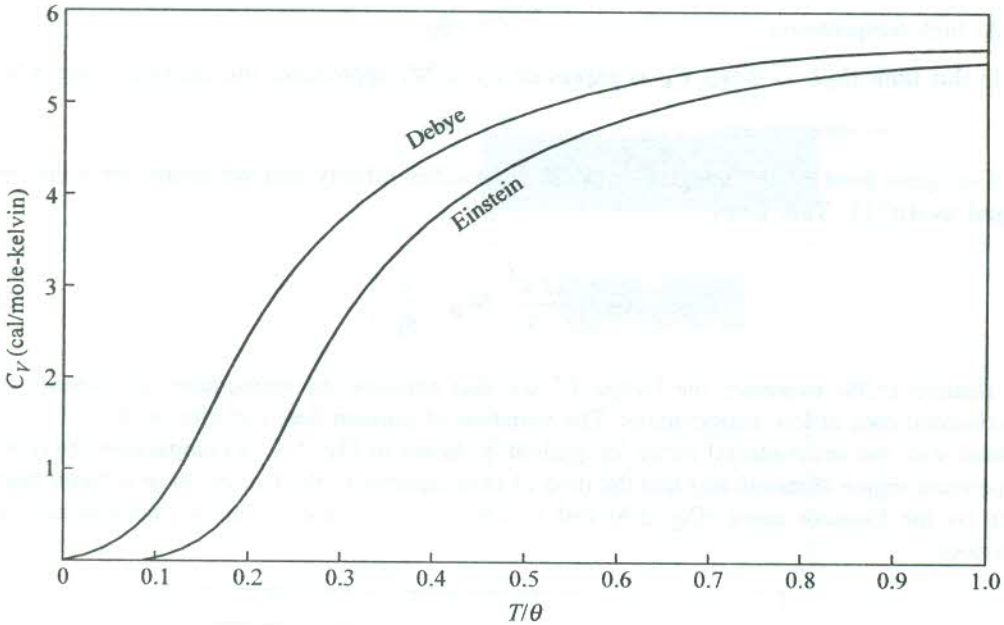


FIG. 5.5 A comparison of the heat capacity in the Debye and Einstein models. In the Debye model, the drop at low temperatures is faster. Curves have been normalized to approach the classical value (Dulong–Petit) of 5.96 cal/mole-K at high temperatures. The θ is either θ_D or θ_E depending on which curve is under consideration.

behaviour has been brought to focus by accurate measurements at low temperatures. The characteristic temperature θ_D that determines completely the phonon heat capacity of a solid plays the role of central importance in the Debye model. The value of θ_D is usually estimated by feeding the measured value of heat capacity in relation (5.39). In T^3–law, θ_D is treated as a constant which does not sound realistic because of the variation in the value of θ_D estimated over a small range of low temperatures. For example, in sodium chloride on cooling from 20 K to 10 K, θ_D increases by 20 K; in lithium it decreases by 28 K when the temperature drops from 30 K to 15 K. This data jeopardizes the soundness of the Debye continuum model even at low temperatures, though only marginally.

For a comparative study of the phonon heat capacity of solids, the knowledge of the respective characteristic temperatures is most useful. The θ_D values of some materials are given in Table 5.1.

Table 5.1 Debye characteristic temperatures (θ_D) in K of some selected materials*

Solid	θ_D	*Solid*	θ_D	*Solid*	θ_D
Li	344	Hg	71.9	C	2230
Na	158	Pb	105	Si	645
Cs	38	Fe	467	Ge	374
In	108	Ni	450	ZnS	315
Te	153	Mo	450	LiF	732
Cu	343	W	400	LiCl	422
Ag	225	Ar	93	NaCl	321
Au	165	Kr	72	KCl	235

*From *American Institute of Physics Handbook,* 3rd ed., McGraw-Hill, New York, 1972.

5.3 ANHARMONIC EFFECTS

While dealing with lattice vibrations, the harmonic approximation has been the basis of discussion so far. Notwithstanding its success in handling most of the solids at temperatures well below the melting point, it would be totally misleading to conclude that corrections to the approximation make sense only in calculations where high precision is required. Keeping aside the exceptions like solid helium in which the harmonic approximation is not applicable, we come across a number of interesting physical phenomena that cannot be explained without relaxing the approximation. Some of these are:

(i) Thermal expansion

(ii) Thermal conductivity has a finite value

(iii) The temperature dependence of elastic constants

(iv) The failure of Debye heat capacity to attain the classical value at high temperatures ($T >> \theta_D$)

(v) The line broadening or the measurable width of one phonon peaks in the neutron inelastic scattering pattern

The cause of failure of the harmonic approximation in explaining the above phenomena rests in the questionable relevance of its two basic assumptions:

1. Vibrations of atoms may be treated as small oscillations in which the displacement of atoms from their respective equilibrium positions are small.
2. Only the leading non-vanishing term in the expansion of potential energy about its equilibrium value is retained.

The validity of the first assumption is always questionable at high temperatures where the amplitude of vibration is large. And when the displacements become large, we can ill-afford to ignore the higher order terms in displacement beyond the quadratic term in the potential energy expansion. In this situation both the assumptions become irrelevant. This is indicated by the shape of the potential energy curve at large interatomic separations (Fig. 5.6) where it is more asymmetric, showing thereby deviation from the harmonic nature. This is understandable as atoms can never oscillate like independent harmonic oscillators in a real crystal because the motion of adjacent atoms is always correlated. In general, we must accept the presence of anharmonicity to a certain degree in the vibrations of a solid. The physical phenomena stated earlier are explained successfully when corrections on the requirement of anharmonicity are included in the calculations.

The anharmonic effects are most reasonably accounted when the potential energy expansion is truncated not before the quartic term. Dropping the constant equilibrium term, we write the potential energy as

$$U(x) = fx^2 - gx^3 - hx^4 \tag{5.40}$$

where x is the deviation from the equilibrium separation at absolute zero and all the coefficients f, g and h are positive.

The first term in (5.40) is the usual harmonic component and the other two refer to the anharmonic effects. The cubic term stands for the asymmetry of the mutual repulsion of the atoms and the quartic term represents the softening of the vibration at large amplitudes.

As the applications of the anharmonic potential, we will deal with the first two of the phenomena cited earlier as examples. While the thermal expansion refers to an equilibrium property, the thermal conductivity is a well-known transport phenomenon.

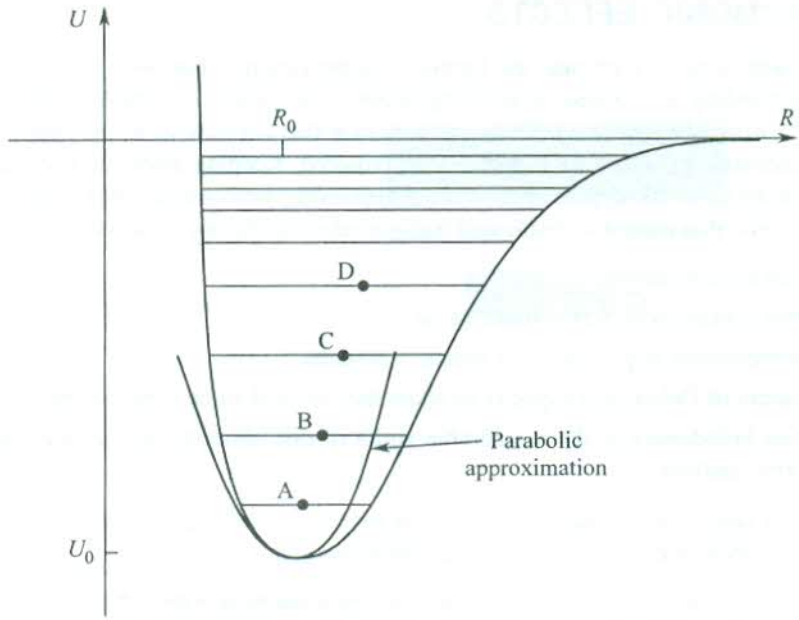


FIG. 5.6 An exaggerated potential energy curve based on Lennard–Jones potential for real crystals. The asymmetrical nature of the curve at higher energies does not allow the motion to be interpreted in parabolic approximation. The increase in the interatomic separation R at high energies (excited at high temperatures) produces thermal expansion. The R_0 is the interatomic distance in the ground state. Because of anharmonicity, the vibration quantum (the spacing between vibrational levels) does not remain constant. It decreases with increase in energy.

5.3.1 Thermal Expansion

It is an established fact that solids expand on heating. In the harmonic approximation, all atoms vibrate about their equilibrium position within a perfectly symmetric well (parabola) of interatomic interaction. But the potential energy curve (see Fig. 5.6), as derived from a Lennard–Jones potential type for real solids, matches this behaviour only in the range of low thermal excitations that correspond to low temperatures. We may thus conclude that the harmonic approximation is strictly valid only at low temperatures.

The potential well gets asymmetrical for larger interatomic separations (see Fig. 5.6) occurring at higher energies. The mean interatomic separations at a few vibrational energies taken in increasing order are denoted by points A, B, C, D respectively. The values represented by these points are in increasing order as shown by the trend of shift toward larger values relative to the mean equilibrium separation R_0 in the ground state. It is then imperative that at higher temperatures when higher vibrational states are sufficiently populated, solids would show expansion. The range of vibration frequencies of solids is such as can be excited by the thermal energy. These arguments explain the thermal expansion of solids. Since this property follows from the anharmonic nature of the potential energy curve, the thermal expansion is attributed to the presence of anharmonicity in atomic vibrations.

Now we use (5.40) to calculate the thermal expansion in terms of the mean displacement $\langle x \rangle$ of the atoms in a solid. It is non-zero because of the cubic term. A crude way of determining $\langle x \rangle$ is to use the condition,

$$\frac{\partial U}{\partial x} = 0$$

and ignore the term in h. This gives

$$\langle x \rangle = \frac{3g \langle x^2 \rangle}{2f} \tag{5.41}$$

If the mean square displacement is calculated classically in the harmonic approximation, we get

$$\langle x^2 \rangle = \frac{\int_{-\infty}^{+\infty} x^2 \exp(-fx^2/k_B T)\, dx}{\int_{-\infty}^{+\infty} \exp(-fx^2/k_B T)\, dx} = \frac{\frac{k_B T}{4f}\sqrt{\frac{\pi}{f/k_B T}}}{\frac{1}{2}\sqrt{\frac{\pi}{f/k_B T}}}$$

or

$$\langle x^2 \rangle = \frac{k_B T}{2f} \tag{5.42}$$

Substituting (5.42) in (5.41), we express the mean displacement as

$$\langle x \rangle = \frac{3g\, k_B T}{4f^2} \tag{5.43}$$

The relation (5.43) shows an increase in the interatomic separations and, therefore, a thermal expansion of the solid.

We may define the coefficient of linear expansion α as

$$\alpha = \frac{\partial \langle x \rangle}{\partial T} \tag{5.44}$$

Using (5.43), we get

$$\alpha = \frac{3g\, k_B}{4f^2} \tag{5.45}$$

Though changes in size because of the thermal expansion are small in solids, the knowledge of their expansion coefficients is nonetheless of great practical value in industry where even simple jobs like making a permanent joint between two materials have their tactical importance.

Another striking feature, that is associated with the thermal expansion, has reference to the change of the characteristic frequency in a solid. The levels of vibrational energy come closer at high energy values (see Fig. 5.6). In the harmonic approximation the levels are expected to remain evenly spaced and unaffected by the change of temperature. The spacing can alter only if the energy quantum $\hbar\omega$ changes. By taking anharmonic considerations into account with the use of (5.40), it can be shown that the characteristic frequency ω decreases with the increase in temperature. This is consistent with the observation of closer energy levels at high energies to which the thermal excitation can be made by raising the temperature. We must appreciate that the anharmonic contributions at these values of energy and temperature are at their maximum.

The above frequency effect may as well be interpreted as the change in characteristic frequencies in a solid because of the change in its volume. It is conventionally put in the form of an assumption

that allows both the characteristic frequencies and the volume V to suffer the same relative change. That is,

$$\frac{\Delta\omega}{\omega} \propto \frac{\Delta V}{V}$$

or

$$\frac{\Delta\omega}{\omega} = \gamma \frac{\Delta V}{V} \tag{5.46}$$

where the proportionality constant γ is called the Grüneisen constant. It is a measure of the anharmonic coupling.

This anharmonic effect, though small, may be studied by neutron scattering measurements that provide the change in characteristic frequencies with temperature.

An exact treatment of the anharmonic effects is unfortunately not easy. We have to take recourse to some approximate method like perturbation method of quantum mechanics. As a first approximation to the true solution, we start with the solutions of harmonic potential (the phonons). Obviously, the phonons are not the exact eigensolutions to the equations of motion. The description of a state of motion in an anharmonic solid in terms of a phonon or a plane wave changes with time and the accuracy of the description drops progressively. For an accurate description with time, it is required to recognize the status of a spectrum of some other phonons. It means that the anharmonic phonons, unlike the harmonic ones, have only a limited lifetime after which they merge or decay to produce new phonons. In the next section, we discuss these phonon processes as they enable us to explain the basis for thermal resistance and related features.

5.3.2 Phonon Collision Processes

An exact quantum mechanical treatment based on the first order perturbation theory shows that:

(a) The cubic term in the potential energy accounts for the following processes:
  (i) one phonon decays into two
  (ii) two phonons merge into one

(b) The quartic term in the potential energy accounts for the following processes:
  (i) one phonon decays into three
  (ii) two phonons get converted into two others
  (iii) three phonons merge into one

We observe that the cubic term is related to the three-phonon processes and the quartic to the four-phonon processes. It is now necessary to examine the probable collision processes, especially in the context of thermal resistance which is on our plan for discussion. Let us consider the collision involving three phonons. The use of the term collision should not be objectionable as phonons are treated here as particles. If two phonons of wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$ collide or merge to produce a new phonon with wavevector $\mathbf{k}_3$, the momentum conservation requires that

$$\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3 \tag{5.47}$$

The total energy of the initial phonons is completely held by the final phonon. The net phonon momentum is written as

$$\mathbf{j} = \sum_{\mathbf{k}s} n_{\mathbf{k}s}\, \hbar\mathbf{k}_s \tag{5.48}$$

where $n_{\mathbf{k}s}$ is the number of phonons with wavevector $\mathbf{k}$ in the branch s.

The net phonon momentum **j** constitutes the phonon thermal current in a solid between whose two ends a difference of temperature is maintained. In view of (5.47), the net phonon momentum is a constant and, therefore, the phonon current **j** remains unchanged. The distribution of phonons (or the distribution function) at any stage when $\mathbf{j} \neq 0$ tends to approach its equilibrium value (condition of zero temperature). But in the present case it is not permitted as **j** has to remain unaffected. Any distribution of phonons flows down the solid unhindered as if the phonon mean free path were infinite. Consequently, the thermal resistance, a measure of the rate at which the equilibrium distribution is approached, becomes zero. This gives an infinite value of thermal conductivity—an unacceptable result. Therefore, the collision processes described by (5.47) do not give rise to thermal resistance. These processes are called 'normal processes' or simply 'N-processes'. Figure 5.7(a) shows a normal process of collision in a two-dimensional square lattice.

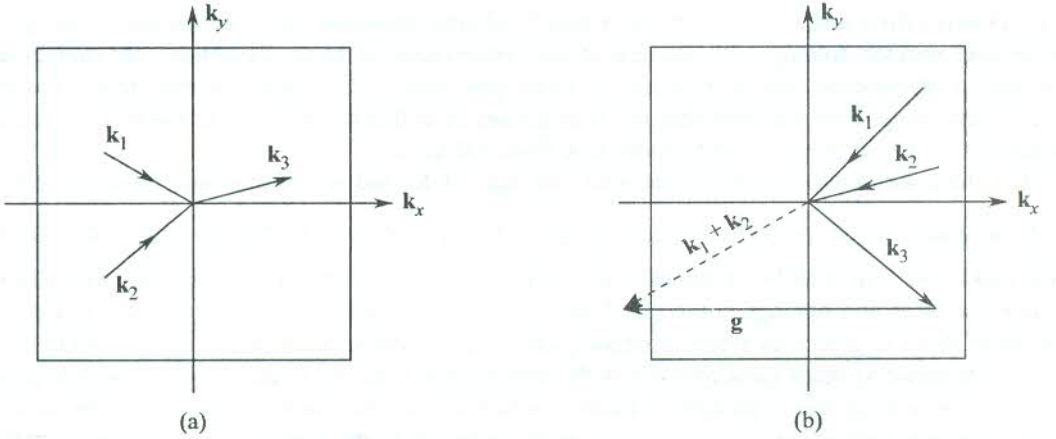


FIG. 5.7 Phonon scattering processes in k-space for a two-dimensional square lattice: (a) normal process, (b) Umklapp process—two phonons with wavevector $\mathbf{k}_1$ and $\mathbf{k}_2$ merge into $\mathbf{k}_3$ with the help of a reciprocal lattice vector $\mathbf{g}$ ($= 2\pi/a$ in this case). Note that the direction of the x-component of the phonon flux has been reversed.

It should be, however, surprising to note that N-processes explain the conduction of heat in gases where we deliberately impose the condition that no mass transport of particles takes place. In this picture, the hot particles while moving along one direction lose their energy to the cold particles moving in the opposite direction [Fig. 5.8(a)]. But these ideas are not tenable in the case of phonon gas where there is no net particle conservation in the true picture. The phonons that carry heat from the hot end are destroyed at the cold end [Fig. 5.8(b)].

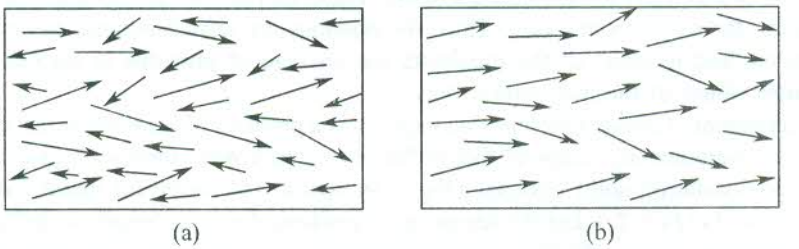


FIG. 5.8 (a) Conduction of heat in a gas: hot electrons moving in one direction lose their energy to cold electrons moving in the opposite direction. During the conduction process the number of electrons is conserved. (b) Conduction of heat by phonons in a solid: There is no net particle conservation. Phonons carrying heat from the hot end get destroyed at the cold end.

The phonons could also suffer collisions with the immobile imperfections and the crystal bounding surface. But these collisions are known to be elastic in nature, meaning thereby that the collisions do not change the frequency of the individual phonons. This makes impossible for an equilibrium distribution of phonons to be established locally.

The puzzle of thermal resistance is solved by a different kind of collisions expressed as

$$\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3 + \mathbf{g} \tag{5.49}$$

where $\mathbf{g}$ is a reciprocal lattice vector.

Since in a periodic lattice the energy of a phonon with wavevector $(\mathbf{k}_3 + \mathbf{g})$ is the same as that for a phonon with wavevector $\mathbf{k}_3$, the phonon $\mathbf{k}_3$ in (5.49) must be carrying the total energy. The special feature of (5.49) is that it destroys the momentum $\hbar\mathbf{g}$ and changes the direction of energy flow as shown in Fig. 5.7(b). The x-components of $\mathbf{k}_1$ and $\mathbf{k}_2$ are directed opposite to the x-component of $\mathbf{k}_3$. Peierls called such collisions by the name 'Umklapp processes' or 'U-processes'. Umklapp is the German term for 'folding over'. Because of non-conservation of the net momentum, the equilibrium distribution of phonons can be restored by these processes at a certain rate that determines the thermal resistance. This confirms that the U-processes of collisions give rise to thermal resistance, limiting the value of thermal conductivity to a finite value.

For the condition (5.49) to be satisfied, the sum of $\mathbf{k}_1$ and $\mathbf{k}_2$ must extend beyond the first Brillouin zone and the individual values should not be less than $\frac{1}{2}\mathbf{g}$. The sum $(\mathbf{k}_1 + \mathbf{k}_2)$ can be translated to the first zone by reciprocal lattice vector $-\mathbf{g}$ [Fig. 5.7(b)]. The $\mathbf{k}_3$ is called the reduced value of $(\mathbf{k}_3 + \mathbf{g})$ and belongs to the set of unique k-values defined by (4.9) in Chapter 4. We will talk more about U-processes while discussing the effect of temperature on thermal conductivity.

It is desirable to make some remarks on the terminology followed in the above discussion. In the refined theoretical picture, a phonon collision is referred to as the scattering of phonons because of phonon–phonon interaction. This is only to avoid confusion to the reader as we may inadvertently be using sometimes a mixed terminology.

5.3.3 Phonon Thermal Conductivity

Metals are good conductors of heat and electricity. The energy is transported mainly by free electrons. But it must not be construed that thermal conductivity depends entirely on the number of available free electrons. Some of the insulators, for example, the crystalline sapphire (Al_2O_3) and quartz (SiO_2) possess higher thermal conductivity than copper at low temperatures. The maximum value of thermal conductivity of sapphire is 200 W cm^{-1} K^{-1} as against the corresponding value, 100 W cm^{-1} K^{-1}, of copper. The observations indicate to the involvement of some other carriers in addition to the free electrons for the heat transport. These carriers are phonons as confirmed by the theory explaining the experimental data. Thus, the phonons are unambiguously the main carriers of heat in insulators, and in some of the insulators the passage of phonons is smooth enough so as to result in large values of thermal conductivity.

An exact treatment of phonon thermal conductivity is a problem of great mathematical complexity. We give below an elementary classical theory based on the kinetic theory of gases.

Consider an insulating solid cylinder (Fig. 5.9) whose two ends are maintained at different temperature T_1 and T_2 $(T_2 > T_1)$. Let the temperature gradient down its length be defined as $-\partial T/\partial x$. At steady state, there is a steady flow of heat from the hot end to the cold end maintaining the temperature difference $(T_2 - T_1)$ at a constant value. It is observed that the thermal energy flowing

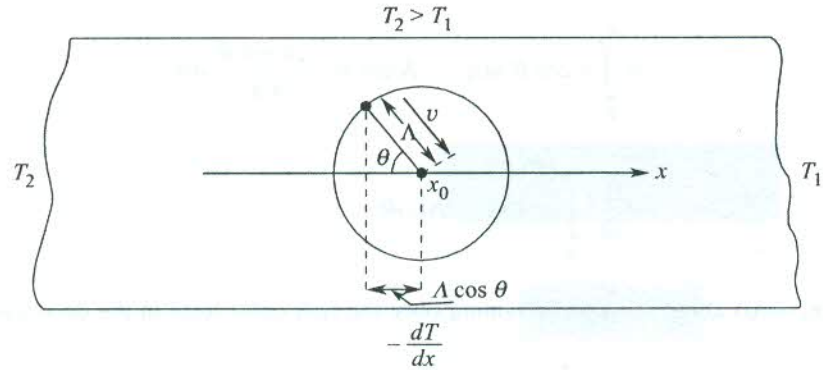


FIG. 5.9 Conduction of heat by phonons in an insulating solid cylinder along its length when a temperature gradient is maintained between the two ends of the cylinder. Phonons approaching the point x_0 at an angle θ with the cylinder's axis (x-direction) and moving with velocity v make a collision at the point. The temperature at the point is T.

down the cylinder per unit of its circular cross-section per unit time j is proportional to the temperature gradient $-\partial T/\partial x$. That is,

$$\mathbf{j} \propto -\frac{\partial T}{\partial \mathbf{x}} \tag{5.50}$$

or

$$\mathbf{j} = -K_{\text{ph}} \left(\frac{\partial T}{\partial \mathbf{x}} \right) \tag{5.51}$$

where K_{ph} is called the thermal conductivity and $|\mathbf{j}|$ is the thermal flux per unit time.

The flow of heat thus cannot be linear, as in that case $\mathbf{j}$ would have been proportional to T. The heat energy diffuses from the hot end to the cold end through a random process. As we consider here only the phonons as the heat carriers, the random process involves mainly the phonon–phonon collisions. The extra phonons created at hot end are destroyed at the cold end. This picture of phonon gas is better suited to be treated by the kinetic theory of molecular gases.

Frequent collisions among phonons limit the phonon mean free path. In a rigorous theory the mean path is treated as a function of the temperature and the phonon frequency. But we take an average value (Λ) over the length of the solid cylinder. Let the non-equilibrium energy density contributed by phonons coming out of collisions at the point x be denoted by $u(x)$. If T be the temperature at x, $u(x)$ is supposed to be proportional to the equilibrium energy density $u[T(x)]$. Then, the contribution to thermal current density j from a single phonon is $v_x u(x)$, where v_x is the x-component of the phonon velocity v.

Figure 5.9 shows that the collision at x_0 involves the phonons approaching it at angle θ with the x-axis. The immediate last collision of these phonons on the average should have occurred at a point whose x-coordinate is shorter by $\Lambda \cos \theta$, i.e. at $x = (x_0 - \Lambda \cos \theta)$. The net thermal current j can be calculated by using the fact that j is proportional to the product $v_x \cdot u(x_0 - \Lambda \cos \theta)$ averaged over all solid angles. Accordingly,

$$j = \langle v_x \cdot u(x_0 - \Lambda \cos \theta) \rangle$$

$$= \int_0^{\pi} v \cos\theta \, u(x_0 - \Lambda \cos\theta) \frac{2\pi \sin\theta}{4\pi} \, d\theta$$

$$= \frac{1}{2} \int_{-1}^{1} vtu(x_0 - \Lambda t) \, dt \tag{5.52}$$

Expanding $u(x_0 - \Lambda t)$ about $u(x_0)$ and retaining only the first order term in the derivative, we have

$$j = \frac{1}{2} \int_{-1}^{1} vt \left[u(x_0) - \Lambda t \frac{\partial u}{\partial x}\bigg|_{x_0} \right] dt$$

$$= 0 - \frac{1}{2} v\Lambda \frac{\partial u}{\partial x} \left[\frac{t^3}{3} \right]_{-1}^{1}$$

$$= -\frac{1}{3} v\Lambda \frac{\partial u}{\partial x}$$

or

$$j = \frac{1}{3} v\Lambda \frac{\partial u}{\partial T} \left(-\frac{\partial T}{\partial x} \right) \tag{5.53}$$

Comparing (5.53) with (5.51), the contribution of phonons to the thermal conductivity may be expressed as

$$K_{ph} = \frac{1}{3} v\Lambda \frac{\partial u}{\partial T}$$

or

$$K_{ph} = \frac{1}{3} C_V \Lambda v$$

$$= \frac{1}{3} C_V v^2 \tau \tag{5.54}$$

where C_V denotes the phonon heat capacity and τ is the relaxation time ($v\tau = \Lambda$).

We can define the phonon collision frequency as τ^{-1}. The variation of thermal conductivity with temperature in terms of the parameters Λ and C_V is explained below.

At high temperatures ($T >> \theta_D$)

In this temperature range, the thermal equilibrium phonon occupation numbers given by

$$\langle n_{\mathbf{k}} \rangle = \frac{1}{\exp(\hbar\omega(\mathbf{k})/k_B T) - 1}$$

reduce to

$$\sim \frac{k_B T}{\hbar\omega(\mathbf{k})} \tag{5.55}$$

So, the total number of phonons is proportional to the temperature. Therefore, the higher the temperature, the more will be the collision frequency, resulting in smaller mean free paths. Since C_V approaches the constant Dulong–Petit value, the change in the thermal conductivity is predominantly controlled by the change in mean free path. Thus, the thermal conductivity decreases with increase in temperature. Though this trend is confirmed by experiments, the observed rate of fall is given by

$$K_{ph} \propto \frac{1}{T^m} \tag{5.56}$$

where m lies between 1 and 2.

At low temperatures ($T << \theta_D$)

The phonon occupancy in thermal equilibrium may be taken as ~ exp $(-\hbar\omega(\mathbf{k})/k_B T)$. The probability of occurrence of U-processes [as expressed by (5.49)] is written as

$$\begin{aligned} n_{\mathbf{k}_1} \cdot n_{\mathbf{k}_2} &= \exp\,(-\hbar\omega_{\mathbf{k}_1}/k_B T) \cdot \exp\,(-\hbar\omega_{\mathbf{k}_2}/k_B T) \\ &= \exp\,[-\hbar(\omega_{\mathbf{k}_1} + \omega_{\mathbf{k}_2})/k_B T] \\ &= \exp\,(-\hbar\omega_{\mathbf{k}_3}/k_B T) \end{aligned}$$

which shows that the U-processes are almost frozen at low temperatures and the thermal conductivity exponentially approaches infinity.

But, in practice, all solids have finite conductivity even in the range of lowest temperatures. Since no perfect crystal exists in nature, there may be collisions of phonons with imperfections, impurities or even with the bounding surfaces. At certain temperature the mean free path on account of these collisions becomes so large that it is comparable with the size of the sample. This being the maximum realistic value, the mean free path assumes the *temperature independent behaviour below this temperature.* This is known as the *size effect.* Then (5.54) is replaced by

$$K_{ph} = \frac{1}{3} C_V \upsilon D \tag{5.57}$$

where D stands for the size (diameter in the case of a rod).

The behaviour of thermal conductivity *in this temperature region is mainly* determined by the *Debye heat capacity* which drops as T^3. In fact, a competition between the *exponential* and the T^3 terms sets in at the temperature *below which* the *mean free path is temperature independent.* The drop of T^3 being faster than the exponential increase, the thermal conductivity is limited to a finite value.

The scattering of phonons from crystal walls was suggested by Casimir.* The nature of the curves of Fig. 5.10 below 10 K provides a sound proof to the ideas of Casimir. The larger the cross-sectional area of the sample, the higher the conductivity. From below 10 K as the temperature rises,

* H.B.G. Casimir, *Physica* **5**, 495 (1938).

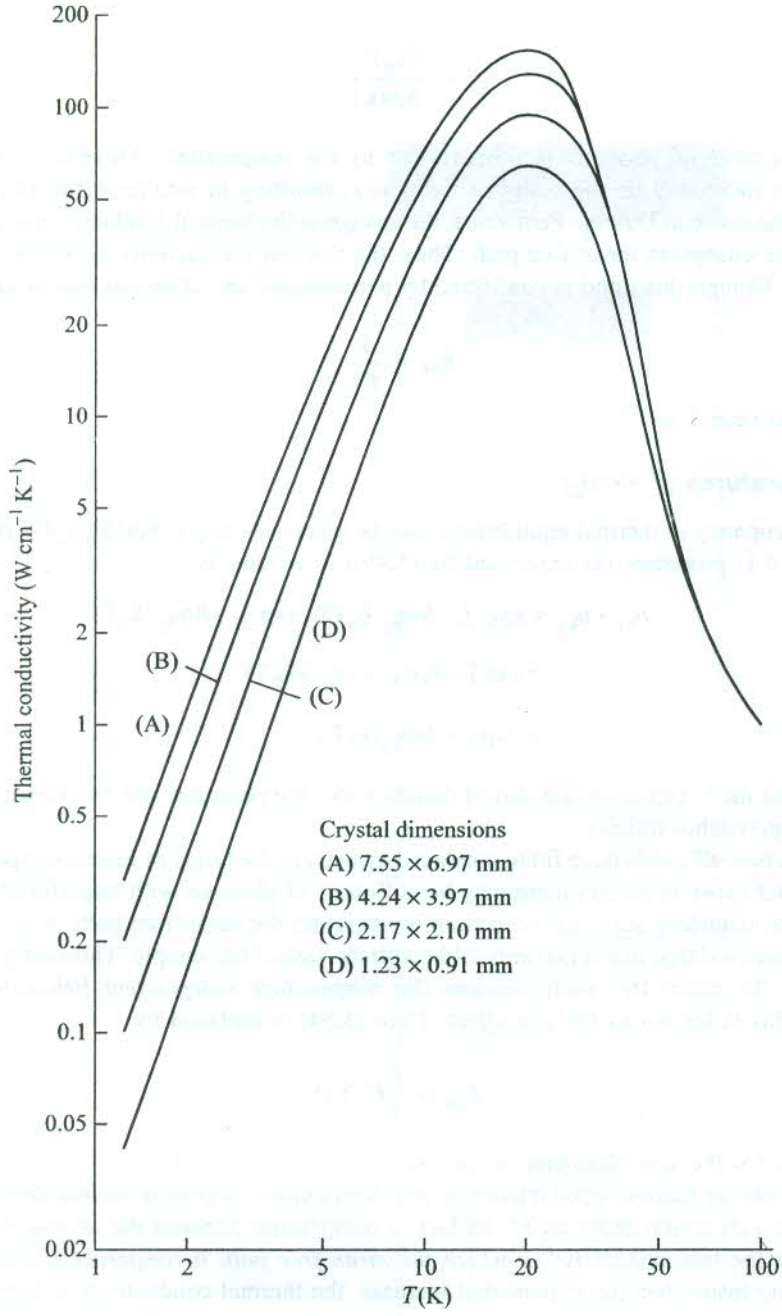


FIG. 5.10 Thermal conductivity of isotopically pure LiF crystals as a function of temperature. The curves are the beautiful demonstration of size effect on thermal conductivity. The larger the cross-sectional area of the sample, the higher the thermal conductivity. [After P.D. Thatcher, *Phys. Rev.*, **156**, 975 (1967).]

the U-processes become more frequent and the conductivity attains its maximum value when the mean free path because of the phonon–phonon scattering becomes comparable with that owing to the scattering from the crystal's surface. With further rise in temperature the conductivity falls because of the increasing phonon–phonon scattering and assumes the expected exponential drop at higher temperatures in the low temperature regime. The fall continues in the highest temperature range but according to (5.56), because the heat capacity here tends to be on level with the constant Dulong–Petit value.

The presence of the rival phonon scattering mechanisms is found to soften the 'sharp conductivity maximum' in most of the samples, irrespective of the material. The scattering from isotopic impurities is well identified among these. Its effect is demonstrated in Fig. 5.11 by the experimental curves of two samples of LiF crystal, one of which contains a mixture of ^{6}Li and ^{7}Li isotopes. On cooling, the rise of conductivity is less steep in the sample having both isotopes than in the sample which is almost free from ^{6}Li isotope. In addition, it has a lower and flatter maximum. The sharp maximum in the pure sample is almost totally contributed by U-processes in the absence of rival isotopic scattering.

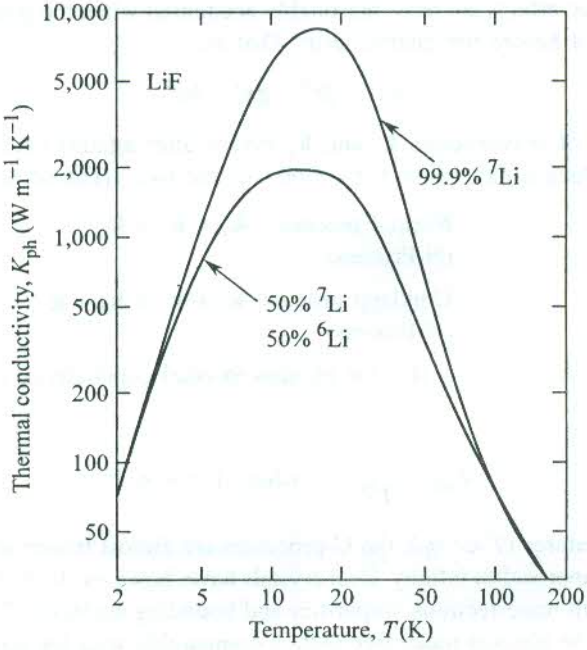


FIG. 5.11 The influence of isotope scattering on the maximum of thermal conductivity. The phonon scattering from isotopes lowers the thermal conductivity as is observed for the sample composed of ^{7}Li and ^{6}Li isotopes. The other sample almost free of ^{6}Li isotope shows a sharper and higher maximum. The presence of the rival phonon scattering mechanisms softens the sharp conductivity maximum. [After R. Berman, *Cryogencis*, **5**, 297 (1965).]

Insulators characterized by large values of thermal conductivity have found wide applications in the field of low temperature physics.

SUMMARY

1. For a harmonic crystal, heat capacity at constant volume, C_V is equal to heat capacity at constant pressure C_P because the difference between the two depends on the square of the temperature coefficient of linear expansion which is zero for a harmonic solid.
2. The achievements of the Einstein model of phonon heat capacity are:
   (i) It explains the behaviour of heat capacity at room temperature and above. In the high temperature limit, it gives the classical value ($3Nk_B$) which the experimental value approaches at these temperatures.
   (ii) It shows that the heat capacity approaches zero as $T \to 0$ K, a fact that agrees with the experiment.
3. The limitation of the Einstein model of phonon heat capacity is that the heat capacity at low temperatures decreases as exp $(-\hbar\omega/k_BT)$ in contrast to the observed T^3 dependence.
   The Debye T^3–law based on a continuum model explains very well the low temperature behaviour. The Debye model enjoys success over the entire useful temperature range.
4. The anharmonic effects are most reasonably accounted when the potential energy expansion is truncated not before the quartic term. That is,

$$U(x) = fx^2 - gx^3 - hx^4$$

5. Two phonons of wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$ may scatter against each other (or collide) and merge to produce a new phonon by either of the two types of scattering processes:

   Normal process: $\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3$
   (N-Process)

   Umklapp process: $\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3 + \mathbf{g}$
   (U-Process)

6. At high temperatures ($T >> \theta_D$), the phonon thermal conductivity (K_{ph}) changes according to

$$K_{ph} \propto \frac{1}{T^m} \qquad \text{where } 1 < m < 2$$

At low temperatures ($T << \theta_D$), the U-processes are almost frozen and K_{ph} for ideal crystals exponentially approaches infinity. Real crystals have, however, finite K_{ph} because of collisions of phonons with imperfections, impurities and bounding surfaces of specimens. At very low temperatures, the phonon mean free path is comparable with the size of the specimen (or a constant), so that

$$K_{ph} = \frac{1}{3} C_V v D$$

where D is the diameter in the case of a rod. Or

$$K_{ph} \propto T^3 \text{ (since } v \text{ and } D \text{ are constant and } C_V \propto T^3)$$

PROBLEMS

5.1 The thermal conductivity maximum of a synthetic sapphire sample of 3 mm diameter is observed at 30 K. The measured maximum value is 2.7×10^4 W m^{-1} K^{-1}. If the speed of sound in sapphire is 10^4 m s^{-1}, calculate the heat capacity of the sapphire at 30 K.

5.2 Obtain expressions for the thermal energy and the phonon heat capacity of
(a) a system of two harmonic oscillators, and
(b) a system with two energy levels.
Explain the difference in the two results.

5.3 Derive an expression for the temperature at which the thermal lattice energy is equal to the zero point energy in the Einstein model. Write down the corresponding condition in the Debye model.

5.4 Using the Debye approximation for a one-dimensional monatomic lattice with atomic spacing a and sound speed v, show that

$$\omega_D = \frac{\pi v}{a} \quad \text{and} \quad \theta_D = \frac{\hbar \omega_D}{k_B}$$

5.5 For Problem 5.4, derive the integral expressions for the thermal energy and phonon heat capacity. Show that

$$C_V = \frac{\pi^2 k_B}{3a}\left(\frac{T}{\theta_D}\right) \text{ per unit length at low temperatures}$$

$$= \frac{k_B}{a} \text{ per unit length at high temperatures.}$$

5.6 Solve the above problem for a two-dimensional crystal and show that at low temperatures $C_V \propto T^2$.

5.7 Consider the dispersion relation (4.4) for a one-dimensional monatomic lattice of N atoms. Show that the density of normal modes is given by

$$D(\omega) = \frac{2N}{\pi} \frac{1}{(\omega_m^2 - \omega^2)^{1/2}}$$

5.8 Show that for temperatures well below θ_D, the mean free path of a phonon can be expressed as

$$\Lambda = \frac{f K_{ph} \theta_D^3}{v T^3}$$

where v is the speed of sound and f is defined in (5.40).

5.9 In the approximation that the heat capacity is temperature independent, prove the following for the variation of thermal expansion coefficient α:

$$\frac{d\alpha}{dT} = -\frac{9gfC_V^2}{8v^4x_0}$$

where x_0 is the equilibrium interatomic separation.

5.10 Show that $u = \sum_{n=1}^{\infty} a_n \exp(in\omega t)$ is an approximate solution multiples of the harmonic frequency $\omega = \sqrt{\frac{f}{M}}$, to the equation of motion

$$M\ddot{u} + fu - \frac{1}{2}gu^2 = 0$$

for an anharmonic oscillator.

5.11 Show that the phonon density of states $D(\omega)$ for a diatomic linear chain diverges at the maximum frequency and at frequencies on either side of the gap while it tends to be a constant as $\omega \to 0$.

SUGGESTED FURTHER READING

Blackman, M., *Handbuch der Physik,* **7**, 325 (Springer, 1955).

Gopal, E.S.R., *Specific Heats at Low Temperatures* (Plenum Press, 1966).

Ho, C.Y. and R.E. Taylor (Editors), *Thermal Conductivity* (Plenum, 1969).

Keesom, P.H. and N. Pearlman, *Handbuch der Physik,* **14**, 282 (Springer, 1956).

Klemens, P.G., in F. Seitz and D. Turnbull (Editors), *Solid State Phys.*, Vol. **7** (Academic Press, 1958).

Launey, de J., in F. Seitz and D. Turnbull (Editors), *Solid State Phys.*, Vol. **2** (Academic Press, 1956).

Ziman, J.M., *Electrons and Phonons*, Chapter 8 (Oxford, 1960).

Chapter 6

Free Electron Theory of Metals

We know that more than a two-third of elements are metals. This shows that the natural tendency of elements is to favour the metallic state. Of the common solids most of which are insulators, metals have found the largest number of applications according them thereby a very special status. The striking physical properties that make them so attractive should emerge as natural consequences of a successful theory. It is easy to develop a theory for metals because of their relatively simple structure. Then, this theory can be extended to other solids with suitable modifications as the situation demands. As no perfect theory exists even today, we make an endeavour in this chapter to give an appropriate account of the main theoretical models proposed in order to explain the characteristic properties of metals.

The theory of metals heavily relies on the conviction that the valence electrons of metal atoms can be treated as free. The valence electrons are so loosely bound to the ion core that in the first approximation they may be considered to have freedom of moving through the volume of the metal. For example, the so called free nature of these electrons readily explains the metallic lustre as for free particles the emission of radiation immediately follows the process of absorption. The concept of free electron motion is indicated by the observed large electrical and thermal conductivities of metals. The free electrons in metals are called '*conduction electrons*' since they are responsible for the electric conduction. The collective body of conduction electrons contributed by individual metal atoms is conventionally interpreted as '*the free electron gas*' that occupies most of the volume of the metal. The fraction of volume belonging to the positive metal ions is mere 15 per cent in sodium metal. To sum up, a metallic crystal may be envisaged as the superposition of a periodic array of positive ion cores and a quasi-uniform negative charge density of the free electron gas.

The formulation of the theory of metals began well before the announcement of the Pauli principle (1926) and the advent of quantum mechanics (1925). Drude was the first to propose a model in 1900, three years after the discovery of electron by J.J. Thomson. Though the model operates in the domain of classical physics, its importance cannot be undermined in view of the amazing success it has achieved in deriving the Ohm's law and the relationship between the electrical and thermal conductivities. Its another feature of quickly demonstrating the behaviour of electrical conductivity further emphasizes the value of its contribution. A befitting account of this theory will be given in this chapter with a final word on the classical models. With the Pauli exclusion principle coming to force, the electrons in metals could no longer be treated by the classical statistics. The requirement that the electrons be subjected to the Pauli principle brought about a remarkable transformation in the theory which is now capable of addressing most of the questions left unanswered by the classical theory. This replaced the Maxwell–Boltzmann statistics by the Fermi–Dirac statistics and renamed the free electron gas as '*the free electron Fermi gas*'. Sommerfeld (1927) won the distinction of applying the related quantum mechanical ideas for the first time to explain the properties of metals. A suitable account of the Sommerfeld theory will lead us to a concluding discussion of this chapter.

6.1 THE DRUDE MODEL

The Drude model is essentially based on the classical kinetic theory of gases. According to Drude the metal must have two types of particles as against only one type in the simplest gases. The discovery of electrons bearing negative charge made it mandatory to accept the presence of positively charged entities (or particles) on the requirement of the charge neutrality condition. It is assumed that when metal atoms come together to form a metal, the valence electrons get liberated and move freely within the metal. The remainder of the atom is a positive ion carrying the major portion of the atomic mass. Drude took these particles as heavy and immobile. Under the action of an external electric field the free electrons referred to as conduction electrons in metals move in the background of immobile positive metal ions. In the absence of a relevant theoretical framework to deal with the free electron gas, Drude took recourse to the methods of kinetic theory of ideal gases for examining metallic conduction without getting deterred by the large electron densities (10^{22} cm^{-3}). It is assumed that the reader is acquainted with the postulates of kinetic theory of gases through a prior exposure, and therefore assumptions specific to only free electron gas are being given below:

1. The electron–electron and electron–ion interactions are neglected. The approach tantamounts to working in the independent electron approximation and the free electron approximation in that order. To be exact, the free electron approximation is not adhered to in strict literal sense because the electrons are considered to remain confined within the metal in the Drude Model and this is possible only if the electron–ion attractive force is active.
2. Under the action of an external electric field, electrons move opposite to the field direction and make collisions with immobile and impenetrable ion cores. An electron bumps† from ion to ion and between successive collisions moves in a straight line as determined by the Newton's equations of motion.
3. All electrons move with the RMS speed of a Maxwell–Boltzmann distribution, representing the random or thermal velocity of electrons. The average electron velocity immediately after the collision is zero.

6.1.1 Electrical Conductivity of Metals

DC Conductivity

We apply Drude theory first to derive an expression for dc conductivity. An external static electric field can affect electron velocity during the time interval between two successive collisions. But the gain in velocity is destroyed each time a collision occurs since the average velocity immediately after the collision is zero. A large influence of electric field is reflected in a larger mean free time or relaxation time τ. The probability of collision per unit time is defined as $1/\tau$. Therefore, the probability that a collision occurs in a small time interval dt is simply dt/τ.

Taking the acceleration of an electron of mass m as eE/m, the mean drift velocity can be written as

$$\mathbf{v}_d = \left(-\frac{e\mathbf{E}}{m}\right)\tau \tag{6.1}$$

† The collision involving physical contact is unrealistic. We, however, refrain from identifying the true scattering mechanism at this stage.

If there are a total of n electrons per m^3 in the metal, all with constant drift velocity $\mathbf{v}_d$, we have the following relation for the net electric current density $\mathbf{j}$:

$$\mathbf{j} = -ne\mathbf{v}_d = \left(\frac{ne^2\tau}{m}\right)\mathbf{E}$$

$$= \sigma\mathbf{E} \tag{6.2}$$

where the positive scalar quantity

$$\sigma = \frac{ne^2\tau}{m} \tag{6.3}$$

is defined as the electrical conductivity (the reciprocal of the electrical resistivity ρ). The form of the expression (6.3) remains the same in all models including those based on quantum physics. The difference lies only in the way n, m, and τ are defined.

We are now beginning to accomplish our aim of highlighting the gains of the Drude model. The first and the foremost of these is the derivation of the Ohm's law contained in (6.2). The law was initially set up purely on an empirical basis.

It must be, however, stated that the electrical conductivity σ is not universally scalar since in some complicated situations $\mathbf{j}$ becomes nonlinear with respect to $\mathbf{E}$ making σ to behave as a tensor. The electrical conductivity is often expressed in terms of the drift mobility of electrons μ as

$$\sigma = ne\mu \tag{6.4}$$

where

$$\mu = \frac{|\mathbf{v}_d|}{\mathrm{E}} \tag{6.5}$$

Drude put (6.3) in a further useful form by exploiting the ideas of kinetic theory such as to define τ in terms of thermal velocity v_{th} expressed by the following relations:

$$\tau = \frac{\Lambda}{v_{th}} \tag{6.6}$$

$$\frac{3}{2}k_B T = \frac{1}{2}mv_{rms}^2 = \frac{1}{2}mv_{th}^2 \tag{6.7}$$

where Λ is the electron mean free path. Using these relations, (6.3) can be written as

$$\sigma = \frac{ne^2\Lambda}{(3mk_B T)^{1/2}} \tag{6.8}$$

The relation (6.8) is another achievement of the Drude theory since it gives the right magnitude of electrical conductivity and correctly describes that conductivity increases with decrease in temperature (see Fig. 6.1). Experimental values at 77, 273 and 373 K are given in Table 6.1 for comparison. But the dependence on $T^{-1/2}$ is not in agreement with the observed T^{-1} dependence in

Table 6.1 Experimental electrical conductivities* of some metals (in 10^8 ohm^{-1} cm^{-1})

Metal	*At* 77 K	*At* 273 K	*At* 373 K
Li	0.96	0.12	0.08
Na	1.25	0.24	melts
K	0.72	0.16	melts
Cu	5.00	0.64	0.45
Ag	3.33	0.66	0.47
Au	2.00	0.49	0.35
Al	3.33	0.41	0.28
Fe	1.52	0.11	0.07
Zn	0.91	0.18	0.13
In	0.56	0.12	0.08
Tl	0.27	0.07	0.04
Pb	0.21	0.05	0.037

*From G.W.C. Kaye and T.H. Laby, *Table of Physical and Chemical Constants* (Longmans Green, 1966).

most of the common metals. The modern theory that resolves this issue attributes this inconsistency to the unrealistic assumption of electron–ion elastic collisions and the use of classical statistics to describe the electrons in the Drude model. The Drude model is further plagued by its inability to account for the low electron heat capacity and the temperature independent behaviour of paramagnetic susceptibility of conduction electrons. As we will see later in this chapter and in Chapter 13, solutions to both of these problems are achieved by applying quantum statistics.

By feeding the measured value of conductivity in (6.3), we obtain the value of relaxation time τ. The estimated value of τ is in the range 10^{-15} to 10^{-14} s. The thermal velocities estimated from (6.7) are around 10^5 m s^{-1}. Using (6.6), we find that the typical values of mean free path Λ are of the order of a few angstroms which compare with the interatomic separations. This in a way supports the electron–ion collisions in Drude theory. But the low temperature measurements on some of the purest and least imperfect crystals show the mean free path as large as a few centimetres. This strongly contradicts the Drude picture in which the electron bumps along from ion to ion. On the other hand, the modern theory tells that electron waves cannot be scattered by a perfectly periodic potential leading to an infinite mean free path in an ideal crystal. The experimental data in no way casts aspersion on the above theoretical view, simply because no body has ever produced any perfect and ideal crystal. Since every real crystal has deviations from periodicity because of the presence of imperfections and even impurities, the value of the mean free path is limited by the scattering from these centres. The measured values of electrical conductivity of a highly pure sample of copper metal is plotted as a function of temperature in Fig. 6.1.

AC conductivity

Here, we need to calculate the current that would flow under the influence of an alternating electric field (time-dependent),

$$\mathbf{E} = \mathbf{E}(\omega)e^{-i\omega t} \tag{6.9}$$

If the time-dependent momentum per electron be denoted by $\mathbf{p}$, we get the following equation of motion for the electron,

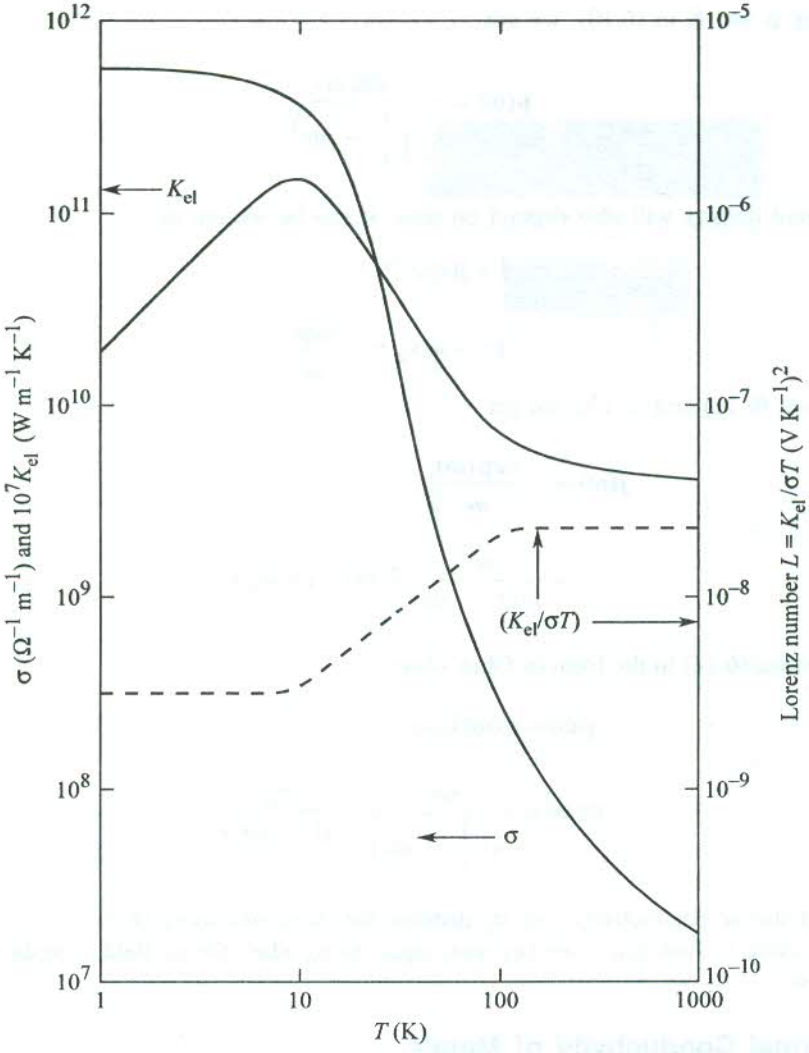


FIG. 6.1 Variations in electrical conductivity σ, electronic thermal conductivity K_{el} and Lorenz number L ($= K_{el}/\sigma T$) with change in temperature for a highly pure sample of copper metal. (Data on σ from G.T. Meaden, *Electrical Resistance of Metals* (Plenum Press, 1965); and K_{el} values from *American Institute of Physics Handbook* (McGraw-Hill, 1971).

$$\frac{d\mathbf{p}}{dt} = -\frac{\mathbf{p}}{\tau} - e\mathbf{E} \tag{6.10}$$

where $\left(-\frac{\mathbf{p}}{\tau}\right)$ represents the damping term following from collisions.

The time-dependent solution of the equation of motion is sought in the form

$$\mathbf{p} = \mathbf{p}(\omega)e^{-i\omega t} \tag{6.11}$$

On substituting **p** and **E** in (6.10), we get

$$\mathbf{p}(\omega) = -\frac{e\mathbf{E}(\omega)}{\left(\frac{1}{\tau} - i\omega\right)} \tag{6.12}$$

Since the current density will also depend on time, it can be written as

$$\mathbf{j} = \mathbf{j}(\omega)e^{-i\omega t} \tag{6.13}$$

But $$\mathbf{j} = -ne\mathbf{v}_d = -\frac{ne\mathbf{p}}{m}$$

Therefore, using (6.11) and (6.13), we get

$$\mathbf{j}(\omega) = -\frac{ne\mathbf{p}(\omega)}{m}$$

$$= \frac{ne^2}{m(\frac{1}{\tau} - i\omega)}\mathbf{E}(\omega) \quad \text{(using 6.12)} \tag{6.14}$$

One usually writes (6.14) in the form of Ohm's law:

$$\mathbf{j}(\omega) = \sigma(\omega)\mathbf{E}(\omega) \tag{6.15}$$

where $$\sigma(\omega) = \frac{ne^2}{m\left(\frac{1}{\tau} - i\omega\right)} = \frac{\sigma_0}{(1 - i\omega\tau)} \tag{6.16}$$

$\sigma(\omega)$ is called the ac conductivity and σ_0 denotes the dc conductivity (6.3).

It is instructive to find that $\sigma(\omega)$ becomes equal to σ_0 when the ac field is replaced with a dc field (i.e. $\omega = 0$).

6.1.2 Thermal Conductivity of Metals

In the Drude model it is assumed that the major contribution to thermal conductivity of metals comes from conduction electrons. This assumption has basis in the general observation that metals are far better conductors of heat than insulators. The participation of metal ions is reflected in the phonon contribution which is neglected in Drude theory on account of its relatively much small measure.

Consider a metal bar whose two ends are maintained at a constant difference of temperature. This situation refers to the steady state when the whole of thermal energy being fed at the hot end is received at the cold end without any net absorption in the bar. Let the temperature gradient along the length of the bar be defined as $-\partial T/\partial \mathbf{x}$, meaning thereby that the temperature decreases down the length being denoted by x. For a small temperature gradient, the thermal current **j**, which is a measure of the thermal energy flowing per unit sectional area of the bar per unit time, is found to be proportional to the temperature gradient and thus

$$\mathbf{j} = K_{el}\left(-\frac{\partial T}{\partial \mathbf{x}}\right) \tag{6.17}$$

where the constant of proportionality, K_{el}, stands for the electronic thermal conductivity.

The relation shows that the thermal energy flows in a direction opposite to that of the gradient.

Let us confine ourselves to the one-dimensional flow of heat energy in which electrons move parallel to the length of the bar. The calculation of **j** proceeds on lines parallel to those discussed in Section 5.3.3 in connection with the phonon current. The main difference between the two cases is that the number of heat carriers in this case (electrons) is conserved. Let us concentrate on a group of n electrons per unit volume. At a point x along the length, half of them ($n/2$) arrive from the hot end and the rest ($n/2$) from the cold end. The electron flowing from the hot end should have met its last collision at $(x - v_x\tau)$ whereas the one from the cold end would have its last collision at $(x + v_x\tau)$. If the thermal energy per electron in thermal equilibrium at T be denoted by $u[T(x)]$, the net thermal current can be written as

$$j = \frac{1}{2} nv_x[u(T(x - v_x\tau)) - u(T(x + v_x\tau))] \tag{6.18}$$

where $T(x)$ is the temperature at x and v_x is the x-component of the average electron velocity. In the approximation that the variation in temperature over a mean free path length Λ is small, (6.18) can be expanded about x giving,

$$j = nv_x^2\tau \frac{\partial u}{\partial T}\left(-\frac{\partial T}{\partial x}\right) \tag{6.19}$$

The factor $n(\partial u/\partial T)$ may be replaced by the electron heat capacity per unit volume C_{el} and v_x^2 by $\frac{1}{3}v^2$ in (6.19) to get

$$j = \frac{1}{3} C_{el} v^2\tau\left(-\frac{\partial T}{\partial x}\right) \tag{6.20}$$

Comparing (6.20) with (6.17), we get

$$K_{el} = \frac{1}{3} C_{el} v^2\tau = \frac{1}{3} C_{el} v\Lambda \tag{6.21}$$

where v is the RMS speed of electrons.

The calculation of C_{el} is made by taking the thermal energy per electron at temperature T as $\frac{3}{2}k_BT$, in accordance with the law of equipartition of energy. This gives the heat capacity per electron as $\frac{3}{2}k_B$, and $C_{el} = \frac{3}{2}nk_B$. Furthermore, the value of v^2 is obtained from (6.7). Making these substitutions in (6.21), we get

$$K_{el} = \frac{3}{2}\left(\frac{nk_B^2\tau}{m}\right)T \tag{6.22}$$

Even before the electrons were known, in 1835 Wiedemann and Franz were able to notice that all good conductors of electricity are good conductors of heat too. They established that the ratio of thermal conductivity to electrical conductivity is proportional to the temperature for a large number of metals. Later in 1881 Lorenz (different from Lorentz) observed that the proportionality constant equalling $K/\sigma T$ has almost the same value for most of the common metals. This constant is known as the Lorenz number and the empirical law became famous as the Wiedemann–Franz law. From (6.3) and (6.22), Drude obtained

$$\frac{K}{\sigma} = \frac{3}{2}\left(\frac{k_B}{e}\right)^2 T \tag{6.23}$$

The coefficient of T in (6.23) is a constant. Drude was thus able to derive the Wiedemann–Franz law which is acknowledged as the biggest success of the Drude model. The Lorenz number L as defined in the theory is given by

$$\begin{aligned} L = \frac{K}{\sigma T} &= \frac{3}{2}\left(\frac{k_B}{e}\right)^2 \\ &= 1.11 \times 10^{-8}\ \text{W}\ \Omega\ \text{K}^{-2} \\ &= 1.11 \times 10^{-8}\ (\text{V K}^{-1})^2 \end{aligned} \tag{6.24}$$

But this number is about half the experimental value given by Table 6.2 and Fig. 6.1. Though Drude theory correctly describes the Wiedemann–Franz law, this disagreement points to certain weak links in the theory. However, it is amusing to learn that Drude obtained the correct Lorenz number because of the mistake of getting half the value of σ given by (6.3). The impressive success as it sounded at that time turns out to be misconceived further because of the errors in the calculation of thermal conductivity. The value of the electron heat capacity $\left(\frac{3}{2} nk_B\right)$ used is exactly half the Dulong–Petit value, which is quite large compared to its negligible contribution observed at room

Table 6.2 Experimental thermal conductivities* and Lorenz number for some metals

Metal	273 K		373 K	
	Thermal conductivity (W cm^{-1} K^{-1})	*Lorenz number* (V K^{-1})^2	*Thermal conductivity* (W cm^{-1} K^{-1})	*Lorenz number* (V K^{-1})^2
Na	1.38	2.12×10^{-8}	melts	
Cu	3.85	2.20	3.82	2.29×10^{-8}
Ag	4.18	2.31	4.17	2.38
Au	3.10	2.32	3.1	2.36
Al	2.38	2.14	2.30	2.19
Fe	0.80	2.61	0.73	2.88
Tl	0.50	2.75	0.45	2.75
Bi	0.09	3.53	0.08	3.35

*From G.W.C. Kaye and T.H. Laby, *Table of Physical and Chemical Constants* (Longmans Green, 1966).

temperature. The actual value is about a hundred times smaller. Another error appears in the magnitude of electrons thermal velocity which, in fact, is a hundred times larger. It is simply a pleasant coincidence that the two errors almost cancel each other in (6.21), giving an overall success to the Drude model.

The experimental values of K and L for some metals at 273 K and 373 K are given in Table 6.2. The observed behaviour of these quantities for a highly pure sample of copper metal is shown in Fig. 6.1 over a wide range of temperature. The pronounced fall in the value of Lorenz number below 100 K (see Fig. 6.1) turns out to be a valid fact for most of the metals to which Drude could not offer an explanation. As will be discussed later in Sections 6.6 and 6.7, the two extreme temperature ranges are characterized by two different mechanisms of electron collisions. These different mechanisms provide different temperature dependences to the mean free path that controls both the electrical and thermal conductivities. This explains why Lorenz number decreases significantly on cooling below 100 K.

The most instructive aspect of the Drude's calculations is the elimination of τ in the derivation of the Wiedemann-Franz law. The astounding success achieved in this respect brings out the inference that the Drude theory is capable of explaining the behaviour of physical quantities that are τ independent. The inference made above was confirmed by Lorentz for Hall coefficient found independent of relaxation time. A quantitative account of the matter is given in the following section without going into the theory of Hall coefficient slated for discussion at a later stage in this chapter.

6.2 LORENTZ MODIFICATION OF THE DRUDE MODEL

Lorentz took up the task of modifying the oversimplified Drude model and constructed his theory on the basis of following points:

1. The assumption that all electrons move with the same thermal velocity is abandoned.
2. The classical Maxwell–Boltzmann velocity distribution is perturbed by the presence of an electric field or a thermal gradient. Both of these tend to displace the equilibrium velocity distribution and distort its symmetry.
3. The approach of Boltzmann transport equation is followed to describe the transport of charge and kinetic energy of electrons by a statistical distribution of mobile electrons constituting the electron gas.

To avoid duplication, the derivation of the expression for electrical conductivity will be taken up at a proper stage for the quantum model of the free electron gas, i.e. for the free electron Fermi gas. Nevertheless, we give below the relation as obtained by Lorentz:

$$\sigma_L = \left(\frac{8}{3\pi}\right)^{1/2} \frac{ne^2\Lambda}{(3mk_BT)^{1/2}} \tag{6.25}$$

If we denote the Drude conductivity of (6.8) as σ_D, then

$$\sigma_D = \left(\frac{3\pi}{8}\right)^{1/2} \sigma_L \tag{6.26}$$

$$= 1.09\ \sigma_L$$

From (6.25) and (6.26) we learn that the Lorentz modifications provide neither any noteworthy change in the quantitative measure nor any change in the temperature dependence of electrical conductivity. Lorentz, however, went on to analyze the effect of a uniform magnetic field on current carrying conductors. This effect is known as the Hall effect. The net force on an electron moving with velocity **v** under the action of a static electric field **E** and a uniform magnetic field **B** is

$$\mathbf{F} = -e[\mathbf{E} + \mathbf{v} \times \mathbf{B}] \tag{6.27}$$

where **F** is called the Lorentz force.

Lorentz solved the Boltzmann transport equation under the above conditions and obtained the following expression for the Hall coefficient,

$$R_H = -\left(\frac{3\pi}{8}\right)\frac{1}{ne} \tag{6.28a}$$

whereas according to Drude theory,

$$R_H = -\frac{1}{ne} \tag{6.28b}$$

The electron density n can be estimated from (6.28) by feeding the measured value† of Hall coefficient in it. These are given in Table 6.3 together with the atomic concentrations in certain metals. The estimate is consistent with the contribution of one conduction electron per atom in sodium. The result is not so good for silver and is worse for other metals. This points to an otherwise established fact that the free electron approximation is best applicable to alkali metals with the lone loosely bound outermost $3s^1$ electron. The subject is effectively treated by the band theory that also accounts for the observed positive Hall coefficient in certain metals.

Table 6.3 Hall coefficient and electron density of some metals

Metal	*Hall coefficient* (R_H)* ($m^3\ C^{-1}$)	*Electron density* (n) *deduced from* R_H** (m^{-3})	*Atomic concentration deduced from lattice constant* (m^{-3})
Na	-2.5×10^{-10}	2.9×10^{28}	2.5×10^{28}
Ag	-8.4×10^{-11}	8.8×10^{28}	5.8×10^{28}
Al	-3.0×10^{-11}	2.5×10^{29}	6.0×10^{28}
Cd	$+6.0 \times 10^{-11}$	Negative	4.6×10^{28}
Fe	$+2.5 \times 10^{-11}$	Negative	8.5×10^{28}
Ni	-6.0×10^{-10}	1.2×10^{28}	9.2×10^{28}

* From *Handbook of Physics*, Ed., E.U. Condon and H. Qdishaw (McGraw-Hill, 1958).

** Using $R_H = \dfrac{3\pi}{8ne}$

† Determined from $R_H = \dfrac{E_y}{j_x B_z}$, whose details can be found in Section 6.6.1.

Other important consequences of the Lorentz solution of Boltzmann transport equation are:

1. *The Thomson effect.* An electromotive force is developed in conductors between whose two ends a temperature gradient is maintained. This effect is treated in Section 6.8.2.
2. *The magnetoresistance effect.* The resistivity of a metal increases when placed in a magnetic field. The Drude model does not predict this effect.

Where the Thomson effect is exploited in several thermoelectric devices, the magnetoresistance effect is immensely useful in the study of electrical transport properties. The magnetoresistance is defined as the fractional increase in the zero field resistivity when the sample is placed in the region of a magnetic field. In the Lorentz model when the Boltzmann equation is solved without ignoring terms in B of higher powers, the magnetoresistance increases as the square of the magnetic field. Experiments have confirmed this square-law dependence. The study of magnetoresistance has proved of tremendous value in the characterization of semiconductors. The knowledge about its behaviour is very crucial in the study of transformation between the metallic and insulating phases of composites. Some of these have interesting properties making them attractive for applications.

6.3 THE FERMI–DIRAC DISTRIBUTION FUNCTION

As already stated, the Sommerfeld theory differs from the classical theories in the replacement of the Maxwell–Boltzmann statistics by the Fermi–Dirac statistics. The distribution of electron thermal speeds in thermal equilibrium at temperature T is described by the function,

$$f(v) = \frac{1}{\exp\left[\left(\frac{1}{2}mv^2 - \mu\right)\Big/k_B T\right] + 1} \tag{6.29}$$

It is more commonly expressed as a function of electron energy,

$$f(\varepsilon) = \frac{1}{\exp\left[(\varepsilon - \mu)/k_B T\right] + 1} \tag{6.30}$$

The $f(\varepsilon)$ is known as the Fermi–Dirac distribution function. It denotes the probability that an orbital of energy ε be populated in an ideal electron gas in thermal equilibrium. The quantity μ is a function of temperature and is called the *chemical potential.* The highest energy level populated at absolute zero is designated as ε_F and is called the Fermi level. Its value is decided so as to get the correct number for the total number of particles (here, electrons), using the distribution function (6.30). At this temperature, none of the electrons have an energy greater than ε_F. All the electrons are confined to levels below ε_F.

The temperature dependence of $f(\varepsilon)$ may be summarized as follows:

At $T = 0$

$$\begin{aligned} f(\varepsilon) &= 0 \quad \text{for } \varepsilon > \varepsilon_F \\ &= 1 \quad \text{for } \varepsilon < \varepsilon_F \end{aligned} \tag{6.31a}$$

At $T > 0$

$$f(\varepsilon) = \frac{1}{2} \quad \text{for } \varepsilon = \varepsilon_F \tag{6.31b}$$

As $T \to 0$

$$\lim_{T \to 0} f(\varepsilon) = 1 \qquad \text{for } \varepsilon < \mu$$
$$= 0 \qquad \text{for } \varepsilon > \mu \tag{6.31c}$$

For these conditions to be true, it is required that

$$\lim_{T \to 0} \mu = \varepsilon_F \tag{6.32}$$

But observations on metals show that to a high degree of precision the chemical potential remains equal to the Fermi energy well up to room temperature. This allows μ to be safely replaced by ε_F in (6.30) while dealing with metals at room temperature and below. This practice, however, should not relent us to ignore the difference of μ from its value at absolute zero in general, since at times it may yield grossly inaccurate results.

The variation of $f(\varepsilon)$ with the change in electron energy ε at certain temperatures is shown in Fig. 6.2. For $(\varepsilon - \mu) >> k_B T$, (6.30) reduces to the form

$$f(\varepsilon) \simeq \exp[(\mu - \varepsilon)/k_B T] \tag{6.33}$$

which is simply the classical Boltzmann distribution. This indicates that for high values of electron energy the Fermi–Dirac function approaches the classical distribution function (see Fig. 6.2).

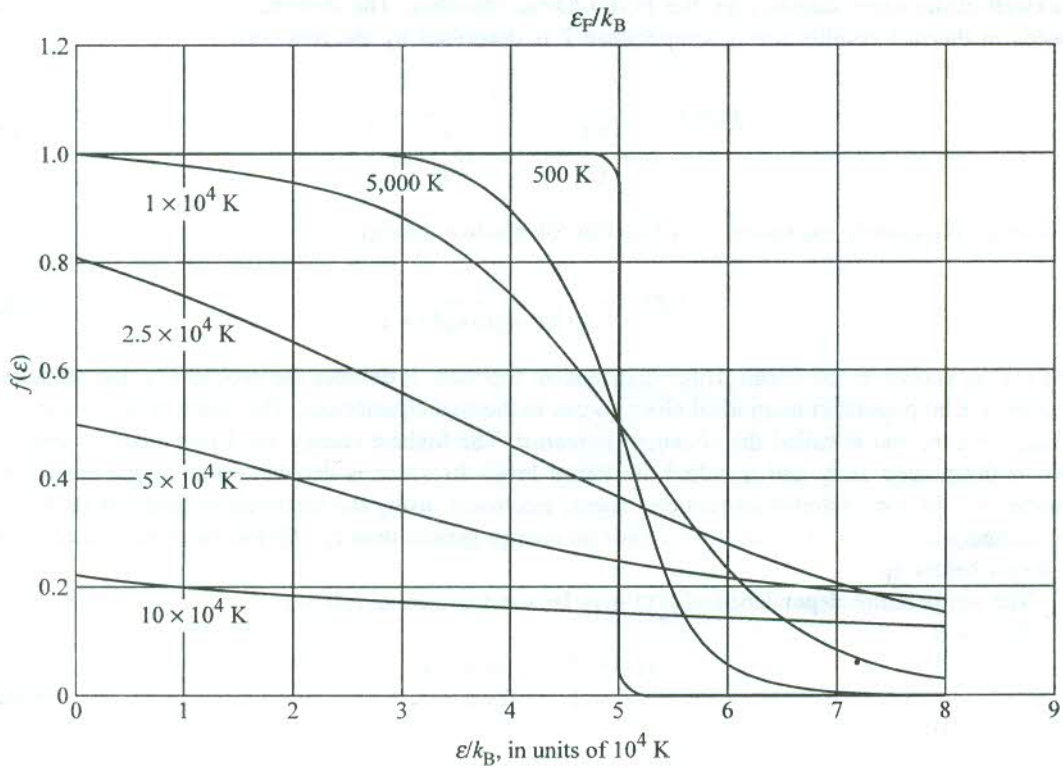


FIG. 6.2 Temperature dependence of the Fermi–Dirac distribution function for a free electron gas in three dimensions, with $T_F = \dfrac{\varepsilon_F}{k_B} = 50{,}000$ K.

Therefore, in this limit the choice between the classical and the quantum statistics is meaningless as both give similar results.

The Fermi energy ε_F is often expressed in terms of temperature by the relation,

$$\varepsilon_F = k_B T_F \tag{6.34}$$

where T_F, called the *Fermi temperature*, turns out to be of the order of tens and thousands of kelvin because the Fermi energy per electron is several electron volts.

6.4 THE SOMMERFELD MODEL

The electron–ion interaction cannot be totally ignored even in metals because the conduction electrons remain confined within the volume of the metal. Since no emission of electrons is observed at room temperature, the potential energy of an electron at rest inside the metal must be lower than that of an electron at rest outside. An electron needs to be excited to a certain level so as to leave the metal and escape to infinity. This level is called the *vacuum level*. The electrons move in the field of the periodic potential of positive ions. Therefore, the potential energy of electrons should be periodic in the crystal lattice. Sommerfeld, however, described the cube of a metal crystal as a three-dimensional potential box with an infinite barrier at the surfaces taking the potential energy as a constant within the box. But it is quite interesting that even such an oversimplified model gives a good agreement with the experiment, though the work function of metals is in the range of 5 eV.

Consider a metal cube of side L. The potential energy of electrons in the metal in the form of an infinite square-well is shown in Fig. 6.3.

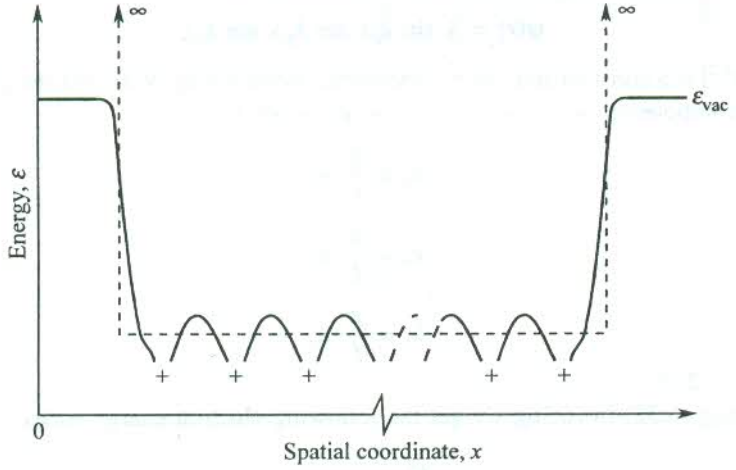


FIG. 6.3 General form of the potential energy of an electron moving in a periodic metallic crystal of positive cores. This is approximated by a square-well potential shown.

Let $V(\mathbf{r})$ denote the potential of an electron at point $\mathbf{r}$ and ε' be its total energy. Then, we can write the time-independent Schrödinger wave equation in the one-electron approximation as

$$\left[\frac{\left(-i\hbar\frac{d}{d\mathbf{r}}\right)^2}{2m} + V(\mathbf{r})\right]\psi(\mathbf{r}) = \varepsilon'\psi(\mathbf{r}) \tag{6.35}$$

where $-i\hbar(d/d\mathbf{r})$ stands for the linear momentum per electron and $\psi(\mathbf{r})$ is the electron wavefunction.

Sommerfeld defined $V(\mathbf{r})$ as

$$\begin{aligned} V(x, y, z) &= V_0, && \text{a constant for } 0 \le x, y, z \le L \\ &= \infty, && \text{otherwise} \end{aligned}$$

Let the kinetic energy per electron be denoted by ε, such that $\varepsilon = \varepsilon' - V_0$, which gives

$$-\frac{\hbar^2}{2m}\nabla^2 \psi(\mathbf{r}) = \varepsilon\, \psi(\mathbf{r}) \tag{6.36}$$

The infinite barrier at the bounding surfaces ($x, y, z = 0$ and L) keeps the electrons confined within the crystal and gives the following boundary conditions for the wavefunction:

$$\begin{aligned} \psi = 0 \quad \text{for} \quad & x = 0 \text{ and } L; \text{ and } 0 < y, z < L \\ & y = 0 \text{ and } L; \text{ and } 0 < z, x < L \\ & z = 0 \text{ and } L; \text{ and } 0 < x, y < L \end{aligned} \tag{6.37}$$

These fixed boundary conditions require the electrons to remain confined within the potential box. Therefore, the free electron plane waves moving along the x-edge of the cube [$\exp(ik_x x)$] have to be reflected back just on reaching the potential barrier. The two oppositely moving plane waves may form standing waves [$\exp(ik_x x) \pm \exp(-ik_x x)$] identified as cosine and sine waves, respectively. Since only sine waves satisfy the boundary condition (6.37), they serve as the proper solutions to the Schrödinger equation (6.36). The solution is expressed as

$$\psi(r) = A \sin k_x x \sin k_y y \sin k_z z \tag{6.38}$$

where A [$= (2/L)^{3/2}$] is a constant and can be determined normalizing $\psi(r)$ over the potential box. The three Cartesian components of the wavevector $\mathbf{k}$ are given by

$$\begin{aligned} k_x &= \frac{\pi}{L} n_x \\ k_y &= \frac{\pi}{L} n_y \\ k_z &= \frac{\pi}{L} n_z \end{aligned} \tag{6.39}$$

with $n_x, n_y, n_z = 1, 2, 3, \ldots$

On substituting (6.38) in (6.36), we get the following electron energy states:

$$\varepsilon_{\mathbf{k}} = \frac{\hbar^2}{2m}\left(k_x^2 + k_y^2 + k_z^2\right) \tag{6.40}$$

$$= \frac{\hbar^2 k^2}{2m} \tag{6.40a}$$

$$= n^2 \frac{1}{2m}\left(\frac{h}{2L}\right)^2 \tag{6.40b}$$

with $n_x^2 + n_y^2 + n_z^2 = n^2$.

For an electron moving in the free space, we use (6.40a) to express the energy. The variation of energy with wave vector as plotted in Fig. 6.4 is known as the free electron dispersion curve. But when a free electron moves in a potential box, as in the present case, the electron energy is meaningfully given by (6.40b).

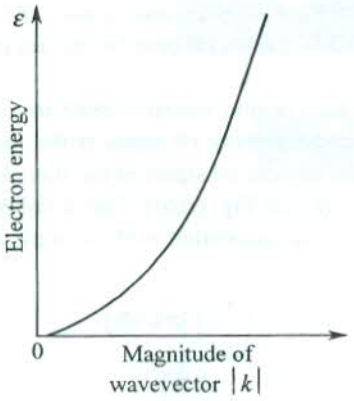


FIG. 6.4 Variation of the free electron energy as a function of the magnitude of its wavevector.

The components (k_x, k_y, k_z) or (n_x, n_y, n_z) together with the spin quantum number m_s are the quantum numbers of the problem. The wavefunctions and the energy states of a free electron in a square-well potential of length L in the x-direction are depicted in Fig. 6.5. The allowed energy values in the three-dimensional space produce constant spherical energy surfaces,

$$\varepsilon_{\mathbf{k}} = \frac{\hbar^2 k^2}{2m} = \text{constant}$$

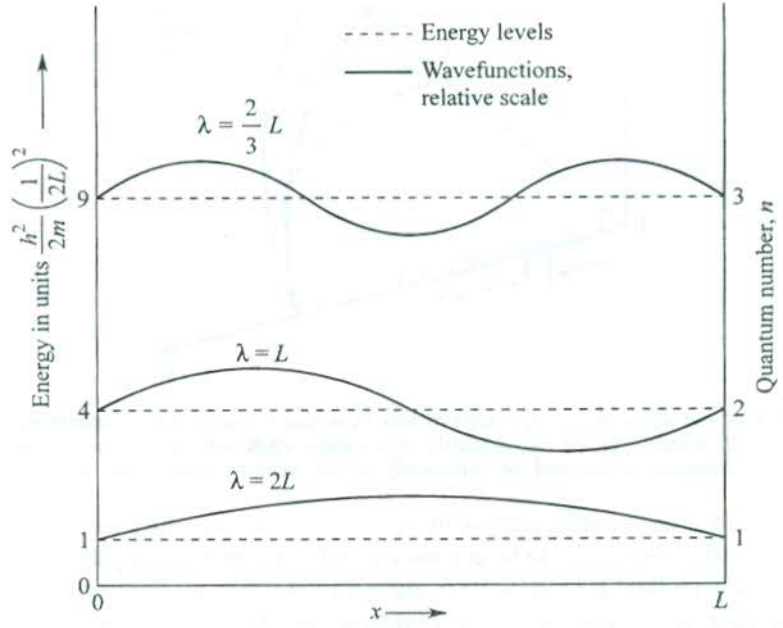


FIG. 6.5 Eigenvalues and eigenfunctions of the first three states of a free electron confined in a potential box of length L.

6.4.1 The Density of States

An insight into the possible values of k_x, k_y, k_z given by (6.39) provides the clue to the determination of density of electron states. We have excluded the zero value of n_x, n_y and n_z since the respective solutions cannot be normalized over the potential box. Further, no new linearly independent solutions are found for the negative values of these integers and as such these values are not considered. Thus for the fixed boundary condition (6.37), the useful possible values of **k** totally lie in the positive octant of the k-space.

Equations (6.39) show that there is one electron state in volume $(\pi/L)^3$ of the k-space. The principle of calculation of the electron density of states is the same as that for phonons described in Section 5.2.3. Let us calculate the number of states in the thin shell of the octant enclosed between the energy surfaces $\varepsilon(\mathbf{k})$ and $\varepsilon(\mathbf{k}) + d\varepsilon$ [see Fig. (6.6)]. This number, obtained by dividing the volume of this part of the shell by the volume associated with a single state, is

$$\frac{\left(\frac{1}{8}\right)\left(4\pi k^2 dk\right)}{\left(\frac{\pi}{L}\right)^3}$$

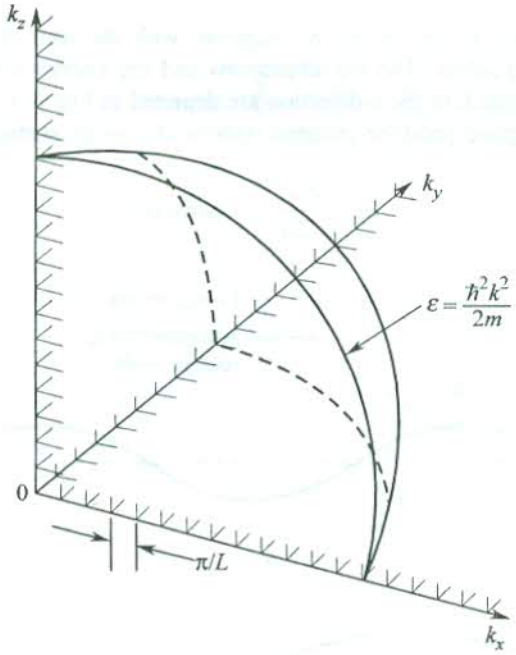


FIG. 6.6 A lattice of the allowed electron wavevectors in **k**-space representing electron states in an infinite square well. Because of the two possible spin values, each point gives two states. The states have a linear separation of π/L and lie completely in one octant for the fixed boundary conditions.

The correct number of possible states is twice this number since the Pauli principle allows two electrons with opposite spins ($\pm$ 1/2) to be accommodated in a single orbital designated by a **k**-value. The two electrons in an orbital are said to be in different states that are degenerate (degenerate states have a common eigenvalue). Therefore, if $D(\varepsilon)$ denotes the density of states,

$$D(\varepsilon)d\varepsilon = 2\frac{\left(\frac{1}{8}\right)(4\pi k^2 dk)}{\left(\frac{\pi}{L}\right)^3}$$

$$= \frac{L^3}{\pi^2} k^2 dk$$

$$= \frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \varepsilon^{1/2} d\varepsilon \qquad \left(\text{using } \varepsilon_{\mathbf{k}} = \frac{\hbar^2 k^2}{2m}\right)$$

Therefore,

$$D(\varepsilon) = \frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \varepsilon^{1/2} \tag{6.41}$$

where V is the volume of the crystal. We may express the density of states per unit volume of the crystal as

$$\frac{D(\varepsilon)}{V} = \frac{1}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \varepsilon^{1/2} \tag{6.42}$$

The density of states as a function of energy is graphically represented in Fig. 6.7.

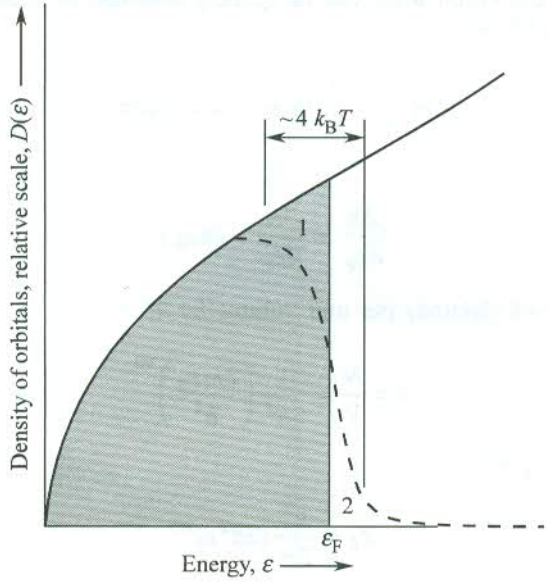


FIG. 6.7 Variation of the density of single-particle states as a function of temperature for a free electron gas in three dimensions. For $k_B T < \varepsilon_F$, the density of the filled orbitals $D(\varepsilon)\, f(\varepsilon, T)$ at finite temperature is given by the dashed curve. The shaded area represents the filled orbitals at 0 K. On heating to temperature T, the average energy increases and the electrons from region 1 (below ε_F) are thermally excited to region 2 (above ε_F).

It is left as an exercise for the reader to show that the periodic boundary conditions,

$$\psi(x + L, y + L, z + L) = \psi(x, y, z)$$

that yield plane wave solutions to the Schrödinger equation (6.36), lead to the same expression for the density of states as given by (6.41).

6.4.2 The Free Electron Gas at 0 K

The state of the free electron gas at the absolute zero temperature forms the basis for studying the properties of metals. We know that at this temperature all the electrons are filled in orbitals below the Fermi level, ε_F. Using this concept, we calculate the total number of electrons,

$$N = \int_0^{\varepsilon_F} D(\varepsilon) f(\varepsilon)\, d\varepsilon$$

$$= \frac{V}{2\pi^2} \left(\frac{2m}{\hbar^2}\right)^{3/2} \int_0^{\varepsilon_F} \varepsilon^{1/2}\, d\varepsilon \qquad (\text{for } \varepsilon < \varepsilon_F,\ f(\varepsilon) = 1)$$

or

$$N = \frac{V}{3\pi^2} \left(\frac{2m\varepsilon_F}{\hbar^2}\right)^{3/2} \tag{6.43}$$

The density of states at the Fermi level can be quickly obtained in terms of the total number of electrons if we rewrite (6.43) as

$$\ln N = \frac{3}{2} \ln \varepsilon_F + \text{a constant} \tag{6.44}$$

Differentiating it, we get

$$\frac{dN}{d\varepsilon_F} = \frac{3N}{2\varepsilon_F} = D(\varepsilon_F) \tag{6.45}$$

If we denote the number of electrons per unit volume by n,

$$n = \frac{N}{V} = \frac{1}{3\pi^2} \left(\frac{2m\varepsilon_F}{\hbar^2}\right)^{3/2} \tag{6.46}$$

giving the Fermi energy ε_F as

$$\varepsilon_F = \frac{\hbar^2}{2m} (3\pi^2 n)^{2/3} \tag{6.47}$$

Thus the Fermi energy may be estimated using (6.47) with the knowledge of the electrons density, n. Since metals have about 10^{28} electrons per m^3, the magnitude of ε_F is several eV per electron. From (6.34) we get a Fermi temperature, T_F in the range of tens of thousands of kelvin which

is enormously higher than the room temperature. On this scale of temperature the room temperature is far more closer to 0 K. Furthermore, experiments show that most of the properties of metals at room temperature are not much different than those at 0 K. This permits us to use the value of the density of states near the Fermi level for all calculations by replacing ε with ε_F in (6.41).

The size of ε_F is also measured in terms of the Fermi wavevector k_F, whose magnitude equals the radius of the spherical Fermi energy surface in the three-dimensional k-space. From (6.47), we have

$$k_F = (3\pi^2 n)^{1/3} \tag{6.48}$$

The electrons move on the Fermi surface with a constant speed, v_F expressed as

$$v_F = \frac{\hbar k_F}{m} = \frac{\hbar}{m}(3\pi^2 n)^{1/3} \tag{6.49}$$

6.4.3 Energy of Electron Gas at 0 K

This estimate is invaluable to several calculations involving free electrons. The energy per unit volume is expressed as

$$U_0 = \frac{1}{V}\int_0^{\varepsilon_F} \varepsilon\, D(\varepsilon)\, d\varepsilon \tag{6.50}$$

Substituting $D(\varepsilon)$ from (6.41), we obtain

$$U_0 = \frac{\varepsilon_F^{5/2}}{5\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \tag{6.51}$$

It is more useful to know the energy per electron, also referred to as the average kinetic energy per free electron. We obtain this value by dividing (6.51) by (6.46). Therefore,

$$\text{Average kinetic energy per electron} = \frac{U_0}{n} = \frac{3}{5}\varepsilon_F \tag{6.52}$$

6.5 THE ELECTRON HEAT CAPACITY

The most glaring thermal anomaly concerning the electron heat capacity, as encountered in the Drude model, was resolved by Sommerfeld with the use of quantum statistics. He applied Fermi–Dirac statistics to study the properties of the free electron gas. The density of states that describes the number of energy states per unit energy interval is shown in Fig. 6.7 as a function of energy. Sommerfeld pointed out that the smeared out region of the distribution function has a width of the order of $4k_BT$. According to Fig. 6.7 the electrons in the energy range $2k_BT$ below the Fermi level undergo excitation to cross it when the metal is heated from absolute zero to a temperature T. The electrons in still deeper levels are unable to scatter to a slightly different energy on account of all levels of comparable energy being already occupied with no room for additional occupancy as dictated by the Pauli principle. From Fig. 6.7 we find that only the uppermost electrons equalling a fraction of roughly $(4k_BT/\varepsilon_F)$ are affected by the presence of a field gradient (thermal or electrical).

Sommerfeld thus concluded that electrons occupying states only close to the Fermi level ε_F contribute to most of the properties of metals. We can make a rough estimate of the electron heat capacity based on the above picture. Taking the rise in thermal energy per electron as $k_B T$ on heating the system from 0 K to T K, the change in the net thermal or internal energy per unit volume may be given by

$$U = n\left(\frac{4k_B T}{\varepsilon_F}\right) \cdot k_B T$$

$$= \frac{4nk_B T^2}{T_F} \qquad (\text{since } \varepsilon_F = k_B T_F) \tag{6.53}$$

Therefore, the electron heat capacity C_{el} is written as

$$C_{el} = \frac{\partial U}{\partial T} = 8nk_B\left(\frac{T}{T_F}\right) \tag{6.54}$$

It is definitely a very small number since T_F is around 10^5 K. The order of magnitude is consistent with the experimental observation that there is hardly a deviation from the Duloug–Petit value of heat capacity at room temperature. This confirms absurdity of the Drude's estimate of $\frac{3}{2}nk_B$ which is exactly half the Dulong–Petit value.

Now, we take up a proper calculation of the heat capacity contributed by the conduction electrons. The net increase in internal energy per unit volume as the temperature rises from 0 K to a temperature T is written as

$$U = \int_0^\infty \varepsilon D(\varepsilon) f(\varepsilon)\, d\varepsilon - \int_0^{\varepsilon_F} \varepsilon D(\varepsilon)\, d\varepsilon \tag{6.55}$$

with the total number of electrons per unit volume given by

$$n = \int_0^\infty D(\varepsilon) f(\varepsilon)\, d\varepsilon \tag{6.56a}$$

So, we write

$$\varepsilon_F \cdot n = \varepsilon_F \int_0^\infty D(\varepsilon) f(\varepsilon)\, d\varepsilon \tag{6.56b}$$

On differentiating (6.55) and (6.56b), we get

$$C_{el} = \frac{\partial U}{\partial T} = \int_0^\infty \varepsilon D(\varepsilon) \frac{\partial f}{\partial T}\, d\varepsilon$$

$$0 = \varepsilon_F \frac{\partial n}{\partial T} = \int_0^\infty \varepsilon_F D(\varepsilon) \frac{\partial f}{\partial T} d\varepsilon$$

Subtracting the latter from the former of the above equations, we get

$$C_{el} = \int_0^\infty (\varepsilon - \varepsilon_F) D(\varepsilon) \frac{\partial f}{\partial T} d\varepsilon \tag{6.57}$$

An examination of Fig. 6.2 and Fig. 6.7 shows that at temperatures of interest the derivative $\partial f/\partial T$ has significant values in the region of $\pm 2k_BT$ about ε_F. The variation of $D(\varepsilon)$ in this region is not much and, therefore, we can use $D(\varepsilon_F)$ for $D(\varepsilon)$ in the first approximation getting,

$$C_{el} = D(\varepsilon_F) \int_0^\infty (\varepsilon - \varepsilon_F) \frac{\partial f}{\partial T} d\varepsilon \tag{6.58}$$

Differentiating the Fermi–Dirac function (6.30), we obtain

$$\frac{\partial f}{\partial T} = \frac{\varepsilon - \varepsilon_F}{k_B T^2} \frac{\exp[(\varepsilon - \varepsilon_F)/k_B T]}{\{\exp[(\varepsilon - \varepsilon_F)/k_B T + 1]\}^2} \tag{6.59}$$

Putting $\dfrac{\varepsilon - \varepsilon_F}{k_B T} = x$, we get

$$C_{el} \approx k_B^2 TD(\varepsilon_F) \int_{-\varepsilon_F/k_B T}^\infty \frac{x^2 e^x}{(e^x + 1)^2} dx \tag{6.60}$$

The e^x factor in the integrand is negligible for $x \leq -\varepsilon_F/k_BT$, i.e. at low temperatures. Therefore, the lower limit of integration may be safely extended to $-\infty$ transforming the integral into a standard form:

$$\int_{-\infty}^\infty \frac{x^2 e^x}{(e^x + 1)^2} dx = \frac{\pi^2}{3} \tag{6.61}$$

When the integral in (6.60) is substituted with this value, obtained from tables of standard integrals, we get

$$C_{el} = \frac{\pi^2}{3} D(\varepsilon_F) k_B^2 T \tag{6.62}$$

Using (6.45), we express the density of states per unit volume at Fermi energy as

$$D(\varepsilon_F) = \frac{3n}{2\varepsilon_F} \tag{6.63}$$

Substituting $D(\varepsilon_F)$ from (6.63) in (6.62), we obtain

$$C_{el} = \frac{\pi^2}{2}\frac{nk_B^2 T}{\varepsilon_F} = \frac{\pi^2}{2} nk_B \left(\frac{T}{T_F}\right) \tag{6.64}$$

The measured values of electron heat capacity of metals at low temperature are in consistency with the magnitude and the linear temperature dependence expressed by (6.64). Also, it can be checked that with the use of (6.64) for the value of heat capacity and the Fermi velocity v_F for the electrons thermal speed, we get a Lorenz number $K/\sigma T$ which is in excellent agreement with the experimental value. It is another improvement on the Drude theory. It may, however, be further observed that with the use of Fermi–Dirac statistics the heat capacity has been degenerated or degraded from its large classical value by a factor of about $(\varepsilon_F/3k_BT)$. It is in this sense that the free electron gas is referred to as the *degenerate electron gas*. We will see that the use of *degenerate* and *non-degenerate* in respect of the semiconductors has a different meaning.

At low temperatures below the Debye temperature and the Fermi temperature, the heat capacity of a metal is expressed as

$$C_V = \gamma T + \alpha T^3 \tag{6.65}$$

where the first term represents the electron contribution and the second term is the usual Debye term contributed by phonons. The phonon or ionic contribution dominates at high temperature but drops spectacularly at low temperatures where it becomes even smaller than the electron contribution. In metals at low temperatures the linear term dominates. The parameters γ and α can be determined from the conventional graph between C_V/T and T^2 using the measured values of C_V. The graph is a straight line and drawn for copper in Fig. 6.8. The intercept with the ordinate axis is the measure of γ whereas the slope gives the value of α. The experimental values of γ together with their ratio to the theoretical values for a number of metals are given in Table 6.4.

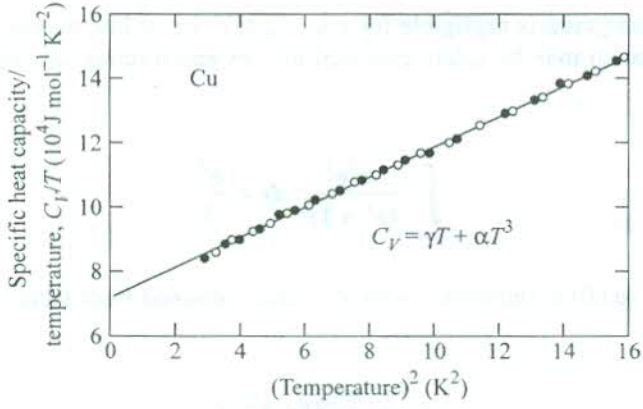


FIG. 6.8 Variation of C_V/T as a function of T^2 for copper. Two sets of experimental points refer to two separate measurements. [After C.A. Bailey, P.L. Smith, *Phys. Rev.*, 114, 1010 (1959).]

Table 6.4 Comparison of the experimental and free electron values of heat capacity constant γ for some metals

Metal	γ_{exp} (mJ mol^{-1} K^{-2})	$\gamma_{exp}/\gamma_{theo.}$
Li	1.7	2.3
Na	1.7	1.5
K	2.0	1.1
Cu	0.69	1.37
Ag	0.66	1.02
Au	0.73	1.14
Al	1.35	1.6
Fe	4.98	10.0
Co	4.98	10.3
Ni	7.02	15.3
Zn	0.64	0.85
Pb	2.98	1.97

We observe a reasonably good agreement for most of the metals shown in Table 6.4 except for the transition metals, Mn, Co and Ni where the estimates from the free electron model show large deviations. This can be interpreted as a result of the strong localized nature of d-electrons occupying a partially filled shell. Since there is little overlap among electron waves, the problem cannot be solved in the free electron model. The solution is provided by the band theory which is the subject of the next two chapters. The point that can be appreciated here is the irrelevance of using the electron rest mass for calculations because the d-electrons behave as heavier ones on account of their tight binding. This fact is effectively considered in the band theory.

6.6 THE SOMMERFELD THEORY OF ELECTRIC CONDUCTION IN METALS

Sommerfeld followed Lorentz approach for calculating the electrical conductivity of metals. This method is based on the use of Fermi–Dirac distribution function for solving the Boltzmann transport equation. A complete rigorous treatment of electrical conduction following this approach is fairly lengthy and complicated. A simplified treatment is outlined in this section to explain the principles involved.

In the absence of a field or thermal gradient, the energy and momentum distribution of the free electron gas is described by the equilibrium distribution function,

$$f^0(p) = \frac{1}{\exp\left[\left(\frac{p^2}{2m} - \varepsilon_F\right)\Big/ k_B T\right] + 1} \tag{6.66}$$

As a simplification it is assumed that the metal is homogeneous, so that f^0 is independent of the spatial coordinates. The application of any field deforms this function leading to a changed function. In transport phenomena there are two opposite mechanisms that compete with each other. In electric

conduction, the electric field distorts the distribution function from its equilibrium form whereas at the same time the scattering of electrons by phonons and defects tries to restore the equilibrium form of the function. If there is no thermal gradient in the system, the net time rate of change of the distribution function can be expressed as

$$\frac{df}{dt} = \left(\frac{\partial f}{\partial t}\right)_{\text{field}} + \left(\frac{\partial f}{\partial t}\right)_{\text{scatt}} \tag{6.67}$$

where the field term may be treated in general sense so as to include the effect of both electric and magnetic fields depending on the situation. In the presence of a thermal gradient there is an additional term on the RHS of (6.67) contributed by the diffusion process and (6.67) is rewritten as

$$\frac{df}{dt} = \left(\frac{\partial f}{\partial t}\right)_{\text{field}} + \left(\frac{\partial f}{\partial t}\right)_{\text{scatt}} + \left(\frac{\partial f}{\partial t}\right)_{\text{diff.}} \tag{6.68}$$

The steady state condition requires the net rate of change of the function to vanish, i.e. $df/dt = 0$, giving

$$\left(\frac{\partial f}{\partial t}\right)_{\text{field}} + \left(\frac{\partial f}{\partial t}\right)_{\text{scatt}} + \left(\frac{\partial f}{\partial t}\right)_{\text{diff.}} = 0 \tag{6.69}$$

The above equation is the required Boltzmann transport equation. The steady state must not be confused with the equilibrium state that refers to the state when no field or thermal gradients are present.

Now, let us apply (6.69) to electronic conduction in a metal assuming that all points within its volume remain at a constant temperature. Then, (6.67) reduces to the form

$$\left(\frac{\partial f}{\partial t}\right)_{\text{field}} + \left(\frac{\partial f}{\partial t}\right)_{\text{scatt}} = 0 \tag{6.70}$$

Under the influence of the electric field $\mathbf{E}_x$ that acts along the negative x-direction, the velocity distribution of electrons would be drifted towards the positive x-direction. The magnitude of drift velocity v_x depends on the strength of the field. The drift of the spherical Fermi distribution of velocities is shown in Fig. 6.9. As a result the equilibrium distribution f^0 approaches a deformed distribution f. Again, on the ground of simplicity we consider the consequences of a uniform field $\mathbf{E}_x$, so that the spatial derivative of $(f - f^0)$ is zero.

The force on an electron owing to the electric field $\mathbf{E}$ along the x-direction is

$$-e\mathbf{E}_x = \frac{\partial \mathbf{p}_x}{\partial t} \tag{6.71}$$

where $\mathbf{p}_x$ denotes the electron's linear momentum along the x-direction.

In view of (6.71), we have

$$\left(\frac{\partial f}{\partial t}\right)_{\text{field}} = \frac{\partial \mathbf{p}_x}{\partial t} \cdot \frac{\partial f}{\partial \mathbf{p}_x} = (-e\mathbf{E}_x)\frac{\partial f}{\partial \mathbf{p}_x} \tag{6.72}$$

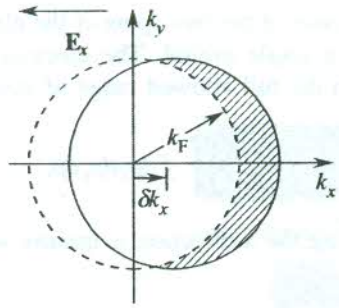


FIG. 6.9 The effect of a steady electric field $\mathbf{E}_x$ on the k-space distribution of quasi-free electrons. The Fermi distribution of equilibrium distribution (dashed) centred at (0, 0, 0) is displaced in the stationary state by an amount $\delta k_x = -e\tau E_x/\hbar$.

The structure of the scattering term in (6.70) is quite complex and cannot be discussed here in completeness as the related theory involves concepts at an advanced level. This term is modelled in the so-called relaxation time approximation. Lorentz made the postulate that the rate at which f returns to the equilibrium distribution f^0 because of scattering is proportional to the deviation of f from f^0, i.e.

$$\left(\frac{\partial f}{\partial t}\right)_{\text{scatt}} \propto f(\mathbf{p}) - f^0(\mathbf{p})$$

which is reduced to

$$\left(\frac{\partial f}{\partial t}\right)_{\text{scatt}} = \frac{f(\mathbf{p}) - f^0(\mathbf{p})}{\tau} \tag{6.73}$$

It should be noted that the relaxation time τ is defined for each location in the momentum space. Assuming that f does not depend on position (i.e. $\nabla_{\mathbf{r}} f = 0$) and using (6.70), (6.72) and (6.73), we can express the stationary non-equilibrium distribution function by

$$f(\mathbf{p}) = f^0(\mathbf{p}) + eE_x\,\tau\left(\frac{\partial f}{\partial \mathbf{p}_x}\right) \tag{6.74}$$

The problem is now reduced to solving this differential equation for f. For this, we make use of the Heisenberg uncertainty principle relating uncertainties in the position (x, y, z) and the momentum (p_x, p_y, p_z) giving

$$\Delta x \cdot \Delta y \cdot \Delta z \simeq \frac{h^3}{\Delta p_x\ \Delta p_y\ \Delta p_z} \tag{6.75}$$

where h is the Planck constant.

Since there is one state in volume $\Delta x \cdot \Delta y \cdot \Delta z$, the number of states per unit volume of the crystal in the momentum range $p + dp$ and $p = \dfrac{\Delta p_x \cdot \Delta p_y \cdot \Delta p_z}{h^3} \times 2$.

A factor of 2 is brought in, because of the two spins of the electron (± 1/2). The values of spin create two degenerate states from a single orbital. The electron density that expresses the total number of states per unit volume in the full allowed range of momenta may thus be written as

$$n = \frac{2}{h^3} \iiint f \, dp_x dp_y dp_z \tag{6.76}$$

The electric current density along the x-direction is usually written as

$$j_x = -nev_x$$

$$= -\frac{2e}{h^3} \iiint \left[f^0 + eE_x \tau \left(\frac{\partial f}{\partial p_x} \right) \right] v_x dp_x dp_y dp_z \qquad \text{[using (6.74) and (6.76)]}$$

$$= -\frac{2e^2}{h^3} E_x \iiint \tau \left(\frac{\partial f}{\partial p_x} \right) v_x dp_x dp_y dp_z \tag{6.77}$$

since f^0 is spherically symmetric and does not contribute to the current.

The volume element $dp_x\, dp_y\, dp_z$ in (6.77) equals the volume of a thin spherical shell in the momentum space centred at the sphere of radius p, i.e.

$$dp_x dp_y dp_z = 4\pi p^2 dp \tag{6.78}$$

Further,

$$\frac{\partial f}{\partial p_x} = \frac{\partial f}{\partial \varepsilon} \cdot \frac{\partial \varepsilon}{\partial p_x} \tag{6.79}$$

with the electron energy

$$\varepsilon = \frac{p^2}{2m} = \frac{p_x^2 + p_y^2 + p_z^2}{2m}$$

which gives

$$\frac{\partial f}{\partial p_x} = \left(\frac{\partial f}{\partial \varepsilon} \right) v_x \tag{6.80}$$

Making use of the relations (6.78) to (6.80), j_x from (6.77) can be expressed as

$$j_x = \frac{8\pi e^2 E_x}{h^3} \int_0^\infty \tau v_x^2 \left(-\frac{\partial f}{\partial \varepsilon} \right) m (2m\varepsilon)^{1/2} \, d\varepsilon \tag{6.81}$$

In order to carry out the integration we make the following assumptions to simplify the treatment:

(i) τ depends only on the magnitude of the electron velocity and not on the direction of motion.

(ii) For small fields,

$$\frac{\partial f}{\partial \varepsilon} = \frac{\partial f^0}{\partial \varepsilon}$$

and

$$v_x^2 \simeq \frac{1}{3} v^2$$

(iii) Since $-\partial f/\partial \varepsilon$ has an appreciable value only in an energy range of a few $k_B T$ about the Fermi level ε_F (see Fig. 6.10), the values of τ and ε that matter in calculations are those at the Fermi energy (ε_F). Therefore, to a good approximation τ and ε under the integral sign may be replaced by $\tau(\varepsilon_F)$ and ε_F in that order making them thus independent of energy.

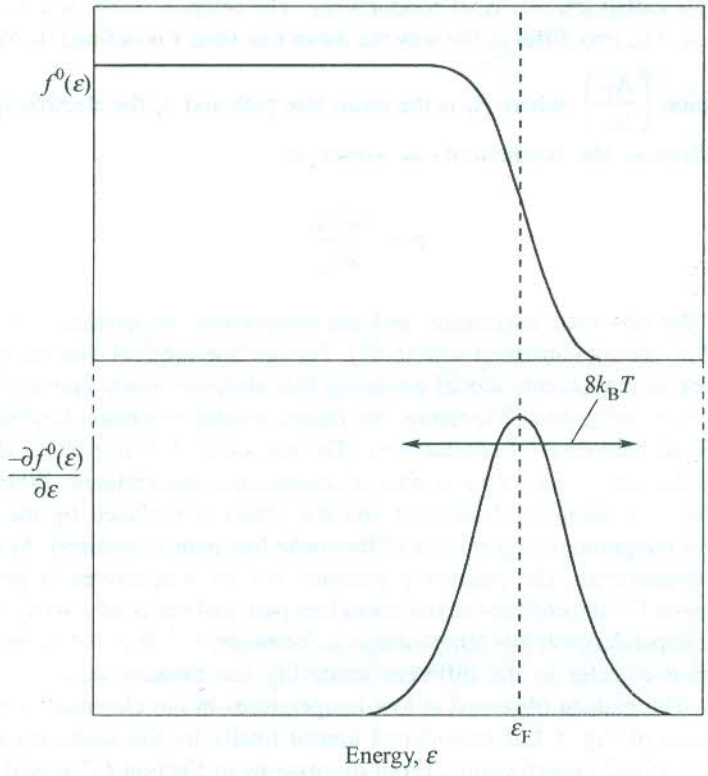


FIG. 6.10 Variation in the Fermi–Dirac distribution function (the occupancy factor) and in its derivative with change in energy.

Applying these assumptions, (6.81) is transformed to:

$$j_x = \frac{16\pi e^2 E_x \tau(\varepsilon_F) \varepsilon_F^{3/2}}{3h^3} (2m)^{1/2} \int_0^\infty \left(-\frac{\partial f^0}{\partial \varepsilon}\right) d\varepsilon \tag{6.82}$$

Using (6.47) and $\int_0^\infty \left(-\frac{\partial f^0}{\partial \varepsilon}\right) d\varepsilon = 1$, we have

$$j_x = \frac{ne^2\tau(\varepsilon_F)}{m} E_x$$

or

$$j_x = \sigma E_x \tag{6.83}$$

where

$$\sigma = \frac{ne^2\tau(\varepsilon_F)}{m} \tag{6.84}$$

The quantity σ is called the electrical conductivity. The relation (6.84) is similar in form to the Drude's relation (6.3). The two differ in the way the mean free time τ is defined. In Sommerfeld model it is defined as the ratio $\left(\frac{\Lambda_F}{v_F}\right)$, where Λ_F is the mean free path and v_F the electron speed at the Fermi level. With this definition, the conductivity is written as

$$\sigma = \frac{ne^2\Lambda_F}{mv_F} \tag{6.85}$$

Unfortunately, the observed magnitude and the temperature dependence of conductivity (see Table 6.1 and Fig. 6.1) are not consistent with (6.85). Though Sommerfeld used the quantum statistics, he continued to work in the Lorentz model assuming that electrons made elastic collisions with ion cores, which is far from being true. Therefore, the theory needed extension beyond this assumption to explain the observed behaviour of conductivity. The variation of σ in (6.85) is determined by the change in Λ_F since the magnitude of v_F is almost temperature independent. When the assumption of electron–ion elastic collision is abandoned and the effect is replaced by the electron–phonon scattering, the correct temperature dependence of the mean free path is obtained. At temperatures well above the Debye temperature, the phonon population for all frequencies is proportional to the temperature. This gives T^{-1} dependence to the mean free path and the conductivity in agreement with the experiment. The dependence at low temperatures is, however, T^{-5} (Bloch–Grüneisen law). The two temperature dependences refer to the different scattering mechanisms active in the two extreme temperature ranges. The plateau observed at low temperatures in the electrical conductivity and the Lorenz number curves of Fig. 6.1 is contributed almost totally by the scattering of electrons from impurities and static crystal imperfections. From discussions in Section 6.7 it will be clear that this scattering is temperature independent. In the high temperature range the electron–phonon collisions dominate over other processes and account alone for the observed behaviour.

Nevertheless, the Sommerfeld model has a unique status in the theory of metals since it identified for the first time the difference between a large number of 'free electrons' and 'the much smaller number of conduction electrons'.

6.6.1 The Hall Coefficient (R_H)

The calculation of the Hall coefficient by Sommerfeld theory is more of academic interest as the estimates from Lorentz theory are also consistent with the measured values. The use of quantum

statistics is of no advantage simply because the expression for Hall coefficient does not finally involve the mean free time. The phenomenon of Hall effect, to which the reader is supposed to have been introduced in an earlier course, needs be discussed for the calculation of Hall coefficient.

Consider a metal crystal in the form of a strip as shown in Fig. 6.11. A steady current of density $\mathbf{j}_x$ flows owing to an electric field $\mathbf{E}_x$ applied along the x-direction in the strip. When a uniform static magnetic field $\mathbf{B}$ directed along the z-direction is switched over the region of the crystal, a small difference of potential (~μV) develops across the crystal's faces along the y-direction. This happens because of the deflection of electrons by the Lorentz force (6.27). These electrons create an electric field along the y-direction in the crystal. As soon as the force owing to this field on electrons becomes equal to the Lorentz force, the deflection of electrons stops resulting in the saturation of the field $\mathbf{E}_y$, known as the Hall field. The effect is known as the Hall effect.

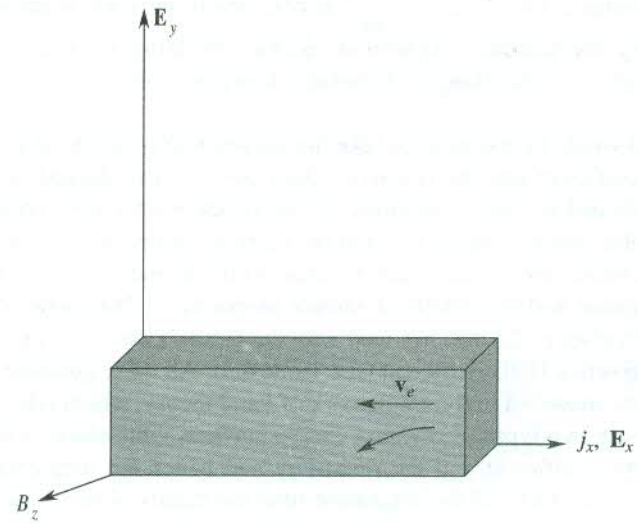


FIG. 6.11 Geometry of the Hall effect. The applied electric field $\mathbf{E}_x$ is along the x-direction. The $\mathbf{E}_y$ denotes the developed Hall voltage. The $\mathbf{v}_e$ represents the electron velocities.

The solution of the Boltzmann transport equation in the above problem is tedious and complex and as such is given here. We straightaway quote the solution. Under the combined influences of $\mathbf{E}_x$ and $\mathbf{E}_y$, the general steady-state solution to the first order in $\mathbf{B}$ is

$$\mathbf{E} = \frac{\mathbf{j}}{\sigma} + \frac{\omega_c \tau}{\sigma |B|} \mathbf{B} \times \mathbf{j} \tag{6.86}$$

There are two components of $\mathbf{E}$, one parallel to $\mathbf{j}$, i.e. $\mathbf{E}_x$ and the other perpendicular to $\mathbf{j}$, i.e. $\mathbf{E}_y$.

$$|E_x| = \frac{|j_x|}{\sigma} \tag{6.87}$$

$$|E_y| = R_H |\mathbf{B}| |\mathbf{j}_x| \tag{6.88}$$

with

$$R_H = \frac{\omega_c \tau}{\sigma B_z} = -\frac{1}{ne} = \frac{E_y}{j_x B_z} \tag{6.89}$$

where ω_c is the cyclotron frequency given by eB_z/m.

Relation (6.89) is the usual expression derived on the basis of much simple arguments. The experimental determination of R_H is based on this relation. The quantities E_y, J_x and B_z are all actually measured. The method of measurement follows from the geometry of Fig. 6.11, described above. Whereas B_z is obtained from the calibration curve of the electromagnet's field for the used pole spacing, E_y and J_x are derived from the measured voltage and current respectively along the y and x directions with the knowledge of sample dimensions. For a certain concentration of electrons, the calculated value of R_H $\left(= -\frac{1}{ne}\right)$ as obtained from (6.89) is about 15 per cent lower than the value given by the Lorentz expression (6.28a), bringing it closer to the measured value. The flagrant discord for the transition metals, however, obscures this marginal success (see Table 6.3).

Furthermore, it would be too naive to take the accuracy of relations (6.86) and (6.89) for granted. In fact, the Hall coefficient and the resistivity both are found to depend on magnetic field and are not constant as indicated by these equations. In the Drude model one expects the Hall field (E_y) to be proportional to the magnetic field (B_z) and the current density (j_x). This keeps the Hall coefficient (R_H) constant. It is rather unbelievable that R_H should not depend on the relaxation time that strongly depends on temperature and the details of sample preparation. The carrier concentration as derived from the measured values of R_H matches well with the Drude's estimates only for alkali metals. Some metals show even positive Hall coefficient (see Table 6.3). All these questions including the magnetic field dependence are answered in the framework of band theory, which relies heavily on the fact that in real solids there are two types of constant energy surfaces—the electron and hole surfaces. In this model both the Hall coefficient and the resistivity and hence the magnetoresistance are shown to depend on the magnetic field and the relaxation time the details of which can be found in Chapter 9. The discussion at this stage, however, cannot be closed without stating that the Hall coefficient and the magnetoresistance do saturate at high magnetic fields under certain conditions to be specified in Chapter 9. An elaborate theory demonstrates that for many metals, this limiting value of Hall coefficient is precisely the same as given by Drude theory.

6.7 MATTHIESSEN'S RULE

The knowledge of different mechanisms of electron collisions is most crucial to the calculation of transport properties of solids. Electrons are mostly scattered by lattice phonons and electrons; and defects involving the impurities and the immobile crystal imperfections. The experimental behaviour of most of the metals reveals that the former processes dominate at room temperature and the latter at low temperatures. The two distinct mechanisms, as we identify, to a good approximation may be treated as independent. This allows us to write the net rate at which the momentum distribution would relax back to the equilibrium distribution after the electric field is switched off as

$$\frac{1}{\tau} = \frac{1}{\tau_l} + \frac{1}{\tau_i} \tag{6.90}$$

Using (6.84), we get

$$\frac{1}{\sigma} = \frac{1}{\sigma_l} + \frac{1}{\sigma_i}$$

or

$$\rho = \rho_l + \rho_i \tag{6.91}$$

where ρ denotes the resistivity with ρ_l and ρ_i referring to the contributions from the lattice and the impurities, respectively.

Therefore, in the presence of several scattering mechanisms the resistivity is simply the sum of the resistivities that we would have if each were present alone.

The conductivity plot of Fig. 6.1 and the relative resistance plots of Fig. 6.12 show a plateau below 10 K where the electron collisions with thermal phonons and electrons are almost frozen. The resistivity in this temperature range, that appears as temperature independent, is totally contributed by electron collisions with the impurities and the static imperfections. The value of ρ_i is estimated by extrapolating the curve back to 0 K in Fig. 6.12. The values of ρ_l at different temperatures are obtained by subtracting ρ_i from the measured values of ρ (6.91). The temperature dependence of ρ_l is in accordance with the theory of collisions of electrons with phonons and electrons. For small concentrations, ρ_l shows no dependence on the defect concentration.

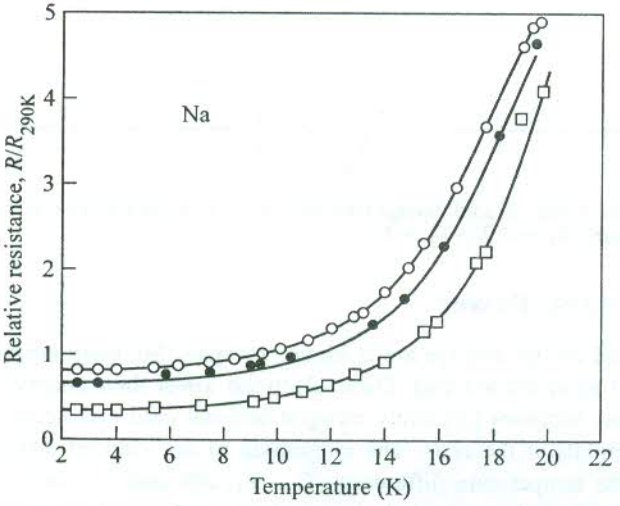


FIG. 6.12 Behaviour of electrical resistance of sodium compared to the value at 290 K with change in temperature. The experimental points marked as ○, ●, □ belong to three different samples with differing defect concentrations. [After D.K.C. McDonald, K. Mendelssohn, *Proc. R. Soc.*, Edinburgh Sect., A 202, 103 (1950).]

The observation that ρ_i is temperature independent is established as the Matthiessen's rule. Matthiessen observed that the resistivity of a metal increases with the increase in defect concentration. This is confirmed by the experimental curves of sodium in Fig. 6.12 where the curve with higher intercept with the resistance axis refers to the sample of higher defect content.

The assumption that mechanisms of the two collision processes under consideration are independent is by and large vindicated by observations on whose basis the Matthiessen's rule is founded. But it is not difficult to show that the Matthiessen's rule breaks down even in the relaxation

time approximation which depends on the electron wavevector **k**. On the other hand, the realistic picture of collisions warrants the assumptions of the relaxation time approximation to be set aside, as a result of which the two collision mechanisms have a rare chance of being independent and the Matthiessen's rule is in general rendered invalid. The presence of the two competing collision mechanisms can seriously affect the configuration of electrons that control the actual collision rate. It could only be fortuitous that the distribution function in the presence of each separate mechanism be the same and therefore, the failure of the Matthiessen's rule be prevented.

6.8 THERMOELECTRIC EFFECTS

The thermoelectric behaviour of metals is described mainly in terms of three parameters: (i) the thermoelectric power, (ii) the Thomson coefficient, and (iii) the Peltier coefficient. These parameters will be defined in this section on the basis of the thermoelectric response of a conducting rod whose ends are maintained at different temperatures T_1 and T_2 ($T_1 < T_2$), as shown in Fig. 6.13.

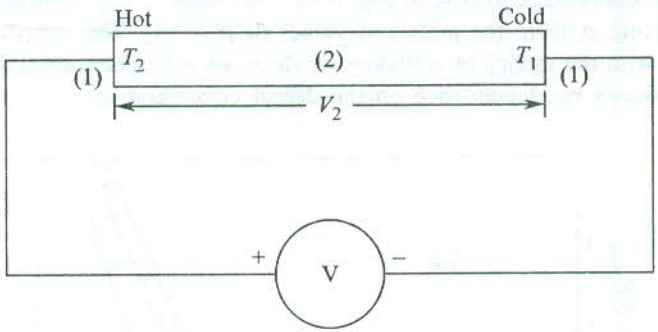


FIG. 6.13 Measurement of the induced voltage (thermo e.m.f.) in a conducting rod whose ends are maintained at temperatures T_1 and T_2 ($T_2 > T_1$).

6.8.1 Thermoelectric Power

Electrons at the hot end on the average are at higher energies. So, there are more electrons in levels above the Fermi level ε_F at the hot end. These electrons lower their energy by drifting to the cold end.† The hot end thus becomes positively charged and the cold end negatively charged, creating different electric potentials at the ends. The magnitude of the induced difference of potential (V_2) depends entirely on the temperature difference ($T_2 - T_1$) and remains unaffected by any change in the temperature distribution along the rod. A voltmeter connected to measure the induced voltage shows no deflection if the connection wires (1) are of the same material as that of the rod. If this material is different, the induced potential difference between the ends of the wires (1), in contact with the ends of the rod, is no more V_2 but some other value, say V_1. The voltmeter reading in this case will be ($V_1 - V_2$), where V_1 and V_2 refer to the difference of potentials in open circuit.

The difference of potential ($V_1 - V_2$) denoted by V_{12}, rises monotonically with the temperature difference ($T_2 - T_1$). The rate of change of the induced voltage with the temperature difference defines

† Even in metals a part of heat is conducted by phonons, which are the main carriers of heat in insulators [see Section 5.3.3].

the thermoelectric power S_{12} of the junction (1, 2). If ΔV_{12} be the rise in V_{12} because of the small increase ΔT in $(T_2 - T_1)$, then

$$S_{12} = \frac{dV_{12}}{dT} = \frac{dV_1}{dT} - \frac{dV_2}{dT} = S_1 - S_2 \tag{6.92}$$

where S_1 and S_2 are the bulk thermoelectric powers of materials to which the wires and the rod, respectively, belong. The V_{12} is often referred to as the *Seebeck potential.* Relation (6.92) most significantly emphasizes that the thermoelectric power of a junction is not a property of the junction. It is immaterial whether the junction is soldered, brazed, spot-welded or fused; S_{12} depends only on the bulk properties S_1 and S_2 of the materials in contact at the junction.

6.8.2 The Thomson Effect

Let an electric current be forced into the rod (Fig. 6.14) between its ends. When the conventional current flows down the rod from the hot end to the cold end, the electrons reaching the hot end raise their potential energy by absorbing heat from the hot end. Similarly, for the reverse conventional current, the electrons lower their potential energy at the cold end by emitting heat as they arrive there from the hot end. Thus there occurs an absorption or evolution of heat on the passage of current through a rod whose ends are maintained at different temperatures. This amount of heat is in addition to the power loss caused by Joule's heating and the heat flowing because of the presence of a temperature gradient. This absorption or evolution of heat is known as the *Thomson effect.* It does not depend on the nature of the junction involved.

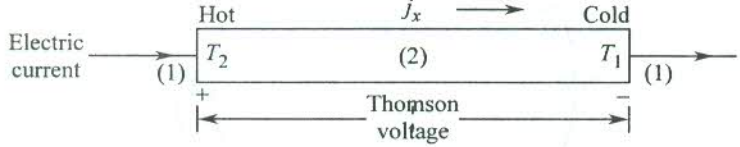


FIG. 6.14 Schematic of the Thomson effect. The Thomson voltage produced because of the temperature gradient is in addition to the Ohm's law voltage.

It is observed that the heat evolved per unit volume per unit time dQ/dt is proportional to the temperature gradient $-dT/dx$ for a constant current density j_x down the length of the rod (see Fig. 6.14). Also, under the condition of a fixed temperature gradient the dQ/dt is found to vary linearly with j_x. These variations may accordingly be expressed as

$$\frac{dQ}{dt} \propto -\frac{dT}{dx}, \text{ with current density } j_x \text{ maintained at a constant value.}$$

$$\frac{dQ}{dt} \propto j_x, \text{ with } -\frac{dT}{dx} \text{ maintained at a fixed value.}$$

Therefore, we may express dQ/dt as

$$\frac{dQ}{dt} = -\mu_T j_x \left(\frac{dT}{dx}\right) \tag{6.93}$$

where the constant of proportionality μ_T is called the *Thomson coefficient.*

The sign of Q, the heat evolved, is taken as positive here.

6.8.3 The Peltier Effect

An electric current density j_x through a conductor down its length (in the x-direction) at constant temperature always has a heat current density j_Q associated with it such that

$$j_Q \propto j_x$$

or

$$j_Q = \Pi j_x \tag{6.94}$$

where Π is called the *Peltier coefficient.*

Relation (6.94) implies that the heat current density will undergo a change at a junction of two conductors 1 and 2 when an electric current flows across the junction because Π is different for the two conductors. The change in heat flux results in absorption or evolution of heat at the junction, depending on the direction of the electric current. When an electric current is forced into a thermocouple, the heat is absorbed at one junction and evolved at the other because the direction of the electric current between conductor 1 and conductor 2 is reversed at the other junction (Fig. 6.15). This is known as the *Peltier effect.* The Peltier effect essentially involves the pumping of heat energy from one junction to the other by making an electric current flow through a thermocouple.

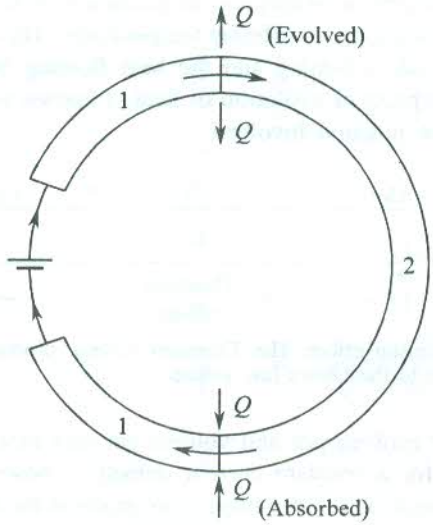


FIG. 6.15 Schematic of the Peltier effect. The heat is shown to be evolved at a thermocouple junction 1–2 when a current flows from the conductor 1 to the conductor 2. The heat is absorbed at the junction 2–1 when the current flows from the conductor 2 to the conductor 1 (the current direction is reversed).

It is appropriate to denote the Peltier coefficient at a junction by Π_{12} and define it as the reversible heat evolved or absorbed per unit time per unit electric current flowing from conductor 1 to conductor 2 at their junction. According to this definition, we have

$$\Pi_{12} = -\Pi_{21} \qquad \text{with} \qquad \Pi_{12} = \Pi_1 - \Pi_2$$

Like the Thomson effect, neither the Peltier effect nor the Seebeck effect depends on the nature of the junction.

6.8.4 Kelvin (Thomson) Relations

Thomson (famous as Lord Kelvin) used the principles of classical thermodynamics to show that the three thermoelectric parameters S, μ_T and Π are not independent and that only one is needed to specify the other two. He derived the following relations, commonly known as Kelvin relations:

$$\mu_T = T\frac{dS}{dT} \tag{6.95}$$

$$\Pi = TS \tag{6.96}$$

In the above relations, S stands for the absolute thermoelectric power and not the entropy. Although the laws of classical thermodynamics are not directly applicable to thermoelectricity, the Kelvin relations stand the test of experiments extremely well. If either S or μ_T is known as a function of temperature, the remaining two parameters can be obtained from (6.95) and (6.96). In the latter case, relation (6.95) is integrated to get the absolute value of S at any temperature, i.e.

$$S(T) = \int_0^T \frac{\mu_T}{T}\,dT \tag{6.97}$$

Here, we have ignored $S(0)$ because the third law of thermodynamics implies that all thermoelectric effects vanish at absolute zero.

A few striking gains notwithstanding, the limitations of the free electron theory are pronounced by its varying degree of success in accounting for several properties of metals. Some of these might not have even figured in this chapter. Satisfactory explanations to all such properties and those of solids in general are provided by a less naive theory which forms the subject of the next two chapters.

SUMMARY

1. According to the Drude model, the d.c. electrical conductivity of metals is

   $\sigma = \dfrac{ne^2\Lambda}{(3m\,k_B T)^{1/2}}$, where Λ is the electron mean free path.

   Experimentally, it is observed that

   $$\sigma \propto T^{-1}.$$

2. The Drude model correctly describes the Wiedemann–Franz law, but gives a wrong magnitude of the Lorenz number L. The value of L as defined in the Drude theory is about half the measured value and is given by

   $L = \dfrac{K}{\sigma T} = \dfrac{3}{2}\left(\dfrac{k_B}{e}\right)^2$, where K is the thermal conductivity.

3. The dc electrical conductivities in the Drude model and the Lorentz model are related as

$$\sigma_D = \left(\frac{3\pi}{8}\right)^{1/2} \sigma_L, \quad \text{D: Drude}$$

L: Lorentz

4. According to the Lorentz model, the Hall coefficient of metals is given by

$$R_H = -\left(\frac{3\pi}{8}\right)\frac{1}{ne}$$

The measured values of the Hall coefficient are close to the values given by

$$R_H = -\frac{1}{ne}$$

R_H is measured using the formula:

$R_H = \dfrac{E_y}{j_x B_z}$, where the Hall field E_y, the current density j_x and the magnetic field B_z are mutually perpendicular along the y, x and z directions, respectively.

5. The Fermi–Dirac distribution function is

$$f(\varepsilon) = \frac{1}{\exp[(\varepsilon - \mu)/k_B T)] + 1}$$

where μ is the *chemical potential,* defined by $\lim_{T \to 0} \mu = \varepsilon_F$; ε_F is the Fermi energy.

6. In the Sommerfeld model for metals:

(i) The density of states is

$$D(\varepsilon) = \frac{V}{2\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2} \varepsilon^{1/2} \left(= \frac{3N}{2\varepsilon_F} \text{ at } \varepsilon = \varepsilon_F\right)$$

with $\varepsilon_F = \dfrac{\hbar^2}{2m}(3\pi^2 n)^{2/3}$, where $n = \dfrac{N}{V}$.

(ii) The electron heat capacity is

$$C_{el} = \frac{\pi^2}{2} n k_B \left(\frac{T}{T_F}\right)$$

(iii) The dc electrical conductivity is

$$\sigma_0 = \frac{ne^2\tau(\varepsilon_F)}{m} = \frac{ne^2\Lambda_F}{m v_F}$$

and the ac conductivity is,

$$\sigma = \frac{\sigma_0}{(1 - i\omega\tau)}$$

7. The resistivity of a metal is expressed as

$$\rho = \rho_l + \rho_i$$

where l and i distinguish the lattice (or phonon) and impurity contributions.

The observation that ρ_i is temperature independent is established as the Matthiessen's rule.

8. Heat absorbed or evolved in the Thomson effect is in addition to the energy loss caused by Joule's heating (produced owing to the electric current, forced externally) and the heat flowing by virtue of the temperature gradient.

9. The commonly known Kelvin relations are:

$$\mu_T = T\frac{dS}{dT}$$

$$\Pi = TS$$

where μ_T is the Thomson coefficient and S and Π denote the thermoelectric power and the Peltier coefficient, respectively.

PROBLEMS

6.1 Assuming Drude model:

(a) Show that the probability for an electron, picked at random at any time, not to have made a collision during the preceding t is given by $e^{-t/\tau}$.

(b) Show that at any moment, the time between the last and the next collision averaged over all electrons is 2τ. In view of this result, comment on the Drude's mistake of a factor of 1/2 in the calculation of electrical conductivity.

6.2 Consider a spherical body of mass m and charge q moving through a viscous fluid with a constant velocity ($\propto q\mathbf{E}_x$) under the action of an electric field $\mathbf{E}_x$. A small magnetic field $\mathbf{B}_z$ is now applied. Obtain equations of motion in the xy-plane and find the radius of curvature of the trajectory.

6.3 (a) Plot $f(\varepsilon)$ as a function of $(\varepsilon - \varepsilon_F)$ in units of k_BT at room temperature in the range $-3k_BT \leq \varepsilon \leq 3k_BT$.

(b) Examine at what values of ε, the $f(\varepsilon)$ differs appreciably from unity.

6.4 Give a comparison of the equilibrium distribution functions representing Maxwell–Boltzmann and Fermi–Dirac statistics. Explain why only the former depends on the particle density.

6.5 Calculate the density of states for a two-dimensional gas of free electrons in a so-called quantum well. The boundary conditions for the electronic wave function are: $\psi(x, y, z) = 0$ for $|x| > a$, where a is of atomic dimensions.

6.6 Show that the fraction of electrons within k_BT of the Fermi level is equal to $3k_BT/2\varepsilon_F$, if $D(\varepsilon) \sim \varepsilon^{1/2}$.

6.7 The magnitude of the electron wavevector corresponding to the top most level filled in a system of free electrons is equal to $(2\pi^2)^{1/3}/a$, where a is the interatomic separation. If each level is three-fold degenerate, calculate the number of conduction electrons per atom.

6.8 Derive relations for the electron density at which the Fermi surface first touches the zone boundary in Na and Cu metals.

6.9 Use the free electron theory to calculate the Fermi energy of Na and Al metals. Their lattice constants are 4.3 Å and 4.0 Å, respectively.

6.10 Calculate the Fermi energy and Fermi temperature of liquid He^3 whose density near 0 K is 81 kg m^{-3}. (He^3 is a fermion with atomic spin 1/2.)

6.11 At what temperature does the electron heat capacity become larger than the phonon heat capacity? Express this temperature in terms of the Debye temperature and the electron density.

6.12 Calculate the electron heat capacity at 1000 K in Na, Al and Cu metals using the following data:

	Electron density $n(\times 10^{28}\ m^{-3})$	*Fermi energy* ε_F(eV)
Na	2.5	3.1
Al	18.0	11.7
Cu	8.5	7.1

Estimate the fraction it forms of the total heat capacity, assuming that 1000 K is well above the Debye temperature in each case.

6.13 The electrical conductivity of a pure gold crystal is 5×10^7 ohm^{-1} m^{-1} at 273 K. Gold has a close-packed FCC structure with atomic radius as 1.44 Å. Each atom contributes one conduction electron. If a gold crystal has 0.1 per cent of randomly distributed vacancies, estimate the electron mean-free path as determined by scattering from vacancies alone. Calculate the mean-free path for phonon-scattering at 273 K from the given value of conductivity. What is the revised value of conductivity according to Matthiessen's rule?

SUGGESTED FURTHER READING

Ashcroft, N.W. and N.D. Mermin, *Solid State Physics* (Saunders College, 1988).

Ibach, H. and H. Lüth, *Solid State Physics* (Springer-Verlag 1995).

Jain, J.P., in *Solid State Physics,* Vol. 5, F. Seitz and D. Turnbull (Eds.) (Academic Press, 1957).

Lorentz, H.A., *The Theory of Electrons* (Dover, 1952).

Meaden, G.T., *Electrical Resistance of Metals* (Plenum Press, 1965).

Resenberg, H.M., *Low Temperature Solid State Physics* (Oxford, 1963).

Ziman, J.M., *Principles of the Theory of Solids* (Cambridge, 1972).

Chapter 7

Electron Energy Bands

In Chapter 6, we gave an account of the degree of success achieved by the free electron theory in explaining some properties of metals. At the same time we also came across observations to which even the quantized free electron theory did not have the answers. To name a few, the positive Hall coefficient and the wrong hint at the number of conduction electrons per atom (Table 6.3) in some metals exposed weak links in the theory. In addition, there are more vital questions like why some materials are metals while others are good insulators. Their electrical conductivities may differ by a magnitude of order 10^{30} which is unusually high for any physical property. Complex transport phenomena in the presence of a magnetic field too, do not find satisfactory explanation in the frame of the free electron theory. The continuous parabolic energy curve (see Fig. 6.4) renders us helpless in finding a clue to the sharp resonance-like structures observed in the optical spectra of solids. This implies that the discreteness of the atomic energy levels must be accounted for in the theory.

Solutions to the above problems may be sought by correcting the free electron theory for certain simplifications that form its basis. Luckily, efforts made in this direction bring about a remarkable transformation of the theory in terms of success. The theory that emerges as a result of pruning the simplifications in the free electron model has come to be known as the *band theory of solids.* To begin with, the constant potential energy in the Sommerfeld theory is replaced by a non-vanishing potential energy term that is periodic in the crystal, i.e. the rectangular potential box in Section 6.4 is replaced by an infinite periodic potential. This ignores the surface effects and requires the electron to observe the periodicity of the crystal lattice.

In this chapter, we solve the one electron time-independent Schrödinger wave equation to obtain the allowed electron energy states. The states appear as bunches of closely spaced levels forming energy bands. Two successive energy bands are separated by a region of forbidden energy, known as the *forbidden energy gap* or simply the *band gap.* One of the most special features of the band theory is that it distinguishes between electrons in the outermost shell and those in the inner shells by treating the latter as heavier. This proves enormously profitable in dealing with properties that are predominantly controlled by electrons in a certain single shell.

7.1 CONSEQUENCES OF PERIODICITY

In light of the above arguments, we consider a periodic potential energy function as

$$V(\mathbf{r}) = V(\mathbf{r} + \mathbf{t}_n) \tag{7.1}$$

where $\mathbf{t}_n$ is an arbitrary translation vector in the direct crystal lattice usually defined by (1.2) as

$$\mathbf{t}_n = n_1\mathbf{a} + n_2\mathbf{b} + n_3\mathbf{c}$$

Being periodic in the crystal lattice, the potential $V(\mathbf{r})$ may be expanded in a Fourier series and expressed as

$$V(\mathbf{r}) = \sum_{\mathbf{g}} V_{\mathbf{g}} \exp\ (i\mathbf{g} \cdot \mathbf{r}) \tag{7.2}$$

where $\mathbf{g}$ denotes a reciprocal lattice vector.

Our first aim is to look for an appropriate solution to the following one electron time-independent Schrödinger equation:

$$H\Psi(\mathbf{r}) = \left[-\frac{\hbar^2}{2m} \nabla^2 + V(\mathbf{r}) \right] \Psi(\mathbf{r}) = \varepsilon \Psi(\mathbf{r}) \tag{7.3}$$

Let us try a general plane-wave expansion,

$$\Psi\ (\mathbf{r}) = \sum_{\mathbf{k}} C_{\mathbf{k}} \exp\ (i\mathbf{k} \cdot \mathbf{r}) \tag{7.4}$$

as the solution to (7.3). Substituting (7.2) and (7.4) in (7.3), we get

$$\sum_{\mathbf{k}} \left(\frac{\hbar^2 k^2}{2m} \right) C_{\mathbf{k}} \exp\ (i\mathbf{k} \cdot \mathbf{r}) + \sum_{\mathbf{k}} \sum_{\mathbf{g}} C_{\mathbf{k}} V_{\mathbf{g}} \exp\ [i\ (\mathbf{k} + \mathbf{g}) \cdot \mathbf{r}] = \varepsilon \sum_{\mathbf{k}} C_{\mathbf{k}} \exp\ (i\mathbf{k} \cdot \mathbf{r}) \tag{7.5}$$

Equating the coefficients of $\exp\ [i\mathbf{k} \cdot \mathbf{r}]$ on both sides, we obtain

$$\left(\frac{\hbar^2 k^2}{2m} - \varepsilon \right) C_{\mathbf{k}} + \sum_{\mathbf{g}} V_{\mathbf{g}} C_{\mathbf{k}-\mathbf{g}} = 0 \tag{7.6}$$

In analogy with phonons (Section 4.2.1), the total number of allowed k-values for the electrons waves, when subjected to Born–von Karman periodic boundary condition, will be equal to the total number of unit cells, say N. Relation (7.6) represents one of the N algebraic equations which as a set have replaced the wave equation (7.3). Each of the equations (7.6) that corresponds to a certain value of k couples only those expansion coefficients $C_{\mathbf{k}}$ in (7.4) whose k-values differ from each other by a reciprocal vector $\mathbf{g}$, i.e. $C_{\mathbf{k}}$ is coupled to $C_{\mathbf{k}-\mathbf{g}}$, $C_{\mathbf{k}-\mathbf{g}'}$, $C_{\mathbf{k}-\mathbf{g}''}$. The solution to each of the N equations (7.6) may be interpreted as a superposition of plane waves of wavevectors $\mathbf{k} - \mathbf{g}$, $\mathbf{k} - \mathbf{g}'$, $\mathbf{k} - \mathbf{g}''$, etc. The wavefunction with eigenvalue $\varepsilon_{\mathbf{k}}$ may thus be written as

$$\Psi_k(\mathbf{r}) = \sum_{\mathbf{g}} C_{\mathbf{k}-\mathbf{g}} \exp\ [i(\mathbf{k} - \mathbf{g}) \cdot \mathbf{r}] \tag{7.7}$$

or

$$\Psi_k(\mathbf{r}) = \sum_{\mathbf{g}} C_{\mathbf{k}-\mathbf{g}} \exp\ (-i\mathbf{g} \cdot \mathbf{r}) \cdot \exp\ (i\mathbf{k} \cdot \mathbf{r})$$

$$= u_{\mathbf{k}}(\mathbf{r}) \exp(i\mathbf{k} \cdot \mathbf{r}) \tag{7.8a}$$

with

$$u_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{g}} C_{\mathbf{k}-\mathbf{g}} \exp\ (-i\mathbf{g} \cdot \mathbf{r}) \tag{7.8b}$$

From (7.8a) we infer that $u_{\mathbf{k}}(\mathbf{r})$ as defined by (7.8b) modulates the plane waves $\exp\ (i\mathbf{k} \cdot \mathbf{r})$ to a

form that serves as a solution to the wave equation (7.3) for a single electron state k. It is essentially the statement of the Bloch theorem defining the Bloch function $\psi_{\mathbf{k}}(\mathbf{r})$ which is written as

$$\Psi_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r}) \exp(i\mathbf{k} \cdot \mathbf{r}) \tag{7.8c}$$

Relation (7.8c) represents the proper wavefunction that we were looking for. We notice that $u_{\mathbf{k}}(\mathbf{r})$ is periodic in the direct crystal lattice. That is,

$$u_{\mathbf{k}}(\mathbf{r} + \mathbf{t}_n) = u_{\mathbf{k}}(\mathbf{r}) \qquad (\text{since } \exp(-i\mathbf{g} \cdot \mathbf{t}_n) = 1). \tag{7.9}$$

7.1.1 Proof of the Bloch Theorem

The k-value associated with the Bloch function $\Psi_{\mathbf{k}}(\mathbf{r})$ belongs to the set of values (4.9) allowed by the periodic boundary condition. The validity of the Bloch theorem would be proved if this set of k-values is reproduced by subjecting the Bloch function (7.8c) to the periodic boundary condition. For the sake of simplicity we provide the proof for a non-degenerate wavefunction in a monatonic one-dimensional crystal. Consider a crystal of length L along the x-direction having N atoms with interatomic spacing $\mathbf{a}$. The eigenstate of an electron with wavevector $\mathbf{k}$ near an atom positioned at $\mathbf{x}$ is given by the Bloch function,

$$\Psi_{\mathbf{k}}(\mathbf{x}) = u_{\mathbf{k}}(\mathbf{x}) \exp(i\mathbf{k} \cdot \mathbf{x}) \tag{7.10}$$

When the electron moves the distance a along the x-direction, the wavefunction changes to

$$\Psi_{\mathbf{k}}(\mathbf{x} + \mathbf{a}) = u_{\mathbf{k}}(\mathbf{x} + \mathbf{a}) \exp[i\mathbf{k} \cdot (\mathbf{x} + \mathbf{a})]$$

$$= Cu_{\mathbf{k}}(\mathbf{x}) \exp(i\mathbf{k} \cdot \mathbf{x}) \qquad (\text{from } 7.9)$$

$$= C\Psi_{\mathbf{k}}(\mathbf{x}) \tag{7.11}$$

where

$$C = \exp(i\mathbf{k} \cdot a) \tag{7.12a}$$

If we imagine the crystal to be in the form of a ring, N steps of the above displacement or translation will bring the electron back to the same atom at x from where it started. This means that

$$\Psi_{\mathbf{k}}(\mathbf{x} + N\mathbf{a}) = \Psi_{\mathbf{k}}(\mathbf{x} + \mathbf{L}) = \Psi_{\mathbf{k}}(\mathbf{x}) \tag{7.12b}$$

But, the net phase-shift in the wavefunction on the basis of (7.11) requires,

$$\Psi_{\mathbf{k}}(\mathbf{x} + \mathbf{L}) = C^N \Psi_{\mathbf{k}}(\mathbf{x}) \tag{7.13}$$

For (7.12b) to be satisfied,

$$C^N = 1 = \exp(i2\pi n)$$

or

$$C = \exp\left(\frac{i2\pi n}{N}\right) \tag{7.14}$$

Using (7.12a), we have

$$k = \frac{2\pi n}{Na} = n\frac{2\pi}{L} \tag{7.15}$$

with

$$n = 0, \pm 1, \pm 2, ..., \pm\left(\frac{N}{2} - 1\right), \frac{N}{2} \tag{7.16}$$

$$\text{for } n = \pm \frac{N}{2}, \quad k = \pm \frac{\pi}{a}$$

These two values of k differ by a reciprocal lattice vector $2\pi/a$ (shortest for a one-dimensional lattice). We now show immediately below that the Bloch function is periodic in the reciprocal lattice ($\Psi_{\mathbf{k}+\mathbf{g}} = \Psi_{\mathbf{k}}$). Because of this reason the two values of $\mathbf{k}$ correspond to the same wavefunction and, therefore, only one sign is included for this value of n in (7.16). The set of k-values (7.15) takes the following form for a three-dimensional crystal.

$$k_x = n_x \frac{2\pi}{L}; \quad k_y = n_y \frac{2\pi}{L} \quad \text{and} \quad k_z = n_z \frac{2\pi}{L} \tag{7.17}$$

with n_x, n_y, n_z defined by (7.16).

The set (7.17) is identical with (4.9) derived for phonons. This confirms that the Bloch function (7.8c) is an appropriate wavefunction to describe electrons moving in the periodic crystal potential.

7.1.2 The Periodicity of the Bloch Functions and Their Eigenvalues

It is straightforward to show that the Bloch functions are periodic in the reciprocal lattice, the periodicity being given by the reciprocal lattice vectors. When the electron wavevector changes from k to $k + g$, the new wavefunction is

$$\Psi_{\mathbf{k}+\mathbf{g}}(\mathbf{r}) = \sum_{\mathbf{g}'} C_{\mathbf{k}+\mathbf{g}-\mathbf{g}'} \exp(-i\mathbf{g}' \cdot \mathbf{r}) \exp(i(\mathbf{k}+\mathbf{g})) \cdot \mathbf{r}$$

$$= \left(\sum_{\mathbf{g}''} C_{\mathbf{k}-\mathbf{g}''} \exp(-i\mathbf{g}'' \cdot \mathbf{r})\right) \cdot \exp(i\mathbf{k} \cdot \mathbf{r}) \qquad (\text{with } \mathbf{g}' - \mathbf{g} = \mathbf{g}'')$$

or

$$\Psi_{\mathbf{k}+\mathbf{g}}(\mathbf{r}) = \Psi_{\mathbf{k}}(\mathbf{r}) \qquad (\text{using 7.8a}) \tag{7.18}$$

Relation (7.18) states that the wavefunction repeats for a change in $\mathbf{k}$ by a reciprocal lattice vector.

Now, the Schrödinger equation (7.3) for the eigenstate $\Psi_{\mathbf{k}+\mathbf{g}}(\mathbf{r})$ is

$$H\Psi_{\mathbf{k}+\mathbf{g}}(\mathbf{r}) = \varepsilon_{\mathbf{k}+\mathbf{g}} \Psi_{\mathbf{k}+\mathbf{g}}(\mathbf{r}) \tag{7.19}$$

But in view of (7.18), we have

$$H\Psi_{\mathbf{k}}(\mathbf{r}) = \varepsilon_{\mathbf{k}+\mathbf{g}} \Psi_{\mathbf{k}}(\mathbf{r}) \tag{7.20}$$

Comparing (7.20) with (7.3), we get

$$\varepsilon_{\mathbf{k}+\mathbf{g}} = \varepsilon_{\mathbf{k}} \tag{7.21}$$

Relations (7.18) and (7.21) express that the Bloch functions and their eigenvalues are periodic in the reciprocal lattice space. Energy value $\varepsilon_{\mathbf{k}}$ is obtained in terms of $\mathbf{k}$ which may be

interpreted as a quantum number. In view of the periodic character of $\Psi_{\mathbf{k}}(\mathbf{r})$ and $\varepsilon_{\mathbf{k}}$, we need to know the allowed values of the quantum number k only within the first Brillouin zone for a complete description of these functions throughout the k-space. The periodic energy surfaces ε_k give rise to the electron energy bands. We discuss the formation of energy bands and band gaps in the next two sections.

7.2 WAVE MECHANICAL INTERPRETATION OF ENERGY BANDS

A crystal may be regarded to have been formed by bringing a large number of isolated atoms into a close assembly. An isolated atom has a set of discrete energy levels. Let ϕ_A represent the wavefunction of an electron in an atom A and ϕ_B in another atom B of a crystal with monatomic basis. When atoms A and B come closer, the electron shares its presence between the regions of A and B. The electron oscillates with a well-defined frequency between the two atoms. In the language of wave mechanics, we say that the electron waves ϕ_A and ϕ_B overlap. The two states between which the shared electron resonates may be approximately expressed as

$$\Psi(\pm) = \frac{1}{\sqrt{2}} (\phi_A \pm \phi_B) \tag{7.22}$$

If the electron energy in an isolated atom be ε_0, the eigenvalues of the above eigenfunctions are written as

$$\varepsilon(\pm) = \varepsilon_0 \pm \varepsilon_1 \tag{7.23}$$

The two values differ by $2\varepsilon_1$, such that

$$2\varepsilon_1 = \hbar\omega \tag{7.24}$$

where ω is the frequency with which the electron shuttles back and forth between the two states $\psi(\pm)$.

The ε_1 is a measure of the electron's interaction with the core of the two atoms. We must be conscious of the fact that we continue to work in the independent electron approximation. For atoms A and B at large distances (isolated), ε_0 is a doubly degenerate level associated with eigenfunctions ϕ_A and ϕ_B. When a large number of atoms N are assembled to make the crystal, an electron is supposed to be shared among N atoms. This splits the N-fold degenerate ε_0 level into N closely-spaced levels. A similar splitting occurs for each of the other discrete levels of an isolated atom. The group of closely-spaced N levels is called an *energy band* (Fig. 7.1). For a silicon crystal of 1 cm^3, the spacing between the successive levels in a band is of the order of 10^{-23} eV, a negligibly small magnitude. Therefore, the variation of energy within a band may be taken as continuous in the first approximation. The splitting increases with the decrease in the interatomic spacing (see Fig. 7.1).

7.3 THE KRONIG–PENNEY MODEL

In view of the discrete energy level scheme of isolated atoms, it is an unlikely possibility that the energy bands would be infinitely continuous. This is really the case, in general, as there is observed a region of forbidden energies between the two successive bands. However, the bands overlap in some cases. An energy band is almost centred around its parent level, the N-fold degenerate level in an isolated atom of a crystal composed of N atoms. Kronig and Penney (1930) demonstrated, after

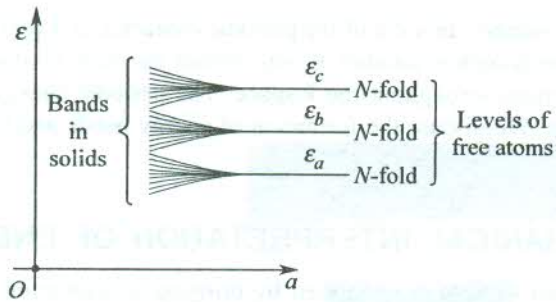


FIG. 7.1 Formation of energy bands in solids. When N atoms are brought together to form a solid, each atomic level is N-fold degenerate for large interatomic separations. The atomic levels ε_a, ε_b, ε_c, etc. spread into bands at a small interatomic separation.

the Bloch theorem (1928) came into existence, that regions of forbidden energies intervene the regions of allowed energies. They accomplished the task using a one-dimensional square-well crystal potential as depicted in Fig. 7.2.

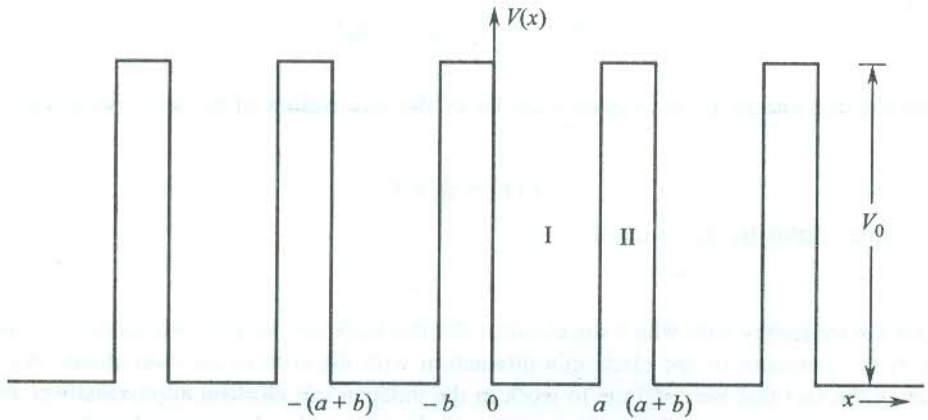


FIG. 7.2 One-dimensional periodic electron potential energy used in the Kronig–Penney model.

There may arise two extreme situations: One in which the barrier is thin, such that bV_0 is vanishingly small and the other when the barrier is thick. In the first case, the electron can move freely through the potential at any electron energy given by the free electron dispersion relation (6.40a). In the second case, the electron with energy less than V_0 has little chance of tunnelling from one atom space to the next and (6.40b) serves as the valid dispersion relation with 'a' in place of L. This extreme case refers to the problem of an electron in the potential box (Section 6.4).

The Kronig–Penney model describes an intermediate situation where V_0 may be large but 'b' is so small that an electron with energy less than V_0 can tunnel the potential barrier. The solution of the time-independent Schrödinger wave equation involving this potential is a simple exercise in quantum mechanics. It is based on the following points:

(i) The wavefunctions in regions I and II match at the inteface and so do their derivatives.

(ii) As V_0 increases, 'b' decreases so as to keep the product bV_0 at a constant value. Even in the limiting case of large V_0, the product remains a finite quantity.

The attendant calculations are, however, tedious and involve complex algebra that is of secondary interest to us. We directly quote the energy–wavevector relation as:

$$P\left(\frac{\sin \alpha a}{\alpha a}\right) + \cos \alpha a = \cos ka \tag{7.25}$$

where α is related to energy by

$$\alpha = \left(\frac{2m\varepsilon_k}{\hbar^2}\right)^{1/2} \tag{7.26}$$

and $\dfrac{(bV_0)\, ma}{\hbar^2}$ has been replaced by P as we pass to the limit, $b = 0$ and $V_0 = \infty$.

The LHS of (7.25) is plotted as a function of α in Fig. 7.3 for $P = 2$. The absolute magnitude of the ordinate crosses unity when α equals any multiple of $\pm\pi$/a. This is found to be consistent and independent of the value of P. Since the ordinate is equal to cos ka, its magnitude cannot cross the

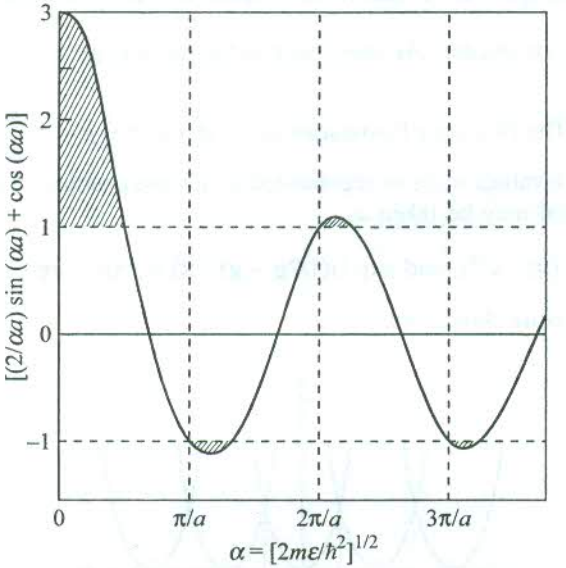


FIG. 7.3 A plot of the left-hand side of (7.25) for $P = 2$ as a function of α which depends on the electron energy in the Kronig–Penney model. The hatched regions correspond to the ranges of forbidden energies.

limits ± 1. The solution to (7.25) yields alternating ranges of real and imaginery k-values. The ranges of imaginary k correspond to the ranges of α in which the absolute magnitude of the ordinate exceeds the permitted limit of unity. Thus the energies that correspond to these α values are forbidden. This results in the alternation of regions of allowed and forbidden energies. The regions of allowed energies refer to the energy bands whereas those of forbidden energies form the band gaps. In the extreme limits $P \to 0$ and $P \to \infty$, we approach the two extreme cases of dispersion relations discussed in the beginning of this section. The elegance with which the Kronig–Penney model predicts the occurrence of band gaps enhances the significance of the model well beyond its historic value.

7.4 THE NEARLY FREE ELECTRON MODEL

The energy bands of solids have generally complex structures. Therefore, it would be appropriate to study the influence of a crystal potential on the electron energy in a certain limiting case and then extend the ideas to any general situation. For example, we may assume that the crystal potential grows slowly from zero value (corresponding to an empty lattice). In the limit of a vanishingly small potential, the Fourier coefficients V_g may be equated to zero in the first approximation. But it would be wrong to ignore totally the effect of periodicity in view of its vital consequences already described. The description of free electron states as given by the energy parabola (Fig. 7.4) is bound to alter in view of (7.21). The allowed energy states are no more confined to a single parabola in the k-space. The states are represented by other parabolae as well, that are displaced by any reciprocal lattice vector **g**:

$$\varepsilon_k = \varepsilon_{k+g} = \frac{\hbar^2}{2m} |k + g|^2 \tag{7.27}$$

On the ground of simplicity, the parabolae for a one-dimensional crystal are drawn in Fig. 7.4. The periodicity in the real space is 'a' and in the k-space it is given by the reciprocal lattice vector $\mathbf{g}\left(= n \cdot \frac{2\pi}{\mathbf{a}}\right)$; n being an integer. At zone boundaries the energy values are degenerate as two parabolae intersect here. The first zone boundaries occur at $k = \pm\frac{1}{2} g = \pm \frac{\pi}{a}$. Therefore, the electron wavefunction with these k-values must be represented by a superposition of at least two plane waves which for a small potential may be taken as

$$\exp(i\mathbf{g} \cdot \mathbf{x}/2); \text{ and } \exp[i(1/2\mathbf{g} - \mathbf{g}) \cdot \mathbf{x}] = \exp(-i\mathbf{g} \cdot \mathbf{x}/2) \tag{7.28}$$

The waves move in opposite directions.

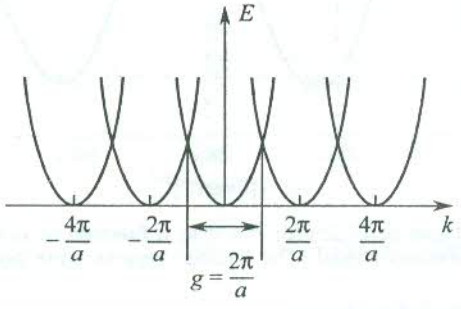


FIG. 7.4 Periodic occurrence of the parabolic energy curves of a free electron in one-dimensional reciprocal space. The periodicity in the real space is 'a'. The electron is supposed to be placed in a periodic lattice with a vanishing potential.

But equation (7.6) requires that reciprocal lattice vectors larger than $2\pi/a$ should also be considered. The value of $C_{\mathbf{k}}$ as determined from (7.6) is appreciably large when $\varepsilon_{\mathbf{k}}$ and $\varepsilon_{\mathbf{k}-\mathbf{g}}$ both approach the value $\left(\frac{\hbar^2 k^2}{2m}\right)$. In this the absolute magnitude of $C_{\mathbf{k}-\mathbf{g}}$ is approximately equal to that

of $C_{\mathbf{k}}$. The two plane waves (7.28) at the first zone boundaries ideally correspond to this condition. Hence, other reciprocal vectors can be ignored in the approximation for the construction of wavefunctions at the zone boundary. The wavefunctions may be expressed as

$$\Psi(+) \sim e^{i\mathbf{g}\cdot\mathbf{x}/2} + e^{-i\mathbf{g}\cdot\mathbf{x}/2} \sim \cos\frac{\pi x}{a} \tag{7.29}$$

$$\Psi(-) \sim e^{i\mathbf{g}\cdot\mathbf{x}/2} - e^{-i\mathbf{g}\cdot\mathbf{x}/2} \sim \sin\frac{\pi x}{a} \tag{7.30}$$

These standing waves appear as a result of the Bragg reflection occurring at $k = \pm g/2 = \pm \pi/a$, as described in Chapter 3. The electron plane waves when Bragg reflected superpose the waves moving towards the same zone edge where the former suffered Bragg reflection. The probability densities of the two sets of standing waves are

$$\Psi^*(+)\,\Psi(+) = |\Psi(+)|^2 \sim \cos^2\frac{\pi x}{a} \tag{7.31}$$

$$\Psi^*(-)\,\Psi(-) = |\Psi(-)|^2 \sim \sin^2\frac{\pi x}{a} \tag{7.32}$$

The electron potential energy in a one-dimensional crystal is drawn in Fig. 7.5. The potential field belongs to the positive ion cores whose valence electrons move in this field. The figure also

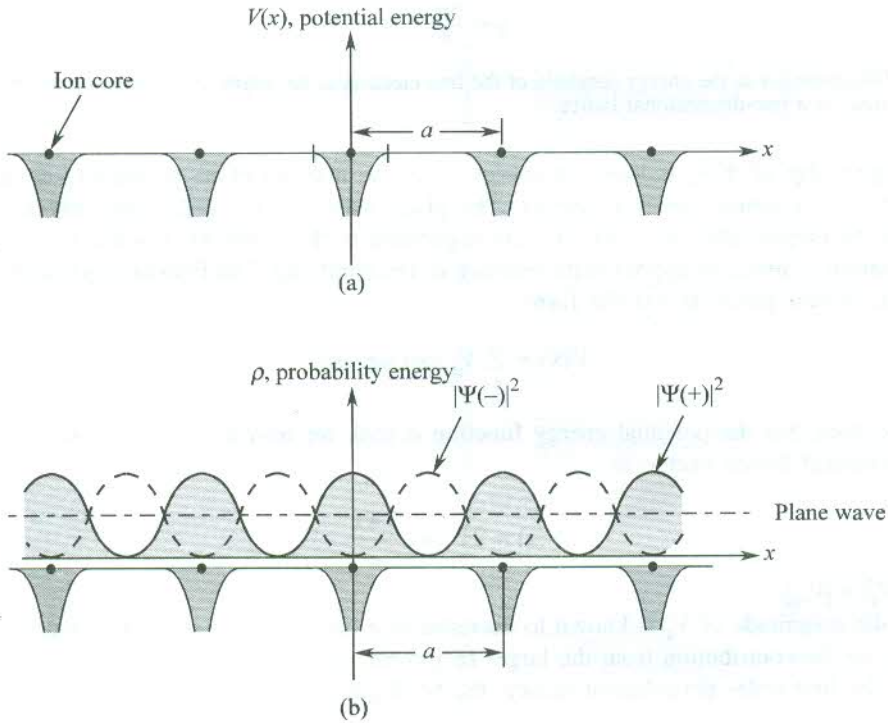


FIG. 7.5 (a) Qualitative form of the electron potential energy $V(x)$ in a one-dimensional crystal lattice. Dark circles denote the positions of the positive ion cores. (b) Distribution of probability density for the standing waves $\Psi(\pm)$ and the plane wave inside the crystal lattice.

depicts the probability distributions of the standing waves and the simple plane waves. The plane waves exp ($i\mathbf{k} \cdot \mathbf{x}$) have the same probability density at all points, since exp ($-i\mathbf{k} \cdot \mathbf{x}$) · exp ($i\mathbf{k} \cdot \mathbf{x}$) = 1. The distribution for $\Psi(+)$ favours the piling of electronic charge exactly above the ion cores. On the other hand, the $\Psi(-)$ waves push the electronic charge away from the ion cores. The eigenvalues of $\Psi(+)$ and $\Psi(-)$ differ, though both correspond to the same k-value (π/a or $-\pi/a$). The energy dispersion curve which is continuous throughout the zone shows a gap with two unequal roots at the zone boundaries (see Fig. 7.6). This explains the origin of the band gaps observed in the energy band structure.

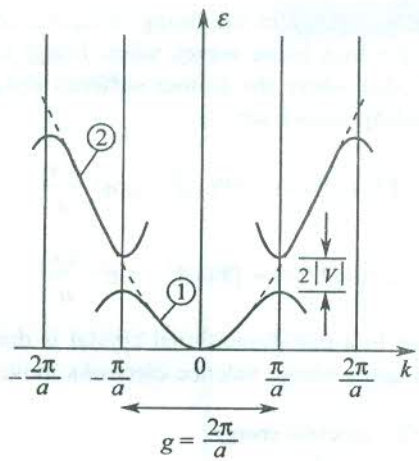


FIG. 7.6 Discontinuity in the energy parabola of the free electron at the edges ($k = \pm \pi/a$) of the first Brillouin zone in a one-dimensional lattice.

The eigenvalue of $\Psi(+)$ is lower in energy since the maxima of its probability density occur at the points of minimum potential energy. The plane wave energy at the zone edges is centred between the two eigenvalues $\varepsilon(+)$ and $\varepsilon(-)$ corresponding to $\Psi(+)$ and $\Psi(-)$ in that order. Now, we are in a position to make an approximate measure of the band gap. The Fourier expansion of a one-dimensional crystal potential has the form

$$V(\mathbf{x}) = \sum_{\mathbf{g}} V_{\mathbf{g}} \exp(i\mathbf{g} \cdot \mathbf{x}) \tag{7.33}$$

Appreciating that the potential energy function is real, we rewrite (7.33) considering only the shortest reciprocal lattice vector as

$$V(x) = 2V \cos \frac{2\pi x}{a} \tag{7.34}$$

with $V = |V_g| = |V_{-g}|$.

Since the magnitude of V_g is known to decrease as g increases, for an approximate calculation we can ignore the contribution from the larger reciprocal vectors.

Using the first order perturbation theory, the band gap is written as

$$\Delta E = \varepsilon(+) - \varepsilon(-) = 2V \int \cos \frac{2\pi x}{a} \left[\Psi^*(+)\, \Psi(+) - \Psi^*(-)\, \Psi(-)\right] dx \tag{7.35}$$

Normalizing the wavefunctions at $k = \pm \pi/a$ over the crystal's length L, we have

$$\Psi(+) \sim \left(\frac{2}{L}\right)^{1/2} \cos \frac{\pi x}{a}$$

$$\Psi(-) \sim \left(\frac{2}{L}\right)^{1/2} \sin \frac{\pi x}{a} \tag{7.36}$$

Using (7.36), we have from (7.35)

$$\Delta E = \frac{2V}{L} \int_0^L \left(1 + \cos \frac{4\pi}{a} x\right) dx$$

or

$$\Delta E = 2V \tag{7.37}$$

Thus the band gap is equal to twice the magnitude of the leading Fourier coefficient of the crystal potential. The range of allowed energy states covered by the dispersion curve (see Fig. 7.6) in the first Brillouin zone constitutes the first energy band. Similarly, the higher bands are identified with the respective dispersion curves in other zones.

7.5 ZONE SCHEMES FOR ENERGY BANDS

There are three zone schemes in which the energy bands are drawn:

(i) The extended zone scheme

(ii) The reduced zone scheme

(iii) The periodic zone scheme

The extended zone scheme

In this scheme different energy bands are drawn in different zones in the k-space. Figure 7.6 gives the plot of dispersion curves in the extended zone scheme. The first band is shown in the first zone ($-\pi/a \leq k \leq \pi/a$) and the next higher in the second zone ($\pi/a \leq k \leq 2\pi/a$ and $-2\pi/a \leq k \leq -\pi/a$), and so on.

The reduced zone scheme

All energy bands are shown in the first Brillouin zone in this scheme. As an example, the free electron parabola (see Fig. 6.4) is shown in this scheme by Fig. 7.7. The curves in the two segments of the second zone are translated to the first zone by reciprocal vectors $2\pi/a$ and $-2\pi/a$, separately. Similarly, the energy band pictures in other zones are translated to the first zone by appropriate reciprocal lattice vectors.

The periodic zone scheme

Every band is drawn in every zone in this scheme.

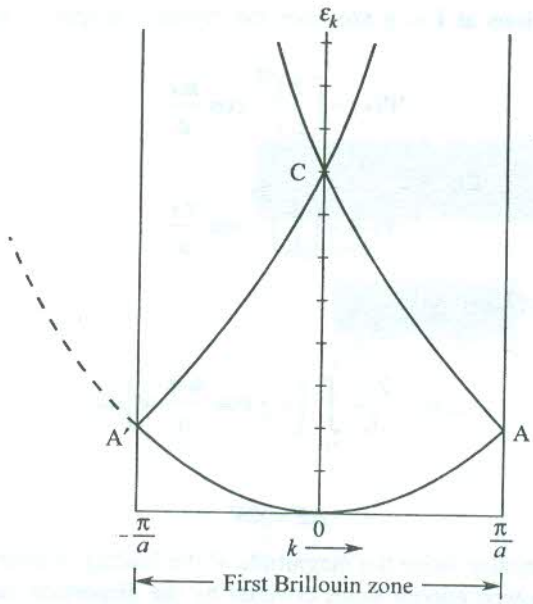


FIG. 7.7 The free electron energy parabola plotted in the reduced zone scheme for a one-dimensional lattice. The curve AC, when translated by $-2\pi/a$, reproduces the usual free electron curve for the negative k-values represented by the dashed curve. Similarly, the translation of A′C by $2\pi/a$ will give the curve for the positive k-values. This often gives a useful description of the band structure of a crystal.

The first three energy bands of a linear crystal as drawn in the three schemes are shown in Fig. 7.8 for the purpose of comparison.

7.6 ENERGY BANDS IN A GENERAL PERIODIC POTENTIAL

Having solved the one-electron Schrödinger wave equation (7.3) in the vanishing limit of a periodic crystal potential (NFE model), we have gained insight into the problem. The wave equation (7.3) in principle can be replaced by (7.6) without any loss of generality. Hence, the process of treating a general periodic potential proceeds in continuation with (7.6), which we rewrite as

$$(\lambda_{\mathbf{k}} - \varepsilon) C_{\mathbf{k}} + \sum_{\mathbf{g}} V_{\mathbf{g}} C_{\mathbf{k}-\mathbf{g}} = 0 \tag{7.38}$$

with

$$\lambda_{\mathbf{k}} = \frac{\hbar^2 k^2}{2m} \tag{7.39}$$

Equation (7.38) is called the *central equation*. It represents a set of simultaneous linear equations that couple the coefficients $C_{\mathbf{k}-\mathbf{g}}$ for all reciprocal lattice vectors. The number of these equations equals the number of reciprocal lattice vectors. The equations are consistent if the determinant of the coefficients of Cs vanishes. Considering only one Fourier component $V_{\mathbf{g}}$, as in the NFE model, we write three consecutive equations which yield the following determinant.

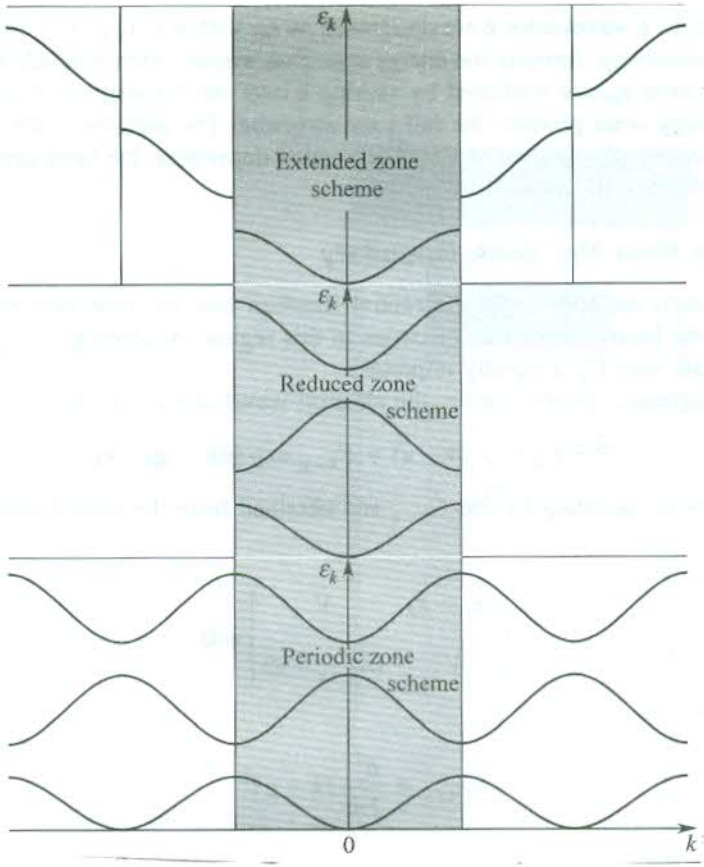


FIG. 7.8 First three energy bands of a one-dimensional crystal in the extended, reduced and periodic zone schemes.

$$\begin{vmatrix} (\lambda_{\mathbf{k}-\mathbf{g}} - \varepsilon) & V & 0 \\ V & (\lambda_{\mathbf{k}} - \varepsilon) & V \\ 0 & V & (\lambda_{\mathbf{k}+\mathbf{g}} - \varepsilon) \end{vmatrix} \tag{7.40}$$

with

$$|V_{\mathbf{g}}| = |V_{-\mathbf{g}}| = V.$$

It is a small portion of a huge determinant that is evolved from the set (7.38). By equating (7.40) to zero and solving for ε we get three roots that fall in three different energy bands at a certain value of k. The size of the determinant is chosen according to the extent of the energy spectrum that is required. But the size factor is no more a consideration now, with the availability of fast computers. On choosing k that differs from a value in the first Brillouin zone by a reciprocal lattice vector, there occurs no change in the energy spectrum as the same set of equations in a different order appear, giving the same roots of energy. Therefore, very often the k-values within the first zone alone are considered.

Roots that refer to a wavevector k are designated as ε_{nk} with $n = 1, 2, 3, \ldots$ for the first, second, third, ..., bands, respectively, forming the energy spectrum whose different levels belong to different energy bands. The roots ε_{nk} are evaluated by varying k over the allowed set of values. These when arranged on an energy scale produce the full band structure. The pictures in the three-dimensional space are fairly complex. On account of k being direction dependent, the band structures, in general, look different in different directions.

7.6.1 Solution Near the Zone Boundary

It is important to study the solution of the central equation near the zone boundary because of the large deviation in the behaviour of free electrons in this region. According to (7.38), if C_{k-g} is an important coefficient, then C_k is equally important.

In the two-component approximation, the electron wavefunction is taken as

$$\Psi = C_{\mathbf{k}} \exp(i\mathbf{k} \cdot \mathbf{x}) + C_{\mathbf{k}-\mathbf{g}} \exp[i(\mathbf{k} - \mathbf{g}) \cdot \mathbf{x}] \tag{7.41}$$

The two equations, coupling C_k and C_{k-g} and obtained from the central equation (7.38), have a solution if

$$\begin{vmatrix} (\lambda_k - \varepsilon) & V \\ V & (\lambda_{k-g} - \varepsilon) \end{vmatrix} = 0$$

with

$$\lambda_{k-g} = \frac{\hbar^2}{2m} |k - g|^2$$

Solving it for ε, we obtain

$$\varepsilon_k(\pm) = \frac{1}{2}(\lambda_{\mathbf{k}-\mathbf{g}} + \lambda_{\mathbf{k}}) \pm \left[\frac{1}{4}(\lambda_{\mathbf{k}-\mathbf{g}} - \lambda_{\mathbf{k}})^2 + V^2\right]^{1/2} \tag{7.42}$$

The roots $\varepsilon_k(\pm)$ when plotted as a function of wavevector give dispersion curves for the first two energy bands as shown in Fig. 7.9.

The corresponding roots at the zone boundary $\left(k = \frac{1}{2}g\right)$ may be written as

$$\varepsilon(\pm) = \lambda \pm V \tag{7.43}$$

where

$$\lambda = \frac{\hbar^2}{2m}\left(\frac{1}{2}g\right)^2 \tag{7.44}$$

The λ is the free electron energy at the zone boundary. We find that the band gap obtained from (7.43) is identical with that given by (7.37).

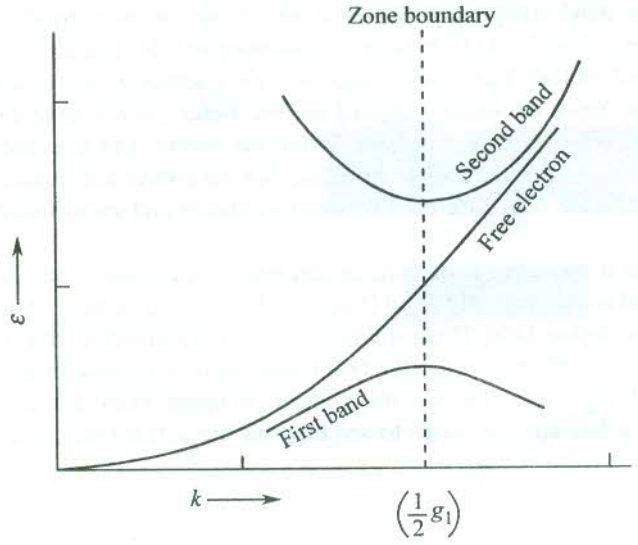


FIG. 7.9 Qualitative form of solutions of (7.42) in the periodic zone scheme near the boundary of the first Brillouin zone. The free electron curve is drawn for comparison.

It is instructive, however, to express the roots near the zone boundary in terms of the wavevector $\tilde{k}$ as measured from the zone boundary. The k and $\tilde{k}$ are related by

$$\tilde{k} = k - \frac{1}{2}g \tag{7.45}$$

Using (7.45), (7.42) can be expressed as

$$\varepsilon_{\tilde{k}}(\pm) = \varepsilon(\pm) + \frac{\hbar^2\tilde{k}^2}{2m}\left(1 \pm \frac{2\lambda}{V}\right) \tag{7.46}$$

Thus in this definition of the wavevector, the dependence of energy on wavevector is similar to that for the free electrons. When V is negative, $\varepsilon_{\tilde{k}}(-)$ corresponds to the upper band.

7.7 INSULATORS, SEMICONDUCTORS AND METALS

The electronic transport in solids is found to be closely controlled by their band structure. With the knowledge of the band structure, it is possible to predict whether a solid is a good or bad conductor of electricity. We discuss below how insulators, semiconductors and metals may be distinguished on the basis of qualitative differences in their band structures.

We know that the number of energy states in a band is equal to the number of primitive cells (say N) in the crystal. Since each state can accommodate two electrons of opposite spins, a number of ($2N$) electrons would be required to fill the band completely. Therefore, if a primitive cell contributed an even number of electrons to a band, the band would be completely filled. For example, in a monatomic crystal in which each atom has one valence electron, the band will be fully occupied if there are two atoms in the primitive cell. With two valence electrons per atom, only one atom

in the cell will fill the band completely. Solids in which one or more bands are completely filled and all others are empty [Fig. 7.10(a)], behave as insulators at 0 K. The lowest allowed empty band is preceded by a region of forbidden energy gap (> 5 eV), across which electrons need be excited for electric conduction. Since the usually applied electric fields are not large enough to provide this excitation, the material behaves as an insulator. When the energy gap is in the range of an eV, the material shows a conductivity intermediate to values for insulators and metals and the material is classified as a semiconductor. The difference between insulators and semiconductors is one of degree and not of type.

On the other hand if the number of valence electrons per primitive cell is an odd number, the top-most occupied band is only partially filled [Fig. 7.10(b)]. Such a material shows the flow of current on the application of an electric field. These solids are thus good conductors of electricity and are called metals. Alkali metals and noble metals are the best examples of this class where one valence electron is contributed by each primitive cell and the highest occupied band is exactly half filled.

Now, we take up a few specific cases to see how far the actual band structures conform to the above principle.

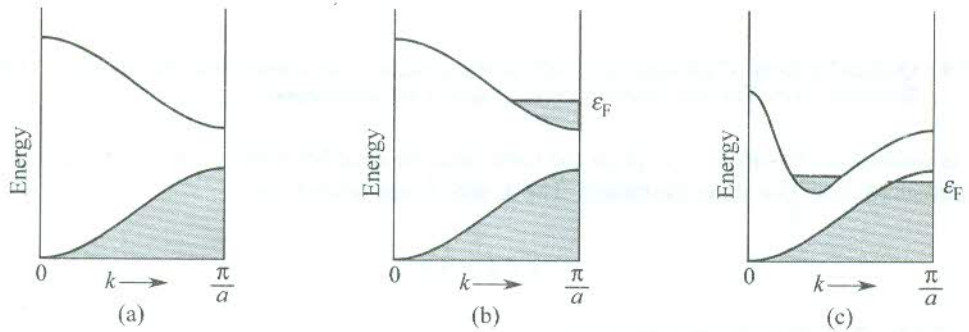


FIG. 7.10 Qualitative band schemes for insulators and metals: (a) insulators—the lower band and all below it are completely filled; all higher bands are empty, (b) metals—a partially occupied band, and (c) metals of relatively low conductivity (semi-metals)—overlap between, a filled band and an empty band.

Sodium. It belongs to the group of alkali metals all having the BCC structure with a rhombohedral primitive cell that contains effectively one atom. The electronic configuration of a sodium atom is $[1s^2\ 2s^2\ 2p^6]\ 3s^1$. Since there is a correspondence between the discrete states of an isolated atom and the Brillouin zones in a solid, we expect that the 10 electrons in the closed inner shells form narrow bands in the sodium metal occupying the first five zones (1 for each s-shell and 3 for the p-shell) in the extended zone scheme. The single outermost ($3s^1$) electron produces the half-filled band in the next zone. Accordingly, the solid sodium is metallic and so are other alkali metals.

Magnesium. It is a member of the alkaline earth metals all of which have two valence electrons per primitive cell, irrespective of the symmetry. The magnesium atom has electronic configuration $[1s^2\ 2s^2\ 2p^6]\ 3s^2$ which would apparently give the isulating behaviour to the solid magnesium, contrary to the observed metallic character. The metallic character arises because of the overlapping of the empty 3p-band with the filled 3s-band. Thus the 3s-electrons can be almost continuously excited to states in the 3p-band. The overlap only saves the alkaline earth metals from being branded insulator. They are not the same good conductors of electricity as the alkali metals. They are rather

classified as semi-metals. A qualitative demonstration of the band overlap is shown in Fig. 7.10(c). The Fermi energy ε_F in magnesium occurs at an energy which fills the 3s band about 90 per cent, with just a small per cent occupancy for the overlapping sp band.

Diamond. The isolated carbon atoms have the electronic configuration $-1s^2\ 2s^2\ 2p^2$. The mixing of 2s and 2p wavefunctions in the tetrahedrally bonded diamond crystal is well known to result in the sp^3 hybridization. The sp^3 hybrid band further splits into two because of the modification of the s- and p-level in the crystal. Each of the two hybrid sub-bands can accommodate four electrons. The four outer electrons belonging to the 2s- and 2p-states fill the lower sub-band, leaving the upper one empty. There exists a forbidden gap E_g of about 5 eV at 0 K between the two sub-bands. These features that account for the insulating property of diamond are shown in Fig. 7.11. Band structures of the semiconductors Ge and Si are characterized by similar features. A proper discussion on semiconductors is deferred to Chapter 9.

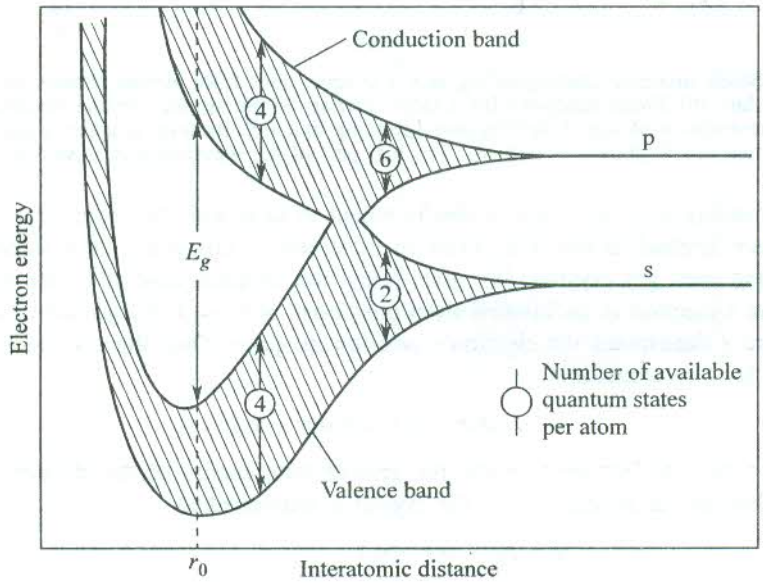


FIG. 7.11 Behaviour of energy bands as a function of interatomic separation for diamond (C), Si and Ge. The band gap E_g is defined at the equilibrium separation r_0. The figure shows that the band gap is not tied to the periodicity of the lattice. Amorphous solids can also show a band gap. [After W. Shockley, *Electrons and Holes in Semiconductors* (van Nostrand, 1950).]

7.8 THE TIGHT-BINDING APPROXIMATION

In the NFE model the plane wave part, $\exp(i\mathbf{k} \cdot \mathbf{r})$, of the Bloch function, $[u_{\mathbf{k}}(\mathbf{r}) \exp(i\mathbf{k} \cdot \mathbf{r})]$, is emphasized and the atomic part is overlooked. But electrons in the low-lying inner core levels of a free atom are strongly localized in space. The property is largely retained by these electrons when atoms are assembled to form the solid. It points to the inadequacy in describing every band structure in terms of quasi-free electrons. In order to deal with the localized electron, an alternative approach is followed in which the atomic part of the Bloch function is stressed. This approach is known as the *tight-binding approximation*. The single electron wavefunction in the crystal is expressed as a

linear combination of the atomic orbitals (LCAO) that the electron occupies in a free atom. The forms of Bloch functions in a linear crystal for $k = 0$ and $k \neq 0$ are drawn in Fig. 7.12.

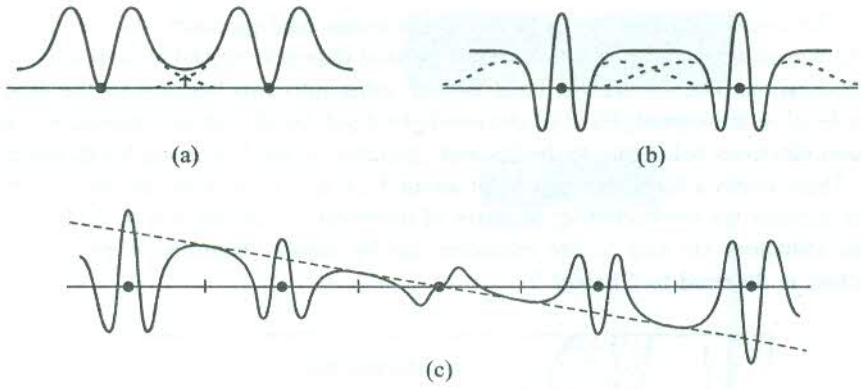


FIG. 7.12 (a) Bloch functions corresponding to $k = 0$ state, built from atomic orbitals (dashed) with small overlap. (b) Bloch functions for a large overlap of the atomic orbitals (dashed). (c) Schematic representation of wavefunctions corresponding to $k \neq 0$. Its form is approximately the product of (*b*) and cos k (shown as dashed). The real part of the wavefunction is shown.

The tight-binding approximation is ideally suited to deal with the inner core electrons. It has successfully been applied to the d-electrons in transition metals and to the valence electrons in diamond like and inert gas crystals. We give below the simplest case of s-state electrons.

Consider an s-electron in an isolated atom positioned at $\mathbf{r}_n$ with the ground state wavefunction $\phi(\mathbf{r} - \mathbf{r}_n)$, where $\mathbf{r}$ determines the electron's position in space. Then the one-electron Schrödinger wave equation for free atoms is

$$H_0\phi(\mathbf{r} - \mathbf{r}_n) = \varepsilon_0\phi(\mathbf{r} - \mathbf{r}_n) \tag{7.47}$$

where H_0 and ε_0 are the Hamiltonian and the ground state energy of the electron in a free atom.

The Hamiltonian for an electron in the crystal is expressed as

$$\begin{aligned} H &= H_0 + v(\mathbf{r} - \mathbf{r}_n) \\ &= \left[-\frac{\hbar^2}{2m}\nabla^2 + v_0(\mathbf{r} - \mathbf{r}_n) + v(\mathbf{r} - \mathbf{r}_n) \right] \end{aligned} \tag{7.48}$$

with

$$v(\mathbf{r} - \mathbf{r}_n) = \sum_{m \neq n} v_0(\mathbf{r} - \mathbf{r}_m) \tag{7.49}$$

In the above equations, $v_0(\mathbf{r} - \mathbf{r}_n)$ denotes the potential energy of an electron when localized at the isolated atom positioned at $\mathbf{r}_n$. The influence of atoms in the vicinity of $\mathbf{r}_n$, where the electron in question is strongly localized on relative terms, is treated as a perturbation on H_0 and represented by $v(\mathbf{r} - \mathbf{r}_n)$. Now, our aim is to look for solutions of the following Schrödinger equation:

$$H\Psi_{\mathbf{k}}(\mathbf{r}) = \varepsilon_{\mathbf{k}}\Psi_{\mathbf{k}}(\mathbf{r}) \tag{7.50}$$

where $\varepsilon_{\mathbf{k}}$ is the electron energy in the crystal and $\Psi_{\mathbf{k}}(\mathbf{r})$ the Bloch wavefunction.

Let us try the solution,

$$\Psi_{\mathbf{k}}(\mathbf{r}) = \sum_{n} \exp(i\mathbf{k} \cdot \mathbf{r}_n)\, \phi(\mathbf{r} - \mathbf{r}_n) \tag{7.51}$$

The above function satisfies the properties of a Bloch function. For example, it is quite simple to check that the function is periodic in the k-space.

According to the perturbation theory, the first order energy is given by

$$\varepsilon_k = \frac{\int \Psi_{\mathbf{k}}^*(\mathbf{r})\, H\, \Psi_{\mathbf{k}}(\mathbf{r})\, dV}{\int \Psi_{\mathbf{k}}^*(\mathbf{r})\, \Psi_{\mathbf{k}}(\mathbf{r})\, dV} \tag{7.52}$$

Using (7.51), we get

$$\int \Psi_{\mathbf{k}}^*(\mathbf{r})\, \Psi_{\mathbf{k}}(\mathbf{r})\, dV = \sum_{n,m} \exp[i\mathbf{k} \cdot (\mathbf{r}_n - \mathbf{r}_m)] \int \phi^*(\mathbf{r} - \mathbf{r}_m)\, \phi(\mathbf{r} - \mathbf{r}_n)\, dV \tag{7.53}$$

For a strongly localized electron, $\phi(r - r_m)$ is significant only in the proximity of r_m. Therefore, we evaluate (7.53) by putting $m = n$ in the first approximation. If there are N atoms in the crystal, we have

$$\int \Psi_{\mathbf{k}}^*(\mathbf{r})\, \Psi_{\mathbf{k}}(\mathbf{r})\, dV = \sum_{n} 1 = N \tag{7.54}$$

Making use of (7.47), (7.48) and (7.54), electron energy ε_k is written as

$$\varepsilon_{\mathbf{k}} \simeq \frac{1}{N} \sum_{n,m} \exp[i\mathbf{k} \cdot (\mathbf{r}_n - \mathbf{r}_m)] \int \phi^*(\mathbf{r} - \mathbf{r}_m)\, [\varepsilon_0 + v(\mathbf{r} - \mathbf{r}_n)]\, \phi(\mathbf{r} - \mathbf{r}_n)\, dV \tag{7.55}$$

For the term containing ε_0, we again neglect the overlap between the nearest neighbours, putting $m = n$. Therefore,

$$\frac{\varepsilon_0}{N} \sum_{n,m} \exp[i\mathbf{k} \cdot (\mathbf{r}_n - \mathbf{r}_m)] \int \phi^*(\mathbf{r} - \mathbf{r}_m)\, \phi(\mathbf{r} - \mathbf{r}_n)\, dV = \varepsilon_0 \tag{7.56}$$

Including the overlap up to the nearest neighbours for the perturbation term, we write

$$\int \phi^*(\mathbf{r} - \mathbf{r}_n)\, v(\mathbf{r} - \mathbf{r}_n)\, \phi(\mathbf{r} - \mathbf{r}_n)\, dV = -\alpha \tag{7.57}$$

(on the same atom)

$$\int \phi^*(\mathbf{r} - \mathbf{r}_m)\, v(\mathbf{r} - \mathbf{r}_n)\, \phi(\mathbf{r} - \mathbf{r}_n)\, dV = -\gamma \tag{7.58}$$

(between the nearest neighbours)

All terms in the summation over n, each of which is evaluated over all m (either the same

atom or the nearest neighbours), are equal in magnitude on the demand of periodicity. Since the summation over n runs over all atoms in the crystal, the sum is simply N times the value of a single term. This factor of N cancels with the factor of N in the denominator. In view of this and relations (7.56), (7.57), (7.58), the electron energy ε_k assumes the form

$$\varepsilon_k \sim \varepsilon_0 - \alpha - \gamma \sum_m e^{i\mathbf{k}\cdot(\mathbf{r}_n - \mathbf{r}_m)} \tag{7.59}$$

The sum in (7.59) is carried only over the nearest neighbours. In a simple cubic crystal with lattice constant 'a', the nearest-neighbour atoms are at

$$\mathbf{r}_n - \mathbf{r}_m = (\pm a, 0, 0); \qquad (0, \pm a, 0); \qquad (0, 0, \pm a)$$

This gives the s-state energy in the crystal as

$$\varepsilon_k \simeq \varepsilon_0 - \alpha - 2\gamma(\cos k_x a + \cos k_y a + \cos k_z a) \tag{7.60}$$

When the atoms are brought together to form a crystal, the single atomic energy level ε_0 broadens to constitute an energy band whose component levels are defined by (7.60). We can determine the bandwidth as follows.

Energy of the band at its bottom:

The bottom lies at $k = 0$, giving

$$\varepsilon_{\text{bott}} \simeq \varepsilon_0 - \alpha - 6\gamma \tag{7.61}$$

Energy of the band at its top:

The top occurs at $k = \pm \pi/a$, giving

$$\varepsilon_{\text{top}} \simeq \varepsilon_0 - \alpha + 6\gamma \tag{7.62}$$

Therefore, from (7.61) and (7.62),

$$\text{the bandwidth} = 12\gamma \tag{7.63}$$

Thus the bandwidth is proportional to γ which represents the overlap of atomic orbitals (7.58). A qualitative illustration of the results of tight-binding calculation for a simple cubic crystal is made in Fig. 7.13. Whereas γ determines the bandwidth, α is interpreted as the lowering of the centre of gravity of the free atomic level on forming the solid. As one proceeds from inner to outer shells in an atom, the width of the respective energy bands goes on increasing because of more and more overlap. This is consistent with the result of the Kronig–Penney model and also confirmed by experiments. It is amply clear from even Fig. 7.13 where the second band appears as much wider.

Furthermore, the effect of tight binding on the energy surfaces of a simple cubic crystal can be examined with the help of (7.60). It is instructive to do this for the limiting k-values in the reduced zone scheme, i.e. near the centre and the boundary of the first zone. Near the centre $ka \leq 1$ and we expand the cosine function to obtain

$$\varepsilon_k \simeq \varepsilon_0 - \alpha - 6\gamma + \gamma k^2 a^2 \tag{7.64}$$

These values refer to the bottom region of the band and conform to the constant spherical energy surfaces. But as the wavevector increases, the shape gets distorted and deviates sufficiently from the spherical nature at large values of k [see Fig. 7.14(a)].

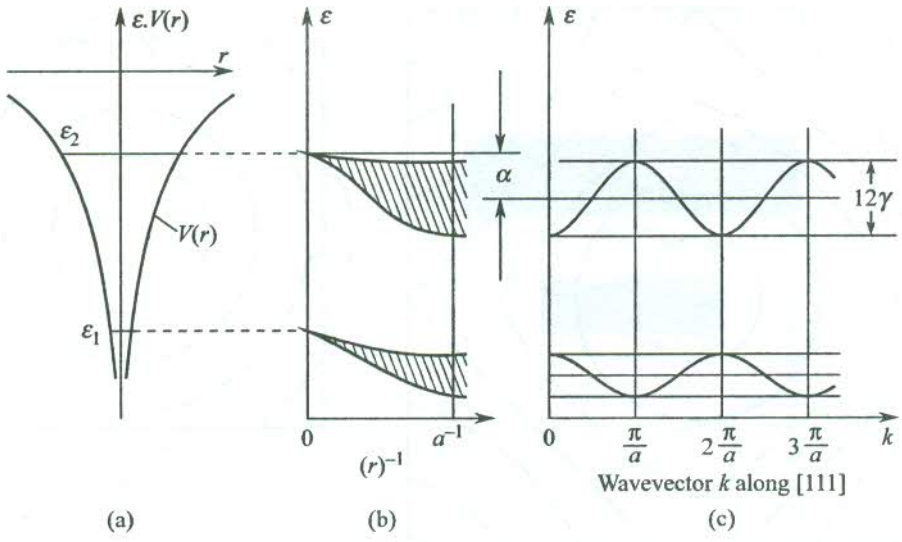


FIG. 7.13 Illustration of the result of a tight-binding approximation for a primitive cubic lattice of lattice constant 'a'. (a) Position of energy levels ε_1 and ε_2 in the potential $V(r)$ of the free atom. (b) Variation in broadening of the levels ε_1 and ε_2 as a function of the reciprocal atomic separation r^{-1}. (c) Variation in one-electron energy ε as a function of the wavevector $k(1, 1, 1)$ in the direction of the body diagonal [111].

In order to investigate the region near the zone boundary, we express k in terms of its value k' as measured from the zone boundary:

$$k = \frac{\pi}{a} - k' \tag{7.65}$$

Substituting (7.65) in (7.60) and appreciating that $k'a \leq 1$ near the zone boundary, we obtain

$$\varepsilon_k \approx \varepsilon_0 - \alpha + 6\gamma - \gamma k'^2 a^2 \tag{7.66}$$

The above result is similar to that of (7.64) with the difference that the spherical surfaces are centred at the corners of the zone [Fig. 7.14(a)]. We may compare the energy surfaces in the TBA with those in the NFE model, shown in Fig. 7.14(b). The main difference lies in the fact that the spherical shape is maintained to much larger values of k in the NFE model. This only shows that the results in the two models are almost similar for small wavevectors. In view of this fact, the k-dependent part of the energy dispersion (7.64) near the zone's centre is comparable with $\hbar^2k^2/2m$ (for the free electrons), and we have

$$\gamma k^2 a^2 = \frac{\hbar^2 k^2}{2m^*} \tag{7.67}$$

giving

$$m^* = \frac{\hbar^2}{2\gamma a^2} \tag{7.68}$$

where m^* is identified as the effective mass of the electron, implying that the electron mass should be treated as variable. Relation (7.68) predicts that m^* is larger for electrons in the inner shells whose

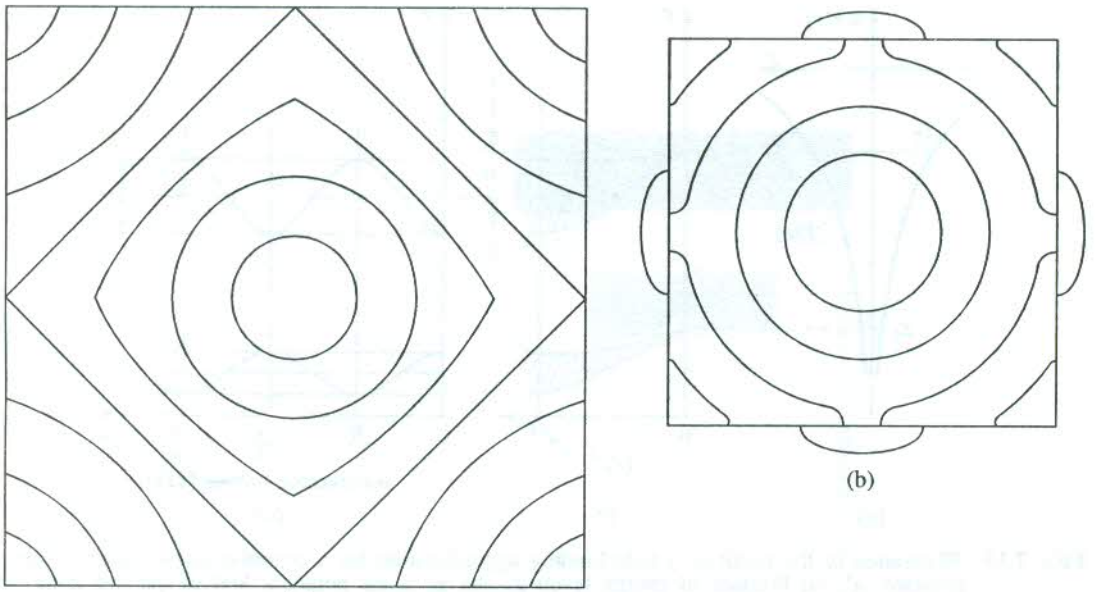


FIG. 7.14 (a) Constant energy curves for the tight-binding approximation for a simple cube crystal in the plane $k_z = 0$. Energy surfaces are spherical around the zone centre for only small k-values. The surfaces are again spherical with centres at the corners of the zone for small k-values, measured from the corners. (b) Constant energy curves in the NFE model. Energy surfaces are spherical up to fairly large values of k. Notice the changes caused by the zone boundary.

overlap is far less. We consider this as an invaluable result of the theory of tight-binding approximation, especially because the concept of effective mass has remarkably improved the understanding of several physical properties ranging from electronic conduction to the complex optical phenomena.

7.9 THE WIGNER–SEITZ CELLULAR METHOD

The significance of the Wigner–Seitz model is substantiated by the impressive success it achieved in accounting for the band structure and the cohesive energy of alkali metals. The first systematic calculation of energy bands appeared in the form of this model. The technique of calculation, referred to as a *cellular method*, is based mainly on the symmetry properties of a certain primitive cell designed by Wigner and Seitz themselves. The method of construction of this cell has already been described in Section 1.2. The alkali metals have the BCC structure for which the Wigner–Seitz cell is a polyhedron.

The entire crystal volume is imagined to have been filled up with identical polyhedra, assuming that there is only one electron in a given polyhedron at a time, together with the positively charged metal ion at the centre. The polyhedra being neutral, the interactions among themselves are initially neglected and only the interaction within each polyhedron is considered. The electron is supposed to move in the spherically symmetric potential field of the ion. The field is assumed not to extend past the boundaries of the polyhedron.

We consider the extreme case of the $\mathbf{k} = 0$ state. Then the Bloch wavefunction $\Psi_{\mathbf{k}}(\mathbf{r})$ has the form

$$\Psi(\mathbf{r}) = u_0(\mathbf{r}) \tag{7.69}$$

This wavefunction itself is periodic in the crystal, i.e. the wavefunction remains unchanged when translated from one face to the opposite face of the cell. It requires that $\frac{\partial \Psi}{\partial \mathbf{n}} = 0$ at the boundary of the polyhedron, with $\mathbf{n}$ as a direction normal to any of the faces of the polyhedron. The above condition in the crystal replaces the free atom boundary condition, $\Psi(r) \to 0$ as $r \to \infty$, by

$$\left(\frac{\partial \Psi}{\partial r}\right)_{r=r_s} = 0 \tag{7.70}$$

where r_s is the radius of a sphere to which the polyhedron may be approximated. Accordingly, the volume of the polyhedron may be given by $4\pi r_s^3/3$.

The solution of the one-electron Schrödinger wave equation is much easier for $k = 0$ than for any general k-value because $u_0(\mathbf{r})$ is non-degenerate and observes the full symmetry of the crystal. Wigner and Seitz gave an accurate estimate of $u_0(\mathbf{r})$. In view of the boundary condition (7.70), the exercise simply reduces to solving the radial Schrödinger wave equation:

$$\left[\frac{1}{\mathbf{r}^2}\frac{\partial}{\partial \mathbf{r}}\left(\mathbf{r}^2\frac{\partial}{\partial \mathbf{r}}\right) + \frac{2m}{\hbar^2}(\varepsilon_0 - V_0(\mathbf{r}))\right]\Psi(\mathbf{r}) = 0 \tag{7.71}$$

where $V_0(r)$ is the potential energy of an s-electron and ε_0 is the energy eigenvalue in the field of a crystalline ion within one of the polyhedra.

However, for a general wavevector we have to solve,

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_0(\mathbf{r})\right] u_{\mathbf{k}}(\mathbf{r}) \exp(i\mathbf{k}\cdot\mathbf{r}) = \varepsilon_{\mathbf{k}} u_{\mathbf{k}}(\mathbf{r}) \exp(i\mathbf{k}\cdot\mathbf{r}) \tag{7.72}$$

But,

$$\begin{aligned}\nabla^2[u_{\mathbf{k}}(\mathbf{r})\exp(i\mathbf{k}\cdot\mathbf{r})] &= \nabla[i\mathbf{k}\,u_{\mathbf{k}}(\mathbf{r})\exp(i\mathbf{k}\cdot\mathbf{r}) + \exp(i\mathbf{k}\cdot\mathbf{r})\nabla u_{\mathbf{k}}(\mathbf{r})] \\ &= \exp(i\mathbf{k}\cdot\mathbf{r})[\nabla^2 u_{\mathbf{k}}(\mathbf{r}) - \mathbf{k}^2 u_{\mathbf{k}}(\mathbf{r}) + 2i\mathbf{k}\cdot\nabla u_{\mathbf{k}}(\mathbf{r})]\end{aligned} \tag{7.73}$$

Putting (7.73) in (7.72), we get

$$\left[-\frac{\hbar^2}{2m}(\nabla^2 + 2i\mathbf{k}\cdot\nabla) + V_0(\mathbf{r})\right] u_{\mathbf{k}}(\mathbf{r}) = \left(\varepsilon_{\mathbf{k}} - \frac{\hbar^2 k^2}{2m}\right) u_{\mathbf{k}}(\mathbf{r}) \tag{7.74}$$

It must be observed that $u_0(\mathbf{r})$ is not an exact solution to (7.74) which is in fact satisfied by $u_{\mathbf{k}}(\mathbf{r})$, the periodic part of the general Bloch function.

We treat the $\mathbf{k}\cdot\nabla$ terms as a perturbation and insist that $u_{\mathbf{k}}(\mathbf{r})$ obey the boundary condition (7.70). This gives electron energies $\varepsilon_{\mathbf{k}}$ in the form

$$\varepsilon_{\mathbf{k}} = \varepsilon_0 + \frac{\hbar^2 k^2}{2m} \tag{7.75}$$

which gives energies in the shape of a band as measured from the level of ε_0.

We must appreciate that these calculations depend only on the atomic volume and are independent of the crystal structure. Therefore, for a solid metal and a liquid of equal density we expect the same results in this model.

The 3s radial wavefunction in sodium metal as estimated in the Wigner–Seitz model at $k = 0$ (the Brillouin zone centre) and at the Brillouin zone boundary is plotted in Fig. 7.15. The 3s atomic wavefunction is also drawn for comparison. When the wavefunction is subject to the boundary condition (7.70), midway between the neighbouring atoms ($r = r_s$), the solution of the radial Schrödinger equation (7.71) for $k = 0$ yields the eigenvalue (ε_0) as – 8.2 eV. This value* is considerably lower than the ground state energy of the free atom (– 5.15 eV), obtained by applying the boundary condition $\Psi(r) \to 0$ as $r \to \infty$. The calculated energy of 3s orbitals at the zone boundary is found to be + 2.7 eV. It should be realized that these orbitals are empty because the 3*s* energy band in sodium is only half-filled (refer to Section 7.7) and the corresponding states are located near the top of the band. The negligible amplitude of the Wigner–Seitz wavefunction (often referred to as the cellular wavefunction) at the zone boundary is consistent with this fact since it gives little probability for the states in this region to be occupied. It is significant to notice that all the three wavefunctions are identical in the core of the metal ion.

The shape of the cellular wavefunction at the zone centre ($k = 0$) carries the most vital information about the behaviour of the 3s electron in sodium metal. The plot in Fig. 7.15 shows the variation of the wavefunction as a function of the distance of the electron from the centre of the atomic polyhedron. The wavefunction is flat over about 90 per cent of the atomic volume. The total charge distribution in the flat region corresponds to the charge on an electron. This takes us to the conclusion that $u_k(r)$ remains constant (u_0) over most of the atomic volume and the plane wave part of the wavefunction alone determines the electron motion in this region. Thus the valence electrons of sodium behave mostly as free electrons. This is found to be true for other alkali metals too. But

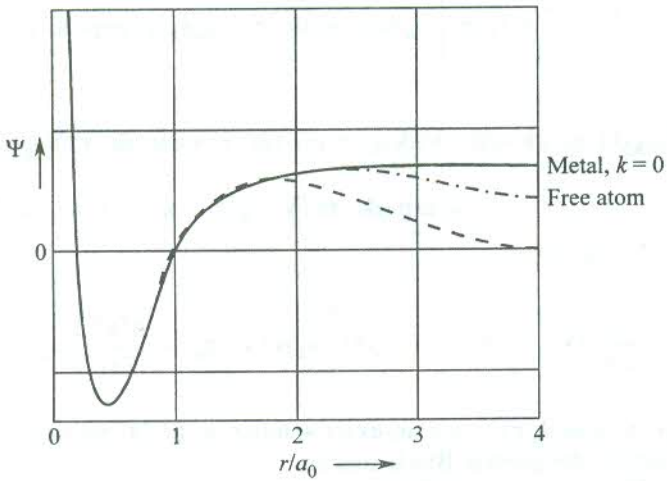


FIG. 7.15 Variation of the radial wavefunction for the 3s electron in sodium. The continuous curve describes the Wigner–Seitz wavefunction (cellular wavefunction) at the Brillouin zone centre ($k = 0$). The lower of the dashed curve represents the cellular wavefunction at the zone boundary. The dashed curve in between the cellular plots denotes the wavefunction in the free atom. The distance r of the electron from the centre of the atomic polyhedron is measured in units of the Bohr radius a_0.

* E.P. Wigner and F. Seitz, *Phys. Rev.*, **43,** 804 (1933); **46,** 509 (1934).

the results for the noble metals, also monovalent, are on the other extreme. The ratio of the ionic to the atomic radius is close to unity, making them act like hard spheres. Because of this reason, the noble metals in contrast to the alkali metals cannot be treated in the framework of the free electron approximation.

7.9.1 Estimation of Cohesive Energy

Concepts of cohesive energy were discussed in Chapter 2. We know that binding of atoms results in the lowering of their ground orbital energy. Thus the ground state energy of an electron in the crystal is lower than that in the free atom. This lowering in energy, taken as a measure of the cohesive energy, is a consequence of replacing the Schrödinger boundary condition in the free atom by the periodic boundary condition (7.70). On the demand of energy conservation, an increase in the binding energy is offset by the Fermi energy contribution to the kinetic energy of valence electrons. The spherical approximation of the Wigner–Seitz theory works satisfactorily in the BCC and FCC crystals. We take sodium metal, a BCC structure, for our discussion.

The electron energy in the crystal is given by (7.75). The value of ε_0 is estimated from the Wigner–Seitz model. The second term in (7.75), denoting the average kinetic energy per electron, is obtained from the free electron theory using (6.52). The ground state energy of the valence electron in the crystal is then,

$$\varepsilon_k = \varepsilon_0 + \frac{\hbar^2 k^2}{2m} = \varepsilon_0 + \frac{3}{5}\,\varepsilon_F \tag{7.76}$$

For sodium, $\varepsilon_0 = -8.2$ eV; $\varepsilon_F = 3.1$ eV (from tables). These values give, $\varepsilon_k = -6.34$ eV.

The above value of ε_k when subtracted from the corresponding value in the free atom (–5.15 eV) gives the cohesive energy as equal to 1.19 eV. The close experimental value (1.13 eV) demonstrates the success of the Wigner–Seitz approximation in alkali metals.

The model provides additional information from the ε_k versus r_s plot. The minimum value of ε_k (7.76) defines the theoretical lattice parameter. For this particular value of r_s, the cohesive energy and compressibility of alkali metals have been calculated and found in good agreement with the experiments.*, †

7.10 METHODS OF BAND STRUCTURE CALCULATION IN USE: A QUALITATIVE VIEW

In the last two sections we discussed two extreme cases of band structure calculation in the form of the tight-binding and Wigner–Seitz approximations. While one overstresses the atomic aspect, the other overstresses the plane wave aspect of the Bloch function. The tight-binding approximation is useful for interpolation and the Wigner–Seitz approximation gives a good account of several properties of alkali metals. But the methods that actually work for a variety of solids are somewhat much different. Though these methods are mathematically tedious, the problem has eased considerably with access to modern computers. No clear-cut prescription can be handed out before beginning the

* F. Seitz, op. cit., p. 365.

† J. Bardeen, *J. Chem. Phys.*, **6**, 367, 372 (1938).

exercise for a certain material. We grow wiser with experience in suitably executing the computer programme which is generally complex.

The scope of this book, however, does not permit us to give a full account of the recent techniques of calculation. Of these the orthogonalized plane wave (OPW) and the augmented plane wave (APW) methods are most prominent and advanced. The pseudopotential method is also often used on account of its ability to predict the energy–wavevector relationship with acceptable accuracy.

We are aware that a single plane wave describes the wavefunction outside the core regions quite well. But a complete wavefunction must be represented by an appropriate combination of a large number of plane waves because of its rapid variation in the vicinity of the core. The knowledge of atomic orbitals at each site readily enables us to distinguish between a plane wave and the complete wavefunction. Based on this idea, Herring formulated the OPW method. The orthogonalized plane waves are in fact linear combinations of plane waves and mixtures of atomic wavefunctions of the occupied states of the cores. This takes care of the electron behaviour both within and outside the core regions. The method has been applied to several metals and non-metals with reasonable success in getting the band shapes.

With a large value of the parameter P in (7.25), the Kronig–Penney model changes over to a one-dimensional form of the APW method. It is another way to improve upon the NFE model by approximating the periodic potential suitably in the regions within and outside the cores. The potential within a sphere around each ion core is taken as the usual atomic potential and assumed constant outside the core regions. The Schrödinger wave equation, when solved in the two regions, yields two separate solutions that are matched on the spherical boundaries between the regions. The wavefunction within the core is expanded in spherical harmonics. But outside the core region it is represented by a combination of plane waves, known as an augmented plane wave. The matching of plane waves onto the atomic functions is the most difficult aspect of the exercise. But with the availability of fast computers these days, the technique is made to work for a large number of metals and semiconductors.

The pseudopotential approach has its roots essentially in the effectiveness of the NFE model (Section 7.4) in many solids. The admixture of core states shows, surprisingly, little effect on the energy of higher states in many metallic materials. In this approach, the periodic potential energy function is replaced by a modified potential energy function with a few Fourier coefficients V_g that refer to only short reciprocal lattice vectors. The modified potential is called the *pseudopotential.* In the well-known Empirical Pseudopotential Method (EPM), the Fourier coefficients are deduced from theoretical fits to the optical reflectance and absorption data of the crystal of interest. The potential is fairly smooth and free from the deep wells of the free potential. Success in the interpretation of band structures notwithstanding, difficulties in the description of some other properties of electron come to surface. It is likely to happen since the method may yield an incorrect wavefunction. Such a problem is resolved by correcting the wavefunction for the due representation of the atomic component.

The band structures as derived from calculations are too complicated to be discussed at this stage. We will refer to these for some simple materials in the next two chapters while gaining more familiarity with the concept of energy surfaces.

SUMMARY

1. The statement of the Bloch theorem is

$$\Psi_{\mathbf{k}}(\mathbf{r}) = u_{\mathbf{k}}(\mathbf{r}) \exp(i\mathbf{k} \cdot \mathbf{r})$$

where $\Psi_{\mathbf{k}}(\mathbf{r})$ denotes an electron wave describing an electron with wavevector $\mathbf{k}$ at $\mathbf{r}$ in the crystal, and $u_{\mathbf{k}}(\mathbf{r})$ is the periodic in the crystal lattice. That is, $u_{\mathbf{k}}(\mathbf{r} + \mathbf{t}_n) = u_{\mathbf{k}}(\mathbf{r})$, where $\mathbf{t}_n$ is an arbitrary translation vector. The function $\Psi_{\mathbf{k}}(\mathbf{r})$ is called the Bloch function.

2. Bloch functions $\Psi_{\mathbf{k}}(\mathbf{r})$ and their eigenvalues $\varepsilon_{\mathbf{k}}$ are periodic in the reciprocal lattice. That is,

$$\Psi_{\mathbf{k}+\mathbf{g}}(\mathbf{r}) = \Psi_{\mathbf{k}}(\mathbf{r})$$

$$\varepsilon_{\mathbf{k}+\mathbf{g}} = \varepsilon_{\mathbf{k}}$$

3. In a crystal of N primitive cells, there are $2N$ independent orbitals in an energy band.

4. Energy bands are separated by regions in which no solutions to electron wave equation exist. These regions are called *band gaps*.

5. The bandwidth is proportional to the overlap of atomic orbitals.

6. The electron effective mass (m^*) is inversely proportional to overlap, i.e. m^* is large for electrons in the inner shells.

7. The cohesive energy of simple metals is estimated by calculating the lowering of the $\mathbf{k} = 0$ orbital in the conduction band. In the calculations, the boundary condition on the wavefunction is changed from Schrödinger ($\Psi \to 0$) as ($r \to \infty$) to Wigner–Seitz $[(d\Psi/dr)_{r=r_s} = 0]$ condition.

PROBLEMS

7.1 Solve the Schrödinger equation for the potential

$$V(x, y, z) = \begin{cases} \infty, & x \le 0 \\ 0, & x \ge 0 \end{cases}$$

and calculate the charge density

$$\rho(x, y, z) = \frac{e}{V} \sum_{k_x} \sum_{k_y} \sum_{k_z} |\Psi_{\mathbf{k}}|^2$$

7.2 For $P \ll 1$ in the Kronig–Penney model, show that the energy of the lowest band at $k = 0$ is given by

$$\varepsilon = \frac{\hbar^2 P}{ma^2}$$

7.3 Make an approximate estimate of the band gap at the corner point of the Brillouin zone for a square lattice in two dimensions, assuming the crystal potential to be given by

$$V(x, y) = -4V_0 \cos\left(\frac{2\pi x}{a}\right)\cos\left(\frac{2\pi y}{a}\right)$$

7.4 Find the energy–wavevector relationship for a one-dimensional monatomic crystal of lattice constant 'a' using the tight-binding approximation. Use this relation to obtain an expression for the effective mass. For which value of k is the electron velocity maximum?

7.5 Using the tight-binding approximation for the overlap up to the nearest neighbours, derive an expression for the energy of an s-band in (i) a BCC crystal and (ii) in an FCC crystal.

7.6 Prove that the density of states in the tight-binding approximation of a simple cubic crystal near the bottom of an s-band is given by

$$D(\varepsilon) = \frac{1}{2\pi^2 a^3 \gamma^{3/2}} (\varepsilon - \varepsilon(0))^{1/2}$$

where $\varepsilon(0)$ is the energy at $k = 0$.

7.7 For a square lattice in two dimensions:

(a) Show that the kinetic energy of a free electron in a corner of the first Brillouin zone is a factor of two larger than that of an electron in the middle of the Brillouin zone edge.

(b) Determine the corresponding factor for a primitive lattice in three dimensions.

(c) Find out if the bands can overlap in a two-dimensional lattice by plotting the following dispersion relations for free electrons in a weak periodic potential.

$\varepsilon(k_y)$ for $k_x = 0$

$\varepsilon(k_y)$ for $k_x = \pi/a$

$\varepsilon(k)$ for $k_x = k_y$

7.8 Explain the variation of forbidden energy gap E_g with temperature in diamond.

SUGGESTED FURTHER READING

Ashcroft, N.W. and N.D. Mermin (Saunders College, 1988).

Brillouin, L., *Wave Propagation in Periodic Structures* (Academic Press, 1960).

Harrison, W.A., *Electronic Structure and the Properties of Solids* (Freeman, 1980).

Madelung, O., *Introduction to Solid State Theory*, Vol. 2 (Springer, 1978).

Ziman, J.M., *Principles of the Theory of Solids* (Cambridge, 1972).

Chapter 8

Mobile Electrons and Fermi Surfaces

Electronic transport properties of metals and semiconductors depend basically on the characteristics of mobile electrons. However, it must be recalled that there are some phenomena which can be explained by the free electron theory only if the existence of positive charge carriers is accepted. The observed positive Hall coefficient for certain metals (see Table 6.3) is an evidence to this effect. The band theory provides a satisfactory explanation to such phenomena by assigning credibility to the concept of positive charge carriers, named as holes. Hence, to be able to account for the transport of charge carriers we need to be familiar with characteristics of holes. In this chapter, we will learn that the motion of charge carriers is strongly controlled by the shapes of the constant energy surfaces. Since the mobile electrons are close to the Fermi surface, a precise knowledge of its shape is a prerequisite to the description of the motion of charge carriers. We begin our task with the discussion on the concept of holes.

8.1 CONCEPT OF HOLES

Let us elaborate on how the concept of a hole evolves. An account of the electronic behaviour, that is linked to the concept, together with the characteristics of holes is given in a series of points as follows.

1. Consider a crystal of unit volume under the action of a steady electric field. Since there can be two electrons in each state, the number of electrons per unit volume of the **k**-space will be equal to $2/8\pi^3$. If $\boldsymbol{v}(\kappa)$ denotes the velocity of an electron with wavevector **k**, the current density owing to electrons in a band occupied up to a certain level can be written as

$$j = -e \int_{\text{occupied}} \frac{\boldsymbol{v}(\mathbf{k})\, d^3\mathbf{k}}{4\pi^3} \tag{8.1}$$

where $d^3\mathbf{k}/4\pi^3$ represents the number of electrons in the small volume element $d^3\mathbf{k}$ around **k**.

Had the unoccupied states in the band been occupied by electrons, their contribution to the current density would be

$$-e \int_{\text{unoccupied}} \frac{\boldsymbol{v}(\mathbf{k})\, d^3\mathbf{k}}{4\pi^3} \tag{8.2}$$

As there is no current in a completely filled band, we have

$$-e \int_{\text{occupied}} \frac{\boldsymbol{v}(\mathbf{k})\, d^3\mathbf{k}}{4\pi^3} - e \int_{\text{unoccupied}} \frac{\boldsymbol{v}(\mathbf{k})\, d^3\mathbf{k}}{4\pi^3} = 0$$

or

$$j = -e \int_{\text{occupied}} \frac{\boldsymbol{v}(\mathbf{k})\, d^3\mathbf{k}}{4\pi^3} = e \int_{\text{unoccupied}} \frac{\boldsymbol{v}(\mathbf{k})\, d^3\mathbf{k}}{4\pi^3} \tag{8.3}$$

Relation (8.3) suggests that the current owing to electrons in the occupied levels is equal to the current that would flow if only the presently unoccupied states in the band were populated, all with positive charges e. These charges are called *holes*. Holes represent essentially the empty states in a band. Holes are thus generally expected to occur in the region close to the top of a band.

2. We refer back to the picture of electron energy bands depicted in Fig. 7.9 and use it to grasp the characteristics of holes. Conventionally, the lower band in the figure is known as the valence band and the upper as the conduction band. Figure 8.1 shows the excitation in which an electron with wavevector $\mathbf{k}_e$ at state E in the valence band has been excited to state Q in the conduction band, on the absorption of a photon of negligible wavevector. This creates a hole at E with a wavevector usually denoted by $\mathbf{k}_h$ in the grapical representation. The relationship between $\mathbf{k}_e$ and $\mathbf{k}_h$ is easily derived with the following arguments. The momentum of a completely filled band is zero. That is,

$$\hbar \sum \mathbf{k} = 0 \tag{8.4}$$

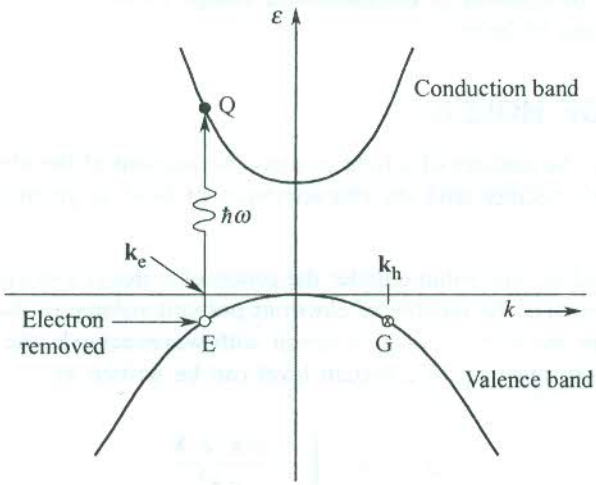


FIG. 8.1 Excitation of an electron with wavevector $\mathbf{k}_e$ from state E in the valence band to state Q in the conduction band by the absorption of a photon of negligible wavevector. The excitation creates a hole with wavevector $\mathbf{k}_h$ $(= -\mathbf{k}_e)$ in the state G of the valence band.

The symmetry of the band in Fig. 8.1 suggests that in a filled band for an electron with wavevector $\mathbf{k}_e$ on the left side of the energy axis there is an electron with wavevector $-\mathbf{k}_e$ on the other side to ensure that (8.4) holds good. When an electron with wavevector $\mathbf{k}_e$ is excited to the conduction band on absorption of an optical photon, the momentum of the valence band becomes

$$\hbar[\sum \mathbf{k} - \mathbf{k}_e] = -\hbar \mathbf{k}_e \tag{8.5}$$

which is rightly equal to the momentum of the electron whose counterpart on the other side has been excited to the conduction band. In consistency with our arguments on unoccupied states, the momentum of the band as expressed by (8.5) can be totally associated with the positive charge (hole) occupying the lone empty state of the band. Therefore, if $\mathbf{k}_h$ denotes the hole wavevector, we have

$$\hbar\mathbf{k}_h = -\hbar\mathbf{k}_e$$

or

$$\mathbf{k}_h = -\mathbf{k}_e \tag{8.6}$$

where $\mathbf{k}_e$ is the wavevector of the missing electron from the valence band.

According to (8.6), the wavevector of the hole is equal to that of the unpaired electron at point G in the valence band whose counterpart (the missing electron) has been transferred to the conduction band.

3. The inversion operator in a symmetric band takes $\mathbf{k}_e$ to $-\mathbf{k}_e$ which denotes the hole wavevector (8.6). This way we can locate the tips of the wavevectors of all the holes that would be created by exciting electrons from the valence band. We thus effectively get the dispersion curve of holes. Taking valence band as the electron band, we associate this curve with the hole band. It is simply the electron band turned upside down as shown in Fig. 8.2. This figure is helpful in simulating the motion of a hole. For a symmetric band,

$$\varepsilon_e(\mathbf{k}_e) = \varepsilon_e(-\mathbf{k}_e) = -\varepsilon_h(-\mathbf{k}_e) = -\varepsilon_h(\mathbf{k}_h)$$

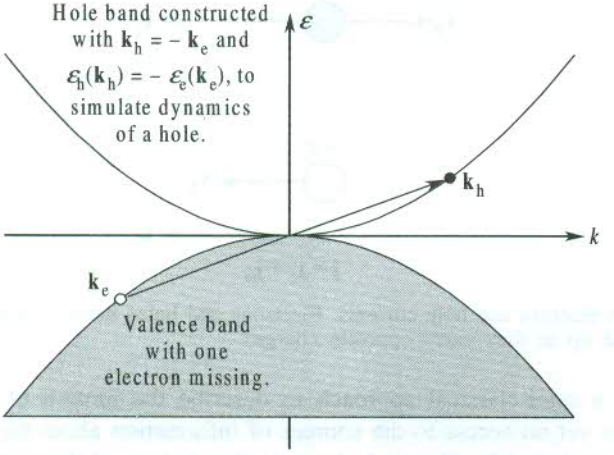


FIG. 8.2 Simulation of hole dynamics by performing the inversion operation in a symmetric band. With the valence band as the electron band, the hole band is constructed using $\mathbf{k}_h = -\mathbf{k}_e$ and $\varepsilon_h(\mathbf{k}_h) = -\varepsilon_e(\mathbf{k}_e)$.

4. In the presence of an electric field the current contribution by the unpaired electron at G is

$$\begin{aligned} -e\mathbf{v}(-\mathbf{k}_e) &= -[-e\mathbf{v}(\mathbf{k}_e)] \\ &= -\text{ (current of the missing electron)} \\ &= \text{current contributed by the hole with charge } +e \end{aligned}$$

Thus the velocity of the hole is equal in magnitude and direction to that of the missing electron, i.e.

$$\mathbf{v}(\mathbf{k}_h) = \mathbf{v}(\mathbf{k}_e)$$

5. It will be shown in the next section that the effective mass is inversely proportional to the curvature, $\frac{d^2\varepsilon}{dk^2}$. Since the curvature of the hole band is opposite to that of the electron band (Fig. 8.2), we get

$$m_h^* = -m_e^* \tag{8.7}$$

where m_h^* and m_e^* denote, respectively, the effective mass of the hole and that of the electron missing from the valence band. For example, we will see that m_e^* is negative near the top of a band. Relation (8.7) implies that m_h^* is positive near the top of the band.

Lastly, we comment on the role of holes in the flow of current. It is sheer for convenience, especially in semiconductors, that we consider the current in the valence band to be carried entirely by holes—the fictitious particles of positive charge—even though the current is actually driven by electrons. In this model, contributions to the total current come from the flow of electrons in the conduction band and from the flow of holes in the valence band. The holes, being centres of the positive charge, move along the direction of the electric field whereas the electrons move opposite to the field direction (Fig. 8.3). Since the direction of the current owing to the two types of carriers, becomes the same, the two components of the current add up to give the net current.

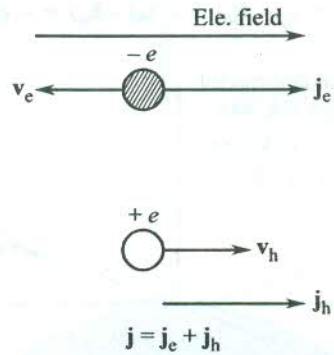


FIG. 8.3 The state of electron and hole currents. Electrons and holes move in opposite directions, but their currents add up as they carry opposite charges.

We will follow a semi-classical approach to describe the motion of electrons and holes in crystals. But we have yet no access to the sources of information about the mass, the velocity and the acceleration for applying the Newton's laws to the motion of these particles. Due caution is demanded in view of the concept of effective mass introduced in Section 7.8. It will be shown below that both the effective mass and the velocity of concern (the group velocity) directly depend on the shape of the Fermi surface.

8.2 EFFECTIVE MASS

The actual motion of an electron in crystals does not conform to its phase velocity, defined as

$$\mathbf{v}_p = \frac{\omega_\mathbf{k}}{\mathbf{k}} \tag{8.8}$$

where $\omega_\mathbf{k}$ and $\mathbf{k}$ denote the angular frequency and the wavevector of the electron, respectively.

The requirement of the value of total wavevector in (8.8) makes it unsuitable for use in the reduced zone scheme which we have adopted for most of the descriptions. On the other hand, the motion of the centre of charge of the wavepacket comprising an electron traces out the actual path of the electron. The wavepacket of energy $\varepsilon_{\mathbf{k}}$ moves around $\mathbf{k}$ in the real space through the crystal at the group velocity, given by

$$\mathbf{v}_g = \frac{\partial \omega}{\partial \mathbf{k}} = \frac{1}{\hbar} \nabla_{\mathbf{k}} \varepsilon \tag{8.9}$$

This definition requires the value of the wavevector as measured from the centre or the boundary of the zone which exactly fits into our scheme. Relation (8.9) shows that the direction of the group velocity is determined by the energy gradient, directed along the surface normal at the point of interest to the constant energy surface. Figure 8.4 shows that the directions of the group velocity $\mathbf{v}_g$ and phase velocity $\mathbf{v}_p$ do not coincide on a non-spherical energy surface. The influence of the periodic crystal potential on electrons causes deviation from the spherical shape of the free electron surfaces.

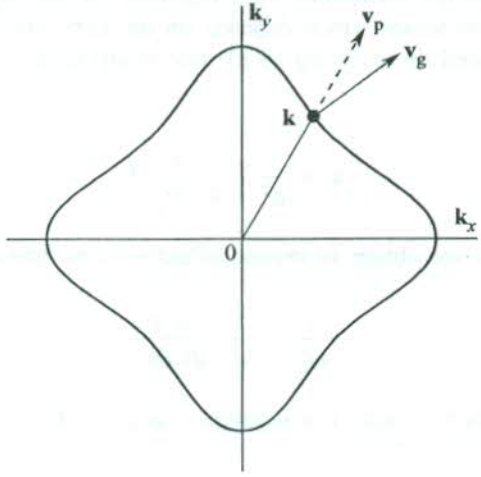


FIG. 8.4 Directions of the phase and group velocities of an electron on a non-spherical constant energy surface.

Under the action of a steady electric field, an electron experiences a force $\mathbf{F}$, equalling $-e\mathbf{E}$. The energy of the electron changes at the rate,

$$\frac{d\varepsilon}{dt} = \mathbf{v}_g \cdot \mathbf{F} \tag{8.10}$$

But

$$d\varepsilon = \left(\frac{d\varepsilon}{dk}\right) \cdot dk = \nabla_{\mathbf{k}} \varepsilon \cdot dk$$

or

$$\frac{d\varepsilon}{dt} = \frac{1}{\hbar} \nabla_{\mathbf{k}} \varepsilon \cdot \hbar \frac{dk}{dt} \tag{8.11}$$

From (8.10) and (8.11), we get

$$\mathbf{F} = \hbar \frac{d\mathbf{k}}{dt} \tag{8.12}$$

Relation (8.12) is an important relation which in accordance with the Newton's second law establishes that the electron momentum is given by $\hbar\mathbf{k}$.

Further, from (8.9) the acceleration of the electron **a** can be written as

$$\mathbf{a} = \frac{d\mathbf{v}_g}{dt} = \frac{1}{\hbar^2} \nabla_{\mathbf{k}} \nabla_{\mathbf{k}} \varepsilon \cdot \hbar \frac{d\mathbf{k}}{dt}$$

$$= \frac{1}{\hbar^2} \nabla_{\mathbf{k}} [\nabla_{\mathbf{k}} \varepsilon \cdot \mathbf{F}] \tag{8.13}$$

The above equation tells us that **F** can produce a change in $\mathbf{v}_g$ in directions other than the direction of **F**. Comparing (8.13) with the Newtonian force equation, we see that the closest thing to a mass of the electron is an inverse tensor which depends on the curvature of the ε versus **k** dispersion curve. This can be determined by resolving (8.13) into components along the three arbitrary axes. Then,

$$\mathbf{a}_i = \frac{1}{\hbar^2} \sum_j \frac{\partial^2 \varepsilon}{\partial k_i \partial k_j} \mathbf{F}_j \tag{8.14}$$

with $i, j = x, y, z$; and thus we obtain an inverse effective mass tensor with components

$$\frac{1}{m_{ij}^*} = \frac{1}{\hbar^2} \frac{\partial^2 \varepsilon}{\partial k_i \partial k_j} \tag{8.15}$$

where m_{ij}^* is called the *effective mass*. For isotropic energy surfaces, the effective mass is simply expressed by

$$\frac{1}{m^*} = \frac{1}{\hbar^2} \frac{d^2 \varepsilon}{dk^2} \tag{8.16}$$

In view of (8.15) or (8.16), the effective mass of an electron in an energy band is positive near the bottom of the band where the curvature is upward or $d^2\varepsilon/dk^2$ is positive. Similarly, it is negative near the top of the band (see Fig. 7.6). The change in sign of the effective mass occurs at the point of inflexion on the dispersion curve. In the two-component approximation of the wavefunction (7.41), the component $\exp[i(k - g)x]$ represents the reflected wave component whose increase owing to an increase in k is a measure of the momentum transferred to the crystal lattice. At the Brillouin zone boundary (the top of band) it exactly cancels the forward moving component $\exp(ikx)$, resulting in Bragg reflection. Near the top of the band, the approach to Bragg reflection causes an overall decrease in the forward momentum of the electron even though the applied electric field is kept on increasing. It means that the electron transfers more momentum to the lattice than what is imparted to it by the field on increasing **k** to $\mathbf{k} + \Delta\mathbf{k}$. This situation demands that the effective mass be treated as negative to satisfy the Newton's second law.

Since the curvature of the hole band is opposite to that of the electron band (see Fig. 8.2), the effective mass of a hole is positive near the top and negative near the bottom of a band. The condition (8.7) is based on the observations made in this section.

8.3 CONSTRUCTION OF THE FERMI SURFACES

As stated in the beginning, the knowledge of the shapes of the Fermi surfaces is helpful in investigating the motion of electrons in crystals. The exercise of deriving the form of the Fermi surface involves the following steps:

(i) Construction of the Brillouin zones

(ii) Construction of the Fermi surface in the extended zone scheme

(iii) Reproducing pictures in the reduced and periodic zone schemes.

The condition of Bragg reflection, $2\mathbf{k} \cdot \mathbf{g} + \mathbf{g}^2 = 0$, determines the boundaries of different Brillouin zones. This condition requires that the tip of the vector $\mathbf{g}/2$ should touch the plane perpendicular to it at the mid-point of the vector $\mathbf{g}$. Considering a square lattice of lattice constant 'a' we define the sets of the three shortest wavevectors as

$$\mathbf{g}_1 = \frac{2\pi}{a}\,\hat{\mathbf{k}}_x$$

and the three equivalent vectors as

$$-\frac{2\pi}{a}\,\hat{\mathbf{k}}_x; \qquad \pm\frac{2\pi}{a}\,\hat{\mathbf{k}}_y \tag{8.17}$$

$$\mathbf{g}_2 = \frac{2\pi}{a}\,(\hat{\mathbf{k}}_x + \hat{\mathbf{k}}_y)$$

and the three equivalent vectors as

$$-\frac{2\pi}{a}\,(\hat{\mathbf{k}}_x + \hat{\mathbf{k}}_y); \qquad \frac{2\pi}{a}\,(-\hat{\mathbf{k}}_x + \hat{\mathbf{k}}_y); \qquad \frac{2\pi}{a}\,(\hat{\mathbf{k}}_x - \hat{\mathbf{k}}_y) \tag{8.18}$$

$$\mathbf{g}_3 = \frac{4\pi}{a}\,\hat{\mathbf{k}}_x$$

and the three equivalent vectors as

$$-\frac{4\pi}{a}\,\hat{\mathbf{k}}_x; \qquad \pm\frac{4\pi}{a}\,\hat{\mathbf{k}}_y \tag{8.19}$$

where $\hat{\mathbf{k}}_x$ and $\hat{\mathbf{k}}_y$ denote unit vectors along the $\mathbf{k}_x$ and $\mathbf{k}_y$ axes, respectively.

The perpendicular bisectors of $\mathbf{g}_1$ and its three equivalent vectors by symmetry enclose the area of the first Brillouin zone. These bisectors join the similar bisectors of $\mathbf{g}_2$ and its allies to bound four separate parts of the second zone. Continuing the exercise for (8.19), we find that the third zone appears as eight separate pieces (Fig. 8.5).

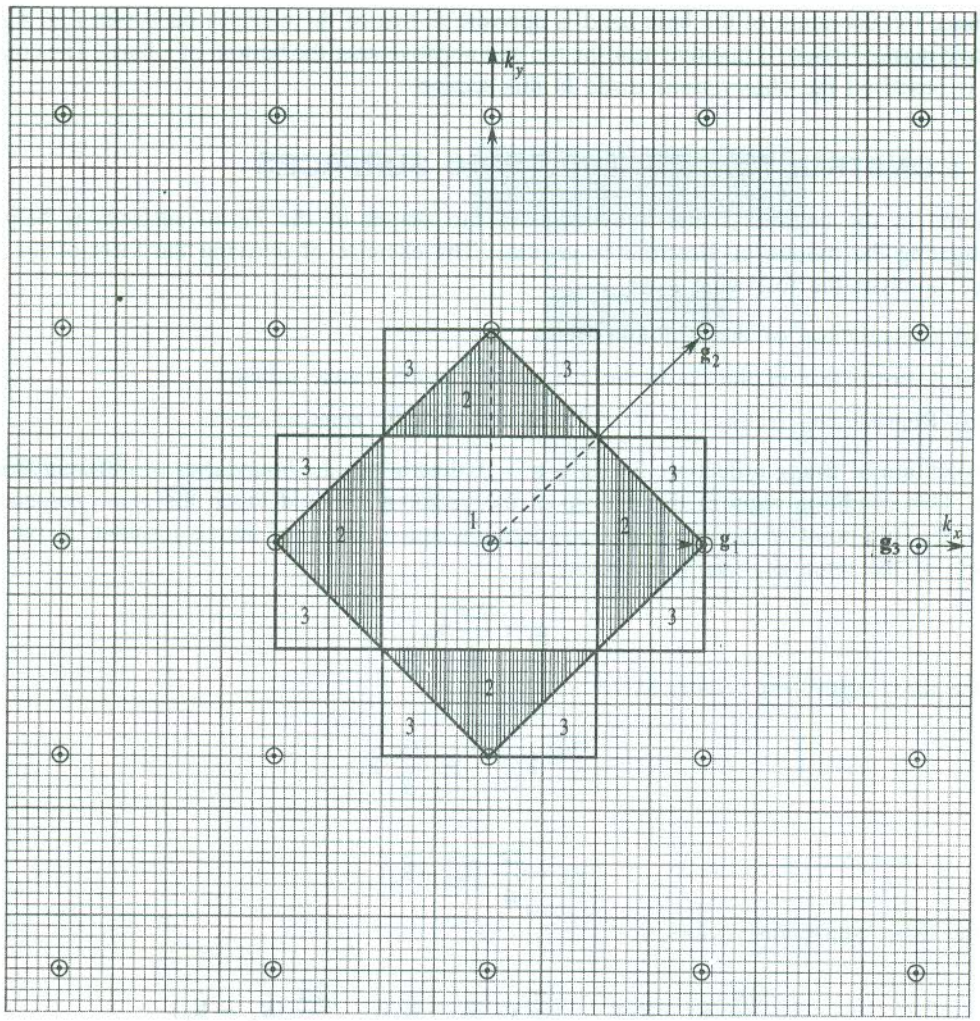


FIG. 8.5 Boundaries and locations of the first three Brillouin zones of a square lattice. $\mathbf{g}_1$, $\mathbf{g}_2$ and $\mathbf{g}_3$ are the first three shortest reciprocal lattice vectors whose perpendicular bisectors form the boundaries of zones. The second and third zones appear in separated pieces.

The Fermi surfaces are constructed either by calculations or by experimental methods to be described later in this chapter. We consider at this stage the case of free electrons whose Fermi surfaces are the simplest to construct by using the Harrison's method. After determining the reciprocal lattice points of the crystal (ideally a metal) of interest, spheres around each lattice point (as centre) are drawn, taking the radius as a proportionate measure of the electron density. An occupied state in the first zone is represented by a point in the k-space that lies within at least one sphere. Similarly, the points lying within at least two spheres correspond to the occupied states in the second zone and the interpretation continues for points in three or more spheres.

The free electron Fermi surface for a square lattice is shown in Fig. 8.6. It extends into the third Brillouin zone. It is straightforward to show that parts of the second and third zones can be translated to the first zone by reciprocal lattice vactors, each with the same magnitude ($2\pi/a$). For example, the

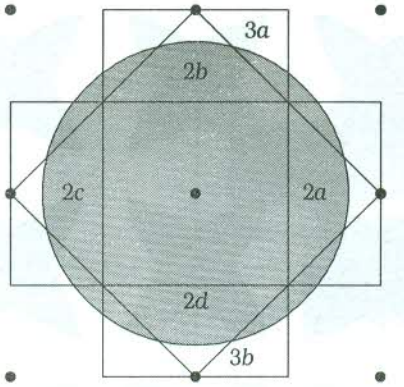


FIG. 8.6 The constant energy surface (the Fermi surface) of free electrons for a certain electron concentration of a square lattice with its first three Brillouin zones. The Fermi surface encloses only the first zone completely. The symbols *a*, *b*, *c*, *d* refer to the different parts of a zone.

part $2a$ is translated by $-2\pi\hat{\mathbf{k}}_x/a$. The first three zones in the reduced zone scheme are drawn in Fig. 8.7. The Fermi surfaces as viewed in the reduced zone scheme are pictured in Fig. 8.8. The parts of the Fermi surface in the third zone are still not connected. When drawn in the periodic zone scheme (Section 7.5), these parts join to form a lattice of rosettes that symbolizes the Fermi surface in this zone (Fig. 8.9).

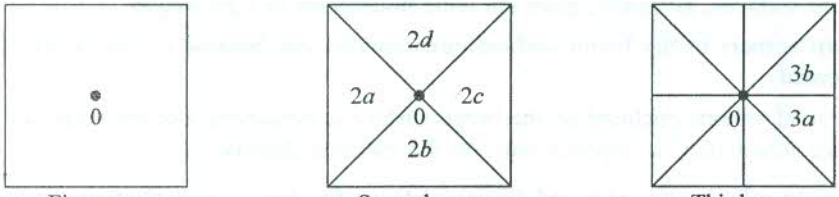


FIG. 8.7 The first, second and third Brillouin zones as pictured in the reduced zone scheme. Parts of the second and third zones are put together in a square on translating them by reciprocal lattice vectors of appropriate size.

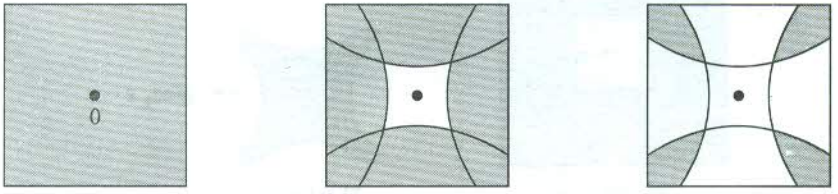


FIG. 8.8 The free electron Fermi surface of Fig. 8.6 in the reduced zone scheme. The occupied electron states are represented by shaded areas. Note that only the first zone is completely occupied.

Taking a realistic step forward, we explore how the Fermi surfaces of nearly free electrons can be derived. The proposal closely corresponds to the case of alkali metals which respond to the NFE model most positively. The Fermi surfaces of nearly free electrons are derived from those of the free electrons by taking the following into consideration:

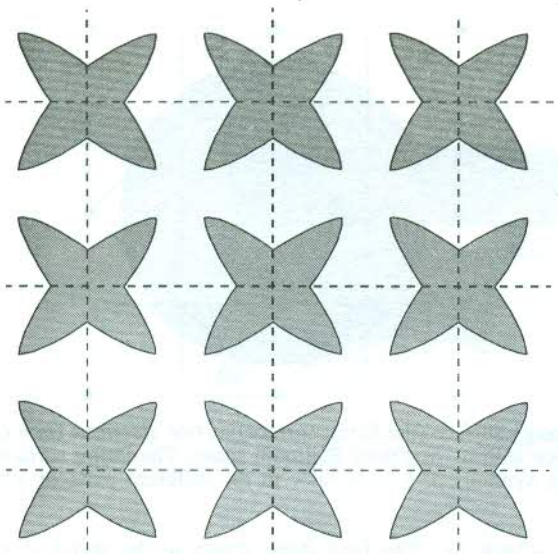


FIG. 8.9 The third zone picture of the Fermi surface in the periodic zone scheme.

(i) Energy gaps are created at the zone boundaries because of the interaction of the electron with the periodic crystal potential.

(ii) Fermi surfaces, generally, meet the zone boundaries at right angles.

(iii) Sharp corners in the Fermi surfaces are rounded out because of the effect of the crystal potential.

(iv) The total volume enclosed by the Fermi surface is independent of the details of the electron lattice interaction. It depends only on the electron density.

On qualitative basis, the expected Fermi surfaces for nearly free electrons are as shown in Fig. 8.10. These are only marginally modified from those of the free electrons (see Figs. 8.8 and 8.9).

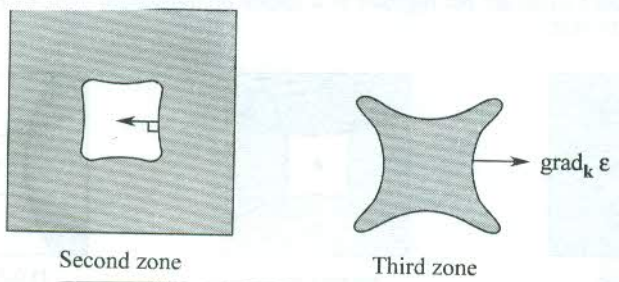


FIG. 8.10 Fermi surfaces in second and third zones for nearly free electrons. The direction of $\text{grad}_k\ \varepsilon$ shows that in the second zone the energy increases towards the interior of the figure and in the third zone the energy increases towards the exterior of the figure. The Fermi surface in the second zone is holelike and in the third zone electronlike.

The Fermi surfaces of alkali metals as determined in the NFE approximatioin are nearly spherical, in agreement with the experiments. To quote, the surface is closely spherical for sodium; but in cesium a deviation of 10 per cent from the spherical shape is observed. In noble metals too,

the shape is spherical except that the surface approaches the surface of the zone most closely along [111] directions (see Fig. 8.22). The divalent metals beryllium and magnesium have weak lattice interactions and their Fermi surfaces are nearly spherical. For the trivalent metal aluminium, the figures in the first three zones are drawn in Fig. 8.11. The section of the Fermi surface in the third zone is funny and known as 'monster', typified by eight tentacles as shown in Fig. 8.11(c). Generally, the predictions of the Harrison free electron model are in good agreement with the features of the Fermi surfaces as revealed by experiments. The monster, however, needs be modified to the form shown in Fig. 8.11(d).

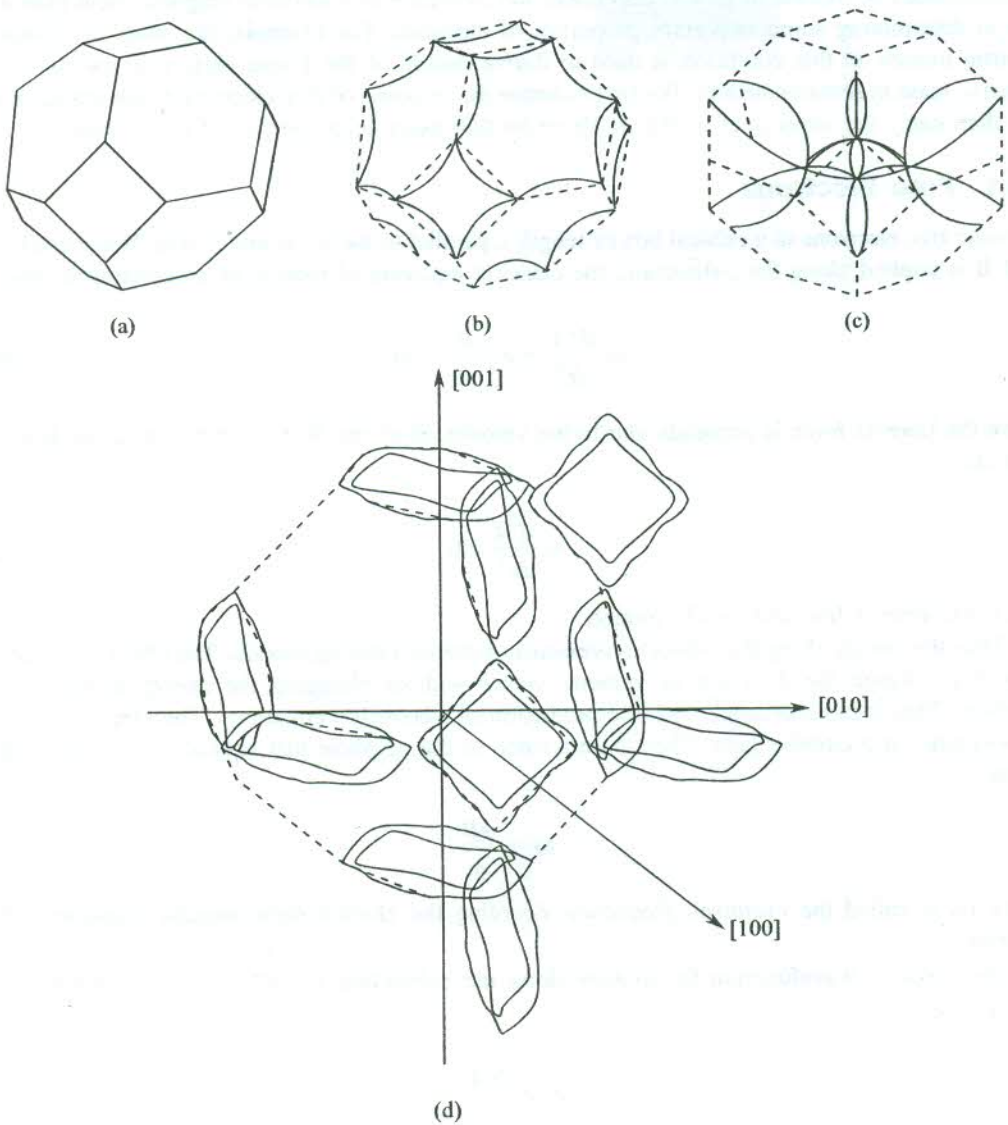


FIG. 8.11 The Fermi surface of aluminium in the (a) first zone, (b) second zone and (c) third zone showing the monster with eight tentacles. Figure 8.11(d) shows the modified form of the monster as derived with the use of experimental data. [After E.P. Volskii, *J.E.T.P.*, **46**, 123 (1963).]

Now, we need to be familiar with the motion of electrons in the presence of magnetic fields as it is crucial to experiments that are conducted to supplement all discussions carried out so far. This will largely enable us to have a picture of the dynamics of electron motion in crystals. Features of the motion are crucial to the determination of the effective mass and the Fermi surface, slated for discussion at a later stage in this chapter.

8.4 ELECTRONS IN A UNIFORM MAGNETIC FIELD

Characteristics of motion of an electron under the influence of a uniform magnetic field play a key role in determining some important properties of electrons. For example, the study of changes in electron motion in this condition is used to derive details of the Fermi surface in metals and the effective mass in semiconductors. We first examine the response of free electrons to the magnetic field and then carry the ideas over to Bloch electrons that describe properties of real solids.

8.4.1 Free Electrons

Consider free electrons in a cubical box of length L parallel to the x-, y- and z-axes. When a magnetic field $\mathbf{B}$ is applied along the z-direction, the classical equation of motion of an electron is given by

$$m \frac{d^2 \mathbf{r}}{dt^2} = e \left(\frac{d\mathbf{r}}{dt} \times \mathbf{B} \right) \tag{8.20}$$

where the Lorentz force is perpendicular to the velocity $d\mathbf{r}/dt$ and $\mathbf{B}$. Its component along $\mathbf{B}$ is zero. That is,

$$m \frac{d^2 \mathbf{z}}{dt^2} = 0 \tag{8.21}$$

which describes a free and steady motion.

Thus the energy along the z-direction remains unchanged during motion. The effect of a magnetic field is to change the direction of velocity vector without changing the energy along the field direction. This is in accordance with the prediction of Maxwell's equations. The electron moves in the xy-palne on a circular path. The Lorentz force in the xy-plane just equals the centripetal force giving,

$$\omega_c = \frac{eB}{m} \tag{8.22}$$

where ω_c is called the *cyclotron frequency*, denoting the characteristic angular frequency of the electron.

The electron wavefunction for motion along the z-direction is $\frac{1}{\sqrt{L}} \exp(ik_z z)$ which has the eigenvalue,

$$\varepsilon_z = \frac{\hbar^2 k_z^2}{2m} \tag{8.23}$$

with k_z taking the same values as in the absence of a magnetic field, i.e.

$$k_z = \frac{2\pi n_z}{L}; \qquad n_z \text{ being any integer} \tag{8.24}$$

But the solution of the one-electron Schrödinger wave equation in the *xy*-plane yields an energy that is quantized in steps of $\hbar\omega_c$ and is not simply $\frac{\hbar^2}{2m}(k_x^2 + k_y^2)$ as expected in the absence of the field. The electron energy in the *xy*-plane is expressed as

$$\varepsilon_\perp = \left(n + \frac{1}{2}\right)\hbar\omega_c \tag{8.25}$$

where n is a positive integer. The total electron energy is

$$\varepsilon_n(k_z) = \varepsilon_z + \varepsilon_\perp$$

$$= \frac{\hbar^2 k_z^2}{2m} + \left(n + \frac{1}{2}\right)\hbar\omega_c \tag{8.26}$$

The relation (8.26) expresses the orbit quantization. We will see later that an energy level with a given n and arbitrary k_z is highly degenerate. Each group of degenerate energy levels that corresponds to a magnetic level as defined by a given pair of quantum numbers n and k_z is called a *Landau level.*

The area of the electron orbit is another quantity which is involved in certain remarkable physical phenomena to be discussed in this chapter. With this purpose, we rewrite (8.20) as

$$\hbar\frac{d\mathbf{k}}{dt} = e\left(\frac{d\mathbf{r}}{dt} \times \mathbf{B}\right) \tag{8.27}$$

The above equation expresses force in terms of the rate of change of the wavevector under the assumption that the electron moves without scattering. We must appreciate that the wavevector changes only in the *xy*-plane ($\perp_r$ to **B**) with k_z remaining constant throughout the motion. The tip of the wavevector ($\mathbf{k}_\perp$) in the *xy*-plane traces out a circle whose area in the **k**-space is $\pi k_\perp^2$. According to quantum mechanics the electron energy in the *xy*-plane ($\varepsilon_\perp$) can be expressed as $\frac{\hbar^2 k_\perp^2}{2m}$. On equating it to the quantized energy as given by (8.25), we get

$$S = \pi k_\perp^2 = \left(n + \frac{1}{2}\right)\frac{2\pi eB}{\hbar} \tag{8.28}$$

where S denotes the area of the orbit in the **k**-space.

The difference between the areas of the consecutive orbits is

$$S_{n+1} - S_n = \left(\frac{2\pi eB}{\hbar}\right) \tag{8.29}$$

Hence the areas enclosed by the electron path in a magnetic field are quantized in units of $2\pi eB/\hbar$. The quantized areas (8.28) are cylindrical in shape. They have come to be known as *Landau tubes.* The magnetic field constantly changes the direction of the velocity vector in such a way that the tips of the allowed **k** vectors touch a set of Landau tubes as shown in Fig. 8.12.

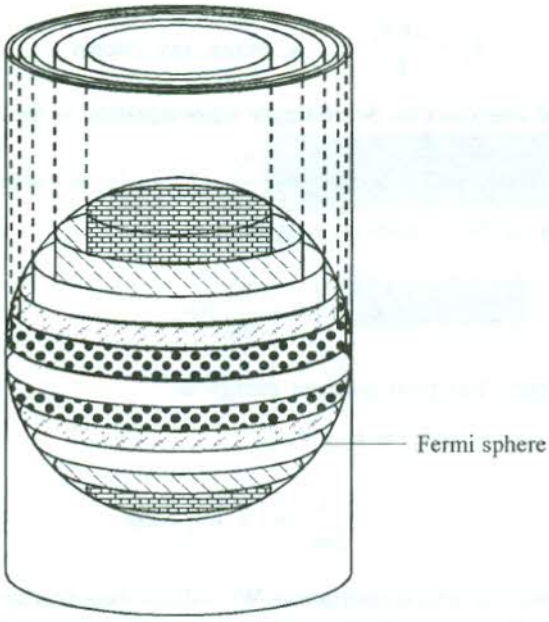


FIG. 8.12 Landau tubes (occupied regions of the **k**-space) for a free electron gas in a magnetic field. The allowed states lie on Landau cylinders, cut off by the Fermi sphere.

We are using the Onsager's generalization of Landau free electron results which is valid for magnetic levels with only high quantum numbers. Thus only for large n is the difference $(S_{n+1} - S_n)$ a constant and independent of k_z. Electrons occupy the surface of Landau tubes. All the allowed states in the **k**-space coalesce onto these tubes as soon as even a small magnetic field is switched on. The distance between the tubes is very small for normal magentic fields. The total number of states remains unchanged as the states for each given k_z that lie between the tubes prior to coalescence migrate onto the tube after coalescence. But it must be borne in mind that in a magnetic field the occupied states lie on only those parts of the Landau tubes that are within the Fermi sphere.

8.4.2 Bloch Electrons

It is simple to extend the above results to electrons in an isotropic band. The motion in the **k**-space is generalized by replacing m by the effective mass m^*. The velocity in (8.27) is taken as $\hbar\mathbf{k}/m^*$. All calculations proceed on the same lines using,

$$\omega_c = \frac{eB}{m^*} \tag{8.30}$$

In order to deal with motion in general bands, we make use of (8.27) which remains valid. The electron orbits may not be circular but remain closed, lying on the surfaces of constant energy in the **k**-space. We write the period of the orbit as

$$T_c = \oint dt \tag{8.31}$$

The integral is evaluated over one complete orbit in the plane perpendicular to **B**. From (8.27), we have

$$\delta t = \frac{\hbar \delta k}{e v_\perp B}$$

$$= \frac{\hbar^2 \delta k}{eB} \frac{\Delta k_\perp}{\Delta \varepsilon} \qquad \text{(using 8.9)} \tag{8.32}$$

With $\oint k_\perp \, dk = S$ (the area of the orbit), we have

$$T_c = \frac{\hbar^2}{eB} \cdot \frac{\partial S}{\partial \varepsilon} \tag{8.33}$$

$$\omega_c = \frac{2\pi eB}{\hbar^2} \cdot \frac{1}{\partial S / \partial \varepsilon} \tag{8.34}$$

Comparing (8.33) with (8.30), we define a *cyclotron mass* as

$$m_c^* = \frac{\hbar^2}{2\pi} \frac{\partial S}{\partial \varepsilon} \tag{8.35}$$

It is emphasized that this effective mass is not necessarily the same as defined in other contexts, e.g. specific heat effective mass.*

Now, we make use of the Bohr's correspondence principle according to which the difference in energy of the two adjacent levels is equal to the product of Planck's constant and the frequency of classical motion at the energy of the levels. As k_z is a constant of the semi-classical motion, we apply this condition to levels with a given k_z.

$$\varepsilon_{n+1} - \varepsilon_n = \frac{h}{T_c} \tag{8.36}$$

Substituting the value of T_c from (8.33), we get

$$(\varepsilon_{n+1} - \varepsilon_n) \frac{\partial S}{\partial \varepsilon} = \frac{2\pi eB}{\hbar} \tag{8.37}$$

Based on the free electron results, the energy difference between the neighbouring Landau levels is expected to be of order $\hbar\omega_c$. Since the energies of levels themselves are at least 10^4 times greater than this, it is an excellent approximation to replace the differential quotient $\partial S/\partial \varepsilon$ with the difference quotient $(S_{n+1} - S_n)/(\varepsilon_{n+1} - \varepsilon_n)$. Doing so in (8.37), we get

$$S_{n+1} - S_n = \frac{2\pi eB}{\hbar} \tag{8.38}$$

*N.W. Ashcroft and N.D. Mermin, *Solid State Physics*, Ch. 2 (Saunders College, 1988).

which is the same result as for free electrons (8.29). By rewriting (8.38) as

$$S_{n+1} = S_n + \frac{2\pi eB}{\hbar}$$

we conclude that for large n the area of the semi-classical orbit may be expressed as

$$S_n = (n + \gamma)\ \frac{2\pi eB}{\hbar} \tag{8.39}$$

where γ denotes the phase correction that is independent of n. It is equal to 1/2 for free electrons. The orbit areas are quantized in units of $2\pi eB/\hbar$. Relation (8.39) is the famous Onsager's quantization condition.

The features of electron dynamics in electric and magnetic fields, as treated so far, have been exploited to plan experiments for evaluating the effective mass, the shape and the size of Fermi surfaces. The method of cyclotron resonance is generally used to determine the effective mass of charge carriers. But wide ranging experimental methods are used for the study of Fermi surfaces. Some of the principles that form the basis of experiments being conducted are: anomalous skin effect, cyclotron resonance, de Haas–van Alphen effect, Shubnikov–de Haas effect, magneto-resistance. On the basis of relative experimental and theoretical significance we have selected the methods of anomalous skin effect, cyclotron resonance and de Haas–van Alphen effect for our discussion.

8.5 ANOMALOUS SKIN EFFECT

In one of the earliest experiments on the determination of the Fermi surfaces, Pippard (1957) studied the reflection and absorption of microwaves in copper. It is observed that the electric field of microwave radiation makes only a limited penetration into the metal. Tendency of a metal to prevent the penetration of an alternating electric field into it is called the *skin effect*. If the microwave frequency is not too high, the depth to which the field penetrates into the metal can be calculated using the classical theory. The penetration depth is called the *classical skin depth* and is given by

$$\delta_0 = \frac{c}{\sqrt{2\pi\sigma\omega}} \tag{8.40}$$

where c is the speed and ω the angular frequency of the electromagnetic radiation; and σ denotes the electrical conductivity of the metal. δ_0 is also defined as the distance inside the metal over which the amplitude of the electric field of radiation is damped to 1/e of its initial value.

The derivation of (8.40) is valid when the skin depth δ_0 is far more greater than the electron mean free path Λ (i.e. $\delta_0 >> \Lambda$). This condition is met in metals at room temperature and above. The resulting skin effect in these conditions is called the *normal skin effect*. But, at low temperatures, especially when the metal sample is very pure the electron mean free path becomes greater than the skin depth. The skin depth may be about 100 times less than the value predicated by (8.40). Under these conditions the classical theory breaks down and the decrease in the amplitude of radiation field will no longer be exponential with distance. This effect is known as the *anomalous skin effect*. In conditions of the anomalous effect only a small number of electrons that move almost tangentially to the metal surface, interact with the radiation field and complete one free path totally within the skin depth.

The most interesting property of anomalous skin effect is, however, its dependence on the geometry of the Fermi surface. Some features of the Fermi surface are known to depend only on the orientation of the Fermi surface relative to the actual surface of the metal sample. Further, the data on reflectivity and absorption of microwaves show correlation with these features making it possible to derive some information about the Fermi surface. This technique is, however, rarely used on account of the limited information provided by it.

8.6 CYCLOTRON RESONANCE

We saw above that the band electrons circulate with the cyclotron frequency ω_c under the action of a magnetic field. The electrons circulate around the field in a plane perpendicular to it. The electron energy in the plane is quantized in units of $\hbar\omega_c$. Therefore, a resonance in the absorption of energy takes place when the crystal is exposed to a suitably phased electromagnetic radiation of frequency ω_c. This is called *cyclotron resonance* on whose principle a cyclotron operates. The resonant magnetic field is searched by varying the field. The useful range of frequencies of the electro-magnetic radiation falls in the microwave region.

For a sharp resonance, the electron scattering has to be minimum. Our assumption that there are no centres of scattering in the crystal may not be justified in absolute sense. Even the most perfect and the purest available crystals are known to have some imperfections and impurities to scatter electrons. Hence, for the measurement of ω_c, experiments must be performed at low temperatures and high magnetic fields. This suppresses the scattering and thermal fluctuations. The sharpness of resonance, however, depends primarily on the broadening of energy levels. From the uncertainty principle, the broadening $\Delta\varepsilon$ is given by

$$\Delta\varepsilon \simeq \frac{\hbar}{\tau} \tag{8.41}$$

Therefore, if $\frac{\hbar}{\tau} > \hbar\omega_c$, no resonance features will be observed as the electron would be scattered before it could complete one revolution. Thus the basic condition for the cyclotron resonance is

$$\omega_c\tau >> 1 \tag{8.42}$$

Another important factor of concern is the skin depth available to the microwaves used as the electromagnetic radiation. The larger the conductivity of the crystal, the smaller the skin depth. The skin depth should be larger than both the electron mean free path and the size of the orbit in the resonant magnetic field. This ensures that the electron feels the electromagnetic field once per revolution. In view of this fact, the experimental requirements for semiconductors and metals are different needing separate discussions.

8.6.1 Semiconductors

The cyclotron resonance in semiconductors is conveniently observed because of their low electrical conductivity which facilitates the penetration of the radio frequency field deeper into the sample. The arrangement of the fields is shown in Fig. 8.13. Using (8.30) we make a rough estimate of ω_c. Taking m^* as 10^{-31} kg for a magnetic field of 1 tesla,

$$\omega_c \simeq 10^{12}\,\mathrm{s}^{-1} \quad \text{or} \quad \nu_c \simeq 2 \times 10^{11}\ \mathrm{Hz} = 200\ \mathrm{GHz}$$

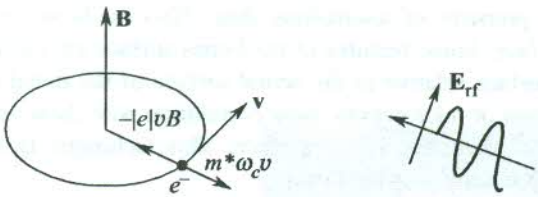


FIG. 8.13 Description of cyclotron resonance for electrons in a semiconductor. The radio frequency field $\mathbf{E}_{rf}$ lies in the plane of the orbit around the static magnetic field **B**. The fields **B** and $\mathbf{E}_{rf}$ are mutually perpendicular. Holes circulate in the clockwise direction in this set-up.

This is an inconveniently high frequency that corresponds to a wavelength of about 1 mm. Generally, the experiment is performed at liquid helium temperature with pure samples to ensure a large value of τ. Then, we can use microwaves of about 1 cm wavelength in a field of 0.1 tesla and satisfy the requirement $\omega_c\tau >> 1$.

The theory of cyclotron resonance is fairly complicated for both semiconductors and metals. Only a simple account of the same is given here. For a parabolic energy surface, the orbit area is proportional to k^2. Notwithstanding its dependence on the direction of **k**, the energy ε_k is thus a linear function of S. The gradient $\partial\varepsilon/\partial S$ in the expression for ω_c (8.34) then remains constant in the whole region of the band. As an example, we take the conduction band of Ge that is a spheroid. When the magnetic field lies in the xz-plane and is directed at angle θ with the z-direction, the form of the ellipse perpendicular to the field direction is given by

$$\varepsilon = \frac{\hbar^2}{2}\left[\frac{k_1^2}{m_t^*} + k_2^2\left(\frac{\cos^2\theta}{m_t^*} + \frac{\sin^2\theta}{m_l^*}\right)\right] \tag{8.43}$$

and

$$\frac{\varepsilon}{S} = \frac{\hbar^2}{2\pi}\left[\frac{\cos^2\theta}{m_t^{*2}} + \frac{\sin^2\theta}{m_t^* m_l^*}\right]^{1/2} \tag{8.44}$$

The effective masses m_t^* and m_l^* are determined with the knowledge of ω_c and θ, available from the experiment. Here t stands for transverse and l for longitudinal.

In practice, the cyclotron resonance is essentially used to determine the effective mass of charge carriers in semiconductors. The resonance absorption spectra give distinct evidence for the presence of electrons and holes. The direction of the hole orbit is, however, opposite to that of the electron orbit. Hence the distinction between electrons and holes can be made by using a circularly polarized high frequency microwave field incident along the axis of **B** field. Figure 8.14 gives the cyclotron resonance spectrum of a germanium crystal where the magnetic field is oriented in the (110) plane at an angle of 60° with the [100] direction. Each peak corresponds to a certain effective mass. The explanation to the observation of light and heavy holes in the spectrum may be sought in the band structure at the valence band maximum where the hole states exist. But, a simple treatment of the band structure, which is rather complicated in this region, fails to meet the objective. The exercise is deferred to the next chapter for a complete answer.

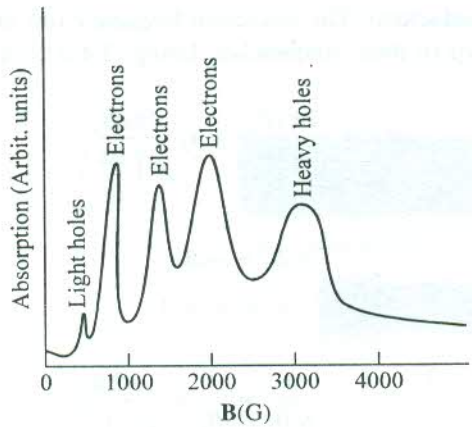


FIG. 8.14 Cyclotron resonance in Ge with magnetic field oriented in a (110) plane at an angle of 60° with the [100] direction. The two valence bands, degenerate at $\mathbf{k} = 0$ with different curvatures, give rise to two peaks marked as 'light' and 'heavy' holes. [After G. Dresselhaus, A.F. Kip, C. Kittel, *Phys. Rev.*, **98,** 368 (1955).]

8.6.2 Metals

The presence of a large number of mobile electrons endows metals with a high electrical conductivity. As a result the skin depth available to the incident mocrowaves is smaller than both the electron mean free path and the dimensions of the electron's real space orbit in the magnetic field corresponding to ω_c. Due to this reason in this case a different geometry is used, as proposed by Azbel and Kaner and shown in Fig. 8.15. The special feature of the geometry is that the steady magnetic field **B** is parallel to the specimen's surface. It makes the electrons available in the skin depth region while performing their spiral orbits. These electrons absorb energy from microwaves if they see the electric field in right direction once during each revolution. On account of small skin depth, the resonance is observed only at harmonics of the cyclotron frequency:

$$\omega = n\omega_c$$

$$= n\frac{2\pi eB}{\hbar^2\left(\frac{\partial S}{\partial \varepsilon}\right)} \qquad \text{(from 8.34)} \tag{8.45a}$$

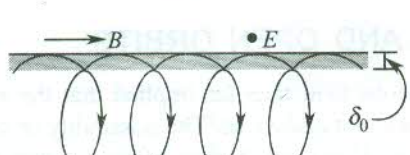


FIG. 8.15 Parallel-field Azbel–Kaner geometry for cyclotron resonance study in metals.

In a normal experimental set-up, a microwave source of fixed frequency is used and the magnetic field is varied to locate the resonance. Suppose two successive resonances are observed for the magnetic fields B_1 and B_2. It implies that B_1 and B_2 must correspond to fields that could have shown resonance for two separate fundamental frequencies of microwaves in a material with

large skin depth (e.g. semiconductors). The resonance frequency (ω) in the present case represents two harmonics (say n_1 and n_2) of these frequencies. Using (8.45a), we can write

$$\omega = n_1 \frac{2\pi e B_1}{\hbar^2 \left(\frac{\partial S}{\partial \varepsilon}\right)} = n_2 \frac{2\pi e B_2}{\hbar^2 \left(\frac{\partial S}{\partial \varepsilon}\right)} \tag{8.45b}$$

or
$$n_1 B_1 = n_2 B_2$$

If
$$n_2 = n_1 + 1$$

$$\frac{1}{B_2} = \frac{1}{B_1} + \frac{1}{n_1 B_1} = \frac{1}{B_1} + \frac{2\pi e}{\hbar^2 \left(\frac{\partial S}{\partial \varepsilon}\right)\omega}$$

or
$$\delta\left(\frac{1}{B}\right) = \frac{1}{B_2} - \frac{1}{B_1} = \frac{2\pi e}{\hbar^2 \left(\frac{\partial S}{\partial \varepsilon}\right)\omega}$$

or
$$= \frac{e}{m_c^* \omega} \quad \text{(from 8.35)} \tag{8.46}$$

Thus we see that the resonant absorption is periodic in $(1/B)$ with a period of $\left(\frac{1}{B_2} - \frac{1}{B_1}\right)$. With the knowledge of B_1 and B_2 in the relation (8.46), the estimation of the cyclotron effective mass (m_c^*) becomes straightforward.

The study of surface impedance Z in this set-up provides another example of the above periodicity. Considering the microwave source as the voltage source, the surface impedance can be measured at different values of the magnetic field. It is observed that the field derivative of the impedance $\partial Z/\partial \mathbf{B}$ oscillates with the period $\delta(1/B)$. Figure 8.16 shows this periodic effect in copper.

8.7 CLOSED ORBITS AND OPEN ORBITS

Although not stated, the discussions held thus far implied that the energy surface of electrons, or holes or of both are closed. But it is not always so. The possibility of open orbits must be appreciated before the interpretation of the de Hass–van Alphen effect is attempted.

Consider a Fermi surface shown in Fig 8.17(a). The Fermi surface touches the zone boundary. An electron on this plane section starts moving from point D′ towards point A. Its arrival at point A is the same as reaching point A′ which is equivalent to point A in the **k**-space on account of being translationally connnected by a reciprocal lattice vector. Since the same is true for B and B′, C and C′, and D and D′, the electron goes round a closed path—from A′ to B (or B′), from B′ to C (or C′) and finally from C′ to D (or D′).

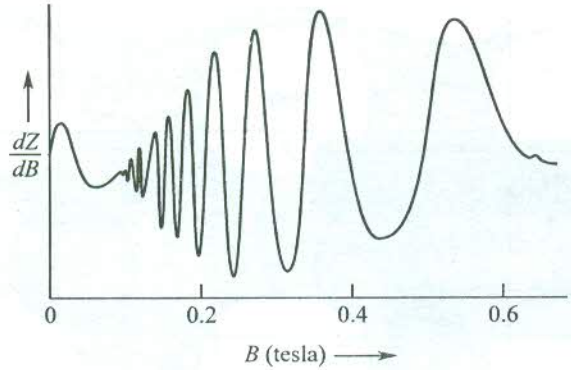


FIG. 8.16 Oscillation of the field derivative of the surface impedance dZ/dB in copper as a function of the magnetic field at 24 GHz. [After A.F. Kip, D.N. Langenberg, T.W. Moore, *Phys. Rev.*, **124**, 359 1961).]

The orbit in the reduced zone scheme [Fig. 8.17(a)] shows discontinuities at the zone boundaries, which is only illusory. The electron has the same wavefunction at points A and A′, since they are equivalent in the **k**-space. The orbit is actually continuous or closed as is clear from the picture in the periodic zone scheme [Fig. 8.17(b)]. Without having to shift the representative point through a reciprocal lattice vector each time the electron crosses a zone boundary, the continuity in path is maintained by providing another cell of the reciprocal lattice on the other side. Since the orbit encloses the empty region of the **k**-space, it is befitting to name it as a *hole orbit*. Its characteristic cyclotron frequency is defined accordingly. A three-dimensional view of a multiply-connected Fermi surface as depicted in Fig. 8.18 shows that on sectional planes both electron orbits and hole orbits can be found.

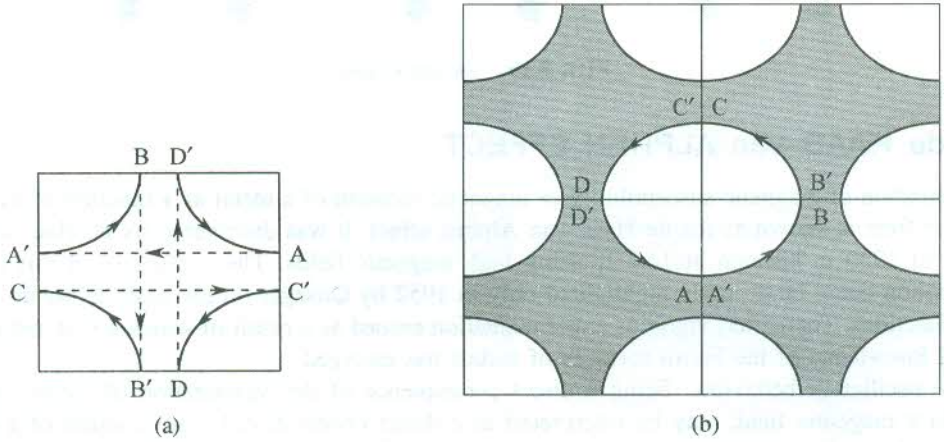


FIG. 8.17 A closed orbit in the (a) reduced zone scheme and (b) periodic zone scheme.

The possibility of an open orbit is exhibited in Fig. 8.19. It is clear that orbits in a plane, whose section with the Fermi surface is not a closed curve, are open. The representative point would not be brought back to the starting point by a magnetic field. In this case the electron continues to move with wavevector increasing towards infinitely large values in the periodic zone scheme.

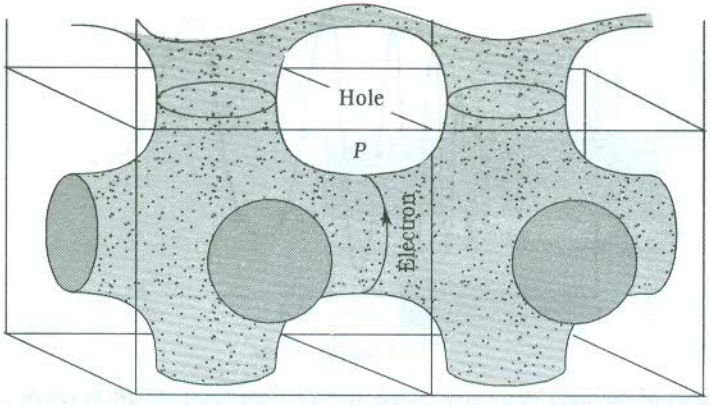


FIG. 8.18 Electron and hole orbits on a multiply-connected Fermi surface.

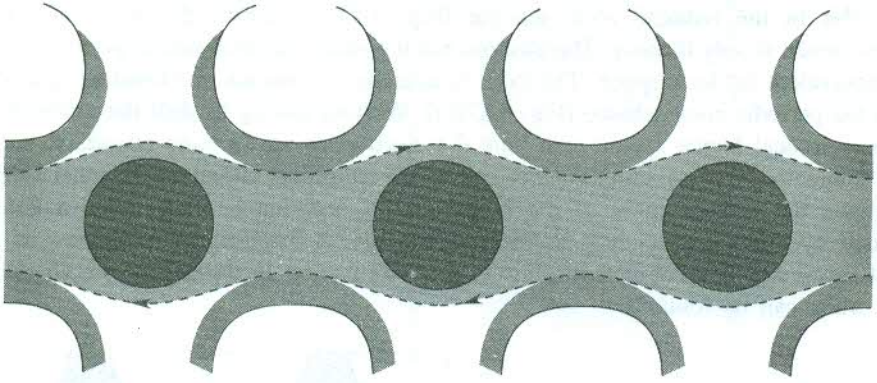


FIG. 8.19 An open orbit.

8.8 de HAAS–van ALPHEN EFFECT

The oscillation of magnetic susceptibility or magnetic moment of a metal as a function of the static magnetic field is known as the de Haas–van Alphen effect. It was discovered by de Haas and van Alphen in 1930 in bismuth at 14.2 K using high magnetic fields. The usefulness of this curious phenomenon could be properly highlighted only in 1952 by Onsagar on the basis of his theory for Bloch functions. Thereafter, vigorous experimentation ensued as a result of which a vast and precise body of knowledge of the Fermi surfaces of metals has emerged.

The oscillatory behaviour, being a direct consequence of the quantization of closed electron orbits in a magentic field, may be interpreted as a direct observational manifestation of a purely quantum phenomenon.

In order to get well-defined oscillations, the experiments are performed with pure crystals at low temperatures where the Fermi surface is sharp and $k_B\, T << \hbar\omega_c$, keeping the electron scattering at minimum. We make use of the Onsager's theory, developed in Section 8.4.2, to explain the dHvA effect. Our arguments, that are strictly valid at 0 K, are tenable as the experiments are conducted actually at very low temperatures.

Let us rewrite the quantization condition (8.38) as

$$\Delta S = \frac{2\pi eB}{\hbar} \tag{8.47}$$

It tells us that as the magnetic field is increased the Landau tubes (see Fig. 8.12) expand. When a Landau tube touches the Fermi surface, the average energy of electrons is maximum because the electrons are drawn up to the Fermi level, the highest occupied level at 0 K. On increasing the field further, the Landau tube moves beyond the Fermi surface and gets depopulated resulting in the decrease of the average energy. The average energy is minimum when the contact contour of the Fermi surface is exactly halfway in the space between two successive Landau tubes, since the highest occupied states fall below the Fermi level. The total average energy of mobile electrons (the electron gas) in this situation is lower than that in the absence of a magnetic field by an amount of the order of $\frac{1}{2}\hbar\omega_c$ per electron at the Fermi level. Thus a slow variation of the magnetic field causes oscillations in the average energy. A quantitative treatment of this oscillatory behaviour has been provided by Ziman.† It involves unusually large and complicated integrals whose evaluation requires the use of certain tricky methods of pure mathematics. The limited scope of the book, obviously, does not allow us to go into the details of this treatment.

It is rather easy to specify the period of oscillation if we appreciate that orbits belonging to different quantum numbers can be manipulated to have a common value of the area by adjusting the magnetic field. The areas of the two successive orbits are:

$$S_n = (n + \gamma)\frac{2\pi eB_1}{\hbar}$$

$$S_{n+1} = (n + 1 + \gamma)\frac{2\pi eB_2}{\hbar} \tag{8.48}$$

If B_1 and B_2 are such that

$$S_n = S_{n+1} = S \tag{8.49}$$

then

$$\left(\frac{1}{B_2} - \frac{1}{B_1}\right) = \frac{2\pi e}{\hbar S}$$

or

$$\delta\left(\frac{1}{B}\right) = \frac{2\pi e}{\hbar S} \tag{8.50}$$

Since $n \propto \frac{1}{B}$ [from (8.48)], $B_2 < B_1$. Thus by decreasing B slowly, we can proceed to orbits of higher quantum numbers. The orbit area changes with variations in B in such a way that for every change of $2\pi e/\hbar S$ in the value of $1/B$, the orbit area is repeated. This is the same as the oscillation

† J.M. Ziman, *Principles of the Theory of Solids,* 2nd ed., Ch. 9 (Cambridge, 1972).

of the orbit area with a period $2\pi e/\hbar S$. These periodic quantum oscillations in principle are expected to be observed in all solid state properties. Similar oscillations of electrical and thermal conductivities in a strong magnatic field are some examples to quote. The phenomenon involving the electrical conductivity is known as de Haas–Shubnikov effect.

The oscillation of magnetic moment in dHvA effect follows from the oscillatory behaviour of the free energy as discussed above. Being a magnetic field derivative of the free energy, the magnetic moment too oscillates as a function of the static magnetic field. The orbit area S in (8.50) is essentially a measure of the cross-sectional area of the Fermi surface. Periods of the sections of a Fermi surface of general shape at different values of the wavevector along **B** are different. The contributions from all sections add up and account for the net response. But, the dominant contribution to $\delta(1/B)$ in (8.50) comes from orbits whose time periods remain unaffected by small changes in the magnitude of **k** along **B**. These orbits are known as *extremal orbits*. The extremal orbits on a general Fermi surface for two orientations of **B** are shown in Fig. 8.20. If the magnetic field be applied along the z-direction, the extremal areas in (8.50) would be represented by those cross-sectional areas at whose contours on the Fermi surface the condition,

$$\frac{dS}{dk_z} = 0 \tag{8.51}$$

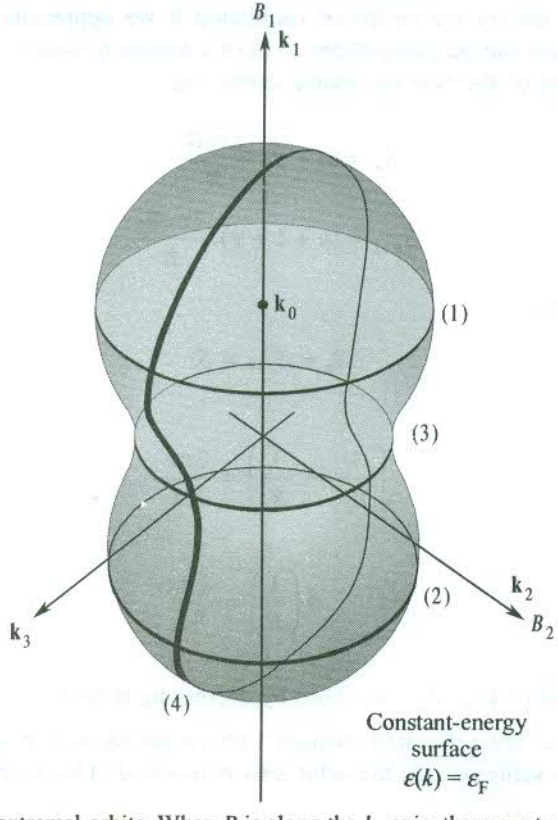


FIG. 8.20 Illustration of extremal orbits. When B is along the k_1-axis, there are two maximum extremal orbits (1) and (2), and one minimum extremal orbit (3). For B along the k_2-axis, only one extremal orbit (4) is observed.

is satisfied. The condition may be met for several values of k_z (at different heights of the Fermi surface along **B**). It should be abundantly clear (see Fig. 8.20) that maximum and minimum cross-sections both refer to the extremal orbits.

The oscillations occur even when only the magnetic field orientation is varied and the strength kept constant. It is so because for some directions the Landau tubes touch the Fermi surface along the contact contour while for other directions they do not. The oscillations disappear for field directions, giving open orbits. The complete theoretical structure of dHvA effect is based on the existence of closed orbits with definite periods. Thus by changing the magnetic field direction all the extremal cross-sections of the Fermi surface can be mapped out, making the reconstruction of the Fermi surface feasible.

Experimental set-up

In a typical set-up, a few mm long single crystal wire is mounted in a small pick-up coil. There are two basic methods of measurement. In one method the magnetic field is produced by discharging a large capacitor through a solenoid cooled to liquid nitrogen temperature. Within a few milliseconds the field increases to about 10 T and drops to zero. This method is known as the impulsive method. With changing magnetic field, an emf is induced which is proportional to

$$\frac{dM}{dt} = \frac{dM}{dB} \cdot \frac{dB}{dt} \tag{8.52}$$

where M is the magnetic moment.

Since metals are only weakly diamagnetic, the difference between the applied field and the average field experienced by electrons in a metal may be ignored. This point can be properly appreciated after having lessons on magnetism (Chapters 13 and 14).

In accordance with (8.52) the induced emf produces an oscillatory signal which is proportional to the differential susceptibility dM/dB.

The second basic method employs superconducting coils† that provide highly stable steady fields of the order of 5 T. Once the solenoids get energized, a persistent current is set up with no external source of emf (because of zero resistance). The steady field is modulated at a convenient frequency using auxiliary coils. With the modulation being done at a fixed frequency, we have the advantage of using phase sensitive detection techniques for improving the signal to noise ratio.

Example

The original spectrum†† showing the oscillation of magnetic susceptibility of gold is reproduced in Fig. 8.21. The picture of the Fermi surface, as constructed by making use of the information derived from this measurement, is shown in Fig. 8.22. The d-states in gold are occupied and the Fermi level is placed in the region of the sp-band. So, the Fermi surface appears approximately spherical. The picture shows that the surface in one Brillouin zone is connected to that in the next by a neck along the <111> direction showing distortions along this direction. Therefore, with a magnetic field along

† The superconducting property is observed at low temperatures. This implies that coils used here are dipped in a low temperature both. This property will be discussed in Chapter 15. The d.c. resistance of a superconductor is zero.

†† B. Lengeler, *Springer Tracts Mod. Phys.*, **82**, 1, p. 1 (Springer, 1978).

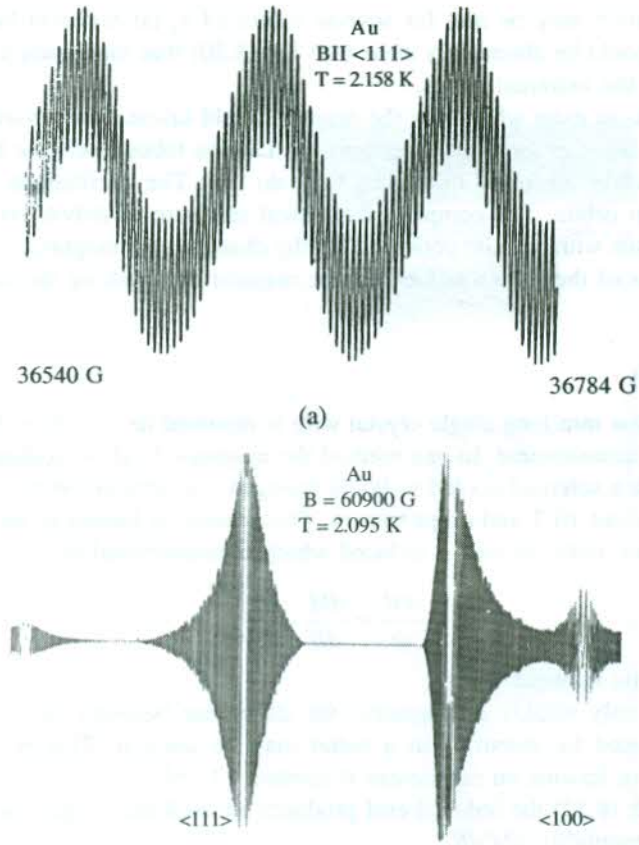


FIG. 8.21 The oscillation of the magnetic susceptibility of gold: (a) as a function of the magnetic field, (b) with variation in the direction of the magnetic field. These quantum oscillations can be observed only for very sharp Fermi distribution ($k_B T < \hbar\omega_c$).

the <111> direction there are two types of closed orbits along which the Landau tubes peel off from the Fermi surface on increasing the magnetic field slowly. These are identified as neck and belly orbits. By comparing the periods (1:29) in the oscillogram [Fig. 8.21(a)], we directly obtain the ratio of the cross-sectional areas of these orbits. The absolute magnitudes are calculated from the data by making use of (8.50). These calculations yield the values of areas as 1.5×10^{19} m^{-2} and 4.3×10^{20} m^{-2} for the neck and belly orbits, respectively.

It is useful to make an estimate of the Fermi surface of gold by approximating its surface to the free electron spherical Fermi surface. The value of the Fermi wavevector $\mathbf{k}_F$ is taken from tables. Using $k_F = 1.2 \times 10^{10}$ m^{-1}, we get an extremal area of 4.5×10^{20} m^{-2}. Interestingly, this is in excellent agreement with the experimental value for the extremal belly orbit that represents the average shape and size of the Fermi surface of gold.

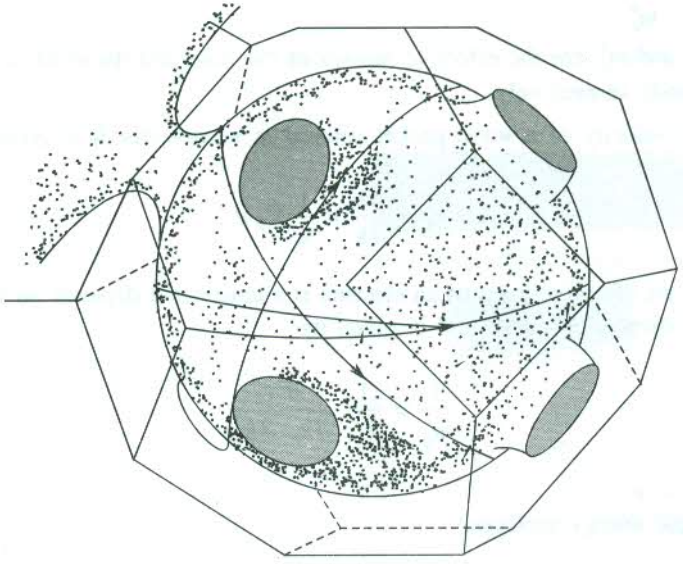


FIG. 8.22 Model of the Fermi surface of noble metals Cu, Ag and Au. The surface makes contact with the Brillouin zone boundary along the [111] directions. There are two extremal orbits perpendicular to the [111] directions.

SUMMARY

1. An empty electron state in an otherwise filled band of a crystal is identified with a hole. If there is only one hole in a band, the properties of the hole are those of $(N - 1)$ electrons, where N is the number of electrons in the completely filled band.

2. Properties of holes:
   (i) Current density contributed by holes is

$$j = -e \int_{\text{occupied}} \frac{v(\mathbf{k})\, d^3\mathbf{k}}{4\pi^3} = e \int_{\text{unoccupied}} \frac{v(\mathbf{k})\, d^3\mathbf{k}}{4\pi^3}$$

   (ii) If an electron with wavevector $\mathbf{k}_e$ is missing from a band, the hole thus created will have a wavevector $\mathbf{k}_h$ such that

$$\mathbf{k}_h = -\mathbf{k}_e$$

   (iii) Taking the energy of a filled band as zero of the energy scale, the energy of the hole is positive:

$$\varepsilon_h(\mathbf{k}_h) = -\varepsilon_e(\mathbf{k}_e)$$

   (iv) The velocity of the hole created in the valence band owing to the excitation of an electron from the valence band to the conduction band is equal in magnitude and direction to the velocity of the electron now missing from the valence band.

$$\mathbf{v}_h(\mathbf{k}_h) = \mathbf{v}_e(\mathbf{k}_e)$$

(v) $m_h^* = -m_e^*$

where m_h^* and m_e^* are the effective masses of the hole and the electron (missing from the valence band), respectively.

3. The group velocity of a wave packet centred at wavevector **k** is given by

$$\mathbf{v}_g = \frac{\partial \omega}{\partial \mathbf{k}} = \frac{1}{\hbar} \nabla_{\mathbf{k}} \varepsilon$$

4. Generally, the effective mass of an electron is a tensor that depends on the curvature of the ε versus k dispersion curve and is given by

$$\frac{1}{m_{ij}^*} = \frac{1}{\hbar^2} \frac{\partial^2 \varepsilon}{\partial k_i \, \partial k_j}$$

with $i, j = x, y, z$.
For isotropic energy surfaces,

$$\frac{1}{m^*} = \frac{1}{\hbar^2} \frac{\partial^2 \varepsilon}{\partial k^2}$$

5. The constant energy surface in the **k**-space that separates the filled states from the empty states at absolute zero is called the Fermi surface. The total volume enclosed by the Fermi surface depends only on the electron concentration and is independent of the details of the lattice interaction.

6. The total electron energy in the presence of a magnetic field applied along the z-direction is written as

$$\varepsilon_n(k_z) = \frac{\hbar^2 k_z^2}{2m} + \left(n + \frac{1}{2}\right) \hbar \omega_c$$

where ω_c is the cyclotron frequency (eB_z/m^*).

Each group of degenerate energy levels that corresponds to a magnetic level as defined by a given pair of quantum numbers n and k_z is called a *Landau level*.

7. The basic condition for the cyclotron resonance is

$$\omega_c \tau >> 1.$$

8. The de Haas–van Alphen effect shows periodicity of $\delta\left(\frac{1}{B}\right)$ given by

$$\delta\left(\frac{1}{B}\right) = \frac{2\pi e}{\hbar S}$$

where S is the extremal cross-sectional area of the Fermi surface in the **k**-space.
The study of the dHvA effect thus provides **S** that is considered as perpendicular to **B**.

PROBLEMS

8.1 (a) Show that the group velocity is twice the phase velocity for any state of an ideal free electron gas.

(b) Assuming that electron velocities in a system are represented by spherical distribution and ε–k relationship as, $\varepsilon = ak^2 - bk^4$, find the group and phase velocities in the system.

What is the value of energy in terms of a and b at which the two types of velocities are equal?

8.2 Construct the first two Brillouin zones of a primitive rectangular two-dimensional lattice with axes a, b (= $2a$).

If the lattice belongs to a metal in which there is one atom of valency one in a primitive cell, calculate the radius of the free electron Fermi sphere in cm^{-1}. Take a = 2 Å, b = 4 Å. Draw the Fermi sphere to scale on a drawing of the first Brillouin zone.

8.3 In potassium at 68 GHz three consecutive cyclotron resonances are observed at magnetic fields of 0.74 T, 0.59 T and 0.49 T. What is the cyclotron resonance mass of electrons in potassium?

8.4 In a cyclotron resonance set-up, a klystron radiation of 2.4×10^{10} Hz is used. For a sample, the resonance absorption occurs at a magnetic field of 8.6×10^{-2} T. Calculate the effective mass of the respective charge carriers. Determine the range of relaxation time over which a resonance is observable.

8.5 (a) Apply the Onsager's quantization condition (8.39) to the orbits of free electron levels and show that it leads directly to the free electron levels (8.26) if $\gamma = 1/2$.

(b) Derive an expression for the degeneracy of Landau levels in the case of a two-dimensional free electron gas. Show that the degeneracy is equal to the number of zero-field free electron levels with a given k_z, and with k_x and k_y within a planar region of area ΔS given by (8.29).

8.6 An electron in sodium at the Fermi level moves initially in the xy-plane. Calculate the radius of its cyclotron orbit when a magnetic field of 1 T is applied. What is the ratio of the orbit area in real space to that in **k**-space?

8.7 The expression for the Fermi energy in a non-isotropic solid is

$$\varepsilon = (\alpha k_x^2 + \beta k_y^2)$$

Describe the rate at which **k** changes during an extremal cyclotron orbit at the magnetic field $\mathbf{B}_z$. Assume that α and β are unequal. How is the orbit in real space?

8.8 Make an estimate of the expected value of period $\delta(1/B)$ in a dHvA effect study of potassium on the free electron model. What is the order of the maximum temperature below which the experiment should be done to obtain good results in a magnetic field of strength 1 T?

SUGGESTED FURTHER READING

Falikov, L.M., *Fermi Surface Studies: Electrons in Crystalline Solids* (IAEA, Vienna, 1973).

Harrison, W.A., *Electronic Structure and the Properties of Solids* (Freeman, 1980).

Shoenberg, D., *Magnetic Oscillations in Metals* (Cambridge, 1984).

Ziman, J.M., *Principles of the Theory of Solids*, 2nd ed. (Cambridge, 1972).

Chapter 9

Semiconductors

As mentioned in the beginning of Chapter 7, the values of electrical conductivity of solids are spread over almost the widest range for any common physical property. Hence these form a basis for classifying solids. Solids characterized by extremely high and extremely low values of electrical conductivity are identified as metals and insulators, respectively. A pure metal at 1 K may have a conductivity of the order of 10^8 ohm^{-1} m^{-1} against a low of 10^{-20} ohm^{-1} m^{-1} for an extreme insulator. Materials with conductivity values intermediate to these extreme orders of magnitude are called *semiconductors*. Typical conductivity values for semiconductors lie in the range from 10^{-7} to 1 ohm^{-1} m^{-1}. The most useful feature of semiconductors is that their electrical conductivity generally decreases with increasing purification in contrast to metals where the conductivity always increases with increasing purification.

Notwithstanding the enormous technological importance of semiconductors, their study is even more crucial to the understanding of electronic properties of solids. We will see that it is possible to apply the Maxwell–Boltzmann statistics to deal with charge carriers in semiconductors. This gives exact analytical solutions of many problems that can be solved only by approximate or numerical methods in metals where the Fermi–Dirac statistics has to be used. Furthermore, with remarkable progress having been made in the technology of growing semiconductor crystals, the degree of purity and perfection achieved in growing single crystals of semiconductors is much higher than that in metals and insulators. It is a matter of absolute importance to the study of electronic properties, some aspects of which are obscured by effects arising because of the presence of impurities and crystalline imperfections. It simply amounts to say that the quantitative studies of some phenomena that are generally difficult or cannot be made accurately in metals, are easily carried out in semiconductors with the required precision. Thus the study of semiconductors is helpful in interpreting the electronic properties of solids in general. In this chapter we concentrate on shaping the basic theoretical ideas in the framework of the band theory of solids. Nevertheless, a brief account of the structure and the principle of working of some simple and popular devices is included in Section 9.10.

9.1 CLASSIFICATION OF SEMICONDUCTORS

The classification of solids made on the basis of their band structures gives a lead to exploiting the band theory for the interpretation of many solid state properties. We learnt in Chapter 7 that completely filled and completely empty bands do not contribute to the flow of current. The highest filled band (the valence band) in metals is only partially filled and the flow of current occurs on account of almost continuous excitation of valence band electrons to empty states of the band under the influence of an electric field. Therefore, a material that has only completely filled and completely empty bands behaves as a perfect insulator at absolute zero when there can hardly be found any electrons in the lowest empty band (the conduction band) as a result of the thermal excitation of

electrons in the valence band. But if the difference between the upper edge of the valence band and the lower edge of the conduction band is not large and less than 2 eV, there is a finite probability for a small fraction of electrons occupying the uppermost states of the valence band to be thermally excited to the conduction band at moderate and high temperatures. At these temperatures the width of the energy range over which the Fermi distribution function rapidly changes is relatively substantial. This enables the consequences of the change in the distribution function easily observable. A representative band scheme for metals, semiconductors and insulators is drawn in Fig. 9.1. The figure portrays a qualitative difference in the band structures of these solids.

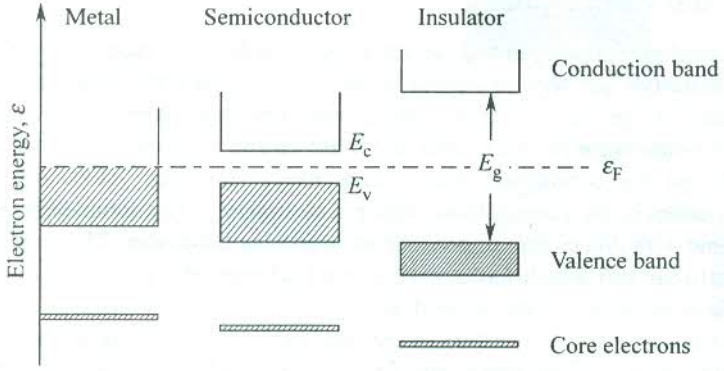


FIG. 9.1 Comparative energy band schemes for metals, semiconductors and insulators at $T = 0$ K. The valence band is completely filled in semiconductors and insulators but only partially filled (shaded) in metals. A relatively much smaller band gap E_g in semiconductors distinguishes them from insulators.

There is a long list of semiconductor materials. Only a few of them are elements. Si, Ge, grey Sn and grey Se are some examples. The first three are in Group IV and Se is in Group VI of the periodic table. Rest of them are mostly binary compounds, mainly of two types. In one type (e.g. GaAs, InSb, GaP) one element from Group III (e.g. B, Al, Ga, In) is combined with an element from Group V (e.g. N, P, As, Sb). The other type of binary compounds are formed with one element from Group II (e.g. Zn, Cd, Pb) and the other from Group VI (e.g. S, Se, Te). Some important examples of this type are ZnS, CdSe, PbTe. SiC and SiGe are the only known binary semiconductors whose both components are from Group IV. Some oxides also show semiconducting behaviour—TiO_2, Cu_2O and ZnO are prominent examples of this class. A list of technologically important semiconductor materials is given in Table 9.1 which also contains the relevant data on these materials.

Semiconductors are primarily of two types—*intrinsic* and *extrinsic*. The intrinsic semiconductors are usually pure monatomic or diatomic solids. An intrinsic material is converted into the extrinsic type (or the impurity type) by adding traces (~ 1 part in one million) of a suitable impurity with the aim to enhance the level of the charge carrier density. Values of electrical conductivity of intrinsic semiconductors lie far below the range that is useful for the purpose of applications. But semiconductors are gifted with the unique quality that their electrical conductivity can be increased by several orders of magnitude by mixing with them suitable impurities in small concentration. This has enormously increased their technological importance. Thus it is mostly the extrinsic semiconductors that form the basis for semiconductor devices. The details of formation and model of extrinsic semiconductors are given in Section 9.4.

9.2 EXAMPLES OF BAND STRUCTURE

We know from Chapter 7 that the band structure of a solid is closely related with its crystal structure. Not many crystalline structures are favourable to the semiconducting behaviour. Most of the thoroughly investigated semiconductors have a diamond type lattice. We take examples of Si, Ge and GaAs which are distinguished for their large number of applications. The Si and Ge crystals have diamond structure. The crystal lattice of GaAs is zinc blende type which is only a modified diamond lattice, as discussed in Chapter 1. We describe below the band structures of these crystals.

9.2.1 Silicon and Germanium

The outer electron configurations of Si and Ge are $3s^2 3p^2$ and $4s^2 4p^2$, respectively. The origin of band structure of these materials has been discussed in Section 7.7. The tetrahedral bonding orbitals (sp^3) are formed because of the mixing of s- and p-wavefunctions. Near the bonding distance, at equilibrium, these orbitals were shown to split into bonding and antibonding orbitals which constitute the valence band and the conduction band, respectively (see Fig. 7.11). All the four s- and p-electrons occupy states in the valence band, filling it completely. The completely empty conduction band should combine with this picture to produce an insulating behaviour. This is really the case with the diamond crystal (carbon) which has a similar band scheme. But on account of small band gap (E_g), Si and Ge show semiconducting properties.

Figure 7.11 reveals an important feature of the band gap regarding its dependence on temperature. The observation that the size of the energy gap (or the splitting) between the valence and conduction bands decreases with increase in the interatomic separation, indicates that the gap is smaller at higher temperatures where the interatomic separation becomes larger because of thermal expansion. This fact is confirmed by the measured values of the band gap E_g at different temperatures (see Table 9.1).

Band structures are calculated by fitting the measured physical quantities, such as the band gap, the positions of points of high symmetry in the Brillouin zone (critical points), and the curvature of energy surfaces (the effective mass). The calculated band structures of Si and Ge are shown in Fig. 9.2. The features of the two band schemes appear quite different in contrast to the qualitative similarity as expected on the basis of Fig. 7.11. This is obviously the effect of the difference in electron wavefunctions associated with $3s^2 3p^2$ and $4s^2 4p^2$ configurations.

The symbols, Γ, X and L stand for positions of certain points of high symmetry in the Brillouin zone. They refer to the points at the zone centre (000), $\frac{2\pi}{a}$ (100) and $\frac{2\pi}{a}\left(\frac{1}{2}\frac{1}{2}\frac{1}{2}\right)$, respectively, where '$a$' is the lattice constant. The valence band maximum occurs at the zone centre, i.e. $k = 0$ for both Si and Ge. But the conduction band minimum occurs for k along the [100] direction in Si and for k along the [111] direction in Ge. This amounts to saying that electrons of the lowest energy in the conduction band have their wavevectors oriented along the [100] direction in Si and along the [111] direction in Ge. In both materials the valence band maximum and the conduction band minimum thus occur at different values of k. Semiconductors having this type of band structure are called the *indirect gap* semiconductors and those for which the maximum and minimum in question fall at the same value of k are referred to as the *direct gap* semiconductors. The discussion on consequences of the indirect gap nature of band structure will be taken up in Section 9.2.3.

The first Brillouin zone of Si and Ge crystals is a truncated octahedron appropriate to the FCC symmetry of their unit cell. In parabolic approximation (i.e. retaining terms up to the order k^2 in the expression for ε_k), the surfaces of constant energy are ellipsoids as confirmed by cyclotron resonance studies (see Section 8.6.1).

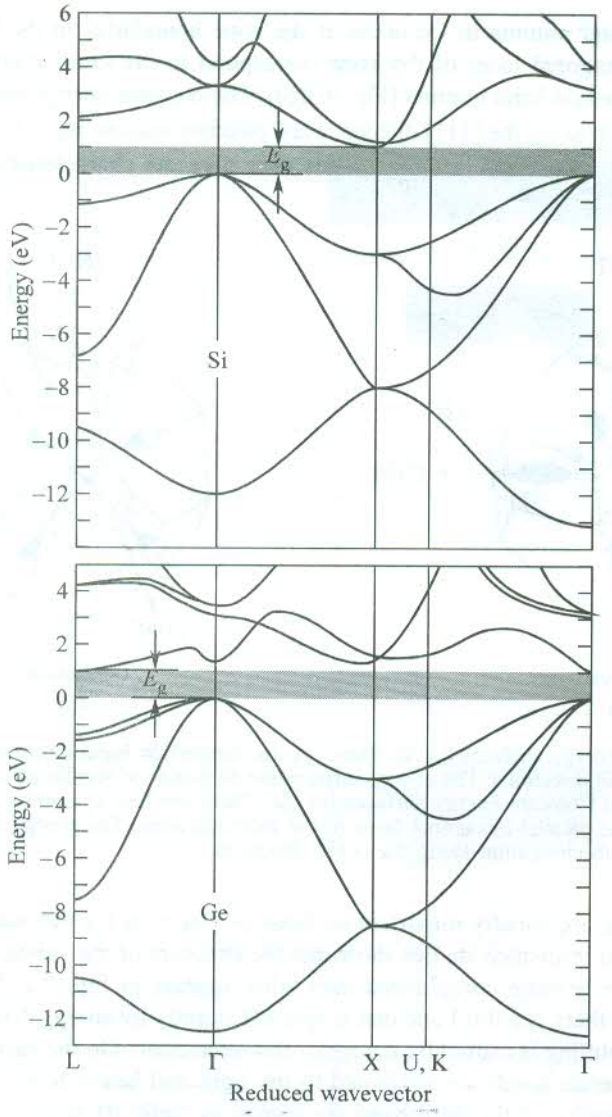


FIG. 9.2 Calculated band structures of Si and Ge [After J.R. Chelikowsky, M.L. Cohen, *Phys. Rev.* **B14**, 556 (1976)]. The structures confirm the indirect gap nature for both Si and Ge. For Ge, the spin-orbit splitting is also considered.

In Si there are six symmetry-related minima of the conduction band at points in the [100] directions. Each of the six ellipsoids is an ellipsoid of revolution about a cubic axis by symmetry. They appear as cigars elongated along the cube axes as shown in Fig. 9.3(a). The electron has two effective masses—the longitudinal m_l^* (along the axis) and the transverse m_t^* (perpendicular to the axis). Their values in terms of the free electron mass m are given as—$m_l^* = 0.98m$ and $m_t^* = 0.19m$. The valence band shows two degenerate maxima both located at $k = 0$ with spherical symmetry within validity of the ellipsoidal expansion (Fig. 9.2). The two effective masses are $0.49m$ and $0.16m$.

The conduction band minima in Ge occur at the zone boundaries in the [111] directions. The minima on parallel hexagonal faces of the zone correspond to the same energy levels giving four symmetry-related conduction band minima [Fig. 9.3(b)]. The constant energy surfaces are ellipsoids of revolution with elongation along the [111] directions and effective masses, $m_l^* = 1.57m$ and $m_t^* = 0.082m$. The two degenerate valence band maxima in this case give the characteristic effective masses of $0.28m$ and $0.44m$.

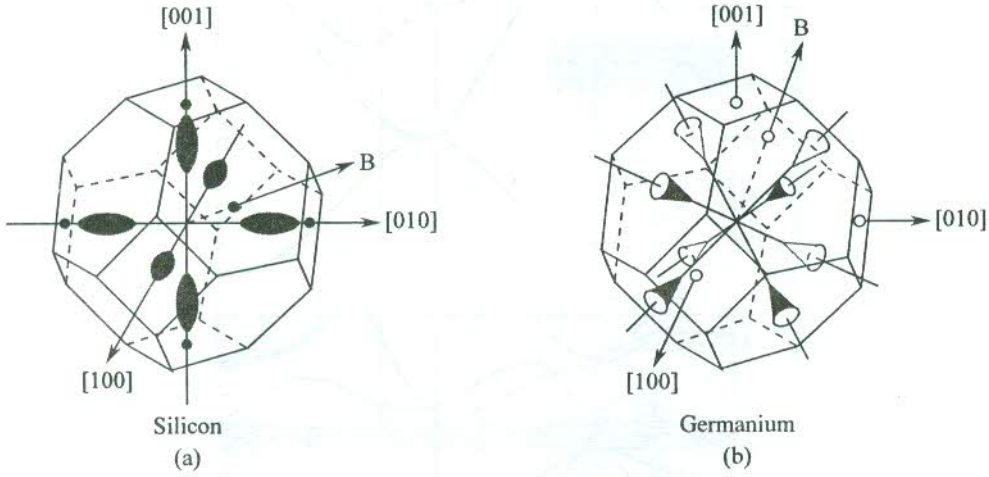


FIG. 9.3 (a) Constant energy surfaces for Si: There are six conduction band minima of cigar shape at points along the [100] directions. The energy surfaces are ellipsoids of revolution with elongation along the cube axes. (b) Constant energy surfaces for Ge: There are four symmetry related conduction band minima on the parallel hexagonal faces of the Brillouin zone. The energy surfaces are ellipsoids of revolution with elongation along the [111] directions.

Properties of holes are equally important as those of electrons for the study of semiconductors. Details of the cyclotron resonance studies show that the structure of the valence band maximum near $\Gamma(k = 0)$ in Si and Ge is more complicated than what appears in Fig. 9.2. In addition to the two degenerate bands at Γ, there is a third one that is split off slightly by energy Δ towards a lower energy band (Fig. 9.4). The splitting is caused by the spin-orbit interaction. On the basis of the effective mass values, the two degenerate bands are attributed to the light and heavy holes (Fig. 8.14). The holes associated with the identity of the third band are named as 'split-off holes'. The split-off energy Δ has been estimated at 0.044 eV in Si and 0.29 eV in Ge.

9.2.2 Gallium Arsenide

GaAs crystal enjoys a special status, firstly on account of its direct-gap and secondly because of its band-gap energy being closely below the energy range of visible radiation. These properties render it most suitable for the fabrication of efficient optical devices.

The crystal has zinc blende (ZnS) structure. It is an example of the mixed ionic and covalent bonding. The chemical bonding is interpreted as the superposition of these two extreme cases of bonding. In ionic bonding, an electron is transferred from Ga to As to give the ionic structure $Ga^+ As^-$. On the other hand in the second extreme case, with the displacement of an electron from As to Ga, the number of electrons in the outer shell of both Ga and As atoms becomes four which results in the

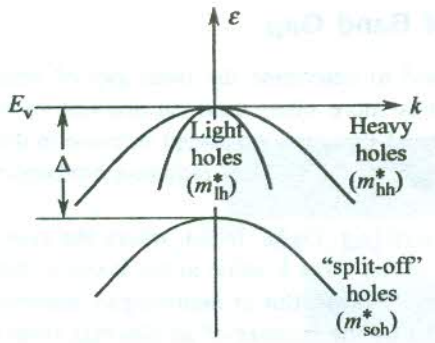


FIG. 9.4 Structure of the valence band (qualitative) in Si or Ge near the top showing the hole bands with the spin-orbit interaction taken into account. The split-off hole band is separated from the top at $k = 0$ by Δ, contributed by the spin–orbit interaction.

sp^3 hybridization as in the case of Si and Ge. The observed tetrahedrally coordinated ZnS structure of GaAs serves as a certain proof to the effect that the effects of covalent bonding dominate.

The band structure of GaAs is shown in Fig. 9.5. All the valence band maxima and conduction band minima occur at $\Gamma(k = 0)$ showing its direct gap nature with a gap of 1.43 eV at 300 K. The constant energy surfaces are accordingly spherical. The conduction band effective mass m_e^* is $0.07m$. There are three distinct valence bands similar in form to those for Si and Ge at Γ. The three respective effective masses are given as $m^*_{lh} = 0.12m$, $m^*_{hh} = 0.68m$ and $m^*_{soh} = 0.2m$ with $\Delta = 0.34$ eV.

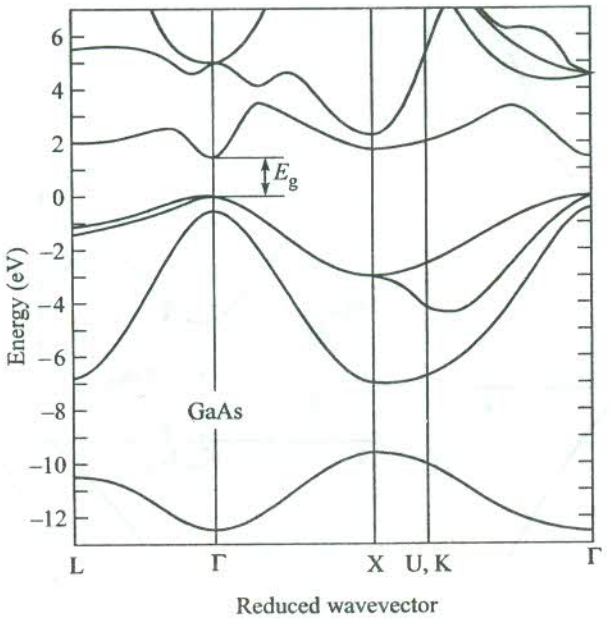


FIG. 9.5 Calculated band structure of GaAs, a representative of III–V semiconductors. The conduction band minima and the valence band maxima occur at $k = 0$, showing the direct gap behaviour. [After J.R. Chelikowsky, M.L. Cohen, *Phys. Rev.*, **B14**, 556 (1976).]

9.2.3 Determination of Band Gap

A number of methods are used to determine the band gap of semiconductors. The technique of continuous optical absorption is more often used on account of its accuracy and the important information gained about the band structure. An abrupt increase in the absorption of optical radiation takes place as soon as the energy of the incident radiation $\hbar\omega$ becomes large enough to exceed the energy gap.

In direct-gap semiconductors (e.g. GaAs, InSb), where the conduction band minimum and the valence band maximum occur at the same **k**-value in the **k**-space, the optical threshold at $\omega(= E_g/\hbar)$ directly gives the band gap [Fig. 9.6(a)]. But in indirect-gap materials (e.g. Si, Ge, GaP), the direct photon absorption accompanied by the transfer of an electron from the top of the valence band to the bottom of the conduction band would not conserve the crystal momentum because the initial and final points of the transition in the **k**-space do not have the same k-value. Hence such a direct transition is not allowed. The transition process will have to be indirect in which the absorption of an optical photon must be accompanied by some other process with whose involvement the condition of momentum conservation may be satisfied. The intensity of a continuous absorption spectrum gets sufficient contribution from phonons. With the involvement of a phonon in the indirect transition under discussion, the sum of wavevectors before the transition becomes equal to their sum after the transition, showing the momentum conservation. There can be two possibilities—one in which a phonon is emitted (created) after the transition and the other in which a phonon is absorbed (destroyed) along with the optical photon to materialize the transition.

Let us denote the wavevector of the optical photon by $\mathbf{K}_{op}$ and that of the phonon by $\mathbf{K}_{ph}$. If

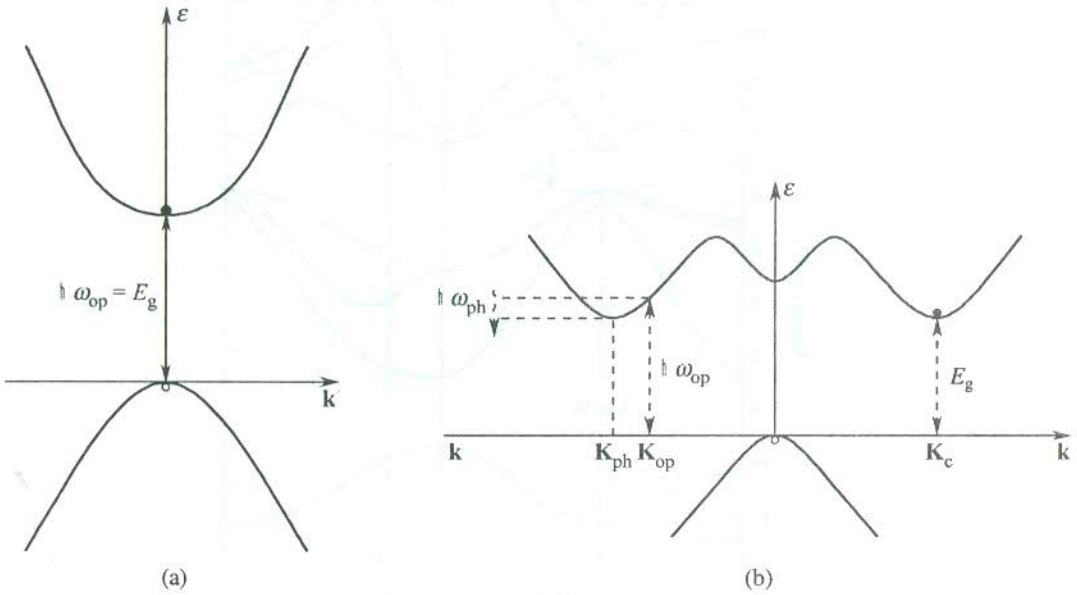


FIG. 9.6 (a) Absorption of a photon in a direct-gap semiconductor, where an electron from the valence band is lifted into the conduction band and the crystal momentum is conserved without the involvement of any other process. (b) The photon absorption in an indirect-gap material where it is shown that a phonon is emitted on the photon absorption to satisfy the momentum conservation. The $\mathbf{K}_{op}$ and $\mathbf{K}_{ph}$, respectively, denote the wavevectors of the absorbed optical photon and the emitted phonon.

the wavevector at the conduction band minimum as measured from the valence band maximum be represented by $\mathbf{K}_c$, then in the case of phonon emission [Fig. 9.6(b)], we have

$$\underset{\text{(before the transition)}}{\mathbf{K}_{op}} = \underset{\text{(after the transition)}}{\mathbf{K}_c + \mathbf{K}_{ph}} \tag{9.1}$$

But the photon wavevectors $\mathbf{K}_{op}$ are negligibly small in the range of energies concerned. Therefore, from (9.1),

$$\mathbf{K}_c \simeq -\mathbf{K}_{ph} \tag{9.2}$$

Relation (9.2) states that the increase of crystal momentum by $\hbar\mathbf{K}_c$ during the transition is offset by an equal amount owing to the emission of a phonon with wavevector $(-\mathbf{K}_c)$. Since the phonon takes away a part of the energy of the incident photon (though very small), the optical threshold energy is greater than E_g [see Fig. 9.6(b)]. In this case,

$$\hbar\omega_{op} = E_g + \hbar\omega_{ph} \tag{9.3}$$

At high temperatures a good number of phonons are present in the crystal. With the absorption of a phonon during the transition, the momentum conservation is given by

$$\underset{\text{(before the transition)}}{\mathbf{K}_{op} + \mathbf{K}_{ph}} = \underset{\text{(after the transition)}}{\mathbf{K}_c} \tag{9.4}$$

This gives the optical threshold as

$$\hbar\omega_{op} = E_g - \hbar\omega_{ph} \tag{9.5}$$

So, the optical threshold is lower than E_g. The change in optical threshold in these processes is generally of little consequence because the phonon energies are characteristically small (~ a few hundredths of an eV) as compared to the size of the band gap. This change, however, is of considerable value in semiconductors having a small band gap.

Since the indirect transitions generate heat when phonons are created, the direct-gap materials are preferred in order to have efficient devices. A mixture of the two types of materials is a desired manipulation in some devices such as injection lasers where it is practised on account of the demand for higher power.

Another method commonly used for determining the band gap is based on the temperature-dependent study of the electrical conductivity σ. We will see later in this chapter that electrical conductivity of intrinsic semiconductors depends on temperature according to the following proportionality:

$$\sigma \propto \exp(-E_g/2k_BT) \tag{9.6}$$

An approximate value of E_g is given by the $\ln(\sigma)$ versus $1/T$ graph. The gap is also determined from the intrinsic carrier concentrations derived from the experimental values of Hall coefficient. The values of band gap and nature of gap for a number of important semiconductors are given in Table 9.1.

9.3 INTRINSIC CARRIER DENSITIES

The difference between the electrical conductivity of semiconductors and metals in the form of

Table 9.1 Data on the band gap of some important semiconductors (D = direct gap, I = indirect gap)

Crystal	*Type of gap*	E_g (eV)	
		0 K	300 K
Si	I	1.17	1.11
Ge	I	0.75	0.67
grey Sn	D	0.00	0.00
GaAs	D	1.52	1.43
GaSb	D	0.81	0.68
GaP	I	2.32	2.25
InSb	D	0.23	0.17
InAs	D	0.43	0.36
InP	D	1.42	1.27
AlSb	I	1.65	1.6
PbS	D	0.29	0.34–0.37
PbSe	I	0.17	0.27
PbTe	I	0.19	0.29
CdS	D	2.58	2.42
CdSe	D	1.84	1.74
CdTe	D	1.61	1.44
ZnO	—	3.44	3.2
ZnS	—	3.91	3.6
SnSe	D	0.3	0.18
TiO_2	—	3.03	—
Cu_2O	D	2.17	—

former's strong dependence on temperature indicates that thermal excitations control the conductivity of semiconductors in a big way. The thermal excitation of an electron from the valence band across the forbidden energy gap E_g to the conduction band creates a hole in the valence band. The number of these carriers that contribute to the flow of electric current increases more and more with a continued thermal excitation. In this section we calculate the density of these carriers in the state of thermal equilibrium in an intrinsic semiconductor considered as highly pure such that the contribution to the carrier concentrations from impurities may be neglected. Calculations based on appropriate statistics show that the electron and hole concentrations equal in the present case, are strongly temperature dependent which thus accounts for the conductivity behaviour. On the basis of what we learnt in Chapter 6, the conductivity of a semiconductor may be expressed as

$$\sigma = |e|(n\mu_n + p\mu_p) \tag{9.7}$$

where n and p are electron and hole densities (per unit volume), respectively, and μ_n and μ_p represent the corresponding mobilities. The contributions from electrons and holes in (9.7) simply add up because of the opposite sign of their charge and opposite directions of their drift velocities.

It is easy to appreciate that in excitations of our interest the carriers near the band edges or the parabolic part of the valence and conduction bands are involved. In this region the effective mass can be treated as constant in the first approximation. This takes care of the neglect of energy dependence of mobility in (9.7). For calculations in semiconductors the chemical potential μ appearing

in the Fermi distribution function (6.30) is replaced by the Fermi energy ε_F. But the level of chemical potential always lies in the region of the forbidden energy gap where no single-electron energy levels exist. The very definition of the Fermi level which is considered as the highest occupied level at absolute zero, thus becomes redundant since no single-electron energy level is available to coincide with the Fermi level. Therefore, in the fitness of things the Fermi level in semiconductors should be interpreted as a synonym to the chemical potential.

The occupancy of energy levels in semiconductors must be described by the Fermi–Dirac distribution function $f(\varepsilon, T)$ as in other solids:

$$f(\varepsilon, T) = \frac{1}{1 + \exp[(\varepsilon - \varepsilon_F)/k_B T]} \tag{9.8}$$

If $D_c(\varepsilon)$ and $D_v(\varepsilon)$ denote the density of states in the conduction and valence band, respectively, the charge carrier densities are usually written as

$$n = \int_{E_c}^{\infty} D_c(\varepsilon) f_e(\varepsilon, T)\, d\varepsilon \tag{9.9}$$

with f_e as the electron occupancy, given by (9.8) and

$$p = \int_{-\infty}^{E_v} D_v(\varepsilon) f_h(\varepsilon, T)\, d\varepsilon = \int_{-\infty}^{E_v} D_v(\varepsilon)\,[1 - f_e(\varepsilon, T)]\, d\varepsilon \tag{9.10}$$

with f_h as the hole occupancy.

Here E_c and E_v refer to the energy values at the edges of the conduction and valence band, respectively (Fig. 9.7). The electron energy as measured from the bottom (or edge) of the conduction band is $(\varepsilon - E_c)$, where ε is the absolute value of electron energy. Similarly, the hole energy when measured from the top (or edge) of the valence band is equal to $(E_v - \varepsilon)$. Applying the parabolic approximation, i.e. assuming the effective mass to remain constant, the two density of states are written using (6.41) as

$$D_c(\varepsilon) = \frac{1}{2\pi^2}\left(\frac{2 m_e^*}{\hbar^2}\right)^{3/2} (\varepsilon - E_c)^{1/2} \tag{9.11}$$

and

$$D_v(\varepsilon) = \frac{1}{2\pi^2}\left(\frac{2 m_h^*}{\hbar^2}\right)^{3/2} (E_v - \varepsilon)^{1/2} \tag{9.12}$$

When electron and hole energies are such that $|(\varepsilon - \varepsilon_F)| >> k_B T$, the respective distribution functions are effectively represented by Boltzmann distribution functions as expressed below:

$$f_e \approx \exp\left[\frac{-(\varepsilon - \varepsilon_F)}{k_B T}\right] \tag{9.13}$$

(in the conduction band)

and

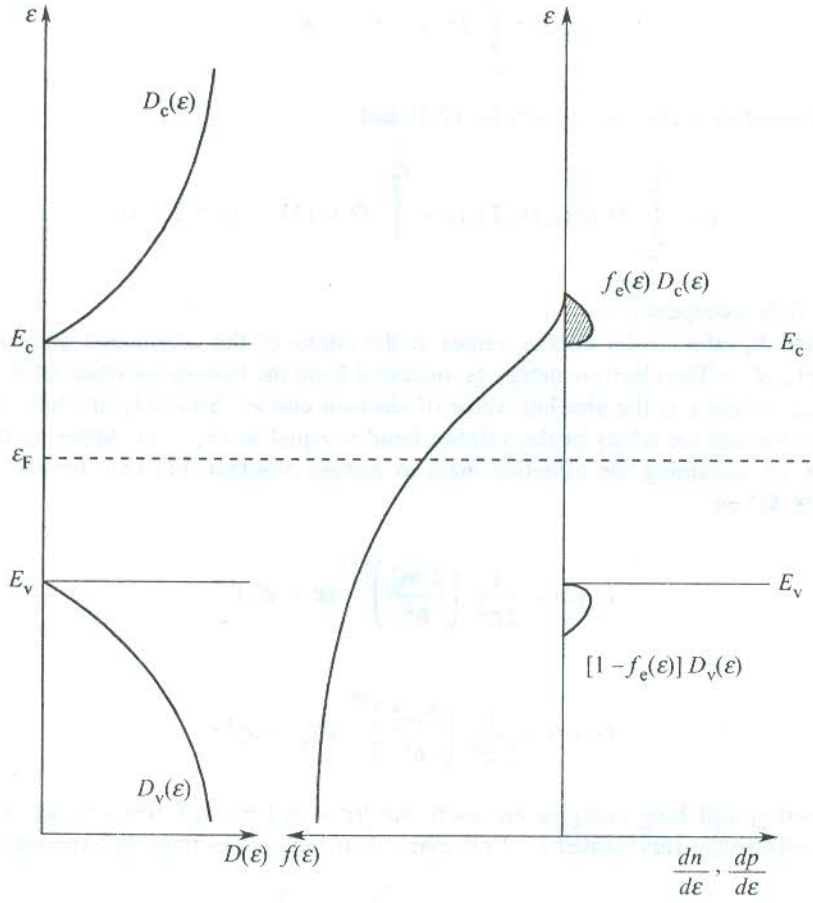
$$f_h = 1 - f_e \approx \exp\left[\frac{-(\varepsilon_F - \varepsilon)}{k_B T}\right] \tag{9.14}$$

(in the valence band)

In this condition the carrier density is not large and the semiconductor is called a non-degenerate semiconductor.

Substituting (9.13) in (9.9), the electron density in the state of thermal equilibrium at temperature T has the form,

$$n = \frac{1}{2\pi^2}\left(\frac{2m_e^*}{\hbar^2}\right)^{3/2} \exp\left(\frac{\varepsilon_F}{k_B T}\right) \int (\varepsilon - E_c)^{1/2} \exp\left(-\frac{\varepsilon}{k_B T}\right) d\varepsilon \tag{9.15}$$



(a)

FIG. 9.7 Contd.

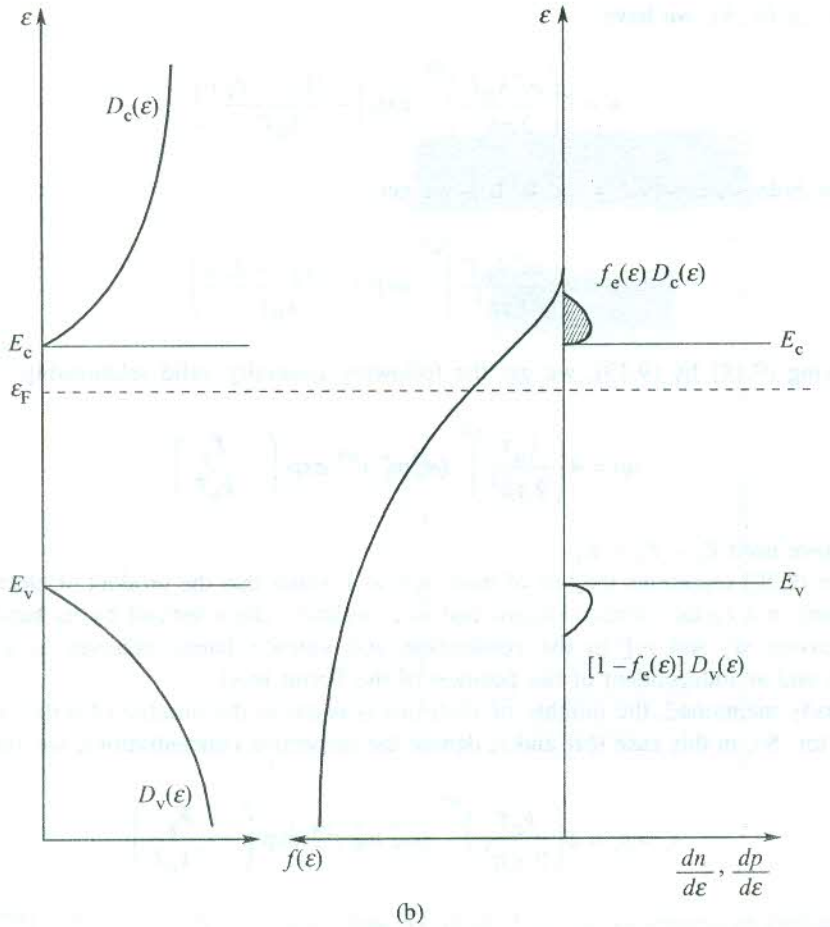


FIG. 9.7 A schematic behaviour of the Fermi function $f(\varepsilon)$, the density of states $D(\varepsilon)$, and electron (n) and hole (p) densities in the conduction and valence bands when: (a) the densities of states in the conduction and valence bands are equal and (b) the densities of states in the conduction and valence bands are not equal.

Using the substitution $\left(\frac{\varepsilon - E_c}{k_B T}\right) = x^2$, we get

$$n = \frac{1}{\pi^2}\left(\frac{2\, m_e^*}{\hbar^2}\right)^{3/2} (k_B T)^{3/2} \exp\left(\frac{-(E_c - \varepsilon_F)}{k_B T}\right) \int_0^\infty x^2 \exp(-x^2)\, dx \tag{9.16}$$

But,

$$\int_0^\infty x^2 \exp(-x^2)\, dx = \frac{\sqrt{\pi}}{4} \tag{9.17}$$

(from standard tables)

Using (9.17) in (9.16), we have

$$n = 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2} \exp\left(\frac{-(E_c - \varepsilon_F)}{k_B T}\right) \tag{9.18}$$

Similarly for holes, using (9.14) and (9.10), we get

$$p = 2\left(\frac{m_h^* k_B T}{2\pi\hbar^2}\right)^{3/2} \exp\left(\frac{-(\varepsilon_F - E_v)}{k_B T}\right) \tag{9.19}$$

By multiplying (9.18) by (9.19), we get the following generally valid relationship:

$$np = 4\left(\frac{k_B T}{2\pi\hbar^2}\right)^3 (m_e^* m_h^*)^{3/2} \exp\left(-\frac{E_g}{k_B T}\right) \tag{9.20}$$

where we have used $E_c - E_v = E_g$.

Relation (9.20) represents the law of mass action. It states that the product of electron and hole concentrations in a certain semiconductor that is completely characterized by its band gap E_g and effective masses m_e^* and m_h^* in the conduction and valence bands behaves as a function of temperature and as independent of the position of the Fermi level.

As already mentioned, the number of electrons is equal to the number of holes in an intrinsic semiconductor. So, in this case if n_i and p_i denote the respective concentrations, we get from (9.20)

$$n_i = p_i = 2\left(\frac{k_B T}{2\pi\hbar^2}\right)^{3/2} (m_e^* m_h^*)^{3/4} \exp\left(-\frac{E_g}{2 k_B T}\right) \tag{9.21}$$

The numbers as determined at 300 K in Si, Ge and GaAs are 2.4×10^{13}, 1.5×10^{10} and 5×10^7 cm^{-3}, respectively. In intrinsic materials the Fermi level at a certain temperature adjusts its position that is required to maintain the charge neutrality condition consistent with (9.18) and (9.19). This demands that

$$n = p$$

giving

$$\exp\left(\frac{2\varepsilon_F}{k_B T}\right) = \left(\frac{m_h^*}{m_e^*}\right)^{3/2} \exp\left(\frac{E_v + E_c}{k_B T}\right) \tag{9.22}$$

or

$$\varepsilon_F = E_v + \frac{1}{2}E_g + \frac{3}{4}k_B T \ln\left(\frac{m_h^*}{m_e^*}\right) \tag{9.23}$$

Thus the Fermi level shows a weak temperature dependence. But if the electron and hole effective masses are equal, $\varepsilon_F = E_v + \frac{1}{2}E_g$ and the Fermi level lies exactly in the middle of the gap. This is true at all temperatures.

But in general when the density of states of conduction and valence bands and the two effective masses are unequal, the Fermi level is asymmetrically placed with respect to the positions of E_c and E_v. It is demonstrated in Fig. 9.7(b) where the two density of states are unequal.

9.4 GENERAL FEATURES OF EXTRINSIC SEMICONDUCTORS

As mentioned in Section 9.1, the extrinsic semiconductors were developed because the electrical conductivity of intrinsic semiconductor is, generally, not large enough to meet the requirement of devices. Besides, it is hard to imagine of an absolute intrinsic material. Even the purest single crystals have impurity contents to a certain degree. But these impurity contents normally do not increase the carrier concentration to a useful level. GaAs is, however, an exception. The purest available single crystals of GaAs show a carrier density of about 10^{16} cm^{-3} which is enormously high compared to the intrinsic value (5×10^7 cm^{-3}). In general, the standard method of increasing the conductivity of an intrinsic material is to add to it a suitable impurity or electrically active element in small concentration. The method is known as *doping*. Impurities that enhance the carrier density by contributing additional electrons to the conduction band are called *donors* and those which create additional holes in the valence band are known as *acceptors*.

For example, let us consider the electrically active elements suitable for doping in Si and Ge crystals. When pure crystals of Si and Ge are doped with any element from Group V such as P or As, we get *n*-type semiconductors. On doping with an element from Group III (e.g. B, Al, In), *p*-type semiconductors are formed.

9.4.1 The *n*-type Semiconductors

Suppose a pure Ge crystal is doped with As, an immediate neighbour to the right of Ge in the periodic table. The arsenic atom may enter the germanium crystal lattice either by replacing a germanium atom (i.e. substitutionally) or by occupying a position where no germanium atom is supposed to be located in a pure and perfect crystal (i.e. interstitially). Data on lattice constant measurements show that the arsenic atom enters the crystal substitutionally. The electronic configurations in the outermost shells of germanium and arsenic atoms are $3s^2\,3p^2$ and $4s^2\,4p^3$, respectively. The germanium crystal has a diamond structure in which each atom forms tetrahedral bonds with its four neighbours. When an arsenic atom that substitutes a germanium atom finds itself surrounded by four germanium atoms, its four electrons get engaged in tetrahedral bonds with all four neighbouring atoms as depicted in Fig. 9.8. The arsenic atom is left with an extra loosely bound valence electron which may be easily freed and made available in the conduction band for the purpose of conduction. On account of having these additional electrons for conduction, the doped crystal is called an *n*-type semiconductor where *n* stands for electrons.

According to a model proposed to interpret the replacement of a germanium atom by an arsenic atom, the arsenic atom is pictured as an occasional germanium atom with an additional positive charge of *e* fixed at its core to which an additional electron is bound. Thus the donor impurity can be described as a hydrogen-like centre in which the Coulomb attraction between the core and the valence electron is screened by the neighbouring germanium electrons. The centre being a bound system is characterized by a set of quantized energy levels whose scheme is qualitatively similar to that of the hydrogen atom. The ionization energies of the hydrogen and the donor atoms are expressed as

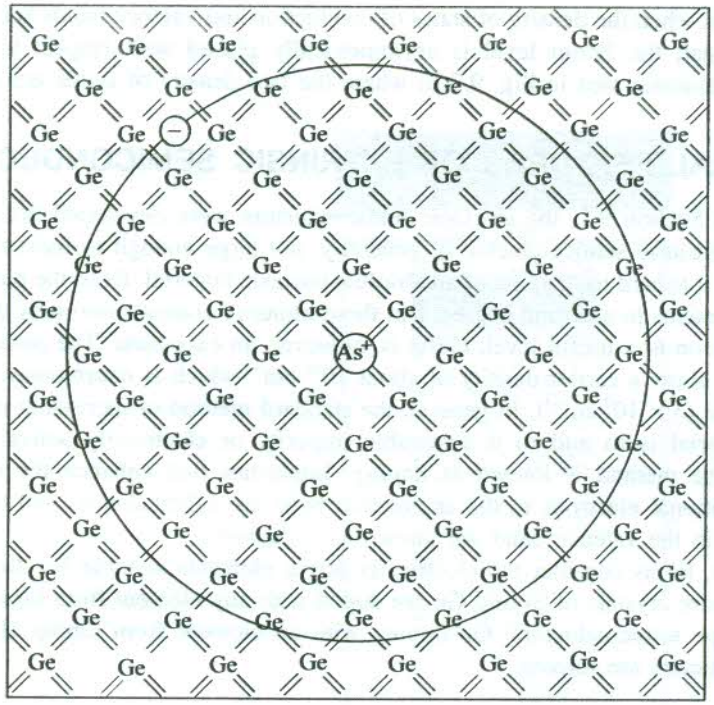


FIG. 9.8 A donor impurity atom As is shown to substitute a Ge atom in the germanium crystal. The fifth electron of As, that is not engaged in a tetrahedral bond (also interpreted as the bound valence electron of donor impurity centre), is smeared over thousands of Ge sites in the crystal. The lattice constant and the Bohr radius of the impurity centre are not drawn to scale.

$$E_H = \frac{me^4}{2(4\pi \epsilon_0 \hbar)^2} \tag{9.24}$$

for the hydrogen atom

and

$$E_d = \frac{m_e^* e^4}{2(4\pi \epsilon_m \hbar)^2} = \frac{m_e^* E_H}{m \epsilon_s^2} \tag{9.25}$$

for the donor atom

Here ϵ_0 is the permittivity of vacuum space and ϵ_s is the relative permittivity or the static dielectric constant of the medium of the germanium crystal. The constant $\epsilon_m (= \epsilon_0 \epsilon_s)$ represents the permittivity of the medium of the crystal. The values of the static dielectric constant of some important semiconductors are given in Table 9.2.

For calculating E_d we take the effective mass of the conduction electron in germanium as $m_e^* = 0.12m$ where m is the free electron mass. Using this substitution in (9.25), we get

$$E_d = \frac{(0.12)\, E_H}{\epsilon_s^2} \tag{9.26}$$

Table 9.2 Values of static dielectric constant ϵ_s for some semiconductors

Crystal	ϵ_s
Si	11.7
Ge	15.8
AlAs	10.1
AlSb	10.3
GaAs	13.13
GaSb	15.69
InSb	17.88
InAs	14.55
InP	12.37
SiC	10.2

With $E_H = 13.6$ eV and $\epsilon_s = 15.8$ for Ge, E_d is found out to be about 6.5 meV. For Si, $m_e^* = 0.3m$ and $\epsilon_s = 11.7$ and, therefore, E_d comes to about 30 meV. The position of the donor's ground energy level E_D with respect to the conduction and valence band edges is shown in Fig. 9.9(a). Its energy as measured from the conduction band edge is E_d. The energy continuum of the donor energy level scheme begins at the conduction band edge. Therefore, the electron from a donor atom can be transferred to the conduction band by simply ionizing the donor atom at the expense of only a few meV (~ 6.5 meV in Ge and ~ 30 meV in Si) and not of a few eV of energy as is the case with the intrinsic material. This makes abundantly clear why the conductivity of an *n*-type semiconductor is several orders bigger than that of the corresponding intrinsic material.

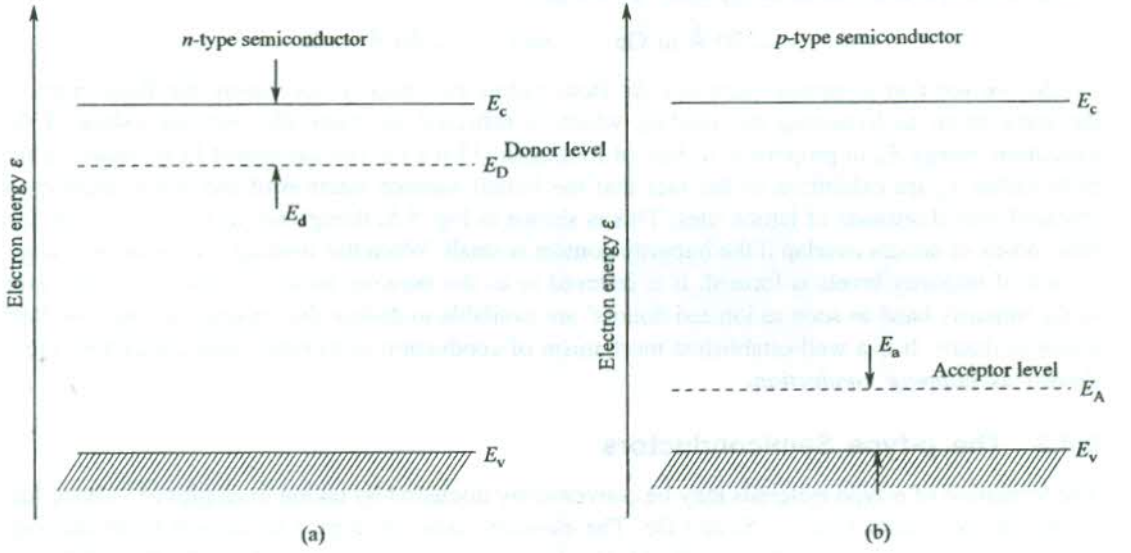


FIG. 9.9 (a) A qualitative picture showing the position of the ground donor level E_D relative to the conduction and valence band edges E_c and E_v. The E_d denotes the ionizaiton energy of the donor atom. (b) A qualitative picture showing the position of the ground acceptor level E_A relative to the conduction and valence band edges E_c and E_v. The E_a denotes the ionization energy of the acceptor atom.

An odd feature of the model described above is that it gives the same value of E_d for all donor impurities such as P, As and Sb in a semiconductor host. The experimental values, however, show a slight variation (see Table 9.3). The description of screening in terms of the dielectric constant is rather crude and it restricts the operational domain of the model. The limitations are reflected in the model's failure to account for some of the subtle consequences of the atomic effects. The effect of screening on the Bohr radius is further revealing. The Bohr radius of the donor impurity atom is written as

$$r_d = \frac{4\ \pi \epsilon_0 \epsilon_s\ \hbar^2}{m_e^* e^2} = \left(\frac{m \epsilon_s}{m_e^*}\right) r_0 \tag{9.27}$$

where r_0 is the hydrogen Bohr radius (0.53 Å).

Table 9.3 Ionization energies for a few donors and acceptors in silicon and germanium (E_d: ionization energy of donors and E_a: ionization energy of acceptors)

Impurity	E_d (meV)		*Impurity*	E_a (meV)	
	Si	Ge		Si	Ge
P	45	12	B	45	10.4
As	49	12.7	Al	57	10.2
Sb	39	9.6	Ga	65	10.8
			In	16	11.2

Using the respective values of ϵ_s and m_e^*, we get

$$r_d \simeq 70 \text{ Å in Ge} \qquad \text{and} \qquad \simeq 20 \text{ Å in Si}$$

So we see that screening increases the Bohr radius enormously. Increasing the Bohr radius is the same thing as loosening the binding which is reflected in drastically reduced values of the ionization energy E_d in proportion to that of hydrogen (13.6 eV). The calculated large values of the Bohr radius r_d are exhibition of the fact that the bound valence electron of the donor impurity is smeared over thousands of lattice sites. This is shown in Fig. 9.8, though not to scale. Thus the first Bohr orbits of donors overlap if the impurity content is small. When this overlap is considerably large, a band of impurity levels is formed. It is referred to as the *impurity band*. Conduction occurs even in the impurity band as soon as ionized donors† are available to initiate the hopping of electrons from donor to donor. It is a well-established mechanism of conduction in extrinsic semiconductors and is known as *hopping conduction*.

9.4.2 The *p*-type Semiconductors

The formation of *p*-type materials may be conveniently discussed by taking examples of valence four elemental semiconductors, i.e. Si and Ge. The elements used for doping to convert these materials into *p*-type are from Group III (e.g. B, Al, Ga, In) with three electrons in the outermost shell. The

† Some ionized donors always exist, firstly because of the presence of some acceptors whose affinity for an electron may ionize donors and secondly because the average thermal energy per electron ($\sim k_B T$) around room temperature is ~26 meV which is enough to ionize some donors.

substitutional impurity in this case completes only three of the four characteristic tetrahedral bonds of the host crystal. The doped impurity atom lacks one electron to complete the bonding with its all the neighbours. A valence electron of the host material may meet this requirement by ionizing the impurity atom negatively and creating a hole in the vicinity. Since the impurity atom is willing to accept an electron, the impurity atom is called an *acceptor*. The acceptor model is similar to the donor model. For example, a boron atom substituted in the germanium crystal is pictured as a germanium atom with a charge of $(-e)$ fixed at its core and a hole of charge (e) bound to it. The impurity acts as the centre of a bound system whose energy level scheme is similar to that of the hydrogen atom. The ground acceptor level E_A lies close to the valence band edge as shown in Fig. 9.9(b). The ionization energy E_a as measured from the valence band edge is again very small as compared to the size of E_g. The values of E_a for some acceptors in Si and Ge are given in Table 9.3. Arguments concerning large values of conductivity and hopping conduction follow the same lines as those for n-type materials. With the latest methods of doping, the lowest impurity concentrations that can be obtained in semiconductors are of the order of 10^{12} cm^{-3}. Hence Si that has an intrinsic concentration of 1.5×10^{10} cm^{-3} at 300 K does not show intrinsic conductivity though Ge does, on account of its higher intrinsic concentration (2.4×10^{13} at 300 K).

9.5 POPULATION OF DONOR AND ACCEPTOR LEVELS IN THE STATE OF THERMAL EQUILIBRIUM

Let us first define the symbols that will be used in our future descriptions:

N_D/N_A : density of all available donors/acceptors

N_D^0/N_A^0 : density of the neutral donors/acceptors

N_D^+/N_A^- : density of ionized donors/acceptors

n_D/p_A : density of electrons bound to donors/density of holes bound to acceptors

n : density of electrons in the conduction band

p : density of holes in the valence band.

For small impurity concentrations (i.e. for non-degenerate semiconductors) the occupancy of the conduction and valence bands is as usual described by the Boltzmann distribution function. Therefore, at these concentrations the law of mass action (9.20) which was derived on the basis of the above conditions must apply even to extrinsic semiconductors. But in the extrinsic case the value of the Fermi energy, not figuring in the law of mass action, depends on a more complicated charge neutrality condition. For homogeneous doping, the neutrality condition can be expressed as

$$n + N_A^- = p + N_D^+ \tag{9.28}$$

with

$$N_D = N_D^0 + N_D^+ \tag{9.29a}$$

$$N_A = N_A^0 + N_A^- \tag{9.29b}$$

The above terminology may be appreciated better with the help of Fig. 9.10. Let us first consider the calculation of population of the donor levels. At low impurity concentrations (up to 10^{17} cm^{-3}), the interaction between electrons bound to separate donors may be neglected. We further assume for

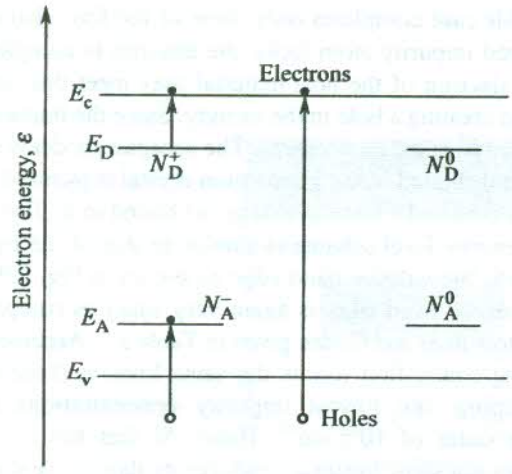


FIG. 9.10 Electrons in the conduction band and holes in the valence band made available either by inter-band excitation or by impurity ionization.

simplicity of calculations that the impurity introduces only a single one-electron level. Under these approximations the level could either be empty, or contain one electron of either spin, or contain two electrons with opposite spins. The double occupation is not favoured on account of its high energy arising from the Coulomb repulsion between two localized electrons. In the state of thermal equilibrium, the mean number of electrons in a system is expressed as

$$\langle n \rangle = \frac{\sum_j N_j \exp\left[-(\varepsilon_j - \varepsilon_F N_j)/k_B T\right]}{\sum_j \exp\left[-(\varepsilon_j - \varepsilon_F N_j)/k_B T\right]} \tag{9.30}$$

where the sum is over all states of the system; the ε_j and N_j denote the energy and the number of electrons in state j. With a single impurity we have just three states: one with no electrons, thereby making no contribution to the energy, and two distinguishable spin states with a single electron of energy E_D. Hence from (9.30), we have

$$\langle n \rangle = \frac{2 \exp\left[-(E_D - \varepsilon_F)/k_B T\right]}{1 + 2 \exp\left[-(E_D - \varepsilon_F)/k_B T\right]}$$

$$= \frac{1}{1 + \frac{1}{2}\left[(E_D - \varepsilon_F)/k_B T\right]} \tag{9.31}$$

The factor of 1/2 serves to modify the Fermi-Dirac distribution in this case. However, this factor is generally ignored which amounts to treating two single electron spin states of equal energy as one state. The population density of the donor level in the state of thermal equilibrium is given by

$$n_D = N_D \langle n \rangle$$

or

$$n_D = \frac{N_D}{1 + \exp[(E_D - \varepsilon_F)/k_B T]} \qquad \text{(ignoring the factor of 1/2)} \tag{9.32}$$

Also note that,

$$n_D = N_D^0 \tag{9.33}$$

Similarly, it can be shown that the population density of the acceptor level in thermal equilibrium is given by

$$p_A = \frac{N_A}{1 + \exp[(\varepsilon_F - E_A)/k_B T]} \tag{9.34}$$

and also,

$$p_A = N_A^0 \tag{9.35}$$

9.6 EXTRINSIC CARRIER DENSITIES

The case of an extrinsic semiconductor in which both donors and acceptors are present is difficult to handle. Because of this reason we consider a pure *n*-type semiconductor and calculate its carrier concentration that is contributed by donors alone. Generally, every *n*-type material has a few acceptors and every *p*-type material has a few donors because of practical limitations on growing 100 per cent pure crystals. But the concentration of naturally present impurities in a pure crystal is expected to be negligibly small compared to the concentration of the impurity doped. Therefore, the relation (9.28) in the present case is rewritten as

$$n = N_D^+ + p_i \tag{9.36}$$

The net electron density n has contributions of electrons from donors and from the valence band (see Fig. 9.10). The density of the latter type (n_i, being intrinsic) is equal to the density of holes (p_i). Normally $N_D^+ >> n_i$ (or p_i). For example, in Si this condition is satisfied even at low levels of the donor concentration since the intrinsic concentration is mere 1.5×10^{10} cm^{-3} around room temperature. Therefore for our *n*-type material, p_i in (9.36) may be dropped giving,

$$n \simeq N_D^+ = N_D - N_D^0 \qquad \text{[using (9.29a)]} \tag{9.37}$$

or

$$n = N_D - n_D \qquad (\text{since } n_D = N_D^0)$$

$$= N_D \left(1 - \frac{1}{1 + \exp[(E_D - \varepsilon_F)/k_B T]}\right) \qquad \text{[using (9.32)]}$$

$$= \left(\frac{N_D}{1 + \exp[-(E_D - \varepsilon_F)/k_B T]}\right) \tag{9.38}$$

As already mentioned in the beginning of Section 9.5, the relations for n, p and np (the law of the mass action) as derived in Section 9.3 are applicable to extrinsic materials with low doping levels. We rewrite the relation (9.18) as

$$n = N(\mathrm{c}) \exp\left[\frac{-(E_\mathrm{c} - \varepsilon_\mathrm{F})}{k_\mathrm{B}T}\right] \tag{9.39a}$$

with

$$N(\mathrm{c}) = 2\left(\frac{m_\mathrm{e}^* k_\mathrm{B} T}{2\pi\hbar^2}\right)^{3/2} \tag{9.39b}$$

and (9.19) as

$$p = N(\mathrm{v}) \exp\left(\frac{-(\varepsilon_\mathrm{F} - E_\mathrm{v})}{k_\mathrm{B}T}\right) \tag{9.40a}$$

with

$$N(\mathrm{v}) = 2\left(\frac{m_\mathrm{h}^* k_\mathrm{B} T}{2\pi\hbar^2}\right)^{3/2} \tag{9.40b}$$

From (9.39a), we have

$$\exp\left(\frac{\varepsilon_\mathrm{F}}{k_\mathrm{B}T}\right) = \frac{n \exp(E_\mathrm{c}/k_\mathrm{B}T)}{N(\mathrm{c})} \tag{9.41}$$

Eliminating ε_F in (9.38) with the help of (9.41), we get

$$n = \frac{N_\mathrm{D}}{1 + \dfrac{n \exp(E_\mathrm{d}/k_\mathrm{B}T)}{N(\mathrm{c})}} \qquad (\text{with } E_\mathrm{c} - E_\mathrm{D} = E_\mathrm{d})$$

or

$$n^2\left(\frac{\exp(E_\mathrm{d}/k_\mathrm{B}T)}{N(\mathrm{c})}\right) + n - N_\mathrm{D} = 0 \tag{9.42}$$

Using the substitution,

$$\frac{\exp(E_\mathrm{d}/k_\mathrm{B}T)}{N(\mathrm{c})} = X \tag{9.43}$$

we get

$$Xn^2 + n - N_\mathrm{D} = 0 \tag{9.44}$$

The physically meaningful solution to the above equation is

$$n = \frac{-1 + \sqrt{1 + 4N_\mathrm{D}X}}{2X} \tag{9.45}$$

Rationalizing (9.45), we get

$$n = 2N_\mathrm{D}\left[1 + \sqrt{1 + 4N_\mathrm{D}X}\right]^{-1}$$

Using (9.43), we have

$$n = 2N_D\left[1 + \sqrt{1 + 4N_D \frac{\exp(E_d/k_BT)}{N(c)}}\right]^{-1} \tag{9.46}$$

There are three limiting cases of the above expression as discussed below:

Case I. At low temperatures,

$$4\left(\frac{N_D}{N(c)}\right)\exp\left(\frac{E_d}{k_BT}\right) >> 1 \tag{9.47}$$

This transforms (9.46) to the following form,

$$n \simeq \sqrt{N_D N(c)}\exp\left(\frac{-E_d}{2k_BT}\right) \tag{9.48}$$

At these low temperatures a large number of donors stay in the unionized state. The range of temperatures over which this condition exists is known as the *freeze-out range*. We can notice the similarities between (9.48) and (9.21) that expresses the intrinsic carrier density. Both depend exponentially on temperature since the exponential dependence dominates the other dependence entering through $N(c)$. In the donors' case a much smaller quantity E_d appears in the exponential as against E_g in the intrinsic case. This accounts for a larger carrier concentration in the n-type material.

Case II. For temperatures at which

$$4\left(\frac{N_D}{N(c)}\right)\exp\left(\frac{E_d}{k_BT}\right) << 1 \tag{9.49}$$

the relation (9.46) reduces to

$$n \simeq N_D \tag{9.50}$$

i.e. a constant.

This means that at these temperatures all donors are ionized and the electron density reaches its maximum with the excitation of electrons from the valence band considered negligible in the first approximation. This is referred to as the *saturation range* or *exhaustion range* of carriers.

Case III. When temperatures are still higher, the concentration of conduction electrons contributed by the valence band becomes appreciable. Since the concentration of donor electrons in the conduction band no more increases on account of saturation, the intrinsic electron density overtakes the density of donor electrons at some stage. In this condition an n-type semiconductor is characterized by intrinsic behaviour and we speak of the *intrinsic range* of carriers. The variations of electron density and energy as functions of temperature are sketched in Fig. 9.11, identifying the three ranges.

The calculation of electron density in the conduction band as carried out above is based on the assumption that there are no acceptor impurities in the n-type semiconductor under consideration. Experiments, however, confirm that traces of acceptors are always present. Therefore, quantitative estimates made on the basis of the above theory may differ slightly from the experimental values.

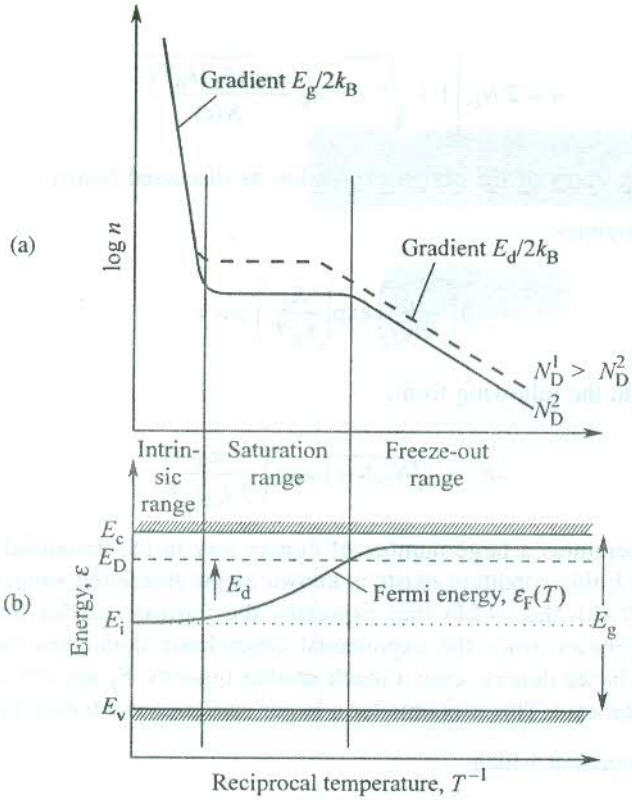


FIG. 9.11 (a) A qualitative illustration of the variation of electron concentration (n) in the conduction band with change in temperature in two samples of different donor concentrations N_D^1 and N_D^2. (b) A qualitative illustration of the temperature dependence of the Fermi energy ε_F in the same semiconductor. The E_i represents the Fermi energy in the instrinsic material. All other symbols have their usual meaning.

There are several applications in which there is a need to monitor the electron and hole concentrations. Since at low doping levels the law of mass action (9.20) is valid, it is exploited to manipulate the carrier concentration in non-degenerate† semiconductors. When the electron concentration n is increased by adding the trace of a suitable donor impurity, the hole concentration p goes down so that the product (np) remains constant at a temperature in accordance with the law of mass action. This way it is possible to reduce the sum ($n + p$) considerably. This method of reduction is in wide practice and known as *compensation*. The sign of majority carriers is quickly determined by detecting the sign of Hall voltage.

† In a degenerate semiconductor the number of impurity atoms is considerably large, due to which the number of energy levels in an impurity band is large. At these large densities the interatomic separation between impurity atoms becomes smaller resulting in the large overlap of the atomic orbitals. Because of this enhanced overlap, the width of the impurity band is increased. Under these circumstances some of the impurity levels extend into the region of the conduction band or valence band depending on whether the levels belong to the donor or acceptor impurity band. The energy value (eigenvalue) of any such impurity level is thus common to two eigen states, one belonging to the host atom and the other to the impurity atom. Hence the level in question is degenerate in the literal sense. The semiconductor with this band scheme is called a *degenerate semiconductor.*

9.7 TEMPERATURE DEPENDENCE OF ELECTRICAL CONDUCTIVITY

The electrical conductivity does not show the same type of temperature dependence over the extended and the practically useful range of temperatures. Above a certain temperature, the conductivity is not modified by the impurities in the crystal. The range extending above this temperature is called *the intrinsic temperature range* and the conductivity is referred to as the intrinsic conductivity (σ_i). Substituting the values of carrier densities (9.21) in (9.7), the intrinsic conductivity can be written as

$$\sigma_i = 2|e|\left(\frac{k_B T}{2\pi\hbar^2}\right)^{3/2} (m_e^* m_h^*)^{3/4}\, e^{-E_g/2k_B T}\,(\mu_n + \mu_p) \tag{9.51}$$

As we will see below, the mobilities depend on temperature only as a simple power law over an appropriate region. In view of this fact, the temperature dependence of conductivity will be dominated by the exponential dependence of carrier concentrations, giving

$$\sigma_i \propto e^{-E_g/2k_B T} \tag{9.52}$$

or

$$\rho_i = A e^{E_g/2k_B T} \tag{9.53}$$

which gives

$$\ln \rho_i = \ln A + \frac{E_g}{2k_B T}$$

The above relation shows that the intrinsic resistivity is a linear function of the reciprocal of temperature which is confirmed by experiments.

However, for a proper analysis of the temperature dependence of electrical conductivity, the temperature variation of the carrier densities and that of the mobilities must be weighed against each other.

The conduction in semiconductors differs from the metallic conduction in the sense that electrons near the Fermi level contribute to the metallic conductivity whereas in semiconductors electrons near the bottom of the conduction band and holes near the top of the valence band take part in the process. There are no charge carriers near the Fermi level in semiconductors. In the approximation where the Fermi statistics can be replaced by Boltzmann statistics for non-degenerate semiconductors, the Boltzmann equation is solved to get to the expression for mobility of charge carriers. On demand of simplicity we present here only a qualitative description of the scattering processes in which electrons and holes by and large behave similarly. Several simplifications of the exact expression for mobility lead to the result,

$$\mu \propto \tau \tag{9.54}$$

where τ is the relaxation time defined in Section 6.1.1.

Since τ is proportional to the average time between successive collisions,

$$\frac{1}{\tau} \propto \langle v \rangle \Sigma \tag{9.55}$$

where Σ denotes the scattering cross-section of electrons and holes at a scattering centre. Relation (9.55) gives a measure of the scattering probability with $\langle v \rangle$ treated as thermal average over all electron or hole velocities in the lower conduction band or upper valence band. The use of Boltzmann statistics in semiconductors gives

$$\langle v \rangle \propto \sqrt{T} \tag{9.56}$$

Phonons happen to be the prominent source of carrier scattering in crystals. It is simpler to calculate the phonon scattering cross-section Σ_L for acoustic phonons, ($T >> \theta_D$), where θ_D is the Debye temperature. This gives the dependence,

$$\Sigma_L \propto T \qquad \text{(L stands for lattice or phonons)} \tag{9.57}$$

Making use of (9.55), (9.56) and (9.57), we have from (9.54)

$$\mu_L \propto T^{-3/2} \tag{9.58}$$

In semiconductors, centres of ionized donors and acceptors serve as another important source of scattering. As an electron or a hole approaches such a centre, it experiences a Coulomb force and suffers scattering similar to Rutherford scattering. A rigorous treatment shows that

$$\Sigma_I \propto \langle v \rangle^{-4} \tag{9.59}$$

where I stands for ionized impurity. Therefore, with the use of (9.55), (9.56) and (9.59) in (9.54), we get

$$\mu_I \propto T^{3/2} \tag{9.60}$$

For one type of carriers, say electrons, we can write contributions of phonons and ionized impurities to resistivity as

$$\rho_L = \frac{1}{\sigma_L} = \frac{1}{ne\mu_L} \quad \text{and} \quad \rho_I = \frac{1}{\sigma_I} = \frac{1}{ne\mu_I} \tag{9.61}$$

Applying the Matthiessen's rule, the total resistivity is written as

$$\rho = \frac{1}{ne\mu} = \frac{1}{ne}\left[\frac{1}{\mu_L} + \frac{1}{\mu_I}\right] \tag{9.62}$$

or

$$\frac{1}{\mu} = \frac{1}{\mu_L} + \frac{1}{\mu_I} \tag{9.63}$$

A qualitative display of the temperature dependence of mobility in an extrinsic semiconductor is made in Fig. 9.12.

The variation of conductivity σ as a function of temperature is plotted for six samples of n-type germanium in Fig. 9.13. The conductivity shows a maximum in the exhaustion range where the electron concentration n becomes almost constant. The behaviour of σ in this range is governed by the characteristic temperature dependence of mobility as shown in Fig. 9.12. In the intrinsic range (at high temperatures) and in the freeze-out (at low temperatures), the variation in σ is effectively controlled

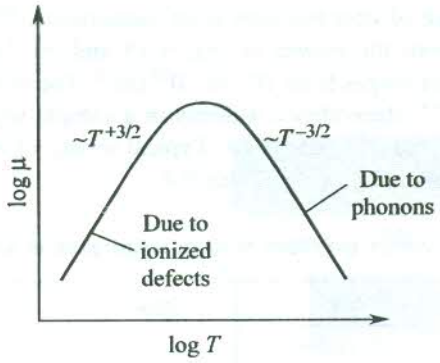


FIG. 9.12 Qualitative temperature dependence of the mobility contributions from charged impurities and phonons.

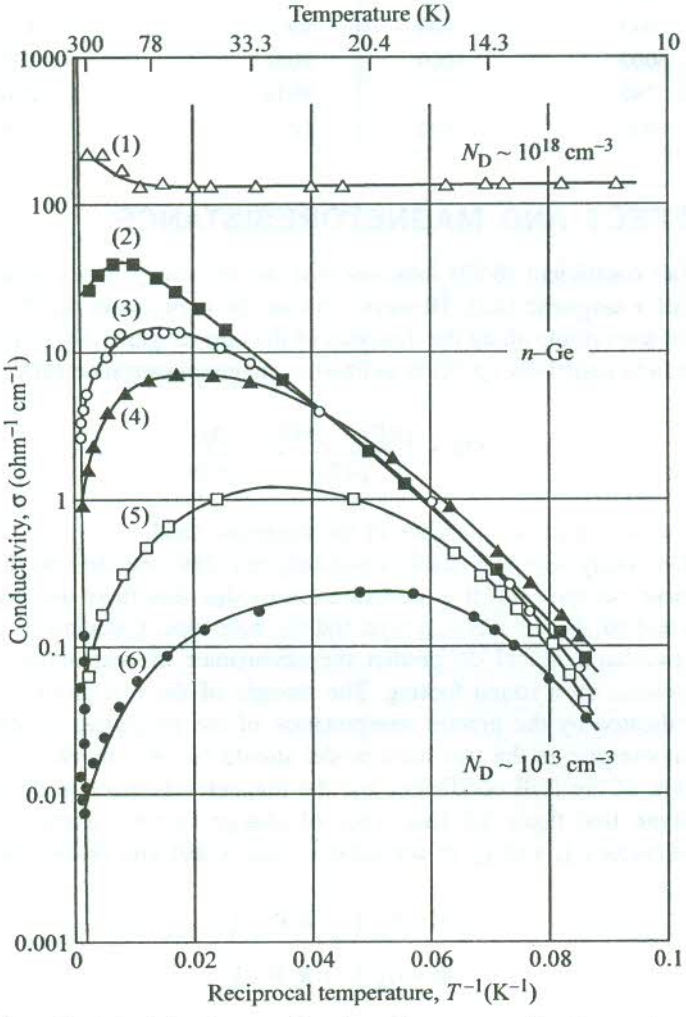


FIG. 9.13 Observed conductivity behaviour as a function of temperature for six samples of *n*-type germanium. The donor concentration N_D ranges from 10^{13} to 10^{18} cm^{-3}. [After E.M. Conwell, *Proc. IRE*, **40,** 1327 (1952).]

by the exponential dependence of concentration n on temperature. The temperature dependence of mobility has been derived from the curves of Fig. 9.13 and the Hall measurements. The donor concentration N_D in six samples ranges from 10^{13} to 10^{18} cm^{-3}. The mobility in the purest crystal with $N_D \simeq 10^{13}$ cm^{-3} shows ideal $T^{-3/2}$ dependence whereas in a sample with increasing values of N_D the mobility approaches the expected $T^{3/2}$ behaviour. Typical orders of electron and hole mobilities in some important semiconductors are given in Table 9.4.

Table 9.4 Typical order of carrier mobilities at room temperature in some important semiconductors

Crystal	*Mobility* (cm^2/V · s)		*Crystal*	*Mobility* (cm^2/V · s)	
	Electrons	*Holes*		*Electrons*	*Holes*
Si	1350	480	InSb	800	450
Ge	3600	1800	InAs	30,000	450
GaAs	8000	300	InP	4500	100
GaSb	5000	1000	PbSe	1020	930
AlAs	280	—	PbTe	2500	1000
AlSb	900	400	SiC	100	10–20

9.8 HALL EFFECT AND MAGNETORESISTANCE

The relation for Hall coefficient (6.89) indicates that the resistivity of a conductor is not affected by the application of a magnetic field. However, during the experiment on Hall effect it is noticed that the resistance of the sample along the direction of the current generally increases. The fractional change in the zero-field resistivity ($\rho(0)$) is defined as magnetoresistance (MR):

$$MR = \frac{\rho(B) - \rho(0)}{\rho(0)} = \frac{\Delta\rho}{\rho(0)} \tag{9.64}$$

where $\rho(B)$ is the resistivity in the presence of the magnetic field.

It is important to clarify here that (6.89), predicting zero MR, was derived for an isotropic single band model of charge carriers. But it is common knowledge that there are two types of constant energy surfaces in real solids: the electron type and the hole type. Calculations of galvanomagnetic properties in the two-band model do predict the occurrence of magnetoresistance, putting the experimental observation on a sound footing. The thought of the two groups of charge carriers is unambiguously vindicated by the precise interpretation of the properties of semiconductors on its basis. Therefore, an exercise in the two-band model should be very much in order. We undertake below the calculation of the Hall coefficient and the magnetoresistance in this approach.

When we assume that there are two types of charge carriers (electrons and holes), their respective current densities $\mathbf{j}_e$ and $\mathbf{j}_h$, in accordance with (6.86) and (6.89) are given by

$$\begin{aligned} \mathbf{E} &= \sigma_e^{-1}\mathbf{j}_e + R_e\mathbf{B}\times\mathbf{j}_e \\ \mathbf{E} &= \sigma_h^{-1}\mathbf{j}_h + R_h\mathbf{B}\times\mathbf{j}_h \end{aligned} \tag{9.65}$$

where σ and R denote the conductivity and the Hall coefficient respectively.

The solution of the Boltzmann transport equation for single band model in the presence of magnetic field gives

$$\mathbf{j} = a_1\mathbf{E} + a_2\mathbf{E}\times\mathbf{B} + a_3(\mathbf{B}\cdot\mathbf{E})\mathbf{B} \tag{9.66}$$

where $a_1 = \dfrac{\sigma_0}{1+(\omega_c\tau)^2}$, $a_2 = \dfrac{\sigma_0\left(\dfrac{\omega_c\tau}{B}\right)}{1+(\omega_c\tau)^2}$ and $a_3 = \dfrac{\sigma_0\left(\dfrac{\omega_c\tau}{B}\right)^2}{1+(\omega_c\tau)^2}$ with σ_0 as the conductivity in zero-field.

In the experiment of Hall effect, **E** is perpendicular to **B**. Therefore, the total current density in the two-band model according to (9.66) is,

$$\begin{aligned}\mathbf{j} &= \mathbf{j}_e + \mathbf{j}_h \\ &= \left(\frac{\sigma_0(e)}{1+(\omega_c\tau_e)^2} + \frac{\sigma_0(h)}{1+(\omega_c\tau_h)^2}\right)\mathbf{E} + \left(\frac{R_e\sigma_0^2(e)}{1+(\omega_c\tau_e)^2} + \frac{R_h\sigma_0^2(h)}{1+(\omega_c\tau_h)^2}\right)\mathbf{E}\times\mathbf{B}\end{aligned} \tag{9.67}$$

To simplify calculations, we assume that electrons (with density, n) and holes (with density, p) occupy two separate isotropic energy surfaces, have the same effective mass, m^* and same relaxation time, τ but have equal and opposite cyclotron frequencies $\mp\omega_c$. With these considerations, the current density has the form:

$$\mathbf{j} = \frac{\sigma_0}{1+(\omega_c\tau)^2}\,\mathbf{E} + \left(\frac{p-n}{p+n}\right)\frac{\sigma_0\left(\dfrac{\omega_c\tau}{B}\right)}{1+(\omega_c\tau)^2}\,\mathbf{E}\times\mathbf{B} \tag{9.68}$$

with $\sigma_0 = \dfrac{(n+p)e^2\tau}{m^*}$, the conductivity in absence of the magnetic field.

Calculation of Hall coefficient

The Hall coefficient (R_H) can be obtained by inverting Equation (9.68) and expressing **E** in form (6.86). Comparing the equation thus obtained with (6.86) and using (6.89), one finds that the coefficient of ($\mathbf{B}\times\mathbf{j}$) in the equation is equal to the Hall coefficient having the form:

$$R_H = \frac{(p-n)(1+(\omega_c\tau)^2)}{e\left[(p+n)^2 + (p-n)^2(\omega_c\tau)^2\right]} \tag{9.69}$$

It gives $R_H = 0$ when $n = p$. In low magnetic fields for closed orbits $\omega_c\tau << 1$ and in this limit, the value of the Hall coefficient is given by,

$$R_H = \frac{(p-n)}{(p+n)^2 e} = \frac{R_e\sigma_0^2(e) + R_h\sigma_0^2(h)}{(\sigma_0(e)+\sigma_0(h))^2} = \frac{p\mu_p^2 - n\mu_n^2}{e(p\mu_p + n\mu_n)^2} \tag{9.70}$$

The sign of R_H is determined by the competing opposite signs for electrons and holes. The sign determined experimentally is consistent with it. The occurrence of Hall effect in samples of *n*-type and *p*-type semiconductors is illustrated in Fig. 9.14. The scene in a real crystal is the one of the superimpositions of the two pictures shown in the figure.

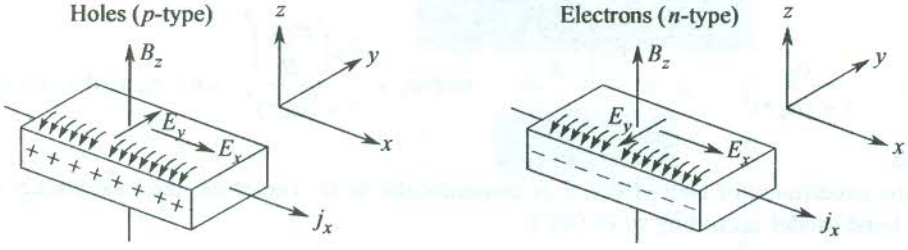


FIG. 9.14 Directions of the applied electric and magnetic fields (**E** and **B** respectively) relative to the direction of current flow in two samples of a semiconductor. The current density j is along the x-direction. Note that the polarity of the Hall voltage is opposite in p- and n-type samples.

Calculation of magnetoresistance

The magnetoresistance behaves as a tensor quantity. Only two components of MR are of special interest: the longitudinal (**j** parallel to **B**) and the transverse (**j** perpendicular to **B**). For isotropic two-band energy surfaces, the longitudinal component vanishes. In order to calculate the transverse *MR*, we are required to determine the component of **E** along **j**. Then, the resistivity in the presence of a magnetic field is expressible as

$$\rho(B) = \frac{\mathbf{j}\cdot\mathbf{E}}{|\mathbf{j}|^2} = \frac{\dfrac{1+(\omega_c\tau)^2}{\sigma_0}}{\left[1+\left(\dfrac{p-n}{p+n}\right)^2(\omega_c\tau)^2\right]} \tag{9.71}$$

This gives

$$MR = \frac{\rho(B)-\rho(0)}{\rho(0)} = \frac{4(\omega_c\tau)^2 np}{(n+p)^2+(n-p)^2(\omega_c\tau)^2} \tag{9.72a}$$

which for an intrinsic material ($n = p$) will have the simple form

$$MR = \omega_c^2\tau^2 \tag{9.72b}$$

The above pair of relations demonstrate that the magnetoresistance is a positive number. We also observe that it is proportional to B^2 (since $\omega_c \propto B$) in intrinsic materials. On increasing the magnetic field (increasing ω_c), the *MR* for closed orbits first increases gradually, then more slowly and finally saturates. When $n \neq p$, the approximate value at saturation is given by

$$MR = \frac{4np}{(n-p)^2} \tag{9.73}$$

One of the mysterious aspects of *MR* is the evidence of a healthy longitudinal *MR* in many semiconductors. Ideally, it should be zero in an isotropic two-band model. Its origin is attributed to the anisotropy in effective mass tensor or simply to the non-spherical shape of Fermi surfaces. As stated in Chapter 8, the measurement of *MR* is useful in deriving information about the shape of Fermi surfaces. If there are several types of charge carriers, the technique, however, fails to provide any worthwhile information.

9.9 THE *p-n* JUNCTION

The comparative ease with which the electrical characteristics of semiconductors can be monitored by controlling the traces of impurities being doped is simply striking. Because of this property, semiconductors have emerged as the indispensable material base for the development of solid state electronics. The dramatic and extensive technological consequences of this property have given a tremendous boost to the commercial interest in semiconductors.

It is of further significance to note that semiconductor devices generally exploit the characteristics of inhomogeneous semiconductors in which the donor and acceptor concentrations are not uniform. A common form in which they are used contains two separate n and p type regions with an abrupt partition boundary (a junction) within a single crystalline sample. This junction is known as the p-n junction. There can be more than one junction in a device. Characteristics of the p-n junction are crucial to the fabrication and working of most of the semiconductor devices. It is an essential component of the present electronics circuitry ranging from a simple rectifier circuit to transistor circuits and integrated circuits used in sophisticated appliances like modern computer. On the other hand, optical applications such as photo cells, LEDs and injection lasers (laser diodes) are equally important today in view of the rapidly upcoming field of optoelectronics. A proper description of all of these is beyond the scope of this book. Nevertheless, salient features of the working of a selected few devices will be presented in Section 9.10. This section is devoted to the description of the theoretical basis for the conduction behaviour of a p-n junction.

Consider a p-n junction, for example of a Si crystal whose left half is p-type and the right side is n-type. A p-n junction is never made by simply joining the two types of a material. It is assumed that the reader has been exposed to the methods of forming p-n junctions through a course in electronics. Nonetheless, the supposition that a piece of n-type and a piece of p-type material are brought into intimate contact helps in determining the electrostatic conditions present at a junction. When the two regions are treated as isolated, the levels of Fermi energy in the two regions are at different positions on the common energy scale [Fig. 9.15(a)]. Since the two regions are parts of the same crystal, the Fermi energy (representing the electrochemical potential in semiconductors) must have the same value in both halves in the state of thermal equilibrium. Near $T \simeq 0$ K the position of the Fermi level is near the acceptor level in the p-region and near the donor level in the n-region. Because the impurity levels are at the extremes of the gap, the Fermi level would not have maintained a constant level unless the bending of conduction and valence band edges occurred in the transition zone. [Fig. 9.15(b)]. Immediately after the junction comes into existence, electrons from the n-region begin to diffuse into the p-region and combine with the holes present in the vicinity of acceptors. This leaves behind the positively charged donors in the n-region and produces negatively charged acceptors in the p-region. The immobile ionized impurities are considered to form a charged double layer of space charges across the jump in the doping profile [Fig. 9.15(c)]. The charged donors and acceptors create an electric field across the transition region whose direction is such as to oppose the diffusion of free charge carriers. Therefore, in the absence of any external field the diffusion

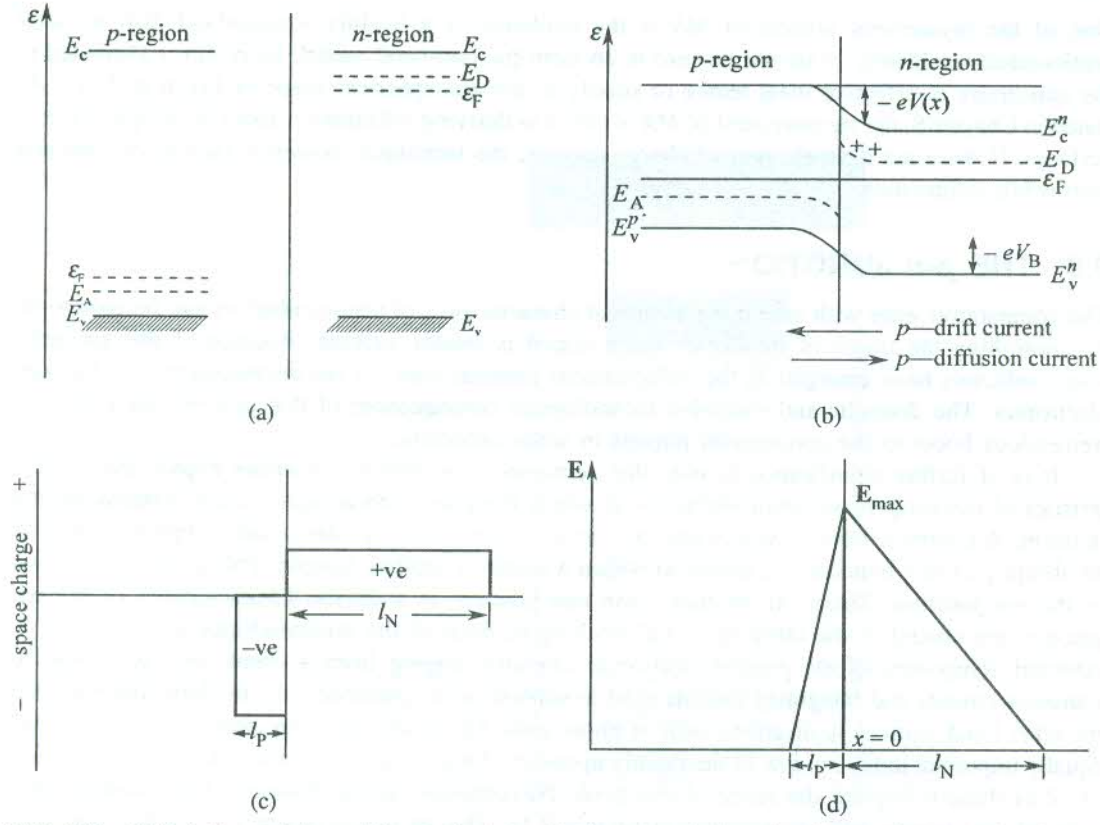


FIG. 9.15 (a) Relative positions of the conduction and valence band edges, the donor and acceptor levels; and the Fermi level in a p-n junction in the state of thermal equilibrium when p- and n-type regions are treated as isolated. (b) All the above positions when the p- and n-type regions are brought in contact. (c) The space charge distribution at a p-n junction. When the effect of the mobile carriers is neglected, i.e. $l_N N_D = l_P N_A$. (d) Variation of the space charge electric field **E** across the junction plane ($x = 0$).

current stops as soon as the field of ionized impurities becomes big enough just to keep the diffusion in complete check. The electric field in question is basically interpreted as the consequence of a macroscopic potential $V(x)$ varying over the transition region. In a one-dimensional model with an abrupt change from p-type to n-type at $x = 0$, the potential $V(x)$ is related to the space charge by the Poisson equation,

$$\frac{d^2V}{dx^2} = \frac{-e}{\epsilon_0 \epsilon_s}(N_D^+ + p_n - n_n) \qquad \text{for } 0 < x < l_N \text{ (n-type region)}$$

$$\frac{d^2V}{dx^2} = \frac{e}{\epsilon_0 \epsilon_s}(N_A^- + n_p - p_p) \qquad \text{for } -l_P < x < 0 \text{ (p-type region)} \tag{9.74}$$

where the subscripts n and p refer to concentrations in the n- and p-type regions, respectively, and $-l_P < x < l_N$ defines the space charge region.

The exact solution of (9.74) is almost impossible, since the carrier concentrations depend on position. To make the solution feasible, we simplify (9.74) by assuming that the electric field is strong enough to keep all mobile free carriers away from the space charge region. The variation of electric field **E** across the junction is shown in Fig. 9.15(d). It is maximum at $x = 0$. Thus the density of space charge is given by simply N_D^+ in the n-region and by N_A^- in the p-region. This approximation is good except at the boundary of the space charge region for a small current across the junction. In this assumption the space charge region is depleted of free carriers and, therefore, is also identified as the *depletion layer*. The relation (9.74) is solved under the boundary conditions:

$$\frac{dV}{dx} = 0 \text{ at } x = l_N \quad \text{and} \quad x = -l_P$$

with $\frac{dV}{dx}$ continuous at $x = 0$ and $V(l_N) - V(-l_P) = V_B$, where V_B is the height of the potential step, known as the *barrier height*. It is also called the *diffusion voltage*.

Well outside the space charge zone, N_D^+ and N_A^- are compensated by equally large free carrier concentrations n_n and p_p, respectively. In accordance with the type of doping, electrons serve as majority carriers in the n-type region and holes as majority carriers in the p-type region. There may always be present small concentration of holes in the n-region (p_n) and of electrons in p-region (n_p), contributed by the naturally occurring impurities in the host crystal.

An estimate of V_B is made in terms of carrier concentrations in the state of thermal equilibrium. These concentrations are given by (9.39a) and (9.40a) which we rewrite below:

$$n_n = N(\text{c}) \exp\left[\frac{-(E_c^n - \varepsilon_F)}{k_B T}\right] \tag{9.75a}$$

$$p_p = N(\text{v}) \exp\left[\frac{-(\varepsilon_F - E_v^p)}{k_B T}\right] \tag{9.75b}$$

Their relationship with the intrinsic concentration n_i is then given by

$$n_i^2 = n_n p_n = N(\text{v})\, N(\text{c}) \exp\left[\frac{-(E_c^n - E_v^n)}{k_B T}\right] \tag{9.76}$$

Here E_c^n, E_v^n, E_v^p are conduction band edge (in the n-region) and valence band edges (in the n- and p-regions), respectively [see Fig. 9.15(b)].

From (9.75b), we have

$$E_v^p - \varepsilon_F = k_B T \ln\left(\frac{p_p}{N(\text{v})}\right) \tag{9.77}$$

Analogically,

$$E_v^n - \varepsilon_F = k_B T \ln\left(\frac{p_n}{N(\text{v})}\right) \tag{9.78}$$

Subtracting (9.78) from (9.77) and using (9.76), we get

$$E_v^p - E_v^n = k_B T \ln\left(\frac{p_p n_n}{n_i^2}\right) \tag{9.79}$$

But V_B is equal to the difference between the maximum and minimum of the macroscopic potential $V(x)$. Hence,

$$eV_B = -(E_v^n - E_v^p) = k_B T \ln\left(\frac{p_p n_n}{n_i^2}\right) \qquad \text{[using (9.79)]}$$

or

$$V_B = \frac{k_B T}{e} \ln\left(\frac{p_p n_n}{n_i^2}\right) \tag{9.80}$$

Other important properties of a p-n junction are the maximum value of the electric field $\mathbf{E}_x$ and the space charge width $l(= l_N + l_P)$. The expressions for them can be obtained from (9.74).

When an external steady voltage is applied across a p-n junction, the barrier step height changes. In a forward-biased configuration† it decreases to value,

$$V_{BF} = V_B - V_{ext} \tag{9.81}$$

where V_{ext} denotes the applied voltage.

The 'flow of current changes' in the band edges and the Fermi level are depicted in Fig. 9.16(a). On the other hand, for a reverse-biased connection in contrast to the forward-biased case, the

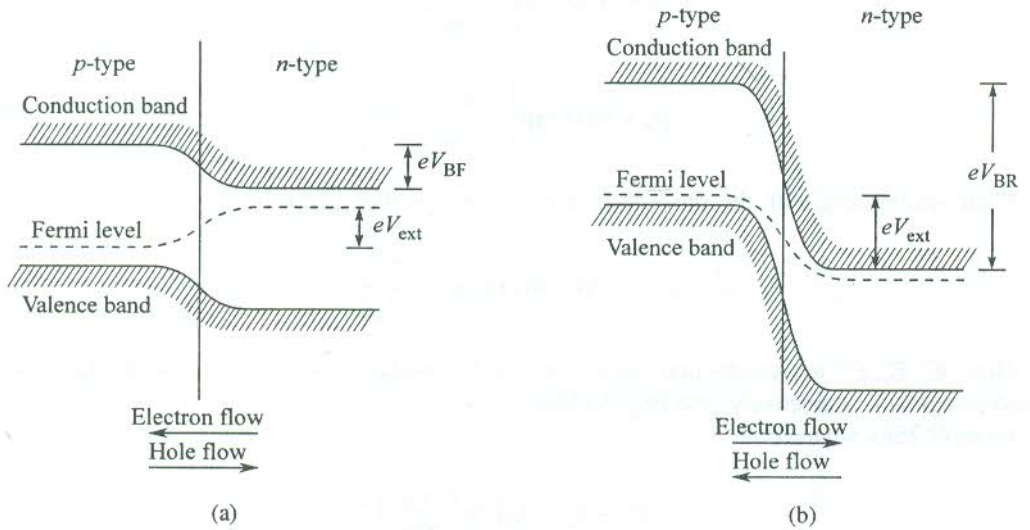


FIG. 9.16 Qualitative behaviour of the Fermi level and the electron potential energy at points near the p-n junction when (a) a forward bias of V_{ext} volt is applied, (b) a reverse bias of V_{ext} volt is applied. Note the direction of electron and hole currents in the two cases.

† *Forward biased.* The free end of the p-region is connected to the positive terminal and that of the n-region to the negative terminal of the voltage source.
Reverse biased. The polarities of the p- and n-free ends are opposite to those in the forward-biased arrangement.

direction of the applied field does not favour the flow of majority carriers across the junction, thus increasing the barrier height to

$$V_{BR} = V_B + V_{ext} \tag{9.82}$$

The electronic conduction and changes in the band scheme structure are illustrated in Fig. 9.16(b).

In the forward-biased configuration the current is carried by majority carriers and is of the order of a few mA which is much higher than that in the reverse-biased case (~ μA). The current in a reverse-biased junction is low as it is contributed by the flow of minority carriers that are relatively very small in number and are required to cross a higher potential barrier. The property that a *p-n* junction favours the flow of current in one direction across the junction forms the basis for using it as a rectifier. The knowledge of voltage-current characteristics of a *p-n* junction is crucial in order to understand the behaviour of the *p-n* junction for the desired application. The case of a lightly doped *p-n* junction is the subject matter of an elementary course in basic electronics. Hence it is omitted from here for brevity. An important device employing a heavily doped *p-n* junction is, however, treated below in Section 9.10.

9.10 EXAMPLES OF *p-n* JUNCTION-BASED DEVICES

Several excellent books are available on semiconductor devices. A couple of them considered suitable to meet our requirement are cited at the end of this chapter. It may again be reminded that there are numerous devices that make use of homogeneous semiconductors. Perhaps, the simplest of them is a thermistor which is used as a standard temperature measuring device. On the other hand, the *p-n* junction based devices have acquired such a tremendous technological potential that even their listing is quite a formidable task. As examples for discussion, our choice falls on (i) tunnel diode and (ii) injection laser, on the merit of their role in the development of modern electronics and photonics.

9.10.1 The Tunnel Diode

Tunnelling is a well-founded concept in quantum mechanics. The electron tunnelling at a *p-n* junction implies that electrons reach the other side of the junction by penetrating through the potential wall at the junction. It should not be confused with the act of crossing or jumping over the potential hill. It is purely a quantum statistical phenomenon based on the concept that the electron has a finite probability of being found on either side of the junction.

For tunnelling to occur, the width of the space charge layer (or the depletion layer) should be of the order of, or less than, the electron mean free path. The width of the space charge can be reduced to this value by doping both regions of the *p-n* junction heavily. A *p-n* junction that is doped heavily enough to make the tunnel current greater than the usual diffusion current under certain conditions is called a *tunnel diode*. The width of the space charge layer in these diodes is usually less than 100 Å. The impurity concentrations are in the range 10^{19}–10^{20} cm^{-3}, whereas in ordinary *p-n* junctions they vary between 10^{14} and 10^{18} cm^{-3}.

The material of a tunnel diode behaves as a degenerate semiconductor. The Fermi level ε_F no more lies in the gap. Instead, it lies within the valence band of the *p*-region and the conduction band of the *n*-region. We say that the Fermi level lies in a hybrid impurity-intrinsic band. This follows from arguments advanced in Section 9.6. Though a small width for the space charge layer is essential, it is not a sufficient condition for tunnelling. There must be an unoccupied state on the other side of the junction into which an electron could tunnel at constant energy.

Figure 9.17(a) shows the state of an unbiased junction where the Fermi level has the same energy in both the regions of the junction. Let us first examine tunnelling in a reverse-biased junction as shown in Fig. 9.17(b). When a small reverse bias voltage V_{ext} is applied, the height of the barrier step becomes much more than V_B. On account of this, the Fermi level has different values in the regions and at large values of V_{ext} a large number of occupied states of the valence band on the p-side of the depletion region lie opposite even to a larger number of empty states of the conduction band on n-side. The band picture is similar to that in a metal. When l_T is of the order of the electron wavelength or less, the junction has ideal conditions for electron tunnelling. At higher values of V_{ext} the tunnel current can be quite large because of the enhanced level of overlap between the occupied and empty states as referred above. The probability of tunnelling is given as

$$P_t \sim \exp(-kl_T) \tag{9.83}$$

where k is the electron wavevector.

For a simple band structure of geometry as shown in Fig. 9.17(b),

$$l_T \simeq \frac{E_g}{eE} \tag{9.84}$$

Substituting this value of l_T in (9.83), we get

$$P_t \simeq \exp\left[\frac{-E_g(2m^*\varepsilon)^{1/2}}{eE\hbar}\right] \tag{9.85}$$

Here E is an average value of the electric field at the junction which generally does not deviate much from E_{max}. Owing to the tunnel current, there is a quick onset of a critical reverse voltage. This property enables a reverse-biased tunnel diode to be used as a voltage regulator. These tunnel diodes are known as Zener diodes. The relation (9.85) provides a good description of tunnelling in a Zener diode.

Now, we proceed to describe the other important tunnel diode that is used in the forward-biased configuration [Fig. 9.17(c)]. This is again a heavily doped p-n junction, known as the Esaki diode. If one approaches the forward-biased state from the reverse-biased condition, by decreasing V_{ext}, the tunnel current is on decrease and becomes zero at $V_{ext} = 0$. The decrease in current occurs because of the reducing level of overlap at lower values of V_{ext} between the regions of occupied states in the valence band of the p-side and the unoccupied states in the conduction band of the n-side [Fig. 9.17(b)]. When unbiased ($V_{ext} = 0$) there is no current at $T = 0$. At non-zero temperatures in this condition almost an equal small number of electrons tunnel from both sides of the junction keeping the current at zero value. The V–I characteristics of a tunnel diode are drawn in Fig. 9.17(e). The portion AO refers to a reverse biased junction.

The characteristic curve beyond the point O describes the V–I relationship in an Esaki diode. As the forward voltage grows from zero to a certain value, the current keeps on increasing and approaches a maximum value. During this rise in the applied voltage, the Fermi level ε_F^p dips towards the conduction band edge of the n-side (E_c^n) and ε_F^n rises towards the valence band edge on p-side (E_v^p) as shown in Fig. 9.17(c). In the state of the maximum overlap of interest, the maximum tunnel current results. With a further increase in the applied voltage, the Fermi levels maintain the trend of dipping and rising in p- and n-regions, respectively. For the present band structure, the above change

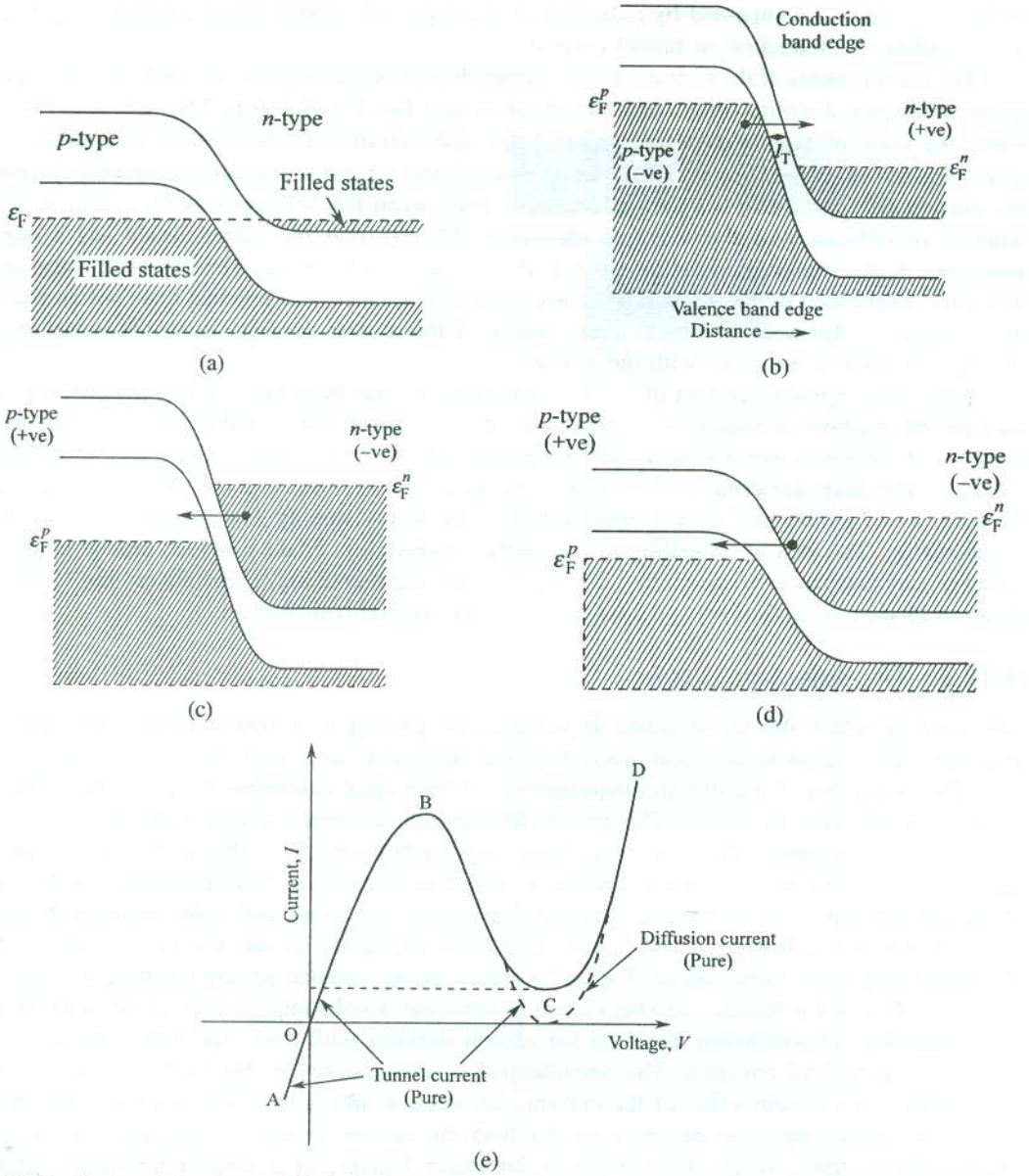


FIG. 9.17 (a) An unbiased tunnel diode in the state of thermal equilibrium. Because of heavy doping the Fermi level ε_F lies in the conduction band of the *n*-type region and in the valence band of the *p*-type region with a common value of energy. There is no current through the junction. (b) A reverse-biased tunnel diode with different values of Fermi energy on the two sides of the junction. A tunnel current flows across the junction because of the substantial overlap between the unoccupied states of the conduction band in the *n*-region and the occupied states of the valence band in the *p*-region. A tunnel diode in this configuration functions as a voltage regulator (the Zener diode). (c) A forward-biased tunnel diode (Esaki diode). The increase in tunnel current to a maximum on increasing the bias voltage is explained on the basis of the increasing degree of overlap between the occupied and unoccupied states as referred in case (b). (d) The decrease in tunnel current (Esaki current) on increasing the bias voltage further refers to the region of the negative resistance of the Esaki diode. The decrease in current is due to the decreasing degree of overlap of occupied and unoccupied states affected by the increasing applied voltage. (e) Voltage-current characteristics of a tunnel diode. The portion AO represents the reversed-biased state; OB—current increases when forward biased; BC—negative resistance region when forward biased; CD—purely diffusion current.

in Fermi levels is accompanied by reduction in the degree of overlap under question [Fig. 9.17(d)] and therefore the reduction in tunnel current.

The main features of the forward-biased tunnel diode (the Esaki diode) are that the tunnel current passes through a sharp maximum and then drops to zero [see Fig. 9.17(e)]. The current is maximum when the forward voltage is about $(\varepsilon_F^p + \varepsilon_F^n)/2e$ and vanishes at about twice this value. In the *V–I* characteristic curve [Fig. 9.17(e)], the portion OB represents the region of increasing current and the portion BC shows that the current decreases even when the voltage is being further increased. This is an indication of the negative resistance. We say that the tunnel diode has a negative resistance in this region. Beyond the point C there is no tunnel current. The observed rising current is totally contributed by the normal diffusion current. The current in practice does not drop to zero but a valley is observed. The most likely source of this excess current lies in the losses incurred during collisions of electrons with the lattice.

The negative resistance part of the *V–I* characteristics has been exploited to use an Esaki diode as a power amplifier or source. In conjunction with a capacitor and an inductance it can be made to work as an oscillator and a switch. Switching times are limited by high capacitance C_j of the thin junction. The times are usually of the order of nanosecond (10^{-9} s). The switching time RC_j can be brought down to the picosecond range (10^{-12} s) by heavy doping that lowers the resistance R considerably. The idea of tunnelling as originally conceived by Esaki provided crucial guidance to Giaever and Josephson in accounting for superconducting tunnelling. Esaki shared the 1973 Nobel prize in Physics with Giaever and Josephson for his original contribution.

9.10.2 The Injection Laser

The laser in which the lasing action is achieved by passing a current through a forward-biased *p-n* junction is called an *injection laser*. It is also referred to as a laser diode. The injection lasers have the distinction of meeting the requirements of fibre optic communication systems in the most convenient and effective manner. This has created tremendous interest in their study and development.

The principle behind the emission of light from a semiconductor is that of the recombination of electrons and holes at a *p-n* junction when a current is injected through the diode. As the current is passed through a forward-biased junction, the injected electrons and holes increase the density of electrons in the bottom of the conduction band and holes at the top of the valence band, simultaneously in the same region of space. A spontaneously emitted photon resulting from electron-hole recombination is fedback into the system by cleaving or polishing the ends of the junction diode. Under the state of population inversion the photon interacts with electrons in the conduction band to produce stimulated emission. The amplification is accomplished by the multiple reflection at the diode ends. At a certain value of the current, the electron and hole densities are so large that the rate of stimulated emission becomes greater than the rate of absorption and thus the process of amplification starts. When the current is increased further, at a certain threshold value the amplification exceeds the cavity losses and a coherent radiation is obtained in the output.

The threshold current density in the *p-n* junction based lasers is inconveniently high (~50,000 A cm^{-2}). This is drastically reduced in practical lasers that employ a double heterojunction. A heterojunction has two different semiconductor materials on its two sides. In a double heterojunction a lasing semiconductor is sandwiched between two wider-gap semiconductors of opposite doping. In a heterojunction laser developed by Kresel and Butler*, the active layer is an undoped

* H. Kressel and J.K. Butler, *Semiconductor Lasers and Heterojunction LEDs* (Academic Press, 1977).

$Ga_{1-x}Al_x$ As ($x = 0.05 - 0.1$) which is embedded between n- and p-type layers of Ga_{1-y} Al_yAs ($y = 0.3 - 0.4$). The n- and p-type layers have the same band gap and the same refractive index. The band gap of the active layer is lower and characterized by a higher value of the refractive index. This ensures that the radiation emitted in the active layer is confined within this layer because of the total internal reflection occurring at its interfaces with the n- and p-type materials. Furthermore, the electrons from the n-side and the holes from the p-side that enter the active layer are not allowed by the potential barriers at the junctions to escape into the p- and n-sides (regions), respectively. This confinement of the light and the carriers results in amplification and wave guidance. For the high efficiency, materials only with the direct band gap are selected for fabrication. One of the biggest advantages with this heterostructure is that the frequency of the laser output can be finely monitored by varying the composition of the compound. The reader interested in complete details may consult the book authored by Chai Yeh,* and that by Ghatak and Thyagarajan.**

For the present heterostructure there are two quasi-Fermi levels ε_F^c and ε_F^v referred to the conduction and valence band edges. The ε_F^c in the optically active layer is on level with the ε_F^n in the n-region and ε_F^v with ε_F^p in the p-region. The population inversion is achieved by applying a forward voltage that exceeds the voltage equivalent of the band gap of the active layer. Once the electrons and holes enter the active layer from n- and p-type regions respectively, they remain confined in the active layer as explained in the preceding paragraph. The emission of radiation occurs in the plane of the active layer when an electron from the conduction band combines with a hole in the valence band as shown in Fig. 9.18. The process of stimulated emission and amplification proceeds the same way as in a p-n junction based device. The frequency of the emitted radiation is temperature dependent. A typical GaAs laser emits in the near infrared around 8380 Å.

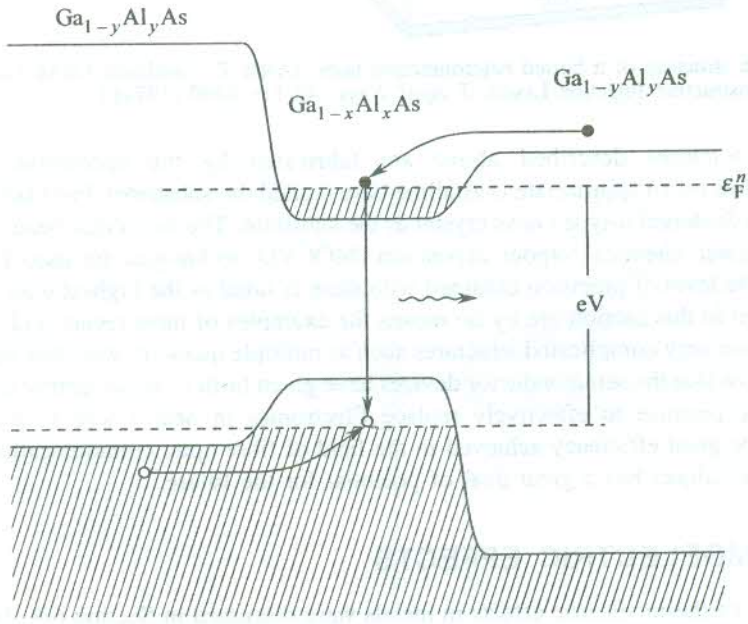


FIG. 9.18 Double heterojunction GaAs-Al laser. The optically active layer has a higher refractive index (lower band gap) with $x = 0.05 - 0.1$. The material of p- and n-type regions has a wider band gap (lower refractive index) with $y = 0.3 - 0.4$. The confinement to carriers in the active layer is provided by the potential barriers at the two junctions.

* Chai Yeh, *Handbook of Fiber Optics: Theory and Applications* (Academic Press, 1990).

** A.K. Ghatak and K. Thyagarajan, *Optical Electronics* (Cambridge, 1989).

In the above device structure the carriers and the light are confined only along one direction. The confinement can also be provided in the lateral direction by surrounding the optically active layer with higher band gap materials on all sides, leaving a window for the output. Such a heterostructure laser is known as the *buried heterostructure laser* and shown in Fig. 9.19. The required threshold current densities (2000–4000 A cm^{-2}) are achieved in this structure for a significantly reduced value of current (< 50 mA), whereas for a stripe-geometry laser with a standard size cavity the threshold current shoots a little over one ampere.

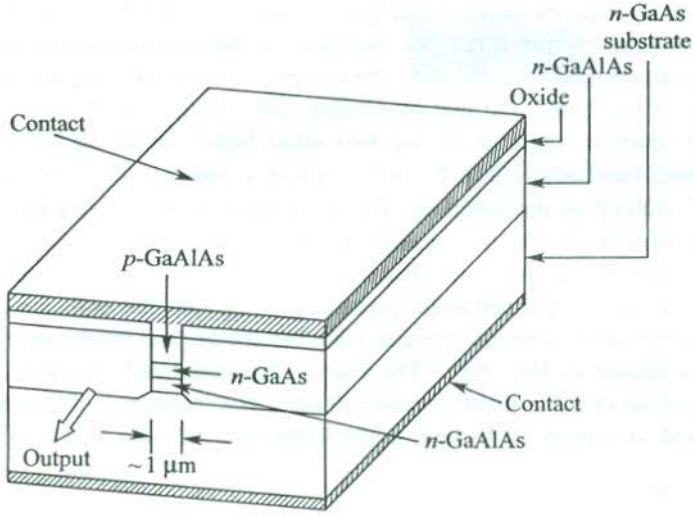


FIG. 9.19 Design structure of a buried heterojunction laser. [After T. Tasukada, GaAs–$Ga_{1-x}Al_xAs$ Buried Heterostructure Injection Lasers, *J. Appl. Phys.*, **45**(11), 4899 (1974).]

The heterostructures described above are fabricated by the successive deposition of multiepitaxial layers on an appropriate crystalline base (called the substrate). For GaAs laser, Kressel and Butler used a Si-doped *n*-type GaAs crystal as the substrate. The molecular beam epitaxy (MBE) and the metal organic chemical vapour deposition (MOCVD) techniques are used for the epitaxial growth because the level of precision obtained with them is rated as the highest toady. The injection lasers as described in this section are by no means the examples of most recent and versatile lasers, some of which have very complicated structures such as multiple quantum well structures. Our object was limited to show that the semiconductor devices have given birth to an altogether new field (called *Photonics*) whose promise to effectively replace Electronics in near future to a large extent is exemplified by the great efficiency achieved in the field of fibre optic communication. Though still in its infancy, the subject has a great deal of potential for the future.

9.11 THERMOELECTRIC EFFECTS

Essential features of thermoelectric effects in metals were discussed in Section 6.8. The treatment is being extended here to semiconductors whose thermoelectric response is many times stronger. The three main thermoelectric parameters, namely the thermoelectric power S, the Thomson coefficient μ_T and the Peltier coefficient Π in semiconductors may be positive or negative, depending on whether

the semiconductor is of *p*-type or *n*-type. We may recall that these parameters can have only negative sign in ideal conductors. Consider a semiconductor rod whose one end is hot and the other cold. As far as the algebraic sign is concerned, the case of an *n*-type semiconductor is similar to that of ideal conductors. But electrons at the hot end in a *p*-type semiconductor occupy acceptor levels on being excited from the valence band. This enables electrons at the cold end near the top of the valence band to lower their energy by moving into holes created in the valence band at the hot end. Thus the hot end becomes negatively charged and the cold end positively charged. This situation is opposite to that described in Section 6.8. Hence in a *p*-type semiconductor the parameters S, μ_T and Π are all positive. These arguments lead to a quick method of determining whether a specimen is of *n*-type or *p*-type. All that is needed is to measure the sign of the voltage across the given specimen whose one end is hot.

Now, we first consider the above-mentioned semiconductor rod in an open circuit, ensuring that a constant temperature difference is maintained between its ends. In this condition a constant difference of potential is observed between its end. This potential difference can be interpreted to be associated with an electric field,

$$E = \frac{dV}{dx} = \frac{dV}{dT}\frac{dT}{dx} = S\frac{dT}{dx} \tag{9.86}$$

Next, we consider a semiconductor rod kept at a constant temperature in a closed circuit. If an electric current is forced into it at one end from a certain external source, the current density in an *n*-type semiconductor may be written as

$$j_e = n(-e)(-\mu_n)E = ne\mu_n E \tag{9.87}$$

where μ_n is the electron mobility in the direction opposite to that of the electric field E from the external source.

It is appropriate to measure the electron energies relative to the Fermi energy ε_F because the two materials in contact have a common Fermi level. Therefore, the average energy transported by electrons is equal to

$$(E_c - \varepsilon_F) + \frac{3}{2}k_B T \tag{9.88}$$

where E_c denotes the energy at the conduction band edge.

The heat current density j_Q associated with the electric current density j_e is given by

$$j_Q = -n\left(E_c - \varepsilon_F + \frac{3}{2}k_B T\right)\mu_n E \tag{9.89}$$

But we know from (6.94) that

$$j_Q = \Pi_e j_e$$

Therefore, the Peltier coefficient of an *n*-type semiconductor is

$$\Pi_e = -\frac{E_c - \varepsilon_F + \frac{3}{2}k_B T}{e} \tag{9.90}$$

Similarly, for a *p*-type semiconductor, we have

$$j_h = pe\mu_p E \tag{9.91}$$

$$j_Q = p\left(\varepsilon_F - E_v + \frac{3}{2}k_B T\right)\mu_p E \tag{9.92}$$

and

$$\Pi_h = \frac{\varepsilon_F - E_v + \frac{3}{2}k_B T}{e} \tag{9.93}$$

The Peltier coefficient can be easily determined by measuring the absolute thermoelectric power to which it is linked by the Kelvin relation (6.96), $\Pi = ST$. The Π versus T plots for *n*-type and *p*-type specimens of silicon are displayed in Fig. 9.20. The significant observation is that the specimens behave as intrinsic semiconductors above 600 K.

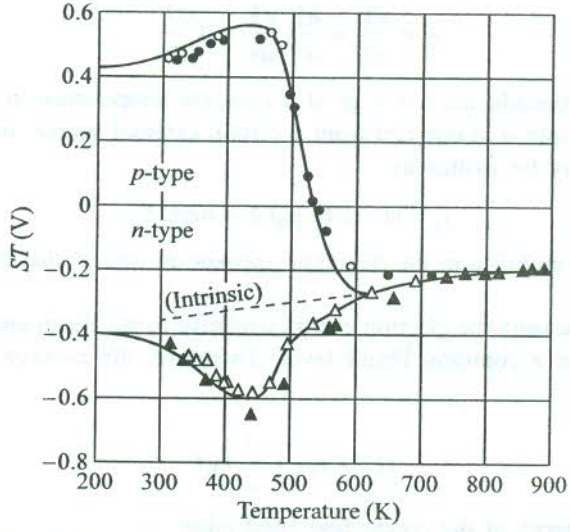


FIG. 9.20 Variation of the Peltier coefficient (= ST, with S as the absolute thermoelectric power) of *n*-type and *p*-type specimens of silicon as a function of temperature. The two specimens show intrinsic behaviour above 600 K. The curves are based on theoretical estimates. The points show the experimental data. [After T.H. Geballe and G.W. Hall, *Phys. Rev.*, **98**, 940 (1955).]

Being many times intense than in metals, the thermoelectric effects in semiconductors are of added interest. For example, the thermoelectric powers of semiconductors are two orders of magnitude larger than those for non-ferromagnetic metals. The reasons ascribed to this huge quantitative difference in response are mainly two-fold. Firstly, the carrier density of semiconductors is sensitive to temperature owing to which the hot end has more conduction electrons or holes per unit volume, depending on whether the material is of *p*-type or *n*-type. Secondly, and more importantly the presence of a forbidden energy gap in semiconductors proves to be a major cause of the observed behaviour.

To elaborate on the second reason we examine the quantity of Peltier heat evolved or absorbed at a metal semiconductor junction. Let a current be flown from a metal to an n-type semiconductor in contact forming a junction as shown in Fig. 9.21(a). In an n-type semiconductor the majority carriers (electrons) are in the conduction band placed well above the Fermi level ε_F. On the other hand, most of the electrons in a metal are near the Fermi level. Because of being in contact, the Fermi energy would be at the same level in the metal and the semiconductor. Figure 9.21(a) clearly indicates that the average energy of conduction electrons in the electric current decreases as the same electric

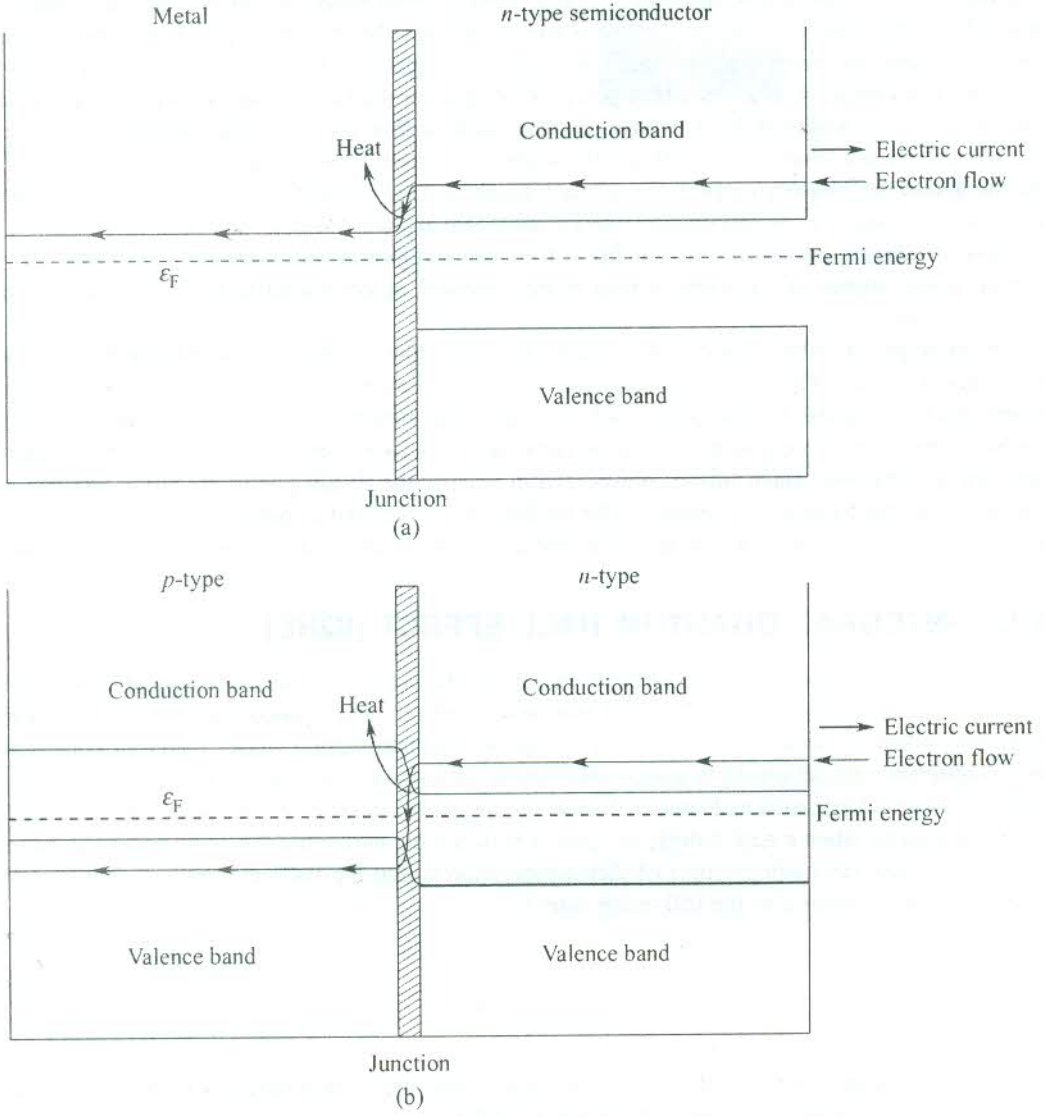


FIG. 9.21 (a) Peltier effect at a metal–semiconductor junction. Heat is evolved at the junction when an electric current flows from a metal to an n-type semiconductor. (b) Peltier effect at a p-n junction. Relatively large amount of heat is evolved at the junction (compared to that in (a)) in the case of an electric current flowing from the p-region to the n-region.

current enters the semiconductor. The difference of electron energies in the two regions is lost in the form of heat at the junction. Similarly, when a *p*-type material is in contact with a metal, heat would be absorbed for the shown direction of electric current because the majority carriers (holes) in the semiconductor would occupy energy states in the valence band that are much below the Fermi level (the approximate level of electric current in metal).

Application

It is obvious from the above discussion that a thermocouple junction can function as a heat pump or a refrigerator when an electric current is forced through the thermocouple. Under the flow of an electric current the thermocouple pumps heat from one junction to the other. We show below that a *p-n* junction can serve as a better heat pump or refrigerator than any common thermocouple junction consisting of two semimetals or even a junction such as that shown in Fig. 9.21(a).

Consider a *p-n* junction through which an electric current flows as demonstrated in Fig. 9.21(b). On the *n*-side the majority carriers are at much higher energy levels than the Fermi level whereas on the *p*-side the majority carriers occupy energy states that are much below the Fermi level. Hence, with the flow of electric current in this condition from the *n*-region to the *p*-region a relatively far more energy would appear in the form of heat at the junction, as compared to that in a junction shown in Fig. 9.21(a).

In principle, a Peltier heater can be more efficient than an electrical resistance heater. For the requirement of each kW of heat, an electrical resistance heater must consume one kW of electrical power that is dissipated. On the other hand, an ideal Peltier heater needs electrical power only to pump the heat energy from one junction (cold) to the other (hot) like a refrigerator or heat engine and thus operates at a relatively much smaller power. But in practice the Peltier pumps are not as efficient as the conventional mechanical heat pumps. The inefficiency is attributed partly to the heat loss through the thermocouples from the hot to the cold side and partly to the Joule heating of the thermocouples.

9.12 INTEGRAL QUANTUM HALL EFFECT (IQHE)

Although the treatment of Hall effect given in Section 9.8 is based on purely classical considerations, it gives a good account of the electrical transport in metals and semiconductors. But the classical magnetoconducting scenario undergoes a spectacular transformation under quantum conditions of temperature and magnetic field in a two-dimensional conductivity channel. K. von Klitzing, Dorda and Pepper* observed that such a channel is formed at the oxide interface in a metal-oxide-semiconductor (MOS) transistor when a gate voltage is applied between the metal and the semiconductor, as shown in Fig. 9.22. The fascinating aspect of their observation is that the Hall resistance ρ_H varies with the magnetic field according to the following rule:

$$\rho_H = \frac{h}{ie^2} \tag{9.94}$$

where i is an interger (= 1, 2, 3, . . .).

The phenomenon expressed by this rule, where the Hall conductance is quantized in units of e^2/h, is called the *Integral Quantum Hall Effect* (IQHE).

* K. von Klitzing, G. Dorda and M. Pepper, *Phys. Rev. Lett.*, **45**, 494 (1980).

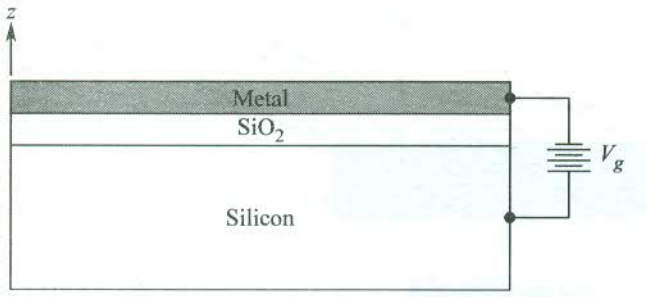


FIG. 9.22 A metal-oxide-semiconductor (MOS) transistor. An accumulation of charge carriers occurs at the oxide interface on the application of a gate voltage between the metal and the semiconductor. The oxide interface (the *xy*-plane) behaves as a two-dimensional conductivity channel.

In the original experiment a constant current of 1 μA was forced to flow between the source and the drain in the presence of a magnetic field of 18 tesla at 1.5 K. The results of this experiment are graphically displayed in Fig. 9.23 and the experimental geometry is illustrated in Fig. 9.24(a). The Hall field versus magnetic field plot in this case is not a straight line (as in the classical Hall effect). Instead, it shows plateaus with steps in between at certain values of the magnetic field [see Fig. 9.24(b)].

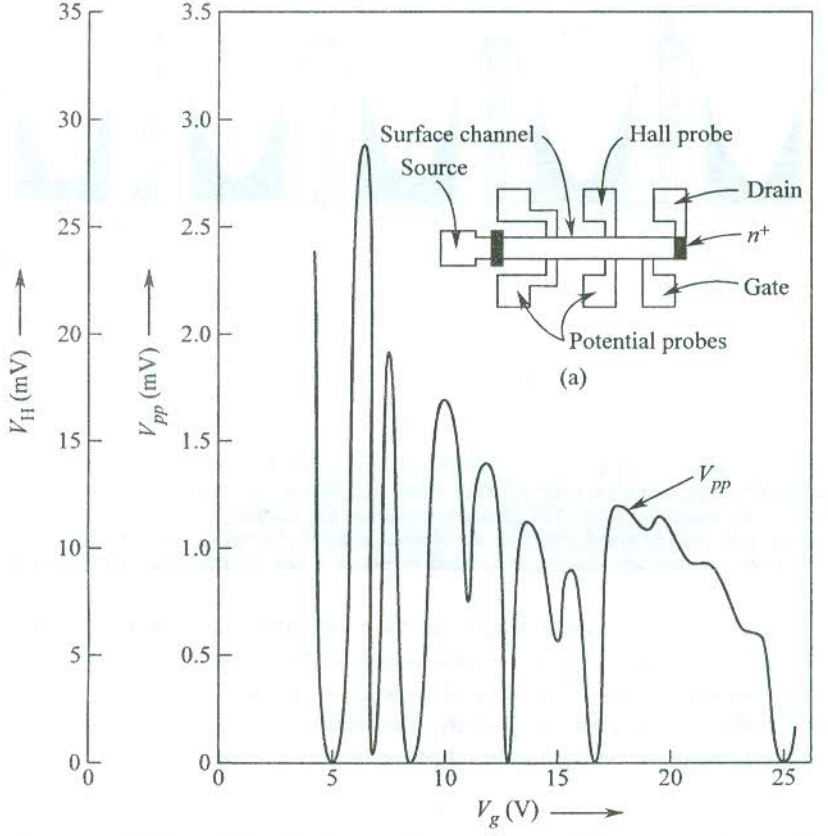


FIG. 9.23 The IQHE voltage V_{pp} and the Hall voltage V_H as functions of the gate voltage V_g. The MOS transistor set-up for the IQHE measurements is shown above the plots.

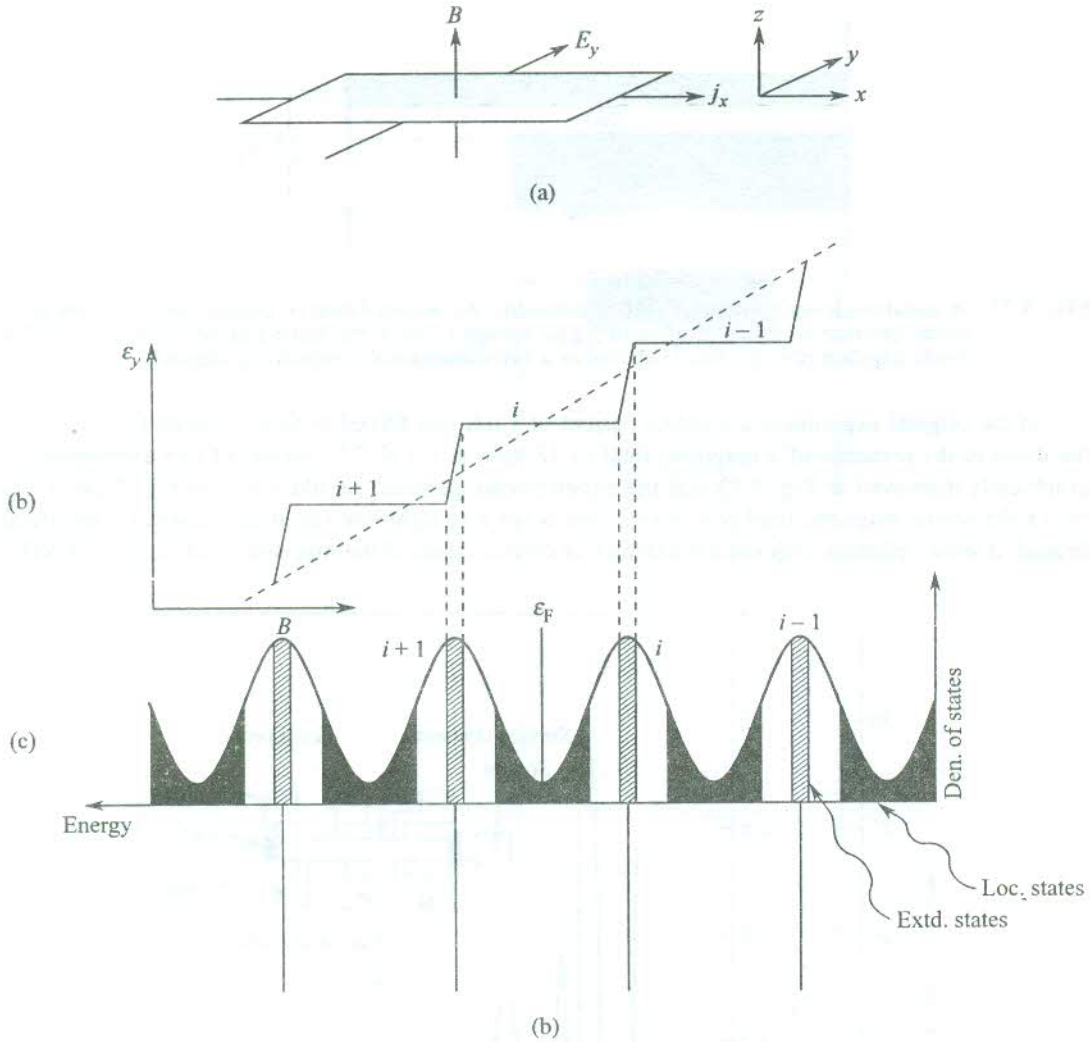


FIG. 9.24 (a) The IQHE geometry. (b) The IQHE plateaus in the Hall field for a fixed system current j_x. The dotted line represents the classical Hall effect. (c) Density of states in a two-dimensional real system in a strong magnetic field. The disorder broadens the Landau levels with the extended states in the middle and the localized states in the shaded regions, known as the mobility gaps. Vertical lines shown below indicate the sharp Landau levels in a two-dimensional ideal system.

First we show how the observed minima in the longitudinal voltage V_{pp} can be explained and relation (9.94) be derived, using only a crude model. Then, we give a qualitative analysis of real two-dimensional systems in the framework of a general theory.

Consider a surface current density $\mathbf{j}_x$ in the x-direction defined as the current crossing a line of unit length in the y-direction on the oxide interface (the xy-plane):

$$I_x = j_x L_y \tag{9.95}$$

If n is the total electron density per unit area of the xy-plane,

$$j_x = nev_d \tag{9.96}$$

with v_d as the drift velocity of electrons in the x-direction.

In an alternative description of the Hall effect, the current in a specimen with mobile charge carriers is produced when the specimen is placed in a region of mutually crossed (perpendicular) electric and magnetic fields. The flow of current in a direction orthogonal to both the fields is detected on closing the circuit. If the electric field ($\mathbf{E}_y$) and the magnetic field $\mathbf{B}$ act along y- and z-directions, respectively

$$\mathbf{j}_x = \sigma_{xx}\mathbf{E}_x + \sigma_{xy}\mathbf{E}_y = \sigma_{xy}\mathbf{E}_y \quad (\text{since } \mathbf{E}_x = 0) \tag{9.97}$$

and

$$v_d = \frac{E_y}{B} \tag{9.98}$$

Here σ_{xy} denotes the conductivity tensor in the plane of the two-dimensional channel. Interpreting the resistivity ρ_{xy} as the Hall resistance ρ_H,

$$\rho_H = \frac{V_H}{I_x} = \frac{E_y L_y}{j_x L_y} = \frac{B}{ne} \tag{9.99}$$

As follows from relation (9.95), ρ_H represents the resistance of a channel of unit thickness. In the commonly used geometry for the Hall effect, a current is flown along the x-direction in the presence of a magnetic field along the z-direction and the Hall voltage V_H measured along the y-direction in the specimen. The description contained in relation (9.99) is consistent with this experimental geometry.

It is now required to exploit relation (8.38) which states that the areas of the successive electron orbits in the k-space in the presence of a magnetic field B differ by an amount $2\pi eB/\hbar$. Considering a square of side L on the xy-plane, the number of states in this area are estimated as

$$\left(\frac{2\pi eB}{\hbar}\right)\left(\frac{L}{2\pi}\right)^2 = \frac{eL^2B}{h} \tag{9.100}$$

It gives the number of electron levels that coalesce into a single magnetic level as soon as even a small magnetic field is switched on. This magnetic level defines a Landau level whose energies in the xy-plane are quantized as $(i + 1/2)\hbar\omega_c$ with $\omega_c = eB/m^*$ (the cyclotron frequency). Relation (9.100) gives essentially the measure of degeneracy of a Landau level. In the present context we define the degeneracy per unit area of the xy-plane as

$$D(B) = \frac{eB}{h} \tag{9.101}$$

Let the applied magnetic field be so strong that $\hbar\omega_c >> k_B T$. It is then reasonable to talk in terms of completely filled or completely empty Landau levels. Suppose B_i is the critical field at which no level is partly filled and i is the magnetic quantum number of the highest occupied level. When the

electron density on the oxide interface is adjusted by varying the gate voltage so that the Fermi level coincides with level i,

$$n = i \cdot D(B_i) \tag{9.102}$$

Under these conditions electrons can undergo neither elastic nor inelastic collisions. The elastic collisions would involve scattering of electrons from one state to the other in the same Landau level. But this is not permitted by the Pauli principle, since all possible final states of equal energy are occupied. The inelastic collisions can be possible with the scattering of electrons to a vacant Landau level by absorbing the required energy from some source, most likely phonons. But in the experimental conditions of low temperature and $\hbar\omega_c >> k_B T$ as established here, there are hardly any phonons whose energy could compare with the large energy interval $\hbar\omega_c$. Therefore, the inelastic collisions are too ruled out and the electron mean free paths are greatly enhanced. This results in the occurrence of the voltage minima in V_{pp} (or the longitudinal resistance minima).

Placing the value of $D(B)$ from (9.101) in (9.102), we obtain (9.94):

$$\rho_H = \frac{h}{ie^2} \text{ (in ohms)}$$

Analysis in real systems

The experimental evidence for the Hall resistance being accurately quantized at h/ie^2 ohms might apparently suggest that IQHE is independent of purity and crystallinity, simply because the theory predicting this quantization does not take these aspects into consideration. But the presence of impurities or microcrystallinity produces disorder, rendering the crystal potential irregular as a result of which the sharp levels in ideal systems are broadened into bands in real systems [Fig. 9.24(c)]. This affects the Hall resistance to such an extent that its linear variation changes to develop plateaus. In a two-dimensional system that concerns the present discussion all the electron states are predicted as localized at any disorder.* To the credit of this prediction the IQHE actually approaches this limit as the magnetic field goes to zero. Therefore, it is logical to believe that there can exist both the extended and localized types of carrier states in a band. As per the latest concept of localization, the extended and localized states cannot coexist at the same energy. The localized states occupy the region of the lowest density of states forming the mobility gap and do not contribute to the flow of electronic current. The extended states, on the other hand, appear around the peaks of the density of states [see Fig. 9.24(c)].

In light of his thought experiment on a two-dimensional system, Laughlin** has interpreted the IQHE in real systems as a consequence of the principle of gauge invariance. By analogy with the flux quantization in a superconductor (where the unit of charge is $2e$), the flux quantization in the IQHE (with e as the unit of charge) is discussed. For a certain increase δB in the magnetic field, there is an addition of one flux quantum that enhances the degeneracy of each Landau level by one. Suppose i denotes the magnetic quantum number of the highest completely filled Landau level. If all electrons cannot be accommodated up to this level at absolute zero, the Fermi level ε_F will coincide with the level $(i + 1)$ which is only partially filled.†

---

* E. Abraham, P.W. Anderson, D.C. Licciardello and T.V. Ramakrishnan, *Phys. Rev. Lett.*, **42**, 673 (1979).

** R.B. Laughlin, *Phys. Rev.*, **B23,** 5632 (1981).

† The statement is valid for temperatures (~1 K) at which the IQHE experiment is conducted.

With an increase of one flux quantum, each of the Landau levels will have one additional level in its subband. Consequently, the level $(i + 1)$ will be vacated, since the electrons in this level can now be accommodated in the newly created lower energy levels (i in total). Over a fixed range δB (equivalent to one flux quantum), during which the Fermi level remains in a level, a plateau in the Hall resistance is observed. The increase δB equals a magnitude for which at the stage under consideration all the states in the level $(i + 1)$ are vacated and the Fermi level drops to coincide with the level i. Thus for each increase of δB, a jump to the next plateau takes place. The next plateau refers to a level with the next lower magnetic quantum number [see Fig. 9.24(c)].

It is important to assert the role of disorder and localization in the IQHE for real systems. Given next is a brief discussion conducted in this approach.

The Hall electric field in an ideal system is

$$E_y(B) = j_x B_z R_H = v_d B_z = \frac{D(B)\, v_d h}{e} \tag{9.103}$$

where R_H denotes the Hall coefficient.

In a disordered system the degeneracy of states and the electron densities may be split as

$$D(B) = D^E(B) + D^L(B) \tag{9.104}$$

$$n = n^E(B) + n^L(B) \tag{9.105}$$

where E and L refer to the extended and localized states, respectively.

Then, the Hall field in a disordered system may be expressed as

$$\varepsilon_y(B) = \frac{D^E(B)\, v_d(B)\, h}{e} \tag{9.106}$$

where $v_d(B)$ stands for the drift velocity in the disordered system.

As observed in a typical IQHE experiment, the current density j_x carried by electrons in the extended states must remain unchanged at the value nev_d. Hence,

$$v_d(B) = \left(\frac{n}{n^E(B)}\right) v_d \tag{9.107}$$

It implies that n^E electrons per unit area carry the current (of density nev_d) with a higher drift velocity to compensate for the loss of current because of the localization of n^L electrons per unit area.

The $D^E(B)$ in a particular band always increases with B, though non-monotonically. It increases by one, only when a δB-increase of B creates an extended state in this band. We can check that this occurs with probability $1/(\nu + 1)$, where

$$\nu = \frac{D^L(B)}{D^E(B)}$$

From (9.98) we see that $\varepsilon_y(B)$ can remain unchanged (forming a plateau) with B whenever $v_d(B)$ decreases, provided

$$D^E(B)\; v_d(B) = \text{a constant} \tag{9.108}$$

But it should be observed that $v_d(B)$ increases as well as decreases with B depending on where the Fermi level ε_F is located. When ε_F lies in the mobility gap, we will have either (i) $v_d(B + \delta B) < v_d(B)$, when an extended state is produced and D^E increases in the subbands below ε_F or (ii) $v_d(B + \delta B) = v_d(B)$, in between the above events where the Hall field $\varepsilon_y(B)$ remains unchanged and the plateau occurs.

On the other hand, when ε_F falls in the band of extended states we always have $v_d(B + \delta B) > v_d(B)$ because D^E decreases on account of the downward movement of ε_F. For a complete treatment, the reader is referred to calculations made elsewhere.* The calculations show that a plateau in the Hall field $\varepsilon_y(B)$ is formed within an accuracy of a few parts in $10^6 D(B)/(\nu + 1)$.

In the extreme quantum conditions, i.e. at extremely low temperatures and extremely high magnetic fields a QHE has been observed** for fractional values of i in relation (9.94). In this limit the lowest Landau level is only partially occupied and the IQHE is not expected. Some of the fractional values of i for which the Hall resistance has been observed to be quantized are $\frac{1}{3}, \frac{2}{3}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}$ and $\frac{2}{7}$. At these occupancies the longitudinal Hall resistance ρ_{xx} is found to vanish. This phenomenon is called the *Fractional Quantum Hall Effect* (FQHE).

SUMMARY

1. The electron concentration n_i and the hole concentration p_i in an intrinsic semiconductor are given by

$$n_i = p_i = 2\left(\frac{k_B T}{2\pi\hbar^2}\right)^{3/2} (m_e^* m_h^*)^{3/4} \exp\left(-\frac{E_g}{2k_B T}\right)$$

2. The electrical conductivity of intrinsic semiconductors depends on temperature according to

$$\sigma_i \propto \exp(-E_g/2k_B T)$$

3. The Fermi energy ε_F of a semiconductor is expressed as

$$\varepsilon_F = E_v + \frac{1}{2} E_g + \frac{3}{4} k_B T \ln\left(\frac{m_h^*}{m_e^*}\right)$$

The Fermi level is centred in the forbidden energy gap, when $m_e^* = m_h^*$.

4. In an extrinsic semiconductor,

$$n = 2\left(\frac{m_e^* k_B T}{2\pi\hbar^2}\right)^{3/2} \exp\left[\frac{-(E_c - \varepsilon_F)}{k_B T}\right]$$

---

* Vipin Srivastava, *Found. Phys Lett.*, **11(6)**, 561 (1998).

** D.C. Tusi, H.L. Stormer and A.C. Gossard, *Phys Rev. Lett.*, **48**, 1559 (1982).

$$p = 2\left(\frac{m_h^* k_B T}{2\pi\hbar^2}\right)^{3/2} \exp\left[\frac{-(\varepsilon_F - E_v)}{k_B T}\right]$$

where E_v and E_c denotė the valence band and conduction band edges, respectively.

5. The electrical conductivity of a semiconductor is

$$\sigma = |e|\,(n\mu_n + p\mu_p)$$

where μ_n and μ_p are the electron and hole mobilities, respectively.

6. The magnetoresistance is defined as

$$\frac{\Delta\rho}{\rho(0)} = \frac{\rho(B) - \rho(0)}{\rho(0)}$$

Here $\rho(B)$ is the resistivity in the presence of a magnetic field **B** and $\rho(0)$ is the zero-field resistivity. In the case of intrinsic materials

$$\frac{\Delta\rho}{\rho(0)} = \omega_c^2\tau^2 \qquad \text{with } \omega_c = \frac{eB}{m^*}$$

where τ is the relaxation time of carriers.

7. In a heterojunction (p-n), p and n regions are derived from different semiconductor materials.

8. The Hall resistance in IQHE is

$$\rho_H = \frac{h}{ie^2}, \text{ where } i = 1, 2, 3, \ldots$$

PROBLEMS

9.1 Assuming the resistivity of an intrinsic Ge crystal as 47 ohm cm and electron and hole mobilities as 0.39 and 0.19 $m^2\ V^{-1}\ s^{-1}$ respectively, calculate its intrinsic carrier density at room temperature.

9.2 A p-type semiconductor has an acceptor density of 10^{18} cm^{-3}. If $E_D - E_v = 0.2$ eV and $m_e^* = m_h^* = m$ (the free electron mass),

(a) show that the intrinsic conduction in the crystal is negligible at room temperature;

(b) estimate the conductivity of the crystal at room temperature, assuming the hole mobility to be equal to 0.01 $m^2\ V^{-1}\ s^{-1}$ at room temperature.

9.3 Plot the intrinsic carrier concentration as a function of $1/T$ on a semilog graph paper over 10–500 K for:

(a) GaAs ($E_g = 1.5$ eV, $m_e^* = 0.1m$, $m_h^* = 0.4m$)

(b) InSb ($E_g = 0.22$ eV, $m_e^* = 0.013m$, $m_h^* = 0.18m$)

For which electron and hole concentrations is the conductivity of a semiconductor minimum? To which value of the net impurity content $|N_D - N_A|$ does this correspond?

9.4 In an InSb crystal, $E_g = 0.23$ eV: the dielectric constant $(\epsilon_s) = 18$ and $m_e^* = 0.015m$. Calculate (i) E_d and (ii) the radius of the ground state orbit. What should be the minimum donor concentration so as to produce an appreciable overlap between the orbits of adjacent impurity atoms?

9.5 The effective mass of electrons at the lower conduction band edge of a semiconductor is three times higher than that of holes at the upper valence band edge. How far is the Fermi level located from the middle of the forbidden energy gap, assuming that the semiconductor is intrinsic? Explain why E_g should be greater than $8k_BT$ for your calculation.

9.6 Calculate $N(c)$ and $N(v)$ in silicon at room temperature. In the conduction band of silicon there are six minima with $m_l^* = 0.97m$, $m_t^* = 0.19m$. The valence band can be approximated by two independent bands coincident at the band edge with $m_1^* = 0.5m$ and $m_2^* = 0.16m$.

9.7 In a two-dimensional intrinsic semiconductor the density of states is defined as

$$D_c(\varepsilon) = \frac{4\pi m_e^*}{h^2}, \qquad \varepsilon > E_c$$

$$D(\varepsilon) = 0, \qquad E_v < \varepsilon < E_c$$

$$D_v(\varepsilon) = \frac{4\pi m_h^*}{h^2}, \qquad \varepsilon < E_v$$

Use Fermi–Dirac statistics without making any approximation to show that

$$\varepsilon_F = E_c - \frac{1}{2}E_g + \frac{k_BT}{2}\ln\frac{8}{3} - k_BT \ln\cos\frac{\phi}{3}$$

with $\phi = \tan^{-1}\left[\left(\frac{32}{27}\right)\exp\left(-E_g/k_BT\right) - 1\right]^{1/2}$

9.8 A semiconductor cuboid crystal (1 cm × 5 mm × 1 mm) has a resistivity of 12.5 ohm cm. A Hall voltage of 5 mV across the 5 mm width is measured when the applied magnetic field is 2000 gauss and a current of 1 mA flows along the length of the crystal. Calculate the carrier concentration and Hall mobility in the crystal.

9.9 Hall measurements are made on a p-type semiconductor bar 500 μm wide and 20 μm thick. The Hall contacts A and B are displaced 2 μm with respect to each other in the direction of current flow of 3 mA. The voltage between A and B with a magnetic field of 10 kG pointing out of the plane of the sample is 3.2 mV. When the magnetic field direction is reversed the voltage changes to –2.8 mV. What is the hole concentration and the mobility?

9.10 In an acceptorless n-type semiconductor with 10^{13} donors cm^{-3}, the donor ionization energy is 1 meV. Taking the effective mass as 0.01 m and assuming that $E_g >> k_BT$, calculate (a) the density of conduction electrons at 4 K and (b) the Hall coefficient.

9.11 Show that the Hall coefficient of a p-type semiconductor is zero when the excess density of acceptors over donors is

$$N_A - N_D = n_i \left(\frac{b^2 - 1}{b} \right)$$

where n_i is the intrinsic density and $b = \mu_n/\mu_p$.

9.12 Can the longitudinal magnetoresistance of n-type germanium be zero for any orientation of the magnetic field? If yes, find the direction.

9.13 Prove that the minimum conductivity of an extrinsic semiconductor is given by

$$\sigma = 2n_i e(\mu_n \mu_p)^{1/2}$$

Show that the conductivity minimum occurs when

$$N_A - N_D = n_i \left[\left(\frac{\mu_n}{\mu_p} \right)^{1/2} - \left(\frac{\mu_p}{\mu_n} \right)^{1/2} \right]$$

9.14 Prove that the barrier step height of an unbiased p-n junction is given by

$$eV_B = k_B T \ln \left(\frac{n_n}{n_p} \right)$$

where all of the symbols have their usual meaning.

9.15 Calculate the built-in voltage V_B for a p-n junction formed by diffusing boron ($n_p = 10^{15}$ cm^{-3}) into one end of an n-type silicon chip ($n_n = 3.87 \times 10^{16} T^{3/2} \exp(-E_g/2k_B T)$ cm^{-3}, $E_g = 1.1$ eV) at room temperature and 127°C.

9.16 Explain why Esaki diodes do not show high electrical conductivity in spite of having very large carrier concentrations.

SUGGESTED FURTHER READING

Dalven, R., *Introduction to Applied Solid State Physics* (Plenum Press, 1981).

Elliott, R.J. and A.F. Gibson, *Introduction to Solid State Physics and its Applications* (Macmillan, 1978).

Long, D., *Energy Bands in Semiconductors* (Wiley-Interscience, 1968).

Lüth, H., *Surfaces and Interfaces of Solid Materials* (Springer, 1995).

Milnes, A.G., *Deep Impurities in Semiconductors* (Wiley-Interscience, 1973).

Phillips, J.C., *Bands and Bonds in Semiconductors* (Academic Press, 1973).

Seeger, R., *Semiconductor Physics*, 2nd ed. (Springer, 1982).

Streetman, B.G., *Solid State Electronic Devices* (Prentice-Hall, New Delhi, 1997).

Sze, S.M., *Physics of Semiconductor Devices* (Wiley-Interscience, 1981).

Wolf, H.F., *Semiconductor* (Wiley-Interscience, 1971).

Chapter 10

Theory of Dielectrics: Applications to Plasmons, Polaritons and Polarons

In accordance with their behaviour in the presence of electric fields, the structure of any insulating solid can be viewed as a body of bound positively and negatively charged particles (atomic/molecular ions). In this context, insulators are referred to as *dielectric materials* or simply *dielectrics*. We begin this chapter with the study of response of insulators to the external electric fields. Both the static and alternating fields produce interesting effects that are exemplified by a good number of physical situations. The effects are described under two approaches: the microscopic and the macroscopic. The interaction of the electromagnetic radiation with matter is perhaps the best example where the difference between the two approaches can be made out easily. In the microscopic approach the properties are interpreted on the basis of the absorption of a photon with the creation of a phonon or an electron-hole pair. On the other hand, the macroscopic approach concerns the use of the Maxwell's equations, given below (in the SI system):

$$\text{div } \mathbf{D} = \rho \tag{10.1a}$$

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{10.1b}$$

$$\text{div } \mathbf{B} = 0 \tag{10.1c}$$

$$\text{curl } \mathbf{H} = \mathbf{j} + \frac{\partial \mathbf{D}}{\partial t} \tag{10.1d}$$

These classical equations are of central importance in the macroscopic description of electromagnetic effects in solids. Vectors **E** and **D** denote the electric field and the electric displacement vectors, respectively. Vectors **B** and **H** are the analogously defined quantities for the description of magnetic effects: the magnetic induction and the intensity of magnetic field in that order. ρ represents the volume density of quasi-free charge carriers (or of charges from an external source) and **j** is the electric current density.

10.1 POLARIZATION

When a dielectric is placed in the region of an electric field, the centres of positive and negative charges within its constituent entities (atoms/molecules) separate out if they had coincided in the

absence of the field. The typical order of the displacement is, however, less than the atomic diameter. Thus the individual constituent entities begin to behave as electric dipoles, giving a net induced dipole moment to the dielectric medium. Nevertheless, we must be aware that asymmetrical molecules such as HCl and H_2O have permanent electric dipole moment by virtue of an appreciable difference in electronegativities of the atoms they are composed of. The dipole moments of a single HCl and a single H_2O molecule are 1.0 and 1.9 debye, respectively (1 debye = 10^{-18} esu-cm = 3.3×10^{-30} C m). The effect of the electric field in these materials is to orient the dipoles, tending to align them along the field direction.

The significance of electric dipoles may be emphasized in two ways. Firstly, the magnitude of the electric field inside a dielectric cannot be accounted without taking the electric field owing to dipoles into consideration. Secondly, the orientation of electric dipoles under the action of an external electric field creates an induced charge density on the surface of the dielectric.

The electric field owing to the internal sources of charge is entirely contributed by electric dipoles that store the bound charges. The field is calculated by replacing the point charges in Coulomb's law with the electric dipoles since there are no free point charges in a dielectric. In a standard problem of electrostatics we show that the electric field owing to a dipole with moment $\mathbf{p}$ at distance $\mathbf{r}$ from its centre is given as

$$\mathbf{E}(\mathbf{r}) = \frac{3(\mathbf{p} \cdot \mathbf{r})\mathbf{r} - r^2\mathbf{p}}{4\pi \epsilon_0 r^5} \tag{10.2}$$

with

$$\mathbf{p} = q\mathbf{a} \tag{10.3}$$

where q is the magnitude of the positive and negative charges forming the dipole and $\mathbf{a}$ is the distance between the centres of the positive and negative charges (or simply the size of the dipole).

For a certain value of the electric field the dipoles align themselves parallel to the field direction (Fig. 10.1) as a result of which the charges surface on the opposite faces along this direction with

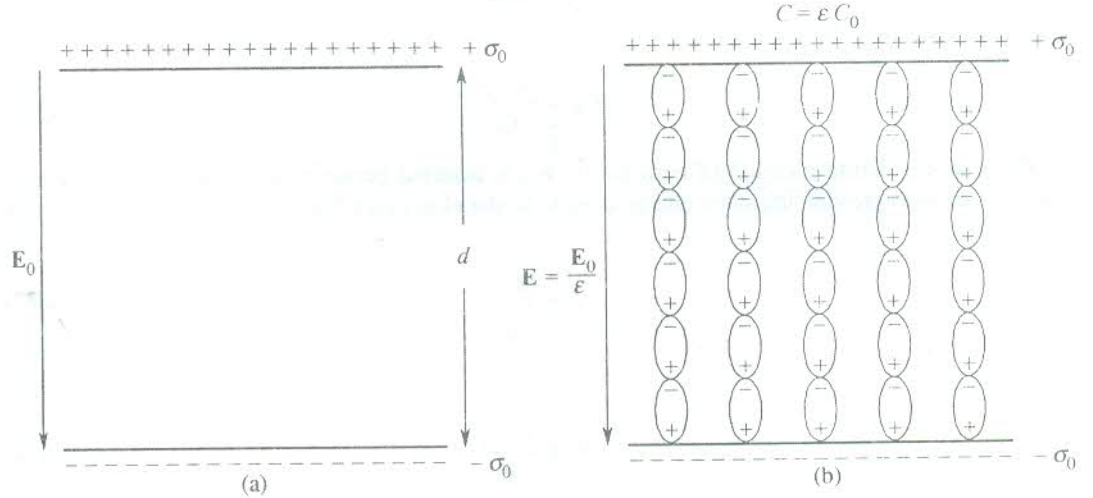


FIG. 10.1 An illustration of the effect of tiny electric dipoles (of atomic dimensions) on the macroscopic electric field: (a) A parallel plate capacitor charged to the surface charge density σ_0 with free space between the plates. The field in the region between the plates is $\mathbf{E}_0$. (b) The capacitor filled with a dielectric and charged to the same charge density σ_0. The applied field polarizes the dielectric, creating tiny dipoles which align themselves along the field direction.

the positive charges on one face and the negative on the other. These are essentially the induced charges as the dielectric on the whole continues to be neutral. In this situation the dielectric is said to be polarized and the phenomenon is termed *polarization*. The extent of polarization is given by the electric dipole moment per unit volume (**P**):

$$\mathbf{P} = \sum_n \mathbf{p}_i = \sum_n q_i \cdot \mathbf{a}_i \tag{10.4}$$

Here **P** is the total dipole moment (including the induced and permanent) and n is the number of dipoles per unit volume.

Polarization **P** has the same units as the surface charge density (C m^{-2}). This equivalence is substantiated by the fact that electric field induces charges on the surface of the dielectric and the density of charges is a measure of the extent of polarization.

10.2 DIELECTRIC CONSTANT

Consider a parallel plate capacitor with plate area A and plate separation d. When the capacitor is charged at a constant difference of d.c. potential, the electric field in the region between the plates can be treated uniform if the size of the plates is considerably larger than their separation. When the space enclosed within the capacitor is evacuated and the two plates acquire the surface charge densities $+\sigma_0$ and $-\sigma_0$, the magnitude of the electric field over the region within the capacitor is given by

$$E_0 = \frac{\sigma_0}{\epsilon_0} \tag{10.5a}$$

where ϵ_0 denotes the permittivity of the vacuum space, and E_0 is related to the difference of potential V_0 between the plates as

$$V_0 = \mathbf{E}_0 d \tag{10.5b}$$

The capacitance may be expressed as

$$C_0 = \frac{A\sigma_0}{V_0} \tag{10.6}$$

When an insulating medium of permittivity ϵ_m is inserted between the plates and the capacitor is charged so as to provide the same charge density to the plates [see Fig. 10.1(b)], the above relations transform to

$$E = \frac{\sigma_0}{\epsilon_m} \tag{10.7a}$$

$$V = Ed \tag{10.7b}$$

$$C = \frac{A\sigma_0}{V} \tag{10.7c}$$

From the above two sets of relations, we get

$$\frac{C}{C_0} = \frac{V_0}{V} = \frac{E_0}{E} = \frac{\epsilon_m}{\epsilon_0} = \epsilon \tag{10.8}$$

where ϵ is a constant known as the *relative permittivity* or the *macroscopic dielectric constant* of the medium between the plates. It is one of the most valuable constants of a solid since it is involved in the description of all of its dielectric and optical properties.

We may rewrite (10.8) as

$$C = \epsilon C_0 \tag{10.9a}$$

$$E = \frac{E_0}{\epsilon} \tag{10.9b}$$

$$V = \frac{V_0}{\epsilon} \tag{10.9c}$$

The relations (10.9) state that on introducing an insulator or a dielectric medium between the plates the capacitance increases to ϵ times the original value; and both the electric field and the potential difference across the plates are reduced to $1/\epsilon$th of their respective original values. The reduced value of electric field E in the dielectric medium would occur in vacuum for a lower value of the charge density, say, $|\sigma'|$. With this interpretation coupled to the polarization $\mathbf{P}$ being perceived as the reduction in the surface charge density, we have

$$|\mathbf{P}| = |\sigma_0| - |\sigma'| \tag{10.10}$$

And in the SI system,

$$\begin{aligned}\mathbf{P} &= \epsilon_0 \mathbf{E}_0 - \epsilon_0 \mathbf{E} \\ &= \epsilon_0 \epsilon \mathbf{E} - \epsilon_0 \mathbf{E}\end{aligned}$$

or

$$\mathbf{P} = (\epsilon - 1)\, \epsilon_0 \mathbf{E} \tag{10.11}$$

We rewrite (10.11) as

$$\frac{\mathbf{P}}{\epsilon_0 \mathbf{E}} = \epsilon - 1 = \chi_E \tag{10.12}$$

or

$$\epsilon = 1 + \chi_E \tag{10.13}$$

where χ_E is called the *electric susceptibility* of the dielectric medium.

Relation (10.11) is more often written in the following form,

$$\begin{aligned}\epsilon_0 \mathbf{E} + \mathbf{P} &= \epsilon_0 \epsilon\, \mathbf{E} \\ &= \epsilon_m \mathbf{E} \\ &= \epsilon_0 \mathbf{E}_0\end{aligned} \tag{10.14}$$

The combination $(\epsilon_0 \mathbf{E} + \mathbf{P})$ is customarily given a special name—the electric displacement vector—and denoted by $\mathbf{D}$. That is,

$$\epsilon_0 \mathbf{E} + \mathbf{P} = \mathbf{D} \tag{10.15}$$

Relations (10.11)–(10.15) are written in the form appropriate for an isotropic medium since $\mathbf{P}$, $\mathbf{E}$ and $\mathbf{D}$ can be in different directions for an anisotropic medium. The susceptibility and the dielectric constant in that case are tensors interpreted with the help of the following relations:

$$P_\mu = (\chi_E)_{\mu\nu} E$$
$$\epsilon_{\mu\nu} = 1 + (\chi_E)_{\mu\nu} \tag{10.16}$$

The displacement vector **D**, defined above, appears in Maxwell's equations (10.1a) and (10.1d). This highlights the importance of this vector in the description of the electrostatic and the electromagnetic properties of matter.

10.3 LOCAL ELECTRIC FIELD

It is often required to know the effective electric field at an atomic site for giving a proper account of dielectric properties. This electric field is termed *local electric field*. The effective electric field at a site where an electric dipole may be located equals the externally applied field only in the case of a single electric dipole. In the presence of several dipoles in a system, the neighbouring dipoles also contribute to the electric field at the atomic site of interest. The effect can be ignored only in dilute systems but surely not in solids where the dipole density is considerably large.

We see below that the local field $\mathbf{E}_{loc}$ is different from the macroscopic field **E**. The method normally employed to calculate the local electric field is due to Lorentz, according to which the atomic site of concern is considered to be located at the centre of a fictitious cavity. In this model, the local electric field is written as

$$\mathbf{E}_{loc} = \mathbf{E}_0 + \mathbf{E}_{dep} + \mathbf{E}_L + \mathbf{E}_{dip} \tag{10.17}$$

where

$\mathbf{E}_0$ is the field owing to fixed charges external to the dielectric body.

$\mathbf{E}_{dep}$ is the depolarization field caused by the uniform polarization of the dielectric. This field, having its origin in induced surface charges, opposes $\mathbf{E}_0$.

$\mathbf{E}_L$ is the field at the centre of the cavity created by charges induced on the imaginary surface bounding the fictitious cavity. This field is also called the *Lorentz field.*

$\mathbf{E}_{dip}$ is the field at the centre of the cavity contributed by dipoles within the cavity.

The value of $\mathbf{E}_{dep}$ calculated in terms of **P** depends on the shape of the dielectric specimen. For symmetrical specimens, say in the form of sphere, cylinder, disc, etc. the medium is uniformly polarized and $\mathbf{E}_{dep}$ in these cases has been calculated by Osborn* and Stoner.** For a sphere placed in a uniform electric field, it is found to be given by

$$\mathbf{E}_{dep} = -\frac{1}{3\epsilon_0}\mathbf{P} \tag{10.18}$$

The macroscopic field **E** is customarily interpreted as

$$\mathbf{E} = \mathbf{E}_0 + \mathbf{E}_{dep} \tag{10.19}$$

Hence $\mathbf{E}_{loc}$ may be expressed as

$$\mathbf{E}_{loc} = \mathbf{E} + \mathbf{E}_L + \mathbf{E}_{dip} \tag{10.20}$$

---

* J.A. Osborn, *Phys. Rev.*, **67**, 351 (1945).
** E.C. Stoner, *Phil. Mag.*, **36**, 803 (1945).

Calculation of E_L

We consider the case of a fictitious cavity of spherical shape for simplicity (Fig. 10.2). For regions external to the cavity, the distances to the cavity centre are so large that we can treat the dipole distribution as continuous, resulting in a macroscopic polarization **P**. The polarization produces a field that can be described in terms of polarization charges on the surface of the cavity. In view of the

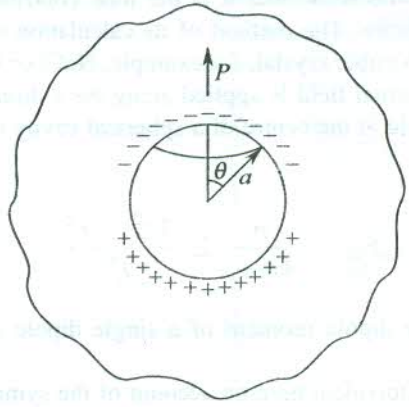


FIG. 10.2 The macroscopic polarization and the induced charges on the surface of a fictitious spherical cavity within the dielectric.

interpretation of polarization **P**, given in Section 10.2, the fictitious charge density† σ on the surface of the imaginary sphere is written as

$$\sigma = - \mathbf{P}_n \tag{10.21}$$

where $\mathbf{P}_n$ stands for the normal component of **P** (see Fig. 10.2). Then, a circular element at polar angle θ bears the charge,

$$dq = -P \cos\theta \cdot 2\pi a \sin\theta \cdot a \, d\theta \tag{10.22}$$

The electric field produced by this charge at the centre of the cavity is

$$dE_L = - \frac{1}{4\pi\epsilon_0} \frac{dq}{a^2} \cos\theta \tag{10.23}$$

Therefore, the field caused by the total charge on the surface of the cavity is

$$\mathbf{E}_L = \frac{\mathbf{P}}{2\epsilon_0} \int_0^{\pi} \cos^2\theta \sin\theta \, d\theta$$

$$= \frac{1}{3\epsilon_0} \mathbf{P} \tag{10.24}$$

† Referred to the electric field in vacuum.

Hence, for a specimen of spherical shape [from (10.18) and (10.24)], we get

$$\mathbf{E}_{dep} + \mathbf{E}_L = 0 \tag{10.25}$$

Calculation of $\mathbf{E}_{dip}$

The nature of $\mathbf{E}_{dip}$ is special in the sense that it is the lone contribution to the local field which depends on the crystalline symmetry. The method of its calculation in non-cubic crystals is not so straightforward. Let us consider a cubic crystal, for example, NaCl or CsCl in which the atomic sites have cubic symmetry.† If the external field is applied along the z-direction, then in accordance with (10.2) the z-component of the field at the centre of a spherical cavity owing to all dipoles within the cavity is written as

$$\mathbf{E}^z_{dip} = \frac{p}{4\pi\epsilon_0} \sum_i \frac{3z_i^2 - r_i^2}{r_i^5} \tag{10.26}$$

where p is the magnitude of the dipole moment of a single dipole and all dipoles are parallel to the z-axis.

Since x, y, z directions are equivalent here on account of the symmetry of the crystal and of the sphere, we have

$$\sum_i \frac{x_i^2}{r_i^5} = \sum_i \frac{y_i^2}{r_i^5} = \sum_i \frac{z_i^2}{r_i^5} \tag{10.27}$$

Therefore, for a cubic crystal with spherical cavity,

$$\mathbf{E}_{dip} = 0 \tag{10.28}$$

Since the cavity is only imaginary, this result also holds good for a cubic crystal of spherical shape.

At the same time we must appreciate that the net contribution to the local field from all the dipoles in a dielectric can be expressed as

$$\sum_i \frac{3(\mathbf{p}_i \cdot \mathbf{r}_i)\mathbf{r}_i - r_i^2 \mathbf{p}_i}{r_i^5} = \mathbf{E}_{dep} + \mathbf{E}_L + \mathbf{E}_{dip} \tag{10.29}$$

Relations (10.25) and (10.28) show that for a cubic crystal of spherical shape, the sum on the RHS of (10.29) is zero.†† So for these specimens,

$$\mathbf{E}_{loc} = \mathbf{E}_0 \tag{10.30}$$

and

$$\mathbf{E} = \mathbf{E}_0 - \frac{1}{3\epsilon_0}\mathbf{P} \tag{10.31}$$

showing different values of the two fields for the same specimen.

---

† Every atomic site in some cubic crystals may not have cubic symmetry (e.g. O^{2-} site in $BaTiO_3$).

†† A proper discussion can be found in *Classical Electrodynamics*, p. 116, J.D. Jackson (Wiley, 1962).

The $\mathbf{E}_{loc}$ at an atomic site having cubic symmetry is generally expressed in the form,

$$\mathbf{E}_{loc} = \mathbf{E} + \frac{1}{3\epsilon_0}\mathbf{P} \tag{10.32}$$

This is known as the *Lorentz relation*. It is very well satisfied by experimental data on cubic ionic crystals.

The $\mathbf{E}_{loc}$ at an atomic site in a medium of cubic symmetry is expressed in another useful form by substituting the value of $\mathbf{P}$ from (10.11) in the Lorentz relation (10.32):

$$\mathbf{E}_{loc} = \left(\frac{\epsilon + 2}{3}\right)\mathbf{E} \tag{10.33}$$

10.4 DIELECTRIC POLARIZABILITY

The electric dipole moment of an atom is found to be related to the electric field at its site by

$$\mathbf{p} = \alpha\, \mathbf{E}_{loc} \tag{10.34}$$

where α is defined as the *polarizability* of the atom concerned. This relation is valid for both CGS and SI systems with

$$(\alpha)_{SI} = 4\pi\epsilon_0(\alpha)_{CGS} \tag{10.35}$$

We can relate polarizability α, which is an atomic property, to the macroscopic property, polarization $\mathbf{P}$ by

$$\mathbf{P} = \sum_j N_j \alpha_j \mathbf{E}^j_{loc} \tag{10.36}$$

where

N_j is the number of jth type of atoms in unit volume of the crystal

α_j is the polarizability of an jth type atom

$\mathbf{E}^j_{loc}$ is the electric field at the site of an jth type atom

Then, using (10.33) the polarization of a cubic crystal is written as

$$\mathbf{P} = \left(\frac{\epsilon + 2}{3}\right)\mathbf{E}\sum_j N_j\alpha_j \tag{10.37}$$

And using (10.11), we get

$$\frac{\epsilon - 1}{\epsilon + 2} = \frac{1}{3\epsilon_0}\sum_j N_j\alpha_j \tag{10.38}$$

This is the famous Clausius–Mossotti relation. It is valid for dielectrics that crystallize in a cubic lattice. It gives the relationship between the dielectric constant and the polarizability of atoms that arises due to the displacement of the electron shell relative to the nucleus in an atom. The polarization of a hydrogen atom under the action of a static electric field is shown in Fig. 10.3. Thus the polarizability appearing in (10.38) is essentially contributed by the electronic polarization. The α_j is a tensor if the respective jth type atom is non-spherical.

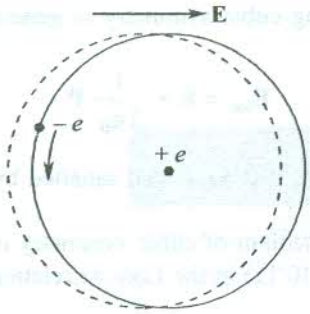


FIG. 10.3 Illustration of the displacement of the electron orbit relative to the nucleus in a hydrogen atom under the influence of an external field **E**.

10.5 SOURCES OF POLARIZABILITY

The sources that contribute to the total polarizability of a dielectric crystal are:

(i) Electronic polarization
(ii) Ionic polarization
(iii) Orientational polarization

We discuss below these polarizabilities and the related physical phenomena.

10.5.1 Theory of Electronic Polarizability and Optical Absorption

We were introduced to electronic polarizability in Section 10.4 through the Clausius–Mossotti relation (10.38). In the presence of an electric field the centre of the electron shell and that of the nucleus in an atom remain no more coincided and the atomic system gets polarized as shown for a hydrogen atom in Fig. 10.3. The interaction of optical radiation with matter offers itself, perhaps, as the best example where the consequences of electronic polarization can be grasped in a simple way. We treat this problem below in the framework of a classical theory. When the optical radiation carrying oscillating electric field strikes an atom, the atom experiences an alternating electric field which we consider as the Lorentz local field $\mathbf{E}_{loc}$ expressed as

$$\mathbf{E}_{loc} = \mathbf{E}_0 \exp(-i\omega t) \tag{10.39}$$

where $\mathbf{E}_0$ represents the amplitude of the local field.

An electron in an atom polarized by this field can be treated as a harmonic oscillator experiencing a restoring force that originates from the Coulomb attraction between the electron and the nucleus. The complete equation of motion of the electron, taken as an elastically bound particle, has the form

$$m\frac{d^2\mathbf{x}}{dt^2} + m\gamma\frac{d\mathbf{x}}{dt} + m\omega_0^2\,\mathbf{x} = e\mathbf{E}_{loc} \tag{10.40}$$

where ω_0 is the natural angular frequency of the electron such that $\omega_0 = \left(\frac{f}{m}\right)^{1/2}$, with f denoting the force constant referred to the restoring force. It must be emphasized here that the strength of the internal field $\mathbf{E}_{loc}$ is not the same as that of the field vector of the incident optical radiation. Only in gases where the density is low, the two can be equal in strength.

The second term on the LHS of (10.40) represents the damping force* on the electron characterized by dissipation constant γ with the dimensions of frequency. The dissipation results from the fact that the electron emits radiation as a consequence of its acceleration.

The solution to the equation of motion (10.40) is

$$\mathbf{x} = \frac{e\mathbf{E}_{\text{loc}}}{m(\omega_0^2 - \omega^2 + i\gamma\omega)} \tag{10.41}$$

It varies sinusoidally as dictated by the local field, the driver of these displacements. With this solution the following expression for the electronic polarizability α_e is obtained showing that α_e is a complex quantity:

$$\alpha_e(\omega) = \frac{e\mathbf{x}}{\mathbf{E}_{\text{loc}}} = \frac{e^2}{m(\omega_0^2 - \omega^2 + i\gamma\omega)} \tag{10.42}$$

If we ignore the damping effect, (10.42) would read as

$$\alpha_e(\omega) = \frac{e^2}{m(\omega_0^2 - \omega^2)} \tag{10.43}$$

which shows a singularity in the polarizability at the natural resonance frequency ω_0. But, practically, the value of the dielectric constant, a corresponding macroscopic physical quantity, never approaches either $+\infty$ or $-\infty$. It indicates to the dissipation of energy or the dielectric loss that prevents the occurrence of singularity. This is perhaps the simplest way we can substantiate the need for having a damping term in the equation of motion of an electron which is a member of the oscillating electronic shell.

The model which we have used to derive (10.42) is distinctly crude. Nevertheless, the utility of this model is underlined by the success of (10.42) in showing that the electronic polarizability is independent of the frequency of optical radiation ω when $\omega << \omega_0$ (the natural frequency of the electronic shell) and the polarizability equals its static value,

$$\alpha_e^s = \frac{e^2}{m\omega_0^2} \tag{10.44}$$

This result is supported by more accurate quantum mechanical calculations.

The values of electronic polarizability, also referred to as the atomic polarizability, in some atoms and ions are given in Table 10.1. The typical order of these values is 10^{-24} cm^3 in the CGS system and 10^{-40} F m^2 in the SI system.† Substituting this value in (10.44), we get the natural frequency of the electronic shell as $\nu_0 = \dfrac{\omega_0}{2\pi} \simeq 10^{15}\,\text{s}^{-1}$ which is about ten times higher than the frequencies in the visible range of the optical radiation. Thus, up to the beginning of the ultravoilet region, the electronic polarizability may be considered to remain constant.

---

* Proof given by R. Becker, *Theorie der Elekrizitat*, 6th ed., Teubner (Leipzig, 1933).

† Multiply the value in CGS units by $\left(\dfrac{1}{9 \times 10^{15}}\right)$ to obtain the value in SI units.

Table 10.1 Electronic polarizabilities α_e of some selected ions and rare gas atoms[a], in 10^{-24} cm^3

Ions	α_e	*Ions*	α_e	*Rare gas atoms*	α_e
Li^+	0.03	F^-	1.2	He	0.2
Na^+	0.2	Cl^-	3	Ne	0.4
K^+	0.9	Br^-	4.5	Ar	1.6
Rb^+	1.7	O^{2-}	3.88	Kr	2.5
Cs^+	2.5	S^{2-}	10.2	Xe	4.0

[a]From: A. Dalgarno, *Adv. Phys.*, **11,** 281 (1962).
L. Pauling, *Proc. R. Soc. London*, A **114,** 181 (1927).

Now, equating the value of α_e from (10.42) to the one obtained from the Clausius–Mossotti relation (10.38) and solving for $\epsilon(\omega)$, we get

$$\epsilon(\omega) = 1 + \frac{N_e e^2}{m\epsilon_0 (\omega_0^2 - \omega^2 + i\gamma\omega) - \frac{1}{3} N_e e^2} \tag{10.45}$$

where N_e denotes the density of polarizable electrons. So, the dielectric constant turns out to be a complex function when examined in the presence of an alternating electric field. When $\omega = 0$, from (10.45), we have

$$\epsilon_s = 1 + \frac{N_e e^2}{m\epsilon_0 \omega_0^2 - \frac{1}{3} N_e e^2} \tag{10.46}$$

which is real and known as the *static dielectric constant.*

We can separate the RHS of (10.45) into real (ϵ') and imaginary (ϵ'') parts and rewrite it as

$$\epsilon(\omega) = \epsilon'(\omega) + i\epsilon''(\omega) \tag{10.47}$$

This gives

$$\epsilon'(\omega) = 1 + \frac{\frac{N_e e^2}{m\epsilon_0}\left[\left(\omega_0^2 - \frac{N_e e^2}{3m\epsilon_0}\right) - \omega^2\right]}{\left[\left(\omega_0^2 - \frac{N_e e^2}{3m\epsilon_0}\right) - \omega^2\right]^2 + \gamma^2\omega^2} \tag{10.48}$$

and

$$\epsilon''(\omega) = \frac{\frac{N_e e^2}{m\epsilon_0}\gamma\omega}{\left[\left(\omega_0^2 - \frac{N_e e^2}{3m\epsilon_0}\right) - \omega^2\right]^2 + \gamma^2\omega^2} \tag{10.49}$$

Relations (10.48) and (10.49) show that there occurs a resonance at frequency,

$$\omega_r = \left(\omega_0^2 - \frac{N_e e^2}{3\, m \in_0} \right)^{1/2} \tag{10.50}$$

The resonance frequency is shifted away from the natural frequency, obviously because of the presence of a large density of polarizable electrons N_e. We say that a resonant absorption occurs at this frequency of the optical radiation. Later, when we describe the variation of the polarizability as a function of frequency, it will be clear that this absorption occurs in the ultraviolet region.

In the situation described above each electron behaves as a harmonic oscillator, possessing an induced dipole moment. Therefore, the oscillating collective electron shell may be viewed as a dipole harmonic oscillator. The terms $\in'(\omega)$ and $\in''(\omega)$ for such a dipole oscillator are plotted as functions of frequency in Fig. 10.4. For finite damping, we get only the approximate values of ω_t and the frequency of the active longitudinal mode ω_L, as obtainable from the zero crossings of $\in'(\omega)$. The dissipation constant γ is alternatively identified as the damping factor. It is shown as a measure of the width at half the maximum of the absorption line appearing in the $\in''(\omega)$ curve. If there were no damping, no absorption of radiation would take place. Such an absorption is termed *resonance absorption*. When the dielectric constant is a complex function, we will show, as a general case, in Section 10.6 that the energy absorbed per unit volume is proportional to $\in''(\omega)$. The constant $\in'(\omega)$ depends on frequency in the absorption region and accounts for the dispersion.

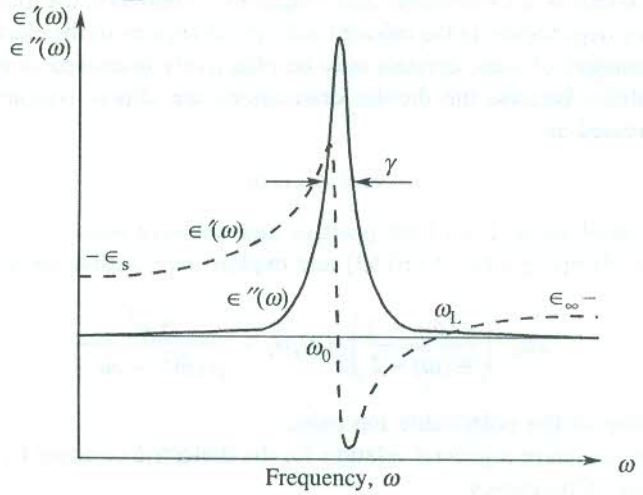


FIG. 10.4 The behaviour of $\in'(\omega)$ and $\in''(\omega)$ for a dipole oscillator. In the presence of damping, the zeros of $\in'(\omega)$ correspond to the approximate values of ω_0 and ω_L (the frequency of the longitudinal mode).

10.5.2 Ionic Polarization

The polarizability of ionic crystals cannot be accounted for without appreciating the contribution from induced ionic dipoles created by the displacement of ions under the action of an electric field. The ionic polarizability α_i on account of having its origin in the displacement of ions is also referred to as the *displacement polarizability*. The problem is treated in the same fashion as for the electronic

polarization. Let us consider an ionic crystal with masses of cations and anions being denoted by M^+ and M^-. The applied electric field produces a relative displacement of ions in an ion pair. If this displacement is $\mathbf{x}(t)$ at any time t, the equation of motion of the ion pair in terms of its reduced mass $\left(\mu = \dfrac{M^+ M^-}{M^+ + M^-}\right)$ can be written in the form

$$\mu \left[\frac{d^2\mathbf{x}}{dt^2} + \gamma \frac{d\mathbf{x}}{dt} + \bar{\omega}^2 \mathbf{x}\right] = e\mathbf{E}_{\text{loc}} \tag{10.51}$$

Here γ stands for the damping constant and $\bar{\omega}$ is the natural frequency obtained with the knowledge of force constant $f(= \mu\bar{\omega}^2)$.

Similar to the electronic response, the solution to (10.51) is found to be a complex function at frequencies of the order of $\bar{\omega}$. Accordingly, the ionic polarizability is

$$\alpha_i(\omega) = \frac{e^2}{\mu(\bar{\omega}^2 - \omega^2 + i\gamma\omega)} \tag{10.52}$$

Since the frequency $\bar{\omega}$ is characteristic of lattice vibrations, in the long wavelength limit, $\hbar\bar{\omega} \simeq \hbar\omega_D$ which may lie in the range of 0.01 to 0.1 eV. This leads to a frequency that may be 10^2 to 10^3 times smaller than ω_0 (the atomic or the electronic shell frequency). Therefore, the ionic polarizability shows a significant frequency dependence in the infrared and optical regions of the electromagnetic spectrum.

The dielectric constant of ionic crystals may be effectively discussed in terms of the ionic and electronic polarizabilities because the dipolar orientations are almost remote. Therefore, the total polarizability is expressed as

$$\alpha = \alpha_e + \alpha_i$$

where α_i includes contributions from both positive and negative ions.

If we neglect the damping term in (10.52) and exploit the Clausius–Mossotti relation, we get

$$3\epsilon_0 \left(\frac{\epsilon(\omega) - 1}{\epsilon(\omega) + 2}\right) = N_e\alpha_e + \frac{N_i e^2}{\mu(\bar{\omega}^2 - \omega^2)} \tag{10.53}$$

where N_i is the density of the polarizable ion pairs.

Now, we attempt to derive a general relation for the dielectric constant by evaluating (10.53) in the two limiting cases of frequency.

When $\omega \to 0$ or $\omega << \bar{\omega}$, we get

$$3\epsilon_0 \left(\frac{\epsilon_s - 1}{\epsilon_s + 2}\right) = N_e\alpha_e + \frac{N_i e^2}{\mu\bar{\omega}^2} \tag{10.54}$$

When ω approaches the optical frequencies, i.e. $\omega >> \bar{\omega}$, we have

$$3\epsilon_0 \left(\frac{\epsilon_\infty - 1}{\epsilon_\infty + 2}\right) = N_e\alpha_e \tag{10.55}$$

From (10.54) and (10.55), we obtain

$$\frac{N_i e^2}{\mu \bar{\omega}^2} = 3\epsilon_0 \left[\frac{\epsilon_s - 1}{\epsilon_s + 2} - \frac{\epsilon_\infty - 1}{\epsilon_\infty + 2} \right] \tag{10.56}$$

Making use of (10.55) and (10.56) in (10.53), we get

$$\frac{\epsilon(\omega) - 1}{\epsilon(\omega) + 2} = \frac{\epsilon_\infty - 1}{\epsilon_\infty + 2} + \frac{1}{\left(1 - \dfrac{\omega^2}{\bar{\omega}^2}\right)} \left[\frac{\epsilon_s - 1}{\epsilon_s + 2} - \frac{\epsilon_\infty - 1}{\epsilon_\infty + 2} \right] \tag{10.57}$$

Solving (10.57) for $\epsilon(\omega)$, we get

$$\epsilon(\omega) = \epsilon_\infty + \frac{(\epsilon_s - \epsilon_\infty)\,\omega_T^2}{\omega_T^2 - \omega^2} \tag{10.58}$$

where

$$\omega_T^2 = \bar{\omega}^2 \left(\frac{\epsilon_\infty + 2}{\epsilon_s + 2} \right) = \bar{\omega}^2 \left(1 - \frac{\epsilon_s - \epsilon_\infty}{\epsilon_s + 2} \right) \tag{10.59}$$

with ω_T as the natural resonant frequency defined by the pole of $\epsilon(\omega)$ [$\epsilon(\omega) = \infty$]. It is the long wave limiting frequency of transverse optical modes of the crystal. Relation (10.58) is an extremely useful one and exploited for describing optical dispersion in ionic crystals which is the subject of discussion in Section 10.9.

10.5.3 Polarization from Dipole Orientation

As already mentioned, an asymmetrical molecule composed of atoms with a difference of electronegativity is characterized by a permanent dipole moment. Under the action of an electric field the permanent dipoles tend to align themselves with the field. The process of orientation is partially frustrated by thermal agitations in solids. The rotation of permanent dipoles contributes to the polarization and, therefore, to the polarizability of the solid. The concept of rotatory motion is, however, more pertinent to the cases of gases and liquids. In solids, the polar molecules are generally not completely free to rotate when acted upon by an electric field. For molecules like HCl and H_2O, several stable orientations are known to exist for each molecule in the solid at high temperatures. Therefore, it is appropriate to describe the orientational motion in a polar solid as a jump in the angular displacement of molecules from one stable set of orientations to another stable set in which the dipoles are expected to be aligned more nearly along the field direction. A set of orientations of polar molecules in a crystal is stable only if allowed by the crystal symmetry.

Notwithstanding the free rotation of polar molecules in solids being scarce, the polarization contributed by the free rotation has been calculated. Such a calculation was made initially for the unhindered rotation of magnetic dipoles while developing the theory of paramagnetic susceptibility. We are giving below the results for permanent electric dipoles in paraelectric materials. For details, the reader should consult Chapter 13. This would enable him to obtain the relations given below without much ado.

The polarization contributed by the unhindered rotation of permanent dipoles measures as

$$P_{\text{dip}} = \frac{Np^2E}{3k_BT} \tag{10.60}$$

where N denotes the number of dipoles per unit volume at temperature T.

The resulting relation for the susceptibility is

$$\chi_{\text{dip}} = \frac{P_{\text{dip}}}{\epsilon_0 E} = \frac{Np^2}{3\epsilon_0 k_BT} \tag{10.61}$$

At low frequencies ($\omega \to 0$), where the contribution from the induced ionic dipoles is meaningful, the paraelectric properties are discussed in terms of the static dielectric constant expressed as

$$\epsilon_s = 1 + \chi_E \qquad \text{[using (10.13]} \tag{10.62}$$

Here

$$\chi_E = \chi_{\text{dip}} + \chi_i + \chi_e \tag{10.63}$$

From the last three relations it is obvious that the dipole moment p can be determined by measuring the static dielectric constant as a function of temperature making use of the fact that χ_i and χ_e are temperature independent.

The behaviour of low frequency (5 kHz) dielectric constant of H_2S crystal from just above the melting point to well below the melting point is depicted in Fig. 10.5. It describes the practical picture within a polar solid. On cooling, the part of the dielectric constant having its origin in the dipole orientation increases significantly in an irregular way. The irregularity is the reflection of a limited freedom available for the free orientation of dipoles. The dipoles, however, jump from one set of orientations to another as may be permitted by the crystal symmetry. An almost constant level below 100 K demonstrates that the jumps are frozen on cooling below this temperature.

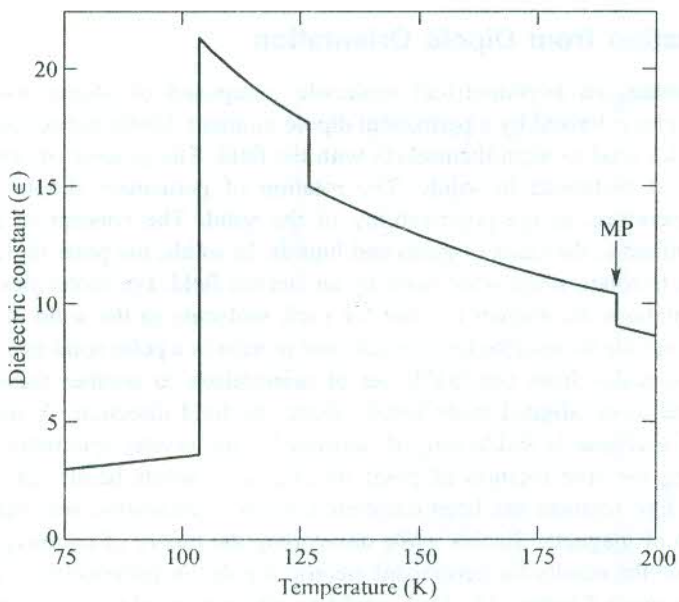


FIG. 10.5 Variation of the dielectric constant of H_2S with change in temperature, as measured at 5 kHz. [After C.P. Smyth and C.S. Hitchcock, *J. Am. Chem. Soc.*, **56**, 1084 (1934).]

The case of a real polar crystal is treated adequately in terms of the major jump of the angular displacement of molecules in going from one stable set of orientations to the other. The rate, at which the jump occurs, is characterized by the relaxation time τ. Using the Boltzmann distribution function, the ratio of occupation number densities in (1) parallel and (2) antiparallel to the field orientations is

$$\frac{N_1}{N_2} = \exp\left(\frac{2Ep}{k_B T}\right) \tag{10.64}$$

The fraction of excess dipoles that align parallel to the field is then given by

$$\frac{N_1 - N_2}{N_1 + N_2} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \tanh x \simeq x \tag{10.65}$$

in the limit, $x = \dfrac{Ep}{k_B T} << 1$.

In this approximation, the bulk polarization is expressed as

$$P_{dip} = Npx = \frac{Np^2 E}{k_B T} \tag{10.66}$$

Apart from a factor, (10.66) is identically the same as (10.60) derived from the freely rotating dipoles.

A short while ago we discussed the case of low frequencies. If we consider the higher limit of frequencies at which the jumping response of permanent dipoles may be endangered but the induced ionic and electronic polarizations may still be able to catch up with the time dependence of the electric field, Debye showed for the first time (1929) that the dielectric constant describing the properties in this situation is a complex function such that its real part $\in'(\omega)$ and imaginary that $\in''(\omega)$ are given by

$$\in'(\omega) = A + \frac{B}{1 + \omega^2\tau^2} \tag{10.67}$$

$$\in''(\omega) = \frac{\omega\tau B}{1 + \omega^2\tau^2} \tag{10.68}$$

where A and B are some constants, and τ the relaxation time.

Relations (10.67) and (10.68) are named as the Debye equations. The real part is related to the polarization which is in phase with the applied field. The imaginary part represents the dielectric loss inherent in a mechanism that lags behind in phase relative to that of the applied field. The dielectric loss occurs for frequencies at which the dispersion occurs. The real and imaginary parts of the dielectric constant of a medium are related to each other by an extremely important pair of integral equations known as the Kramers–Kronig dispersion relations.*

* H.A. Kramers, *Collected Scientific Papers* (North-Holland, 1956).

It might surprise the reader as to why no attempt has been made in this section to obtain an expression for the polarization as is usually done by invoking the Clausius–Mossotti relation. The answer to this question follows from the fact that the Lorentz method of calculating $\mathbf{E}_{loc}$ applies only to the induced dipoles which align themselves completely with the applied field. Since the permanent dipoles in a polar crystal are oriented along a number of directions, the usual method of calculation becomes irrelevant in the case of orientational polarization.

Within solids, that contain permanent dipoles, all three sources of polarization discussed in this section contribute to the polarizability. The general behaviour of the real part of the total polarizability over the frequency range from microwave to ultraviolet region is shown in Fig. 10.6. It shows that the contributions from α_e, α_i and α_{dip} fall in different frequency ranges of the electromagnetic radiation.

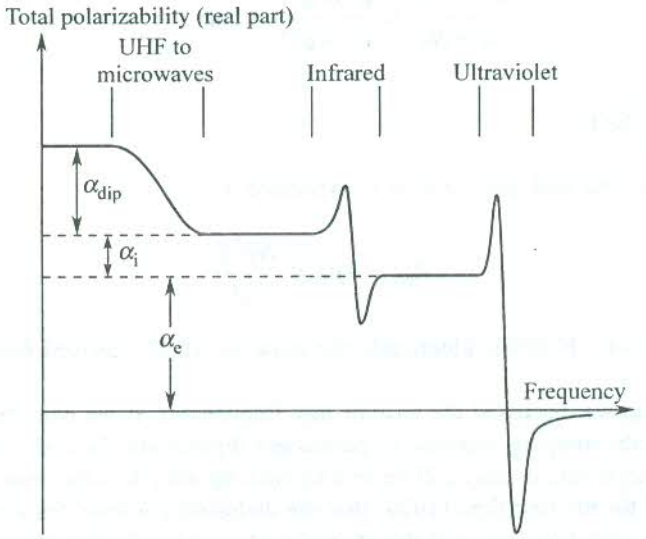


FIG. 10.6 Effectiveness of different sources of polarizability. The specific frequency range in which a certain source contributes in a major way can be identified.

10.6 DIELECTRIC LOSSES

In this section we show that the energy absorbed per second per unit volume (or the energy loss) in a dielectric medium is proportional to the imaginary part $\epsilon''(\omega)$ of the dielectric constant. The relationship among vectors **E**, **P** and **D** clearly indicates that on the application of an alternating electric field in a dielectric, vectors **P** and **D** would both vary periodically. They may generally be lagging behind in phase relative to that of **E**. Defining vectors' magnitudes as

$$E(t) = E_0 \exp(-i\omega t) \tag{10.69}$$

$$D(t) = D_0 \exp[-i(\omega t - \delta)] \tag{10.70}$$

where δ is the phase angle, giving the measure of phase lag.

In view of (10.14) and (10.15) we express the dielectric function in the following form, being aware that it is a complex quantity in the present situation:

$$\epsilon(\omega) = \epsilon'(\omega) + i\epsilon''(\omega) = \frac{D(t)}{\epsilon_0 E(t)} \tag{10.71}$$

On substituting E and D from (10.69) and (10.70), respectively, in (10.71) and then rationalizing the obtained relation for $\epsilon(\omega)$, we get

$$\epsilon'(\omega) = \frac{D_0 \cos \delta}{\epsilon_0 E_0} \tag{10.72}$$

$$\epsilon''(\omega) = \frac{D_0 \sin \delta}{\epsilon_0 E_0} \tag{10.73}$$

or

$$\tan \delta = \frac{\epsilon''(\omega)}{\epsilon'(\omega)} \tag{10.74}$$

Relation (10.74) establishes the frequency dependence of the phase angle.

Let us now take the example of a parallel plate capacitor filled with a dielectric material and bearing a surface charge density $\sigma(t)$ on its plates at any time t. Then the current density in the capacitor at that moment of time is

$$j(t) = \frac{d\sigma(t)}{dt} = \frac{dD(t)}{dt}$$

$$= \omega(-D_0 \cos \delta \sin \omega t + D_0 \sin \delta \cos \omega t) \tag{10.75}$$

Since $j(t)$ is a real physical quantity, only the real part of $D(t)$ is considered in (10.75).

The energy dissipated per unit time in one cubic metre of the dielectric is equal to

$$W = \frac{\omega}{2\pi} \int_0^{2\pi/\omega} j(t)\, E(t)\, dt \tag{10.76}$$

Using (10.75) and (10.69) (taking the real part as W is real), we obtain

$$W = \frac{1}{2} \omega \epsilon_0 E_0^2 \epsilon''(\omega) \tag{10.77}$$

showing thereby that the energy losses in the dielectric are proportional to $\epsilon''(\omega)$.

Relation (10.77) can also be put in the form

$$W = \frac{1}{2} \omega E_0 D_0 \sin \delta \tag{10.78}$$

The tan δ, given by (10.74), is often referred to as the *loss factor*. But this terminology is relevant only when δ is small, so that tan $\delta \simeq \sin \delta \simeq \delta$, and the usage may thus be held justified.

10.7 OPTICAL PHENOMENA

We know that some solids are transparent and some opaque to the light. Also, some solid surfaces are known to be highly reflecting whilst others bend the electromagnetic waves incident on them.

These properties are investigated by measuring certain physical quantities that have a close relationship with the dielectric constant ($\in(\omega)$) of the solid. This dielectric constant represents the behaviour of all electrons, those in filled and those in unfilled bands. It must be distinguished from the electrical conductivity ($\sigma(\omega)$) that takes the effect of electrons in the partially filled bands (conduction band) only. Mathematical relations applied for the study of optical properties of solids follow from the theory of propagation of electromagnetic waves. An account of the same is summarized below.

The electric field associated with the electromagnetic waves travelling along z-direction in a medium is expressed as

$$E = E_0 \exp\left[-i\omega(t - \bar{n}(\omega)z/c)\right] \tag{10.79}$$

where $\bar{n}(\omega)$ is the complex index of refraction, described as

$$\bar{n}(\omega) = n + ik = \sqrt{\in(\omega)} \tag{10.80}$$

with the usual definition

$$\bar{n}(\omega) = \frac{c}{v} \tag{10.81}$$

where v is the velocity in the medium.

According to (10.71), complex dielectric constant is written as

$$\in(\omega) = \in'(\omega) + i\in''(\omega)$$

Then using (10.80) we get

$$\in'(\omega) = n^2 - k^2$$

and

$$\in''(\omega) = 2nk \tag{10.82}$$

Here k is called the extinction coefficient, while n and k are the two optical constants which one generally quotes.

Determination of Optical Constants

The optical constants, n and k (and hence the dielectric constant, $\in(\omega)$) are estimated by exploiting their relationship with the reflectance and the absorption coefficient, measured independently. The exercise involving the measurements at normal incidence is simple. Suppose a plane wave of light from vacuum falls on a crystal at normal incidence. The intensities of the reflected and incident light waves are measured to obtain the reflectance (R) which is defined as the ratio of the former to the latter. The reflectance corresponds with another quantity called reflectivity (r) that is written as

$$r = \frac{E_{\text{refl}}}{E_{\text{inc}}} \tag{10.83}$$

where E stands for the electric field.

By requiring the components of the electric field vector (**E**) and the magnetic field vector (**B**) parallel to the crystal surface to be continuous, it can be shown that

$$r = \frac{n + ik - 1}{n + ik + 1} \tag{10.84}$$

By its definition, the reflectance is given by

$$R = \frac{E^*_{\text{refl}} E_{\text{refl}}}{E^*_{\text{inc}} E_{\text{inc}}} = r^* \cdot r = \frac{(n-1)^2 + k^2}{(n+1)^2 + k^2} \tag{10.85}$$

One can follow either of the following two approaches to extract the values of the two optical constants:

1. In one approach, relation (10.85) is generalized to non-normal angles of incidence (say, θ) and a second expression for reflectance is obtained at angle of incidence θ, in terms of θ, $n(\omega)$, $k(\omega)$, and the polarization of the incident radiation. When one compares this with the measured reflectance at angle θ, another equation containing $n(\omega)$ and $k(\omega)$ is formed and two constants can be readily estimated.
2. In the other approach, one makes use of the Kramers–Kronig relations through which $n(\omega)$ and $k(\omega)$ are connected:

$$n(\omega) = 1 + P \int_{-\infty}^{\infty} \frac{d\omega'}{\pi} \frac{k(\omega')}{\omega' - \omega}$$

$$k(\omega) = -P \int_{-\infty}^{\infty} \frac{d\omega'}{\pi} \frac{n(\omega') - 1}{\omega' - \omega} \tag{10.86}$$

where P stands for a principal value integral.

The required input for the numerical analysis, employing these equations, is the knowledge of $r(\omega)$ at all frequencies. Thus the method is practically disadvantaged by the need of making measurements over a vast range of frequencies so as to have reliable extrapolations over the full range of frequencies on the demand of (10.86). The procedure, though fraught with complications, is ideal in principle. It is in fact preferred because of the high accuracy achieved in the estimation of the optical constants.

10.8 APPLICATION TO PLASMA

In this section we apply the theory developed so far to study the properties of a plasma.

10.8.1 Plasma Oscillations

A medium of equal concentrations of positive and negative charges, with at least one charge type being mobile, is known as a *plasma*. A metallic crystal may be described as an assembly of immediate positive ion cores (atomic nuclei) and conduction electrons that are nearly free to move over the whole of the crystal. The condition of charge neutrality is maintained because of the balance struck between the negative charge on conduction electrons and an equal concentration of positive charge on ion cores. Thus a metal serves as a good example of plasma.

An idealized plasma in which both particles carrying positive and negative charges, respectively, are motionless is referred to as *cold plasma*. But every real plasma at temperatures of interest is characterized by thermal agitations which at times are random in nature. A random motion may be a momentary fluctuation in the equilibrium position of an electron, caused by the average electrostatic field of all other electrons. The position fluctuation would create a charge imbalance in the region of that electron, as a result of which other electrons would rush into that region in order to restore the condition of charge neutrality. At any finite temperature, electrons being very light particles, move with fairly large speed relative to ions which we consider to be at rest. The electrons rushing into the region of the electron that suffered a fluctuation in its equilibrium position are unable to stop at the desired positions and overshoot their mark on account of the large kinetic energy which by the way represents the total energy there. As soon as the energy goes totally electrostatic, electrons turn around and attempt again to approach the wanted locations in the region of the misbehaving electron. The repetition of this process constitutes the collective oscillatory motion. These collective oscillations are called *plasma oscillations.*

The plasma oscillations, in principle, need to be interpreted in terms of the collective modes of the electron gas. For a proper theoretical treatment, extra degrees of freedom associated with the collective motion of the electron gas must be identified in addition to those associated with the motion of individual electrons. In this description, the electrons beyond a certain distance (the screening radius) from a given electron act cooperatively on that electron. Only for distances shorter than this, are the independent-particle aspects important. Therefore, an accurate description of plasma oscillations can be found only in the framework of many body theory which is too an advanced subject to find place in this book. Nevertheless, we present below a rather simple picture of the screening mechanism as given for the first time by Debye and Huckle.*

Consider a metal plasma in which a negative charge is introduced at a point chosen as the origin. Let the volume charge density at the point be denoted by ρ. If the charge imbalance because of a momentary fluctuation in charge density lowers the electron concentration at that point by δN, then the resulting charge density at the point is $(\rho - e\delta N)$. The effective electrostatic potential ϕ is related to this charge density by the Poisson equation

$$\nabla^2\phi = \frac{1}{\epsilon_0}(\rho - e\delta N) \tag{10.87}$$

When the excitation energy $e\phi$ is very much smaller than $k_B T$ and the electron concentration N is so low as to permit the use of classical statistics, the decrease in electron concentration can be written as

$$\delta N = N_0\left[\exp\left(\frac{e\phi}{k_B T}\right) - 1\right]$$

$$\simeq N_0\left(\frac{e\phi}{k_B T}\right) \tag{10.88}$$

where N_0 is the equilibrium electron concentration (in the absence of fluctuation in the equilibrium position). Putting the value δN from (10.88) in (10.87), we have

* P.P. Debye and E. Huckel, *Z. Physik,* **24**, 185 (1923).

$$\left(\nabla^2 + \frac{N_0 e^2}{\epsilon_0 k_B T}\right)\phi = \frac{\rho}{\epsilon_0} \tag{10.89}$$

which has the solution,

$$\phi = \frac{\rho}{r}\exp\left(-\frac{r}{\lambda_D}\right) \tag{10.90}$$

with

$$\lambda_D = \left(\frac{\epsilon_0 k_B T}{N_0 e^2}\right)^{\frac{1}{2}} \tag{10.91}$$

Relation (10.90) explains the screening of a given electron from other electrons via the fluctuation of the electron charge density. λ_D is known as the Debye length. The effect of screening is such that when other electrons are at a distance λ_D from the given electron, the normal Coulomb potential is suppressed by the factor (1/e). The screening effect closely controls the behaviour of the dielectric constant* which is exploited to explain many interesting phenomena involving electron–electron, electron–photon and electron–phonon interactions.

We follow a classical approach to calculate the dielectric constant of a plasma in the absence of an external source of perturbation. Let us consider the two-dimensional model of a metal plasma as shown in Fig. 10.7(a). It is assumed that the configuration of the plasma particles does not depend on their position on the third spatial axis. In view of this fact, we suppress this direction in our calculations. On account of a random fluctuation in the equilibrium positions of electrons, two adjoining regions of positive and negative charge densities are created, meaning thereby that the charge neutrality condition is destroyed in these regions. In order to calculate the frequency of a given plasma electron, we work with a picture that is alternative to the one used earlier for describing the plasma oscillations.

Let a small volume element enclosed by surface **S** in the negatively charged region have a charge q. The electric field in this small volume is given by Gauss' theorem:

$$\int_S \mathbf{E} \cdot d\mathbf{S} = \frac{q}{\epsilon_0} \tag{10.92}$$

And suppressing the third spatial axis, we have

$$\int_S \mathbf{E} \cdot d\mathbf{S} = -lE \qquad \text{[see Fig. 10.7(b)]} \tag{10.93}$$

Also,

$$q = -lxN_0 e \tag{10.94}$$

where x denotes the displacement or the overshoot in the direction of E.

---

* J.M. Ziman, *Principles of the Theory of Solids*, Chap. 5 (Cambridge Univ. Press, 1972).

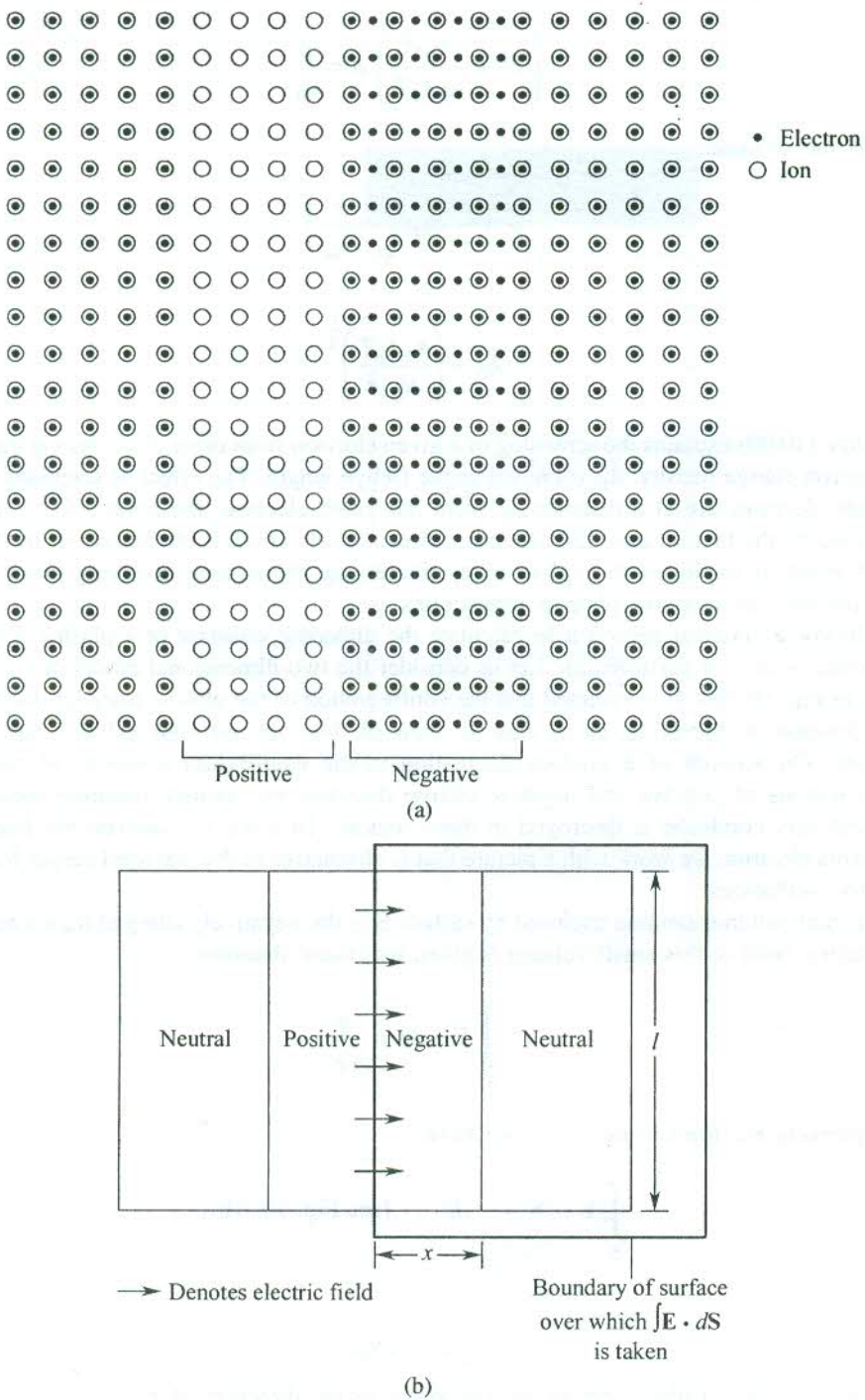


FIG. 10.7 (a) Displacement of electrons in a two-dimensional plasma. (b) Electric field in a two-dimensional plasma stands for the electric field. The dark line shows the boundary of surface over which $\int \mathbf{E} \cdot d\mathbf{S}$ is evaluated.

From the preceding three relations, we have

$$\mathbf{E} = \frac{N_0 e\mathbf{x}}{\epsilon_0} \tag{10.95}$$

The electric field **E** serves as a perturbation and drives an electron into oscillatory motion. The equation of motion has the form

$$m\ddot{\mathbf{x}} = -e\mathbf{E} \tag{10.96}$$

Making use of (10.95), we get

$$\ddot{\mathbf{x}} = -\left(\frac{N_0 e^2}{\epsilon_0 m}\right)\mathbf{x} \tag{10.97}$$

This relation describes a simple harmonic motion of the characteristic frequency

$$\omega_p = \left(\frac{N_0 e^2}{\epsilon_0 m}\right)^{1/2} \tag{10.98}$$

where ω_p is known as the *plasma frequency*.

In the presence of an external field, both **E** and **x** are bound to a common oscillatory character represented by the time-dependent perturbation ~ exp(– $i\omega t$), where ω denotes the angular frequency of the perturbation force. We may now easily calculate the dielectric constant of the plasma in which the positive ion cores are at rest. Under the influence of **E**, an electron has a dipole moment and the bulk polarization of the plasma is

$$P(\omega) = -N_0 ex \tag{10.99a}$$

with

$$m\ddot{x} = -m\omega^2 x = -eE(\omega) \tag{10.99b}$$

But,

$$\epsilon(\omega) = \frac{D(\omega)}{\epsilon_0 E(\omega)} = 1 + \frac{P(\omega)}{\epsilon_0 E(\omega)} \tag{10.100}$$

Using (10.99a) and (10.99b), we have

$$\epsilon(\omega) = 1 - \frac{N_0 e^2}{\epsilon_0 m\omega^2} \tag{10.101}$$

or

$$\epsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \qquad \text{[using (10.98)]} \tag{10.102}$$

Relation (10.102) obviously expresses the dielectric constant of the free electron gas. When the frequency of the position fluctuation ω matches the plasma frequency ω_p, the dielectric constant $\epsilon(\omega) = 0$. This condition refers to the longitudinal plasma oscillations whose wavevector is taken as very nearly zero. The energy quanta $\hbar\omega_p$ are commonly known as *plasmons*. By the geometry of a

longitudinal polarization wave, a depolarization field $\mathbf{E} = -\mathbf{P}/\epsilon_0$ is created. This leads to $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P} = 0$ and $\epsilon(\omega) = 0$. During a longitudinal plasma oscillation, the electron gas is moved as whole relative to the positive ion background. Figure 10.8 shows how the regions of negative charge and positive charge emerge out of a neutral charge distribution within a thin metal slab.

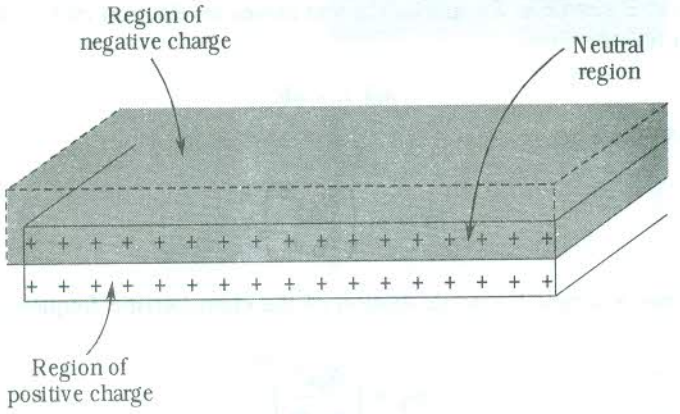


FIG. 10.8 The negative charges (electrons) against the background of positive ions in the thin slab of a metal.

Typical values of plasmon energy lie between 3 and 20 eV. They can be calculated if the electron number density and an appropriate value for the effective mass of electrons are known. The calculated values are found to be in good agreement with the observed ones in several metals and dielectrics. For example, they are 5.71 and 5.95 eV respectively in sodium. The plasmons may be excited by passing electrons through a thin metal film. The reflection of electrons or photons is also used for the purpose. During the process the charge of an electron couples with the electrostatic field fluctuations of plasma oscillations and the reflected or transmitted electron shows a loss of energy that equals an integral multiple of $\hbar\omega_p$. Figure 10.9 shows the energy loss spectrum for a 20 keV electron beam, incident on an aluminium foil. In this case, the characteristic energy loss peaks are all due to the excitation of one or more plasmons by the electrons passing through the foil.

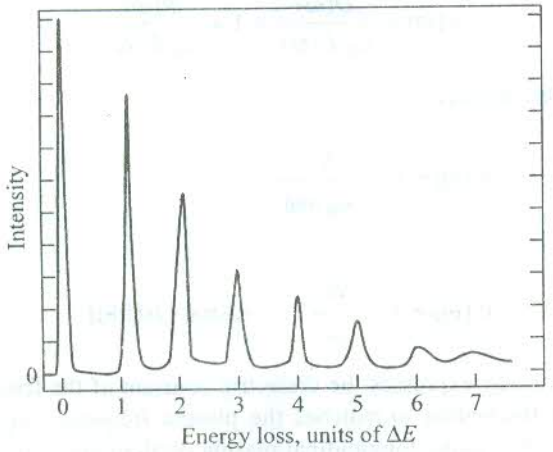


FIG. 10.9 Energy loss spectrum for a beam of 20 keV primary electrons passing through an aluminium foil of thickness 2580 Å. The unit of energy loss, ΔE, is about 1.5 eV, the plasmon excitation energy in aluminium. [After L. Marton, J.A. Simpson, H.A. Fowler and N. Swanson, *Phys. Rev.*, **126**, 182 (1962).]

If we treated ionic motion independently, we would get a much smaller frequency of oscillation for ions due to their heavier mass. The dielectric constant of the positive ion core background may be defined by (10.102) with the frequency of ion cores coming in for ω_p. It remains practically constant (a little below unity) up to the frequencies of perturbation ω well above ω_p. On account of its correspondence with high frequencies, it is denoted by ϵ_∞. Accordingly, we rewrite (10.102) as

$$\epsilon(\omega) = \epsilon_\infty - \frac{\omega_p^2}{\omega^2} = \epsilon_\infty \left[1 - \frac{\omega_p^2/\epsilon_\infty}{\omega^2}\right]$$

or

$$\epsilon(\omega) = \epsilon_\infty \left(1 - \frac{\Omega_p^2}{\omega^2}\right) \tag{10.103}$$

with

$$\Omega_p = \frac{\omega_p}{\sqrt{\epsilon_\infty}} \tag{10.104}$$

The frequency Ω_p expresses the uniform collective longitudinal oscillation of the electron gas against a background of the fixed positive ions. It also denotes the low cutoff for the propagation of transverse electromagnetic waves in plasma because at this frequency $\epsilon(\omega) = 0$ (defining the longitudinal waves).

10.8.2 Transverse Optical Modes in Plasma

We first derive the dispersion relation for the electromagnetic waves, taken as model transverse waves. Then, using the theory developed above we obtain a relation for the positive real values of the dielectric constant. This relation when compared with the dispersion relation for the electromagnetic waves will be found to describe the electromagnetic waves that propagate through the plasma, establishing thereby the excitation of the transverse optical modes of the plasma oscillation.

The electromagnetic wave equation in a non-magnetic isotropic medium is given by

$$\mu_0 \epsilon_0 \epsilon(\omega, \mathbf{k}) \frac{\partial^2 \mathbf{E}}{\partial t^2} = \nabla^2 \mathbf{E} \tag{10.105}$$

where $\mathbf{k}$ is the wavevector of the electromagnetic wave. With the electric field vector $\mathbf{E}$ varying as $\sim \exp(i\mathbf{k} \cdot \mathbf{r}) \exp(-i\omega t)$, we get the following dispersion relation:

$$\epsilon(\omega, \mathbf{k}) \mu_0 \epsilon_0 \omega^2 = \mathbf{k}^2 \tag{10.106}$$

Substituting the value of the dielectric constant from (10.103) into (10.106), we obtain

$$\epsilon(\omega)\omega^2 = \epsilon_\infty \omega^2 \left(1 - \frac{\Omega_p^2}{\omega^2}\right) = \frac{k^2}{\mu_0 \epsilon_0} \tag{10.107}$$

For $\omega < \Omega_p$, the above relation gives a negative value of the dielectric constant and an imaginary value of the wavevector. It gives the k-dependence of waves, say along the x-direction as $\exp(-|k|x)$

in the frequency range $0 < \omega \leq \Omega_p$. These waves cannot propagate through the medium and suffer complete reflection.

But at frequencies such that $\omega > \Omega_p$, (10.107) gives a positive real value of the dielectric constant and real k-values. Thus the medium is transparent to these waves. We rewrite (10.107) as

$$\omega^2 = \Omega_p^2 + \frac{k^2}{\mu_0 \epsilon_0 \epsilon_\infty} \tag{10.108}$$

This relation is similar to (10.106) and it represents the dispersion relation of the transverse electromagnetic modes in plasma. The dispersion curve is drawn in Fig. 10.10.

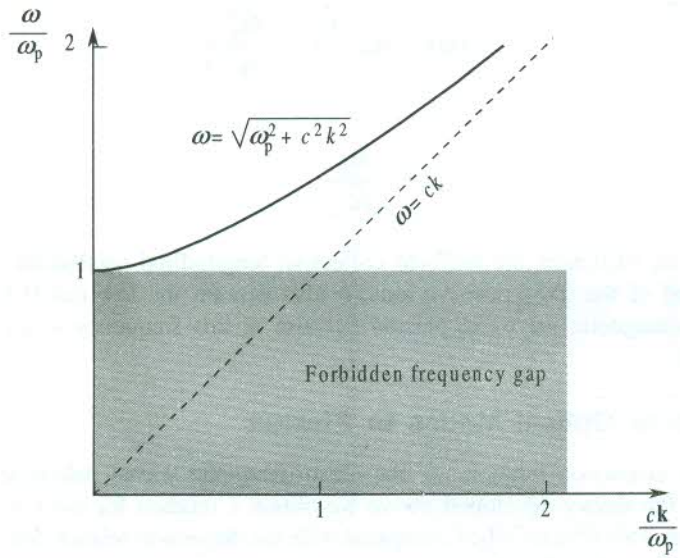


FIG. 10.10 The dispersion curve for transverse electromagnetic waves in a plasma.

The significance of the transverse optical modes in plasma is underlined by their close involvement in controlling the optical properties of crystals, especially metals. For example, the metallic lustre is attributed to the total reflection of visible light to which metals are not transparent. The reason being that the frequency of visible radiation is less than Ω_p. The transparency of alkali metals in the ultraviolet region strongly vindicates the theory that requires an electromagnetic radiation to have a certain minimum frequency or a wavelength not longer than a certain value to be able to pass through a plasma. The limit is defined in terms of ω_p or the corresponding wavelength in vacuum $\lambda_p (= 2\pi c/\omega_p)$. The theoretical and observed values of λ_p in alkali metals are given in Table 10.2. The two sets of values are within a reasonable limit of agreement.

Table 10.2 Longest electromagnetic waves λ_p, allowed for transmission in alkali metals (in Å)

P	Li	Na	K	Rb	Cs
Observed	1550	2100	3150	3400	—
Calculated	1550	2090	2870	3220	3620

It is equally interesting to apply our free electron gas model to semiconductors. The practice of doping facilitates the variation of the free electron concentration over a wide range. This is extremely helpful in varying the plasma edge of reflectivity. For example, Sn-doped In_2O_3 layers are almost totally transparent to the visible light but serve as effective barrier to the propagation of infrared radiation. Figure 10.11 shows the reflectivity and the transmissivity of one such layer. The coatings of Sn-doped In_2O_3 are applied in sodium vapour lamps and used as heat reflecting windows.

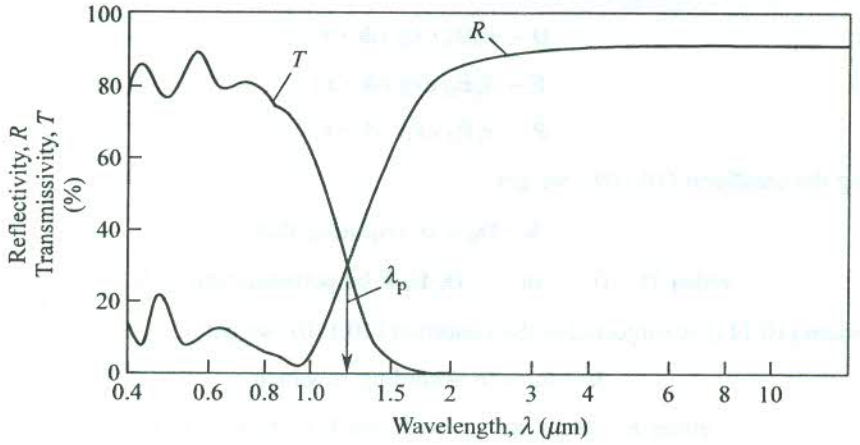


FIG. 10.11 Variations of reflectivity R and transmissivity T with change in the wavelength for a Sn-doped In_2O_3 layer. Thickness of the layer = 0.3 μm and the electron concentration is 1.3×10^{21} cm^{-3}. [After G. Frank, E. Kauer, H. Köstlin, *Phys. Blätter*, **34**, 106 (1978).]

10.9 APPLICATION TO OPTICAL PHONON MODES IN IONIC CRYSTALS

The dispersion relations for the normal modes of an ionic crystal were treated in Chapter 4. The general technique is, however, fraught with difficulties in dealing with the long ranging interior electrostatic interactions. If we limit our attention to long wavelength optical modes we can avoid the use of complicated methods, developed for the purpose and work in an electrostatic approximation where the Maxwell's equation for curl **E** is relaxed to ignore the effect of $-\partial \mathbf{B}/\partial t$ at optical frequencies.

At long wavelengths ($\mathbf{k} \simeq 0$) during the optical mode of vibration in an ionic crystal, the positive and negative ions suffer displacements in opposite directions giving rise to a net polarization **P**. The macroscopic electric field **E** and the displacement vector **D** associated with it are given by the well-known relationship (10.15):

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$$

Since there are no charges to have been introduced by an external source (or simply no free charges), we have

$$\text{div } \mathbf{D} = 0 \tag{10.109}$$

Further, the microscopic field $\mathbf{E}^{\text{micro}}(\mathbf{r})$ rapidly varying over a range of atomic distances is connected with the corresponding potential $\phi^{\text{micro}}(\mathbf{r})$ by

$$\mathbf{E}^{\text{micro}}(\mathbf{r}) = -\text{grad}_{\mathbf{r}}\, \phi(\mathbf{r})$$

Because the macroscopic field **E** represents the average of $\mathbf{E}^{micro}$, the above relation applies in that case too. As stated in the beginning, we ignore the effect of the $-\partial\mathbf{B}/\partial t$ term in the Maxwell equation and write

$$\text{curl } \mathbf{E} = \text{curl } (-\text{ grad}_r\ \phi) = 0 \tag{10.110}$$

In ionic crystals owing to the cubic symmetry, ϵ is not a tensor and therefore **D**, **E** and **P** are parallel (10.15). We take the spatial variation of these vectors as

$$\begin{aligned} \mathbf{D} &\sim R_e \mathbf{D}_0 \exp(i\mathbf{k}\cdot\mathbf{r}) \\ \mathbf{E} &\sim R_e \mathbf{E}_0 \exp(i\mathbf{k}\cdot\mathbf{r}) \\ \mathbf{P} &\sim R_e \mathbf{P}_0 \exp(i\mathbf{k}\cdot\mathbf{r}) \end{aligned} \tag{10.111}$$

On imposing the condition (10.109), we get

$$\mathbf{k}\cdot\mathbf{D}_0 = 0, \text{ requiring that}$$

$$\text{either } \mathbf{D} = 0 \quad \text{or} \quad \mathbf{D}, \mathbf{E}, \mathbf{P} \text{ be perpendicular to } \mathbf{k}. \tag{10.112}$$

Similarly, when (10.111) is subjected to the condition (10.110), we get

$$\mathbf{k}\times\mathbf{E}_0 = 0, \text{ according to which}$$

$$\text{either } \mathbf{E} = 0 \quad \text{or} \quad \mathbf{D}, \mathbf{E} \text{ and } \mathbf{P} \text{ be parallel to } \mathbf{k}. \tag{10.113}$$

We see immediately below that one of the requirements under (10.112) combines with another, the compatible one under (10.113) to define a certain type of optical mode (longitudinal or transverse).

10.9.1 The Longitudinal Optical Mode

In a longitudinal mode the polarization **P** is parallel to the direction of propagation of the wave, i.e. **k**, as is one of the two allowed possibilities under (10.113). One of the two possibilities under (10.112) that must be considered to define this mode completely is $\mathbf{D} = 0$, since $\mathbf{P} \perp \mathbf{k}$ is already ruled out. Therefore, for a longitudinal optical mode,

$$\mathbf{E} = -\frac{\mathbf{P}}{\epsilon_0} \quad \text{and} \quad \epsilon = 0 \qquad \text{(from 10.15)} \tag{10.114}$$

since,

$$\epsilon = \frac{\mathbf{D}}{\epsilon_0 \mathbf{E}} \quad \text{and} \quad \mathbf{D} = 0$$

10.9.2 The Transverse Optical Mode

In this case $\mathbf{P} \perp \mathbf{k}$, therefore, for a transverse optical mode,

$$\mathbf{E} = 0 \text{ and } \epsilon = \infty \tag{10.115}$$

We must not forget here that (10.115) has been derived in the electrostatic approximation of the Maxwell's equation, curl $\mathbf{E} = -\partial\mathbf{B}/\partial t$. Therefore, for an accurate description of the transverse optical modes, the condition $\epsilon = \infty$ must be replaced by a more general condition.

The transverse modes with angular frequency ω and wavevector $\mathbf{k}$ propagate, provided

$$\epsilon(\omega) = \frac{k^2 c^2}{\omega^2} \qquad \text{[from (10.106)]} \tag{10.116}$$

In view of this, the condition $\epsilon = \infty$ is acceptable only if the wavevectors of the optical modes are in accordance with $kc >> \omega$. Since the magnitude of the optical phonon wavevector matches the order of the Debye wavevector, we have

$kc >> k_D S$, where S is the speed of sound in the crystal

or

$$k >> k_D \times \left(\frac{S}{c}\right) \tag{10.117}$$

Because the $S/c \simeq 10^{-4}$ to 10^{-5} and k_D compares with the size of the Brillouin zone, it is concluded that the electrostatic approximation is applicable to those optical modes whose wavevectors are long enough and never be as small as a small fraction of a per cent of the Brillouin zone size when measured from $\mathbf{k} = 0$.

10.10 THE INTERACTION OF ELECTROMAGNETIC WAVES WITH OPTICAL MODES

We refer back to Section 10.5.2 dealing with the ionic polarizability of a diatomic ionic crystal in the presence of an electric field $\mathbf{E}_{loc}$ [$\sim \mathbf{E}_0 \exp(-i\omega t)$]. This field can as well come from the one associated with an electromagnetic wave. Therefore, the theory of Section 10.5.2 forms the basis for the discussion in this section. At the outset it must be clear that the longitudinal optical modes cannot have any coupling with the transverse electromagnetic waves in the bulk of a crystal. We are concerned here with the coupling between the transverse optical phonons and the photons.

Let us first study the structure of the transverse modes all the way down to $k = 0$. This is done by making use of the relation (10.58) for the dielectric constant derived in Section 10.5.2:

$$\epsilon(\omega) = \epsilon_\infty + \frac{(\epsilon_s - \epsilon_\infty)\,\omega_T^2}{\omega_T^2 - \omega^2}$$

We plot ϵ versus ω using the above relation in Fig. 10.12. The intersection with the ω-axis (on which $\epsilon = 0$) gives the value of ω_L and the pole of $\epsilon(\omega)$ determines ω_T. It is significant that ϵ is negative between ω_T and ω_L, making k imaginary (10.106). This means that no radiation can propagate in the crystal over this range of frequencies. The region corresponds with the spread of the reststrahlen band which appears owing to strong reflection of the radiation. As mentioned in Section 10.5.2, there occurs a resonance at frequency ω_T where the frequencies and the wavevectors of the optical phonon and the photon are equal. In this condition the phonon–photon coupling becomes so substantial that it completely changes the nature of propagation and a forbidden gap is created which in no way can be tied to the periodicity of the crystal. The consequence of this coupling is implicitly contained in the dielectric function (10.58). The quantum of the coupled phonon–photon transverse wave field is commonly known as a *polariton*.

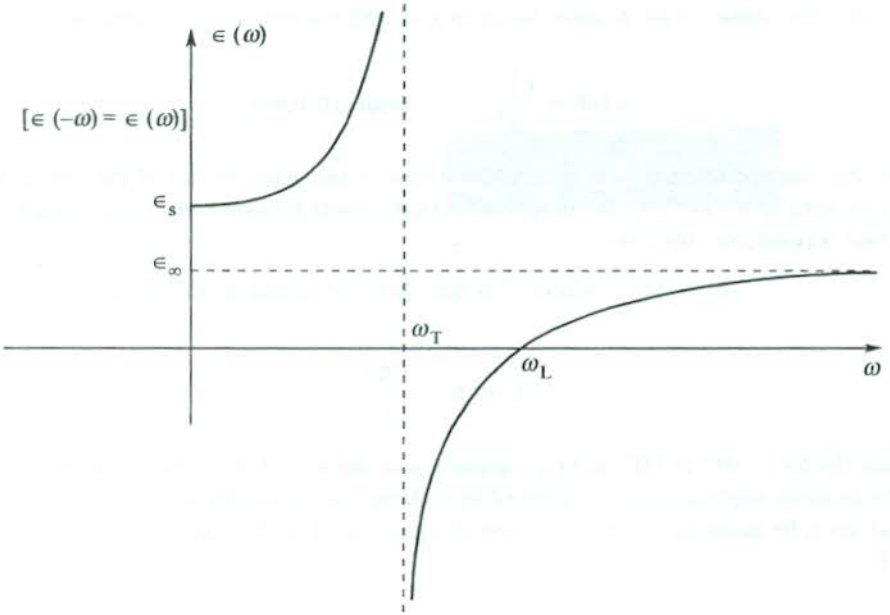


FIG. 10.12 Behaviour of the dielectric constant with change in frequency in a diatomic ionic crystal.

We give below a simple account of the transverse optical (TO) phonon–photon coupling and calculate the polariton frequency at $k = 0$.

The dispersion relation for the photons is [see (10.106)]

$$\in(\omega)\omega^2 = \frac{k^2}{\mu_0 \in_0}$$

or

$$\left(\frac{\in_0 \mathbf{E} + \mathbf{P}}{\in_0 \mathbf{E}}\right)\omega^2 = \frac{k^2}{\mu_0 \in_0}$$

or

$$\mathbf{P}\omega^2 + \left(\in_0 \omega^2 - \frac{k^2}{\mu_0}\right)\mathbf{E} = 0 \tag{10.118}$$

If we do not consider the electronic polarization and neglect the damping term in (10.52), the polarization within the crystal can be written as

$$\mathbf{P} = \frac{N_i e^2}{\mu(\bar{\omega}^2 - \omega^2)}\mathbf{E}$$

or

$$\mu(\bar{\omega}^2 - \omega^2)\mathbf{P} - N_i e^2 \mathbf{E} = 0 \tag{10.119}$$

For the relations (10.118) and (10.119) to have a solution, the determinant of the coefficients of **P** and **E** in these relations must vanish. Therefore,

$$\begin{vmatrix} \omega^2 & \left(\epsilon_0 \omega^2 - \dfrac{k^2}{\mu_0}\right) \\ \mu(\bar{\omega}^2 - \omega^2) & - N_i e^2 \end{vmatrix} = 0 \tag{10.120}$$

For $\mathbf{k} = 0$, it gives two roots:

$$\omega = 0 \text{ for the photon}$$

and

$$\omega^2 = \bar{\omega}^2 + \frac{N_i e^2}{\epsilon_0 \mu} \tag{10.121}$$

To remind, N_i is the density of ion pairs and μ is the reduced masss of an ion pair. The $\bar{\omega}$ is related to ω_T by (10.59) when the electronic polarization is also considered. Therefore, the polariton frequency at $\mathbf{k} = 0$ may be given by

$$\omega^2 = \omega_T^2 \left(\frac{\epsilon_s + 2}{\epsilon_\infty + 2}\right) + \frac{N_i e^2}{\epsilon_0 \mu} \tag{10.122}$$

But we must appreciate that the resonance occurs for a finite value of **k**, where the coupling between the TO phonon and the photons is the strongest. It can be determined by the intersection of the photon dispersion curve (10.106) and the dispersion curve of TO phonons. This value of **k** turns out to be very small compared with the extent of the Brillouin zone. This region of resonance is often referred to as the *crossover region.*

Solutions to the dispersion relation for the transverse optical modes propagating in a diatomic ionic crystal are drawn in Fig. 10.13. There are two branches, one lying completely below ω_T and the other completely above ω_L. The appropriate dielectric constant to be used for frequencies below ω_T is ϵ_s, and ϵ_∞ for frequencies above ω_T (see Fig. 10.12). The corresponding dispersions (the dashed lines at $\mathbf{k} = 0$) are represented accordingly. In the linear regions of each of the continuous curves, one mode is clearly phonon like. The curved regions are mixed in character and describe the dispersion of polaritons.

The theory discussed in this section also provides the relationship between ω_L and ω_T. When $\omega = \omega_L$, $\epsilon = 0$. Imposing this condition on (10.58), we have

$$\omega_L^2 = \frac{\epsilon_s}{\epsilon_\infty} \omega_T^2 \tag{10.123}$$

Relation (10.123) is known as the Lyddane–Sachs–Teller relation (addressed popularly as the LST relation). This relation has been tested by comparing the values of ω_L and ω_T, obtained from the neutron scattering experiment, with the measured values of the dielectric constant and index of refraction ($\bar{n} = \sqrt{\epsilon_\infty}$). Woods et al.* have observed excellent agreement between ω_L/ω_T and

* A.D.B. Woods et al., *Phys. Rev.*, **131,** 1025 (1963).

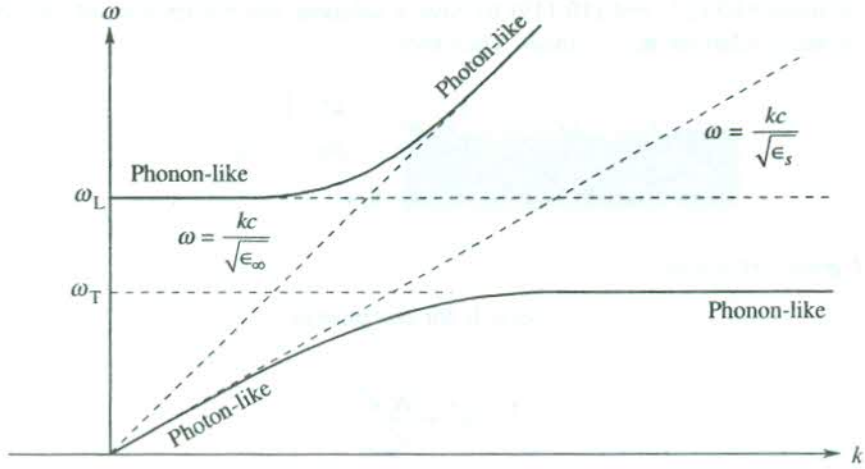


FIG. 10.13 Dispersion curves of coupled TO phonon and optical photon in a diatomic ionic crystal (in thick continuous lines). The curves for the uncoupled TO phonons and optical photons are shown by the dashed lines. There are two linear regions in each branch of the thick curves. In these regions one mode is clearly phonon-like and the other is clearly photon-like. The curved portions of the curve truly represent mixed modes, i.e. polaritons.

$(\epsilon_s/\epsilon_\infty)^{1/2}$ in NaI and KBr crystals. The mismatch is well within the limits of the experimental error. We will be back with the LST relation in Chapter 11, where it will be found useful in explaining the occurrence of ferroelectricity.

10.11 APPLICATION TO THE MOTION OF ELECTRONS IN POLAR CRYSTALS

Imagine an electron inside a polar (partially or completely ionic) crystal. The electrostatic field owing to the electron displaces each ion to a small extent. Thus a spatial distribution of lattice polarization is set up about the electron. The role of the neighbouring positive ions is more significant because they surround the electron with a positive screening cloud. As the electron moves, its polarization cloud must accompany it. Of course, the same ions cannot move along with the electron; the cloud is made up of just the neighbouring ions and these ions change as the electron moves. This way a deformation or a strain field propagates through the crystal with the moving electron. This quasielectron which is a system of the real electron plus the polarization cloud or the deformation field is called a *polaron*.

The coupling between an electron and its deformation field is more strong in ionic crystals than that in covalent crystals where the interaction of the electron with the neutral atoms is very weak. The polaron behaves as a heavier particle than the electron since the motion of a polaron may be interpreted as the motion of an electron that drags the heavy ion cores along with it.

The creation of a polaron is viewed as a consequence of the electron–lattice interaction which in analytical terms amounts to the electron–phonon interaction. The relatively low frequency longitudinal optical phonons are relevant to the theory of the polaron. The response of the lattice is limited dynamically by their frequency ω_L, that refers to the natural oscillations of dielectric polarization of the crystal. When the energy of the electron is more than $\hbar\omega_L$, an optical phonon is created with the electron getting decelerated and strongly scattered. Even for the slow moving electrons, we must

consider the virtual excitation and re-absorption of quanta in the optical modes, subscribing to the picture in which the electron is surrounded by a cloud of virtual phonons. The strength of the electron–phonon coupling is measured by a dimensionless interaction parameter,

$$\alpha = \frac{2(\text{the deformation energy})}{\hbar\omega_L} \tag{10.124}$$

where $\alpha/2$ can be visualized as the number of phonons that surround a slow moving electron in the crystal.

For small values of α, perturbation theory gives the following relation for the effective mass of the polaron:

$$m^* = m\left(1 + \frac{1}{6}\alpha\right) \tag{10.125}$$

Here m is the effective mass of the electron in the undeformed crystal. The value of m^* is determined from the cyclotron resonance studies. The values of α and m^*/m in a number of crystals have been given by Kittel.† In KCl the values are 3.97 and 2.5, respectively.

The electron associated with a polaron, with properties described above, moves in a band. Such a polaron is identified as a large polaron. In a small polaron, the associated electron remains mostly localized or trapped on a single ion. Both electrons and holes may be self-trapped by the asymmetric local deformations induced by them in the crystal lattice. This is feasible in a polar crystal when the band edge is degenerate. Since the valence band edge is more often degenerate than the conduction band edge, the holes are more likely to be self-trapped. They may, however, be released by thermal excitation. Such a small polaron is so sensitive to the graininess of the lattice that it moves only by hopping from cell to cell. The concept of a valence band in an ionic crystal thus becomes meaningless.

SUMMARY

1. The polarization **P** in terms of the macroscopic electric field **E** of the Maxwell's equation is given by

$$\mathbf{P} = (\epsilon - 1)\epsilon_0\mathbf{E}$$

2. The macroscopic field **E** is equal to the average of the electric field within the volume of the dielectric.

3. The electric field at an atomic site is

$$\mathbf{E}_{loc} = \mathbf{E}_0 + \mathbf{E}_{dep} + \mathbf{E}_L + \mathbf{E}_{dip}$$

where

$\mathbf{E}_0$ is the external field

$\mathbf{E}_{dep}$ is the depolarization field

$\mathbf{E}_L$ is the Lorentz field, the field at the centre of the fictitious cavity owing to the induced charges on the imaginary surface of the cavity.

$\mathbf{E}_{dip}$ is the field at the centre of the cavity owing to dipoles within the cavity.

† C. Kittel, *Introduction to Solid State Physics*, 7th ed., p. 298 (John-Wiley, 1996).

4. The macroscopic field

$$\mathbf{E} = \mathbf{E}_0 + \mathbf{E}_{dep}$$

with $\mathbf{E}_{dep} = -\frac{1}{3\epsilon_0}\mathbf{P}$ for a sphere.

5. $\mathbf{E}_L = \frac{1}{3\epsilon_0}\mathbf{P}$, for spherical cavity

6. $\mathbf{p} = \alpha \mathbf{E}_{loc}$
where
$\mathbf{p}$ is the atomic dipole moment
α is the atomic polarizability

7. In electronic polarization the resonance frequency is shifted from the natural frequency of oscillation of the electronic shell (towards a lower value) because of the presence of a large density of polarizable electrons.

8. For a longitudinal optical mode

$$\mathbf{E} = -\frac{1}{\epsilon_0}\mathbf{P} \quad \text{and} \quad \epsilon = 0$$

For a transverse optical mode

$$\mathbf{E} = 0 \quad \text{and} \quad \epsilon = \infty$$

9. The dielectric constant of a cubic ionic crystal is

$$\epsilon(\omega) = \epsilon_\infty + \frac{(\epsilon_s - \epsilon_\infty)\,\omega_T^2}{\omega_T^2 - \omega^2}$$

10. Power dissipated per cubic metre of the dielectric is given by

$$W = \frac{1}{2}\omega\epsilon_0 E_0^2 \epsilon''(\omega)$$

$$= \frac{1}{2}\omega E_0 D_0 \sin\delta; \text{ with } \tan\delta = \frac{\epsilon''(\omega)}{\epsilon'(\omega)}$$

11. The complex index of refraction is given by

$$\bar{n}(\omega) = n + ik = \sqrt{\epsilon(\omega)}$$

n, k are the optical constants.

12. The real and imaginary parts of the dielectric constant are written as

$$\epsilon'(\omega) = n^2 - k^2; \qquad \epsilon''(\omega) = 2nk$$

13. The reflectivity at normal incidence is given as

$$r = \frac{n + ik - 1}{n + ik + 1}$$

and the reflectance at normal incidence is of the form:

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

14. The uniform collective longitudinal oscillations of the electron gas against a background of fixed positive ions is given by

$$\Omega_P = \frac{\omega_p}{\sqrt{\epsilon_\infty}} \quad \text{with} \quad \omega_p = \left(\frac{N_0 e^2}{\epsilon_0 m}\right)^{1/2}, \text{ the plasma frequency}$$

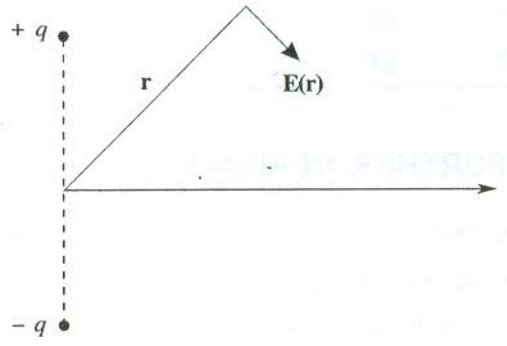
15. A quantum of coupled TO phonon–photon mode is called a *polariton.*

16. The relation $\frac{\omega_L^2}{\omega_T^2} = \frac{\epsilon_s}{\epsilon_\infty}$ is known as the LST relation.

17. The effective mass of the polaron (the combination of an electron and its strain field) is more than that of the electron in an undeformed crystal because of the additional mass associated with the deformation field being dragged by the moving electron in the form of a polarization cloud.

PROBLEMS

10.1 Find the components of the electric field resulting from charges $+q$ and $-q$ in the following figure. Add them vectorially to show that the resultant is given by



$$\mathbf{E}(\mathbf{r}) = \frac{3(\mathbf{p}\cdot\mathbf{r})\mathbf{r} - r^2\mathbf{p}}{4\pi\epsilon_0 r^5}$$

10.2 Assuming that an atom consists of a uniform sphere of negative charge with radius R surrounding a point positive charge, show that the polarizability is equal to $4\pi\epsilon_0 R^3$. Assume that the negative charge remains uniform even under the action of a field. If the diameter of argon atom be 3 Å, calculate the index of refraction of argon gas at NTP.

10.3 An infinite dielectric medium with dielectric constant ϵ has a spherical cavity of radius R cut within it. If a very small conducting sphere of radius r is placed at the centre of the cavity and a uniform electric field E_0 is produced within the main body of the dielectric, what will be the induced dipole moment of the sphere?

10.4 The resonance absorption exhibited by a medium shows a single absorption line at 6000 Å (in vacuum). When a beam of light of this wavelength travels through 2.5 cm in the medium, the intensity of light drops to 1/e of its initial value. Calculate the maximum value of the imaginary part of the index of refraction.

10.5 The complex index of refraction of a metal for infrared radiation ($\omega\tau << 1$) is expressed as

$$\sqrt{\epsilon(\omega)} = (n + ik) = 1 + 4\pi i\sigma_0/\omega \qquad \text{(C.G.S)}$$

where σ_0 is the electrical conductivity for static fields. Check that σ_0 has the unit s^{-1} in e.s.u. Assuming $\sigma_0 >> \omega$ (leading to $n \approx k$), show that the reflectance of the metal approximately equals

$$[1 + (2\omega/\pi\sigma_0)^{1/2}]$$

10.6 Assume that the electrons and the nucleus of an atom form a spherical cloud of uniform density and radius r_0. Taking r_0 as the Bohr radius, calculate the polarizability of hydrogen atom based on this model. How does it compare with the true polarizability?

10.7 Complete the following table, which lists the typical parameters for electrons in four different types of plasma.

	N_0 (m^{-3})	T (K)	ω_p (rad s^{-1})	λ_D (m)
Solar atmosphere	10^{18}	10^4		
Solar corona	10^{13}	10^6		
Ionosphere	10^{10}	10^3		
Tokamak	10^{19}	10^8		

SUGGESTED FURTHER READING

Aschroft, N.W. and N.D. Mermin, *Solid State Physics*, Chap. 27 (Saunders College, 1988).

Ibach, H. and H. Lüth, *Solid State Physics*, Chap. 11 (Springer, 1995).

Ziman, J.M., *Electrons and Phonons*, Chaps. IV, V (Oxford Univ. Press, 1967).

Ziman, J.M., *Principles of the Theory of Solids*, Chaps. 5, 8 (Cambridge Univ. Press, 1972).

Chapter 11

Ferroelectric Crystals

The natural primitive cells of certain dielectric crystals can possess nonvanishing electric dipole moment, even in the absence of applied electric field. The crystal as a whole will thus have a polarization implying that it is spontaneously polarized. The spontaneously polarized dielectric crystals are called *pyroelectric crystals.* Though a proper interpretation for the spontaneous polarization is not straightforward, simple details of crystal structure surprisingly come handy in predicting if a crystal would be pyroelectric. When the centre of positive charge does not coincide with the centre of negative charge in a primitive cell, the primitive cell will possess an electric dipole moment even when there is no external electric field. This is, perhaps, the simplest way to look at the spontaneously polarized nature of pyroelectric crystals.

The shifting of the centre of positive charge from the centre of negative charge is exhibited in the lack of centre of inversion symmetry in the crystal. Out of 32 crystal point groups, 21 point groups do not have a centre of symmetry. Excepting one, which is very highly symmetric, the rest 20 point groups represent an extremely useful class of materials, commonly known as *piezoelectrics.* The piezoelectric crystals show electric polarization on being externally strained and conversely, show deformation when placed under the influence of an applied electric field. If it is confirmed that a crystal belongs to the class of 20 point groups under discussion, it can be predicted that the crystal would be piezoelectric. Among these 20, there are only 10 such groups in which the crystals are spontaneously polarized to be identified as pyroelectrics. At this point we have to examine the consequences of the following requirement on pyroelectric crystals. While maintaining the crystalline properties, the symmetry operations of a pyroelectric crystal must preserve the direction of polarization vector **P**. This imposes severe restrictions on the point group symmetries as a result of which only 10 point groups are found to meet the conditions of pyroelectric crystals. The rotation is allowed about only one axis that is parallel to **P** and there cannot exist mirror planes perpendicular to this axis. The structural scrutiny of crystal point groups (see Table 1.2) reveals that only the following point groups meet the restrictions of pyroelectric crystals:

$$C_n,\ C_{nv}\ (n = 2, 3, 4, 6),\ C_1 \text{ and } C_{1h}$$

Thus the pyroelectric property too, like piezoelectricity, is solely determined by the symmetry properties of crystals.

The polarization in pyroelectric crystals is usually masked by surface charges that accumulate on the surface from the atmosphere and subsequently neutralize the layers of ions. But, when the temperature of the crystal is altered, the masking is no longer complete as the polarization changes because of the thermal expansion or contraction of the crystal. Owing to the thermal effect on polarization, these crystals are named pyroelectric (pyro means fire). The thermal effect accompanying deformations thus corroborates the piezoelectric property of these crystals. This only confirms that all pyroelectric crystals are piezoelectric, though the converse is not true. The description of pyroelectric

crystals remains grossly incomplete without the mention of ferroelectric crystals which are only part of the group of pyroelectrics. The ferroelectric crystals, however, have an additional property that the polarization in them can be changed and even reversed by an external electric field. On the other hand, the same is not possible in pyroelectrics even with the maximum electric field that may be applied without causing electrical breakdown. The additional feature of ferroelectrics that distinguishes them as a special class of pyroelectrics does not follow from the characteristics of crystal structure. It is established only on the basis of dielectric measurement.

Furthermore, the additional feature of ferroelectrics mentioned above converts the usual linear relationship between polarization and applied electric field into a hysteresis loop. The hysteresis effect finds a valid interpretation in the multiple domain theory according to which a bulk crystal is constituted of regions with varying direction of polarization. These regions are called *ferroelectric domains* (analogous to ferromagnetic domains).

The ferroelectric property is observed only below a certain characteristic temperature, called the *Curie point.* On heating the crystal above its Curie point, it shows paraelectric behaviour (normal dielectric behaviour). The Curie point is taken as a critical temperature or a transition temperature T_c that characterizes a phase transition. The theory of structural changes that bring about this transition forms the core of this chapter. At this stage it may suffice to treat the ferroelectric state simply as an ordered state of induced dipoles such that the total energy of the crystal in this state is lower than when the crystal shows no spontaneous polarization since the thermal energy at these temperatures is not strong enough to frustrate the ordering of dipoles.

The role of electric dipoles may further be appreciated by referring to an interesting class of solids which may apparently display permanent electric dipole moments. These substances are known as *electrets.* The materials, in fact, are polarized at high temperatures in an electric field, cooled in the field and then removed from the field. The process leaves the polarization 'frozen in'. As the moments thus produced may at the most survive a few years, the stable state of electrets would be unpolarized. Certain types of waxes, plastics and ceramics are some examples of electrets. Their applications are restricted because of the accumulation of surface charges that are attracted by the dipole field.

11.1 REPRESENTATIVE CRYSTAL TYPES OF FERROELECTRICS

The ferroelectric crystals may be classified into certain representative groups. Such a broad classification is given in Table 11.1. The table gives the Curie point T_c and the spontaneous polarization P_s for a number of common ferroelectric crystals. The electric susceptibility χ_E in the paraelectric phase is related to temperature by the Curie–Weiss law:†

$$\chi_E = \frac{C}{T - T_c} \tag{11.1}$$

where C is the Curie constant.

The ferroelectric crystals are also distinguished on the basis of the oscillatory nature of the atomic displacements that destroy the ferroelectric dipole order above the Curie temperature. In the ferroelectric phase of some crystals, the atomic displacements can be viewed as oscillations about a polar site. These oscillations take place about a non-polar site in the paraelectric phase. The phase

† The law gives a similar temperature dependence as that for the paramagnetic susceptibility χ_p. It will be derived in Chapter 13, giving due importance to magnetism which was discovered earlier than ferroelectricity.

Table 11.1 Data† on some representative ferroelectric crystals

Group	*Crystal*	T_c (K)	P_s $Cm^{-2}(10^{-2})$	*At T* (K)
Ilmenites and	GeTe	670	—	—
Perovskites	$LiNbO_3$	1480	71	296
	$KNbO_3$	710	30	600
	$BaTiO_3$	393	26	300
	$SrTiO_3$	32	3	4.2
KDP type	KH_2PO_4 (KDP)	123	4.7	100
	KD_2PO_4	213	5.5	100
	RbH_2PO_4	147	5.6	90
	KH_2AsO_4	97	5.0	78
TGS type	$(NH_2CH_2COOH)_3 \cdot H_2SO_4$ (Triglycine sulphate)	322	2.8	275
Rochelle salt type	$NaKC_4H_4O_6 \cdot 4H_2O$ (Rochelle salt)	296 (upper) 255 (lower)	0.25	275

† *Sources:* F. Jona and G. Shirane, *Ferroelectric Crystals*, Pergamon (1962); E.C. Subbarao, *Ferroelectrics*, **5**, 267 (1973).

transition that brings about this transformation in the nature of oscillations is called a *displacive phase transition*. These crystals are accordingly identified as *displacive type*. The well-known examples of this class are ionic crystals with ilmenite and perovskite structures. GeTe is the simplest ferroelectric crystal having the ilmenite structure (i.e. NaCl structure). We propose to take $BaTiO_3$ as the representative of perovskites for a discussion of its ferroelectric behaviour in Section 11.1.2.

There is another very interesting class of crystals in whose non-ferroelectric state the potential energy function around certain atomic sites is double-well or multiple-well shaped. On the transition to the ferroelectric state the atomic displacements about those sites are executed as oscillations in an ordered subset of the referred potential wells. It involves an order–disorder type of phase transition. Common examples of these crystals, classified as order–disorder type, are some hydrogen bonded solids, namely KDP type crystals. The motion of protons engaged is linked to the ferroelectric behaviour. Neutron diffraction experiments confirm that the phosphate groups are bound together by what is known as a hydrogen bond. The proton prefers one end of the bond in the ferroelectric state because the proton distribution is found to be asymmetrically contracted in this state. The replacement of hydrogen by deuterium in KDP type crystals raises the Curie point in an amazing proportion. Though the increase in the molecular weight is less than 2 per cent, the T_c rises from 123 K to 213 K in the deuterated KDP and from 96 K to 162 K in KD_2AsO_4. This abnormal effect does not have explanation in the domain of classical physics. It is believed to be a consequence of the change in the de Broglie wavelength, affected by the variation in mass.

The above brief account certainly does no justice to the cause of ferroelectric crystals, a large number of which have found important applications. Unfortunately, a complete and befitting treatment demands the description of each crystal separately since there exists no comprehensive theory that may treat them in a collective fashion. For specific description we have selected Rochelle salt and $BaTiO_3$ as the two representative compounds whose ferroelectric properties are uniquely different.

11.1.1 Properties of Rochelle Salt

Rochelle salt ($NaKC_4H_4O_6 \cdot 4H_2O$) has the distinction of being the first compound that showed ferroelectric property. So, it is of historical importance as well. It was first prepared in 1672 by a pharmacist Seignette who lived in Rochelle. It represents the tartaric group of salts whose other well-known members are lithium ammonium tartrate and lithium tantalum tartrate. The most noteworthy characteristic of Rochelle salt is that it is ferroelectric between two temperatures (255 K and 296 K). On account of its two transition temperatures, Rochelle salt becomes a special and peculiar example of ferroelectrics.

The crystal structure of Rochelle salt is somewhat complex. Above 296 K and below 255 K the structure is orthorhombic. It has a monoclinic symmetry in the ferroelectric phase such that the angle β (between the c- and a-axes) differs from 90° and the spontaneous polarization is along the original orthorhombic a-axis. The original orthorhombic a-axis acts as the lone polar axis with two directions ($\pm$) along this axis.

Halblützel* has measured the dielectric constant of Rochelle salt along the three crystal axes over the whole useful range of temperatures. Figure 11.1 gives a logarithmic plot of these values. The Curie–Weiss law applies above 296 K and below 255 K. With the help of the experimental data it is easy to confirm that the two regions have different values of Curie constants. The dielectric constant

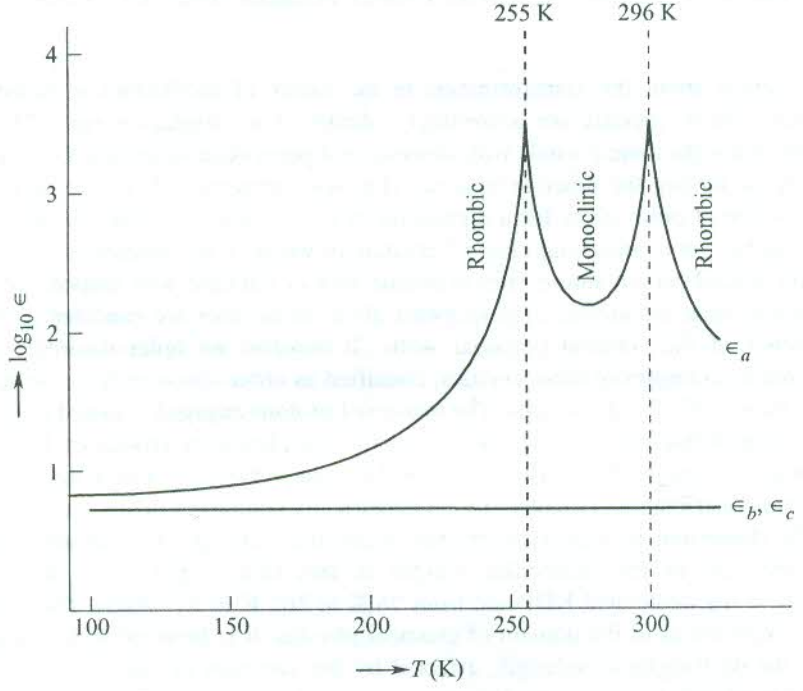


FIG. 11.1 Variation of the dielectric constants of Rochelle salt along a-, b- and c-axes with change in temperature*.

measured along the polar axis ϵ_a peaks at both the transition temperatures, assuming a value as high as 4000. The behaviour of spontaneous polarization* as a function of temperature is shown in Fig. 11.2. The lower curve represents Rochelle salt and the upper curve belongs to the deuterated salt.

* J. Halblützel, *Helv. Phys. Acta*, **12,** 489 (1939).

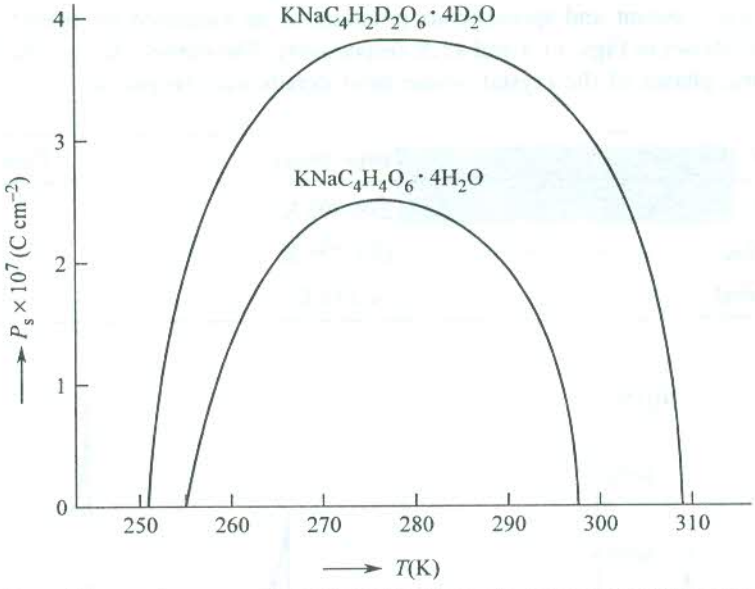


FIG. 11.2 Variation of the spontaneous polarization with change in temperature. The lower curve represents Rochelle salt and the upper curve the deuterated salt.

From the remarkable increase in polarization in the ferroelectric phase of the deuterated crystal, we might be tempted to infer that certain hydrogen bonds are involved in the mechanism of the transition. But the latest structural studies rule out this possibility. No definite theory has been developed so far to account for the ferroelectric properties of Rochelle salt and, therefore, the details of the transition mechanism remain obscure.

11.1.2 Properties of $BaTiO_3$

$BaTiO_3$ is the most important and most completely investigated representative of the perovskite type ferroelectrics. In the non-ferroelectric state (i.e. above 393 K) it has cubic symmetry as shown in Fig. 11.3(a). The Ba^{2+} ions are positioned at the corners, the O^{2-} ions at the centre of the faces and the Ti^{4+} ion is located at the centre of the cube. It has an arrangement of highly polarizable oxygen ions in the form of an octahedron with a small titanium ion at the centre [see Fig. 11.3(b)].

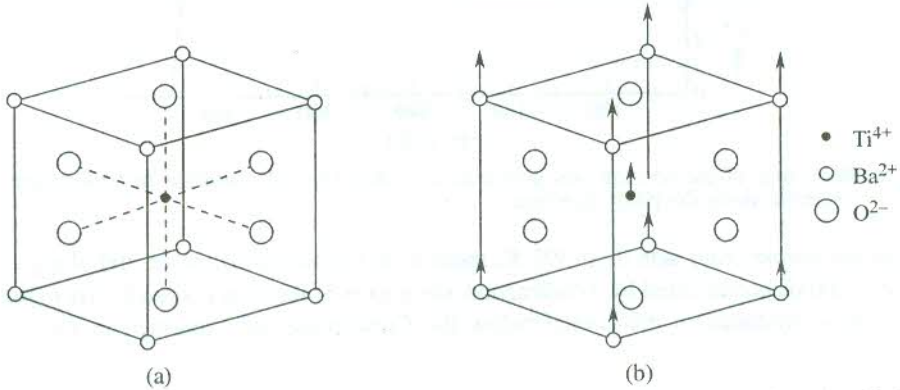


FIG. 11.3 (a) Unit cell of $BaTiO_3$ (perovskite structure). (b) Main distortion in $BaTiO_3$ unit cell that gives rise to ferroelectricity.

The dielectric constant and spontaneous polarization as measured by Merz* over a range of temperatures are shown in Figs. 11.4 and 11.5, respectively. The curves clearly indicate that there are three ferroelectric phases of the crystal whose brief details may be put as:

Crystalline symmetry	*Temp. range*	*Direction of* P_s
Tetragonal	278–393 K	[001]
Orthorhombic	193–278 K	[011]
Rhombohedral	< 193 K	[111]

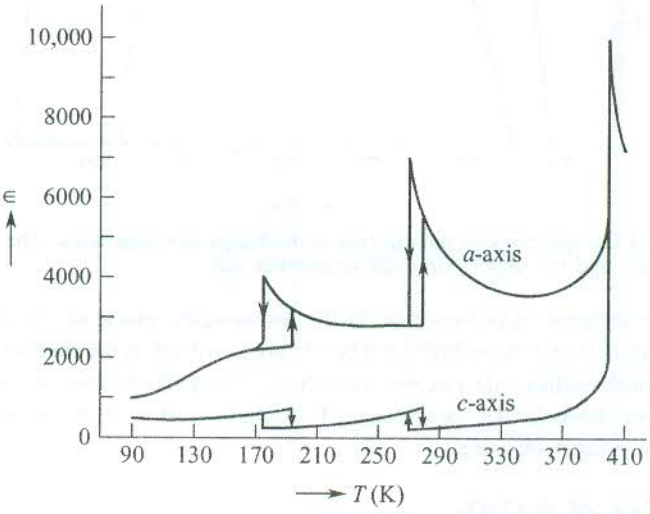


FIG. 11.4 Variation of the dielectric constant of $BaTiO_3$ with change in temperature.*

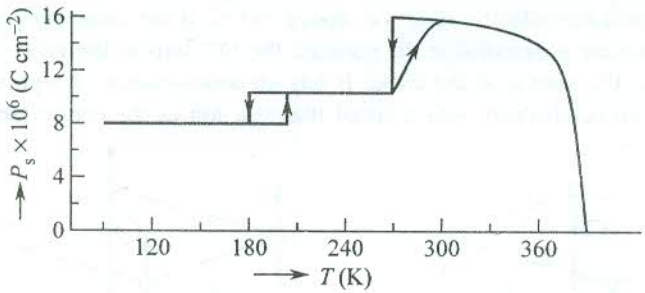


FIG. 11.5 Behaviour of the spontaneous polarization of $BaTiO_3$ with variation in temperature.* The P_s is measured along the [001] direction.

When the dipole order sets in at 393 K, there is an expansion of the crystal along one pseudo-cubic axis (*c*-axis) accompanied by a contraction along each of the axes perpendicular to this direction. The distortions produced in the crystal below the Curie point are explained in Fig. 11.3(b). The

* W.J. Merz, *Phys. Rev.*, **76**, 1221 (1949).

sublattice of all the Ba^{2+} and Ti^{4+} ions is shifted with respect to the sublattice of the O^{2-} ions, the displacement d being barely ~ 0.1 Å at room temperature. This leads to the dipole moment per unit cell p, given by

$$p = 6e \cdot d = 0.96 \times 10^{-29} \text{ C m}$$

The dipole moment p can alternatively be estimated by multiplying P_s (as obtained from Fig. 11.5 at room temperature) by the unit cell volume. Treating the unit cell as a simple cube of edge 4 Å even in the ferroelectric state, we get $p = 0.15(4 \times 10^{-10})^3 = 0.96 \times 10^{-29}$ C m. Thus we find that this value is in perfect agreement with that obtained on the basis of the observed deformation of the unit cell. The order of magnitude gives a measure of the ferroelectric effect in $BaTiO_3$. The effect, however, is fairly large in some other perovskites (e.g. $LiNbO_3$).

The fact that Fig. 11.5 shows P_s along the [001] direction, warrants our further attention. This implies that we must multiply the values shown in the figure by $\sqrt{2}$ and $\sqrt{3}$ to obtain the actual values in the regions 193 K < T < 278 K and T < 193 K, respectively, because the direction of P_s in these regions is along the [011] and [111] directions, respectively. It is then quite interesting to note that spontaneous polarization (same as the saturation polarization) remains almost constant below 300 K.

11.2 THEORY OF THE FERROELECTRIC DISPLACIVE TRANSITIONS

Although we plan to discuss formal thermodynamic theory later in this chapter, the theory that gives a good account of these transitions in perovskite type crystals merits a separate treatment on account of having stood the test of vast experimental data. These crystals generally undergo a displacive transition at the Curie point. We can follow two approaches in the pursuit of finding interpretation to a displacive transition. In one approach we talk of polarization catastrophe that refers to an unusual situation in which the polarization becomes infinitely large. In this condition the force exerted by the local electric field is greater than the elastic restoring force. This produces an asymmetric shift in the positions of the positive and negative ions. The shift is, however, limited to a finite displacement by the anharmonic restoring forces. We can alternatively follow the soft mode approach in which a transverse optical (TO) mode is frozen, i.e. its frequency vanishes at some point in the Brillouin zone below the Curie temperature. This TO mode is known as a soft mode. When $\omega_T = 0$, the crystal becomes unstable because of the absence of an effective restoring force.

11.2.1 Polarization Catastrophe

The Clausius–Mossotti relation (10.38) can be rearranged in the form

$$\epsilon = 1 + \frac{3(N_i\alpha_i + N_e\alpha_e)}{3\epsilon_0 - (N_i\alpha_i + N_e\alpha_e)} \tag{11.2}$$

where all of these symbols have their meaning as defined in Chapter 10. When

$$(N_i\alpha_i + N_e\alpha_e) = 3\epsilon_0 \tag{11.3}$$

the dielectric constant becomes infinite [from (11.2)], giving the state of polarization catastrophe.

Further,

$$\mathbf{P} = (N_i\alpha_i + N_e\alpha_e)\,\mathbf{E}_{loc}$$

$$= (N_i\alpha_i + N_e\alpha_e)\left(\mathbf{E} + \frac{\mathbf{P}}{3\epsilon_0}\right) \tag{11.4}$$

for a cubic crystal (using the Lorentz expression for $\mathbf{E}_{loc}$).

If $\mathbf{E} = 0$, then from (11.4) we have

$$\mathbf{P}\left(\frac{N_i\alpha_i + N_e\alpha_e}{3\epsilon_0} - 1\right) = 0 \tag{11.5}$$

But, when the polarization catastrophe occurs, the quantity within the brackets equals zero [from (11.3)]. This requires that

$$\mathbf{P} \neq 0 \tag{11.6}$$

for (11.5) to be true.

The result (11.6) is obtained with the condition that the macroscopic field $\mathbf{E}$ ($= \mathbf{E}_0 + \mathbf{E}_1$) be zero which is possible only if the applied field $\mathbf{E}_0$ is zero, since then $\mathbf{E}_1$ (the depolarization field, caused by the induced charges) would also be zero.

In order to get an improved physical understanding of the above situation, we consider a highly polarizable ionic crystal of cubic symmetry. Let α denote the total polarizability and p the dipole moment of an ion pair. We assume that some transient stray field starts polarizing the ion pairs. The ion pairs will keep on polarizing until some resistance develops to stop the process. The resistance that finally stops the process of polarization exists in the form of anharmonic restoring forces. The dipole moment of a single ion pair with ion separation $\mathbf{x}$ is

$$\mathbf{p} = q \cdot \mathbf{x} = \alpha\mathbf{E}_{loc} = \left(\frac{\alpha\mathbf{F}}{q}\right) \tag{11.7}$$

where $\mathbf{F}$ is the restoring force that tends to bring the positive and negative ions together and q is the charge on each ion.

The work required to create N such dipoles in the unit volume of the crystal is

$$\varepsilon_1 = N\int \mathbf{F}\cdot d\mathbf{x} = \frac{Nq^2}{\alpha}\int \mathbf{x}\cdot d\mathbf{x} = \frac{Np^2}{2\alpha} \qquad \text{[using (11.7)]} \tag{11.8}$$

$$= \frac{P^2}{2N\alpha}$$

On the other hand, the energy density associated with the electrical displacement brought about by $\mathbf{E}_{loc}$ is given by

$$\varepsilon_2 = \int \mathbf{E}_{loc}\cdot d\mathbf{P}$$

$$= \int \left(\mathbf{E} + \frac{\mathbf{P}}{3\epsilon_0} \right) \cdot d\mathbf{P}$$

$$= \frac{P^2}{6\epsilon_0} + \int \mathbf{E} \cdot d\mathbf{P} \tag{11.9}$$

Since ε_1 is set against ε_2, the net energy density of a polarized dielectric is

$$\varepsilon_2 - \varepsilon_1 = \frac{P^2}{2N\alpha} \left(\frac{N\alpha}{3\epsilon_0} - 1 \right) + \int \mathbf{E} \cdot d\mathbf{P} \tag{11.10}$$

This shows that even when $\mathbf{E} = 0$, $\varepsilon_2 > \varepsilon_1$, provided that

$$N\alpha \geq 3\epsilon_0 \tag{11.11}$$

The above condition in a general case is written in the form

$$\sum_j N_j \alpha_j \geq 3\epsilon_0 \tag{11.12}$$

where N_j stands for the density of the jth type of particles (ions/electrons) in the crystal and α_j denotes the polarizability of a single particle of this type.

The sign of equality in (11.12) describes the condition of polarization catastrophe (11.3) with

$$\sum_j N_j \alpha_j \equiv N_i \alpha_i + N_e \alpha_e \tag{11.13}$$

From relation (11.10) it follows that the energy of the crystal is smaller in the presence of induced dipoles than it is without them. The minimum value of $\sum_j N_j \alpha_j$ for which the ferroelectricity would occur is $3\epsilon_0$. Unfortunately, a simple situation that exactly corresponds to the polarization catastrophe has not been found in any real ferroelectric crystal. However, a small deviation in the value of $\sum_j N_j \alpha_j$ from $3\epsilon_0$ changes the value of ϵ (11.2) by a large amount. When we express $\sum_j N_j \alpha_j$ as

$$\sum_j N_j \alpha_j = 3\epsilon_0 - 3\beta \tag{11.14}$$

with $\beta << 1$ and using (11.2), we get

$$\epsilon \propto \frac{1}{\beta} \tag{11.15}$$

If we assume that β is a linear function of temperature near the Curie point and given by

$$\beta = \frac{T - T_c}{\eta} \tag{11.16}$$

η being a constant, then

$$\epsilon \propto \frac{1}{T - T_c} \tag{11.17}$$

The temperature dependence of ϵ as given by this relation is in excellent agreement with the observed behaviour in several perovskite crystals (Fig. 11.6).

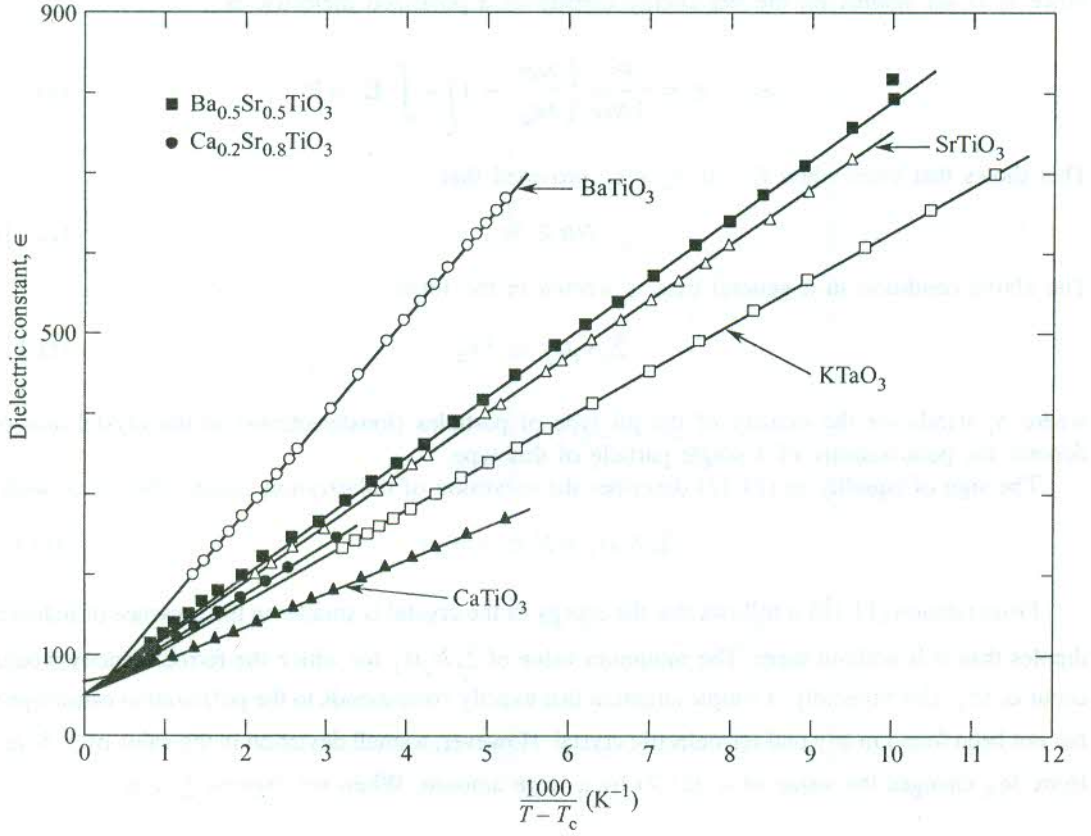


FIG. 11.6 Plot of the measured values of the dielectric constant versus $1/(T - T_c)$ in the non-ferroelectric phase ($T > T_c$) for certain perovskite crystals. [After G. Rupprecht and R.O. Bell, *Phys. Rev.*, **135**, A 748 (1964).]

The occurrence of ferroelectricity in perovskite crystals is understood conclusively to a large extent in view of the following two points, made in respect of barium titanate:

1. **The titanium ion motion.** The barium ions situated at the cube corners leave a big hole at the body centre position. Being smaller than the barium ion, the titanium ion is unable to fill the hole and is free to rattle around in the hole. Because the ionic polarizability is a measure of the ease of displacement, its value is greatly increased. In this case the potential well, in which the titanium ion moves, is somewhat flatter (Fig. 11.7) than the one under the condition of ordinary ionic bonding.

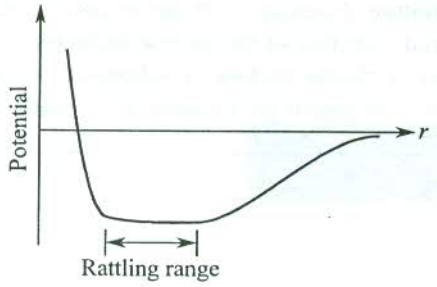


FIG. 11.7 Rattling of titanium ion in a potential well over a large volume, giving rise to a poorly defined equilibrium position of the titanium ion.

2. **The non-cubic symmetry around oxygen ions.** Unlike the barium and titanium ions, the oxygen ions are in the non-cubic environment. An oxygen ion has only two nearest neighbours in the form of titanium ions. Because of this reason, $\mathbf{E}_{loc}$ is not given by the Lorentz expression $(\mathbf{E} + \mathbf{P}/3\epsilon_0)$ and instead found to be greater than this value.

A larger value of α envisaged under point 1 leads to a smaller value of the deformation energy ε_1 or the work required to create induced dipoles. Similarly, a larger value of $\mathbf{E}_{loc}$ as expected under point 2 implies that the dipolar attraction energy will be larger. Thus both the directions of change in α and $\mathbf{E}_{loc}$ are favourable to the onset of ferroelectricity.

11.2.2 'Frozen in' Transverse Optical Phonons

As stated earlier, a ferroelectric state can be regarded as a 'frozen in' TO phonon. According to the LST relation (10.123)

$$\frac{\omega_{TO}^2}{\omega_{LO}^2} = \frac{\epsilon_\infty}{\epsilon_s}$$

showing that when the static dielectric constant ϵ_s increases, ω_{TO} decreases. Thus in the case of an infinitely large ϵ_s, which happens at the Curie point (T_c), ω_{TO} may even be zero. In practice, however, ϵ_s remains finite on approaching T_c. The TO modes in question are called *soft modes*. Such TO modes have surprisingly low frequencies. For example, $BaTiO_3$ has a soft mode of frequency 12 cm^{-1} at 297 K which is low for a TO mode. We are not concerned here with LO phonons whose frequency is higher for the same value of the wavevector. At the transition point T_c when ω_{TO} approaches the zero value, the crystal becomes unstable and anharmonic elastic forces come into play. In the presence of anharmonic forces, ω_{TO} may show a temperature dependence of the form

$$\omega_{TO}^2 \propto (T - T_c) \tag{11.18}$$

On assuming that ω_{LO} are temperature independent, the LST relation in view of (11.18) gives

$$\frac{1}{\epsilon_s} \propto (T - T_c) \tag{11.19}$$

Experiments on several perovskite ferroelectrics strongly support the view that a large static dielectric constant (ϵ_s) is associated with a low frequency TO phonon (the soft mode). In view of

(11.18) and (11.19) the temperature dependence of the square of energy of a low frequency TO phonon can be directly compared with that of the inverse dielectric constant, as shown in Fig. 11.8 for a $KTaO_3$ crystal.* To have a clearer picture, a schematic representation of the temperature dependence of $\epsilon_s^{-1}, \omega_{TO}^2$ and the saturation polarization P_s is made in Fig. 11.9.

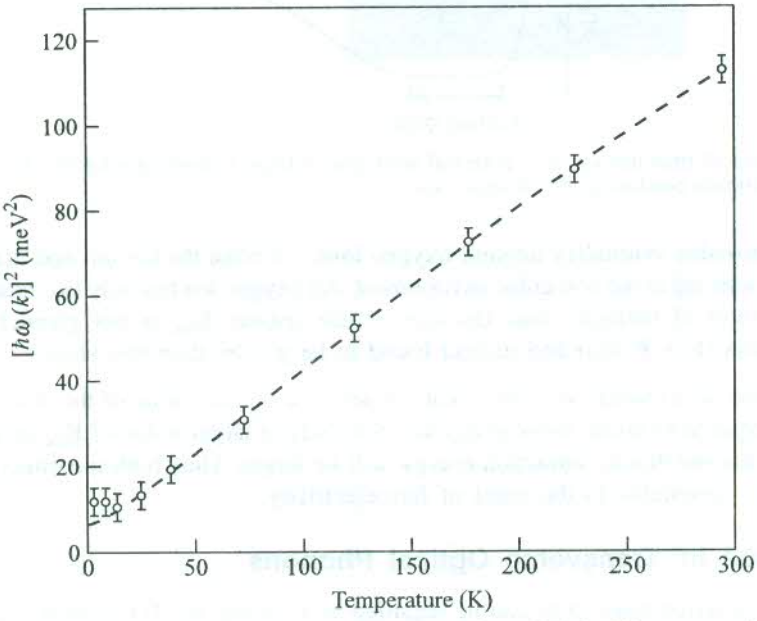


FIG. 11.8 Temperature dependence of a low frequency TO mode in $KTaO_3$. The square of the phonon energy (points) is compared with the reciprocal of the dielectric constant (dashed line).

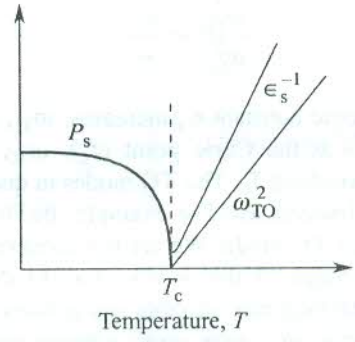


FIG. 11.9 Schematic view of the temperature dependence of the spontaneous polarization P_s, the reciprocal dielectric constant ϵ_s^{-1} and the square of the TO mode frequency ω_{TO}^2 for a ferroelectric crystal.

11.3 THERMODYNAMIC THEORY OF THE FERROELECTRIC TRANSITION

The features of a general thermodynamic theory as developed by Devonshire† are described in this

* G. Shirane et al., *Phys. Rev.*, **157**, 393 (1967).

† A.F. Devonshire, 'Theory of Ferroelectrics' in *Adv. Phys.*, **3**, 85 (1954).

section. Let x denote the relative displacement of the centres of the positive and negative charges in the crystal during a particular mode of vibration. If F_0 be the free energy of the unpolarized crystal, the free energy of the polarized crystal F is a function of the even powers of x. That is,

$$F - F_0 = \phi_2 x^2 + \phi_4 x^4 + \phi_6 x^6 + \ldots \tag{11.20}$$

The constants ϕ are functions of all other displacements and given by their thermal average values. They are thus functions of temperature. Since the electric polarization P is proportional to the displacement x, we have

$$F - F_0 = \frac{1}{2}\lambda_2 P^2 + \frac{1}{4}\lambda_4 P^4 + \frac{1}{6}\lambda_6 P^6 + \ldots \tag{11.21}$$

The constants λ are automatically the functions of temperature. The numerical factors are introduced to facilitate calculations.

Consider first the paraelectric phase of the crystal, i.e. for $T > T_c$. If a small electric field E is applied in the absence of any external pressure, the following thermodynamic relation holds good:

$$dF = -S\,dT + E\,dP \tag{11.22}$$

where S represents the entropy of the crystal.

Since for small E, P will also be small, and hence we retain only the first term in (11.21) neglecting all other terms in the first approximation. Then, using (11.22) we have

$$E = \left(\frac{\partial F}{\partial P}\right)_T = \lambda_2 P \tag{11.23}$$

We may define the electric susceptibility χ_P in the paraelectric phase by

$$1/\chi_P = \epsilon_0 \left(\frac{dE}{dP}\right) = \epsilon_0 \lambda_2 \qquad \text{[from (11.23)]} \tag{11.24}$$

(see 10.12)

Using the Curie–Weiss law (11.1), we have

$$\epsilon_0 \lambda_2 = \frac{T - T_c}{C}$$

or

$$\lambda_2 = C_1(T - T_c) \tag{11.25}$$

where C_1 is another constant.

Relation (11.25) shows that λ_2 increases linearly with the rise in temperature. As a result of this temperature dependence, λ_2 varies from positive values to negative values as the temperature is lowered from above T_c to below T_c.

In the state of thermal equilibrium, the free energy is minimum which requires that

$$\left(\frac{\partial F}{\partial P}\right)_T = 0$$

Applying this condition to (11.21) in the absence of the applied electric fields, we have

$$\lambda_2 P + \lambda_4 P^3 + \lambda_6 P^5 + \ldots = 0 \tag{11.26}$$

The spontaneous polarization is bound to satisfy (11.26) and

$$P_s(\lambda_2 + \lambda_4 P_s^2 + \lambda_6 P_s^4 + \ldots) = 0 \tag{11.27}$$

We find that $P_s = 0$ is always a root of (11.27). For this solution the free energy has a minimum provided λ_2 is positive $\left(\frac{\partial^2 F}{\partial P^2} = \lambda_2\right)$. However, if λ_2, λ_4 and λ_6 are all positive and higher order terms are neglected, the condition (11.27) is satisfied only for $P_s = 0$. Thus $P_s = 0$ corresponds to the only minimum of the free energy and the paraelectric phase exists for the positive sign of λ_2, λ_4 and λ_6.

When the temperature is lowered through the transition point, λ_2 goes from positive to negative values while passing through $\lambda_2 = 0$ at the transition point. There are two interesting situations that are identified in terms of the signs of λ_2, λ_4 and λ_6. These characterize the two different types of transitions, namely the second-order and first-order types.

11.3.1 Second-order Transitions

When λ_2 varies from positive to negative as the temperature is lowered and λ_4, λ_6, . . . are all positive, the free energy changes as shown in Fig. 11.10(a). Assuming that the terms beyond the second term in (11.27) are negligible, we get

$$P_s^2 = -\frac{\lambda_2}{\lambda_4} = \frac{C_1(T_c - T)}{\lambda_4} \tag{11.28}$$

Hence P_s is a continuous function of temperature and falls continuously to zero at $T = T_c$ as shown in Fig. 11.10(b).

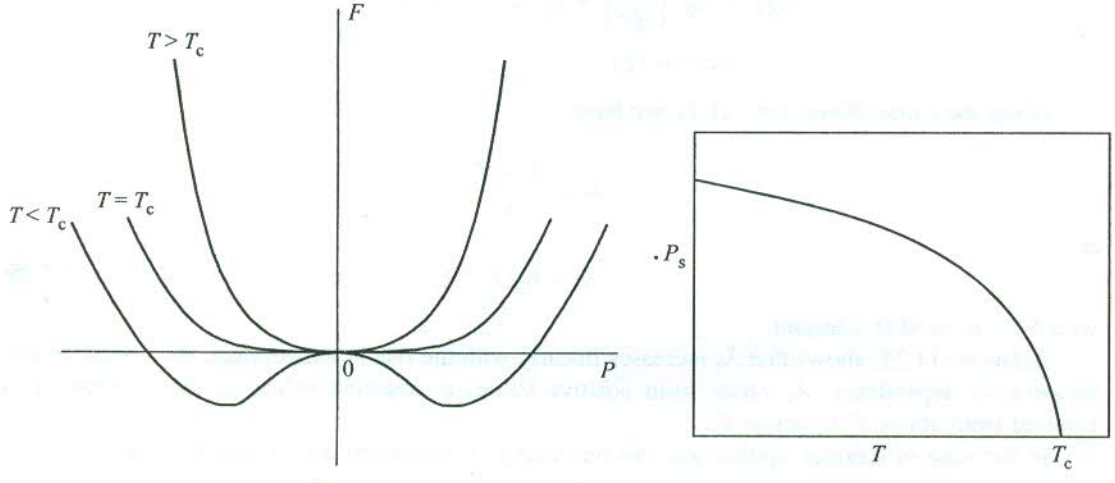


FIG. 11.10 (a) Free energy as a function of polarization as the temperature is varied near a second-order phase transition. (b) Temperature dependence of the spontaneous polarization (the order parameter) below the transition temperature T_c in a second-order phase transition.

It is useful to examine the spontaneously polarized state in terms of the frequency of normal modes. From the forms of the free energy (11.20) and (11.21), it follows that

$$\phi_2 \quad \text{or} \quad \lambda_2 \propto \omega_i^2(k)$$

and hence in view of (11.25),

$$\omega_i^2(k) \propto (T - T_c) \tag{11.29}$$

where $\omega_i(k)$ is the frequency of the normal mode i (actually a TO mode). The transition takes place when $\omega_i(k) \to 0$. This decrease in the mode frequency is called *softening*. It means that the harmonic restoring forces are becoming very weak, permitting a large displacement which is limited solely by the anharmonic forces. A small positive value of $\omega_i^2(k)$ or λ_2 means that the crystal lattice is soft and close to instability. Below T_c, λ_2 is negative and so is $\omega_i^2(k)$ which implies that the unpolarized lattice is unstable and the crystal is in the spontaneously polarized ferroelectric state.

The understanding of the above type of transition remains incomplete without a discussion on the free energy in the state of thermal equilibrium. We saw a little earlier that the free energy minimum in the ferroelectric phase occurs for a non-zero value of P_s, given by (11.28) and not by $P_s = 0$. At $P_s = 0$, there is a local maximum (since $\lambda_2 < 0$) and not the minimum as in the paraelectric phase. This means that the minimum of the free energy in the ferroelectric state is shifted to a point where $P_s \neq 0$. In other words, the minimum occurs for a non-zero value of the relative displacement (say $\bar{x}$). This implies that the distortion sets in below the Curie temperature and increases proportionally to $(T_c - T)^{1/2}$. Note that both F and P_s are continuous at $T = T_c$, but the slope $\dfrac{\partial P_s}{\partial T}$ is discontinuous.

The free energy minimum refers to the value at the thermal equilibrium F_e which we obtain by putting the value of P_s^2 from (11.28) in (11.21). This gives

$$F_e = F_0 - \frac{C_1^2 (T_c - T)^2}{2\lambda_4} \tag{11.30}$$

The heat capacity is given by

$$C_V = \frac{\partial}{\partial T} T^2 \frac{\partial}{\partial T}\left(\frac{F_e}{T}\right)$$

$$= \frac{C_1^2 T}{\lambda_4} \tag{11.31}$$

It falls discontinuously to zero at $T = T_c$ (see Fig. 11.11). But there is no latent heat at the transition. Since the discontinuities are in the second derivatives of the free energy, such a transition is called a *second-order transition*. In the terminology of the Landau theory of phase transitions the spontaneous polarization P_s serves here as an order parameter. The continuous fall of the order parameter to zero at the transition point [Fig. 11.10(b)] is another significant feature of the second-order transitions.

The transitions in Rochelle salt, KH_3PO_4 and $LiTaO_3$ are some examples of the second-order transition. The transition to the superconducting state is the most popular example of this type of transition.

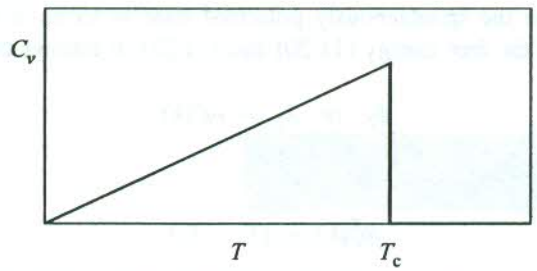


FIG. 11.11 Temperature dependence of specific heat showing anomaly at a second-order phase transition.

11.3.2 First-order Transitions

We saw above that when λ_2 is negative and λ_4 positive, the transition is of the second-order type. We now examine a situation where λ_4 is negative. Positive values of λ_6 are considered to restrain the free energy from going to minus infinity. In accordance with (11.25), λ_2 varies from positive to negative as the crystal is cooled through the Curie point. The corresponding free energy curves are depicted in Fig. 11.12(a). The thermal equilibrium condition, $\partial F/\partial P = 0$, in the absence of the applied electric field gives

$$\lambda_2 P_s + \lambda_4 P_s^3 + \lambda_6 P_s^5 = 0 \tag{11.32}$$

which implies that either $P_s = 0$, or

$$\lambda_2 + \lambda_4 P_s^2 + \lambda_6 P_s^4 = 0 \tag{11.33}$$

At $T = T_c$, the free energy in the paraelectric state is equal to that in the ferroelectric state, i.e.

$$F_0(T_c) = F(T_c) \tag{11.34}$$

Using (11.34) in (11.21), we have

$$0 = \frac{1}{2}\lambda_2 P_s^2(T_c) + \frac{1}{4}\lambda_4 P_s^4(T_c) + \frac{1}{6}\lambda_6 P_s^6(T_c) + \ldots \tag{11.35}$$

Then, using (11.33), we get

$$\lambda_2 + \lambda_4 P_s^2(T_c) + \lambda_6 P_s^4(T_c) = 0 \tag{11.36}$$

Substituting the value of λ_2 from (11.36) in (11.35) and solving for $P_s^2(T_c)$, we get

$$P_s^2(T_c) = -\frac{3}{4}\left(\frac{\lambda_4}{\lambda_6}\right) = \frac{3}{4}\frac{|\lambda_4|}{\lambda_6} \tag{11.37}$$

and with

$$\lambda_2 = \frac{3}{16}\left(\frac{\lambda_4^2}{\lambda_6}\right) \tag{11.38}$$

$$P_s^4(T_c) = \frac{3\lambda_2}{\lambda_6} \tag{11.39}$$

Note that in Fig. 11.12(a) at the transition point there are two minima of free energy with equal value; one at $P_s(T_c) = 0$ in the paraelectric phase and the other for the value of $P_s(T_c)$ given by (11.37) in the ferroelectric phase. Thus there is a jump [see Fig. 11.12(b)] in the value of P_s at T_c, meaning thereby that the spontaneous polarization (the order parameter) drops discontinuously to zero at $T = T_c$ when a ferroelectric crystal is heated slowly. Such transitions are called the *first-order transitions.* The other important property of these transitions is that there is a latent heat at the transition. A well-known example of this type of transition is the upper transition in a $BaTiO_3$ crystal.

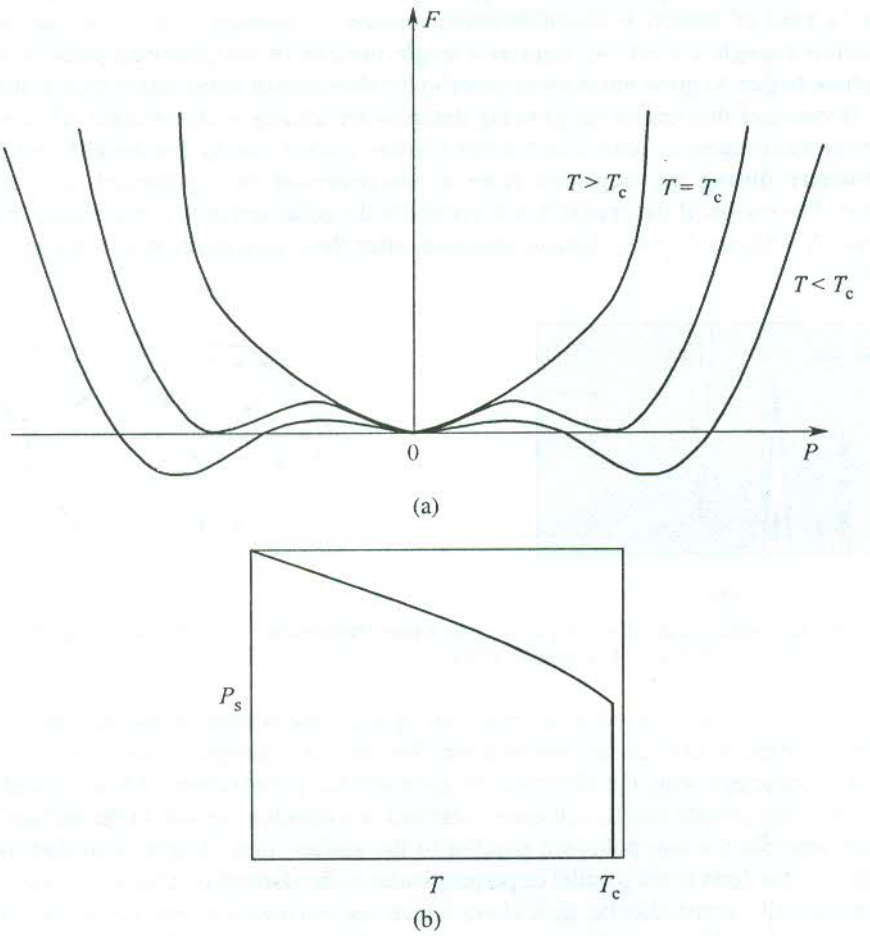


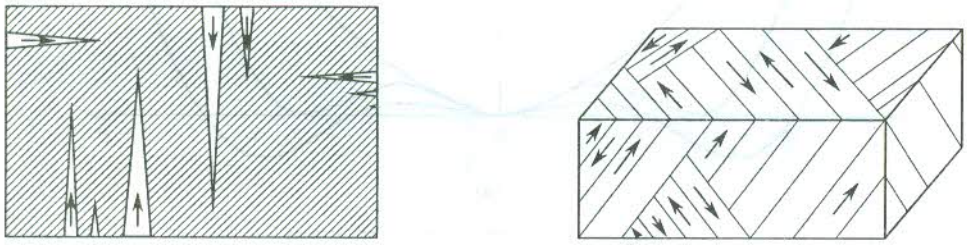
FIG. 11.12 (a) Free energy as a function of polarization as the temperature is varied near a first-order phase transition. (b) Fall of the spontaneous polarization (the order parameter) below the transition point T_c in a first-order phase transition.

11.4 FERROELECTRIC DOMAINS

A ferroelectric crystal need not necessarily show an observable evidence of spontaneous polarization. When the crystal is cooled from the paraelectric phase through the Curie temperature, the polarized phase may be nucleated at several points in the crystal. These nuclei normally differ in the direction of polarization since there may be several equivalent crystallographic directions in which the

spontaneous polarization can occur. For example, in $BaTiO_3$ the spontaneous polarization may occur along one of the three edges, giving six possible directions for the spontaneous polarization. Thus, as the nuclei grow through the crystal in the ferroelectric phase, they form several regions or domains differing in their direction of polarization. The vector sum of these polarizations may not be always big enough to show up macroscopically.

Polarization is accompanied by some distortion of the unit cell and the domain walls are consequently in a state of strain. The dimensional changes, however, are relatively small. Though the domain walls act as interruptions in the regularity of the crystal, they are not regarded as grain boundaries (a kind of defect) between different crystals. A domain wall is, instead, treated as a subgrain within a single crystal. As soon as a single nucleus of the polarized phase is formed, the polarized phase begins to grow much more faster in the direction of polarization than in the transverse directions. Because of this reason the growing domains are usually wedge-shaped. This was revealed by the optical birefringence studies on barium titanate carried out by Forsbergh* and Merz.† The domain structure during the transition from a non-polarized to a polarized state is shown in Fig. 11.13(a). The region of the crystal that is yet not in the polarized state is represented by the cross-hatched area. A schematic of the domain structure, after the transformation is complete, is given in Fig. 11.13(b).



(a) (b)

FIG. 11.13 (a) Schematic of the domain growth into a non-ferroelectric (paraelectric) phase. (b) Schematic of the domain structure in a single crystal.

As referred above, it is possible to make an optical observation of the ferroelectric domains because the ferroelectric crystals are birefringent. We take the example of barium titanate in which the optic axis coincides with the direction of spontaneous polarization. When viewed through a microscope between crossed nicols, a domain polarized in a direction normal to the surface of a crystal appears dark whereas the one polarized parallel to the surface looks bright, provided the direction of polarization of the light is not parallel or perpendicular to the domain polarization. As a consequence of the three mutually perpendicular axes along which the polarization may occur, the directions of polarization in the neighbouring domains are either at 90° or at 180°. These directions have given origin to the concept of 90° and 180° walls. A wall separating regions of antiparallel polarization is called a 180° wall. A 90° wall may be defined similarly.

The ferroelectric domains are regarded as the electrical analogues of the ferromagnetic domains despite the fact that there are some interesting differences in their origin and growth. On application of an electric field to a ferroelectric crystal, the number and size of domains that are polarized in the

* P.W. Forsbergh, Jr., *Phys. Rev.*, **76,** 1187 (1949).
† W.J. Merz, *Phys. Rev.*, **88,** 421 (1952).

field direction increase. As a result of this effect, upon the reversal of the field direction a hysteresis in the P versus E curve is observed. The size of ferroelectric domains increases because of the growth in the forward direction (the direction of polarization) and not sideways. Practically, there is little sideways motion of the domain walls. But in ferromagnetic domains it is just the reverse. A change in the direction of magnetization is achieved because of the growth of favourably-oriented domains such that the growth is facilitated mainly by a sideways motion of the domain walls. Another significant difference between the two sets of domains, lies in the thickness of a domain wall. It is estimated by calculating the energy per unit area of the wall and then minimizing it with respect to the thickness of the wall. The thickness of a ferromagnetic domain wall is found to be about 300 lattice constants in contrast to a considerably small value of the order of a few lattice spacings for a ferroelectric domain wall. Accordingly, the energy associated with a wall in the former case is relatively much small.

11.5 ANTIFERROELECTRICITY

Like ferroelectrics there is another group of solids which has induced, ordered electric dipoles below a characteristic temperature but shows no spontaneous bulk polarization. In these crystals the neighbouring lines of atoms are associated with antiparallel polarizations because of which the bulk polarization of the crystal vanishes. Crystals exhibiting this property are called *antiferroelectric* crystals and the property is known as *antiferroelectricity.* The structural requirements for the ferroelectric and antiferroelectric phases being common, a number of well-known antiferroelectric crystals are found to be isomorphous with some ferroelectrics. For example, ammonium dihydrogen phosphate (ADP) which is antiferroelectric is isomorphous with potassium dihydrogen phosphate (KDP).

Perovskite type crystals are known to be susceptible to several types of deformation with almost equal energy difference between them. In many of them the coupling through the oxygen octahedra causes adjacent lines of basic cells to be polarized in opposite directions. Below a certain temperature the resultant deformation is such that the total energy in the antiparallel arrangement of adjacent lines of dipoles is lower, when compared separately to that in the state of fully parallel arrangement of dipoles and to that in the state with no induced dipoles. Lead zirconate ($PbZrO_3$) is a notable example of these perovskites. It shows two antiferroelectric phases, one each ferroelectric and paraelectric phase over different ranges of temperature.

11.6 PIEZOELECTRICITY

Requirements for a crystal to show piezoelectric behaviour were discussed in the beginning of the chapter. The piezoelectric crystals get electrically polarized when stressed and conversely get strained when polarized (placed in the electric field). This phenomenon is called *peizoelectricity* and it is of great technological importance. We showed earlier that all ferroelectrics are piezoelectrics and that its converse is not true. For example, quartz is piezoelectric but it does not possess the ferroelectric property.

The foremost condition for a crystal to be piezoelectric is the absence of the centre of symmetry. Figure 11.14(a) shows the array of a simple two-dimensional ionic crystal with no centre of symmetry. It is evident that a compressive force **F** [Fig. 11.14(b)] decreases the electric dipole moment (hence the polarization) and a tensile force **F** [Fig. 11.14(c)] increases the same. This is essentially the piezoelectric effect. We must appreciate that the displayed crystal [Fig. 11.14(a)] could well be a ferroelectric crystal.

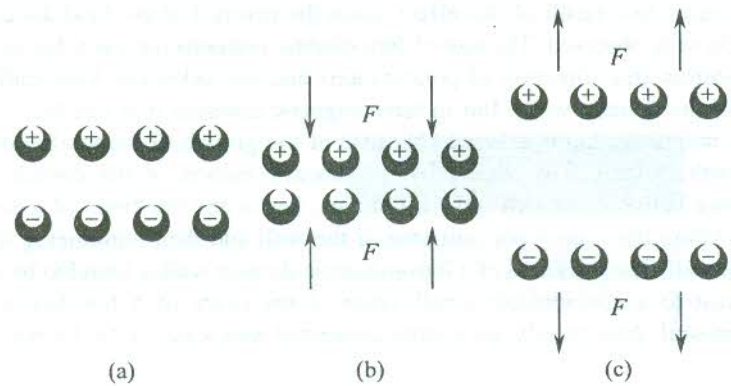


FIG. 11.14 (a) A two-dimensional ionic crystal with no centre of symmetry. (b) Compression under the action of force F decreases the polarization. (c) Extension of structure under the action of force F increases the polarization.

Next we take up another example to show how the symmetry of a non-centrosymmetric crystal controls firstly the magnitude and direction of polarization when the crystal is stressed and secondly the crystal dimensions when the crystal is polarized. Consider a molecule of a hypothetical ionic solid which at equilibrium has three electric dipoles of equal magnitude distributed over 360° at an interval of 120° as shown in Fig. 11.15(a). The molecule belongs to the point group 3 m and its net dipole moment is zero. But if the molecule together with the crystal is stressed or compressed along a direction parallel or antiparallel to one of the three directions of the dipole moment, a net dipole moment would appear [see Fig. 11.15(b) and (c)]. Similarly, a molecule may be distorted by an electric field applied along one of the three arrows shown in the Fig. 11.15(a). The electric field produces an elongation or contraction of the crystal along the field direction and a length change of opposite sign in the lateral direction. An applied field that is perpendicular to one of the three dipole directions in Fig. 11.15(a) finds itself perpendicular to a mirror plane of symmetry and, therefore, is rendered ineffective in changing the crystal dimensions.

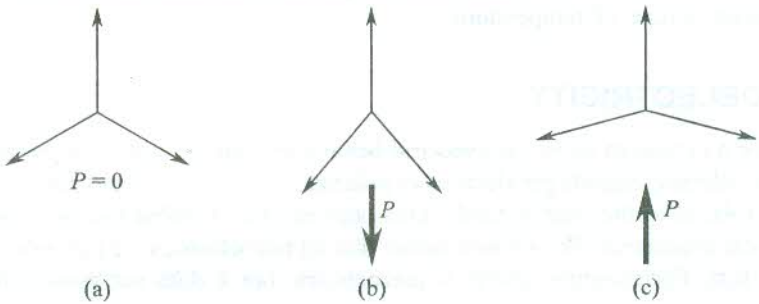


FIG. 11.15 Response of a piezoelectric molecule to strain: (a) Directions of polarization (in accordance with symmetry) in a molecule within an undistorted crystal in the state of equilibrium. The net polarization of the molecules is zero. (b) A vertical tension or a horizontal compression causing a net polarization. (c) A vertical compression or a horizontal tension causing a net polarization.

Because of lack of centre of inversion and complex structure of piezoelectrics, their electrical behaviour under strain or strain behaviour under an electric field is not isotropic in nature. Nevertheless,

a simple picture of the phenomena can be presented in a schematic one-dimensional notation by the following equations:

$$P = \sigma d + \in_0 E\chi; \qquad e = \sigma s + Ed \tag{11.40}$$

where P is the polarization, σ the stress, d the piezoelectric strain constant, $\in_0$ the permittivity of free space, E the electric field, χ the dielectric susceptibility, e the strain and s the elastic compliance constant.

In real crystals, however, the tensile, compressional or shear strains produced by an electric field may develop in different directions and depend on the crystal orientation and the field direction. In view of this fact the piezoelectric strain constants, that form a third rank tensor, are defined as

$$d_{ik} = \left(\frac{\partial e_k}{\partial E_i}\right)_\sigma$$

where $i \equiv x, y, z$ and $k \equiv xx, yy, zz, xy, yz, zx$.

Depending on the application and the desired behaviour, a crystal is cut so as to have the parallel faces of the crystal in a specific orientation. An X-cut is defined as a section cut from the crystal such that the x-axis of the crystal is perpendicular to the parallel crystal faces. The axes in some common piezoelectric crystals are shown in Fig. 11.16. In order to obtain certain desirable properties the crystals are sometimes given oblique cut, i.e. cut at angles different from 90° with the principal axes.

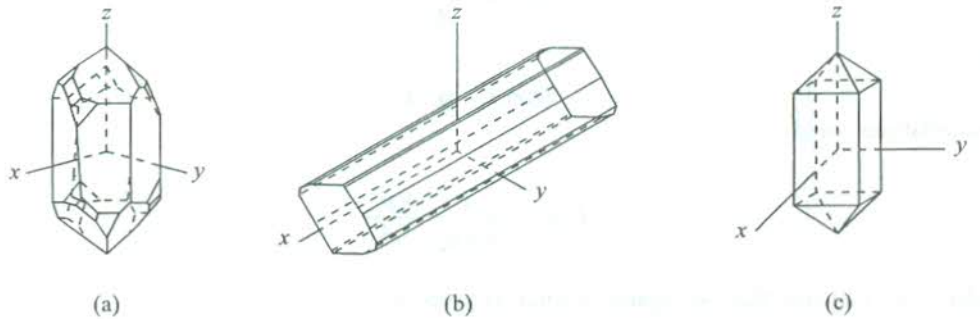


FIG. 11.16 Orientations of axes that are used to describe some common piezoelectric crystals: (a) Quartz, (b) Rochelle salt, and (c) ADP.

11.7 ELECTROSTRICTION

It may not be out of place to discuss a more universal phenomenon of deformation in crystals that is caused by an applied electric field. It refers to the deformations in ionic crystals and the effect is commonly known as *electrostriction*. In the first approximation, the deformation of a piezoelectric crystal is proportional to the applied electric field and the stress induced polarization varies linearly with the strain produced (11.40). But in ionic crystals, which do not have to be necessarily piezoelectric, the strain is much smaller and proportional to the square of the electric field. We can understand the origin of electrostriction by appreciating that the dipoles created by the applied electric field would interact with each other (Fig. 11.17). The in-line dipoles attract each other with a repulsive force acting perpendicular to the direction of polarization.

Let p denote the moment of a dipole, and r the separation between two neighbouring in-line dipoles. The value of the electric field caused by a dipole at its in-line neighbour may be written as

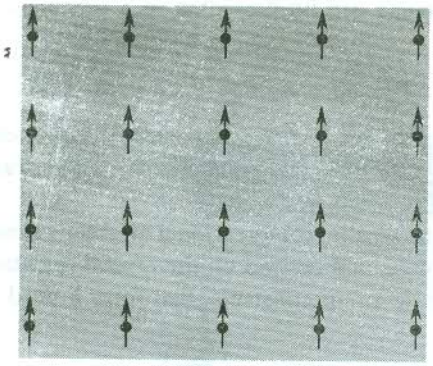


FIG. 11.17 Production of electrostriction by attraction among the in-line dipoles. This attraction exceeds the repulsion from the side-by-side dipoles.

$$E = \frac{1}{4\pi \epsilon_0} \frac{2p}{r^3} \tag{11.41}$$

The energy of a dipole in this field $U(r)$ and the corresponding attractive force F are related as

$$F = -\frac{dU(r)}{dr}$$

and

$$U(r) = -\mathbf{p} \cdot \mathbf{E}$$

These relations yield

$$F = -\frac{1}{4\pi \epsilon_0} \frac{6p^2}{r^4} \tag{11.42}$$

Similarly, we can find that the repulsive force is given by

$$F = \frac{1}{4\pi \epsilon_0} \frac{3p^2}{r^4} \tag{11.43}$$

Since $p = \alpha E$, the attractive force can be expressed as

$$F = -\frac{1}{4\pi \epsilon_0} \left(\frac{6\alpha^2}{r^4}\right) E^2 \tag{11.44}$$

To a first approximation the strain or deformation u may be assumed to follow the Hooke's law and then

$$u = -\frac{F}{f}$$

where f is the usual force constant in the direction of the in-line dipoles.

With the use of (11.44) in the above relation, we get

$$u = \frac{1}{4\pi \epsilon_0} \left(\frac{6\alpha^2}{fr^4} \right) E^2 \tag{11.45}$$

Thus there will occur a compression in the field direction and an extension perpendicular to the field direction. The above treatment holds good for permanent dipoles as well on account of the effective dipole moment being proportional to the electric field.

11.8 APPLICATIONS OF PIEZOELECTRIC CRYSTALS

It may be recalled that all ferroelectrics are piezoelectrics, though the converse is not true. As a result, ferroelectric materials have been frequently used in many applications that are based on the principle of piezoelectricity. But, because of the importance of properties such as mechanical and thermal strength the use of certain piezoelectric crystals (e.g. quartz which is non-ferroelectric) becomes inevitable. There are a very large number of piezoelectric applications whose complete listing at this place is a formidable task. We describe below only a few of them which span all the significant areas of applications:

1. Crystals shaped to have a prescribed mechanical resonance frequency are used as narrow band electrical filters. Only those electrical signals whose frequency coincides with the mechanical vibrational frequency pass through the crystal and all others are rejected.
2. When an a.c. voltage at one of the resonant frequencies of the crystal is impressed perpendicular to a certain pair of parallel faces, the amplitude of oscillation at this frequency grows. The constancy of the elastic constants of the crystal controls the stability of the oscillator. Specially-cut quartz discs are used for this purpose. Due to a very low value of the coefficient of thermal expansion for quartz, a fairly high level of stability is achieved.
3. Piezoelectric oscillators are used to convert mechanical pulses into electrical ones and vice versa. The crystal in these devices works as a transducer. The acoustic pulses are used in underwater search (sonars) and other applications. The acoustic pulses are generated by the piezoelectric transducers, excited by electric fields in almost all such cases. The generation of ultrasonic waves is invariably accomplished by exploiting the above principle.
4. The highest possible efficiency is a desirable attribute of every device. Rochelle salt, being a sensitive and inexpensive transducer, has been used in phonographs. A shaped ceramic block of barium strontium titanate or lead zirconium titanate (PZT) should, however, be preferred since these materials are very sensitive and resistive to heat and moisture.
5. The piezoelectric effect in synthetic polyvinylidenfluoride (PVF_2) is about five times stronger than that in quartz. Being flexible and easy to handle like ultrasonic transducers, the PVF_2 films are frequently used in applications such as monitoring blood pressure and respiration.
6. The piezoelectric materials are used as delay lines. When an electrical signal is converted into an acoustic one at one end of a quartz rod, the signal passes along the rod as an acoustic wave, travelling at the velocity of sound. At the other end, the acoustic wave may be converted into an electrical signal. The initial signal is thus delayed. Such an arrangement is often used in communication devices.

SUMMARY

1. Pyroelectric, ferroelectric and piezoelectric crystals lack centre of inversion symmetry.
2. A pyroelectric crystal belongs to one of the following point groups:

$$C_n,\ C_{nv}\ (n = 2, 3, 4, 6),\ C_1 \text{ and } C_{1h}$$

3. All ferroelectric crystals are piezoelectric, but the converse is not true.
4. Polarization catastrophe occurs when

$$\sum_j N_j \alpha_j = 3\epsilon_0$$

5. The soft modes are low frequency TO modes (e.g. 12 cm^{-1} in $BaTiO_3$ at 293 K). The soft mode frequency tends to zero as the transition temperature is approached.
6. The free energy of a ferroelectric crystal is given by

$$F - F_0 = \frac{1}{2}\lambda_2 P^2 + \frac{1}{4}\lambda_4 P^4 + \frac{1}{6}\lambda_6 P^6 + \ldots$$

where
F_0 is the free energy of the unpolarized crystal
F is the free energy of the polarized crystal
P is the electric polarization
$\lambda_2, \lambda_4, \lambda_6, \ldots$ are constants that depend on temperature.

7. A transition is of the second-order type if λ_2 is negative and $\lambda_4, \lambda_6, \ldots$ are positive. The order parameter (P_s in ferroelectrics) falls continuously to zero at the transition point. The specific heat drops discontinuously to zero at the transition point.
8. A transition is of the first-order type if λ_4 is negative and λ_2, λ_6 are positive. The order parameter falls discontinuously to zero at the transition point. There is a latent heat at the transition.

PROBLEMS

11.1 Explain why ZnS could be piezoelectric but not diamond, although the atomic arrangement is the same in both the crystals.

11.2 Two dipoles, each having a polarizability α, are at a fixed distance a. If they form a ferroelectric state, set up the relationship between a and α.

11.3 Consider an ionic crystal with cubic symmetry at every lattice point. Calculate its ferroelectric Curie point taking the total polarizability of the crystal as given by

$$\alpha = \left(0.5 + \frac{100}{T}\right) \times 10^{-30}\ m^3$$

11.4 Atoms of polarizability α are arranged in a line with a as the interatomic separation. Prove that the array can polarize spontaneously if

$$\alpha \geq \frac{a^3}{4\sum n^{-3}}$$

where the sum is carried out over all positive integers.

11.5 The $SrTiO_3$ crystal has the perovskite structure, but is not ferroelectric. Determine a likely reason on the basis of a comparison between the barium and strontium ions.

11.6 The resonant frequency of a quartz crystal is given by

$$\nu_0 = \frac{1}{2l}\sqrt{\frac{Y}{\rho}}$$

where l is the dimension that determines the mode of oscillation of the crystal, Y is the Young's modulus and ρ is the density of quartz. Determine the useful sizes of the crystal for oscillators in kHz and MHz ranges (for quartz, $Y = 10^{10}$ N m^{-2}, $\rho = 2500$ kg m^{-3}).

SUGGESTED FURTHER READING

Bruce, A.D. and R.A. Cowley, *Structural Phase Transitions* (Taylor and Francis, 1981).

Fröhlich, H., *Theory of Dielectrics* (Oxford, 1986).

Lines, M.E. and A.M. Glass, *Ferroelectrics and Related Materials* (Oxford, 1977).

Toledano, J.-C. and P. Toledano, "Landau Theory of Phase Transitions: Application to structural, incommensurate, magnetic and liquid crystal systems" (*World Scientific*), 1987.

Zheludev, I.S., *Physics of Crystalline Dielectrics* (Plenum, 1971).

Chapter 12

Imperfections in Crystals

It is a well-established fact that there exist no perfect crystals. Every crystalline structure shows some deviation or the other from the regular atomic arrangement, as prescribed by the symmetry and structure of the respective unit cell. These deviations from the ideal crystal structure are called *imperfections*. Imperfections could be of several types, as we will see below. The presence of imperfections or defects is not accounted by the translational symmetry of the perfect crystal though it forms the basis of most of the interpretations in perfect crystals. We may obviously be curious to know how the translational symmetry in a crystal is consistent with laws of thermodynamics since these laws are applied to describe the growth of crystals all of which have some imperfections. We know that the Helmholtz free energy

$$F = U - TS \tag{12.1}$$

must be minimum in the state of equilibrium at a certain temperature. In (12.1) U stands for the internal energy and S for the entropy. We take the advantage of the following statistical statement of entropy,

$$S = k_B \ln W \tag{12.2}$$

where W is the number of possible ways in which the elements of a system may be distributed.

Now, we take the case of a defect that can be localized at a lattice point. In a perfect crystal there can be only one way (i.e. $W = 1$) to arrange atoms at different sites and, therefore, the entropy in this case would be zero [see (12.2)]. On the other hand, a defect at a site within a unit cell makes the unit cell look different from others. In this case there can be as many ways of arranging the defect as the number of sites within the unit cell. The entropy is given by

$$S = k_B \ln N \tag{12.3}$$

where N is the number of sites in the unit cell.

Relation (12.3) expresses the contribution of the defect to the entropy of the crystal. Thus any increase in the defect concentration raises the entropy which in turn lowers the free energy at a finite temperature. In the equilibrium state there is a finite concentration of imperfections in the crystal. In the above example we considered only one type of defect. But, as a necessary consequence of the inherent disorder associated with finite temperature, all kinds of imperfections (one can imagine) could be present, though some of them might be very small in number. The concentration of a particular type of imperfection depends on the type of the crystal lattice, the binding energy of the lattice and the structure of the imperfection itself. The imperfections are crucial to the interpretation of several properties of crystals that are not accounted by the translational symmetry. To name a few: colour of crystals, enhancement of conductivity of pure semiconductors, plasticity, strength of crystals, luminescence and diffusion of atoms in solids are some such significant examples.

Imperfections play a pivotal role even in the very existence of the solid state. It will be clear through discussions on the strength of crystals and the process of crystal growth, taken up towards the end of the chapter.

For a systematic description of imperfections, we classify them broadly as follows:

1. Point imperfections
2. Line imperfections
3. Planar imperfections

12.1 POINT IMPERFECTIONS OR POINT DEFECTS

The well-established forms in which point defects are found in crystals are the following:

(a) **Vacancy.** A normal lattice site from where the atom or ion is missing, is known as *Schottky defect* (Fig. 12.1).

(b) **Interstitial atom.** An atom located at a position that is not a normal lattice site (Fig. 12.1), known as *Frenkel defect.*

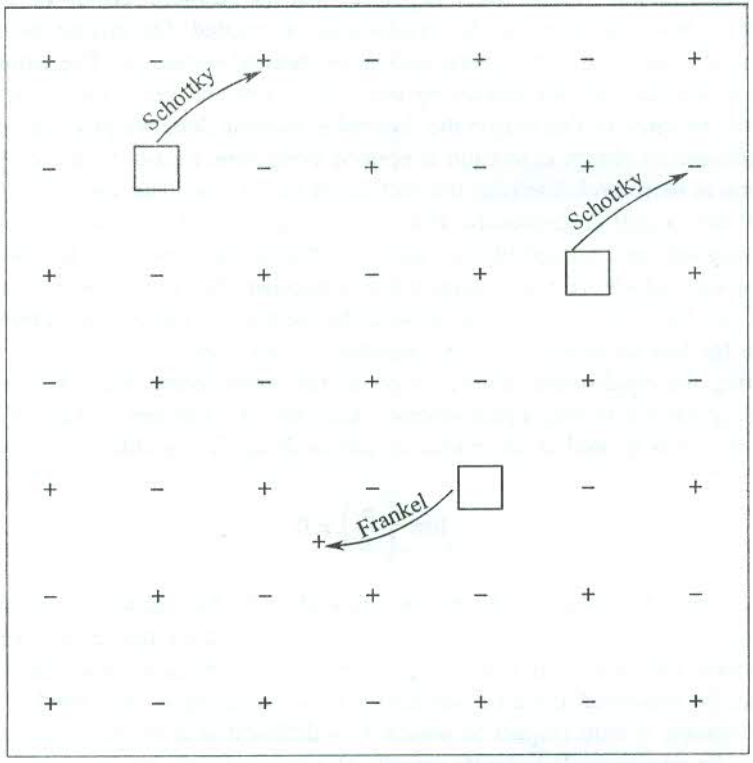


FIG. 12.1 Formation of Schottky and Frenkel defects in a diatomic ionic crystal—arrows indicate the direction of displacement of ions. The creation of a Schottky defect results in the production of a pair of positive ion and negative ion vacancies to meet the charge neutrality requirement. The formation of a Frenkel defect implies the creation of a pair of a lattice vacancy and an interstitial position. The symbol □ denotes a vacancy.

(c) **Colour centre.** A bound system results from a charge compensation mechanism. This bound system, known as *colour centre*, has a quantized scheme of energy levels and absorbs visible light in some crystals. These crystals thus show colour which are otherwise colourless.

(d) **Excitons.** Quantized electron–hole pair excitations.

(e) **Polarons.** Quantized polarization waves (introduced in Section 10.10)

A detailed account of these point defects is given below.

Schottky defects

A Schottky defect as described above is a vacancy at a normal lattice site which is otherwise occupied by an atom or ion in a perfect crystal. We base our discussion in this subsection on ionic crystals with NaCl structure which are easy to handle theoretically on account of their simple structure. In this case there can be two kinds of Schottky defects—one type representing the cation vacancies and the other corresponding to the anion vacancies. The two types of vacancies must be equal in number to maintain the charge neutrality of the crystals. Therefore, if a Schottky defect in a unit cell is a cation vacancy, there must be another unit cell within the crystal which has an anion vacancy representing the Schottky defect of the other type.

Suppose we have found a crystal which is perfect near the absolute zero of temperature. We may be curious to know how vacancies in this crystal can be created. On raising the temperature the thermal agitations are excited and the crystal undergoes thermal expansion. The whole of the thermal energy supplied is not used up for increasing the lattice constant, but a part of it is consumed in vacancy formation. Because of this reason the thermal expansion does not proceed continuously with rise in temperature and an abrupt expansion is noticed every time a pair of vacancies is formed. The vacancy formation is more probable near the surface of the crystal. The vacancies from the surface flit in and out of the crystal continuously. But as the temperature rises, more vacancies diffuse in, than those diffusing out until an equilibrium state is achieved. It is easier to follow the diffusion of vacancies than of ions which are many times more in number. The present situation is analogous to that of a nearly filled band where it is easier to describe the transport of charge in terms of the motion of holes that are far less in number than the number of electrons.

For calculating the equilibrium concentration at any finite temperature, it is assumed that the number of vacancy pairs n is very much smaller than the total number of sites N available in the crystal. The number n is treated as an extensive thermodynamic variable satisfying the condition

$$\lim_{N \to \infty} \left(\frac{n}{N} \right) \neq 0$$

implying that the limit does not vanish, but is required to be less than unity since $n << N$. This condition is always met for point defects in crystals. As the vacancy pair concentration approaches its equilibrium value, the Helmholtz free energy F touches a minimum value. Thus the equilibrium concentration can be evaluated thermodynamically by minimizing F. We need to express F as a function of the variable n with respect to which F is differentiated in the process of applying the condition for F to be minimum. If E_v be the energy of a vacancy pair, the increase in internal energy owing to production of n pairs is

$$U = nE_v \tag{12.4}$$

For calculating entropy S we have to determine the number of ways of picking n molecules from N sites. For the first molecule there are N ways, for the second $(N-1)$, for the third $(N-2)$ and for the nth $(N-n+1)$. This gives the number of ways as

$$N(N-1)\,(N-2)\,\ldots\,(N-n+1)$$

Since the order of taking molecules does not matter, we divide the above expression by $n!$ making the number of ways to be equal to

$$\frac{N(N-1)(N-2)\ldots(N-n+1)}{n!} = \frac{N!}{(N-n)!\,n!} \tag{12.5}$$

It should be marked that we have not accounted for the number of ways of picking the components of the pair. These are N for the first pair, $(N-1)$ for the second and so on. Therefore, each factor in the estimate (12.5) must be squared to get the total number of ways W, which is then expressed as

$$W = \left[\frac{N!}{(N-n!)n!}\right]^2 \tag{12.6}$$

and, therefore, from (12.2), we get

$$S = 2k_B \ln\left(\frac{N!}{(N-n)!\,n!}\right) \tag{12.7}$$

Relation (12.7) gives the configurational entropy which is different from thermal entropy. The former is determined by the number of ways the atoms can be distributed over the lattice sites whereas the latter is determined by the number of ways the thermal energy of the crystal can be distributed among the modes of vibration. Vacancies may also have thermal entropy because of the changes in the frequency spectrum or change in volume of the crystal. This contribution being relatively very small is ignored here.

On substituting U and S from (12.4) and (12.7), respectively, into (12.1), we obtain

$$F = nE_v - 2k_BT \ln\left(\frac{N!}{(N-n)!\,n!}\right) \tag{12.8}$$

Using Stirling formula†: $\ln X! \cong X[\ln X - 1]$ in (12.8), we have

$$F = nE_v - 2k_BT\,[N \ln N - n \ln n - (N-n) \ln (N-n)]$$

and

$$\frac{\partial F}{\partial n} = E_v + 2\,k_B\,T \ln\left(\frac{n}{N-n}\right) \tag{12.9}$$

For F to be minimum, $\dfrac{\partial F}{\partial n} = 0$. This gives

† Valid for large values of X.

$$E_v = 2k_B T \ln\left(\frac{N-n}{n}\right)$$

or

$$\left(\frac{n}{N-n}\right) = \exp\left(-\frac{E_v}{2k_B T}\right) \tag{12.10}$$

Since $N >> n$, the above relation reduces to

$$\frac{n}{N} = \exp\left(-\frac{E_v}{2k_B T}\right) \tag{12.11}$$

An estimate of these concentrations near the melting point can be made in alkali halides by using $T = 10^3$ K and $E_v = 1$ eV. This gives the value of n/N as 10^{-5}, i.e. about $10^{24}/m^3$. It is implied in these calculations that the number of cation vacancies is equal to the number of anion vacancies. In no way it is meant that the vacancy pairs are tied together. On account of having an effective charge associated with them, they tend to pair up or move together. But it is not necessary that they exist adjacent to each other. In pure alkali halides the most common vacancies are the Schottky defects.

Frenkel defects

The process of an interstitial occupation is accompanied by the creation of a vacancy. If N_i be the number of interstitial positions in the crystal, the different ways W in which n Frenkel defects can be produced is given by the relation

$$W = \frac{N!}{(N-n)!\,n!} \cdot \frac{N_i!}{(N_i-n)!\,n!} \tag{12.12}$$

The Holmholtz free energy in this case is then expressed as

$$F = nE_i - k_B T \ln\left[\frac{N!}{(N-n)!\,n!} \cdot \frac{N_i!}{(N_i-n)!\,n!}\right] \tag{12.13}$$

where E_i denotes the energy required for removing an atom from a lattice site to an interstitial position.

Proceeding on lines parallel to those for Schottky defects and taking $n << N$ and $n << N_i$, we obtain the following relation

$$n = (NN_i)^{1/2} \exp\left(-\frac{E_i}{2k_B T}\right) \tag{12.14}$$

In pure silver halides, the most common vacancies are the Frankel defects.

It can be shown that vacancies bear the same relation to ionic conduction as holes do to electronic conduction. Lattice vacancies are created in alkali halide crystals with the addition of divalent elements while the growth of the crystal is in progress. For example, when a triatomic molecule like $CaCl_2$ is added to a KCl crystal, a cation vacancy is produced. The Ca^{2+} occupies a

normal K^+ site and the two Cl^- enter the Cl^- sites. On requirement of charge neutrality, one K^+ site near the Ca^{2+} site must be left vacant during the crystal growth (Fig. 12.2).

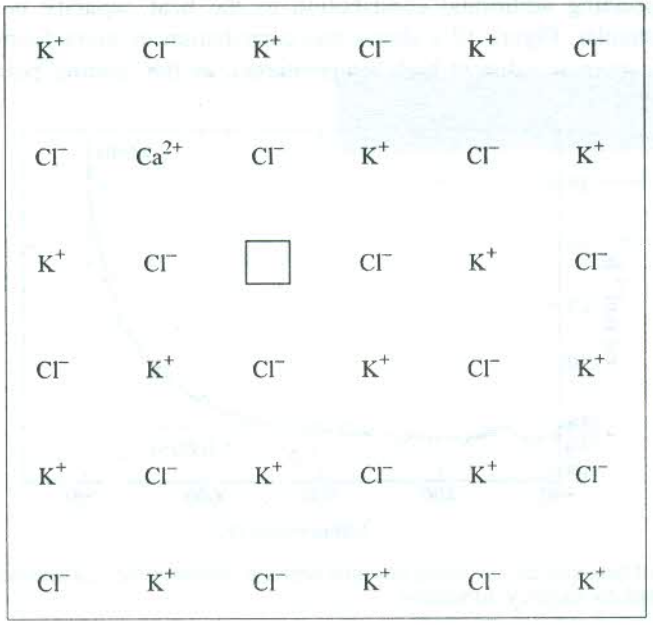


FIG. 12.2 Creation of a positive ion vacancy caused by the addition of traces of $CaCl_2$ in KCl during crystal growth.

We might expect the density of the crystal to increase on account of Ca^{2+} being heavier than K^+. But this can happen only if no vacancies are present. The density in fact is lowered because of the presence of vacancies that effectively increase the volume of the unoccupied space within the crystal. Lattice defects are generally investigated by measuring the ionic conductivity. The data at not too high temperature show that the ionic conductivity of alkali halide and silver halide crystals increases linearly with increase in the amount of divalent element added. While the current flows the monovalent metal deposits at the cathode, establishing that the role of divalent addition is limited to the production of vacancies. The vacancies enhance the diffusion of monovalent atoms forming the basis for current flow. The diffusion of a lattice vacancy is equivalent to the diffusion of the metal ion in the opposite direction. Basic mechanisms of diffusion are shown in Fig. 12.3.

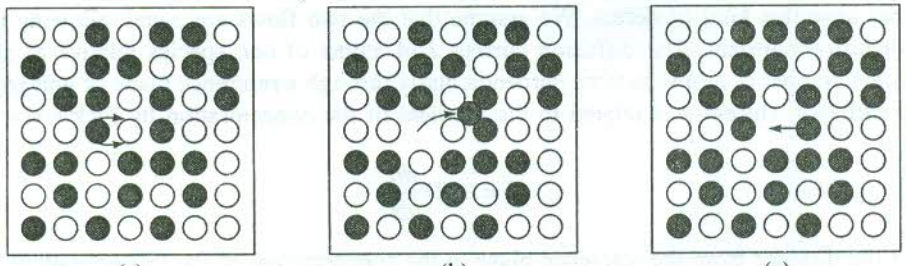


FIG. 12.3 Basic mechanisms of diffusion: (a) Atoms interchange positions by rotation about a midpoint. (b) Atoms move through interstitial sites. (c) Atoms exchange positions with vacant lattice sites.

As vacancies require energy for their formation, they should contribute to the heat capacity. Their formation can be regarded as an extra degree of freedom, in addition to the $3N$ vibrational modes. The corresponding additional contribution to the heat capacity is noticeable near the melting point, in particular. Figure 12.4 shows this contribution in silver bromide as a sudden rise over the value $3Nk_B$ (normal value at high temperatures), as the melting point is approached.

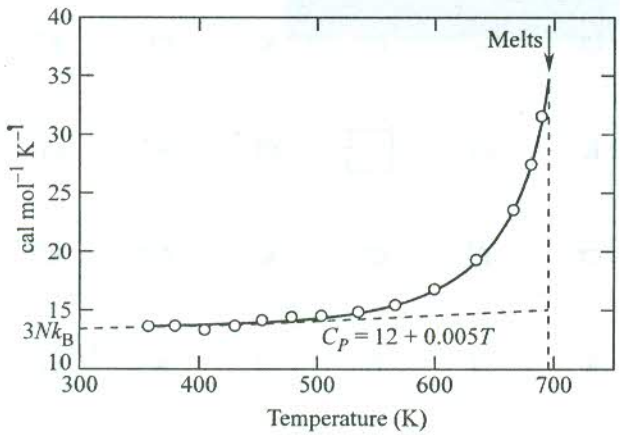


FIG. 12.4 Variation of heat capacity of silver bromide near the melting point. Excess heat capacity (over $3Nk_B$) is attributed to vacancy formation.

Diffusion

In the absence of electric field, lattice defects move in a random fashion. But, on applying a field the defects move with a velocity that depends upon the magnitude of the field. This introduces the concept of mobility μ, defined as the drift velocity per unit electric field. The mobility is related to a quantity known as the *diffusion constant*, D, by the Einstein relation

$$\mu k_B T = qD \tag{12.15}$$

where q is the charge on a defect or an atom whose diffusion may be under study.

The so-called diffusion laws connect the rate of flow of an atom with the concentration or the thermodynamic potential that causes the flow. In general, the diffusion has at least two components, each diffusing in opposite directions. This means that in an interstitial diffusion the empty interstitial spaces form the second component. It would be analogous to the flow of vacant lattice sites in crystals that have this kind of defect. We assume that the two flows are equal, allowing us to use a single diffusion constant. The diffusion current J of atoms of one species relative to another is defined as the number of atoms passing perpendicularly through a reference plane of unit area in unit time. The diffusion current J is related to the gradient of the concentration by Fick's law

$$J = -D\frac{dn}{dx} \tag{12.16}$$

where x is the distance from the reference plane, n the concentration of the diffusing atom and t the time. Thus units of D are m^2/s. The minus sign indicates that diffusion occurs away from regions of higher concentration. It is further assumed that D is independent of concentration in a given medium at a given temperature. This assumption may, however, be true only as a first approximation.

Let us consider the diffusion of impurity atoms between interstitial sites. The treatment of the diffusion of lattice vacancies proceeds on identical lines. For diffusion, an atom is required to migrate over the potential barrier provided by its nearest neighbours. The probability of each try that the atom jumps over the potential hill is $\exp\left(-\frac{E_a}{k_B T}\right)$, E_a being the height of the barrier (or the activation energy). If the characteristic atomic vibrational frequency is ν_0, the atom will make ν_0 tries in one second to cross over the potential barrier. Therefore, the probability per second or the number of times the atom actually crosses the barrier in one second is

$$\nu = \nu_0 \exp(-E_a/k_B T) \tag{12.17}$$

where ν is also known as the *jump frequency*. The other process that may contribute to diffusion is quantum mechanical tunnelling. But it is significant only in the diffusion of the light nuclei.

Consider two adjacent parallel planes having impurity atoms in interstitial sites. When the two planes are separated by one unit of lattice constant a and the total concentration of impurity atoms is denoted by n, the number of these atoms per unit area on one plane will be $na\left(=\frac{na^3}{a^2}\right)$. The corresponding number on the other plane can be written as $\left[na + a\frac{d(na)}{dx}\right]$. Therefore, the number of atoms that move between the two planes in unit time through their unit area is

$$(\nu a^2)\frac{dn}{dx}$$

Thus we get another statement of the Fick's law, written as

$$J = (\nu_0 a^2)\exp\left(-\frac{E_a}{k_B T}\right)\frac{dn}{dx} \tag{12.18}$$

On comparing (12.18) with (12.16), we find that

$$\begin{aligned} D &= (\nu_0 a^2)\exp(-E_a/k_B T) \\ &= D_0 \exp(-E_a/k_B T) \end{aligned} \tag{12.19}$$

This relation gives the temperature dependence of the diffusion constant. Its value shows a great variation for various impurities at a given temperature. For common metals and Ge and Si, its value lies in the range 10^{-50} to 10^{-20} m^2/s. The constant D_0 is experimentally found not to vary much with temperature. Its value for carbon impurity in alpha iron is about 2×10^{-6} m^2/s. The representative values of D_0 and E_a for a number of impurities in various host crystals have been given by Kittel.*

From the ionic conductivity data obtained at different temperatures it is possible to estimate the values of energy of formation of a vacancy pair and the jump activating energy. The conductivity is expressed as

$$\begin{aligned} \sigma &= nq\mu \\ &= \left(\frac{nq^2\nu_0 a^2}{k_B T}\right)\exp\left(-\frac{E_a}{k_B T}\right) \qquad \text{[using (12.15) and (12.19)]} \end{aligned} \tag{12.20}$$

* C. Kittel, *Introduction to Solid State Physics*, 7th ed., Chap. 18, p. 545 (John Wiley, 1996).

In the temperature range over which the number of vacancies is controlled by the number of divalent metal ions, the vacancy concentration is independent of temperature. The slope of ln σ versus $(1/k_B T)$ plot then gives the jump activation energy of positive ion vacancies E_a^+. There is another range of temperature in which thermal activation energy determines the defect concentration and the proportion of vacancies is given by $\exp\left(-\frac{E_v}{2k_B T}\right)$, E_v being the energy of formation of a vacancy pair. The slope of ln σ versus $1/k_B T$ plot in this case provides the value of $\left(E_a^+ + \frac{1}{2}E_v\right)$. Thus the conductivity data over a wide range of temperatures enable us to determine the jump activation energy and the energy of formation of a vacancy pair.

Colour centres

The band gap of ionic crystals is about 6 eV which is equivalent to the energy of a photon of wavelength 2000 Å (i.e. in the ultraviolet region). The ionic polarizability is found to resonate at 50 microns, falling in the far infrared region. Therefore, these crystals are expected to be transparent in the visible region, as is actually the case with alkali halides. But with certain treatment, these crystals can be coloured or made to show colour.

On the requirement of charge neutrality, vacancies of one constituent of a diatomic ionic crystal are balanced either by an equal number of interstitials of the same constituent (Frenkel) or by an equal number of vacancies of the other constituent (Schottky). The missing charge of a negative ion vacancy may also be compensated by an electron localized or trapped† in the vicinity of the point defect whose charge the electron is replacing. The negative ion vacancy together with the trapped electron can be regarded as a bound system that is characterized by a quantized set of energy levels. Similar to the case of single isolated atoms, these levels give rise to a series of absorption lines in the optically forbidden band between $\hbar\omega_{TO}$ and $\hbar\omega_{LO}$, as discussed in Chapter 10. The band covers the visible region. The bound system as described above and other such defect–electron structures, that absorb in the visible region imparting colour to an otherwise transparent perfect crystal, are called *colour centres*.

Colour centres have been extensively studied in alkali halides. They are usually coloured in the following ways:

(i) Adding chemical impurities

(ii) Heating the crystal in the vapour of the alkali metal and then cooling it quickly

(iii) Irradiating the crystal with x-rays, γ-rays, neutron or electron beams

(iv) Conducting electrolysis

By heating the crystal in the vapour of the alkali metal, an excess number of alkali atoms are incorporated into the crystal and as such we could expect the mass density to increase. But the density practically decreases in proportion to the concentration of excess alkali atoms. When an alkali atom from the vapour diffuses into the crystal, it occupies a pre-existing cation vacancy. Since the equilibrium concentration must be maintained, a new cation vacancy is formed. But to maintain charge neutrality an anion vacancy must also be created simultaneously. Therefore, the net result of putting one alkali atom into the crystal is to increase the size of the crystal by one alkali atom and by one anion vacancy. This increases the vacancy concentration causing a decrease in density.

† A negative ion vacancy acts as a centre of a positive charge and, therefore, it can trap an electron.

F-centres

We saw above that the excess alkali atoms enter the crystal at the pre-existing cation vacancies. While diffusing from the surface the atom is ionized and then occupies a cation vacancy. The liberated valence electron is trapped at an anion vacancy and forms a hydrogen atom-like bound system (see Fig. 12.5). This bound system is known as an *F*-centre (*F* stands for Farbe, a German word that means colour). The above model of *F*-centre was originally conceived by de Boer and confirmed by the electron spin resonance studies. Pohl* was the first to study the experimental properties of *F*-centres in detail. Figure 12.6 shows the central absorption band caused by *F*-centre in several alkali halide crystals.

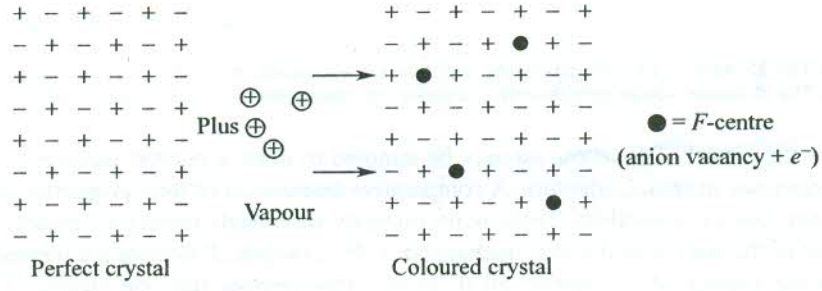


FIG. 12.5 An excess of alkali ions being produced in a crystal by heating it in the vapour of the alkali metal. Highly localized excess electrons occupy the corresponding negative ion vacancies.

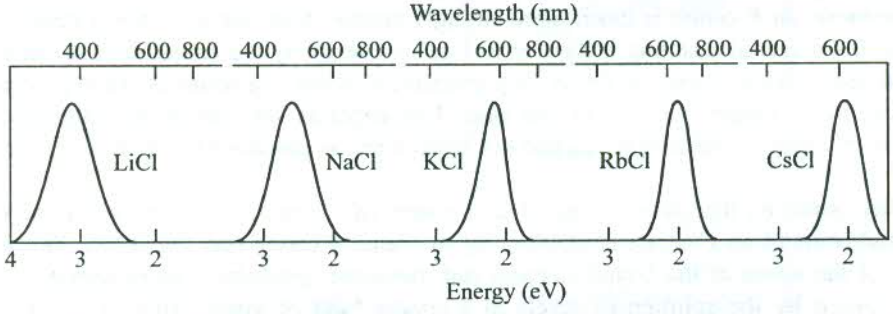


FIG. 12.6 *F*-centre absorption bands in certain alkali halide crystals. [After R.W. Pohl, *Proc. Phys. Soc.*, **49**, 3 (1937).]

The qualitative features of the spectra produced by *F*-centres are strikingly similar to those of ordinary atomic spectra in many respects. However, the basic difference lies in the symmetry around the *F*-centre which is cubic and not spherical as is the case with isolated atoms. The cubic crystal field splits the orbitally degenerate energy levels making the spectra of *F*-centres relatively complicated. A complete analysis using the crystal field theory demonstrates that *F*-centre is not the only way the electrons and vacancies can combine to colour the crystal. There are two other possibilities:

* R.W. Pohl, *Proc. Phys Soc.*, **49**, 3 (1937).

(a) The *M*-centre [Fig. 12.7(a)], in which two electrons are bound to two neighbouring anion vacancies in (100) plane.

(b) The *R*-centre [Fig. 12.7(b)], in which three electrons are bound to three neighbouring anion vacancies in (111) plane.

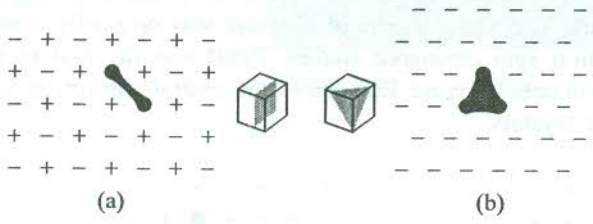


FIG. 12.7 (a) The *M*-centre—two neighbouring negative ion vacancies in a (100) plane binding two electrons. (b) The *R*-centre—three neighbouring negative ion vacancies in a (111) plane binding three electrons.

In view of the model of *F*-centre we may be tempted to draw a parallel between *F*-centres and bound donor electrons in semiconductors. A comparative assessment of their properties demonstrated that they are not exactly equivalent. There is no impurity in crystals having *F*-centres. The F-band is characteristic of the host crystal and is unchanged if, for example, *F*-centres are formed by heating the crystal in the vapour of a different alkali metal. This implies that the electron is no longer associated with the atom which released it in the process of ionization. This contrasts the behaviour of a donor electron that is bound to its parent impurity atom. The electron in an *F*-centre is bound to a different type of imperfection, i.e. the vacancy.

Furthermore, an *F*-centre is much more strongly bound. Also, the *F*-centre is strongly coupled to the rest of the crystal enabling it thereby to lose its energy by the emission of phonons. On the other hand, the isolated atoms can decay only radiatively which is a relatively slow process, i.e. the excited state has a longer lifetime† in this case. This explains why the optical absorption spectra produced by *F*-centres consist of bands and not sharp lines, as produced by the excitation of isolated atoms.

Another centre F_A that may be treated as a cousin of *F*-centre is formed when one of the six nearest neighbours of an *F*-centre is replaced by an impurity cation (see Fig. 12.8). This lowers the symmetry of the levels of the bound electron and, therefore, produces a more complex absorption spectrum caused by the splitting of levels in a crystal field of lower symmetry. It is, however, interesting to note that the formation of an F_A-centre is often energetically favourable.

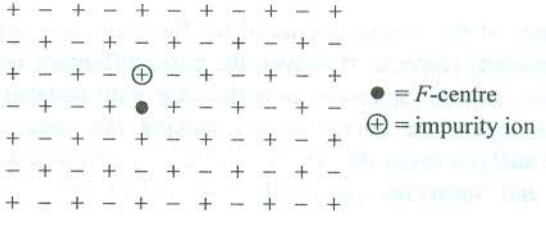


FIG. 12.8 The F_A-centre—an impurity ion replacing one of the six nearest-neighbour positive ions that surround an *F*-centre.

† The linewidth is inversely proportional to the lifetime of the excited state of concern.

The *M*- and *R*-centres described earlier, are examples of more complex centres. They may be regarded as members of the *F*-centre family since they represent the grouping of two and three *F*-centres, respectively.

Trapped-hole centres

By heating an alkali halide crystal in a halogen gas, it should, in principle, be possible to introduce alkali metal vacancies. A hole then could be bound to such a vacancy, producing an antimorph to the *F*-centre. But no such antimorphs are observed in alkali halides. Though holes can be trapped by point imperfections, there exists no evidence for these imperfections to be cation vacancies. A V_k-centre is the best known trapped-hole centre in which a hole binds two neighbouring anions (see Fig. 12.9). The optical absorption spectrum of the V_k-centre is interpretable as that of a negative halogen molecular ion in a crystal (say, Cl_2^- in KCl). Another example is the *H*-centre which is thought to result from the binding by a hole of an interstitial chlorine ion to a symmetrically positioned lattice ion (Fig. 12.10).

FIG. 12.9 The V_k-centre—two neighbouring negative ions bound by a hole.

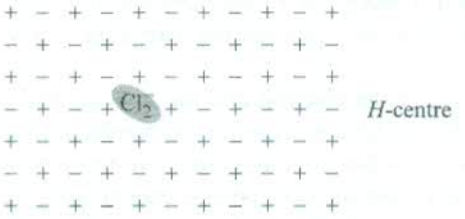


FIG. 12.10 The *H*-centre—an interstitial chlorine ion bound by a hole to a symmetrically situated lattice chlorine ion. The resulting singly-ionized chlorine molecule occupies a single chlorine ion site.

The structure and properties of colour centres are investigated by electron paramagnetic resonance and optical absorption studies. A typical absorption spectrum of colour centres in a KCl crystal is shown in Fig. 12.11. The peaks attributed to different centres are marked accordingly.

Excitons

We described above some of the most common point defects. There is a subtle possibility when an atom in a perfect crystal can be regarded as a point defect. This refers to the situation when only one of the atoms in a perfect crystal is in an excited state and the rest are in the ground state. In this picture the single excited atom is treated as a defect and known as a *Frenkel exciton*. The excitation may be transferred from atom to atom because of the coupling between atoms. On being excited, a valence electron is transferred to the conduction band leaving behind a hole in the

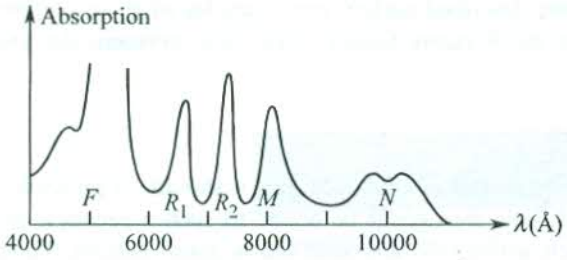


FIG. 12.11 Absorption spectrum of KCl showing peaks caused by various colour centres.

valence band. The electron and the hole may be bound by the Coulomb attractive force acting between them. The bound electron–hole pair is called an *exciton*. The exciton being neutral does not transport charge, though it does carry energy as it moves through the crystal. This means that by the movement of excitons the energy is transported through the crystal without the atoms themselves having to move out from their equilibrium positions. This results in fast movement of excitons. Excitons, in fact, move much more faster than vacancies, interstitial or substitutional impurities.

Two limiting cases of the electron-hole binding give rise to two types of excitons—the Frenkel and Mott–Wannier excitons. Before we take up separate discussions on these excitons, we describe below the general features of exciton levels.

In reflection and absorption spectra of insulators, there is enough evidence to suggest that there exist energy levels immediately below the lower conduction band edge, though the crystals are expected to be transparent in this region. The discrete energy level structure belongs to the exciton that represents a bound electron–hole pair. When a photon of energy greater than the forbidden gap is incident on an insulating crystal, a free electron and a free hole are likely to be produced as already described. The threshold photon energy in a direct absorption process equals the band gap. But in an indirect phonon assisted absorption process the threshold is lower by the phonon energy, as discussed in section 9.2.3.

When the group velocities of the electron and the hole created are equal, i.e. $d\varepsilon_c/d\mathbf{k} = d\varepsilon_v/d\mathbf{k}$ (ε_c and ε_v refer to the electron and hole energies in the conduction and valence band, respectively), the two particles may be bound by the Coulomb attractive force between them and thus an exciton formed. In the process a part of the incident photon energy goes as the binding energy, bringing thereby the exciton energy to a level that is lower than the lower conduction band edge even when the threshold is equal to the band gap E_g. A schematic of the exciton levels is shown in Fig. 12.12 which also illustrates the concept of the exciton binding energy E_x. The usual range of variation of the exciton binding energy is from 1 meV to 1 eV. Its value in a number of semiconductor and ionic crystals has been given by Kittel.* Excitons are unstable and prone to a recombination process in which the electron drops into the hole in the valence band with the emission of a photon or phonons.

Frenkel excitons

Excitons are examined in two limiting cases: one in which the exciton is small and tightly bound and the other in which it is weakly bound and its size can be equal to many lattice distances. The former is called a *Frenkel exciton* while the latter is known as *Mott–Wannier exciton*. In a Frenkel exciton the excitation is strongly localized on or near a single atom. Both the electron and the hole are on the same atom with no restriction on the whereabouts of the pair within the crystal. Frenkel

* C. Kittel, *Introduction to Solid State Physics*, 7th ed., p. 314 (John Wiley, 1996).

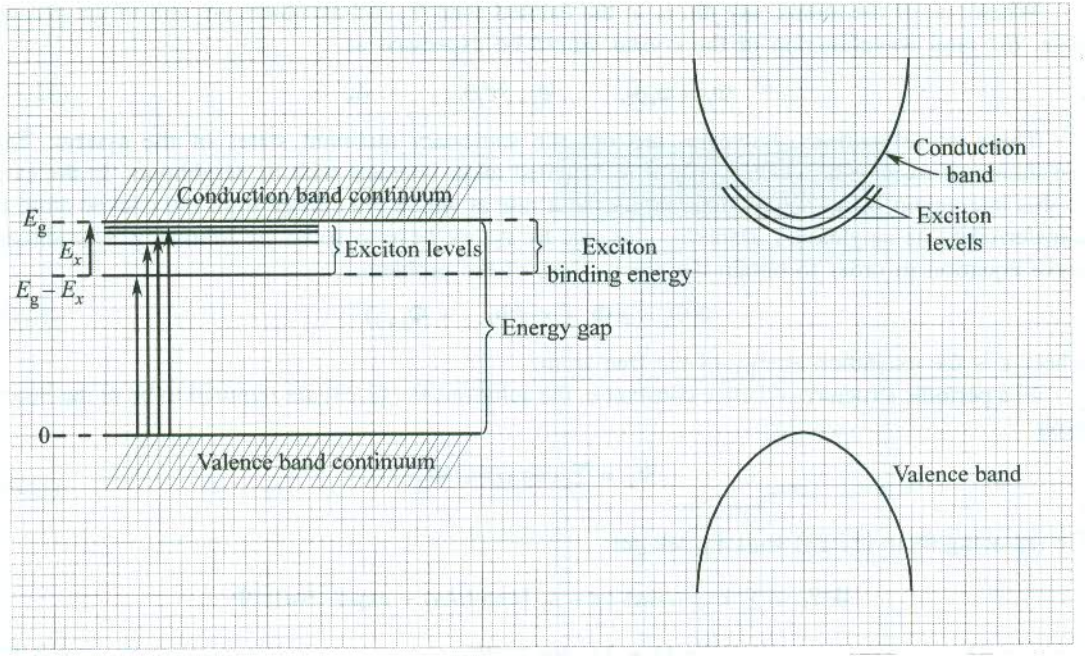


FIG. 12.12 Exciton levels of a solid of simple band structure with both conduction and valence band edges at $\mathbf{k} = 0$. The exciton binding energy is equal to the depth of the exciton ground state below the conduction band minimum.

envisaged an excited electron describing an orbit of atomic dimension around an atom with a vacant valence state. Since the coupling between the neighbouring atoms facilitates the transfer of excitation from atom to atom, the empty valence state, though instantaneously on one atom, behaves as a mobile hole (see Fig. 12.13).

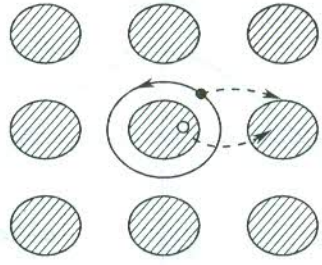


FIG. 12.13 Motion of a Frenkel exciton.

Thus a Frenkel exciton is like an electron bound in a relatively deep impurity state except that it is mobile. The Frenkel picture is applicable to the solidified rare gases and some ionic crystals.

Let us consider the translation of Frenkel excitons in a crystalline ring of N identical atoms. If ϕ_j be the ground state wavefunction of the jth atom in the ring, the ground state of the crystal may be represented by the wavefunction

$$\Phi_0 = \phi_1 \phi_2, \ldots, \phi_{N-1}\ \phi_N \tag{12.21}$$

provided that the interaction between the atoms is ignored.

When one of the atoms, say jth, is in the excited state Ψ_j and all others are still in the ground state, the total wavefunction of the crystal would be expressed as

$$\Phi_j = \phi_1\phi_2, \ldots, \phi_{j-1}\Psi_j\phi_{j+1} \ldots \phi_N \tag{12.22}$$

These wavefunctions, however, are not the stationary quantum states of the system. The translation of exciton is possible only if the excited atom interacts with its neighbours. In the case of this interaction being active, there would be a rate at which the energy from the excited atom is transferred to its neighbours. Denoting this rate by ε_T, we have to solve the following Schrödinger wave equation in order to obtain the eigenvalues of the problem,

$$\mathcal{H}\Phi_j = \varepsilon\Phi_j + \varepsilon_T(\Phi_{j-1} + \Phi_{j+1}) \tag{12.23}$$

where ε is the excitation energy of the free atom.

In a periodic structure of lattice constant a, the solutions to (12.23) are required to be of the Bloch type

$$\Psi_k = \sum_j \exp(ijka)\,\Phi_j \tag{12.24}$$

On replacing Φ_j in (12.23) with Ψ_k, we get

$$\begin{aligned}\mathcal{H}\Psi_k &= \sum_j \exp(ijka)\,\{\varepsilon + \varepsilon_T[\exp(ika) + \exp(-ika)]\}\Phi_j \\ &= [\varepsilon + 2\varepsilon_T \cos ka]\Psi_k\end{aligned} \tag{12.25}$$

Thus we have the eigenvalues

$$\varepsilon_k = \varepsilon + 2\varepsilon_T \cos ka \tag{12.26}$$

The allowed set of k values is the same (4.9) as obtained in Section 4.2.1 by applying periodic boundary conditions. Relation (12.26) shows that the energy of Frenkel excitons is minimum at the zone boundary ($k = \pm \pi/a$) and maximum at the zone centre ($k = 0$) (see Fig. 12.14).

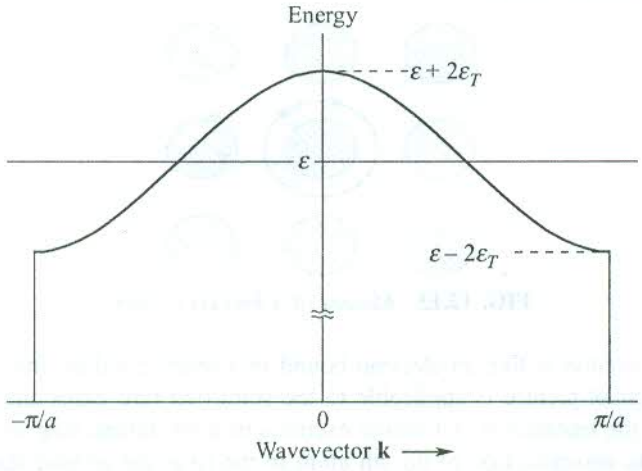


FIG. 12.14 Dispersion curve for a Frenkel exciton, obtained with the positive nearest-neighbour transfer interaction ε_T.

Examples of Frenkel excitons. (1) Solidified inert gases—two lowest transitions in crystalline xenon occur at 8.4 eV and 9.55 eV. These are clearly Frenkel excitons (see Fig. 12.15). (2) Alkali

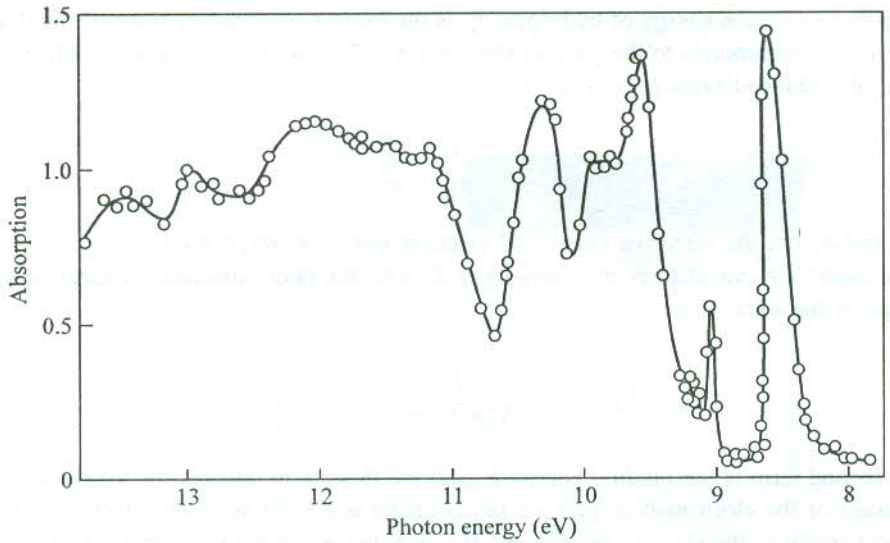


FIG. 12.15 Absorption spectrum of solid xenon between 8.0 and 14.0 eV. [After G. Baldini, *Phys. Rev.*, **128** 1562 (1962)].

halide crystals show Frenkel excitons in the vacuum ultraviolet region. A doublet structure observed in NaBr is a significant example of Frenkel excitons. The splitting is attributed to the effect of spin–orbit coupling. (3) In molecular crystals the covalent bonding within a molecule is much more stronger than the van der Waals coupling that holds molecules together in the crystal. There is no noticeable shift in the energy of excitations of individual molecules when the respective excitations are measured in crystalline samples. These are ideal examples of Frenkel excitons.

Mott–Wannier excitons

These are weakly bound excitons with a separation of many lattice distances between the electron and hole. This model works exceedingly well to account for the excitonic absorption in covalently bonded crystals, such as Group IV semiconductors. It is also applicable to many partially ionic, partially covalent crystals (e.g. Group III–IV semiconductors). We saw in Chapter 9 that the shallow donor or acceptor impurity levels in semiconductors could be described by analogy with hydrogen atoms. The Mott–Wannier picture of exciton is a similar one and its simplest mathematical description is given by the so-called hydrogenic model. The exciton binding energy E_x can be deduced by following the arguments of section 9.4.1 and replacing the effective mass by the reduced mass μ of the electron–hole pair. By analogy with (9.27) and (9.25), we write

$$r_{ex} = n^2 \left(\frac{m \in_s}{\mu} \right) r_0 \tag{12.27}$$

and

$$E_x = -\frac{1}{n^2} \left(\frac{\mu}{m \in_s^2} \right) E_H$$

$$= -\frac{1}{n^2} \frac{\mu e^4}{32 \pi^2 \in_0^2 \in_s^2 \hbar^2} \tag{12.28}$$

where E_H is the ionization energy of hydrogen, r_{ex} is the exciton orbit radius and r_0 is the Bohr radius. The state $n = 1$ corresponds to the ground state and $n = 2, 3, 4$, etc. give a series of excited states of exciton. The reduced mass μ is given by

$$\frac{1}{\mu} = \frac{1}{m_e^*} + \frac{1}{m_h^*}$$

where m_e^* and m_h^* are the effective masses of electron and hole, respectively.

In this model we can identify the energy gap E_g with the photo-ionization energy and write the total energy of the exciton as

$$\varepsilon_{nk} = E_g - \frac{1}{n^2} \frac{\mu e^4}{32\pi^2\hbar^2 \epsilon_0^2 \epsilon_s^2} + \frac{\hbar^2 k^2}{2(m_e^* + m_h^*)} \tag{12.29}$$

where the second term is the binding energy E_x and the third term denotes the kinetic energy of the centre of mass of the electron–hole pair whose inclusion is a must since the electron–hole pair can move almost freely in the crystal (see Fig. 12.16). **k** is the wavevector corresponding to the centre of mass motion of the two particles.

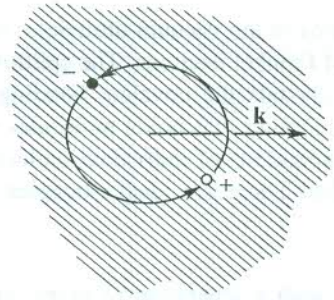


FIG. 12.16 Motion of the Mott–Wannier exciton.

Various relations given above imply that the Coulomb force between the electron and hole is reduced to $1/\epsilon_s$ of the value in vacuum (ϵ_s being the static dielectric constant of the crystal). For semiconductors, ϵ_s is of the order of 10 and therefore the exciton spectrum for $\mathbf{k} = 0$ is a strongly compressed hydrogen spectrum and the binding energies lie below 0.1 eV. On the same account the spatial extent of the Mott–Wannier exciton is much larger than that of the hydrogen atom (~ 10 Bohr radii). This in a way justifies the application of the simple hydrogenic model with the use of a macroscopic dielectric constant.

When excitons are generated by photon absorption and no other particles are involved, the energy conservation demands that the wavevector of the exciton be equal to that of the photon. Since the photon wave vector is small, the third term in (12.29) can be dropped in such a case. This explains why in an optical absorption experiment, exciton states with $\mathbf{k} = 0$ are not seen. As the energy separation of the exciton state from the lower conduction band edge is rather small, the excitonic character of the optical absorption can be investigated only by experiments conducted at low temperatures. The excitonic absorption in GaAs observed at 21 K is shown in Fig. 12.17.

In Fig. 12.15 the two smaller peaks at 9.1 eV and 9.2 eV are identified in the Mott–Wannier model

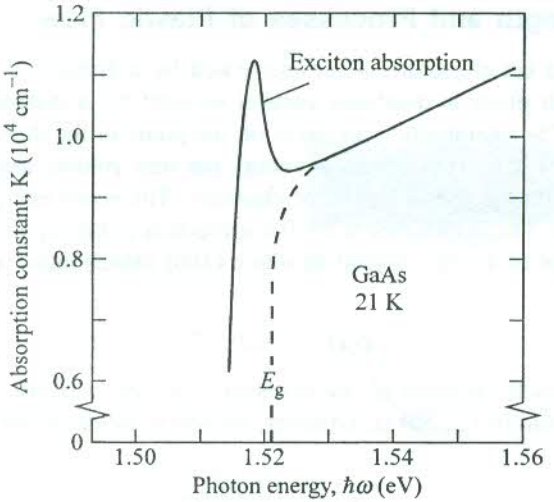


FIG. 12.17 Absorption in a GaAs crystal near the band gap energy E_g at 21 K. The dashed curve shows absorption in the absence of exciton excitation. [After M.D. Sturge, *Phys. Rev.*, **127**, 768 (1962).]

as the $n = 2$ and $n = 3$ lines of a series in which the 8.4 eV line is the lowest. We may recall the 8.4 eV line to have been attributed to a Frenkel exciton. But it must not surprise us to have both types of excitons in one crystal because the condition on orbit size for the Mott–Wannier model in the present case is met for $n \geq 2$ and not for $n = 1$.

Polarons

An electron on being excited to the conduction band of a perfect ionic crystal may often be favoured to move in a spatially localized level with a local deformation in the ionic arrangement being carried along. The deformation is produced by the electron field that effectively polarizes the lattice, as a result of which the electron field is screened and the electrostatic energy lowered. Since perfect crystals are considered to be free of deformations, the quasi-particle composed of an electron and its strain field, known as a *polaron*, is viewed as another point defect which is observed to be fairly mobile. A brief theoretical account of polarons was given in Section 10.10 where they were not treated as defects but as a complication in the theory of electron mobility in ionic crystals.

12.2 LINE IMPERFECTIONS: DISLOCATIONS

Single crystals show far less strength than they would have according to the theory of elasticity. This anomalous behaviour is attributed to the movement of certain line imperfections in crystals under the action of a deforming force. These imperfections are commonly known as *dislocations*. All crystals, except those grown under very special conditions, usually have a high concentration of dislocations. In contrast to a point defect that directly affects one lattice point or a few immediately adjacent lattice points, a dislocation is a line source of imperfection. A general dislocation may follow any curved route through a crystal.

How crucial is the role of dislocations in limiting the strength of a crystal can be appreciated by comparing the measured values with the corresponding theoretical estimates. An approximate estimate of the theoretical elastic limit can be had by calculating the critical stress at which the crystal becomes unstable and the shear occurs.

12.2.1 Shear Strength and Processes of Plastic Flow

Consider a crystal with a set of parallel planes separated by a distance d (Fig. 12.18). Suppose in a shear deformation each plane is displaced parallel to itself by a distance x, with respect to the nearest plane below it. The interatomic separation on the plane in the direction of the displacement in an underformed crystal is a. The additional energy per unit volume associated with the shear of the crystal is obtained with the use of theory of elasticity. This extra energy, say $\phi(x)$, is quadratic in x for small values of x. The displacement $x = 0$ corresponds to the equilibrium state of the crystal. For the shear deformation of a cubic crystal in which (100) planes move along the [010] direction, $\phi(x)$ is related to x by

$$\phi(x) = 2(x/d)^2 \, C_{44} \tag{12.30}$$

where C_{44} is the appropriate stiffness elastic constant. The above relation fails for large x.

The more general form of (12.30) is, however, written in terms of the shear modulus G

$$\phi(x) = \frac{1}{2}\left(\frac{x}{d}\right)^2 G \tag{12.31}$$

The $\phi(x)$ versus x plot is shown in Fig. 12.19(a). As the displacement increases from the zero value, the extra energy per unit volume increases. It becomes maximum at $x = a/2$ where the shear could

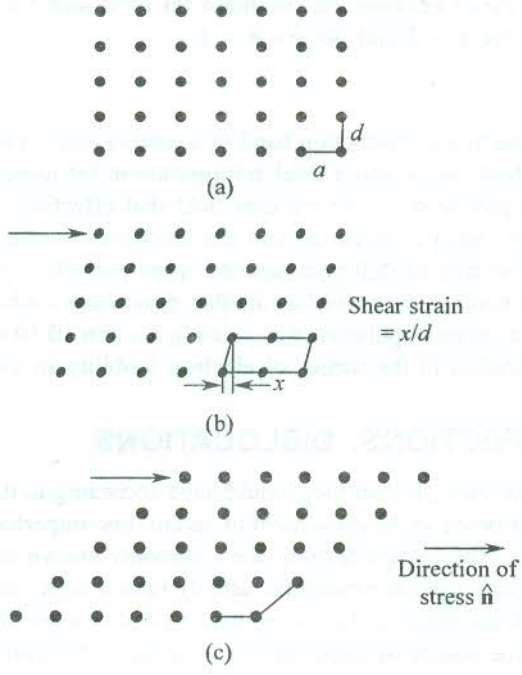


FIG. 12.18 Progressive deformation under increasing shear stress: (a) Perfect crystal. (b) Deformed crystal. (c) Crystal deformed to the limit such that the new interior configuration and the original undeformed crystal appear identical.

go either way and refers to the zero value of stress. At an extreme displacement $x = a$, the interior displacement configuration [see Fig. 12.18(c)] can in no way be distinguished from the underformed crystal if we overlook small surface effects. $\phi(x)$ then obviously drops to zero. It indicates to the fact that $\phi(x)$ is a periodic function of x with period a, i.e. $\phi(x + a) = \phi(x)$. This has no inconsistency with relation (12.31) if $x << a$.

If the crystal has N planes of area S, its volume is $V = NSd$ and, therefore, the additional energy acquired by it in maintaining a relative displacement of x between two successive parallel planes as shown in Fig. 12.18(b) may be written as

$$\Phi(x) = V\phi(x) \tag{12.32}$$

The shear stress σ is defined as the force per unit area on a single plane that maintains the displacement x. We can express it as

$$\sigma = \frac{1}{NS}\frac{d}{dx}(\Phi)$$

$$= d\left(\frac{d\phi}{dx}\right) \qquad \text{[using (12.32)]} \tag{12.33}$$

Relation (12.33) implies that the behaviour of σ must be identical with that of the first derivative of ϕ. The σ versus x curve [Fig. 12.19(b)] in conjunction with the ϕ versus x plot [Fig. 12.19(a)] clearly vindicates this contention. The shear stress becomes maximum (σ_c) at a

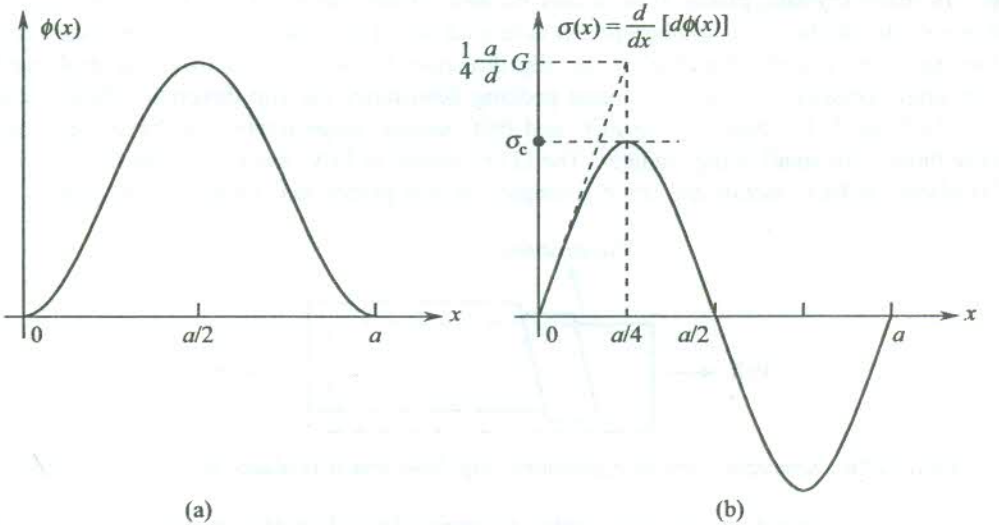


FIG. 12.19 (a) Variation of the additional energy per unit volume, $\phi(x)$, with relative displacement x. (b) Plot of shear stress σ versus the relative displacement x. Note that σ is proportional to the first derivative of $\phi(x)$. The curves are in consistency with this fact.

displacement between 0 and $a/2$. Since our treatment is valid strictly for small values of x and these fall on the linear portion of the curve in Fig. 12.19(b), it is obligatory to extrapolate the linear region out to the point of the maximum of the actual curve ($x = a/4$) and evaluate $d\phi/dx$ at this point for a reasonable, though still approximate, estimate of σ_c. This yields

$$\sigma_c \simeq d\left[\frac{d}{dx}\left(\frac{1}{2}\frac{x^2}{d^2}G\right)\right]_{x=a/4}$$

$$= \frac{1}{4}\left(\frac{a}{d}\right)G \tag{12.34}$$

At shear stress greater than σ_c, one plane slides over another, or in other words the crystal undergoes *slip*. Since G is typically of the order 10^{10} to 10^{11} N m^{-2}, σ_c must be of a similar order of magnitude. On the other hand, the measured values of σ_c even in some best grown pure single crystals are lower by a factor as large as 10^4. In view of this flagrant disparity, attempts were made to improve the theoretical estimate by identifying the true form of intermolecular forces and the available configurations of mechanical stability to the lattice. On these considerations, Mackenzie* was able to show that the ideal shear strength comes down to $G/30$ corresponding to a critical shear angle of about 2 degrees. Since this small improvement does not make any headway towards the resolution of the problem, a suspicion developes over the genuineness of the process of slip on which the estimate of (12.34) is based. Extensive experimental studies have, however, now established that slip occurs by the motion of dislocation lines.

Even though we have not taken to a proper discussion on dislocations so far, a few introductory remarks on *slip* and *twinning* are desirable here because they are the well-known modes of plastic deformation.

Slip. In many crystals, plastic flow occurs because of the sliding of one part of the crystal as a unit relative to another. The sliding process is termed *slip*. The plane on which slip occurs is known as the slip plane and the direction as the slip direction. Figure 12.20 shows a general translation slip in single crystals. The line of closest packing determines the slip direction. These correspond to the [110] and [111] directions in FCC and BCC metals, respectively. The likely slip planes are usually those with small Miller indices. The (111) planes in FCC metals and the (110), (112), and (123) planes in BCC metals are some examples of slip planes that underline this point.

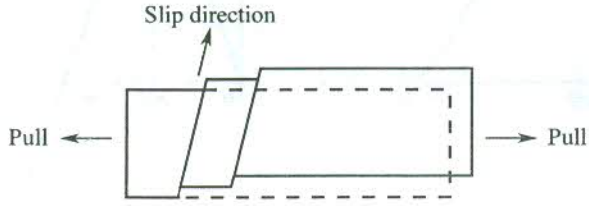


FIG. 12.20 Schematic view of translational slip. Note that it produces an increase in length.

If the displacement or the slip vector measures less than the shortest lattice translation vector, the crystalline order cannot be maintained. This gives rise to *stacking fault* such as the regular sequence of ABC ABC ABC ... of the closest-packed planes in FCC crystals appearing as ABC AB ABC . . . . The slip plane has a large number of dislocations produced by the multiplication of a dislocation segment pinned at each end (called a Frank–Read source) under the influence of an applied shear stress.

* J.K. Mackenzie, *Thesis*, Bristol (1949).

Twinning. It is another mode of plastic deformation observed usually either in HCP or BCC structures. Twinning represents a shearing motion of atomic planes over one another in which the magnitude of translation of each plane is proportional to its distance from a particular plane known as the *twinning plane*. A twinning plane acts as a mirror plane of symmetry at which when one portion of the twinned crystal is reflected the other portion is produced (or constructed).

12.2.2 Dislocation Types

Experiments have confirmed that there are few pure crystals that are not plastic. Germanium and silicon crystals are not plastic at room temperature and fail or yield only by fracture. As mentioned in the preceding section, the weakness and plastic deformation of crystals are attributed to the presence of dislocations and their movement. We describe below two simplest straight-line dislocations—the *edge* and *screw* types. Though there are several of other types, they can generally be interpreted as modifications or combinations of these two.

Edge dislocations

An edge dislocation is formed when either a row of atoms is removed from a crystal lattice or a row of atoms is displaced at unit lattice distance. Figure 12.21(a) illustrates a pure edge dislocation in which one of the planes of atoms terminates at certain level within the crystal and thus resembles a knife blade embedded in a block of cheese with one edge of the knife on level with the top surface of the block. The direction of the line of atoms in the incomplete plane determines the line (or direction) of dislocation. The vector representing the lattice displacement is called the *Burgers vector* **b**. We see that in an edge dislocation the Burgers vector is perpendicular to the dislocation line.

The motion of an edge dislocation is equivalent to slip. In the presence of a shearing stress there is a considerable lattice strain at an edge dislocation, compressive on one side and extensive on the other, facilitating the motion. Figure 12.21(b) explains how the motion of an extra plane across the crystal takes place at the cost of little energy. The motion is equivalent to a slip of one lattice vector. It is analogous to the movement of a wriggle across a rug which is easier than the motion of the rug as a whole. The calculated value of the stress required to move an edge dislocation comes to less than 10^4 N m^{-2}. It shows how easily crystals can be plastically deformed in the presence of dislocations.

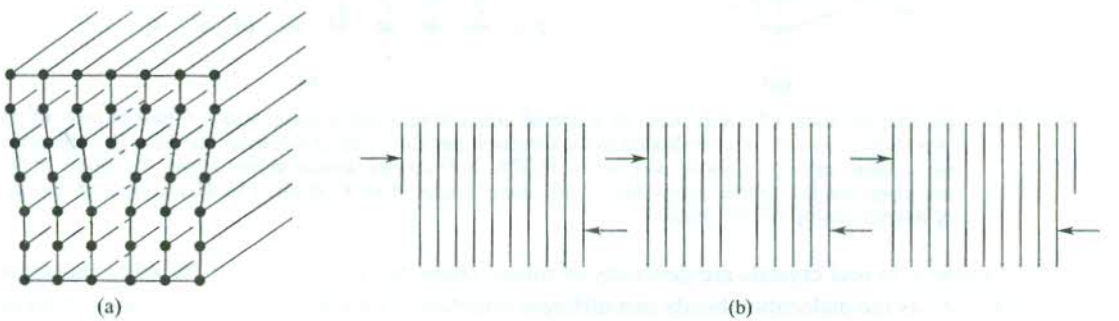


FIG. 12.21 (a) An idealized form of an edge dislocation. (b) Motion of an edge dislocation under applied stress, producing slip.

Screw dislocation

A screw dislocation, as sketched in Fig. 12.22, may be thought to have been accomplished by the following procedure. First, cut a perfect crystal partway through, then force the material on one side of the cut to move up with respect to the material on the other side by one unit of atomic spacing, and finally glue the material on the two sides in this condition. The dislocation marks the boundary between the displaced and undisplaced parts of the crystal. The Burgers vector is again used to describe the displacement. It is parallel to the dislocation line for a pure screw dislocation.

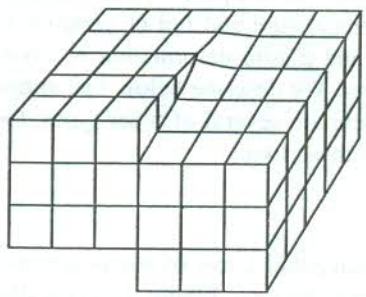
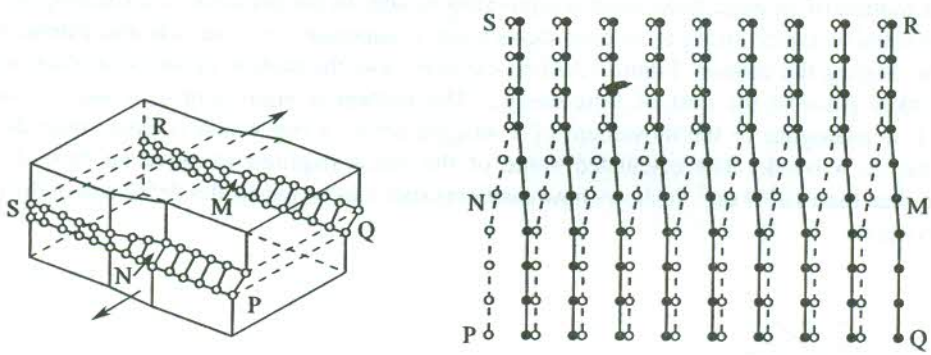


FIG. 12.22 An idealized form of a screw dislocation.

Figure 12.23 shows the relative displacement of atoms in the displaced part of the crystal. A screw dislocation may be pictured as a spiral arrangement of lattice planes in which on going once completely around the dislocation axis we move from one plane to the immediate next.



(a) (b)

FIG. 12.23 (a) Displacement of a slip plane in a crystal containing a screw dislocation. A part PQMN of the slip plane is shown to have slipped in the direction parallel to the dislocation line MN. (b) The atoms (in a cubic structure) above and below PQRS. Full circles denote atoms above the plane PQRS and open circles denote atoms below the plane. [After A.H. Cottrell, *The Mechanical Properties of Matter* (John Wiley, 1964).]

Dislocations in real crystals are generally of mixed character. Proportions of the edge and screw characters vary as the dislocation bends in a different direction. The Burgers vector, however, remains unchanged throughout the length of the dislocation. The dislocation is either a closed or an open loop, terminating only on an outer surface or a grain boundary (to be defined in Section 12.4). The chemical etching is often helpful in locating the termination point of an open dislocation loop since

the etch prefers to attack the strained part revealing thereby an *etch pit*. These studies on tellurium single crystals made by Blakemore et al.† serve as an excellent proof to this effect.

With the application of mechanical stress, dislocations easily sweep through a crystal. A moving dislocation generates point defects until its movement is *pinned* either by impurities or by the path of some other dislocation (work hardening).

Dislocations have energies which do not favour their existence in the state of thermal equilibrium. They are introduced in a non-equilibrium state during the solidification of the crystal, and continue to exist at lower temperatures. A quantity, dislocation density (the number of dislocations crossing per unit area) is used as an indicator of the degree of imperfection of a crystal. The value ranges from 10^{15} or 10^{16} m^{-2} in heavily deformed metals to well below 10^5 m^{-2} in the best crystals of germanium and silicon. Common techniques employed for estimating dislocation densities are: electron microscopy, x-ray transmission and reflection, decoration and etch pits.

12.3 STRESS FIELDS OF DISLOCATIONS

With the knowledge of the magnitudes and directions of stresses introduced into an ideal crystal by the presence of a dislocation, the forces on a dislocation caused by the presence of another one can be calculated. This enables us to give a semi-quantitative description of the behaviour of dislocation networks. Since the calculations for screw dislocations are simpler, we derive the expressions for this case and only quote results for the edge dislocations.

Screw dislocation

Consider a cylindrical crystal that has been sheared in the axial direction. The shearing, as shown in Fig. 12.24, amounts to producing a screw dislocation along the crystal axis with Burgers vector **b**. Since the original length (peripheral) that has undergone displacement is $2\pi r$, the shear strain would be

$$\gamma = \frac{b}{2\pi r} \tag{12.35}$$

Then, the corresponding shear stress is

$$\sigma = G\gamma = \frac{Gb}{2\pi r} \tag{12.36}$$

where G is the shear modulus.

We can express the elastic energy dE_S of the shell of volume dV caused by the presence of the dislocation as

$$dE_S = \frac{1}{2} G\gamma^2 dV$$

† J.S. Blakemore, et al., *J. Appl. Phys.*, **31**, 2226 (1960).

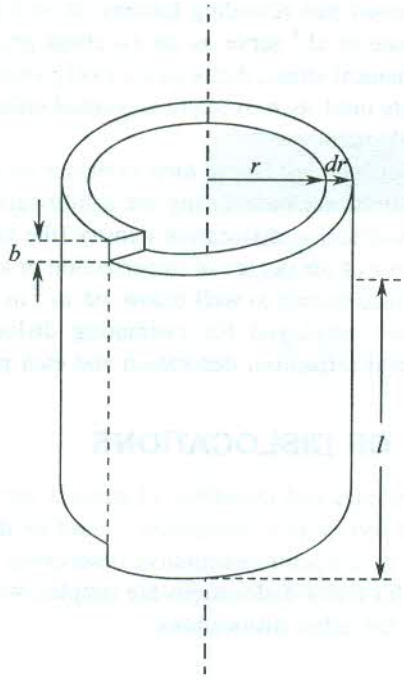


FIG. 12.24 Shell of a deformed cylindrical crystal surrounding a screw dislocation marked by the dotted line.

$$= \frac{1}{2} G \left(\frac{b}{2\pi r} \right)^2 \cdot (2\pi r l dr)$$

$$= \left(\frac{Gb^2}{4\pi} \right) \frac{dr}{r} l \tag{12.37}$$

This gives the elastic energy E_S per unit length of the screw dislocation as

$$E_S = \int_{r_1}^{r_2} \left(\frac{Gb^2}{4\pi} \right) \frac{dr}{r}$$

$$= \left(\frac{Gb^2}{4\pi} \right) \ln \left(\frac{r_2}{r_1} \right) \tag{12.38}$$

where r_1 and r_2 are the lower and upper limits for the radius of the cylinder, strained because of the presence of the dislocation. The choice of r_1 and r_2 is not obvious. Investigations suggest that little error would be introduced if we let $\ln \left(\frac{r_2}{r_1} \right) = 4\pi$ in relation (12.38). It gives

$$E_S \simeq Gb^2 \tag{12.39}$$

Edge dislocation

In this the strain energy E_e per unit length is found as

$$E_e = \frac{Gb^2}{4\pi(1-\nu)} \ln\left(\frac{r_2}{r_1}\right) \tag{12.40}$$

where ν denotes the Poisson ratio which measures approximately ~ 0.3 for metals.

Using $\ln\left(\frac{r_2}{r_1}\right) = 4\pi$, in (12.40), we have

$$E_e \simeq \frac{Gb^2}{1-\nu} \tag{12.41}$$

which is slightly higher than E_S.

The above mathematical exercise reveals two interesting features of dislocations:

(i) The fact that the energy is proportional to b^2, implies that dislocations with minimum possible b will be most stable, i.e. the most stable dislocations occur in directions of close packing.

(ii) The energy increases with the length of the dislocation, meaning thereby that there is a 'line tension' along the dislocation. Since the elastic energy per unit length is equivalent to the force, the tension may be given by

$$T \simeq Gb^2 \tag{12.42}$$

12.4 PLANAR IMPERFECTIONS: GRAIN BOUNDARIES

A junction of two single crystals of different orientations along a common planar surface is called a *grain boundary*. The grain boundary is characteristic of crystalline solids and is not found in amorphous materials. These planar imperfections are very common in polycrystalline materials which consist of a large number of crystals bound to one another along interfaces behaving as grain boundaries. When the difference in orientation of single crystals intervened by a grain boundary is small, we have a low-angle grain boundary. On annealing (cooling slowly), a random array of edge dislocations in a crystal often coalesce to develop low-angle grain boundaries. The movement involved in the process lowers the total energy of the crystal, showing thereby that the crystal prefers to be free of dislocations over the largest possible volume. The two common examples are the *tilt* and *twist* grain boundaries. The former is formed from a linear sequence of edge dislocations (Fig. 12.25) whereas the latter results from a sequence of screw dislocations. Generally, the low-angle grain boundaries are a mixture of these two.

Burgers gave a model for a symmetrical low-angle grain boundary as shown in Fig. 12.25. Denoting the average distance between the dislocations by D, the tilt angle θ is expressed as $\theta = b/D$, where b is the magnitude of the Burgers vector. This model is strongly supported by experiments. Vogel et al.* carried out x-ray and optical investigations of germanium single crystals

* F. L. Vogel, W.G. Pfann, H.E. Corey, and E.E. Thomas, *Phys. Rev.*, **90**, 489 (1953).

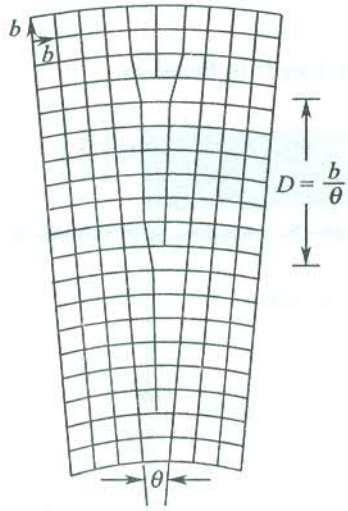


FIG. 12.25 View of a typical low-angle grain boundary.

and concluded that the results verified the model. A germanium crystal grown along the [100] direction from a seeded melt was etched with an acid to reveal the grain boundaries. The grain boundaries were found to consist of regularly spaced conical pits as evidenced in the optical micrograph (Fig. 12.26), recorded under high magnification. Each pit is believed to represent a single dislocation penetrating through the surface. The distance between the pits is obtained by counting and the angle θ determined from the x-ray diffraction measurements. Vogel and co-workers found the calculated and measured values of D and θ to be in a remarkably good agreement.

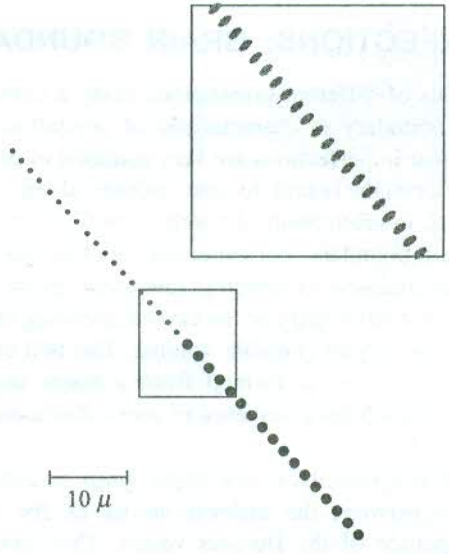


FIG. 12.26 Optical micrograph of lineage boundary in a germanium single crystal when viewed in the face, transverse to the growth direction. [After F.L. Vogel, W.G. Pfann, H.E. Corey, and E.E. Thomas, *Phys Rev.*, **90**, 489 (1953)].

Washburn and Parker† studied the motion of low-angle grain boundaries in zinc single crystals under the action of a suitable stress. They observed that the boundaries move normal to themselves and concluded that the low-angle grain boundaries can be interpreted as arrays of dislocations, supporting the view gathered from other investigations.

12.5 ROLE OF DISLOCATIONS IN CRYSTAL GROWTH

Consider the process of growing a large crystal by exposing a small piece of crystal to a vapour of the same atoms. The deposition of atoms at those lattice sites will be preferred whose surrounding sites are occupied in the original small crystal. This is so because of a relatively strong attractive force of neighbouring atoms in comparison with the weak attractive force to the surface of a perfect crystal. It means that the growth of a seed crystal is faster in the presence of an imperfection in the seed. Theoretical estimates of the growth rate for ideal crystals, especially in the condition of low supersaturation, are lower than the observed growth rates by many orders of magnitude. It is, perhaps, the largest ever recorded disagreement between theory and experiment. The fast observed growth rates are attributed to the presence of dislocations.

Frank* has compiled an excellent review of the phenomena under discussion here. The large disagreement recorded above is attributed, primarily to the difficulty of nucleating a new monolayer on a completed plane of a perfect crystal. But if the crystal contains a screw dislocation, the local helical mould in the growing crystal can wind endlessly about the screw dislocation like a spiral, without necessitating the nucleation of a new plane. A picture of such a growth pattern of a single screw dislocation in SiC crystal is shown in Fig. 12.27.

The development of a spiral step in the growth of a crystal having a screw dislocation is shown in Fig. 12.28. If the growth is taken as independent of the direction of the edge in the plane of the surface, the growth pattern would resemble an Archimedes spiral

$$r = a\theta \tag{12.43}$$

where a is a constant.

The level of supersaturation determines the limiting minimum value of curvature near the dislocation. A too small radius of curvature approaches the equilibrium curvature with the evaporation of an appropriate number of atoms on the curved edge. Each part of a step that is away from the origin acquires new atoms at a constant rate dr/dt. The growth of the crystal proceeds by virtue of rotation of the spiral at an angular speed $d\theta/dt$, which is required to be a constant for a uniform growth.

Whiskers

The kind of crystal growth we discussed above can lead to thin and long crystals wound around a single dislocation. We call them *whiskers*. The presence of a single axial screw dislocation in a whisker is unable to cause yielding, because in bending the crystal the direction of the shear stress on the dislocation is not parallel to its Burgers vectors. Hence, slip cannot occur for this direction of stress, endowing whiskers with elastic properties that are close to theoretical predictions for perfect crystals. Herring and Galt** measured a shear stress of the order of $G/100$ in whiskers of tin

† J. Washburn and E.R. Parker, *J. Metals*, **4**, 1076 (1952).
* F.C. Frank, *Advances in Physics*, **1**, 91 (1952).
** C. Herring and J.K. Galt, *Phys. Rev.*, **85**, 1060 (1952).

FIG. 12.27 Phase contrast micrograph of spiral growth pattern in SiC crystal of 126 R unit cell modification (rhombohedral). Step height, $h = 109 \pm 5$ Å (Mag. × 90). (Courtesy A.R. Verma)

of radius ~1 μm. This value is about 1000 times more than that in bulk tin, giving credence to the theoretical estimates for perfect crystals [~$G/30$, as mentioned in Section 12.2.1].

12.6 STRENGTH OF ALLOYS

There are other important properties of solids that can be manipulated to advantage by exploiting the characteristics of dislocations. Strength of engineering materials is one such example. On account of their wide ranging applications, alloys attract special attention in this regard. Realizing that the motion of dislocations under the action of a deforming force is the main cause of weakness of crystals, the following methods are used to raise the yield strength of alloys:

1. The dislocations are mechanically blocked by introducing tiny particles of a second phase into the crystal lattice. For example, in the hardening of steel, particles of iron carbide are precipitated into iron. Similarly, particles of Al_2Cu are precipitated into aluminium for hardening of aluminium.
2. The dislocations are pinned in solid solutions by solute atoms. Since a foreign atom prefers

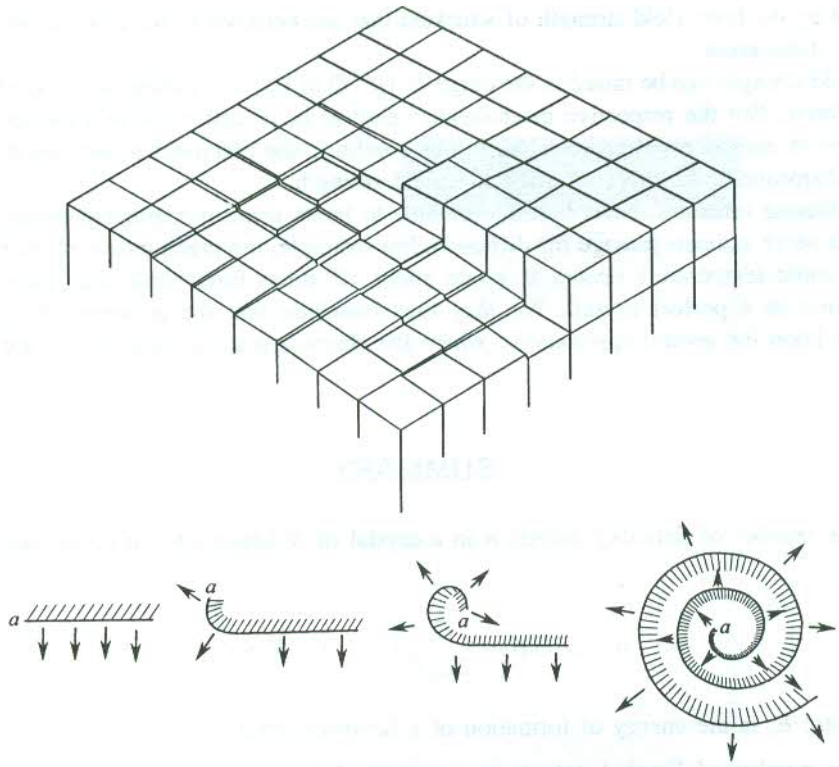


FIG. 12.28 The intersection of a screw dislocation with a free surface to produce a spiral step. [After F.C. Frank, *Discussions Faraday Soc.*, 5, 48 (1949).]

to dissolve itself near a dislocation than elsewhere, the pinning is easily accomplished by this method.

3. Slip can be impeded by short range or local order, found in most of the concentrated solid solutions. When the number of like neighbour bonds in an alloy is different from that expected statistically, the alloy can have only short range order. A disordered phase is first produced by quenching (cooling quickly) from a high temperature and then some ordering is introduced by heat treatment at a lower predetermined temperature. This produces a sizable increase in strength, something impossible in the presence of long range order. In a long range order, atoms of one kind occupy preferred lattice sites and a super lattice is produced where correlation occurs over many lattice sites that is conducive to slip. The order-hardening in 18 carat gold is accomplished by the precipitation of an ordered AuCu phase.
4. Crystalline materials gain in strength even when they are highly enriched with dislocations. Large dislocation densities make it more difficult for a dislocation to move across a slip plane which at these densities is threaded by many dislocations. The method, known as *work-hardening* or *strain-hardening* is, however, useful at temperatures low enough so as not to allow annealing.

In addition, the yield strength may also be increased by preparing dislocation-free crystals. Such a perfection in crystal growth is attainable to some extent only in crystals of small size. This is

exemplified by the high yield strength of whiskers that are believed to have grown around a single axial-screw dislocation.

The yield strength can be raised to the range of 10^{-3} G to 10^{-2} G by using any of the four methods described above. But the respective mechanisms cease to be effective above a certain temperature where diffusion rates become appreciable. In this condition the precipitated particles that hinder the motion of dislocations dissolve, lowering the yield strength.

Fast diffusion rates are, nevertheless, welcome to boost some precipitation reactions in solids. Dislocations serve as open passage for diffusion. For example, the precipitation of tin from lead–tin solution at room temperature occurs at a rate about 10^8 times faster than what may be expected from diffusion in a perfect crystal. We may thus conclude that the presence of dislocations in crystals is a boon for several applications, where the characteristics of dislocations are exploited to advantage.

SUMMARY

1. The number of Schottky defects n in a crystal of N lattice sites at temperature T is given by

$$n = N \exp\left(-\frac{E_v}{2k_B T}\right) \qquad \text{for } N >> n$$

where E_v is the energy of formation of a Schottky defect.

2. The number of Frenkel defects in a crystal of N lattice sites and N_i interstitial sites at temperature T is given by

$$n = (NN_i)^{1/2} \exp\left(-\frac{E_i}{2k_B T}\right) \qquad \text{for } n << N,\ N_i$$

where E_i is the energy required for removing an atom from a lattice site to an interstitial position.

3. Diffusion constant at a temperature T is given by

$$D = (\nu_0\ a^2) \exp\left(-\frac{E_a}{k_B T}\right)$$

where ν_0 is the characteristic atomic vibration frequency, a is the lattice constant and $\mathbf{E}_a$ the activation energy.

4. The energy of a Frenkel exciton is

$$\varepsilon_k = \varepsilon + 2\varepsilon_T \cos ka$$

where ε is the excitation energy of the free atom and ε_T denotes the rate at which the energy from an excited atom is transferred to its neighbours.

5. The binding energy of a Mott–Wannier exciton is

$$E_x = -\frac{1}{n^2}\frac{\mu e^4}{32\pi^2 \epsilon_0^2 \epsilon_s^2 \hbar^2}$$

with $\frac{1}{\mu} = \frac{1}{m_e^*} + \frac{1}{m_h^*}$

6. The theoretical critical shear stress for perfect crystals is given by

$$\sigma_c \cong \frac{1}{4}\left(\frac{a}{d}\right) G$$

where

a is the atomic separation in a plane
d is the interplanar spacing
G is the shear modulus

The more reasonable estimate of the ideal sheer stress is $G/30$, with $G = 10^{10}$ to 10^{11} N m^{-2}.

7. The strain energy per unit length of a screw dislocation, $E_S \simeq Gb^2$ where b is the magnitude of the Burgers vector.

The strain energy per unit length of an edge dislocation is

$$E_e \simeq \frac{Gb^2}{1-\nu}, \text{ where } \nu \text{ is the Poisson ratio.}$$

PROBLEMS

12.1 Calculate the atomic percentage of interstitials and vacancies at the melting point in copper. Take the formation energies, respectively, as 4.5 and 1.5 eV.

12.2 A perfect crystal of KCl is doped with 10^{24} $CaCl_2$ molecules per m^3. Estimate the fractional change of density assuming that (a) no vacancies are introduced owing to doping, (b) one positive ion vacancy is created for each $CaCl_2$ molecule.

12.3 Given that the energy required to remove a sodium atom from the inside of a sodium crystal to the boundary is 1 eV, calculate the number of Schottky vacancies at room temperature. Assuming that a neighbouring sodium atom has to cross over a potential barrier of 0.5 eV and the atomic vibration frequency be 10^{12}, calculate the diffusion coefficient for radioactive sodium in normal sodium at room temperature.

12.4 If the energy required to create a vacancy in a solid be 2 eV, show that the relative density of vacancies to atoms will always be less than 10^{-8} per cent unless the melting point is higher than 1000 K.

12.5 Show that in a crystal having N lattice points and N' possible interstitial positions, the number n of Frenkel defects in equilibrium condition is given by

$$E_i = k_B T \ln \left[\frac{(N - n)(N_i - n)}{n^2} \right]$$

Show further that for $n << N, N_i$

$$n \cong (NN_i)^{1/2} \exp\left(-\frac{E_i}{2k_B T} \right)$$

where E_i denotes the energy required for removing an atom from a lattice site to an interstitial position.

12.6 Apply simple Bohr model to predict the energy of the F-centre absorption in KCl, taking $1s \rightarrow 2p$ as the absorption transition. The refractive index of KCl is 1.49.

12.7 What is the smallest Burgers vector parallel to a [111] direction that a dislocation in an FCC crystal may have?

12.8 A crystalline cube of side L contains an edge dislocation with Burgers vector **b**. It is subjected to a shear stress σ on the upper and lower faces in the direction of slip. On energy balance consideration show that the magnitude of force **F** on the dislocation per unit length is given by $F = b\sigma$.

12.9 A single crystal of copper has a low angle tilt boundary on (010) plane with the tilt axis parallel to the [001] direction. If the spacing of dislocations in the boundary is 1.5×10^{-6} m, determine the tilt angle.

12.10 The elastic displacement round an isolated dislocation of Burgers vector **b** is $\mathbf{b}\theta/2\pi$, where θ is the angular coordinate around the dislocation line. Show that two parallel dislocations with Burgers vectors $\mathbf{b}_1$ and $\mathbf{b}_2$ and a distance r apart in a crystal of shear modulus G repel each other with a force whose magnitude per unit length is approximately given by $Gb_1b_2/2\pi r$.

SUGGESTED FURTHER READING

Farge, Y. and M. Fontana, *Electronic and Vibrational Properties of Point Defects in Ionic Crystals* (North-Holland, 1979).

Flynn, C.P., *Point Defects and Diffusion* (Oxford, 1972).

Friedel, J., *Dislocations* (Pergamon Press, 1964).

Hull, D. and D.J. Bacon, *Introduction to Dislocations*, 3rd ed. (Pergamon Press, 1984).

Markham, J.J., *F-Centres in Alkali Halides* (Academic Press, 1966).

Nabarro, F.R.N., *Theory of Crystal Dislocations* (Dover, 1987).

Schulman, J.H. and W.D. Compton, *Color Centres in Solids* (Pergamon Press, 1962).

Whelan, M.J., *Worked Examples in Dislocations* (Institute of Metals, 1990).

Chapter 13

Diamagnetism and Paramagnetism

Unlike the dielectric response, the magnetic response of most of the solids is dominated by permanent dipoles. The existence of magnetic dipoles, whether permanent or induced, can be explained only on the basis of quantum theoretical considerations. If we regard matter as made up of charged particles, there can be no magnetic moment in the state of thermal equilibrium of a strictly classical system even in the presence of an applied magnetic field. Thus magnetism is essentially a quantum effect. The two fundamental forms of magnetism, *diamagnetism* and *paramagnetism*, have their origin in induced and permanent magnetic moments, respectively. The impossibility of magnetism in the domain of classical physics may be argued in the following way. As the magnetic field is switched on, the electronic circulating currents having an associated magnetic moment are induced in the system. But these currents are destroyed by collisions in tending to bring the system finally to the state of thermal equilibrium.

None of the derivations of relations in classical physics including those for magnetic susceptibility is self-consistent. The permanent atomic magnetism (paramagnetism) cannot be accounted for without restricting the circulating electrons to the discrete stationary orbits as required in the Bohr's quantum theoretical model of the hydrogen atom. In the classical picture, there can be no magnetic moment associated with the current of circulating electrons because electrons in accelerated motion would radiate and finally fall on the nucleus, causing the atomic structure to collapse. Diamagnetism, where the applied magnetic field is pushed out of the system, can be appreciated similarly by realizing that the discrete quantum states occupied by electrons are stable to a certain extent only against external perturbations, like the magnetic field in the present case.

We define below the fundamental physical quantities that concern the magnetic properties of materials. In vacuum, the intensity of the applied magnetic field $\mathbf{H}$ and the magnetic induction $\mathbf{B}$ are related by the equation

$$\mathbf{B} = \mu_0 \mathbf{H} \tag{13.1}$$

where μ_0 is the permeability of free space ($\mu_0 = 4\pi \times 10^{-7}$ V s/A m).

The magnetic state of a system is specified by its magnetization $\mathbf{M}$, defined as the magnetic moment per unit volume. $\mathbf{M}$ is related to $\mathbf{B}$ and $\mathbf{H}$ by

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) \tag{13.2}$$

For convenience in discussions it is a practice to introduce an external induction such that

$$\mathbf{B}_0 = \mu_0 \mathbf{H} \tag{13.3}$$

Mostly, there is a linear relationship between $\mathbf{B}_0$ and $\mathbf{M}$ given by

$$\mu_0 \mathbf{M} = \chi \mathbf{B}_0 \tag{13.4}$$

Since **r** is perpendicular to linear velocity (say, **v**), (13.6) may be rewritten as

$$\boldsymbol{\mu}_l = -\left(\frac{e}{2m}\right)(\mathbf{r} \times \mathbf{p})$$

$$= -\left(\frac{e}{2m}\right)(\text{orbital angular momentum}) \tag{13.7}$$

where e is the electronic charge, m the electron mass, and **p** its linear momentum.

Using simple methods of Newtonian mechanics, we can study the motion of an electron within an atom or ion in the presence of a steady magentic field. The characteristic feature of the motion in this condition is that the plane of the electron orbit precesses about the direction of the magnetic field [see Fig. 13.1(a)]. This amounts to precession of the orbital angular momentum vector $\boldsymbol{l}$ associated with the electron motion around the central nucleus [see Fig. 13.1(b)]. The precession is superposed on the electron's orbital motion in the central field of the nucleus which to a first order in the magentic field remains almost the same as in the absence of the magnetic field. The frequency of precession is, however, much smaller than the frequency of the orbital motion of the electron. The precession is called the 'Larmor precession'.

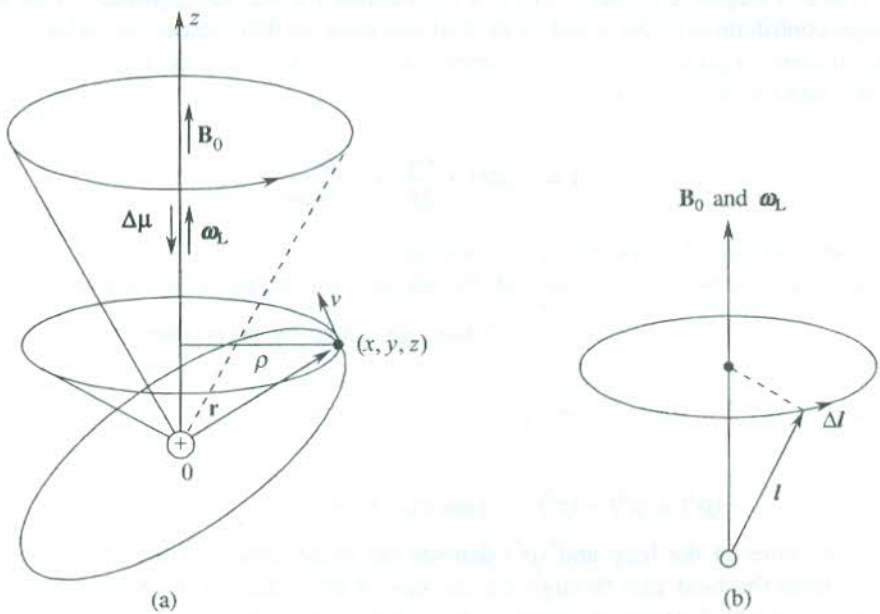


FIG. 13.1 Effect of the magnetic field on orbital motion of an electron. The field $\mathbf{B}_0$ is perpendicular to the orbit plane: (a) The precession of the electron orbit. $\boldsymbol{\omega}_L$ denotes the angular frequency of precession and $\Delta\boldsymbol{\mu}$ denotes the induced magnetic moment. (b) The precession of the angular momentum $\boldsymbol{l}$ in the field.

The magnetic moment associated with the orbital angular momentum $\boldsymbol{l}$ according to (13.7) is

$$\boldsymbol{\mu}_l = -\left(\frac{e}{2m}\right)\boldsymbol{l} \tag{13.8}$$

This dipole, when acted upon by a magnetic field $\mathbf{B}_0$ (actually the magnetic induction), experiences a torque given by

$$\frac{d\boldsymbol{l}}{dt} = \boldsymbol{\mu}_l \times \mathbf{B}_0 = \left(\frac{e}{2m}\right) \mathbf{B}_0 \times \boldsymbol{l} \tag{13.9}$$

The above equation that serves as the equation of motion of $\boldsymbol{l}$ is a well-known equation in mechanics, describing precession of the characteristic angular frequency

$$\omega_L = \frac{eB_0}{2m} \tag{13.10}$$

where ω_L is commonly known as the 'Larmor frequency'.

Alternatively, the vector relationships as depicted in Fig. 13.1(b) give

$$\Delta \boldsymbol{l} = \boldsymbol{\omega}_L \times \boldsymbol{l} \cdot dt \tag{13.11}$$

On comparing (13.11) with (13.9), we reproduce the value of ω_L as given by (13.10).

The precession of the electron orbit produces its own loop current and, therefore, magnetic moment. It is this magnetic moment which is responsible for the diamagnetism of an atom or ion. The average contribution of the usual motion of electrons in their respective orbits is zero in this model, as already explained. The loop current in the atomic or ionic system produced by the precessional motion is

$$I = -(Ze) \times \frac{\omega_L}{2\pi} = -\frac{Ze^2 B_0}{4\pi m} \tag{13.12}$$

where Z is the number of electrons in the atom or ion.

The magnetic moment associated with the above loop current is written as

$$\mu = I \times \text{area of the loop generated by precession}$$

$$= -\frac{Ze^2 B_0}{4m} \langle \rho^2 \rangle \tag{13.13a}$$

with

$$\langle \rho^2 \rangle = \langle x^2 \rangle + \langle y^2 \rangle \qquad \text{[see Fig. 13.1(a)]} \tag{13.13b}$$

Here ρ is the radius of the loop and $\langle \rho^2 \rangle$ denotes the mean square of the perpendicular distance of electrons from the field axis through the nucleus. If the radius of the orbit around the nucleus be r, the mean square distance of electrons from the nucleus is given by

$$\langle r^2 \rangle = \langle x^2 \rangle + \langle y^2 \rangle + \langle z^2 \rangle \tag{13.14}$$

When the charge distribution is spherically symmetric,

$$\langle x^2 \rangle = \langle y^2 \rangle = \langle z^2 \rangle \tag{13.15}$$

Then, using (13.14) and (13.15) in (13.13b), we get

$$\langle \rho^2 \rangle = \frac{2}{3} \langle r^2 \rangle \tag{13.16}$$

Substituting this value of $\langle \rho^2 \rangle$ in (13.13a), we get the following expression for the magnetization

$$\mathbf{M} = -\frac{N Z e^2 \mathbf{B}_0}{6\, m} \langle r^2 \rangle \tag{13.17}$$

where N is the number of atoms per unit volume of the system and Z is the atomic number.

Then, the diamagnetic susceptibility is expressed as

$$\chi = \frac{\mu_0 \mathbf{M}}{\mathbf{B}_0} = -\frac{\mu_0 N Z e^2}{6\ m} \langle r^2 \rangle \tag{13.18}$$

The relation (13.18) is known as the Langevin's relation for the diamagnetic susceptibility. The diamagnetic susceptibility of isolated atoms or ions can be calculated with the knowledge of $\langle r^2 \rangle$ whose appropriate determination is feasible only in a quantum mechanical approach. van Vleck* has discussed some other models of intricate nature, including the diamagnetism of molecules with electron-pair bonds.

For typical densities and orbital sizes of electrons in solids, relation (13.18) gives a susceptibility of approximately -10^{-5}. This is fairly comparable with the measured values which for a few materials are given below:

Materials	Bi	Hg	Ga	Cu	Si
χ	-1.6×10^{-4}	-2.9×10^{-5}	-2.3×10^{-5}	-9.5×10^{-6}	-4.2×10^{-6}

Diamagnetism is an essential property of all substances. On account of being weak, it is masked in paramagnetic materials under the background of their positive susceptibility.

13.2 LANGEVIN'S THEORY OF PARAMAGNETISM

The classical treatment of paramagnetism is also due to Langevin. Consider a solid containing N magnetic atoms per unit volume, each bearing a magnetic moment μ. An assembly of randomly oriented magnetic moments (dipoles) defines the ground state of this solid in the absence of a magnetic field. The energy of a dipole in the magnetic field $\mathbf{B}_0$ is

$$U = -\boldsymbol{\mu} \cdot \mathbf{B}_0 \tag{13.19}$$

Any increase in the magnetic field should produce preferential orientation of the dipoles with the field. This orientation at the same time is resisted by the thermal disorder in the solid. The orientation produces the magnetization $\mathbf{M}$ whose magnitude in the state of thermal equilibrium may be expressed as

$$\mathbf{M} = N\mu\, \overline{\cos\theta} \tag{13.20}$$

where $\overline{\cos\theta}$ is the average over a distribution in thermal equilibrium, θ being the angle between $\boldsymbol{\mu}$ and $\mathbf{B}_0$.

Using the Boltzmann statistical law of distribution, the relative probability of finding a dipole in an element of solid angle $d\Omega$ is expressed as $\exp\left(-\frac{U}{k_B T}\right)$ and

* J.H. van Vleck, *The Theory of Electric and Magnetic Susceptibilities* (Oxford, 1952).

$$\overline{\cos\theta} = \frac{\int \exp\left(-\frac{U}{k_B T}\right)\cos\theta\, d\Omega}{\int \exp\left(-\frac{U}{k_B T}\right) d\Omega} \tag{13.21}$$

Since the limits of integration must include all solid angles,

$$\overline{\cos\theta} = \frac{\int_0^{\pi} 2\pi \sin\theta\cos\theta \exp\left(\frac{\mu B_0 \cos\theta}{k_B T}\right) d\theta}{\int_0^{\pi} 2\pi \sin\theta \exp\left(\frac{\mu B_0 \cos\theta}{k_B T}\right) d\theta}$$

Letting $\cos\theta = x$, and $\frac{\mu B_0}{k_B T} = a$

$$\overline{\cos\theta} = \frac{\int_{-1}^{1} x \exp(ax)\, dx}{\int_{-1}^{1} \exp(ax)\, dx} = \frac{d}{da} \ln \int_{-1}^{1} \exp(ax)\, dx$$

$$= \coth a - \left(\frac{1}{a}\right)$$

$$\equiv \mathrm{L}(a) \tag{13.22}$$

where $\mathrm{L}(a)$ is called the 'Langevin function'.

Then, from (13.20)

$$M = N\mu\, \mathrm{L}(a) \tag{13.23}$$

A schematic plot of the Langevin function $\mathrm{L}(a)$ is given in Fig. 13.2, showing that it saturates when a or $\frac{\mu B_0}{k_B T} >> 1$. It suggests that the saturation can be easily achieved at low temperatures. The initial slope is linear as indicated by the dashed line. In this region, $\frac{\mu B_0}{k_B T} << 1$ or $\mu B_0 << k_B T$ that conforms to the use of low magnetic fields at which the experiments are done. For $B = 1$ T and μ of the order of a Bohr magneton (10^{-23} J/T), $\mu B_0 = 10^{-23}$ J which is about one hundredth of $k_B T$ at room temperature ($\sim 10^{-21}$ J). Thus the initial linear part of the curve, ideally refers to the experimental conditions at room temperature. The value of $\mathrm{L}(a)$ in this limit ($a << 1$) comes to $a/3$.

Therefore, using $\mathrm{L}(a) \simeq \frac{a}{3} = \frac{\mu B_0}{3k_B T}$ in (13.23), we get

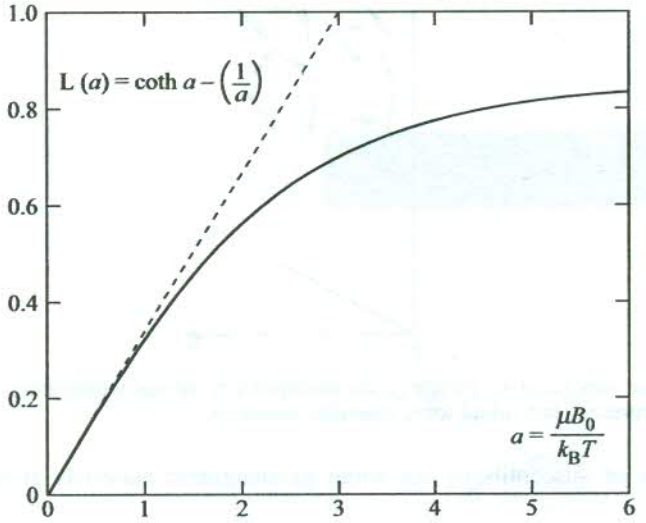


FIG. 13.2 Variation of Langevin function L(a) as a function of $a = \dfrac{\mu B_0}{k_B T}$. The dashed line indicates the initial slope.

$$M = \frac{N\mu^2 B_0}{3k_B T} \tag{13.24}$$

and

$$\chi = \frac{\mu_0 M}{B_0} = \frac{\mu_0 \mu^2 N}{3k_B T} \tag{13.25}$$

or

$$\chi = \frac{C}{T} \tag{13.26}$$

with

$$C = \frac{\mu_0 \mu^2 N}{3\,k_B} \tag{13.27}$$

Relation (13.26) is called the 'Curie law' and C is known as the 'Curie constant'. According to the Curie law, the plot of $1/\chi$ versus T is a straight line as shown in Fig. 13.3. While the law is obeyed in a large number of paramagnetic solids, deviations at low temperatures are observed in some others. The deviations can be justified on the ground that the dipolar orientation is complicated by the rules of quantum mechanics in respect of the space quantization of angular momentum.

With $N = 5 \times 10^{28}$ m^{-3} and $\mu = 1$ Bohr magneton, we have

$$\chi = \frac{0.13}{T} \tag{13.28}$$

The numbers obtained from (13.28) are of the order of the typical measured values. The

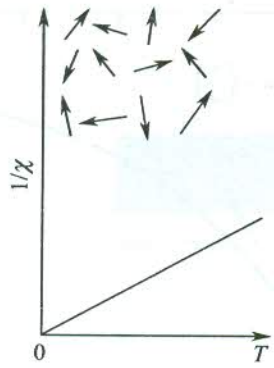


FIG. 13.3 Plot of the reciprocal of paramagnetic susceptibility versus temperature. The randomly oriented arrows represent individual ionic magnetic moments.

experimental values of susceptibility for some paramagnetic materials at room temperature are given below:

Material	Cr	Al	Ca	Na	O_2 (at NTP)
χ	$+2.7 \times 10^{-4}$	$+2.2 \times 10^{-5}$	$+1.9 \times 10^{-5}$	$+9.1 \times 10^{-6}$	$+1.9 \times 10^{-6}$

13.3 THEORY OF ATOMIC MAGNETIC MOMENT

In this section we apply the principles of quantum theory to calculate the atomic magnetic moment. We begin by subjecting the electrons of an atom or ion to the Bohr's quantization condition, according to which their orbital angular momentum is constrained to be a multiple of $\hbar$. This requires (13.8) to be written as

$$\boldsymbol{\mu} = -\left(\frac{e\hbar}{2m}\right)\boldsymbol{l}; \qquad \text{with } (\mathbf{r} \times \mathbf{p}) = \hbar\boldsymbol{l}$$

$$= -\mu_B \boldsymbol{l} \tag{13.29}$$

where $\boldsymbol{l}$ denotes the orbital angular momentum of an electron, and

$$\mu_B = \frac{e\hbar}{2m} = 5.7884 \times 10^{-5} \text{ eV/T} \tag{13.30}$$

$$= 9.2742 \times 10^{-24} \text{ J/T}$$

Here μ_B is called the 'Bohr magneton'. According to (13.29) the magnetic moment associated with the orbital motion is a multiple of the Bohr magneton.

The magnetic moment caused by electron spin can be calculated similarly. It may not be out of place to remind readers that the electron spin is purely a quantum mechanical concept. It is simply impossible to provide an interpretation to it that may conform to any form of motion we come across in our day-to-day life. The spin quantum number $s = \pm 1/2$ characterizes the angular momentum associated with the electron spin. It is a standard practice to express spin magnetic moment as

$$\boldsymbol{\mu}_s = -g_0 \mu_B \mathbf{s} \tag{13.31}$$

where the quantity g_0 is called the electronic g-factor. The above relation is often used to define the Bohr magneton. Because $g_0 = 2.0023$ for a free electron, one Bohr magneton is taken as almost an exact measure of the magnetic moment of a spinning electron.

For the calculation of the magnetic moment of an atom or ion which generally has more than one electron, we require to determine the total angular momentum. This is accomplished by combining vectorially the individual orbital and spin angular momenta by following either the L–S coupling or the J–J coupling scheme, whichever may be valid for the element of the atom under consideration. The two schemes have been tested for all the elements of the periodic table and the response in this respect may be found in standard books on atomic physics. Suppose, we have a case for the L–S coupling (also known as the Russel–Saunders coupling) to be used. In this scheme, the orbital quantum numbers of individual electrons are added vectorially to get the net orbital angular momentum as $\hbar\mathbf{L} \left(= \hbar \sum_i \boldsymbol{l}_i\right)$. Similarly, the vectorial combination of individual spin quantum numbers gives the gross spin angular momentum expressible as $\hbar\mathbf{S} \left(= \hbar \sum_i \mathbf{s}_i\right)$. Then, the total angular momentum is written as

$$\hbar\mathbf{J} = \hbar\mathbf{L} + \hbar\mathbf{S} \tag{13.32}$$

where J denotes the total angular momentum quantum number. The associated atomic or ionic magnetic moment $\boldsymbol{\mu}$ precesses around the direction of $\mathbf{J}$. The $\boldsymbol{\mu}$ is customarily related to $\mathbf{J}$ as*

$$\boldsymbol{\mu} = -\mu_B(\mathbf{L} + g_0\mathbf{S}) \tag{13.33a}$$

$$= -g\mu_B\,\mathbf{J}$$

with

$$\mathbf{L} + g_0\mathbf{S} = g\mathbf{J} \tag{13.33b}$$

where g is called the Landé splitting factor.

For the Russel–Saunders type of coupling† (or L–S coupling), the Landé splitting factor is given by

$$g = 1 + \frac{J(J+1) + S(S+1) - L(L+1)}{2J(J+1)} \tag{13.34}$$

If we take the value of g_0 as 2.0000 instead of 2.0023, to a good approximation the magnetic moment may be expressed as

$$\boldsymbol{\mu} = -\mu_B(\mathbf{L} + 2\mathbf{S}) \tag{13.35}$$

The above discussion suggests that the presence of a permanent magnetic moment is crucial to the magnetic behaviour of a system. This concerns atoms or ions with partially filled shells. The magnetic moment of a system can be estimated by calculating L and S (hence J) of individual atoms or ions of the system. It is accomplished with the help of the Pauli exclusion principle and the

---

* Relation (13.33b) is valid only within the $(2J + 1)$ dimensional set of states that form the degenerate atomic ground state in the absence of a magnetic field. For details, see N.W. Ashcroft and N.D. Mermin, *Solid State Physics*, Ch. 31, p. 654 (Saunders College, 1988).

† We must appreciate that J is always a good quantum number of an atom. But, if L–S coupling (the spin–orbit coupling) is insignificant, L and S can also be good quantum numbers. In this limit, the Hamiltonian of the atom can be taken to commute with $\mathbf{L}$ and $\mathbf{S}$ as well as with $\mathbf{J}$. Hence the states of the atom can be described by quantum numbers L, L_z, S, S_z, J and J_z, indicating that they are eigenstates of the operators $\mathbf{L}^2$, $\mathbf{L}_z$, $\mathbf{S}^2$, $\mathbf{S}_z$, $\mathbf{J}^2$ and $\mathbf{J}_z$ with eigenvalues $L(L + 1)$, L_z, $S(S + 1)$, S_z, $J(J + 1)$, and J_z, respectively.

Hund's rules that describe the manner in which the electrons are assigned to various quantum states within a partially-filled shell of an atom or ion in its ground state.

13.3.1 Hund's Rules

Quantum numbers for the spin, orbital, and total angular momentum in the ground state of an atom or ion with a partially-filled shell are derived by applying the following rules:

The first rule. The electron spins are added so as to get the maximum possible S, as permitted by the Pauli principle.

The second rule. The orbital momenta are combined to get the maximum value for L that is consistent with S determined by the first rule.

The third rule. For a partially-filled shell,

$$J = L - S \text{ for a shell less than half-filled}$$

$$= L + S \text{ for a shell more than half-filled}$$

A few examples are discussed below to illustrate the application of the Hund's rules.

Example 1. Consider the Cr^{3+} ion which is a transition metal ion of the iron group. Its $3d$ shell is partially filled and contains 3 electrons, i.e. short of 7 electrons to be completely occupied. In this case $l = 2$; so the allowed m_l values [$(2l + 1)$ in number] are: 2, 1, 0, –1, –2, characterizing five different degenerate orbitals. For the maximum values of S and L, as required under the first and second rules, the filling of electrons in the orbitals is as follows:

m_l	2	1	0	–1	–2
	↑	↑	↑		

where (↑) denotes an electron with $m_s = +1/2$

The above filling arrangement gives

$L = \Sigma\, m_l = 3$ (the sum is taken over the occupied orbitals)

$S = \Sigma\, m_s = 3/2$

Since the shell is less than half-filled, in accordance with the third rule, we have

$$J = L - S = 3/2$$

The energy states are conventionally designated as

$$M_{X_A}$$

where $M = (2S + 1)$, the multiplicity

$A = J$, the total angular momentum quantum number

	$X = S$	P	D	F	G	$H \ldots$
for	$L = 0$	1	2	3	4	5 . . .

Because our calculations are for the ground state, the ground state of Cr^{3+} ion is $^4F_{3/2}$.

Example 2. Let us take the next example of a rare earth ion, Ho^{3+} in which $4f^{10}$ represents the electronic configuration of the incomplete shell (4f). For this shell, $l = 3$, therefore, the filling of electrons in various allowed orbitals can be shown as

m_l	3	2	1	0	−1	−2	−3
	↓↑	↓↑	↓↑	↑	↑	↑	↑

This gives $L = 6$ and $S = 2$. Since the shell is more than half-filled, $J = 6 + 2 = 8$. Thus the ground state is ${}^5I_8$.

Example 3. We take an interesting case where an ion is short of one electron to make it half-filled. Consider Mn^{3+} or Cr^{2+} whose 3d shell has four electrons with the filling as shown below:

m_l	2	1	0	−1	−2
	↑	↑	↑	↑	

Therefore, $L = 2$ and $S = 2$. According to Hund's third rule, $J = L - S = 0$, giving the ground state as ${}^3D_0$.

Examples 1 and 2 deal with ions having a non-vanishing total angular momentum. These ions, as expected, show paramagnetic behaviour. On the other hand, Example 3 gives a case where an ion with a partially filled shell has no net angular mementum, a situation similar to that with ions having only completely-filled shells. In this case both L and S are independently nonzero unlike in the latter case where both of these are independently zero. As we will see later in Section 13.5, the atoms or ions in Example 3 show a second-order paramagnetism. It is regarded as a correction to the diamagnetic term, though of opposite sign. The susceptibility associated with this correction at low temperatures is found to be temperature independent. It is famous in literature as van Vleck paramagnetism.

13.4 QUANTUM THEORY OF MAGNETIC SUSCEPTIBILITY: A QUANTUM MECHANICAL FORMULATION

Magnetization (also known as the intensity of magnetization) of a quantum mechanical system having N magnetic ions per unit volume at $T = 0$ is defined as

$$M(B_0) = -N \frac{\partial E_0(B_0)}{\partial B_0} \tag{13.36}$$

where $E_0(B_0)$ is the ionic ground state energy in the presence of the field B_0.

In the state of thermal equilibrium at temperature T, the thermal average of the magnetization of each excited state of energy $E_n(B_0)$ gives the measure of magnetization, i.e.

$$M(B_0, T) = \frac{\sum_n M_n(B_0) \exp\left(-\frac{E_n}{k_B T}\right)}{\sum_n \exp\left(-\frac{E_n}{k_B T}\right)} \tag{13.37}$$

where

$$M_n(B_0) = -N \frac{\partial E_n(B_0)}{\partial B_0} \tag{13.38}$$

The thermodynamical definition of magnetization is given as

$$M = -N \frac{\partial F}{\partial B_0} \tag{13.39}$$

where F is the magnetic Helmholtz free energy. The free energy F is defined by the well-known rule of statistical mechanics written as

$$\exp\left(-\frac{F}{k_B T}\right) = \sum_n \exp\left[\frac{-E_n(B_0)}{k_B T}\right] \tag{13.40}$$

The general definition of susceptibility gives

$$\chi = \mu_0 \frac{\partial M}{\partial B_0} = -\mu_0 N \frac{\partial^2 F}{\partial B_0^2} \tag{13.41}$$

But, in most of the cases, M is found to be very accurately linear in B_0 for attainable field strengths. In such a case, the definition of χ reduces to

$$\chi = \frac{\mu_0 M}{B_0} \qquad \text{[see (13.5)]}$$

The theory discussed so far makes it abundantly clear that the determination of energy shifts of various ionic states induced by a magnetic field is a prerequisite to the calculation of susceptibility. The exercise is carried out below in a quantum mechanical approach.

The part of the Hamiltonian operator of the energy of an atomic dipole in a magnetic field $\mathbf{B}_0$ owing to its orbital magnetic moment $\boldsymbol{\mu}_L$ is

$$\Delta \mathbf{H}_L = -\boldsymbol{\mu}_L \cdot \mathbf{B}_0 = \mu_B \mathbf{L} \cdot \mathbf{B}_0 \tag{13.42}$$

Similarly, the interaction energy owing to electron spin gives another term in the Hamiltonian operator as

$$\Delta \mathbf{H}_S = -\boldsymbol{\mu}_S \cdot \mathbf{B}_0 = g_0 \mu_B \mathbf{S} \cdot \mathbf{B}_0 = g_0 \mu_B S_z B_0 \tag{13.43}$$

where

$$\mathbf{S}_z = \sum_i \mathbf{s}_z^i \quad \text{with } \mathbf{s}_z^i = \frac{1}{2} \sigma_i \ (\boldsymbol{\sigma}_i \text{ is a Pauli spin matrix})$$

It is assumed here that the magnetic field is applied along the z-direction. In the presence of a magnetic field, the linear momentum of an electron is given by

$$\mathbf{p}_{\text{field}} = \mathbf{p} + e\mathbf{A}(\mathbf{r}) \tag{13.44}$$

where $\mathbf{p}$ is the linear momentum of the electron in the absence of the field and $\mathbf{A}$ denotes the vector potential of the field to which it is related as

$$\mathbf{B}_0 = \text{curl } \mathbf{A} \quad \text{with div } \mathbf{A} = 0 \tag{13.45}$$

For a homogeneous field, a possible choice of the vector potential is

$$\mathbf{A} = -\frac{1}{2}\,\mathbf{r} \times \mathbf{B}_0 \tag{13.46}$$

As a result, we can write the kinetic energy part of the Hamiltonian as

$$H_{\text{kin}} = \frac{1}{2m}\sum_i \left(\mathbf{p}_i - \frac{e}{2}\,\mathbf{r}_i \times \mathbf{B}_0\right)^2$$

$$= T_0 + \mu_B \mathbf{L}\cdot\mathbf{B}_0 + \frac{e^2}{8m} B_0^2 \sum_i (x_i^2 + y_i^2) \tag{13.47}$$

where T_0 is the kinetic energy in the absence of the field and $\hbar\mathbf{L} = \sum_i \mathbf{r}_i \times \mathbf{p}_i$.

Combining the spin term (13.43) with (13.47), we get the total interaction Hamiltonian (the field dependent) in the form

$$\Delta H = \mu_B(\mathbf{L} + g_0\mathbf{S})\cdot\mathbf{B}_0 + \frac{e^2}{8m} B_0^2 \sum_i (x_i^2 + y_i^2) \tag{13.48}$$

Changes in energy affected by (13.48) even with the strongest magnetic fields that can be produced in a laboratory are very small on the scale of atomic excitation energies. Hence it may be justified to follow the ordinary perturbation approach for calculating the changes in electron energies induced by a magnetic field.

The dependence of susceptibility on the second derivative of energy (13.41) indicates that it would suffice to confine the perturbation calculations to second-order terms. If energy E_n changes by ΔE_n on applying the field, this change according to the standard result of the second-order perturbation theory is expressed as

$$\Delta E_n = \langle \phi_n | \Delta H | \phi_n \rangle + \sum_{n' \neq n} \frac{|\langle \phi_n | \Delta H | \phi_{n'} \rangle|^2}{E_n - E_{n'}} \tag{13.49}$$

where ϕ_n denotes the eigenfunction of the nth energy state.

On substituting ΔH from (13.48) and retaining terms up to those in quadratic in B_0, we obtain

$$\Delta E_n = \mu_B \mathbf{B}_0 \cdot \langle \phi_n | \mathbf{L} + g_0 \mathbf{S} | \phi_n \rangle + \sum_{n' \neq n} \frac{|\langle \phi_n | \mu_B \mathbf{B}_0 \cdot (\mathbf{L} + g_0\mathbf{S}) | \phi_{n'} \rangle|^2}{E_n - E_{n'}}$$

$$+ \frac{e^2}{8m} B_0^2 \langle \phi_n | \sum_i (x_i^2 + y_i^2) | \phi_n \rangle \tag{13.50}$$

This relation serves as the basis for the description of magnetic susceptibility of individual atoms, ions or molecules. It can also be applied to ionic and molecular solids by computing the susceptibility ion by ion, provided the concerned solid may be regarded as a collection of only slightly deformed ions. We show below how it accounts for diamagnetism or paramagnetism at atomic level.

13.4.1 Diamagnetism

Consider the case of a solid composed of ions whose all electronic shells are filled. An ion has zero spin and orbital angular momentum in its ground state represented by the wavefunction ϕ_0, i.e.

$$\mathbf{J}\,|\phi_0\rangle = \mathbf{L}\,|\phi_0\rangle = \mathbf{S}\,|\phi_0\rangle = 0 \tag{13.51}$$

In relation (13.50), only the last term contributes to the field-induced shift in the ground state energy:

$$\Delta E_0 = \frac{e^2}{8m} B_0^2 \langle \phi_0 | \sum_i (x_i^2 + y_i^2) | \phi_0 \rangle$$

$$= \frac{e^2}{12\,m} B_0^2 \langle \phi_0 | \sum_i r_i^2 | \phi_0 \rangle \qquad \text{[under (13.15)]} \tag{13.52}$$

In the state of thermal equilibrium ions are generally in their ground state, excepting the situation at high temperatures. Therefore, the susceptibility of a solid with N atoms or ions per unit volume at room temperature is given as

$$= -\,\mu_0 N \frac{\partial^2 \Delta E_0}{\partial B_0^2}$$

$$= -\,\frac{\mu_0 N e^2}{6m} \langle \phi_0 | \sum_i r_i^2 | \phi_0 \rangle \tag{13.53}$$

If there are Z electrons in an ion, the mean square radius of the ion may be defined by

$$\langle r^2 \rangle = \frac{\langle \phi_0 | \sum_i r_i^2 | \phi_0 \rangle}{Z} \tag{13.54}$$

This leads to

$$\chi = -\,\frac{\mu_0 N Z e^2}{6m} \langle r^2 \rangle \tag{13.55}$$

which is surprisingly the same as (13.18), obtained on the basis of purely classical considerations.

13.4.2 Paramagnetism

The paramagnetism is attributed generally to ions with a partially-filled shell. There arise two cases of partially-filled shells. In one case, $J \neq 0$ and paramagnetism follows from the first term in relation (13.50). Being the leading term, it determines the major paramagnetic effect. Its magnitude is so large that the other two terms can simply be ignored at the first instance. In the other case, $J = 0$ which refers to ions whose partially-filled shell is one electron short of being half-filled. The paramagnetism in this case appears only as a second-order effect. We discuss here the first case and treat the second in the following section.

The ionic ground state in the present case in the absence of a magnetic field is $(2J + 1)$-fold degenerate. This produces difficulty in calculating the Landé splitting factor (g-factor). Furthermore, the relation

$$\mathbf{L} + g_0\mathbf{S} = g\mathbf{J}$$

holds good only within the ground state multiplet of $(2J + 1)$ states in zero field. Only when the separation between the ground state multiplet in zero field and the first excited state multiplet is much greater than k_BT, is there a sizeable contribution from the states of the ground state multiplet to the free energy. This is usually the case in practice. Only in this event, the above relation allows us to treat the first term in (13.50) to be equal to the interaction energy $(-\boldsymbol{\mu} \cdot \mathbf{B}_0)$, with $\boldsymbol{\mu}$ being given by

$$\boldsymbol{\mu} = -\mu_B g\mathbf{J}$$

The calculation of susceptibility, nevertheless, is not straightforward on account of the degenerate character of the ground state. When the field becomes zero, the splitting of the $(2J + 1)$ lowest lying states would be smaller than k_BT, rendering it unjustifiable to substitute the free energy with the ground state energy as is possible in the case of a non-degenerate ground state. Therefore, for the calculation of susceptibility we have to take recourse to a method based on principles of statistical mechanics as detailed below.

Consider a system containing identical ions with angular momentum J. Assuming that only the lowest $(2J + 1)$ states are thermally excited with appreciable probability, we can define the magnetic Helmholtz free energy F by the fundamental statistical mechanical rule (13.40):

$$\exp\left(-\frac{F}{k_BT}\right) = \sum_{J_z=-J}^{J} \exp\left(-\frac{E_J^z}{k_BT}\right)$$

$$= \sum_{J_z=-J}^{J} \exp\left(-\frac{g\mu_B J_z B_0}{k_BT}\right) \tag{13.56}$$

with the applied magnetic field in the z-direction.

The sum is in a geometrical progression and it can be easily evaluated, leading to

$$\exp(-\beta F) = \frac{\exp\left[\beta\gamma B_0\left(J + \frac{1}{2}\right)\right] - \exp\left[-\beta\gamma B_0\left(J + \frac{1}{2}\right)\right]}{\exp\left(\frac{\beta\gamma B_0}{2}\right) - \exp\left(-\frac{\beta\gamma B_0}{2}\right)} \tag{13.57}$$

with $\beta = \dfrac{1}{k_BT}$ and $\gamma = g\mu_B$.

If there are N such ions per unit volume, the magnetization can be written as

$$M = -N\frac{\partial F}{\partial B_0}$$

$$= N\gamma J\, \mathrm{B}_J(x) \tag{13.58}$$

where $B_J(x)$ is called the 'Brillouin function', defined by

$$B_J(x) = \left(\frac{2J+1}{2J}\right) \coth\left(\frac{2J+1}{2J}\, x\right) - \frac{1}{2J} \coth\left(\frac{1}{2J}\, x\right) \tag{13.59}$$

with

$$x = \beta\gamma J B_0 \tag{13.60}$$

$$= \frac{g\mu_B\, JB_0}{k_B T} = \frac{\mu B_0}{k_B T}$$

The experimental plot of the magnetic moment versus B_0/T for certain materials is shown in Fig. 13.4. It shows that as $T \to 0$, the magnetic moment approaches saturation, i.e. each ion approaches complete alignment with the field. This implies that $[J_z]$ acquires its maximum value J which may be possible only if $k_B T << \gamma B_0$. This condition refers to that of the lowest temperatures and highest fields. Such a condition is, however, usually not realized in practice. On the contrary, over a wide

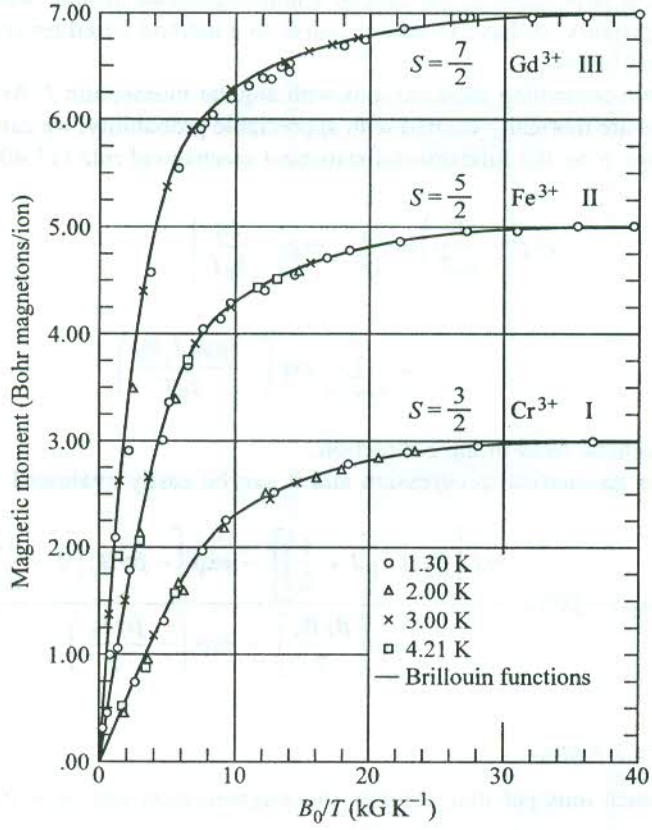


FIG. 13.4 Magnetic moment versus B_0/T plot showing saturation at higher fields for spherical samples: I—potassium chromium alum, II—ferric ammonium alum and III—gadolinium sulphate octahydrate. [After W.E. Henry, *Phy. Rev.*, **88**, 559 (1952).]

range of temperature and fields the condition $k_B T >> \gamma B_0$ is nicely applicable. In this limit x is small and

$$\coth x = \frac{1}{x} + \frac{1}{3}x - \frac{1}{45}x^3 + \ldots$$

$$\simeq \frac{1}{x} + \frac{1}{3}x \tag{13.61}$$

and

$$B_J(x) \simeq \left(\frac{J+1}{3J}\right)x \tag{13.62}$$

Using (13.62) with (13.58), we get

$$\chi = \frac{\mu_0 N J(J+1)\, g^2 \mu_B^2}{3k_B T} \tag{13.63}$$

$$= \frac{\mu_0 N p^2 \mu_B^2}{3k_B T}$$

where p is the 'effective Bohr magneton number', given by

$$p = g[J(J+1)]^{1/2} \tag{13.64}$$

Relation (13.63) can be put in the form of the Curie law $\chi = C/T$, where C is Curie constant, defined by

$$C = \frac{\mu_0 N p^2 \mu_B^2}{3\, k_B} \tag{13.65}$$

The paramagnetic susceptibility (13.63) is greater than the diamagnetic susceptibility (13.55) by a factor of about 500 at room temperature. This confirms that in a solid containing ions that have a partially-filled shell with $J \neq 0$, the contribution from the partially-filled shell to the total susceptibility dominates the diamagnetic contribution from filled shells.

13.4.3 Application to Magnetic Ions in Solids: Effect of the Crystal Field

In this section we examine how well the theory of paramagnetism of free ions is applicable to paramagnetic solids. The Curie law (13.63) is obeyed very well in insulating solids containing rare earth ions. The p-values (the effective Bohr magneton numbers) are derived from the coefficient of $1/T$ in (13.63) using the measured values of χ. These are in exceedingly good agreement with those calculated with (13.64) for all rare earth ions* excepting samarium and europium. For both of these ions the J-multiplet lying just above the ground state is very close in energy as a consequence of which a couple of assumptions made in the derivation of the Curie law remain no more valid:

* N.W. Aschroft and N.D. Mermin, *Solid State Physics*, Table 31.3 (Saunders College, 1988).

(i) The second term in (13.50) that is ignored in the derivation of the Curie law becomes important because the denominators $(E_n - E_{n'})$ are now very small.

(ii) There is an appreciable probability of thermally exciting some ions from the state(s) of lowest J to higher states, contrary to what is assumed for deriving the Curie law.

These observations explain the discrepancy noticed in respect of samarium and europium ions. In addition, our analysis is not applicable to the latter since J is equal to zero in that case. The excellent agreement in respect of all other ions, however, leads to the conclusion that the rare earth ions can be treated as free ions even in solids.

Let us now take up the case of 3d transition metal ions (the iron group). Other transition metal ions (4d and 5d) will not be discussed here in view of their complex behaviour. In the case of 3d transition metal ions, although the Curie law is obeyed, the experimental p-values* are not accounted for by (13.64). The agreement is close only if J is replaced by S in the relation, assuming that L is zero though S will be still given by the Hund's rules. This phenomenon is known as *quenching of the orbital angular momentum* and attributed to the crystal field effect. The crystal field effect is stronger in transition metal ions since their partially filled d-shell (3d in the iron group) happens to be the outermost shell. The electrons in the d-shell are thus directly exposed to the electric field created by ions surrounding the magnetic ion of concern. The coupling between L and S is largely broken so that the states are no longer specified by their J values. Further, the $(2L + 1)$ sublevels belonging to a certain L and degenerate in the free ion may be split by the crystal field. The splitting decreases the contribution of the orbital motion to the magnetic moment.

On the other hand, the crystal field effect for rare earth ions is almost negligible since their partially-filled shell (4f) lies deep inside the ion, sheltered by 5s and 5p shells. This explains why these ions behave as almost free ions even when they are embedded in crystals.

The symmetry of a crystal field is determined by the arrangement of ions around the magentic ion. Since the symmetry is never spherical in a crystal, the basis for the application of the Hund's rules prevails no more. Under the action of an asymmetrical (noncentral) crystal field, the plane of the electron orbit moves about and the precession of the orbital angular momentum sets in. As a result, all the three components of the orbital angular momentum (L_x, L_y, L_z) may average out to zero, though L^2 may still be a constant of motion with the mean value $L(L + 1)$. This contrasts with the behaviour in a central field where in accordance with the quantum theory one component of L, usually L_z, and L^2 are constant. For a magnetic field along the z-direction, the contribution of orbital motion to magnetic moment is proportional to the quantum expectation value of L_z. Therefore, when L_z averages out to zero, the orbital angular momentum, and hence the orbital magnetic moment, is said to be quenched.

13.5 van VLECK PARAMAGNETISM

In this section we discuss the case of a partially-filled shell with $J = 0$, giving a non-degenerate ground state for which calculations are simpler. To repeat, it is the case of ions whose partially-filled shell is one electron short of being half filled (see Example 3 in Section 13.3.1). In a filled shell too, $J = 0$. But the present case is different in the sense that the second term in (13.50) does not vanish here, though it does so for ions having only completely occupied shell, simply because L and S

*N.W. Ashcroft and N.D. Mermin, *Solid State Physics*, Table 31.4 (Saunders College, 1988).

are both independently zero for a completely filled shell. Hence the shift in the ground state energy induced by the magnetic field in the present case is written as

$$\Delta E_0 = \frac{e^2}{8m} B_0^2 \langle \phi_0 | \sum_i (x_i^2 + y_i^2) | \phi_0 \rangle - \sum_n \frac{|\langle \phi_0 | \mu_B \mathbf{B}_0 \cdot (\mathbf{L} + g_0 \mathbf{S}) | \phi_n \rangle|^2}{E_n - E_0} \tag{13.66}$$

If the system has N such ions per unit volume,

$$\chi = -\mu_0 N \frac{\partial^2 \Delta E_0}{\partial B_0^2}$$

$$= -\mu_0 N \left[\frac{e^2}{4m} \langle \phi_0 | \sum_i (x_i^2 + y_i^2) | \phi_0 \rangle - 2\mu_B^2 \sum_n \frac{|\langle \phi_0 | (\mathbf{L}_z + g_0 \mathbf{S}_z) | \phi_n \rangle|^2}{E_n - E_0} \right] \tag{13.67}$$

The first term with a negative sign (depending on the diagonal matrix element) in (13.67) represents the usual Langevin's diamagnetic susceptibility (13.55). The second term is positive because the energy of excited state E_n is always greater than the ground state energy E_0. It indicates the preference of the magnetic moment for aligning with the field direction (z-direction) which is a definite evidence of paramagnetism. This term may be regarded as a correction to the diamagnetic contribution. It must be stressed that the paramagnetism being discussed is fairly weak since it appears through a non-diagonal matrix element as a second-order term. In view of this fact, the overall magnetic behaviour of ions is determined by the balance between two weak terms of opposite sign.

The paramagnetic term is required to be examined in two extreme limits:

1. $(E_n - E_0) << k_B T$

In this limit the excess population in the ground state over the excited state of energy E_n is approximately equal to $\frac{N(E_n - E_0)}{2k_B T}$. This would give a magnetization

$$M = \frac{NB_0 \mu_B^2}{k_B T} \sum_n |\langle \phi_0 | (\mathbf{L}_z + g_0 \mathbf{S}_z) | \phi_n \rangle|^2 \tag{13.68}$$

and

$$\chi = \frac{\mu_0 N \mu_B^2}{k_B T} \sum_n |\langle \phi_0 | (\mathbf{L}_z + g_0 \mathbf{S}_z) | \phi_n \rangle|^2 \tag{13.69}$$

The above contribution is of the usual Curie form with the difference that the magnetization occurs here by polarization of the states of the system and not by the usual redistribution of ions among the spin states. We also observe that χ does not depend on the separation of excited states from the ground state.

2. $(E_n - E_0) >> k_B T$

In this condition, almost all the ions stay in the ground state with little probability of being excited to higher states in the state of thermal equilibrium. Also, in this limit the free energy is just the ground state energy. Treating all the ions to be in the ground state,

$$M = 2NB_0\mu_B^2 \sum_n \frac{|\langle\phi_0|(\mathbf{L}_z + g_0\mathbf{S}_z)|\phi_n\rangle|^2}{E_n - E_0} \tag{13.70}$$

and therefore,

$$\chi = 2\mu_0 N\mu_B^2 \sum_n \frac{|\langle\phi_0|(\mathbf{L}_z + g_0\mathbf{S}_z)|\phi_n\rangle|^2}{E_n - E_0} \tag{13.71}$$

This expression denotes a temperature independent paramagnetism which is commonly known as the van Vleck paramagnetism.

13.6 PAULI PARAMAGNETISM

The contribution of conduction electrons to the magnetic moment of a metal is discussed in this section. It is interesting to find that the conduction electrons, apart from producing diamagnetism on account of their angular momentum, also exhibit a temperature independent paramagnetism, known as 'Pauli paramagnetism'. These electrons can be safely treated in the independent electron approximation as they are neither localized in space like those in filled shells nor favoured by the Pauli principle to be localized on different ions. However, the complex response of the orbital motion to the magnetic field poses serious difficulties for an accurate calculation of the magnetic properties. In view of this fact, we settle for an oversimplified model in which the electron is considered to have a spin magnetic moment but no charge and the response of the orbital motion is completely ignored. As a result the calculations become far more easier and even the free energy does not figure in calculations.

The degeneracy between the electrons of opposite spin that share the same orbital is resolved by a magnetic field. In a metal this causes a redistribution of electrons between the two spin orientations and, therefore, gives rise to a magnetic moment. Each electron with its spin parallel to the field lines contributes a magnetic moment $-g_0\mu_B/2$ and the one with spin antiparallel to the field contributes $g_0\mu_B/2$ to the magnetic moment of the system of conduction electrons. If $n_\pm$ denotes the number of electrons per unit volume with spin parallel (+) or antiparallel (–) to the external induction B_0, the magnetization will be given by

$$M = -\frac{1}{2} g_0\mu_B n_+ + \frac{1}{2} g_0\mu_B n_-$$

$$= -\frac{1}{2} g_0\mu_B (n_+ - n_-) \tag{13.72}$$

The electrons with spin parallel and magnetic moment antiparallel to $\mathbf{B}_0$ are in states whose energy has been increased by $\frac{1}{2} g_0\mu_B B_0$ with respect to the zero field condition. Similarly, for the electrons with spin antiparallel and magnetic moment parallel to the field, the energy is lowered by $\frac{1}{2} g_0\mu_B B_0$. The energy parabola splits into two parabolae (Fig. 13.5) that are separated by $g_0\mu_B B_0$ on the energy axis. The figure shows the separation as $2\mu_B B_0$ under the assumption, $g_0 = 2$, since the conduction electrons can be treated very nearly like free electrons. Henceforth, this assumption shall apply to all our discussions.

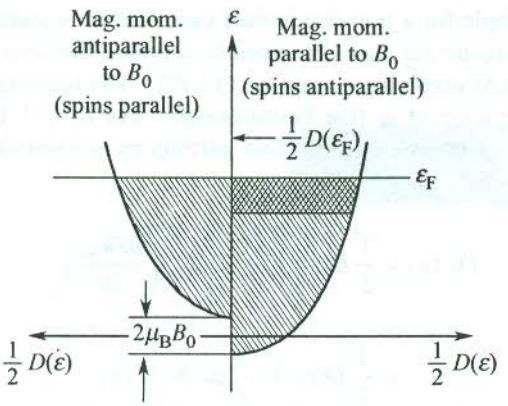


FIG. 13.5 Relative shift energy parabolae referred to parallel and antiparallel spin states of conduction electrons in a magnetic field $\mathbf{B}_0$. The cross-hatched area represents excess (uncompensated) electron spins parallel to $\mathbf{B}_0$, that account for the net magnetic moment.

Suppose $D_{\pm}(\varepsilon)$ represents the density of energy levels at energy ε belonging to the two sets of electrons with densities n_+ and n_- in the presence of the field. Then,

$$D_{\pm}(\varepsilon) = \frac{1}{2}D(\varepsilon) \tag{13.73}$$

where $D(\varepsilon)$ is ordinary density of levels in the absence of the field.

The energy of each electronic level with spin parallel to the field lines is raised by $\mu_B B_0$. This implies that the number of levels with energy ε for this spin species in the presence of the field is the same as with $(\varepsilon - \mu_B B_0)$ in the absence of the field. That is,

$$D_+(\varepsilon) = \frac{1}{2}D(\varepsilon - \mu_B B_0) \tag{13.74a}$$

Similarly,

$$D_-(\varepsilon) = \frac{1}{2}D(\varepsilon + \mu_B B_0) \tag{13.74b}$$

The number of electrons per unit volume for the two sets of spin orientations is

$$n_{\pm} = \int D_{\pm}(\varepsilon) f(\varepsilon)\, d\varepsilon \tag{13.75}$$

where $f(\varepsilon)$ is the Fermi–Dirac distribution function

$$\frac{1}{\exp\left(\dfrac{\varepsilon - \mu}{k_B T}\right) + 1}$$

As usual the chemical potential μ is estimated by requiring the total electron density n to be consistent with

$$n = n_+ + n_- \tag{13.76}$$

The calculations are simple for a non-degenerate case. But the conduction electrons or for that matter metals belong to a degenerate case. We need to estimate the density of levels $D_{\pm}(\varepsilon)$ for the calculation of magnetization M involving n_+ and n_- [(13.72)]. The important variation of the density of levels $D(\varepsilon)$ occurs on the scale of ε_F (the Fermi energy). For $B_0 = 1$ T, $\mu_B B_0 \sim 10^{-4}\ \varepsilon_F$ which is very small compared to the electron energies. This permits us to expand $D_{\pm}(\varepsilon)$ [given by (13.74)] as

$$D_{\pm}(\varepsilon) = \frac{1}{2} D(\varepsilon) \mp \frac{1}{2} \mu_B B_0 \left(\frac{\partial D(\varepsilon)}{\partial \varepsilon}\right)_0$$

$$= \frac{1}{2} D(\varepsilon) \mp \frac{1}{2} \mu_B B_0 D'(\varepsilon) \tag{13.77}$$

Using (13.77) with (13.75), we get

$$n_{\pm} = \frac{1}{2} \int D(\varepsilon) f(\varepsilon)\, d\varepsilon \mp \frac{1}{2} \mu_B B_0 \int D'(\varepsilon) f(\varepsilon)\, d\varepsilon \tag{13.78}$$

Putting these values in (13.76), we find that the overall electron density, as expected for moderate fields, does not change on the application of a magnetic field:

$$n = \int D(\varepsilon) f(\varepsilon)\, d\varepsilon \tag{13.79}$$

Using (13.76) with (13.72) gives the magnetization

$$M = \mu_B^2 B_0 \int D'(\varepsilon) f(\varepsilon)\, d\varepsilon$$

$$= \mu_B^2 B_0 \int D(\varepsilon) \left(-\frac{\partial f}{\partial \varepsilon}\right) d\varepsilon \qquad \text{(on integrating by parts)} \tag{13.80}$$

At $T = 0$, $-\dfrac{\partial f}{\partial \varepsilon} = \delta(\varepsilon - \varepsilon_F)$, therefore, at absolute zero, $D(\varepsilon)$ should be rightfully replaced by $D(\varepsilon_F)$. But in practice we are always concerned with non-zero temperatures. As may be seen in Fig. 6.10, the variation of f at temperatures not too high† ($\sim 10^4$ K) is maximum around ε_F. Since the experiments are performed at temperatures far below 10^4 K, $D(\varepsilon_F)$ determines the all important density of levels and it can replace $D(\varepsilon)$ in (13.80) without involving a meaningful error. Then, (13.80) becomes

$$M = \mu_B^2 B_0 D(\varepsilon_F) \int \left(-\frac{\partial f}{\partial \varepsilon}\right) d\varepsilon$$

$$= \mu_B^2 B_0 D(\varepsilon_F) \tag{13.81}$$

† At $T \neq 0$, corrections to $(-\partial f/\partial \varepsilon)$ are of the order of $(k_B T/\varepsilon_F)^2$. See N.W. Aschroft and N.D. Mermin, *Solid State Physics*, Ch. 2 (Saunders College, 1988).

and

$$\chi = \frac{\mu_0 M}{B_0} = \mu_0 \mu_B^2 D(\varepsilon_F) = \frac{3\mu_0 \mu_B^2 N}{2k_B T_F} \tag{13.82}$$

with

$$D(\varepsilon_F) = \frac{3N}{2\varepsilon_F} = \frac{3N}{2k_B T_F} \qquad \text{[see (6.37)]}$$

where N is the total number of electrons in the system and T_F is the Fermi temperature (~ 10^4 K).

The result (13.82) determines the Pauli paramagnetic susceptibility; a contribution from the conduction electrons that is essentially temperature independent and not accounted by calculations based on classical statistics. As remarked in Section 6.1.1, it is identified as one of the most prominent gains of using quantum statistics. The use of Fermi–Dirac statistics in the above treatment is in conformity with the antisymmetric character of the spinning electrons. The expression is ideally valid at $T = 0$, though applicable in the limit $T << T_F$ (= ε_F/k_B). The Pauli susceptibility is measured by making use of the nuclear magnetic resonance (NMR) technique, to be discussed in Section 13.9.2. The measured values are in fairly good agreement with those calculated for the electrons with (13.82). These are of the order of 10^{-6}, i.e. of the size of typical diamagnetic susceptibilities. Thus the Pauli susceptibilities are about hundreds of times smaller than the paramagnetic susceptibilities of magnetic ions.

In addition to the Pauli paramagnetism, there is a diamagnetic contribution too from the conduction electrons. A magnetic field induces an orbital motion creating a magnetization that is antiparallel to the field lines. The magnetism thus produced is known as 'Landau diamagnetism'. For free electrons,

$$\chi_{\text{Landau}} = -\frac{1}{3}\,\chi_{\text{Pauli}} \tag{13.83}$$

A comparative view of various forms of paramagnetism and diamagnetism is provided by Fig. 13.6 with the help of a χ versus T plot.

13.7 NUCLEAR PARAMAGNETISM

In addition to the orbital motion and the spin of electrons, the nuclear spin also contributes to the magnetic moment of atoms. The nuclear magnetic moment is expressed in units of the nuclear magneton in analogy with the Bohr magneton defined by

$$\mu_n = \frac{e\hbar}{2M_P} = 5.051 \times 10^{-27}\ \text{J/T} \tag{13.84}$$

where M_P is the proton mass.

Comparing (13.84) with (13.30) we see that the nuclear magneton is smaller than the Bohr magneton in the ratio of the proton mass to the electron mass (~10^3). Therefore, the static nuclear paramagnetism is masked by the electron paramagnetism in paramagnetic substances. Solid hydrogen

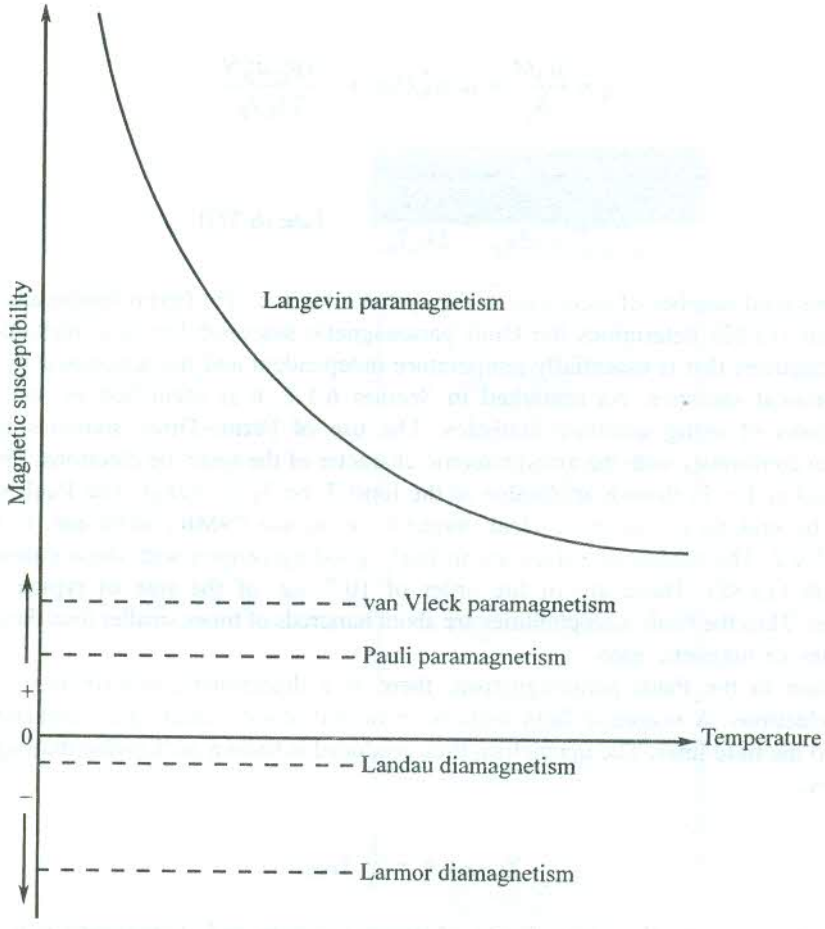


FIG. 13.6 A comparative view of various forms of paramagnetism and diamagnetism, giving χ versus T plot.

shows nuclear paramagnetism, though its electron configuraiton suggests diamagnetism only. The value of proton magnetic moment (2.793 nuclear magnetons) is verified by these measurements. Heavy nuclei are found to possess even smaller magnetic moments. The method of nuclear magnetic resonance (NMR) is used to determine the nuclear magnetic moments. The nuclear magnetic moments, being very small compared to the electronic components, are almost ignored while discussing static magnetization.

13.8 COOLING BY ADIABATIC DEMAGNETIZATION

The cooling of a system, in principle, can be produced by an isothermal reduction of its entropy by some operation, followed by an adiabatic (isentropic) reversal of the operation. All that is required is to find a system whose entropy can be lowered. For cooling below 1 K, for which the adiabatic demagnetization has been used, the selection of an appropriate system is not easy because at these low temperatures there is hardly any entropy left in any system. There may be

only some solids as there are no gases or liquids suitable for the purpose. The lattice entropy† of solids is, however, not encouraging. At 1 K for a solid with $\theta_D = 100$ K it is merely $\sim 10^{-4}\, Nk_B$. The choice falls on paramagnetic solids which possess appreciable spin entropy because of the random orientations of atomic magnetic moments.

A magnetic field is applied to reduce the entropy of a paramagnetic salt immersed in liquid helium. The creation of a more ordered state of spins leads to the reduction of entropy. The liquid helium absorbs any heat liberated in the process. The salt is then removed from the liquid helium bath and the magnetic field is switched on under adiabatic conditions implying that the entropy remains unchanged. The field is switched off slowly to ensure that the system passes through states always in thermal equilibrium. As we will explain below, the temperature will have to fall if the entropy is to remain unchanged even after the magnetic field is completely withdrawn. In order to preserve the cooling thus produced, no heat should flow into the spin system. The most likely source of heat is the lattice entropy. Therefore, it is most important that the lattice entropy of the salt be smaller than its spin entropy to disallow heating. Salts containing rare earth elements because of their larger magnetic moments adequately satisfy the above condition and, therefore, used for adiabatic demagnetization.

We learn from the above discussion that the method of adiabatic demagnetization can be applied in a limited range of temperatures over which the spin entropy dominates its lattice counterpart. In practice, the method is useful for cooling from a few K down, at the best, to 10^{-3} K.

Consider an entirely disordered spin system at high temperatures where the thermal disorder overpowers magnetic interactions that could produce any preferential spin orientations. A spin system of N ions, each of spin J, has $(2J + 1)$ states in total over which the spins are distributed. If W represents the number of possible ways of distribution in a quantized spin system, the entropy S of the system is defined by

$$S = k_B \ln W \tag{13.85}$$

where

$$W = (2J + 1)^N \tag{13.86}$$

Therefore,

$$S = Nk_B \ln (2J + 1) \tag{13.87}$$

showing that the entropy is temperature independent in the absence of a magnetic field.

For $J = 1/2$, the entropy is equal to $Nk_B \ln 2$. It shows that even for a two-level system the spin entropy is far greater than the lattice entropy around 1 K ($\sim 10^{-4}\, Nk_B$). On the application of a magnetic field B_0, the $(2J + 1)$ states are separated in energy and the entropy is lowered when the lower levels gain in population. When the magnetic field is withdrawn adiabatically, the temperature falls so that the entropy may remain unchanged as required in an adiabatic change. We attempt below to find a theoretical basis for this phenomenon.

---

† The lattice entropy is defined as

$$S = \int_0^T C_V \cdot \frac{dT}{T} = \int_0^T \frac{12\pi^4 Nk_B}{5} \left(\frac{T}{\theta_D}\right)^3 \cdot \frac{dT}{T} = 78\, Nk_B \left(\frac{T}{\theta_D}\right)^3$$

with

$$C_V = \frac{12\pi^4 Nk_B}{5} \left(\frac{T}{\theta_D}\right)^3 \qquad \text{[see (5.39)]}$$

The entropy is defined in terms of the Helmholtz free energy F and the internal energy U as

$$S = \frac{U - F}{T}$$

$$= k_B\beta(U - F) \qquad \text{with } \beta = \frac{1}{k_B T} \tag{13.88}$$

Further, we learn from relation (13.40) that for a system of non-interacting paramagnetic ions, βF depends on B_0 only through the product βB_0. This requires F to be of the form

$$F = \frac{1}{\beta} f(\beta B_0) \tag{13.89}$$

where $f(\beta B_0)$ denotes a function of the product (βB_0).

Since U can be expressed as

$$U = \frac{\partial}{\partial \beta}(\beta F) \tag{13.90}$$

we can write the entropy using (13.88) as

$$S = k_B \beta^2 \frac{\partial F}{\partial \beta} \tag{13.91}$$

or

$$S = k_B[-f(\beta B_0) + \beta B_0 f'(\beta B_0)] \qquad \text{[using (13.89)]} \tag{13.92}$$

Relation (13.92) indicates that S, too, depends on the product βB_0. Thus, when S is constant, we would require that βB_0 remains constant, i.e. B_0/T = constant, which we can write as

$$\frac{(B_0)_{\text{initial}}}{T_{\text{initial}}} = \frac{(B_0)_{\text{final}}}{T_{\text{final}}}$$

or

$$T_{\text{final}} = \frac{(B_0)_{\text{final}}}{(B_0)_{\text{initial}}} \times T_{\text{initial}} \tag{13.93}$$

Instead of withdrawing the field completely, if we decrease it adiabatically to a certain value, according to (13.93), then, $T_{\text{final}} < T_{\text{inital}}$.

The lower limit of temperature that can be approached is determiend by the validity of (13.92). In principle, we could even reach the absolute zero, if the relation were absolutely valid. Had it really been true, the zero-field entropy would not have been found to be temperature dependent as shown in Fig. 13.7. The observed temperature dependence leads to the conclusion that the entropy will really drop to zero as the absolute zero is approached, a result that is consistent with the third law of thermodynamics. Therefore, the condition (13.93) must fail at small fields to account for the temperature dependence of the zero-field entropy. In fact, even after the magnetic field is completely withdrawn there remains a magnetic field, though weak, mainly contributed by magnetic interactions between paramagnetic ions. This field in some cases may be as large as 100 gauss around 1 K. When

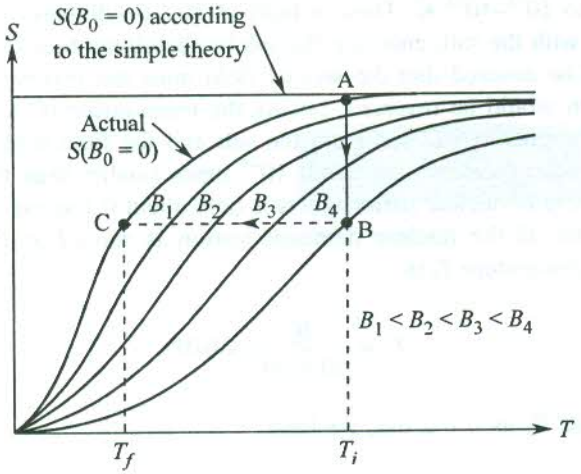


FIG. 13.7 Plot of entropy versus temperature cooling curves for interacting spins at various values of external induction B_0 in an adiabatic demagnetization process. The top horizontal line refers to the constant zero-field entropy $Nk_B \ln (2J + 1)$ of non-interacting spins. The actually observed zero-field entropy shows temperature dependence, represented by the curve passing through the point C.

this aspect and other effects, such as strong crystal field splittings at low temperatures, are taken into consideration, the temperature dependence in question is properly explained using the resultant modified expression for entropy.

The process of cooling by adiabatic demagnetization is explained in Fig. 13.7 with the help of S versus T curves for different magnetic fields. Initially, the paramagnetic salt rests immersed in liquid helium. The vertical line AB represents the first step of operation where the entropy is isothermally reduced from its initial value at A to a lower value at B by applying a magnetic field B_4. Notice that the point A lies on the zero-field S versus T curve and the point B on a curve obtained in the presence of the field B_4. There are other curves for fields lower than B_4—$B_3 > B_2 > B_1$. The S versus T behaviour in the absence of a magnetic field as predicted by simple theory and as observed is shown by two separate curves.

In the next step, the salt is removed from the liquid helium bath and the field reduced under adiabatic conditions (but slowly) through B_3, B_2, B_1 . . . to zero value. The operation is represented by the horizontal line BC. The intersections of BC with various S versus T curves gives the temperature $T_3 > T_2 > T_1$ corresponding to the fields B_3, B_2 and B_1 at three different stages during the process of adiabatic demagnetization. The lowest temperature approached is shown as T_f symbolized by the point C on the observed zero-field curve. As mentioned earlier, the lowest temperatures produced by adiabatic demagnetization are in the range of 10^{-3} K.

For approaching temperature below 10^{-3} K, a cascade process involving nuclear demagnetization is employed. The process consists of two steps. In the first step the cooling is produced by the demagnetization of an electron-spin system. It is followed in the next step by a further lowering of temperature resulting from demagnetization of a nuclear spin system in which the electron spin paramagnet acts as the reservoir.

Kurti et al.* carried out the first experiment on nuclear magnetization using copper as the nuclear spin system. An adiabatic demagnetization is performed in the usual way bringing the salt

* N. Kurti, F.N.H. Robinson, F.E. Simon and D.A. Sophr, *Nature*, **178**, 450 (1956).

and the copper down to 10^{-2}–10^{-3} K. Then, a field of 20–30 kilogauss is applied locally to the copper kept in contact with the salt, enabling the salt to absorb the heat released when the copper is magnetized. It must be ensured that the second field must not involve the salt, otherwise the path of demagnetization would be reversed, raising the temperature (C to B in Fig. 13.7). After being magnetized, the copper is isolated from the salt and the field turned off.

Since nuclear magnetic moments are about 10^{-3} times smaller than the electronic moments, the residual magnetic field in nuclear paramagnets is only about 0.1 gauss in contrast to 100 gauss in electron-spin systems. If the nuclear demagnetization is started at $B_0 = 50$ kilogauss and $T_i = 0.01$ K, the final temperature T_f is

$$T_f = \frac{B_0'}{50 \times 10^3} \times 0.01$$

Taking the residual field B_0' as 0.1 gauss, we have

$$T_f \simeq 10^{-7} \text{ K}$$

In their historic experiment, Kurti et al. started the magnetization at 0.02 K and could approach 1.2×10^{-6} K. According to the latest available reports, the lowest temperature approached so far is 280×10^{-12} K, achieved with the use of rhodium nuclei.†

13.9 MAGNETIC RESONANCE

It was shown in Section 13.3 that every magnetic moment $\boldsymbol{\mu}$ has an angular momentum $\mathbf{J}$ associated with it. When a constant magnetic field $\mathbf{B}_0$ is applied, the field exerts a torque on $\boldsymbol{\mu}$. The torque makes $\mathbf{J}$ to precess around the direction of the field as discussed in Section 13.1. The motion represents effectively the precession of the system, just as the precession of a spinning top, with the Larmor frequency ω_L [given by (13.10) for an electron orbit]. As a result, a magnetic dipole in the presence of a constant magnetic field is characterized with a natural frequency ω_L. This reveals that permanent magnetic dipoles are prone to show resonant behaviour in contrast to the normally observed behaviour of permanent electric dipoles. The induced electric dipoles, nevertheless, do show resonance because they have a natural frequency.

The magnetic resonance is observed by applying a second magnetic field which is alternating and perpendicular to the dc field. The source of this additional field is a radio-frequency radiation. The resonance occurs when the frequency of precession becomes equal to the frequency of rf radiation under conditions, favourable for resonance absorption. In experiment, the rf radiation is fed at a fixed frequency and the matching precession frequency is set by searching (varying) the dc magnetic field. The rf magnetic field produces a torque on the magnetization that is parallel to the dc field. The rf field bends the magnetic moment away from the dc field, causing the amplitude of precession to increase. The precession with an increasing amplitude in the present case produces the favourable condition for resonance. As in any resonant system, the damping comes into play through relaxation processes that limit the amplitude of precession. The above is a classical interpretation of the magnetic resonance. A description based on quantum physics will be taken up after making a few introductory remarks in the following paragraph.

† P.J. Hakonen et al., *Phys. Rev. Lett.*, **70**, 2818 (1993).

In view of the form of magnetism discussed in this chapter, two types of magnetic resonances are of basic importance. They are: electron paramagnetic resonance (EPR) and nuclear magnetic resonance (NMR). The EPR in solids is mostly concerned with spin angular momentum since the crystal field generally quenches the orbital angular momentum. It was shown in Section 13.4.3 that the quenching is nearly complete in crystals containing 3d-transition metal ions. When the quenching is not complete the orbital contribution is reflected in small splittings of states, caused by the crystal field. If the quenching can be treated as complete, the paramagnetic atom or ion appears to have a free spin. Therefore, to be precise, the resonance in such a case should better be referred to as electron spin resonance (ESR). In this section, we discuss ESR and NMR phenomena that are linked with the magnetic moments of spinning electrons and spinning nuclei, respectively.

13.9.1 Electron Spin Resonance (ESR)

We saw in Section 13.3 that the contribution of electron spin to the magnetic moment of an atom or ion is

$$\begin{aligned} \boldsymbol{\mu} &= -\, g_0 \mu_B \mathbf{S} \\ &= -\, g_0 (e/2m) \hbar \mathbf{S} \\ &= \gamma_e \hbar \mathbf{S} \end{aligned} \tag{13.94}$$

where γ_e is a constant known as the *magnetogyric ratio* and the subscript e stands for electron. It is defined as the ratio of the magnetic moment to the angular momentum:

$$\gamma_e = \boldsymbol{\mu}/\hbar\mathbf{S} = -\, g_0 (e/2m) \tag{13.95}$$

with $g_0 = 2.0023$. If we can consider[†] the electrons to be spinning freely and take $g_0 = 2$, then $\gamma_e = -(e/m)$.

According to quantum mechanics, the eigenvalue of the spin operator $\mathbf{S}$ is $\sqrt{S(S+1)}$ and, therefore, the magnitude of magnetic moment may be expressed as

$$\mu = \gamma_e \hbar \sqrt{S(S+1)} \tag{13.96}$$

For simplicity, we now consider the case of an unpaired spinning electron in a homogeneous environment. Its spin angular momentum ($\hbar\sqrt{S(S+1)}$) is calculated using $S = |m_s|$ with $m_s = \pm 1/2$. The value thus obtained is $\sqrt{3}\hbar/2$. Under the influence of a magnetic field $\mathbf{B}_0$, the angular momentum is space quantized and has two allowed orientations in the present case[††] [(since $S = 1/2$), as shown in Fig. 13.8(a)]. The respective components along the field direction are $\hbar/2$ and $-\hbar/2$ with the angular momentum being inclined at an angle $\cos^{-1}(1/\sqrt{3})$ with the field direction.

Further, the magnetic moment interacts with the magnetic field, affecting a change in the electron energy. The electron energy in the presence of the magnetic field can be written as

$$U = U_0 + \Delta U \tag{13.97}$$

---

† In the case where the total spin quantum number $S = 0$ and the orbital motion contributes to the magnetic moment ($L \neq 0$), then $\gamma_e = -\, e/2m$.

†† For magnetic field along the z-direction, the angular momentum $\hbar S$ has $(2S + 1)$ orientations corresponding to S_z values: $-S, -(S-1), \ldots, 0, \ldots, (S-1), S$.

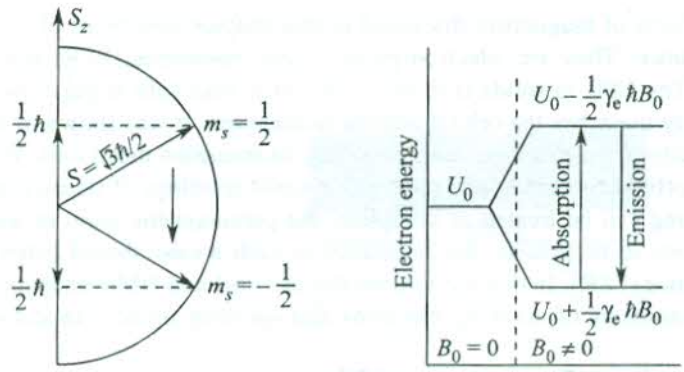


FIG. 13.8 (a) Two possible orientations of the spin angular momentum of an electron relative to the direction of a magnetic field. (b) Splitting of a one-electron energy level in a magnetic field. The two levels differ in energy by $\gamma_e \hbar B_0$. For a photon of this energy, resonance absorption occurs.

where U_0 is the electron energy in the absence of the field and

$$\Delta U = -\boldsymbol{\mu} \cdot \mathbf{B}_0 = -\mu_z B_0 \quad \text{(if the field is along the } z\text{-direction)}. \tag{13.98}$$

Therefore,

$$U = U_0 \pm \frac{1}{2}\gamma_e \hbar B_0 \tag{13.99}$$

$$\left(\text{since } \mu_z = \pm \frac{1}{2}\gamma_e \hbar, \quad \text{for } S_z = \pm \frac{1}{2}\right)$$

Thus an electron level with energy U_0 splits in the presence of a magnetic field B_0 into two levels [see Fig. 13.8(b)] with energies $\left(U_0 - \frac{1}{2}\gamma_e \hbar B_0\right)$ and $\left(U_0 + \frac{1}{2}\gamma_e \hbar B_0\right)$. This is essentially the manifestation of the Zeeman effect, removing the degeneracy of states with $m_s = \pm 1/2$. The two levels are separated in energy by the amount $\gamma_e \hbar B_0$. The frequency of transition between these levels, $\gamma_e B_0$, has reference to the Larmor frequency in classical picture. So, if rf photons of frequency, $\gamma_e B_0$, were fed into the system, a resonant absorption of the radiation would take place. As already stated, in actual experiment, the frequency of the rf radiation is kept fixed at certain value ω_0 and the magnetic field searched for a value B_0, satisfying the condition

$$\hbar\omega_0 = \gamma_e \hbar B_0 \tag{13.100}$$

or

$$\omega_0 = \gamma_e B_0$$

The experimental assembly for a magnetic resonance measurement is given in Fig. 13.9.

For ESR, where the electron spin is involved,

$$\gamma_e = \frac{2\mu_B}{\hbar}, \quad \text{taking} \quad g_0 = 2$$

$$= 17.587 \times 10^{10}\ \text{s}^{-1}\ \text{T}^{-1}$$

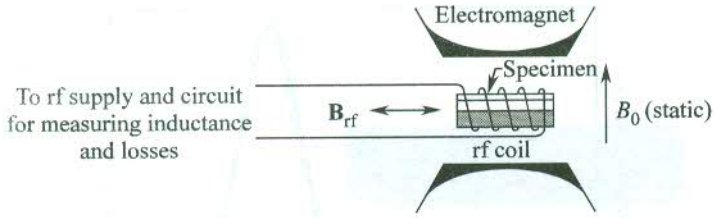


FIG. 13.9 Experimental set-up for a magnetic resonance measurement.

Using this value of γ_e, the frequency of rf radiation that would produce resonance can be expressed as

$$\nu_0(\text{GHz}) = 28.0\, B_0(\text{T}) \tag{13.101}$$

For an induction field of 10 T, $\nu_0 = 28$ GHz (wavelength ~ 1 cm). These frequencies are covered by the microwave region of the e.m. radiation. In experiment, a klystron is used as the source of microwaves. These days frequencies up to 50 GHz and fields up to 1.5 T (15 kilogauss) are very common. The use of a high frequency-high field combination not only provides a higher resolution but also improves the absorption probability.

Let N_1 and N_2 denote the spin population in the lower and upper levels, respectively, in the present system with $m_s = \pm\, 1/2$. Their ratio at the thermal equilibrium is written as

$$\left(\frac{N_2}{N_1}\right)_0 = \exp\left[-\frac{\gamma_e \hbar B_0}{k_B T}\right] \tag{13.102}$$

Generally, for temperatures at which the experiments are performed, $\gamma_e \hbar B_0/k_B << T$ (for $B_0 = 1.0$ T, it is ~1 K). In this limit it is easy to show that the population difference $(N_1 - N_2)$ at thermal equilibrium is given by

$$(\Delta N)_0 \cong N\left(\frac{\gamma_e \hbar B_0}{2 k_B T}\right) = N\left(\frac{h\nu_0}{2 k_B T}\right) \tag{13.103}$$

where N is the total spin density $(N_1 + N_2)$. The $(\Delta N)_0$ is usually very small compared to N. Though relation (13.103) shows that $(\Delta N)_0$ is proportional to ν_0, the experimental setting is often manipulated to make the ESR sensitivity closely proportional to ν_0^2. Also, for a large absorption probability $(\Delta N)_0$ should be reasonably large. These considerations indicate the desirability of performing the experiment at high fields and low temperature. Accordingly, the use of a high frequency high field combination is very much welcome.

The ESR was observed for the first time in $MnSO_4$ by Zavoisky* using a radio frequency in the MHz range and a field less than 0.01 T. The resonance is attributed to the unpaired spins of Mn^{++} ($S = 5/2$, $L = 0$). The resonance could barely be detected because of poor resolution. Zavoisky was later able to get a well resolved resonance (see Fig. 13.10) by increasing ν_0 and B_0.

Substances containing paramagnetic ions show EPR. The EPR is also observed in a few compounds with an even number of electrons including molecular oxygen and organic biradicals.

* E. Zavoisky, *J. Phys.* (USSR), **9**, 211, 245 (1945).

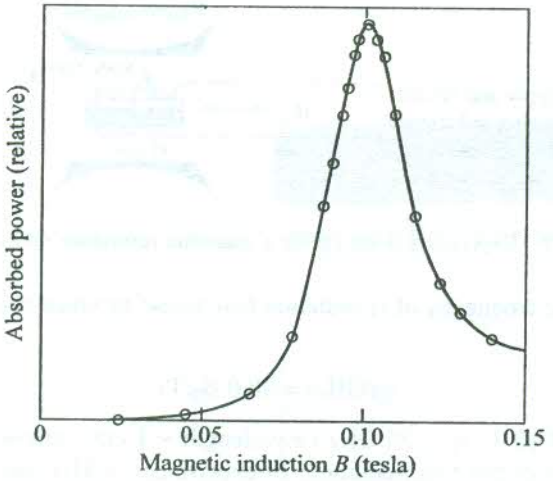


FIG. 13.10 The ESR of Mn^{2+} ions in $MnSO_4$ at room temperature (for $\nu_0 = 2.75$ GHz). The Mn^{2+} ion has five unpaired electrons in the 3d shell ($L = 0$, $S = 5/2$). (After E. Zavoisky, *J. Phys.* (USSR), **10**, 197 (1946)).

The fine structure of the resonance line is especially useful in studying the electronic structure of point defects such as *F*-centre in alkali halides and the donor impurity in semiconductors.

13.9.2 Nuclear Magnetic Resonance (NMR)

The principle of NMR is the same as that of the ESR with modification in expressions affected on account of replacing the electron spin **S** by the nuclear spin **I**. The nuclear energy under the influence of a magnetic field $\mathbf{B}_0$ acting along the *z*-direction changes by an amount expressible as

$$U = -\gamma_n \hbar \mathbf{I} \cdot \mathbf{B}_0$$

$$= -\gamma_n \hbar B_0 I_Z \qquad (13.104)$$

with $I_z = -I, -(I-1), \ldots 0, \ldots, (I-1), I$.

For the simplest case involving a proton,

$$\gamma_n = \frac{2\mu_p}{\hbar} = 2.675 \times 10^8 \text{ s}^{-1} \text{ T}^{-1}$$

(Using $\mu_p = 1.4106 \times 10^{-26}$ J T^{-1})

This value of γ gives the resonance frequencies in the range of MHz:

$$\nu_0(\text{MHz}) = 42.58\, B_0(T) \qquad (13.105)$$

So far as solids are concerned there are three most important aspects of studies for which the NMR technique has been widely exploited: (i) Structure, (ii) Molecular rotation, and (iii) Pauli susceptibility (appearing as the Knight shift in NMR).

Analysis in respect of the first two cases is carried out on the basis of general features of NMR such as the structure, width and shape of the resonance line. Details of the actual analysis may be dispensed with in a book like this that emphasizes the basic principles. Nevertheless, a suitable

account of relaxation processes that affect the features of magnetic resonance in a significant way is given at a later stage in this chapter.

Estimation of the Pauli susceptibility of metals by NMR, on the other hand, is especially significant because there is hardly any other experimental method that meets this purpose. In practice, a measurement of the bulk magnetic moment induced by a magnetic field in metals gives the sum of the Pauli paramagnetic susceptibility, the Landau diamagnetic susceptibility (see Section 13.6) and the Larmor diamagnetic susceptibility. Generally, it is not easy to isolate any of these contributions experimentally. The NMR study, nevertheless, provides an indirect method of estimating the Pauli susceptibility. The basis of measurement lies in the fact that the magnetic moments of the ionic nuclei couple more strongly to the spin magnetic moments of the conduction electrons than to fields produced by the translational motion of electrons.

The magnetic field acting directly on the nucleus determines the frequency of resonance. The field in non-paramagnetic materials differs by a small diamagnetic correction from the actually applied field. At a fixed frequency, the resonance thus occurs for a value of the magnetic field that is shifted from what would be expected in the absence of the diamagnetic correction. This shift is known as the *chemical shift*. There is, however, a stronger correction to the field in metals caused by the magnetic field of the spin magnetic moment of the conduction electrons. The same external magnetic field, in which the nuclei precess, also creates imbalance in the electronic spin populations in the parallel and antiparallel alignments with the field (see Fig. 13.5). We saw in Section 13.6 that the imbalance is responsible for the net magnetic moment and, therefore, for the Pauli paramagnetism of the conduction electron gas. The net electronic magnetic moment has its own field leading to a modified field at the nucleus. At a fixed frequency, the resonance of a spin is observed at a slightly different magnetic field in a metal than in a non-paramagnetic solid containing the metallic element. The observed shift, known as the *Knight shift*, is proportional to the Pauli susceptibility. But this estimate is not precise because the Knight shift is also proportional to the square of the magnitude of the conduction electron wavefunction at the nucleus. The actual Pauli susceptibility is obtained by incorporating this correction which is calculated with the knowledge of wavefunctions that are generally derived from the atomic s-shells.

13.9.3 Spin Relaxation

Consider again the simplest spin systems of an unpaired electron and proton to study the effect of spin relaxation on magnetic resonance. We are concerned with a two-level system in the presence of a dc magnetic field. Both upward and downward transitions are induced by the microwave photons at rates proportional to N_1 and N_2, the population of the lower and upper levels respectively. For the absorption of the microwave power, the population difference $\Delta N(=N_1 - N_2)$ must maintain a positive value during the measurement. It is essential that ΔN remains at the equilibrium value (13.103) while the magnetic field is swept near resonance. This is ensured by restricting the microwave power input to a reasonably low level, usually determined by the mechanicsm of spin relaxation because of which the spin system falls back to the lower energy level.

Spin-lattice relaxation

The crystal lattice, usually represented by lattice vibrations, participates in this kind of spin relaxation. It occurs through mainly three types of processes:

(i) **Direct process.** The spin-phonon coupling induces the relaxation with the emission of a low frequency ($\omega_0 = \gamma B_0$) acoustic phonon. The relaxation time τ_1 that represents the process varies with temperature T as $1/T$.

(ii) **Raman process.** In certain solids this process competes with the direct process effectively at higher temperatures. A phonon of frequency ω, already present in the system at these temperatures, is inelastically scattered to cause relaxation with the emission of a phonon of frequency $(\omega + \omega_0)$. In this case, $\tau_1 \propto T^{-7}$ or T^{-9}.

(iii) **Orbach process.** It is different from the Raman process in the sense that it is a two-stage phonon process in which the system at higher energy first goes to a further higher state by the absorption of a phonon of energy Δ present in the system, and then relaxes to the lower energy state with the emission of a phonon of energy $(\Delta + \hbar\omega_0)$. The temperature dependence of the relaxation time is expressed as $\tau_1 \propto \exp\left(\frac{\Delta}{k_B T}\right)$.

In the absence of any spin relaxation, the equation of motion describing the precession of angular momentum $\hbar\mathbf{J}$ (see Section 13.1) is

$$\hbar \frac{d\mathbf{J}}{dt} = \boldsymbol{\mu} \times \mathbf{B}_0 \tag{13.106}$$

with

$$\mathbf{J} = \mathbf{S} \text{ for a pure ESR}$$
$$= \mathbf{I} \text{ for NMR}$$

Relation (13.106) can be rewritten as

$$\frac{d\boldsymbol{\mu}}{dt} = \gamma\boldsymbol{\mu} \times \mathbf{B}_0 \qquad (\text{since } \boldsymbol{\mu} = \gamma\hbar\mathbf{J}) \tag{13.107}$$

and

$$\frac{d\mathbf{M}}{dt} = \gamma\mathbf{M} \times \mathbf{B}_0 \tag{13.108}$$

where M is the magnetization given by $(N_1 - N_2)\mu$ or $(\Delta N)\mu$, with N_1, N_2 defined as the spin population per unit volume of the lower and upper energy levels, respectively, in the two-level spin system.

When the field is applied along the z-direction, the magnetization at thermal equilibrium will be along the z-direction, i.e.

$$M_z = (\Delta N)_0 \mu \qquad \text{with} \qquad M_x = 0 \qquad \text{and} \qquad M_y = 0$$

In the presence of the spin-lattice relaxation the magnetization M_z at any instant tends to approach the equilibrium value M_0 at the rate given by

$$\frac{dM_z}{dt} = \frac{M_0 - M_z}{\tau_1} \tag{13.109}$$

Integrating (13.109), we get

$$M_z(t) = M_0 \left(1 - \exp\left(-\frac{t}{\tau_1}\right)\right) \tag{13.110}$$

Accordingly, the spin-lattice relaxation modifies the equation of motion (13.108) to the form

$$\frac{dM_z}{dt} = \gamma(\mathbf{M} \times \mathbf{B}_0)_z + \frac{M_0 - M_z}{\tau_1} \tag{13.111}$$

The above equation states that besides precessing around the magnetic field, M_z tends to relax to the equilibrium value M_0.

Spin-spin relaxation

In a static magnetic field $\mathbf{B}_0$ along the z-direction, if the transverse components of magnetization M_x and M_y are not zero to start with, they decay to zero since their values at thermal equilibrium are zero. The decay takes place through another type of relaxation known as the *spin-spin relaxation*. This occurs owing to the influence (as a result of the local field contribution) of one magnetic atom on others. The decay rate of the transverse magnetization is set by the spin-spin relaxation time τ_2. The factor $1/\tau_2$ gives the rate constant for the loss of phase coherence of a set of spins with e.m. field applied in the x- or y-direction. In this sense, τ_2 is also identified as the dephasing time. The loss of phase coherence of a set of spins causes an increased fluctuation in the magnetic field exerted by spins on each other. As a consequence to this, the spins show resonance over a small range of the magnetic field values, providing thus a finite width to the resonance line.

Equations of motion for the transverse components of magnetization are written as

$$\frac{d\mathbf{M}_x}{dt} = \gamma(\mathbf{M} \times \mathbf{B}_0)_x - \frac{\mathbf{M}_x}{\tau_2} \tag{13.112}$$

$$\frac{d\mathbf{M}_y}{dt} = \gamma(\mathbf{M} \times \mathbf{B}_0)_y - \frac{\mathbf{M}_y}{\tau_2} \tag{13.113}$$

The relations (13.111) to (13.113) are called *Bloch equations*. They are solved to derive several important properties of a magnetic system from its magnetic resonance measurement.

The net effect of relaxation is demonstrated by the combined relaxation time τ given by

$$\frac{1}{\tau} = \frac{1}{\tau_1} + \frac{1}{\tau_2} \tag{13.114}$$

τ is effectively determined by the shorter of τ_1 and τ_2. It is usually deduced from the width of the resonance line.

13.9.4 Line Width and Line Shape

For the sake of completeness, let us first have a cursory look at the causes of the line width. The main sources are identified as follows:

(i) **Spontaneous emission.** It is responsible for the natural line width which is negligibly small compared to the width due to other sources.

(ii) **Spin-lattice relaxation.** It is represented by the relaxation time τ_1, as defined in Section 13.9.3.

(iii) **Spin-spin relaxation (Magnetic dipole-dipole interaction).** It is represented by the relaxation time τ_2, as described in Section 13.9.3. It is absent in liquids and gases because of rapid rotation of molecules which effectively averages out the magnetic fields.

(iv) **Inhomogeneous magnetic field (an experimental limitation).** The static magnetic field may vary over the volume of the sample. There is a contribution to the overall line broadening because of the superposition of lines from the spins in the different parts of the sample.

(v) **Electric quadrupole coupling.** It occurs in the NMR of nuclei with $I > 1/2$, where the electric quadrupole moments interact with the electric field gradients. The contribution in this case is relatively much small.

At first sight, the gain from a magnetic resonance measurement is appreciated in the form of knowledge about the effective g-factor that is crucial to the study of magnetic properties of a system. In fact, much more information is derived from this measurement. For example, the width and shape of the resonance line furnish information about the environment of spins. At this stage it must be impressed on the reader's mind that precise information about the symmetry around a paramagnetic centre, especially in crystalline solids, is determined from the fine structure and hyperfine structure of the resonance line. Being topics of advanced level, they are not discussed here.

In a magnetic resonance measurement, the absorbed rf power P can be plotted as a function of the magnetic field B. But a greater sensitivity is achieved with a phase-sensitive detection method, which gives dP/dB versus B plot. Bloch* considered the upward and downward stimulated transitions for a spin system where B/ν_0 is swept at a linear rate through the resonance condition in an NMR measurement. But the theory applies equally well to EPR. Bloch has shown that when the deviation of the value of ΔN from $(\Delta N)_0$ (the value at thermal equilibrium) is negligibly small throughout the sweep of B/ν_0, the resonance line has a Lorentzian shape expressed as

$$P = \frac{P_{max}}{1 + \dfrac{(B - B_0)^2}{\xi^2}} \tag{13.115}$$

where B_0 is the value at resonance and the parameter ξ is determined from the measure of the separation $\Delta B_{1/2}$ between the half-power points. Using simple calculus, we can show that

$$\Delta B_{1/2} = 2\xi \tag{13.116}$$

and

$$\Delta B_{pp} = \frac{2\xi}{\sqrt{3}} = \frac{\Delta B_{1/2}}{\sqrt{3}} \tag{13.117}$$

where ΔB_{pp} denotes the separation between the positive and negative peaks in the dP/dB plot.

The P and dP/dB plots for the EPR of DPPH (solid 1,1-diphenyl-2-picryl-hydrazyl) are shown in Fig. 13.11. The DPPH line is often taken as standard for the calibration of a spectrometer.

---

* F. Bloch, *Phys. Rev.*, **70**, 460 (1946).

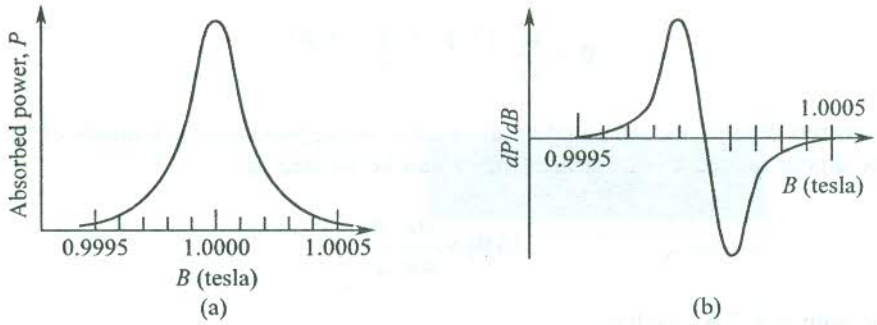


FIG. 13.11 (a) EPR absorption line for DPPH at room temperature (ν_0 = 28 GHz). (b) The corresponding plot of the first derivative of the line.

The line width in units of frequency is given by the reciprocal of the net relaxation time τ which is estimated from

$$\tau \equiv \frac{1.1 \times 10^{-11}}{g\xi} \text{ s} \tag{13.118}$$

when ξ is expressed in units of tesla.

In case the value of ΔN happens to be much different from $(\Delta N)_0$, the resonance line assumes a Gaussian shape

$$P = P_{\max} \exp\left[-\frac{(B - B_0)^2}{\xi^2 \ln 2}\right] \tag{13.119}$$

giving again $\Delta B_{1/2} = 2\xi$ but a larger ΔB_{pp} and a relationship between ξ and τ that differs from (13.118).

The NMR line behaves differently by the way of its width being practically independent of τ_1, even though τ_1 continues to determine the rate of spin transitions induced without creating saturation. In a nuclear spin system, τ_1 can become appreciably large at low temperatures. Pound (1951) reported the value of τ_1 as 300 s at room temperature for ^{7}Li in LiF crystal. Harping on the consequences of this slow response, he along with Purcell* founded the concept of 'population inversion' in a nuclear spin system where the spin population density is greater in a higher energy state than that in the ground state. The spin populations being given by the Boltzmann distribution at thermal equilibrium, the inversion of population must be viewed as accompanied by the cooling of the spin system. Thus, for a spin system at low temperatures, it is very much likely that a negative value of temperature would have to be assigned to the system when in an excited state at thermal equilibrium under the condition of population inversion. The concept of population inversion and negative temperature was exploited later in the same decade in the discovery of masers and lasers.

The magnetic dipolar interaction is usually the main contributor (through a small τ_2) to the large width of an NMR line in a solid. The magnetic field caused by a magnetic dipole† of strength μ at a distance r in a rigid lattice is

---

* E.M. Purcell and R.V. Pound, *Phys. Rev.*, **81**, 279 (1951).

† Relation (13.120) is comparable with the corresponding relation (10.2) for an electric dipole.

$$\mathbf{B} = \frac{\mu_0}{4\pi} \frac{[3(\boldsymbol{\mu} \cdot \mathbf{r})\,\mathbf{r} - r^2\boldsymbol{\mu}]}{r^5} \tag{13.120}$$

If two nearest dipoles are separated by distance a, the approximate magnitude of the magnetic field at one dipole caused by the nearest other can be written as

$$|\Delta \mathbf{B}| \approx \frac{\mu_0}{4\pi} \frac{\mu}{a^3} \tag{13.121}$$

For protons with $a = 2$ Å, we have

$$|\Delta \mathbf{B}| \approx 10^{-7} \times \frac{1.4106 \times 10^{-26}}{8 \times 10^{-30}} \text{ T}$$

$$\approx 2 \times 10^{-4} \text{ T}$$

The observed resonance line reflects the effect of the vector sum of fields mainly caused by all the nearest magnetic dipoles. Relation (13.121) essentially gives the line broadening contributed by the magnetic dipole-dipole interaction.

SUMMARY

1. The (Larmor) diamagnetic susceptibility of N atoms of atomic number Z is given by $\chi = -\frac{\mu_0 N Z e^2}{6m} \langle r^2 \rangle$, where $\langle r^2 \rangle$ is the mean square atomic radius. This relation is known as the Langevin's relation for the diamagnetic susceptibility.
2. According to the classical theory, the relation for paramagnetic susceptibility of atoms with a permanent magnetic moment μ is given by

   $$\chi = \frac{\mu_0 \mu^2 N}{3 k_B T} \quad \text{for} \quad \mu B_0 << k_B T \quad \text{(Curie–Langevin expression)}$$

   where N denotes the number of atoms per unit volume.
3. The magnetic moment of an atom is

   $$\boldsymbol{\mu} = -\mu_B (\mathbf{L} + g_0 \mathbf{S}) = -g \mu_B \mathbf{J}$$

4. When an atom or ion is in its ground state, its electrons in the same shell acquire the maximum value of S in accordance with the Pauli principle and maximum L as may be consistent with the maximum S. The value of J is $(L + S)$ for a shell more than half-filled and $(L - S)$ for a shell less than half-filled.
5. $J = 0$ for completely-filled shells and shells that are one electron short of being half-filled.

6. The magnetic Helmholtz energy F is defined as

$$\exp(-F/k_B T) = \sum_n \exp[-E_n(B_0)/k_B T]$$

where E_n is the electronic energy in the nth state in the presence of field B_0.

7. The magnetic susceptibility of a system is defined in terms of the Helmholtz free energy F as

$$\chi = -\mu_0 N \frac{\partial^2 F}{\partial B_0^2}$$

where N is the number of atoms/ions per unit volume.

8. According to quantum mechanics, the change in the energy of an ion in its nth state when placed in a magnetic field B_0 is

$$\Delta E_n = \mu_B \mathbf{B}_0 \cdot \langle \phi_n | \mathbf{L} + g_0 \mathbf{S} | \phi_n \rangle + \sum_{n' \neq n} \frac{|\langle \phi_n | \mu_B \mathbf{B}_0 \cdot (\mathbf{L} + g_0 \mathbf{S}) \phi_{n'} \rangle|^2}{E_n - E_{n'}}$$

$$+ \frac{e^2}{8m} B_0^2 \langle \phi_n | \sum_i (x_i^2 + y_i^2) | \phi_n \rangle$$

where the first term gives rise to paramagnetism because of the permanent magnetic moment; the second term leads to weak paramagnetism including van Vleck paramagnetism which is temperature independent (for a partially-filled shell with $J = 0$) and the last term refers to Larmor diamagnetism.

9. According to the quantum theory, the paramagnetic susceptibility of ions with $J \neq 0$ (with permanent ionic magnetic moment μ) is given by

$\chi = \dfrac{\mu_0 N p^2 \mu_B^2}{3 k_B T}$ in the limit $\mu B_0 << k_B T$, where p is the 'effective Bohr magneton number' defined as

$$p = g[J(J+1)]^{1/2}.$$

and N is the number of ions per unit volume.

10. The paramagnetic susceptibility of conduction electrons (metals) is independent of temperature for $k_B T << \varepsilon_F$ and expressed as

$$\chi = \frac{3 N \mu_0 \mu_B^2}{2 \varepsilon_F} = \frac{3 N \mu_0 \mu_B^2}{2 k_B T_F}$$

where N is the total number of electrons.

11. The frequency of magnetic resonance is

$$\omega_0 = \gamma B_0, \quad \text{where } \gamma = \frac{\mu}{\hbar J} \text{ (the magnetogyric ratio)}$$

12. The Bloch equations are

$$\frac{d\mathbf{M}_x}{dt} = \gamma(\mathbf{M} \times \mathbf{B})_x - \frac{\mathbf{M}_x}{\tau_2}$$

$$\frac{d\mathbf{M}_y}{dt} = \gamma(\mathbf{M} \times \mathbf{B})_y - \frac{\mathbf{M}_y}{\tau_2}$$

$$\frac{d\mathbf{M}_z}{dt} = \gamma(\mathbf{M} \times \mathbf{B})_z + \frac{\mathbf{M}_0 - \mathbf{M}_z}{\tau_1}$$

13. The shapes of the magnetic resonance line:

$$\text{Lorentzian: } P = \frac{P_{\max}}{1 + \dfrac{(B - B_0)^2}{\xi^2}}$$

$$\text{Gaussian: } P = P_{\max} \exp\left[-\frac{(B - B_0)^2}{\xi^2 \ln 2}\right]$$

$$\text{with } \Delta B_{1/2} = 2\xi.$$

14. Line width due to magnetic dipole-dipole interaction in a rigid lattice is given by

$$|\Delta \mathbf{B}| \simeq \frac{\mu_0}{4\pi} \frac{\mu}{a^3}$$

where a is the separation between the two nearest dipoles.

PROBLEMS

13.1 The wavefunction of the hydrogen atom in its ground state is

$$\psi = \frac{1}{(\pi a_0^2)^{1/2}} \exp\left(-\frac{r}{a_0}\right)$$

Assuming the charge density to be given by

$$\rho(x, y, z) = e|\psi|^2$$

show that

$$\langle r^2 \rangle = 3a_0^2$$

and calculate the molar diamagnetic susceptibility of atomic hydrogen.

13.2 Ampere defined classically the magnetic moment of electron owing to its orbital motion as the average over the orbit of $-e/2(\mathbf{r} \times \mathbf{v})$.

Prove that our definition, $\mu = \partial E/\partial B_0$, reduces to this form by showing from (13.47) that

$$\boldsymbol{\mu} = -\frac{e}{2m} \sum_i \mathbf{r}_i \times \left(\mathbf{p}_i - \frac{e}{2} \mathbf{r}_i \times \mathbf{B}_0 \right)$$

and

$$\mathbf{v}_i = \frac{\partial \mathbf{H}}{\partial \mathbf{p}_i} = \frac{1}{m} \left(\mathbf{p}_i - \frac{e}{2} \mathbf{r}_i \times \mathbf{B}_0 \right)$$

13.3 (a) Show that the following formulae summarize the Hund's rules for a shell of angular momentum l containing n electrons:

$$S = 1/2 \; [(2l + 1) - |2l + 1 - n|]$$

$$L = S \; [2l + 1 - n]$$

$$J = |2l - n| \, S$$

(b) For a given LS-multiplet, verify that

$$(2L + 1)(2S + 1) = \sum_{|L-S|}^{|L+S|} (2J + 1)$$

(c) Find the ground state of

(i) Eu^{2+} in the configuration $4f^7 5s^2 p^6$

(ii) Dy^{3+} in the configuration $4f^9 5s^2 p^6$

(iii) Tm^{3+} in the configuration $4f^{12} 5s^2 p^6$

13.4 A magnetic field B_0, when applied on an atom with a spherically symmetric charge distribution $\rho(r)$, induces a diamagnetic current. Show that the magnetic field produced by the diamagnetic current at the nucleus is

$$\Delta B_0 = -\left(\frac{eB_0}{3m} \right) V_0$$

where V_0 is the electrostatic potential at the nucleus, given by

$$V_0 = \int_0^\infty \frac{\rho(r)}{r} \, 4\pi r^2 \, dr$$

13.5 When a system of electron spins is placed in a magnetic field of 2 Wb/m^2 at a certain temperature, the number of the spins parallel to field lines is twice the number of the spins antiparallel to the field. Determine the temperature.

13.6 The most important contribution to the paramagnetism of copper sulphate comes from the copper ions which have spin 1/2 and may be treated as non-interacting. Show that the magnetization in field B_0 is given by

$$M = N\mu_B = \tanh\left(\frac{\mu_B B_0}{k_B T}\right)$$

where N is the number of ions per unit volume.

13.7 In benzene, the carbon atoms form a regular hexagon of side 1 Å. One outer electron from each carbon atom has a wavefunction that extends round the whole ring of atoms (the other three are in sp^2 atomic orbitals). Make a rough estimate of the contribution of these electrons to the diamagnetic susceptibility of liquid benzene (density 0.88 g cm^{-3}; molecular wt. = 76).

13.8 An ion has a partially-filled shell of angular momentum J with Z electrons in the completely filled shells. Show that the ratio of the Curie law paramagnetic susceptibility to the Larmor diamagnetic susceptibility is

$$\frac{\chi_{\text{para}}}{\chi_{\text{dia}}} = -\frac{2J(J+1)}{Zk_B T}\frac{\hbar^2}{m\langle r^2\rangle}$$

13.9 A magnetic field is applied to a salt containing Cu^{2+} ions. Cu^{2+} has 9 electrons in the 3d shell. What magnetic field must be applied to the salt when at 1 K, so that 99 per cent of the ions are in the lowest energy state?

13.10 For quantized states we cannot use the Langevin function $L(x)$ given by (13.22) to compute susceptibilities, but instead, the Brillouin function $B_J(x)$ as given by (13.59) must be used. Show that $B_J(x) \to L(x)$ as $J \to \infty$, with $x = \mu B_0/k_B T$.

13.11 For temperatures very small compared to the Fermi temperature, show that the temperature dependent correction to the Pauli susceptibility is

$$\chi\ (T) = \chi(0)\left(1 - \frac{\pi^2}{6}(k_B T)^2\left[\left(\frac{D'}{D}\right)^2 - \frac{D''}{D}\right]\right)$$

where D, D' and D'' are the density of levels and its derivatives at the Fermi energy. Show that for electrons it reduces to

$$\chi(T) = \chi(0)\left(1 - \frac{\pi^2}{12}\left(\frac{k_B T}{\varepsilon_F}\right)^2\right)$$

13.12 (a) Show that for a crystal containing N paramagnetic ions with $S = 1/2$ and $g_0 = 2$, the spin entropy can be expressed by

$$S = \frac{N\Delta}{2T}\tanh\left(\frac{\Delta}{2k_B T}\right) + Nk_B \ln(2)\cosh\left(\frac{\Delta}{2k_B T}\right)$$

with $\Delta = 2\mu_B B_0$

(b) For cooling by adiabatic demagnetization, why is it not possible to start with the nuclear demagnetization, without going through the process of electron-spin demagnetization?

SUGGESTED FURTHER READING

Abragam, A., *Nuclear Magnetism* (Oxford, 1961).

Ashcroft, N.W. and N.D., Mermin, *Solid State Physics* (Saunders College, 1988).

Blakemore, J.S., *Solid State Physics* (W.B. Saunders, 1974).

Chakravarty, A.S., *Introduction to Magnetic Properties of Solids* (John Wiley, 1980).

Hudson, R.D., *Principles and Applications of Magnetic Cooling* (Elsevier, 1972).

Kittel, C., *Introduction to Solid State Physics*, 7th ed. (John Wiley, 1996).

Slichter, C.P., *Principles of Magnetic Resonance*, 3rd ed. (Springer, 1990).

van Vleck, J.H., *The Theory of Electric and Magnetic Susceptibilities* (Oxford, 1932).

White, R.M., *Quantum Theory of Magnetism* (Springer, 1982).

Chapter 14

Ferromagnetism, Antiferromagnetism and Ferrimagnetism

A system of non-interacting magnetic ions is expected to be thermally disordered at any temperature in the absence of an applied magnetic field. The vector magnetic moment of each ion would then average to zero as confirmed by (13.20). Nevertheless, there are plenty of solids in which individual magnetic ions have nonvanishing average vector moments below a critical temperature T_c. This points towards a possible correlation or interaction among magnetic ions, leading to the ordering of magnetic moments as revealed by experiments. These solids are regarded as magnetically ordered.

The magnetic ordering may or may not lead to a net magnetization for the solid as a whole. There arise three distinct cases of ordering:

1. The individual localized magnetic moments may combine, even in the absence of an applied magnetic field, to produce a macroscopic bulk magnetization reflecting the existence of microscopic ordering. Such a magnetization is known as *spontaneous magnetization* and the solids exhibiting this property are called *ferromagnets*. An array of local magnetic moments with the same magnitude and average direction describes the simplest model of ferromagnets. It is represented by the first of the three arrays in Fig. 14.1(a) for a linear ferromagnet.
2. The individual local moments may add up to zero, resulting in no bulk magnetization. The solids with this type of ordering have been identified as *antiferromagnets*. Two possible arrays of local moments for a linear antiferromagnet are drawn in Fig. 14.1(b).
3. The primitive cell of some solids showing spontaneous magnetization contains several magnetic ions which may necessarily not be identical. Such a possibility makes it desirable to adopt a rather restrictive definition of ferromagnets, according to which all the local moments are required to have a positive component along the direction of spontaneous magnetization. The solids that show spontaneous magnetization but do not satisfy this requirement are described as *ferrimagnets*. We will show later that the spontaneous magnetization that is acquired by ferrimagnets is relatively much small. The possible orderings for a linear ferrimagnet are shown in Fig. 14.1(c).

14.1 WEISS THEORY OF FERROMAGNETISM

Ferromagnetic materials have been known for many years. Of these only five (Fe, Co, Ni, Gd and Dy)

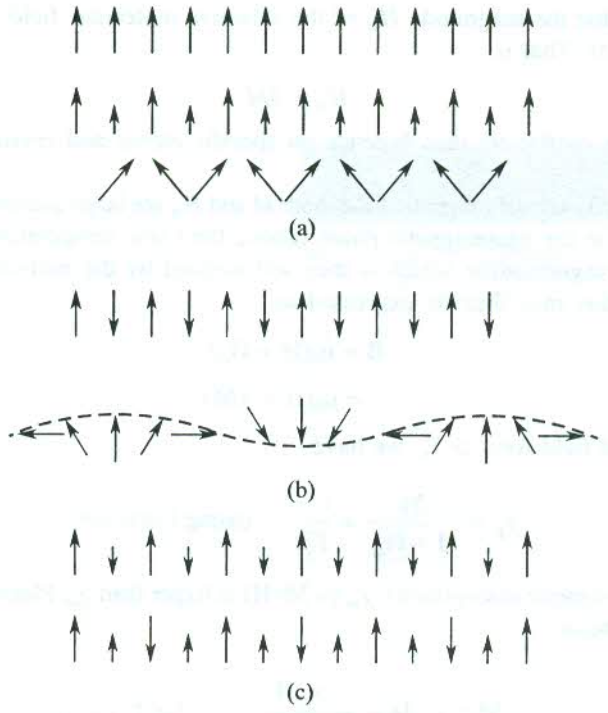


FIG. 14.1 Possible orderings of atomic moments in a linear (a) ferromagnet, (b) antiferromagnet, and (c) ferrimagnet.

are elements (all metals) and the rest are alloys and insulating compounds (see Table 14.1). In similarity with ferroelectrics, ferromagnetic materials exhibit hysteresis in the relationship between magnetic induction B and intensity of the applied magnetic field H. Ferromagnetism is clearly revealed by a macroscopic magnetic field which in one of the simplest approaches is recognized as the mean molecular field. It is apparent from what we said in the beginning that ferromagnetism arises because of the cooperative alignment of permanent atomic dipoles which may be supposed to be caused by the mean molecular field. The concept of mean field was introduced by Pierre Weiss* through his theory that provides a satisfactory explanation to ferromagnetism. The Weiss mean field theory is based on the following two hypotheses.

1. In ferromagnetic solids of macroscopic dimensions, there are a number of small regions, called *domains,* that are spontaneously magnetized.
2. There exists a molecular magnetic field within each domain and the field tends to produce a parallel alignment of the individual localized atomic moments.

While describing the properties of magnetic domains later in this chapter we will find that the explanation to the hysteresis revealed by the B–H curve follows from these hypotheses. It must, however, be remarked here that spontaneous magnetization refers to a single domain whereas the remanent magnetization (for $H = 0$) refers to the specimen as a whole.

* P. Weiss, *J. Phys. Radium*, **4**, 661 (1907).

Weiss observed that the magnitude H_m of the effective molecular field should be proportional to the magnetization M. That is,

$$H_m = \lambda M \tag{14.1}$$

where λ is the Weiss coefficient that depends on specific atoms and crystal structure but not on temperature.

In the absence of an external magnetic field, both M and H_m are large quantities in the ferromagnetic phase of a solid. But in the paramagnetic phase (above the Curie temperature T_c), an external field is needed to initiate magnetization which is then self-assisted by the molecular field it builds. The total magnetic induction may thus be expressed as

$$\begin{aligned} \mathbf{B} &= \mu_0(\mathbf{H} + \mathbf{H}_m) \\ &= \mu_0(\mathbf{H} + \lambda\mathbf{M}) \end{aligned} \tag{14.2}$$

and in a weak applied field for $T > T_c$, we have

$$\chi_p = \frac{\mathbf{M}}{\mathbf{H} + \mathbf{H}_m} = \frac{C}{T} \qquad \text{(using Curie law)} \tag{14.3}$$

The observable magnetic susceptibility χ_m (= $\mathbf{M}/\mathbf{H}$) is larger than χ_p. Placing the value of $\mathbf{H}_m$ from (14.1) in (14.3), we obtain

$$\mathbf{M} = \chi_m \mathbf{H} = \frac{\chi_p \mathbf{H}}{(1 - \lambda\chi_p)} \qquad \text{for } T > T_c \tag{14.4}$$

or

$$\chi_m = \frac{\mathbf{M}}{\mathbf{H}} = \frac{\chi_p}{1 - \lambda\chi_p} = \frac{C}{T - \lambda C} = \frac{C}{T - T_c} \tag{14.5}$$

with

$$T_c = \lambda C \tag{14.6}$$

In view of (14.5), χ_m has a singularity at $T = T_c = \lambda C$. The spontaneous magnetization shows up for $T \leq T_c$, since there can be a finite M for $H = 0$ even when χ_m is infinite. Relation (14.5) states the Curie–Weiss law that is known to explain the observed variation of susceptibility above the Curie temperature T_c (Fig. 14.2) in most of the ferromagnetic solids. Using the value of C from (13.27) in (14.6), the Weiss coefficient can be written as

$$\lambda = \frac{T_c}{C} = \frac{3k_B T_c}{\mu_0 N g^2 S(S+1)\mu_B^2} \tag{14.7}$$

The use of S in place of J in the above relation implies that the contribution of the orbital motion can be treated as negligible, letting the effective Bohr magneton number to be determined entirely by the spin of the unpaired electrons.

Relation (14.7) gives the value of λ for iron as $\sim 10^3$. Using the value of spontaneous magnetization M_s as 1.74×10^6 A m^{-1} (Table 14.1), we obtain $H_m \sim 10^9$ A m^{-1} or the value of magnetic induction $\mu_0 H_m \simeq 10^3$ T. On the other hand, the order of magnitude of the field produced by a magnetic ion regarded as a dipole at a nearest ion at distance a equals $\sim \mu_0\mu_B/4\pi a^3 \simeq 0.1$ T. These small dipolar

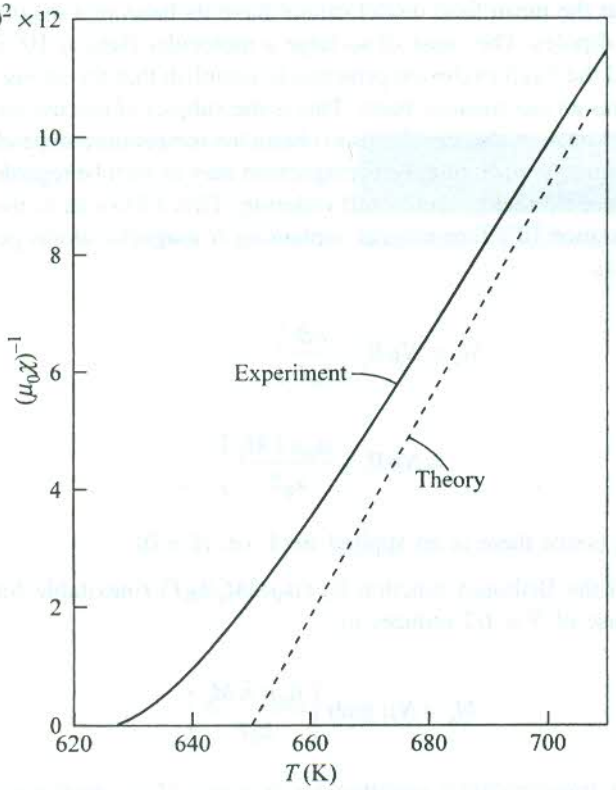


FIG. 14.2 Reciprocal of the susceptibility for nickel as a function of temperature above the Curie temperature on the mean field theory. The experimental curve shows the typical results on nickel. [After P. Weiss and R. Forrer, *Ann. Phys.*, **15**, 153 (1926).]

Table 14.1 Selected ferromagnetic solids† with Curie temperatures T_c and saturation magnetization M_s at $T = 0$ K

Substance	T_c (K)	M_s (gauss)
Fe	1043	1752
Co	1388	1446
Ni	627	510
Gd	293	1980
Dy	85	3000
MnAs	318	870
MnBi	670	675
Au_2MnAl	200	323
Cu_2MnIn	500	613
EuO	77	1910
$GdCl_3$	2.2	550

† *Sources*: F. Keffer, *Handbuch der Physik*, vol. 18, pt. 2 (Springer, New York, 1966); P. Heller, *Rep. Progr. Phys.*, **30** (pt. II), 731 (1967).

fields clearly indicate that the mean field model cannot have its basis in a magnetostatic coupling of the permanent magnetic dipoles. The cause of so large a molecular field as 10^3 T was discovered by Heisenberg* who applied the Pauli exclusion principle to establish that the strong interaction between adjacent magnetic ions has an electrostatic basis. This is the subject of discussion of the next section.

The mean field approximation also enables us to obtain the temperature-dependence of spontaneous magnetization below the Curie temperature. Ferromagnetism may as well be regarded as paramagnetism with a massive molecular field that produces self-ordering. This allows us to use (13.58) to express the spontaneous magnetization of a ferromagnet containing N magnetic atoms per unit volume, each of magnetic moment μ, as

$$M_s = N\mu \mathrm{B}_J\left(\frac{\mu B}{k_B T}\right)$$

$$= N\mu \mathrm{B}_J\left(\frac{\mu_0 \mu \lambda M_s}{k_B T}\right) \tag{14.8}$$

with $B = \mu_0 H_m = \mu_0 \lambda M_s$ (since there is no applied field, i.e. $H = 0$) (14.9)

Replacing J by S in the Brillouin function B_J ($\mu_0 \mu \lambda M_s / k_B T$) (inevitable for transition metals), relation (14.8) for the case of $S = 1/2$ reduces to

$$M_s = N\mu \tanh\left(\frac{\mu_0 \mu \lambda M_s}{k_B T}\right) \tag{14.10}$$

The solution to this transcendental equation for nonzero M_s is determined graphically as in Fig. 14.3 to give a curve for $M_s(T)$ which falls from a saturated value at $T = 0$ to zero at $T = T_c$ (see Fig. 14.4).

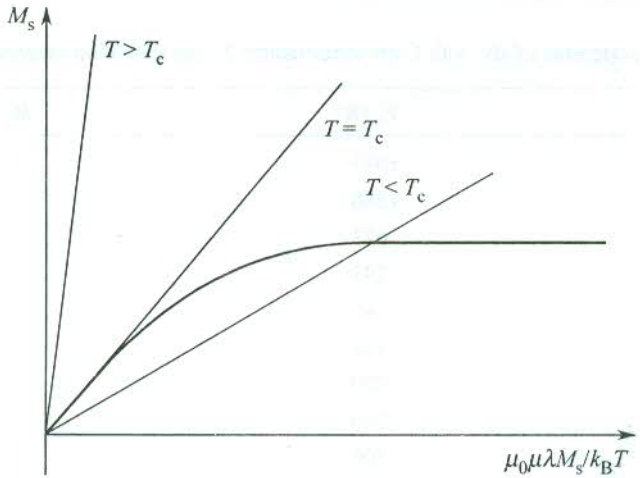


FIG. 14.3 Graphical solution of equation (14.10). For $T > T_c$, there is only one solution at $M_s = 0$. For temperatures below T_c, the curves have a second intersection that gives the other solution.

* W. Heisenberg, *Z. Physik*, **49**, 619 (1928).

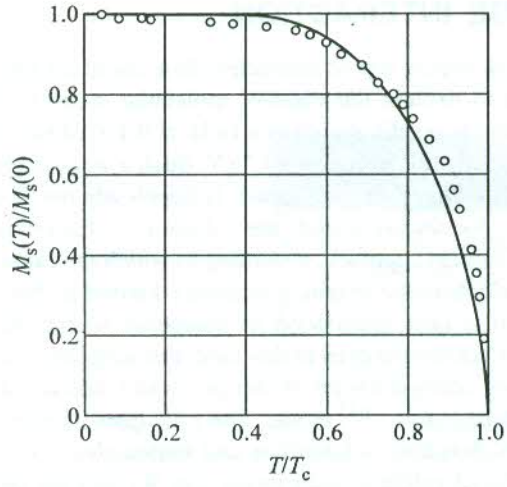


FIG. 14.4 Saturation magnetization of nickel as a function of temperature for $S = 1/2$ on the molecular field model. Open circles denote the experimental points. [After P. Weiss and R. Forrer, *Ann. Phys.*, **15**, 153(1926).]

The energy per unit volume U is written as

$$U = -M_s H_m = -\lambda M_s^2 \tag{14.11}$$

which gives the heat capacity C_V as

$$C_V = -2\lambda M_s \left(\frac{\partial M_s}{\partial T}\right) \tag{14.12}$$

According to (14.12), C_V rises to a maximum at T_c and then falls abruptly to zero as shown in Fig. 14.5. The mean field theory thus predicts an anomaly in the heat capacity at T_c which is confirmed by experiments. The transition from the paramagnetic to the ferromagnetic phase is typically a second-order type as is demonstrated by the continuous fall of spontaneous magnetization, being treated as an order parameter (see Fig. 14.4).

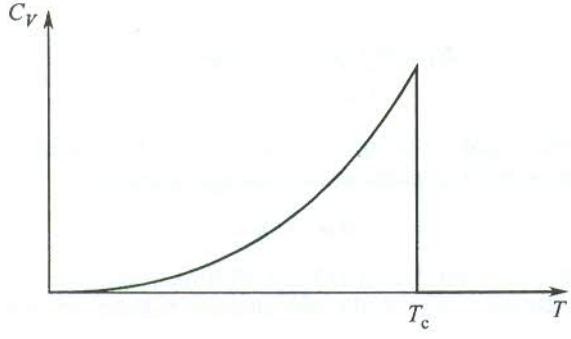


FIG. 14.5 Heat capacity anomaly as predicted on mean field theory.

14.2 THE EXCHANGE INTERACTION

The inadequacy of the magnetic dipole-dipole interaction to account for ferromagnetism is proved by the failure of the interaction to explain the massive molecular mean fields and fairly high Curie temperatures. In iron, an atomic magnetic dipole in a field of 0.1 T caused by the nearest dipole (see Section 14.2) has a magnetic interaction energy ~ 10^{-4} eV. Such a weak magnetic coupling is expected to break down above $(1.6 \times 10^{-23}/k_B)$ K $\simeq$ 1 K, which is hopelessly much smaller than the observed Curie temperature (1043 K). As already stated, the solution to this puzzle has been provided by Heisenberg in a quantum mechanical approach, according to which the cause of the magnetic ordering lies in an interaction of the electrostatic origin, commonly known as the *exchange interaction.*

The exchange interaction is best understood in insulators where the magnetic ions are well separated, though the theory is not simple even in this case. For simplicity, we consider the interaction within a hydrogen molecule regarding it a solid of two atoms and thus avoiding complications arising in a real solid containing approximately 10^{23} atoms. The principles involved are then generalized for application to real solids. It is, however, a laborious and intricate exercise to carry over these ideas to metals because of the required subtle modifications. Let $\mathbf{R}_A$ and $\mathbf{R}_B$ denote the position vectors of the two protons A and B, respectively, and $\mathbf{r}_1$ and $\mathbf{r}_2$ define locations of the two electrons labelled as 1 and 2, respectively. The extent to which the antiparallel ($S = 0$) spin arrangement is more favourable than the parallel is determined by the singlet-triplet energy splitting. We now proceed below to obtain an expression for this splitting. It is quite pleasant to find that the expression obtained justifies the need of viewing the source of magnetic interaction as an exchange interaction.

The two-electron Hamiltonian of a hydrogen molecule can be expressed as the sum of three parts:

$$\mathcal{H} = H_1 + H_2 + H_I \tag{14.13}$$

with

$$H_i = -\frac{\hbar^2}{2m}\nabla_i^2 - \frac{e^2}{4\pi\varepsilon_0}\left[\frac{1}{|\mathbf{r}_i - \mathbf{R}_A|} + \frac{1}{|\mathbf{r}_i - \mathbf{R}_B|}\right]; \qquad i = 1, 2 \tag{14.14}$$

and

$$H_I = \frac{e^2}{4\pi\varepsilon_0}\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} \tag{14.15}$$

When the electron-electron interaction H_I is ignored, the two-electron Schrödinger equation is written as

$$(H_1 + H_2)\,\psi(1, 2) = \xi\psi(1, 2) \tag{14.16}$$
$$(1 \equiv r_1;\ 2 \equiv r_2)$$

The Hamiltonian in this equation comprises one-electron Hamiltonians and solutions to this equation can be derived from the one-electron Schrödinger equation

$$H\psi = \varepsilon\psi \tag{14.17}$$

Let ψ' and ψ'' denote the solutions to (14.17) of lowest energies ε' and ε'' ($\varepsilon' < \varepsilon''$). The symmetric solution of the lowest energy to the two-electron Schrödinger equation (14.16) is

$$\psi_s(1, 2) = \psi'(1)\,\psi'(2) \tag{14.18}$$

with

$$\xi_s = 2\varepsilon'$$

This is clearly a singlet state meeting the requirement of the Pauli principle. The lowest antisymmetric solution is a triplet:

$$\psi_t(1, 2) = \psi'(1)\ \psi''(2) - \psi'(2)\ \psi''(1) \tag{14.19}$$

with

$$\xi_t = \varepsilon' + \varepsilon''$$

We are thus led to the result $\xi_s < \xi_t$ which is in fact a general theorem for the two-electron systems. The ground state energy ($2\varepsilon'$) is determined by the symmetric solution (14.18) which is not expected to be a proper solution to the exact two-electron Schrödinger equation that includes the electron-electron Coulomb interaction as well. Actually, it turns out to be a thoroughly bad solution when the protons are far separated because the solution fails to deal effectively with the electron-electron interaction in this situation mainly because of the structures of ψ' and ψ''. The specialized case of a solid in which $N = 2$ and the protons are well-separated, is most elegantly handled in the tight-binding approximation, generally used for the band structure calculations (see Section 7.8). In this approach, the one-electron stationary-state wavefunctions ψ' and ψ'' are taken as linear combinations of atomic stationary-state wavefunctions ϕ_A and ϕ_B centred at lattice points $\mathbf{R}_A$ and $\mathbf{R}_B$, respectively:

$$\begin{aligned} \psi' &= \phi_A + \phi_B \\ \psi'' &= \phi_A - \phi_B \end{aligned} \tag{14.20}$$

Accordingly, the two-electron wavefunctions (14.18) and (14.19) change to

$$\psi_s(1, 2) = \phi_A(1)\phi_B(2) + \phi_B(1)\phi_A(2) + \phi_A(1)\phi_A(2) + \phi_B(1)\phi_B(2) \tag{14.21}$$

and

$$\psi_t(1, 2) = 2[\phi_B(1)\phi_A(2) - \phi_A(1)\phi_B(2)] \tag{14.22}$$

where $\phi_\alpha(i)$ stands for an orbital of the hydrogen atom α at which the electron i is localized.

The structure of (14.21) reveals that states in which both electrons are located on the same atom ('ionic states') are equally represented in this form of $\psi_s(1, 2)$. The corresponding terms (the last two) describe a hydrogen molecule as a H^- ion and a bare proton which may be acceptable (though far from being accurate), provided we ignore the repulsive Coulomb term (the electron-electron interaction) or at least absorb it into the effective potential of the ions. Therefore, in order to include the effect of the electron-electron interaction, especially when the protons are well-separated, the ionic states be better omitted. This is essentially the Heitler–London approximation according to which (14.21) reduces to

$$\psi_s(1, 2) = \phi_A(1)\phi_B(2) + \phi_B(1)\phi_A(2) \tag{14.23}$$

The singlet-triplet splitting is then estimated from

$$\xi_s - \xi_t = \frac{\langle \psi_s | \mathcal{H} | \psi_s \rangle}{\langle \psi_s | \psi_s \rangle} - \frac{\langle \psi_t | \mathcal{H} | \psi_t \rangle}{\langle \psi_t | \psi_t \rangle} \tag{14.24}$$

where $\mathcal{H}$ is the complete Hamiltonian inclusive of the electron-electron interaction.

With a little bookkeeping, it can be shown that this splitting in the limit of large spatial separation is given by

$$\frac{1}{2}(\xi_s - \xi_t) = \frac{e^2}{4\pi\varepsilon_0}\int\left(\frac{1}{|\mathbf{R}_A - \mathbf{R}_B|} + \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} - \frac{1}{|\mathbf{r}_1 - \mathbf{R}_A|} - \frac{1}{|\mathbf{r}_2 - \mathbf{R}_B|}\right)$$

$$\times \phi_A^*(1)\,\phi_A(2)\,\phi_B^*(2)\,\phi_B(1)\,d\mathbf{r}_1\,d\mathbf{r}_2 \tag{14.25}$$

Since the matrix element represented by the integral is between two states that differ only through the exchange of the coordinates of the two electrons, the singlet-triplet splitting is justifiably referred to as the *exchange splitting*. In the same spirit, the source of splitting, when viewed as the source of magnetic interaction, is called the *exchange interaction* and the integral as the *exchange integral*, customarily denoted by J_{ex}. This implies that

$$(\xi_s - \xi_t) = 2J_{ex} \tag{14.26}$$

14.3 THE HEISENBERG MODEL

The spin-dependence of energy of a spin configuration is implicitly carried through the total wavefunction which is a product of the spatial and spin functions. In this section, we discuss the Heisenberg model for a two-electron system in a hydrogen molecule. We will define a spin-dependent Hamiltonian whose eigenvalues are the same as those of the original Hamiltonian in certain limit as discussed below. This Hamiltonian is called the *spin Hamiltonian.*

When the protons in a hydrogen molecule are well separated, the ground state wavefunction describes two independent hydrogen atoms. Since each electron can have two spin orientations, the ground state is fourfold degenerate. But when the protons are not far apart, ξ_s is no more equal to ξ_t and a splitting of the fourfold degeneracy occurs. The splitting, however, is quite small compared to other excitation energies of the system. If k_BT is small, $\sim (\xi_s - \xi_t)$, then only the lowest four-state manifold can be thermally populated. The analysis of many properties especially at thermal equilibrium becomes far more easier if other excited states are simply overlooked. We can then use a linear combination of these four lowest states to represent the ground state wavefunction of the system. The eigenvalues of the corresponding Hamiltonian (the spin Hamiltonian) operator will then have the same eigenvalues as those of the original Hamiltonian within the energy range of the lowest four states whose spin is contained in the eigenfunctions.

Each individual electron spin operator satisfies $\mathbf{S}_i^2 = \frac{1}{2}\left(\frac{1}{2} + 1\right) = \frac{3}{4}$, requiring the total spin operator $\mathbf{S}$ to satisfy

$$\mathbf{S}^2 = (\mathbf{S}_1 + \mathbf{S}_2)^2 = \frac{3}{2} + 2\mathbf{S}_1 \cdot \mathbf{S}_2 \tag{14.27}$$

Using the result that the eigenvalue of $\mathbf{S}^2$ is $S(S + 1)$ in states of spin S, we obtain the eigenvalue of the operator $\mathbf{S}_1 \cdot \mathbf{S}_2$ from (14.27) as

$$-\frac{3}{4} \text{ in the singlet } (S = 0) \text{ state}$$

and (14.28)

$$+\frac{1}{4} \text{ in the triplet } (S = 1) \text{ state}$$

In view of (14.28), the spin Hamiltonian can be manipulated to have the form:

$$\mathcal{H}^{\text{spin}} = \frac{1}{4}(\xi_s + 3\xi_t) - (\xi_s - \xi_t)\,\mathbf{S}_1 \cdot \mathbf{S}_2 \tag{14.29}$$

Making use of (14.28) together with (14.27), we can easily verify that

$$\mathcal{H}^{\text{spin}}\,\psi_s = \xi_s \psi_s$$

and

$$\mathcal{H}^{\text{spin}}\,\psi_t = \xi_t \psi_t$$

shifting the origin to $\frac{1}{4}(\xi_s + 3\xi_t)$,

$$\mathcal{H}^{\text{spin}} = -2J_{ex}\,\mathbf{S}_1 \cdot \mathbf{S}_2 \qquad \text{[using (14.26)]} \tag{14.30}$$

It is essentially this term in the Hamiltonian that gives rise to the so called Weiss molecular field. The value of J_{ex} is usually obtained from a spin wave resonance measurement.

Relation (14.30) works fairly well for real solids where $\mathbf{S}_1$ and $\mathbf{S}_2$ may be regarded as spins on two neighbouring atoms. It would, however, be wrong to infer that $\mathbf{S}_1$ is directly coupled with $\mathbf{S}_2$. In the quantum mechanical picture, it can be shown that the exchange integral J_{ex} is a negative quantity for a small separation $|\,\mathbf{R}_A - \mathbf{R}_B\,|$ between the neighbouring magnetic atoms. As a consequence of this, the parallel alignment of spins becomes unfavourable because of the resulting positive sign of the spin Hamiltonian that is associated with a positive value of the exchange energy. This is consistent with what we found in the Heitler–London model. For ferromagnetism, the spacing $|\,\mathbf{R}_A - \mathbf{R}_B\,|$ must be larger than a certain critical distance, ensuring a positive value to J_{ex} and, therefore, the parallel alignment of spins. This places a requirement of the minimum spacing of magnetic atoms in a ferromagnet. The ferromagnetic behaviour of Fe, Co and Ni is consistent with this requirement, whereas the consequence of this condition is not observed in other transition metals with smaller atomic numbers. It is equally important to emphasize that ferromagnetism also requires the magnetic neighbours not to be far apart because the exchange integral J_{ex} gradually falls to zero for large values of $|\,\mathbf{R}_A - \mathbf{R}_B\,|$, as is evident in Fig. 14.6. The negative values of J_{ex} relate to the antiferromagnetic behaviour where the spins are ordered antiparallel so as to be in the lower energy configuration.

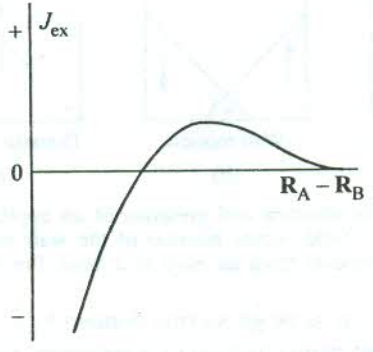


FIG. 14.6 A schematic of the variation of the exchange integral with distance between the two nearest magnetic ions.

While dealing with atoms with electrons of unpaired spins, we must be aware that states are to be filled in accordance with the Pauli principle so as to acquire a minimum electrostatic energy. It is sheer an accident in such a state of energy that a parallel spin alignment (or the spontaneous magnetization) emerges for an appropriate spacing $|\mathbf{R}_A - \mathbf{R}_B|$, producing ferromagnetism. Despite having the maximum possible unpaired spins for iron group elements, manganese in its elemental form (with a half-filled 3d shell) does not show ferromagnetism because of the small Mn-Mn distance, whereas in some of its compounds (e.g. MnSb, MnAs) the Mn-Mn spacing is large enough to make the exchange integral positive, thus letting them behave as ferromagnets. The importance of the Heisenberg theory of exchange interaction is emphasized through its success in accounting for the cause of magnetic ordering or massive molecular fields (and hence the high Curie temperatures). The gains of the theory in no way imply that the Weiss mean field theory be discarded. In fact, the physical theory of ferromagnetism is still often discussed in the framework of mean field theory on the ground of simplicity.

14.4 FERROMAGNETIC DOMAINS

Though pieces of ferromagnetic materials are normally found in the non-magnetized state, they can be magnetized (below T_c) with the application of a magnetic field. In order to explain this property of ferromagnetic solids Weiss, as indicated in Section 14.1, proposed the existence of an array of small magnetic domains that are randomly magnetized in the absence of an applied field. Each domain is magnetized to a saturated value M_s because of the parallel alignment of the atomic magnetic moments over its volume. The individual randomly oriented $\mathbf{M}_s$ vectors may add to zero [Fig. 14.7(a)] in the absence of a field, giving no macroscopic magnetization (or spontaneous magnetization). An applied magnetic field may at the best create an ideal parallel alignment of the magnetic moments over the whole volume of the specimen with the same uniform distribution of parallel moments as within a domain along the direction of whose magnetization the magnetic field is applied. In view of this fact, the saturated value of the magnetic moment per unit volume (recognized as the saturation magnetization) of the specimen as a whole will at the most be equal to M_s and can never exceed this value.

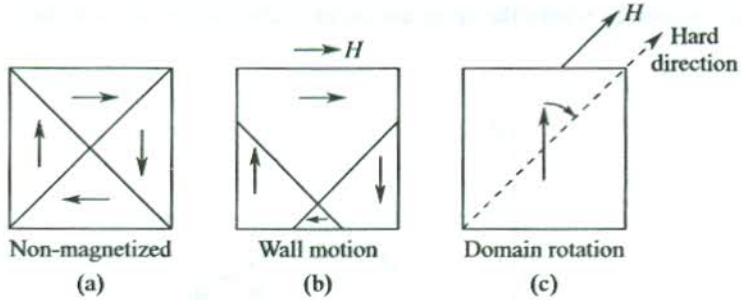


FIG. 14.7 Domain structure in the absence and presence of an applied field: (a) A non-magnetized state. (b) Magnetization by a field occurs because of the wall motion. (c) Magnetization caused by a rotation of domain moments from an *easy* to a *hard* direction.

Each domain is separated from its neighbouring domain by a domain wall called the *Bloch wall*. On applying a magnetic field, the magnetization of a specimen may occur either by the growth of a favourably-oriented (i.e. $\mathbf{M}_s \parallel \mathbf{H}$) domain at the expense of the neighbouring domains, i.e. by the motion of domain walls [Fig. 14.7(b)] or by the rotation of domains [Fig. 14.7(c)].

The first direct evidence for the existence of domains was found by Bitter* in powder patterns. In this technique a drop of colloidal suspension of finely divided ferromagnetic powder is placed on a carefully prepared surface of the ferromagnetic material under study. Because of the presence of strong local magnetic fields near the domain boundaries the particles collect along these directions and the domains can be observed under a microscope (Fig. 14.8). Typically, the dimensions of a single domain are of the order of 0.1 to 1.0 mm. A sample so small that it contained only one domain would exhibit permanent magnetization M_s, regardless of the presence of an applied magnetic field.

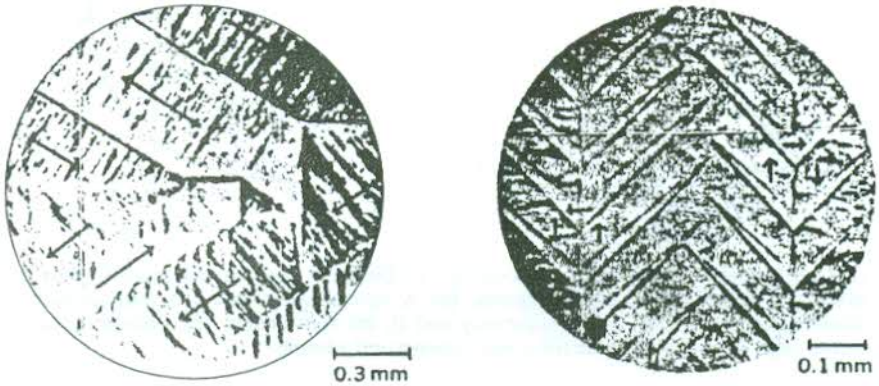


FIG. 14.8 Domain structures under a microscope: (a) For permalloy (78% Ni, 22% Fe) in which stable magnetization occurs along [111] directions and the resulting domains are separated by 180°, 90°, and 71° walls. (b) For iron containing 4% silicon; the cube edges [100] are the easy directions of magnetization and domains form 180° and 90° walls as viewed from a surface that is almost a cube face. The direction of magnetization within a domain is determined by observing the growth or contraction of the domain in a magnetic field. [After S. Chikazumi, *Physics of Magnetism* (Wiley, 1964).]

14.4.1 Technical Magnetization Curve

A representative magnetization curve is given in Fig. 14.9(a), showing dominant magnetization processes in different regions of the curve. The magnetic induction B of a multidomain ferromagnetic solid is plotted as a function of the applied magnetic field H in Fig. 14.9(b). The magnetic field is varied to carry out magnetization along the path indicated by the arrow. The technical magnetization curve traces a hysteresis loop. The hysteresis loop is interpreted easily in terms of the two magnetization processes described in Section 14.4 and the information available on Fig. 14.9(a). We define below two technical terms associated with the hysteresis curve.

Remanence, B_r

When the magnetic field is withdrawn ($H = 0$) after attaining the saturation point, the magnetic induction does not vanish but has a finite magnitude ($\mu_0 M_s$), referred to as the *remanence*.

Coercivity, H_c

Coercivity is defined as the reverse magnetic field that reduces the magnetic induction to zero, starting from saturation. The coercivity is often referred to as the coercive force.

* F. Bitter, *Phys. Rev.*, **38**, 1903 (1931).

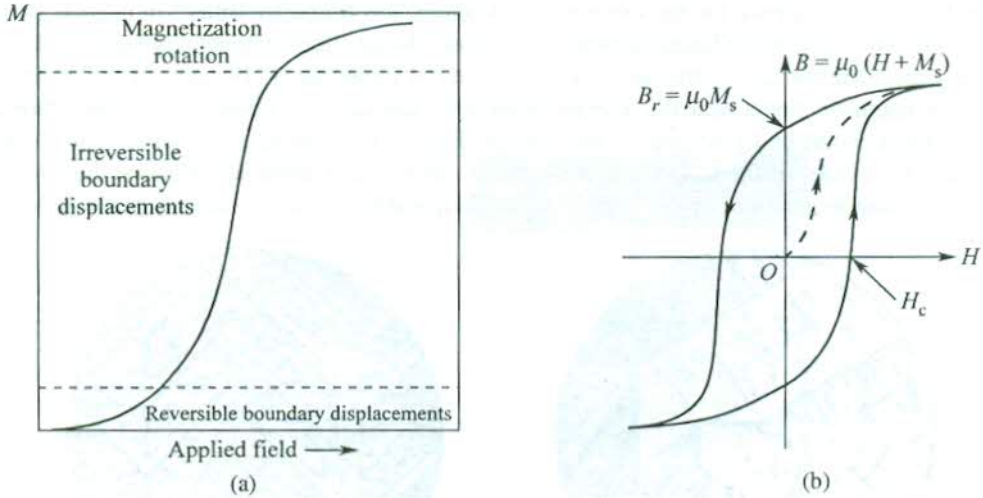


FIG. 14.9 (a) A schematic of the magnetization curve. Different regions are marked with the respective dominant processes of magnetization. (b) A hysteresis loop for a multidomain sample of a ferromagnetic solid; H_c is the coercivity and B_r the remanence. The dashed curve represents the initial path of magnetization for a non-magnetized sample.

The occurrence of hysteresis, regarded as a consequence of the domain structure, is accompanied by a loss of energy during the magnetization cycle. This implies the involvement of the domain structure in determining the efficiency of ferromagnetic devices. For example, a material to be used as a transformer core should have low coercivity in order that it incurs a low energy loss when taken round a cycle. A low coercivity relates to the high permeability which is ensured by using a pure, homogeneous and well-oriented material. The easy movement of domain walls in such a specimen may conform to a value of relative permeability as high as 4×10^6. A permanent magnet is another important application for which a material of high coercivity is required. A high value of coercivity may be achieved by impeding the domain wall movements. This, in principle, can be accomplished in small crystallites whose size must be below a certain characteristic critical size (~ 0.1 mm to 1 mm) so as to exhibit single domain attributes, ruling out the presence of any domain walls. In practice, specimens are made heterogeneous on a fine scale by precipitating in them a second metallurgical phase.

14.4.2 Anisotropy Energy

In the derivation of the spin Hamiltonian (14.30), the spin-orbit coupling was ignored and the exchange coupling between neighbouring spins was taken as perfectly isotropic, depending only on the angle between them. But the spins in a real ferromagnetic solid are coupled to the charge density via spin-orbit coupling (Fig. 14.10). Therefore, the energy of spins will depend to some extent on their absolute orientation with respect to the crystal axes, as well as on their relative orientation with respect to one another. The part of spin energy that depends on the absolute orientation of spins with respect to the crystal axes is known as the *magnetocrystalline* or *anisotropy* energy.

The anisotropy energy directs the magnetization along certain crystallographic axes called directions of easy magnetization. The directions of easy magnetization in iron, which is cubic, are its cube edges, as evidenced in Fig. 14.11. The body diagonals ([111] directions) that appear as hard directions of magnetization in iron are the easy directions in nickel, also a cubic crystal. In cobalt, the hexagonal axis is the only preferred direction of magnetization.

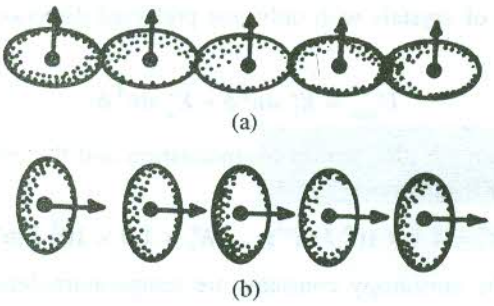


FIG. 14.10 An asymmetric electron charge distribution with different anisotropy energies for two different spin directions relative to the crystal axes. Changes in the exchange energy and electrostatic interaction energy take place with the rotation of the spin direction to which the asymmetry is tied. Due to the spin-orbit interaction, the charge distribution appears as spheroidal which is otherwise taken as spherical.

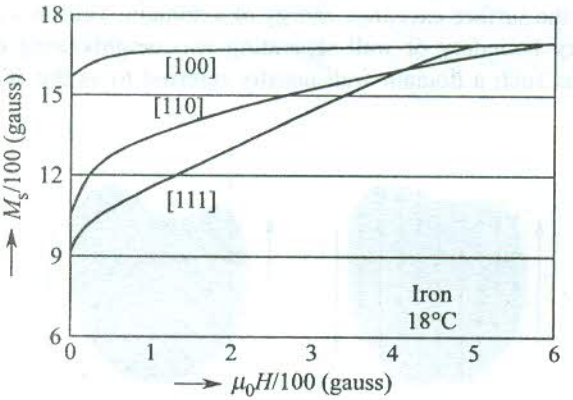


FIG. 14.11 Magnetization curves for a single crystal of iron at 291 K for different directions of the field relative to the crystal axes. [After R.G. Piety, *Phys. Rev.*, **50**, 1173 (1936).]

In order to obtain an expression for the anisotropy energy in terms of the direction of magnetization, the crystalline symmetry must be brought into consideration. For a cubic crystal magnetized in an arbitrary direction, the direction cosines α_1, α_2, α_3 are defined with reference to the cube edges. On account of the cubic symmetry, the expression for the anisotropy energy must be an even power of each α and be invariant under interchanges of αs among themselves. So $(\alpha_1^2 + \alpha_2^2 + \alpha_3^2)$ is the lowest order combination that satisfies the symmetry requirements. Being identically equal to unity, this combination cannot be associated with anisotropy effects. The leading term that accounts for the anisotropy energy is the next order combination $(\alpha_1^2\alpha_3^2 + \alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2)$. Despite the fact that this term by itself explains the experimental results on iron and nickel quite well, one more term of further allowed higher degree, namely $\alpha_1^2\alpha_2^2\alpha_3^2$ is retained and the rest neglected. The anisotropy energy density of a cubic crystal may thus be expressed as

$$U_{\text{anis}} = K_1(\alpha_1^2\alpha_2^2 + \alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2) + K_2\alpha_1^2\alpha_2^2\alpha_3^2 \tag{14.31}$$

The constants K_1 and K_2 are determined from experiment; for iron at room temperature:

$$K_1 = 4.2 \times 10^4 \text{ J m}^{-3}; \qquad K_2 = 1.5 \times 10^4 \text{ J m}^{-3} \tag{14.32}$$

The expression in the case of crystals with only one preferred direction of magnetization, such as cobalt, is simpler:

$$U_{\text{anis}} = K_1' \sin^2\phi + K_2' \sin^4\phi \tag{14.33}$$

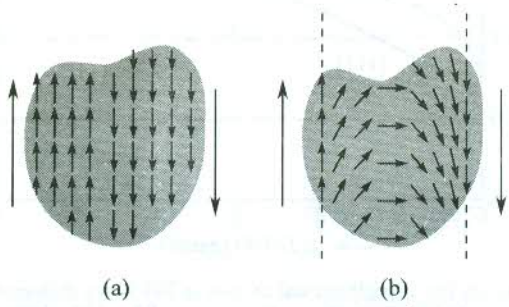
where ϕ is the angle between the direction of magnetization and the easy axis (hexagonal axis, for cobalt). For cobalt, at room temperature

$$K_1' = 4.1 \times 10^5 \text{ J m}^{-3}; \qquad K_2' = 1.0 \times 10^5 \text{ J m}^{-3} \tag{14.34}$$

It may be observed that the anisotropy constants are temperature-dependent.

14.4.3 The Bloch Wall

The change in direction of the magnetization on moving from one domain to its adjacent domain does not take place in a single step, as erroneously indicated by Fig. 14.12(a), since it is unnecessarily costly in exchange energy. The reversal of spin direction, in fact, is spread out over many spins [Fig. 14.12(b)], lowering the surface exchange energy of a domain. Thus in a real ferromagnetic solid there cannot exist a sharp boundary or wall separating two neighbouring domains magnetized in different directions and as such a domain wall usually referred to as the *Bloch wall* would have a finite thickness.



(a) (b)

FIG. 14.12 Change of the magnetization direction (or the spin orientation) across a domain wall (the Bloch wall): (a) An abrupt boundary. (b) A boundary with a finite width over which a gradual change in spin direction takes place. Since the former is more costly in exchange energy, the latter corresponds to the picture of a real crystal.

If the spin reversal is uniformly distributed over n spins, then each spin within the wall will differ in orientation from its immediate neighbour by an angle π/n. In a crude classical interpretation of the spin Hamiltonian (14.30), we can express the exchange energy of successive spin pairs as

$$\varepsilon_{\text{ex}} = -2J_{\text{ex}} S^2 \cos\left(\frac{\pi}{n}\right) \approx -2J_{\text{ex}} S^2 \left[1 - \frac{1}{2}\left(\frac{\pi}{n}\right)^2\right] \tag{14.35}$$

which is more than the minimum value $(-2J_{\text{ex}}S^2)$ for a single pair.

The cost of creating a 180° spin rotation down a line of $(n + 1)$ spins will be

$$n[\varepsilon_{\text{ex}} - (-2J_{\text{ex}} S^2)] = \frac{\pi^2}{2n} (2J_{\text{ex}} S^2) \tag{14.36}$$

This amount is lower than the cost of one-step or abrupt reversal by the factor $\pi^2/2n$.

In the absence of any other consideration, the width of the transition layer, the so-called Bloch wall, would broaden out to a thickness limited only by dipolar interactions. But we must appreciate that there is an anisotropy energy associated with the wall on account of spins in the transition layer being largely directed away from easy directions of magnetization. Since the anisotropy energy acts to reduce the width of the transition layer, the thickness of a Bloch wall, in practice, is determined by a balance between exchange and anisotropy energies.

Let us consider the example of a simple cubic lattice for calculating the energy density of a Bloch wall. Imagine a wall parallel to the cube face separating two oppositely magnetized domains. The energy has contributions from exchange and anisotropy energies and so, the energy per unit area σ of the wall may be expressed as

$$\sigma = \sigma_{ex} + \sigma_{anis} \tag{14.37}$$

Relation (14.36) gives an approximate value of the exchange energy for each line of spins normal to the wall. The number of lines per unit area of the wall is $1/a^2$ where a is the lattice constant. This gives

$$\sigma_{ex} = \frac{\pi^2 J_{ex} S^2}{na^2} \tag{14.38}$$

The anisotropy energy per unit area of the wall is

$$\begin{aligned} \sigma_{anis} &= \text{anisotropy constant} \times \text{wall thickness} \\ &= K \cdot (na) \end{aligned} \tag{14.39}$$

Therefore,

$$\sigma = \frac{\pi^2 J_{ex} S^2}{na^2} + Kna \tag{14.40}$$

The number n is determined by requiring σ to be minimum (i.e. $\partial\sigma/\partial n = 0$) for this number. Applying this condition, we get

$$n = \left(\frac{\pi^2 J_{ex} S^2}{Ka^3}\right)^{1/2} \tag{14.41}$$

Placing this value of n in (14.40), we have

$$\sigma = 2\pi\left(\frac{KJ_{ex} S^2}{a}\right)^{1/2} \tag{14.42}$$

The estimated orders of magnitude for iron are

$$n \approx 300 \quad \text{and} \quad \sigma \approx 10^{-3}\ \text{J m}^{-2}$$

14.4.4 Origin of Domains

The domain structure has its origin in the possibility of lowering the energy of the system. The total energy has contributions from exchange, anisotropy and magnetic energies.

The origin of domains can be understood with the help of structures drawn in Fig. 14.13, each representing a cross-section through a ferromagnetic single crystal. Figure 14.13(a) refers to a specimen with a single domain, i.e. the specimen is in the state of saturation magnetization (the maximum possible value). In such a specimen, the magnetic energy (arising on account of dipolar interaction) has a very high value $(1/2\mu_0) \int B^2 \, dV$. The magnetic energy is substantially reduced in the configuration shown in Fig. 14.13(b), since the spatial extension of the field has become very small. In fact such a subdivision is paid for in exchange energy, since the spins near the boundary of a domain experience unfavourable exchange interactions with misaligned spins in the domain wall. Because of the short range nature of exchange interaction, it is only the spins near the domain boundary whose exchange energy will be raised. In contrast, on the formation of domains the magnetic energy of every spin reduces on account of the long range of dipolar interactions. As we saw in Section 14.4.3, there is an energy associated with a domain wall, meaning thereby that some energy is always spent in creating a domain. Therefore, the subdivision into domains continues until the reduction in magnetic energy is less than the increase in energy to form another domain and its boundary, the Bloch wall.

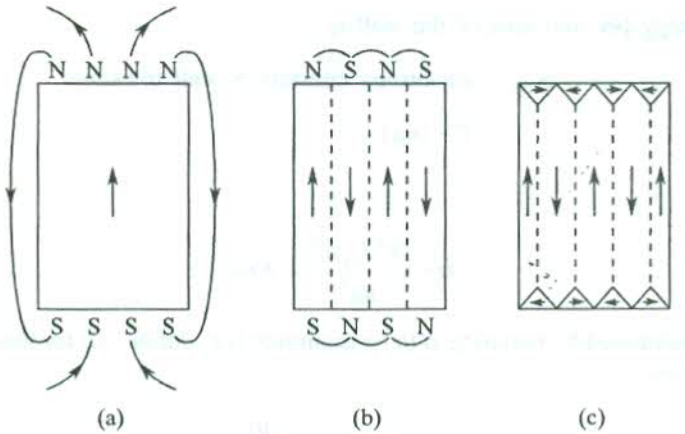


FIG. 14.13 The origin of domains. [After C. Kittel, *Rev. Mod. Phys.*, **17**, 541 (1949).]

Figure 14.13(c) shows a domain structure whose magnetic energy is zero. The presence of prism domains at the top and bottom of the crystal are responsible for the zero magnetic energy. Such domains are called *closure domains*. The wall separating a closure domain and a vertical domain is inclined at an angle of 45° with the magnetization directions in both types of domains. In view of this, the normal component of the magnetization is continuous across the boundary and there is no field (hence no energy) associated with the magnetization because there are no free poles. The energy required to produce a closure domain is essentially determined by the anisotropy energy of the crystal.

Solids showing antiferromagnetic, ferroelectric, antiferroelectric, ferroelastic and superconducting properties also have domain structures. Domain structures are often quite complicated and may not be as simple as described above.

14.5 NÉEL MODEL OF ANTIFERROMAGNETISM

When the exchange interaction between magnetic ions in a solid is such as to favour antiferromagnetic spin alignment within each domain, there is a tendency for an ordering at low temperatures in which up and down spins alternate in the structure. As stated in the beginning of the chapter, these solids with no gross magnetization are identified as antiferromagnetic materials. Néel* used the Weiss mean field theory to demonstrate that this property is observed below a certain critical temperature known as the Néel temperature T_N above which the paramagnetic behaviour is observed. A list of antiferromagnetic materials with their Néel temperatures is given in Table 14.2.

Table 14.2 Selected antiferromagnetic solids† with Néel temperatures T_N

Substance	T_N (K)	*Substance*	T_N (K)
MnO	122	$KFeF_3$	115
MnF_2	67.34	CoO	291
$MnCl_2$	2	CoF_2	37.7
$RbMnF_3$	54.5	$KCoF_3$	125
$KMnF_3$	88.3	NiO	600
FeO	198	Cr	311
FeF_2	78.4	VS	1040

† *Source:* F. Keffer, *Handbuch der Physik*, vol. 18, pt. 2 (Springer, New York 1966).

In the Heisenberg model for an antiparallel spin alignment of two adjacent magnetic ions to correspond to the lowest energy configuration, the exchange integral J_{ex} is required to be a negative quantity. But it is by no means a sufficient condition for a real crystal to possess the antiferromagnetic property. In fact, a three-dimensional array of magnetic ions must be favourably disposed to the creation of magnetic sublattices that may oppose the magnetization of each other. The simplest picture arises when the spins up and down fall on two separate interpenetrating sublattices† of an identical structure. Two examples of antiferromagnetic ordering are displayed in Fig. 14.14. Figure 14.14(a) shows the ordering on a BCC lattice that can be interpreted as two interpenetrating SC lattices, one

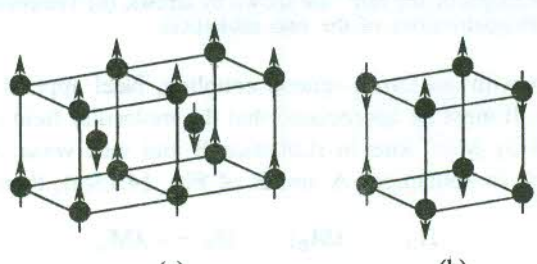


FIG. 14.14 Simple antiferromagnetic spin orders: (a) On a BCC lattice composed of two interpenetrating SC lattices, one with spins up and the other with spins of the same kind down. (b) On an SC lattice in which the spins of the same kind form two interpenetrating FCC lattices.

* L. Néel, *Ann. Phys.*, **18,** 5 (1932); **5,** 232 (1936).

† Four or even more sublattices are required for the distribution of magnetic ions to interpret the antiferromagnetic order of some crystals.

having the spins up and other the spins down. Similarly, the ordering on an SC lattice shown in Fig. 14.14(b) can be imagined as a combination of two FCC lattices with opposite spins. It can be verified for both the examples that the spin magnetic moments in a unit cell sum to zero moment (realizing that the moment at a corner contributes only 1/8th of its value).

The first experimental evidence* of antiferromagnetism was found in MnO which is also a case with two sublattices. The ordering of Mn^{2+} ions in $RbMnF_3$ [Fig. 14.15(a)] is one of the simplest with Mn^{2+} ions on an SC lattice. Let Mn^{2+} ions with spin up be on sublattice A and those with spin down on sublattice B. Figure 14.15(b) shows the temperature variation of magnetization M_A and M_B belonging to the two sublattices. They sum to zero at temperatures below T_N and drop independently to zero at T_N.

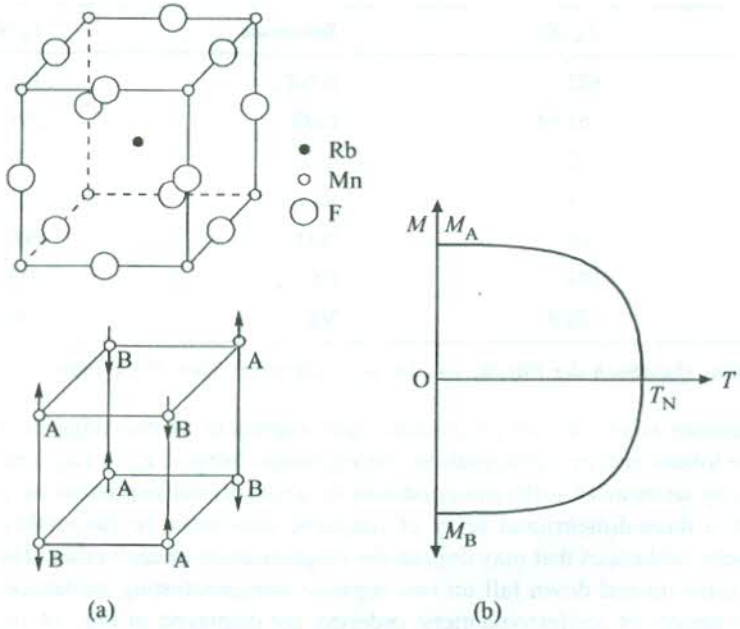


FIG. 14.15 (a) Unit cell structure of $RbMnF_3$ (antiferromagnetic, T_N = 54.5 K). The directions of spin moments of the magnetic ion Mn^{2+} are shown by arrows. (b) Temperature dependence of the equal and opposite magnetizations of the two sublattices.

In a model treatment with antiferromagnetic coupling, Néel applied the mean field theory to each sublattice separately. It must be appreciated that the molecular field experienced by Mn^{2+} ions in sublattice A is created by Mn^{2+} ions in sublattice B and vice versa. If H_A and H_B denote the respective molecular fields in sublattices A and B of Fig. 14.15(a), then

$$\mathbf{H}_A = -\lambda \mathbf{M}_B; \qquad \mathbf{H}_B = -\lambda \mathbf{M}_A \tag{14.43}$$

where λ is positive and the minus sign corresponds to an antiparallel interaction; the interaction of parallel spins in a sublattice with its own molecular field is neglected. Then, by analogy with (14.3) for a ferromagnet, we have

* H. Bizette, C.F. Squire and B. Tsai, *Compt. Rend.*, **207**, 449 (1938).

$$\mathbf{M}_A = \frac{C}{2T}(\mathbf{H} - \lambda \mathbf{M}_B) \tag{14.44a}$$

$$\mathbf{M}_B = \frac{C}{2T}(\mathbf{H} - \lambda \mathbf{M}_A) \tag{14.44b}$$

where **H** is the intensity of the applied magnetic field and C/2 is the appropriate Curie constant for one-half of the magnetic ions on each sublattice.

The saturation magnetizations of sublattices A and B are equal and opposite below the Néel temperature T_N. Equations (14.44) have a nonzero solution for M_A and M_B in the absence of an applied field if

$$\begin{vmatrix} 2T & \lambda C \\ \lambda C & 2T \end{vmatrix} = 0 \tag{14.45}$$

This determines the Néel temperature given by

$$T_N = \frac{\lambda C}{2} \tag{14.46}$$

which is a constant for a material.

Equations (14.44) are solved for M_A and M_B to obtain the susceptibility at $T > T_N$ in the form

$$\chi_m = \frac{\mathbf{M}_A + \mathbf{M}_B}{\mathbf{H}} = \frac{C}{T + \frac{\lambda C}{2}} = \frac{C}{T + T_N} \tag{14.47}$$

But experimental observations are not entirely in agreement with the predictions of (14.47). They are, instead, explained by

$$\chi_m = \frac{C}{T + \theta} \qquad T > T_N \tag{14.48}$$

where the ratio θ/T_N is usually greater than one. The discord between theory and experiment can be accounted for with the use of more complicated models that consider the influence of distant magnetic ions.

The observed temperature variation of the susceptibility of antiferromagnetic solids is shown in Fig. 14.16, together with those for paramagnetic and ferromagnetic solids for the purpose of a comparative study. The susceptibility of antiferromagnetic solids is not infinite at $T = T_N$, but has a weak cusp. It shows considerable anisotropy for a single crystal sample at temperatures below T_N. The susceptibility has different values for the magnetic field applied parallel and perpendicular to the direction of the sublattice magnetization. The behaviour of susceptibility in this temperature range is well accounted by experiments and can be appreciated even qualitatively. For a magnetic field acting perpendicular to the direction of magnetization the spins are tilted easily even at absolute zero, leading to a constant susceptibility as shown in Fig. 14.16(c). But when the field is parallel to the direction of magnetization, it is hard to cause spin flips. The magnetization is opposed by the full molecular

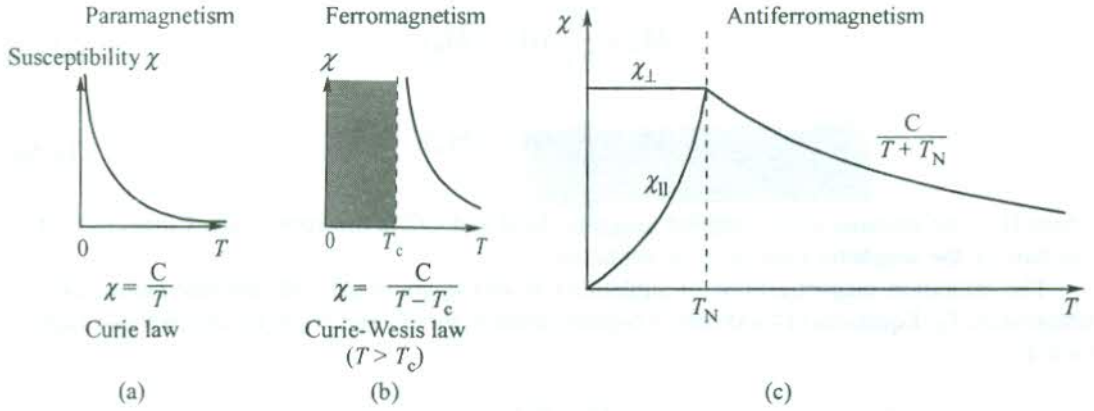


FIG. 14.16 Schematics of the susceptibility as a function of temperature in paramagnetic, ferromagnetic and antiferromagnetic solids.

field in this case and the susceptibility drops to zero as $T \to 0$ K with the two sublattice magnetizations of equal magnitude. On rise of temperature to T_N, the molecular field weakens and the susceptibilities for the two orientations of the magnetic field become comparable.

At temperatures $T \geq T_N$, the spin magnetic moments become entirely free because of the collapse of exchange coupling, and the susceptibility is back to the usual paramagnetic behaviour described by the Curie law.

The transition from an ordered magnetic phase to the disordered paramagnetic phase in all types of magnetic crystals is accompanied by a maximum in the heat capacity at the critical temperature that is much sharper than the break predicted by molecular field theory (see Fig. 14.5). Data from most carefully performed experiments suggest the following variation for the heat capacity:

$$C_V \propto \log | T - T_c | \tag{14.49}$$

which shows singularity at $T = T_c$. But the nature of singularity is such that $\int C_V \, dT$ remains finite, ruling out the involvement of latent heat, i.e. showing the transition as second-order type. The observed anomaly in an antiferromagnetic crystal ($NiCl_2 \cdot 6H_2O$) is depicted in Fig. 14.17.

14.6 NÉEL MODEL OF FERRIMAGNETISM

In view of the Néel sublattice model we need to elaborate on the definition of ferrimagnetism, given earlier in this chapter. Néel explored the consequences of the possibility that the magnetic moments of the two sublattices be unequal and anti-parallel. The corresponding configuration of spins is identified as a ferrimagnetic state that is usually characterized by a relatively weak spontaneous magnetization. A list of the selected ferrimagnetic materials is given in Table 14.3. This happens generally because of the presence of unidentical magnetic ions in the crystal. Néel* applied this concept of ferrimagnetism to explain the magnetic properties of an important group of magnetic oxides, known as *ferrites*. Ferrites are represented by the general formula, $MO \cdot Fe_2O_3$ where M can be any one of the divalent atoms: Fe, Co, Ni, Cu, Zn, Mg and Cd. These include lodestone (Fe_3O_4 = $FeO \cdot Fe_2O_3$) whose spontaneously magnetized pieces were used by ancient navigators as their primitive compasses. Lodestone is commonly known by the name *magnetite*.

* L. Néel, *Ann. Phys.*, **3**, 137 (1948).

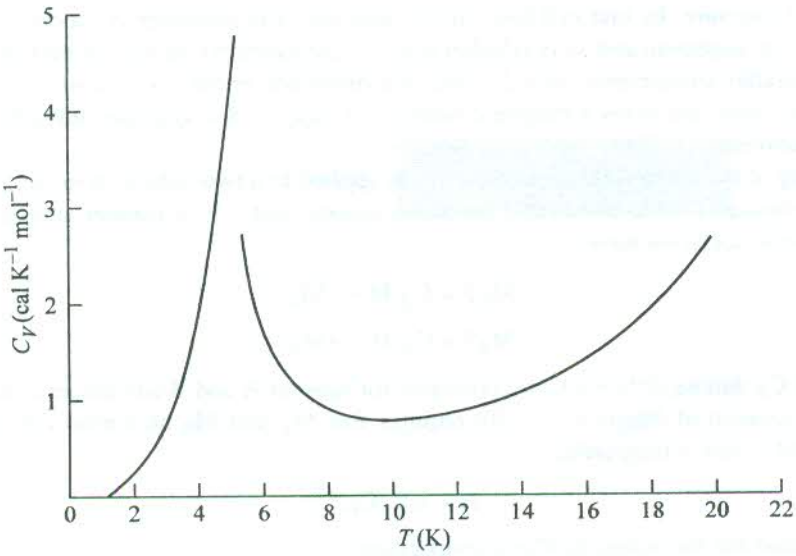


FIG. 14.17 Heat capacity anomaly in an antiferromagnetic crystal ($NiCl_2 \cdot 6H_2O$). [After W.K. Robinson and S.A. Friedberg, *Phys. Rev.*, **117**, 402 (1960).]

Table 14.3 Selected ferrimagnetic solids† with critical temperatures T_f and saturation magnetization M_s at $T = 0$ K

Substance	T_f (K)	M_s (gauss)
Fe_3O_4 (magnetite)	858	510
$CoFe_2O_4$	793	475
$NiFe_2O_4$	858	300
$CuFe_2O_4$	728	160
$MnFe_2O_4$	573	560
$Y_3Fe_5O_{12}$ (YIG)	560	195

† *Source:* F. Keffer, *Handbuch der Physik*, vol. 18, pt. 2 (Springer, New York, 1966).

In a Fe_3O_4 formula unit, there is one Fe^{2+} ion (with $3d^6$ configuration; $S = 2$) and two Fe^{3+} ions (with $3d^5$ configuration; $S = 5/2$). If the spins of all these ions were in a parallel alignment, the magnetic moment per formula unit would be about $14\mu_B$ ($= 4\mu_B + 2 \times 5\mu_B$). But the observed value is only $4.07\mu_B$, suggesting thereby that Fe_3O_4 must not be having an entirely parallel arrangement of spins, as is really confirmed by the neutron diffraction measurements. The magnetic ions are found to be distributed over two different symmetry sites: the tetrahedral and octahedral. These two sets of sites may be regarded to fall on two different sublattices. The spins of all ions at one type of site are in parallel alignment. But if the spins at tetrahedral sites point up, the spins at octahedral sites point down. Further, while Fe^{3+} ions are equally distributed over two sets of sites, the Fe^{2+} ions occupy only octahedral sites. Because of this distribution, only Fe^{2+} ions contribute to the net magnetic moment, equalling about $4\mu_B$ per formula unit. Thus we see that the model gives a value that is in exceedingly good agreement with the experimental value. Magnetite is a cubic ferrite having the

spinel* crystal structure. Its unit cell has 8 tetrahedral sites (on sublattice A) occupied by Fe^{3+} ions in parallel spin arrangement and 16 octahedral sites (on sublattice B), equally shared by Fe^{3+} and Fe^{2+} ions also in parallel arrangement such that the spin directions in the two sublattices are opposite to each other. This structure gives a magnetic moment of $32\mu_B$ ($= 8 \times 4\mu_B$) per unit cell, in agreement with the experimental value.

Continuing in the mean field approximation as applied to a two-sublattice model, we proceed to set up an expression for the observable magnetic susceptibility. In a manner similar to that of an antiferromagnetic solid, we have

$$\mathbf{M}_A T = C_A(\mathbf{H} - \lambda \mathbf{M}_B) \tag{14.50a}$$

$$\mathbf{M}_B T = C_B(\mathbf{H} - \lambda \mathbf{M}_A) \tag{14.50b}$$

where C_A and C_B denote different Curie constants for ions on A and B sublattices in that order. The simultaneous solution of equations (14.50) requires that $\mathbf{M}_A$ and $\mathbf{M}_B$ be nonzero in the absence of an applied field below a temperature,

$$T_f = \lambda(C_A C_B)^{1/2} \tag{14.51}$$

where T_f is called the ferrimagnetic Curie temperature.

Solving (14.50) for M_A and M_B at $T > T_f$, we have

$$\chi_m = \frac{M_A + M_B}{H} = \frac{T(C_A + C_B) - 2\lambda C_A C_B}{T^2 - T_f^2} \tag{14.52}$$

The more complicated form of (14.52) points towards a possible less straightforward behaviour for the susceptibility than what follows from the Curie–Weiss law. The $(1/\chi)$ versus T plot derived from the experimental values for magnetite and shown in Fig. 14.18 furnishes a sufficient proof to this effect. The curvature of the plot is a characteristic feature of ferrimagnetic materials. But the saturation magnetization is found to fall from a maximum at very low temperatures to zero at the ferrimagnetic Curie temperature T_f in a manner generally similar to that of a ferromagnetic solid.

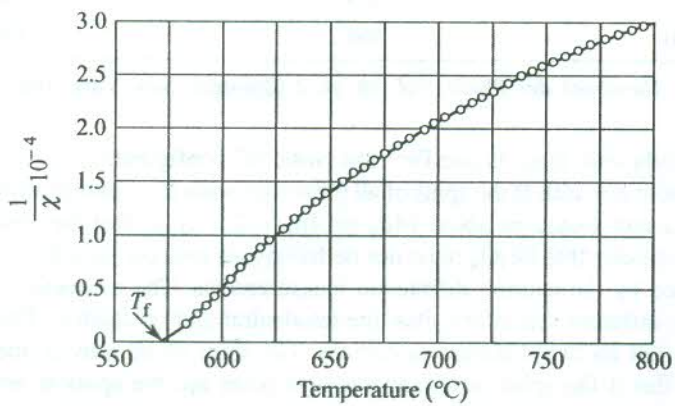


FIG. 14.18 Inverse susceptibility as a function of temperature for magnetite, Fe_3O_4.

* See C. Kittel, *Introduction to Solid State Physics*, 7th ed., p. 460 (John Wiley, 1996).

Ferrites are known to be poor conductors of electricity. Several members of this family, crystallizing in spinel structure have been used as rf transformer cores and toroidal cores serving as primary information storage in most of the computers developed about three decades ago.

There is another group of ferrimagnets that have found important applications. These are iron garnets with general formula, $M_3Fe_5O_{12}$ where M is a trivalent metallic atom such as yttrium or one of the rare-earth elements and Fe is the trivalent ferric ion ($S = 5/2$, $L = 0$). An example of iron garnets is $Y_3Fe_5O_{12}$, known as YIG. These are also electrical insulators and are used in magnetic resonance and spin wave studies.

14.7 SPIN WAVES

Individual elementary magnetic moments in a magnetic crystal are essentially completely ordered at $T = 0$ K. But, on increasing the temperature, the magnetization decreases towards zero at the critical temperature (T_c, T_N or T_f), since the atomic moments acquire enough thermal energy to turn against the effective molecular field. A deviation from the ordered state may be produced at $T = 0$ K as well by providing the spin system with an excitation, achieved with the energy from an external source. The excitations of a spin system are customarily described by treating the completely ordered (presumably at $T = 0$ K) as the ground state of the system. A state in which one or more spins are reversed relative to their orientation at absolute zero, then represents an excited state of the system.

Consider the simplest spin system with all its atomic magnetic moments in parallel alignment at absolute zero (i.e. a ferromagnet) with $S_i^z = S$, where S_i^z denotes the component of the spin vector state of a moment at site i along the magnetization direction (z-direction). An excitation of the spin at site i that may change the value of S_i^z from S to $(S - 1)$ against the effective molecular field or the exchange field B_{ex} would cost an energy $2\mu_B B_{ex}$. Because all the spins are coupled through the exchange interaction, the excitation does not remain localized at the site; it is in fact transferred to the neighbouring spin, setting its onward propagation in the crystal. Deviations in the spin direction at different sites follow a certain pattern such that the tips of the spin vectors convolute into a wave-like envelope (Fig. 14.19). The wave thus generated is called a *spin wave*. Discussions of ferromagnetism in terms of spin waves, however, became a practice only after the spin waves were observed* in the spin wave resonance of thin ferromagnetic films. On symmetry considerations, a spin representing a true excitation is expected to have the same amplitude at every site. The quantized unit of spin wave energy is called a *magnon*. The total magnetization decreases by one unit of spin with the excitation (or creation) of a magnon.

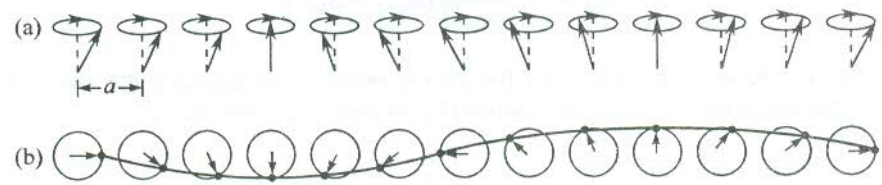


FIG. 14.19 Formation of a spin wave over a line of spins: (a) A perspective view of changing spin orientations. (b) Spin vectors as viewed from above. The curve joining the tips of spin vectors has a wave-like shape. A spin wave is shown here as an oscillation in the relative orientations of spins on a linear lattice.

* M.H. Seavey, Jr. and P.E. Tannenwald, *Phys. Rev. Letters*, 1, 168 (1958).

We attempt in this section a classical treatment of magnons, also referred to as the elementary excitations of a spin system. It must, however, be observed that a quantum mechanical approach is the befitting way to deal with this problem, because spin is purely a quantum concept. Setting all reasons of simplicity and convenience aside, it must be emphasized at the same time that some of the results of a classical calculation are surprisingly the same as those given by a quantum mechanical method. The magnon dispersion relation for a model ferromagnet discussed below is one such example.

14.7.1 Magnons in Ferromagnets

Let us consider a one-dimensional ferromagnet consisting of a linear chain of identical magnetic ions. If there are in total N ions each of spin S, coupled by the exchange interaction with the nearest neighbour spins, the exchange energy of the spin system in the Heisenberg model is given by

$$U = -2J_{ex} \sum_{n=1}^{N} \mathbf{S}_n \cdot \mathbf{S}_{n+1} \tag{14.53}$$

with system in the ground state and spins treated as classical vectors, $\mathbf{S}_n \cdot \mathbf{S}_{n+1} = S^2$. Thus the ground state energy of the given ferromagnet is

$$U_0 = -2NJ_{ex}S^2 \tag{14.54}$$

The exchange energy contributed by the nth spin to (14.53) is

$$U_n = -2J_{ex}\, S_n \cdot (S_{n+1} + S_{n-1})$$

$$= -\boldsymbol{\mu}_n \cdot \left(-\frac{2J_{ex}}{g\mu_B}\right)(\mathbf{S}_{n+1} + \mathbf{S}_{n-1}) \qquad (\text{using } \boldsymbol{\mu}_n = -g\mu_B\mathbf{S}) \tag{14.55}$$

where $\boldsymbol{\mu}_n$ is the magnetic moment associated with the nth spin.

The U_n can also be expressed as

$$U_n = -\boldsymbol{\mu}_n \cdot \mathbf{B}_{ex} \tag{14.56}$$

where $\mathbf{B}_{ex}$ is the effective molecular field.

Compairing (14.55) and (14.56), we have

$$\mathbf{B}_{ex} = \left(-\frac{2J_{ex}}{g\mu_B}\right)(\mathbf{S}_{n+1} + \mathbf{S}_{n-1}) \tag{14.57}$$

The gyroscopic equation of motion of the spin is obtained by equating the rate of change of its angular momentum with the torque experienced by its magnetic moment under the influence of the field $\mathbf{B}_{ex}$:

$$\hbar \frac{d\mathbf{S}_n}{dt} = 2J_{ex}\, \mathbf{S}_n \times (\mathbf{S}_{n+1} + \mathbf{S}_{n-1})$$

or

$$\frac{d\mathbf{S}_n}{dt} = \left(\frac{2J_{ex}}{\hbar}\right)[\mathbf{S}_n \times \mathbf{S}_{n+1} + \mathbf{S}_n \times \mathbf{S}_{n-1}] \tag{14.58}$$

This equation of motion is nonlinear. It can be linearized for small amplitudes by expressing $\mathbf{S}_n$ as

$$\mathbf{S}_n = \mathbf{S} + \boldsymbol{\sigma}_n \tag{14.59}$$

where $\mathbf{S}$ denotes the component of $\mathbf{S}_n$ along the magnetization direction (z-direction) and $\boldsymbol{\sigma}_n$ is a small spin vector in the xy-plane.

The geometrical relationship between the three spin vectors of (14.59) is shown in Fig. 14.20. Note that $\mathbf{S}$ is a constant vector, i.e. $\dfrac{d\mathbf{S}}{dt} = 0$. Placing the value of $\mathbf{S}_n$ from (14.59) in (14.58) and retaining only the first order terms in $\boldsymbol{\sigma}_n$, we get

$$\frac{d\boldsymbol{\sigma}_n}{dt} = \left(\frac{2J_{\text{ex}}}{\hbar}\right) S(\boldsymbol{\sigma}_{n+1} + \boldsymbol{\sigma}_{n-1} - 2\boldsymbol{\sigma}_n) \quad \left(\text{enforcing } \frac{d\mathbf{S}}{dt} = 0\right) \tag{14.60}$$

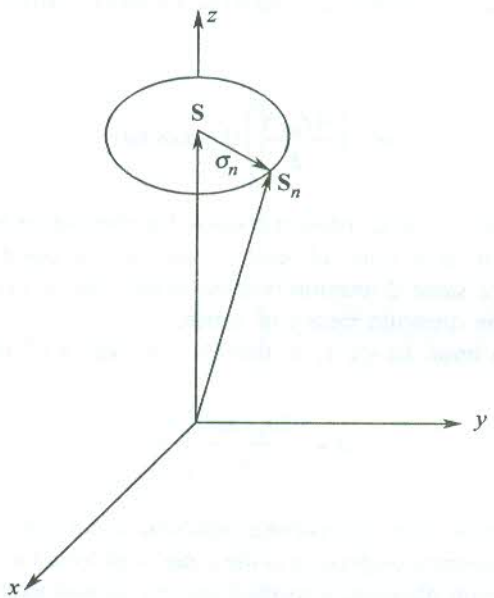


FIG. 14.20 Precession of spin vectors $\mathbf{S}_n$ in a classically interpreted spin wave.

From the above equation, the components of spin motion along the x and y directions may be written as

$$\left(\frac{d\sigma_n}{dt}\right)_x = \left(\frac{2J_{\text{ex}}\,S}{\hbar}\right)(2\sigma_n - \sigma_{n+1} - \sigma_{n-1})_y \tag{14.61}$$

$$\left(\frac{d\sigma_n}{dt}\right)_y = -\left(\frac{2J_{\text{ex}}\,S}{\hbar}\right)(2\sigma_n - \sigma_{n+1} - \sigma_{n-1})_x \tag{14.62}$$

If we construct a complex spin vector

$$\boldsymbol{\sigma}^+ = \boldsymbol{\sigma}_x + i\boldsymbol{\sigma}_y$$

relations (14.61) and (14.62) can be combined to yield

$$\frac{d\sigma_n^+}{dt} = \left(\frac{2J_{ex}\,S}{i\hbar}\right)(2\sigma_n^+ - \sigma_{n+1}^+ - \sigma_{n-1}^+) \tag{14.63}$$

By analogy with phonons we look for travelling wave solutions of (14.63) of the form

$$\sigma_n^+ = u \exp\,[i(kna - \omega t)] \tag{14.64}$$

where a is the interionic spacing in the chain.

Substituting the above solution in (14.63), we obtain

$$\omega = \left(\frac{2J_{ex}\,S}{\hbar}\right)\{2 - [\exp(ika) + \exp(-ika)]\}$$

or

$$\omega = \left(\frac{4J_{ex}\,S}{\hbar}\right)(1 - \cos ka) \tag{14.65}$$

Relation (14.65) is known as the dispersion relation for spin waves in a linear ferromagnet with nearest-neighbour interaction. It is quite pleasant to find that a quantum mechanical method of derivation gives precisely the same dispersion relation in this case as (14.65). This may be verified by looking into any book on quantum theory of solids.

In the long wavelength limit, $ka << 1$, so that $(1 - \cos ka) \simeq 1/2\,(ka)^2$ and the frequency is

$$\omega \approx \left(\frac{2J_{ex}\,Sa^2}{\hbar}\right) \cdot k^2 \tag{14.66}$$

This shows that the magnon dispersion curve is parabolic in the long wavelength limit, contrasting the linear behaviour for the phonon dispersion in the same limit for a linear monatomic crystal lattice (see Fig. 4.4). By analogy with phonons, a method entirely similar to that described in Section 4.8 can be used for magnon creation and annihilation through the inelastic scattering of slow neutrons in a ferromagnetic solid. A dispersion curve obtained by this procedure is shown in Fig. 14.21.

The quantization of spin waves is carried out in a manner exactly similar to that for photons and phonons. If there are n_k magnons in a mode of frequency ω_k, the energy of the mode is expressed as

$$\varepsilon_k = \left(n_k + \frac{1}{2}\right)\hbar\omega_k \tag{14.67}$$

14.7.2 The Bloch $T^{3/2}$ Law

The concept of spin waves was introduced by Felix Bloch (1930) who correlated the thermal excitation of magnons with the decrease in spontaneous magnetization from its saturated value at absolute zero. He went on to set up a law that describes the temperature dependence of the decrease in magnetization.

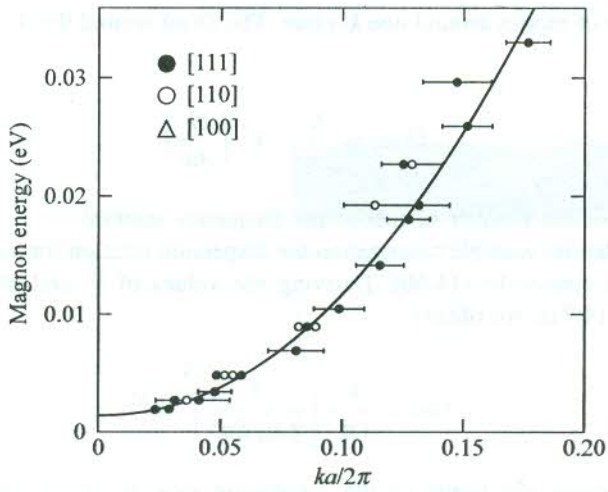


FIG. 14.21 Magnon dispersion relation in a ferromagnetic alloy (92% cobalt and 8% iron) determined from an inelastic neutron scattering experiment at room temperature. The energy is plotted against the dimensionless equivalent of wavevector. [After R.N. Sinclair and B.N. Brockhouse, *Phys. Rev.*, **120,** 1638 (1960).]

To derive the Bloch's law we essentially need to determine the average number of thermally excited magnons at thermal equilibrium. This number gives a measure of the decrease in magnetization, since with the excitation of each magnon there is a decrease in magnetization by one unit of spin moment.

It is evident from the form of (14.67) that the structure of energy levels of magnons is identical to those of a harmonic oscillator. This allows us to treat magnons in a manner similar to that in practice for the radiation oscillators of Planck. Thus the average number of magnons in the mode **k** at thermal equilibrium is given by the Planck distribution function:

$$\langle n_{\mathbf{k}} \rangle = \frac{1}{\exp(\hbar\omega/k_B T) - 1} \tag{14.68}$$

Then, the total number of thermally excited magnons at temperature T is estimated to be equal to

$$\sum_{\mathbf{k}} n_{\mathbf{k}} = \int D(\omega) \langle n(\omega) \rangle \, d\omega \tag{14.69}$$

where $D(\omega)$ denotes the density of the magnon modes.

There is a single polarization for magnons for each value of **k** and all the useful values of **k** (the unique values) lie in the first Brillouin zone, as is the case with phonons. The integral in (14.69) is evaluated by emphasizing the analogy with phonons.

The periodic boundary condition gives the number of modes per unit volume in the **k**-space as $V/(2\pi)^3$, where V is the volume of the crystal. Therefore, the number of modes of the wavevector measuring between k and $(k + dk)$ for a spherical distribution of momentum in the **k**-space is equal to

$$\frac{V}{(2\pi)^3} \times \frac{d}{dk}\left(\frac{4\pi}{3}k^3\right) = \frac{V}{2\pi^2} \cdot k^2 \, dk \tag{14.70}$$

This gives the number of modes around one k-value. The $D(\omega)$ around this k-value can then be given as

$$D(\omega) = \frac{V}{2\pi^2} \cdot k^2 \left(\frac{dk}{d\omega}\right) \tag{14.71}$$

since it is a measure of the number of modes per frequency interval.

It can be shown that for a single polarization the dispersion relation for a cubic lattice in the limit of low frequencies is similar to (14.66). Deriving the values of k^2 and $dk/d\omega$ from (14.66) and substituting them in (14.71), we obtain

$$D(\omega) = \frac{V}{4\pi^2} \left(\frac{\hbar}{2J_{ex} Sa^2}\right)^{3/2} \omega^{1/2} \tag{14.72}$$

At very low temperatures the limits of the integration may be placed at 0 and ∞ because as $\omega \to \infty$, $\langle n(\omega) \rangle \to 0$ exponentially.

We rewrite (14.69) to express the total number of magnons as

$$\sum_{\mathbf{k}} n_{\mathbf{k}} = \frac{V}{4\pi^2} \left(\frac{k_B T}{2J_{ex} Sa^2}\right)^{3/2} \int_0^{\infty} \frac{x^{1/2}}{\exp(x) - 1} \quad \text{with } x = \frac{\hbar\omega}{k_B T}$$

$$= (0.0587) \frac{V}{a^3} \left(\frac{k_B T}{2J_{ex} S}\right)^{3/2} \tag{14.73}$$

For a cubic crystal with N magnetic ions per unit cell, the contribution of a single spin to the saturation magnetization $M(0)$ is

$$\frac{M(0)}{(N/a^3) S} = \frac{M(0)\, a^3}{NS} \tag{14.74}$$

Since the total spin is reduced from its saturation value NS by one unit per magnon, the change in magnetization is given by

$$\Delta M = \frac{\sum n_k}{V} \times \frac{M(0)\, a^3}{NS} = \frac{0.0587}{NS} \left(\frac{k_B T}{2J_{ex} S}\right)^{3/2} \cdot M(0) \tag{14.75}$$

Therefore, the fractional change of magnetization is written as

$$\frac{\Delta M}{M(0)} = \frac{0.0587}{NS} \left(\frac{k_B T}{2J_{ex} S}\right)^{3/2} \tag{14.76}$$

The above relation is famous as the Bloch $T^{3/2}$ law and has been verified by experiments. The heat capacity contributed by thermal excitation of magnons at low temperatures also varies as $T^{3/2}$.

14.7.3 Magnons in Antiferromagnets

Consider a one-dimensional antiferromagnet in which the spins with even indices $2n$ are on sublattice A with $S^z_{2n} = S$ and the spins with odd indices $(2n + 1)$ are on sublattice B with $S^z_{2n+1} = -S$. As for one-dimensional ferromagnets, we use

$$\mathbf{S}_{2n} = \mathbf{S} + \boldsymbol{\sigma}_{2n}$$
$$\mathbf{S}_{2n+1} = -\mathbf{S} + \boldsymbol{\sigma}_{2n+1} \tag{14.77}$$

to linearize the equations of motion for $\mathbf{S}_{2n}$ and $\mathbf{S}_{2n+1}$. The small spin vectors $\boldsymbol{\sigma}_{2n}$ and $\boldsymbol{\sigma}_{2n+1}$ lie in the xy-plane. Following the treatment of Section 14.7.1, the equations of motion for spins in the two sublattices are written as

$$\frac{d\sigma^+_{2n}}{dt} = \left(\frac{2iJ_{ex}S}{\hbar}\right)(2\sigma^+_{2n} + \sigma^+_{2n+1} + \sigma^+_{2n-1}) \tag{14.78}$$

$$\frac{d\sigma^+_{2n+1}}{dt} = -\left(\frac{2iJ_{ex}S}{\hbar}\right)(2\sigma^+_{2n+1} + \sigma^+_{2n} + \sigma^+_{2n+2}) \tag{14.79}$$

with $\sigma^+ = \sigma_x + i\sigma_y$.

Taking travelling wave solutions

$$\sigma^+_{2n} = A \exp i[2nka - \omega t]; \quad \sigma^+_{2n+1} = B \exp i[(2n+1)ka - \omega t] \tag{14.80}$$

and substituting them in (14.78) and (14.79), we get

$$\omega A = -\left(\frac{2J_{ex}S}{\hbar}\right)\{2A + B[\exp(ika) + \exp(-ika)]\} \tag{14.81a}$$

$$-\omega B = -\left(\frac{2J_{ex}S}{\hbar}\right)\{2B + A[\exp(ika) + \exp(-ika)]\} \tag{14.81b}$$

These equations have a solution, if

$$\begin{vmatrix} \left(-\frac{4J_{ex}S}{\hbar}\right) - \omega & \left(-\frac{4J_{ex}S}{\hbar}\right)\cos ka \\ \left(-\frac{4J_{ex}S}{\hbar}\right)\cos ka & \left(-\frac{4J_{ex}S}{\hbar}\right) + \omega \end{vmatrix} = 0 \tag{14.82}$$

which gives

$$\omega^2 = \left(\frac{4J_{ex}S}{\hbar}\right)^2 (1 - \cos^2 ka)$$

or

$$\omega = \frac{4J_{ex}S}{\hbar} \sin ka \tag{14.83}$$

In the long wavelength limit, (14.83) reduces to

$$\omega \approx \frac{4J_{ex}S}{\hbar} |ka| \tag{14.84}$$

since $|ka| << 1$ in this limit.

This is the dispersion relation for magnons in an antiferromagnet. The relation is different from that for a ferromagnet. For an antiferromagnet, ω is proportional to k and not to k^2 in the limit of low frequencies as given by (14.66) for a ferromagnet. The magnon dispersion is measured by inelastic neutron scattering experiments. The experimental magnon dispersion of $RbMnF_3$ crystal is traced in Fig. 14.22, showing a fairly large linear portion in the low frequency limit. The thermal excitation of magnons at low temperatures contributes a term in T^3 to the heat capacity. This term is in addition to the Debye T^3 term for phonons.

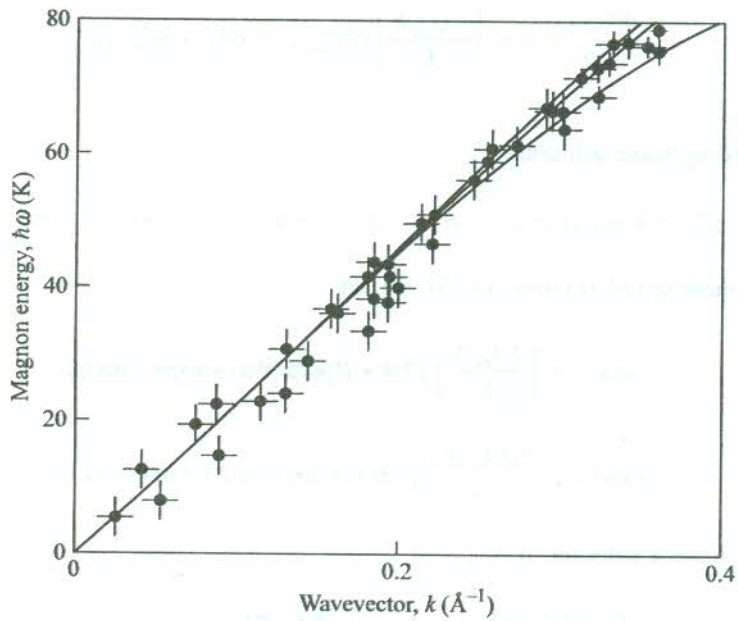


FIG. 14.22 Spectrum of magnon energy versus wavevector in the antiferromagnet $RbMnF_3$ as measured with an inelastic neutron scattering experiment at 4.2 K. [After C.G. Windsor and R.W.H. Stevenson, *Proc. Phys. Soc.* (London) **87**, 501 (1966).]

14.8 DETERMINATION OF MAGNETICALLY ORDERED STRUCTURES

Crystal structures are more often determined by the x-ray diffraction method. But the use of a slow neutron† (low energy) beam has certain distinct advantages, especially while dealing with the

† By virtue of being slow, the magnetic moment of a neutron gets sufficient time for interacting with the atomic moments of the system and thus gets scattered with the maximum possible information in its scattering cross-section.

magnetically ordered materials. The diffracted x-ray photons provide details about the spatial distribution of electronic charge, but carry no information about the atomic magnetic moment vectors in a magnetically ordered structure. On the other hand, a beam of slow neutrons serves as an excellent probe of local moments since the neutron itself has a magnetic moment which couples to the spin of elementary moments in a magnetic crystal. As a result of this coupling there appear peaks in the diffraction pattern in addition to those belonging to the non-magnetic Bragg reflection of neutrons by the atomic nuclei.

The additional lines get weaker in intensity as the temperature rises to the critical temperature at which the magnetic ordering vanishes and the lines disappear. These features easily distinguish them from other lines.

The gains of using a neutron beam are best appreciated in antiferromagnetic solids. The antiparallel ordering of moments within each domain results in zero macroscopic magnetic field which usually reveals the magnetic ordering in solids with a spontaneous magnetization. Therefore, a more subtle method is required to probe the ordering in such materials. As an example, we show the neutron diffraction pattern of MnO below and above the Néel temperature (120 K) in Fig. 14.23. Several lines in the pattern recorded at 80 K are not observed at 293 K, simply because the corresponding Bragg reflections originate from the magnetic ordering which exists no more above 120 K. Nuclear magnetic resonance offers another way to investigate the microscopic spin structure. For determining the size of a unit cell of a crystal having magnetic ordering it becomes imperative to take into consideration not only the equivalence of the sites but also the equivalence of the magnetic moment vectors located

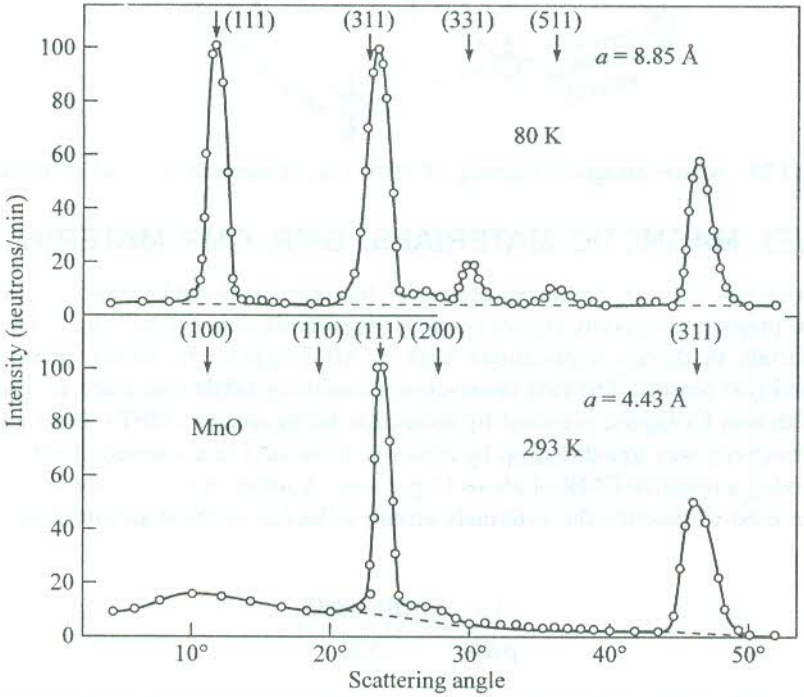


FIG. 14.23 Neutron diffraction patterns for an aniferromagnetic solid (MnO) below and above the Néel temperature (120 K). Reflection indices in the ordered and the disordered phases relate to the unit cells of different lattice constants 8.85 Å and 4.43 Å, respectively. The lines attributed to the ordering of spin moments are missing from the pattern recorded at 293 K since no ordered structure exists at this temperature. [After C.G. Shull, W.A. Strauser, and E.O. Wollan, *Phys. Rev.*, **83**, 333 (1951).]

at those sites. The size of a unit cell for an antiferromagnetic crystal as obtained by the x-ray diffraction is just half the size given by the neutron diffraction. The unit cells determined by the two techniques are referred to as *chemical unit cell* and *magnetic unit cell,* respectively. This point is clarified in Fig. 14.24 with the help of the ordering of moments of Mn^{2+} ions in the unit cell of $RbMnF_3$ ($T_N = 54.5$ K).

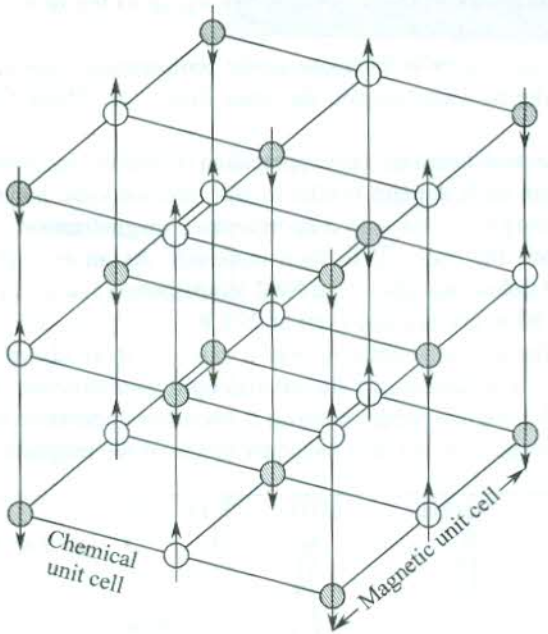


FIG. 14.24 Antiferromagnetic ordering of Mn^{2+} spin moments in a crystal of $RbMnF_3$.

14.9 NOVEL MAGNETIC MATERIALS: GMR–CMR MATERIALS

Recently, an extraordinarily large magnetoresistance* has been exhibited by some magnetic materials. This spectacular property commonly known as *Giant Magnetoresistance* (GMR) has found applications for these materials in device applications such as (i) magnetic recording (memory storage), (ii) actuators and (iii) sensors. The first observation concerning GMR was made in respect of Fe/Cr multilayers† with thin Cr layers, prepared by molecular beam epitaxy (MBE). For Cr layers of 9 Å thickness the resistivity was found to drop by almost a factor of 2 in a magnetic field of 2 T at 4.2 K (Fig. 14.25), giving a negative GMR of about 50 per cent. Another term *Colossal Magnetoresistance* (CMR) is often used to describe the extremely strong influence of the magnetic field. The CMR is defined as

$$\frac{\Delta\rho}{\rho(B)} = \frac{\rho(B) - \rho(0)}{\rho(B)} \tag{14.85}$$

The GMR–CMR effect is observed generally at low temperatures in the presence of large

* For definition refer to relation (9.65).
† M.N. Baibich, et al., *Phys. Rev. Lett.*, **61**, 2472 (1988).

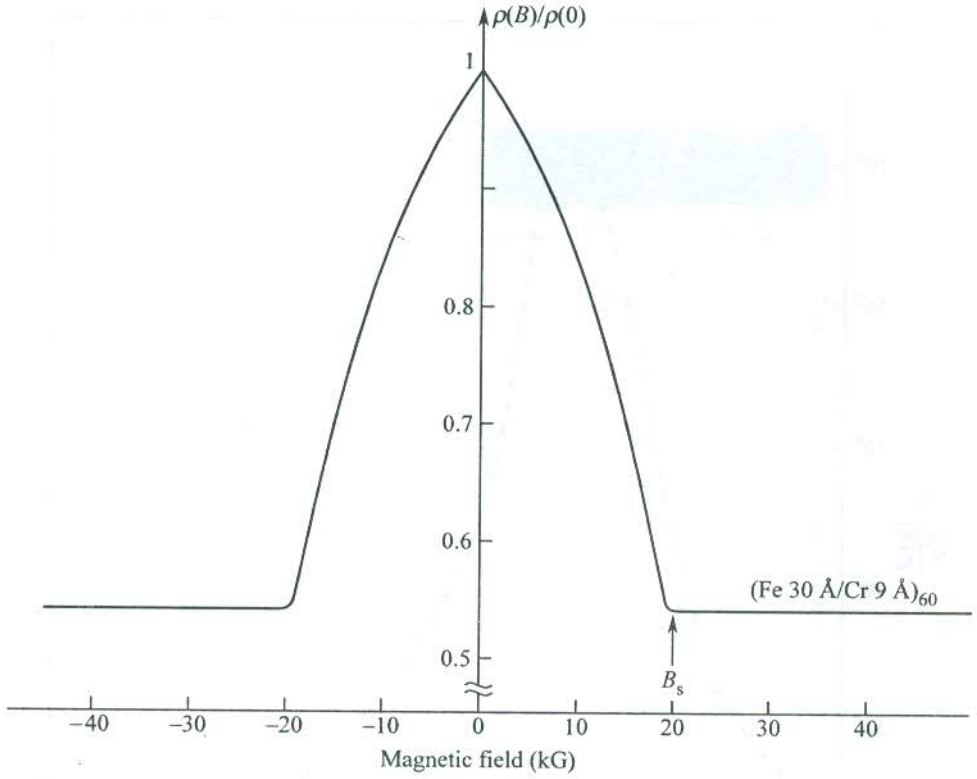


FIG. 14.25 Variation of resistivity of multilayers [(001) Fe 30 Å/(001) Cr 9 Å]$_{60}$ as a function of the magnetic field at 4.2 K. The specimen is a superlattice of 60 bilayers in which the current flows along the [110] direction and the magnetic field is applied in the layer plane (001) along the current direction. At the magnetic field B_s (in the saturated state; B_s = 2 T) the resistivity drops to almost half its zero-field value. [After M.N. Beibich et al., *Phys. Rev. Lett.*, **61**, 2472 (1988).]

magnetic fields (~ a few tesla). But we require to exploit this property ideally near room temperature and at low fields for enhancing the technological viability of GMR–CMR materials. Hence, ever since the discovery of GMR and CMR a relentless activity has been on to develop such materials. Recently, a promising class of magnetic materials with composition $R_{1-x}A_x MnO_3$ (R = La, Nd, Gd, Y; A = Ca, Sr, Ba, Pb) has been identified. These manganites having perovskite structure have the unique distinction of being paramagnetic and semiconducting at high temperatures. They make a transition to the ferromagnetic and metallic state at low temperatures. This is unusual of metal–insulator systems because they are generally metallic at high temperatures. In the metallic state the resistivity is unexpectedly high.

The spin-off during search for ideal GMR–CMR materials occurred with the work of McCormack et al.* In epitaxially grown thin films of La–Ca–Mn–O, MR is found to depend strongly on film thickness and temperature. The CMR reaches its maximum (in excess of 10^6 per cent) at 110 K with the magnetic field at 6 T (Fig. 14.26). The peak occurs just below the Curie temperature. For films thicker than ~ 2000 Å, the MR is reduced by orders of magnitude. The presence of grain boundaries

* M. McCormack, et al., *Appl. Phys. Lett.*, (USA) **64**, 3045 (1994).

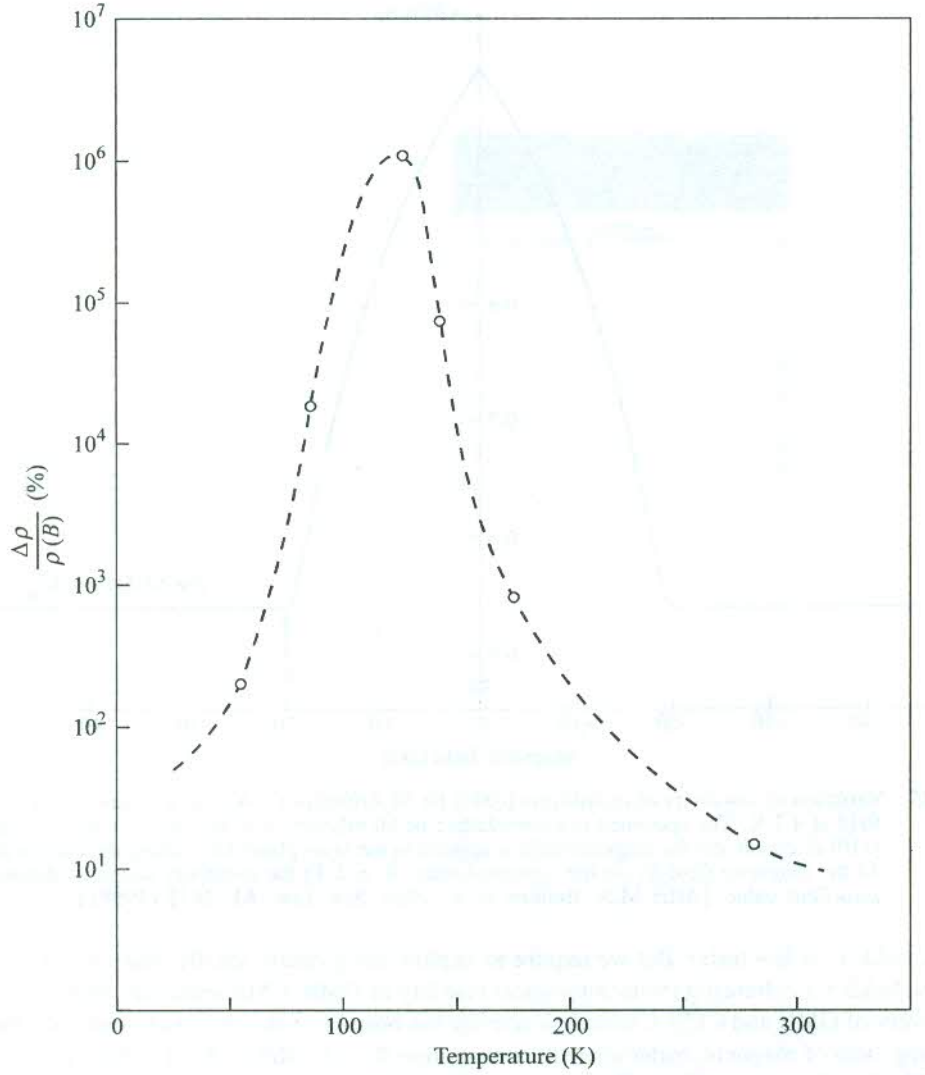


FIG. 14.26 Variation of magnetoresistance of a thin film of La–Ca–Mn–O as a function of temperature. The MR peaks at 110 K in a magnetic field of 6 T. [After S. Jin et al., *Appl. Phys. Lett.*, **67**, 557 (1995).]

leading to lattice strain is detrimental to achieving large MR. The MR improves further on heat treatment.

Several theoretical approaches have been advanced to understand ferromagnetism and the GMR effect in manganites. The ferromagnetism is interpreted in terms of the coupling between charge carriers and the coupling between localized spin moments of Mn ions. The picture of the GMR effect is, however, too complex to be described here. The details fall outside the scope of the present book and, therefore, dispensed with. Nevertheless, we must bring to focus the two most striking features of the GMR–CMR effect in manganites. Firstly, the MR peak can be shifted to occur at room temperature by adjusting the processing parameters. Secondly, the resistivity can be

manipulated by magnetic field to change by orders of magnitude. These keep the material scientists motivated to search tirelessly for ideal GMR–CMR materials and hold the realization of their dream in good stead.

SUMMARY

1. The total magnetic induction inside a ferromagnetic solid is $\mu_0(\mathbf{H} + \lambda\mathbf{M})$, where $\mathbf{H}$ is the intensity of the applied field and λ the Weiss coefficient.
2. The susceptibility of a ferromagnet above the Curie temperature T_c is given by the Curie law

$$\chi = \frac{C}{T - T_c}, \text{ where } T_c = \lambda C$$

3. The interaction energy because of the exchange coupling between spins $\mathbf{S}_1$ and $\mathbf{S}_2$ is equal to

$$-2 J_{ex}\, \mathbf{S}_1 \cdot \mathbf{S}_2$$

where J_{ex} is the exchange integral. The J_{ex} is positive for a ferromagnet and negative for an antiferromagnet.

4. The wall separating the domains magnetized in different directions is called the *Bloch wall*. The energy of the Bloch wall per unit area is $2\pi\left(\frac{KJ_{ex}\, S^2}{a}\right)^{1/2}$, where K is the anisotropy constant and a the lattice constant.
5. In an antiferromagnet, the susceptibility above the Néel temperature varies with temperature according to

$$\chi = \frac{C}{T + \theta}$$

where the ratio θ/T_N is usually greater than one.

6. The observed temperature variation of heat capacity of solids with magnetic ordering near T_c takes place according to

$$C_V \propto \log |T - T_c|$$

where T_c is the critical temperature.

7. The magnon dispersion relation for a one-dimensional monatomic ferromagnet in the limit of low frequencies is

$$\hbar\omega \approx (2J_{ex}Sa^2) \cdot k^2$$

8. The contributions of thermally excited magnons in ferromagnets at low temperatures to the heat capacity and to the fractional magnetization change are both proportional to $T^{3/2}$.

9. The magnon dispersion relation for a one-dimensional monatomic antiferromagnet in the long wavelength limit is

$$\hbar\omega \approx (4J_{ex}Sa) \cdot k$$

10. The heat capacity contributed by thermally excited magnons in an antiferromagnet varies with temperature as T^3. This term is in addition to the Debye T^3 term from phonons.

11. An antiferromagnetic crystal possesses no magnetic moment because of the antiparallel alignment of magnetic moments of equal value. The neutron diffraction study furnishes the size of the magnetic unit cell that is different from the chemical unit cell detected by XRD.

PROBLEMS

14.1 Show that the magnon dispersion relation for a ferromagnetic cubic lattice with nearest neighbour interactions is

$$\hbar\omega = 2J_{ex}S\left[z - \sum_{\delta} \cos(\mathbf{k} \cdot \delta)\right]$$

where the summation is over the z vectors denoted by δ that connect the central atom to its nearest neighbours.

14.2 If the effective fields on the two sublattices of an antiferromagnet are written as

$$\mathbf{H}_A = (-\gamma\mathbf{M}_B - \eta\mathbf{M}_A); \qquad \mathbf{H}_B = (-\gamma\mathbf{M}_A - \eta\mathbf{M}_B)$$

show that

$$\frac{\theta}{T_N} = \frac{\gamma + \eta}{\gamma - \eta}$$

14.3 Apply the formalism of the two-sublattice model to a ferrimagnetic solid if the magnetic moments on the two sublattices are of differing strengths.

14.4 The Brillouin function $B_J(x)$ is of the form $(Ax - Bx^3)$ for small x, where A and B are positive. Use the mean-field approximation to show that the spontaneous magnetization of a ferromagnet vanishes as $(T_c - T)^{1/2}$ as T_c is approached from below.

14.5 Use the mean field theory to derive the following relation for the exchange integral, considering only the nearest neighbour interactions:

$$J_{ex} = \frac{3k_B T_c}{2zS(S+1)}$$

where z is the number of nearest neighbours.

14.6 Assuming that iron in its metallic form has a magnetic moment of 2 Bohr magneton per atom, calculate (a) the Curie constant, (b) the Weiss constant, (c) the saturation magnetization, and (d) the magnitude of the internal field (lattice constant = 2.86 Å; T_c = 1043 K).

SUGGESTED FURTHER READING

Ashcroft, N.W. and N.D. Mermin, *Solid State Physics* (Saunders College, 1988).

Kittel, C., *Introduction to Solid State Physics*, 7th ed. (John Wiley, 1996).

Mattis, D.C., *Theory of Magnetism* (Springer, 1985).

Morrish, A.H., *Physical Principles of Magnetism* (John Wiley, 1966).

Rado, G.T. and H. Suhl (Eds.), *Magnetism* (Academic Press, 1963).

Sinha, K.P. and N. Kumar, *Interactions in Magnetically Ordered Solids* (Oxford, 1980).

Smart, J.S., *Effective Field Theories of Magnetism* (Saunders College, 1966).

White, R.M., *Quantum Theory of Magnetism*, 2nd ed. (Springer, 1983).

Chapter 15

Superconductivity

In the year 1911, Kamerlingh Onnes* observed that the DC resistance of mercury dropped suddenly to an immeasurably small value when cooled to below 4.2 K. The resistance abruptly dropped from 0.084 ohms at 4.3 K to below 3×10^{-6} ohms at 3 K which is ten-millionth of the value at 0°C. The specimen may, therefore, be regarded to have practically lost its DC resistance. The phenomenon showing complete disappearance of the DC resistance below a critical temperature T_c is known as superconductivity. Cooling below T_c is considered to cause a phase transition in the specimen from a state of normal resistivity to a superconducting state.

Several metallic elements (see Table 15.1) and alloys show superconductivity. Semiconductors like Si, Ge, Se and Te become superconductors under high pressure at low temperatures. All but one

Table 15.1 Periodic table showing superconducting elements (light shaded). The dark shaded elements show superconductivity only under pressure. The numerical figures give T_c in K

H																He
Li* 2.4	Be 0.03											B	C	N	O	F	Ne
Na	Mg* 5.5											Al 1.19	Si 6.7	P 4.5-6.1	S	Cl	Ar
K	Ca* 4.3	Sc 0.3	Ti 0.39	V 5.3	Cr	Mn	Fe* 0.3	Co	Ni	Cu	Zn 0.9	Ga 1.09	Ge 5.4	As 0.5	Se 6.9	Br	Kr
Rb	Sr* 3.6	Y 05.27	Zr 0.55	Nb 9.2	Mo 0.92	Tc 7.8	Ru 0.5	Rh 325μ	Pd† 3.2	Ag	Cd 0.55	In 3.4	Sn 3.7;5.3	Sb 3.6	Te 4.5	I	Xe 7.0
Cs 1.5	Ba 1.8;5.1	La 4.8;5.9	Hf 0.13	Ta 4.4	W 0.01	Re 1.7	Os 0.65	Ir 0.14	Pt# 0.1	Au	Hg 4.15 3.95	Tl 2.39 1.45	Pb 7.2	Bi 3.9 7.2;8.5	Po	At	Rn
Fr	Ra	Ac															

Ce 1.3	Pr* 5.0	Nd* 4.6	Pm	Sm	Eu* 3.4	Gd	Tb	Dy	Ho	Er	Tm	Yb* 2.9	Lu 0.1;0.7
Th 1.37	Pa 1.3	U 0.2	Np	Pu	Am 0.6	Cm	Bk	Cf	Es	Fm	Md	No	Lw

* Thin film

† Disordered film

Fine powders

* H. Kamerlingh Onnes, *Akad. van Westenschappen* (Amsterdam), **14,** 113, 818 (1911).

allotropic forms of Bi are superconducting with different transition temperatures. This suggests that superconductivity may depend on the crystal structure. This should not be surprising since the electronic band structure varies with the crystal structure. Significantly, none of the ferromagnetic materials (e.g. Fe, Co, Ni) displays superconductivity. The strong molecular magnetic field inside these materials is believed to be the cause, as strong magnetic fields are known to suppress superconductivity as discussed later in the chapter.

The compound superconductors generally show higher transition temperatures, a list of some selected compound superconductors is given in Table 15.2. Superconductors that have transition temperatures around 20 K or below are called conventional or low T_c superconductors and their properties are well explained by the microscopic theory of Bardeen, Cooper and Schrieffer* (famous as the BCS theory). There is a unique and extremely fascinating class of superconductors belonging to the ceramic class like $YBa_2\ Cu_3\ O_{7-x}$. These materials are distinguished by relatively much high transition temperatures (around or even more than 100 K), and are, as such, commonly known as *high* T_c superconductors. They form the subject of modern research in superconductivity. The theoretical scene in this case is, however, uncertain because it is plagued by difficulties of formulating a general theory to deal with widely different characteristics of these materials.

Table 15.2 Some selected compound superconductors with their transition temperatures T_c

Compound	T_c (K)	*Compound*	T_c (K)
$(ET)_2I_3^*$	1.4	Nb_3Al	17.5
MO_2BC	7	Nb_3Sn	18.05
La_3In	10.4	Nb_3Ge	23.2
$LuRh_4B_4$	11.7	Rb_2CsC_{60}	31.3
NbN	16	$YBa_2Cu_3O_7$	90
V_3Si	17.1	$Tl_2Ca_2Ba_2Cu_3O_{10}$	125

*ET = bis-ethylendithiotetrathiafulvalene.

Like ferromagnetism, superconductivity too is a quantum phenomenon interpreted as a quantum mechanical consequence of the electron-electron interaction. Still, some of the aspects of superconductivity can be understood by making a simple assumption that electrons contributing to superconductivity are different in behaviour from those of an ideal Fermi gas. We organize the plan of this chapter broadly on the following lines:

1. Description of phenomena in which the actual quantization conditions are macroscopically not observable.
2. Energy gap.
3. Description of properties dependent on energy gap.
4. Theoretical basis (isotope effect) for BCS theory: qualitative discussion of the theory and its quantitative predictions.
5. Basic principles of Ginzburg–Landau theory: description of phenomena effectively explained by this theory.
6. A survey of development and properties of high T_c superconductors.

* J. Bardeen, L.N. Cooper and J.R. Schrieffer, *Phys. Rev.*, **108**, 1175 (1957).

15.1 PHENOMENA WITHOUT OBSERVABLE QUANTIZATION

15.1.1 Zero Resistance and Persistent Currents

The disappearance of DC resistance at temperatures below the transition temperature is the foremost dramatic property of superconductors. This is in contrast with the behaviour of normal metallic resistivity that decreases with fall in temperature and approaches a residual value at $T = 0$ K [Fig. 15.1(a)]. The residual resistivity is attributed to the electron scattering from impurities and static imperfections (see Section 6.7).

A schematic of the fall of the DC resistivity of a superconductor is shown in Fig. 15.1(b). In pure materials, the transition at $T = T_c$ takes place very abruptly and the width of the transition region may be less than 10^{-3} K, indicating to a certain cooperative effect involving many electrons that act together. The presence of impurities results in the broadening of the transition width [Fig. 15.1(b)].

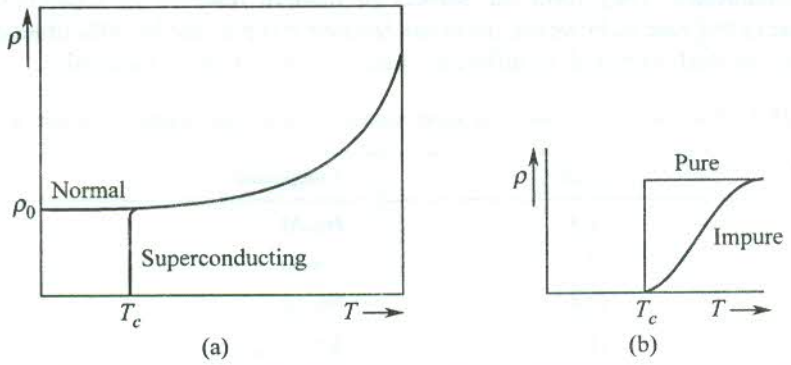


FIG. 15.1 (a) The dc resistivity ρ of a superconductor vanishes abruptly at $T = T_c$. In a normal metal, the resistivity becomes almost constant and levels off at ρ_0, the value at $T = 0$. The ρ_0 arises purely because of electron scattering from static defects. The normal state below T_c is produced by applying a magnetic field greater than a critical value. The reader must bear this in mind for other measurements to follow. (b) The superconducting transition in a pure and an impure sample is shown. In a pure specimen it is very sharp and the transition width is ~ 10^{-3} K whereas it is appreciably broadened in impure specimens.

Even with today's much improved methods of measurements it is impossible to test whether the resistance is zero; only an upper limit to the value can be set. Onnes determined the small resistances that occur below T_c by measuring the decrease in the value of current in a conducting ring. In this method, a magnetic flux is created through the ring with a magnet and the ring is then cooled below T_c, so that it goes to the superconducing state. With the withdrawal of the magnetic field, a current is induced in the ring as a result of the change in magnetic flux. Should the superconducting ring of inductance L have a finite resistance R, the current flowing in it would decline in accordance with

$$I(t) = I_0 \exp\left(-\frac{R}{L}t\right) \tag{15.1}$$

The measurements indicate that the resistance of a metal decreases by at least 14 orders of magnitude when transformed into a superconductor. Because the DC resistance of a superconductor is practically zero, there must be flowing a very large current. But the supercurrent never exceeds a critical value (the critical current) above which the superconductivity is destroyed. The size of the

critical current is typically of the order of 100 A in a 1 mm wire. This supercurrent is expected to persist forever without any need for a driving electromotive force. S.C. Collins† is credited with watching these persistent currents for a record period of $2^1/_2$ years. Experimentalists have concluded that the decay-time of supercurrents is certainly more than 10^5 years.

At $T = 0$ K, a superconductor responds without dissipation (zero resistance) even to an ac electric field up to a limited frequency. But, at a finite temperature below T_c, there is a small ac loss at all frequencies.

15.1.2 Perfect Diamagnetism: Meissner Effect

Meissner and Ochsenfeld (1933) discovered that a bulk superconductor behaves as a perfect diamagnet in a weak magnetic field by the way of showing zero induction ($\mathbf{B} = 0$) in its interior. Placing $\mathbf{B} = 0$ in relation (13.2), we obtain the value of magnetic susceptibility ($\chi = \mathbf{M}/\mathbf{H}$) as -1 (SI), characterizing a perfect diamagnet. When a superconductor is placed in a weak magnetic field, the magnetic lines of force do not penetrate the specimen but rather go round the specimen as shown in Fig. 15.2 for a superconducting sphere. The exclusion of magnetic field is nearly perfect. The penetration occurs over hardly 10^{-6} to 10^{-5} cm from the surface. This phenomenon is, however, not the true representation of the Meissner–Ochsenfeld effect (commonly known as the Meissner effect). In their original experiment, Meissner and Ochsenfeld placed a single crystal of tin in a weak magnetic field at room temperature. The magnetic field penetrated the specimen at this temperature. When they cooled the crystal to its superconducting transition temperature T_c, they detected an abrupt change in the strength of the magnetic field outside the specimen. This amounts to the magnetic field being pushed out from the interior of the specimen on attaining the superconducting state. Hence, the expulsion and not the screening of magnetic field from the interior of a superconductor is genuinely the Meissner effect.

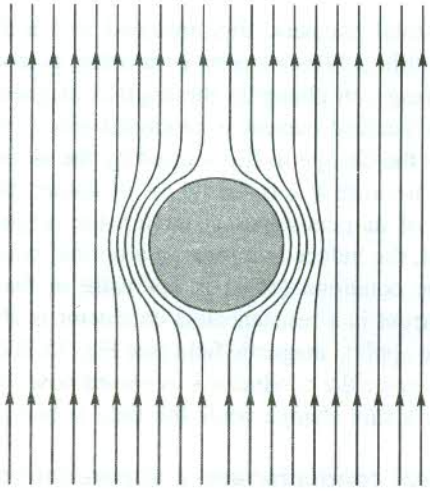


FIG. 15.2 The magnetic flux completely pushed out of the interior of a superconductor (Meissner effect). The material behaves as a perfect diamagnet.

If the magnetic field is increased, maintaining $T < T_c$, to a certain critical value of the field B_c, the superconductor reverts to its normal conducting state and the magnetic field penetrates the specimen. For magnetic field $B \geq B_c$, the specimen shows the normal conducting behaviour.

† Quoted in E.A. Lynton, *Superconductivity* (John Wiley, 1969).

Resistanceless conductor

The study of the behaviour of a resistanceless conductor in a uniform magnetic field is crucial to distinguishing between the screening of a magnetic field and the Meissner effect. Suppose a metal cylinder is placed in a magnetic field of uniform flux density (B) [see Fig. 15.3B(a)]. The total magnetic flux (Φ) enclosed within the cylinder is equal to ($B \times S$), where S is the area of the circular cross-section of the cylinder. If we change the magnetic field, a current will be induced, circulating on the curved surface in a direction so as to create a magnetic flux inside the cylinder that, in accordance with the Lenz's law, opposes the flux change caused by the change in the applied magnetic field. The e.m.f. induced in a normal resistive circuit, while the magnetic field is changing, is expressed as

$$-\frac{d\Phi}{dt} = -S\frac{dB}{dt} = Ri + L\frac{di}{dt}$$

where i is the induced current, and R and L are the resistance and the inductance of the circuit.

When we cool our metal cylinder in the presence of the magnetic field below a certain temperature (T_c) where it practically loses resistance completely, we observe that the flux remains penetrated in the cylinder [see Fig. 15.3B(b)]. The cylinder thus becomes a resistanceless conductor ($R = 0$) and the induced e.m.f. is written as

$$-S\frac{dB}{dt} = L\frac{di}{dt}$$

which, when integrated over the closed current path gives

$$Li + B \cdot S = \text{constant} \tag{15.2}$$

Relation (15.2) gives the total magnetic flux enclosed within the cylinder. It emphasizes that the magnetic flux enclosed within a resistanceless conductor, placed in a uniform magnetic field, is constant and it does not change on changing the applied magnetic field. Due to any change in the applied magnetic field, an induced current is produced whose magnitude is such that the flux generated by it exactly cancels the change in flux caused by the altering applied magnetic field. The induced current flows forever because $R = 0$. Based on its nature, the current is commonly known as persistent current. In view of its persistence character, the original value of flux is maintained indefinitely. On the other hand, the induced currents in a normal conductor quickly die away due to heat dissipation and hence the condition (15.2) is not valid in that case.

The induced persistent current in a resistanceless conductor is able to maintain the internal flux even after the withdrawal of the applied magnetic field [see Fig. 15.3B(c)]. But it is important to notice that the direction of the field outside the conductor is reversed now. It is again explained by applying Lenz's law as there will occur a flux change while the field is being switched off.

Distinction between an ideal conductor and a superconductor

Consider a conductor having been first cooled in the absence of a magnetic field to become an *ideal conductor* below T_c and then a small magnetic field $\mathbf{B}_a$ switched on in the region. Before the field is switched on, the ideal conductor ($T \leq T_c$) has no flux inside or outside its body. Therefore, in accordance with relation (15.2), the flux remains zero in its interior even after applying the field. During the short period of steadying up the magnetic field, the variation of flux induces eddy currents on the surface of the conductor. The magnetic field produced by these currents opposes the applied

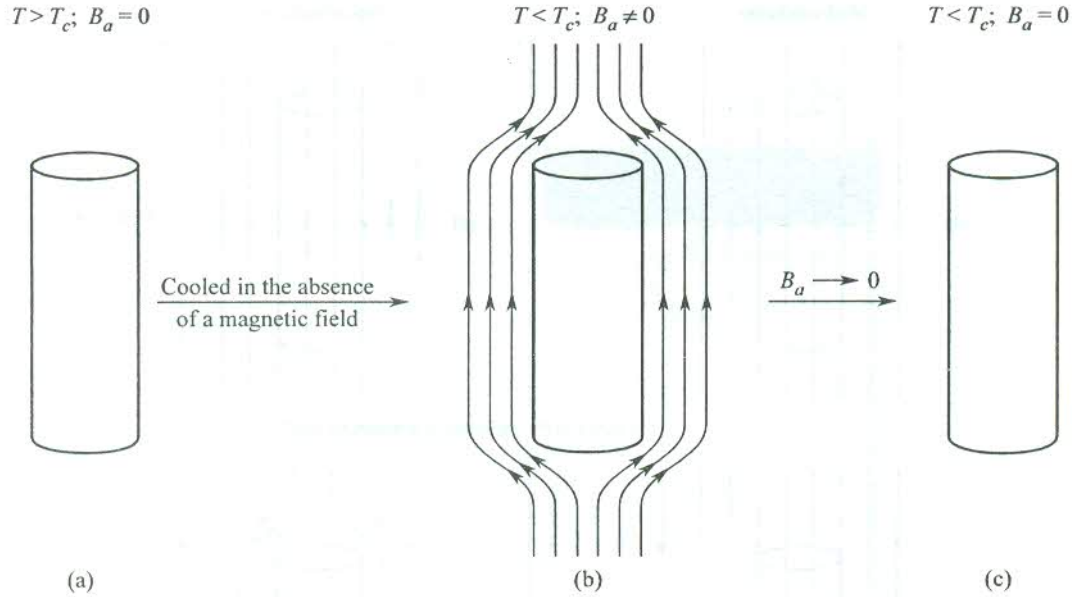


FIG. 15.3A Specimen of an ideal conductor ($R = 0$) or a superconductor cooled in the absence of a magnetic field (specimens of an ideal conductor and a superconductor behave identically under these conditions): (a) The specimen at $T > T_c$ in the absence of the field. (b) The state of the specimen in the presence of a weak magnetic field which is switched on after cooling the specimen through T_c ($T < T_c$). The magnetic flux inside the specimen is zero. (c) The state of the specimen after the magnetic field is switched off. Notice that the flux inside the specimen remains unchanged (zero) in all the states described by (a), (b) and (c).

field inside the conductor and reinforces it outside. This results in zero magnetic induction in the interior of the conductor and the lines of force do not penetrate the conductor. The condition stays on because the eddy currents do not die out and persist almost forever because of the resistanceless nature of the conductor. These persistent currents are commonly known as *screening currents.*

When the specimen of a *superconducting* material is taken through the above process, its response is identical to that of an ideal conductor [Fig. 15.3A(b)]. The eddy screening currents again account for this behaviour. This property is often exploited to shield enclosures from external magnetic fields, by using hollow superconducting cylinders.

Next, consider both the specimens to have been first placed in the region of a weak magnetic field and then cooled through T_c. At temperatures $T > T_c$, the magnetic lines of force penetrate both the types of specimens. But as the temperature decreases to $T < T_c$, the flux remains penetrated only in the specimen of the ideal conductor [see Fig. 15.3B(b)]. On cooling the specimen of the superconductor through T_c, the flux is expelled from its interior (the Meissner effect). While the requirement for the constancy of flux in an ideal conductor explains the observed behaviour in that case, there is no clue to the response of superconductors [Fig. 15.3B(e)]. There is no possibility of even screening currents as the intensity of the magnetic field is held constant throughout.

Another way to distinguish between an ideal conductor and a superconductor is to observe their response to the withdrawal of the magnetic field at temperatures $T < T_c$. If the field is turned on after the cooling, the response is identical and there is no field inside or outside of either of the specimens [see Fig. 15.3A(c)]. But, if the specimens are cooled in the presence of the field, one observes a magnetic field in and around the ideal conductor [Fig. 15.3B(c)], produced by the induced persistent

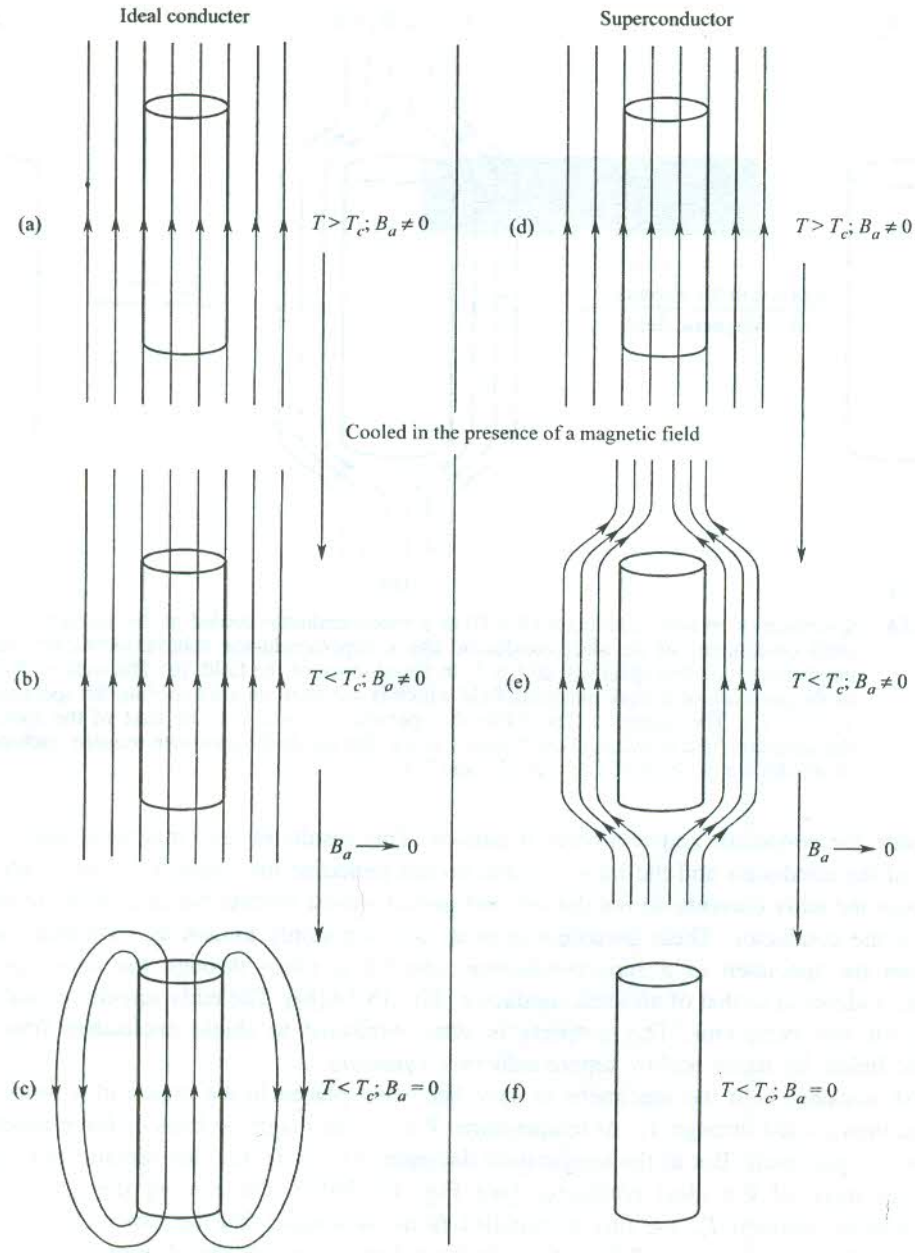


FIG. 15.3B Specimens of an ideal conductor and a superconductor cooled in the presence of a weak magnetic field: (a) The specimen of the ideal conductor at $T > T_c$. (b) The specimen of the ideal conductor when cooled through T_c ($T < T_c$). The flux penetration continues. (c) The state of the specimen of the ideal conductor after the magnetic field is switched off. The net flux inside the specimen remains the same as in (a) and (b). But the direction of the field outside the specimen is reversed. Notice that this state is different from the corresponding state A(c) where the cooling is produced in the absence of the field. (d) The specimen of the superconductor at $T > T_c$. (e) The specimen of the superconductor when cooled through T_c ($T < T_c$). The magnetic flux is expelled from the interior of the specimen (Meissner effect) on reaching T_c. (f) The specimen of the superconductor after the magnetic field is withdrawn. Notice that this state is identical to the corresponding state A(c) where the specimen is cooled in the absence of the field.

currents. Lenz's law demands that the magnetic field produced by these persistent currents support the dying applied field within the body of the ideal conductor. As a consequence the direction of the field produced outside the specimen becomes opposite to the direction of the withdrawn field. It is interesting to find the superconductor in a state identical to the one it acquired when the operations of switching on and switching off the magnetic field followed the operation of cooling through T_c [see Figs. 15.3A(c) and Fig. 15.3B(f)].

The distinction between an ideal conductor and a superconductor as brought out above and sketched in Figs. 15.3A and B clearly shows that the state of an ideal conductor after the withdrawal of the external magnetic field depends on whether the cooling through T_c is carried out in the presence or the absence of the magnetic field. On the other hand, the corresponding state of a superconductor is independent of this aspect. These discussions make it obligatory to regard the Meissner effect as an essential magnetic property of superconductors. In other words, the exhibition of perfect diamagnetism by a resistanceless material which has become resistanceless on cooling in the presence of a magnetic field is taken as the signature of superconductivity. The condition of zero resistance alone does not define superconductivity in complete sense.

15.1.3 London Equations

The zero resistance of the superconductor requires the Maxwell equation for curl **E** to change to curl $\mathbf{E} = 0 = -\dfrac{\partial \mathbf{B}}{\partial t}$ (since, $\mathbf{E} = \rho\mathbf{j}$). It gives the condition,

$$\mathbf{B} = \text{constant} \tag{15.3}$$

leading to the Meissner effect only when the constant is zero.

A phenomenological description of the Meissner effect was first given by F. London and H. London (1935). This description requires a modification of the conventional equations of electrodynamics. Their first attempt was to examine in a quantitative way the expulsion of magnetic flux from the interior of a metal in its superconducting state. Adopting the two-fluid model of Gorter and Casimir (1934) they assumed that at temperatures $T < T_c$, only a fraction n_s/n of the total number of conduction electrons carry the supercurrent. Here n_s denotes the density of the superconducting electrons (the superconducting fluid) and n is the full electron density. The remaining electrons with density $(n - n_s)$ form a 'normal fluid' through which no current can flow without dissipation. The normal current and supercurrent flow in parallel. The current should in fact be entirely the supercurrent because the resistance offered to the flow of supercurrent is zero. Normal electrons are, therefore, ignored in the present discussion.

The equation of motion of superconducing electrons accelerated by a transitory electric field **E** is

$$m\frac{d\mathbf{v}_s}{dt} = -e\mathbf{E} \tag{15.4}$$

where $\mathbf{v}_s$ is the mean velocity of electrons.

The current density of superconducting electrons is

$$\mathbf{j}_s = -n_s e\mathbf{v}_s \tag{15.5}$$

Therefore,

$$\frac{d\mathbf{j}_s}{dt} = \frac{n_s e^2}{m} \cdot \boldsymbol{E} \tag{15.6}$$

which is known as the first London equation.

Taking curl of (15.6), we have

$$\text{curl } \mathbf{E} = \frac{m}{n_s e^2} \frac{d}{dt} (\text{curl } \mathbf{j}_s) \tag{15.7}$$

Using this relation with the Maxwell equation curl $\mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$, we obtain

$$\frac{\partial}{\partial t} \left(\frac{m}{n_s e^2} \text{curl } \mathbf{j}_s + \mathbf{B} \right) = 0 \tag{15.8}$$

The magnetic fields and current densities that can exist in an ideal conductor ($\rho = 0$) are determined by the above equation together with the Maxwell equation

$$\text{curl } \mathbf{B} = \mu_0 \mathbf{j}_s \tag{15.9}$$

assuming that the rate of time variation is so slow that the displacement current can be neglected. The static magnetic field and the static current density are related to each other through (15.9). Any time-independent $\mathbf{B}$ and $\mathbf{j}_s$ could be trivial solutions of (15.8) and, therefore, relations (15.8) and (15.9) are consistent with an arbitrary static magnetic field. This describes the behaviour of an ideal conductor but not the perfect diamagnetism (i.e. the expulsion of a magnetic field) associated with the Meissner effect. If we integrate (15.8) and equate the constant of integration to zero, we get the second London equation:

$$\text{curl } \mathbf{j}_s = -\frac{n_s e^2}{m} \mathbf{B} \tag{15.10}$$

Taking the curl of (15.9), we get

$$\text{curl curl } \mathbf{B} = \mu_0 \text{ curl } \mathbf{j}_s \tag{15.11}$$

Then, using the Maxwell equation, div $\mathbf{B} = 0$ in the identity

$$\text{curl curl } \mathbf{B} = \text{grad div } \mathbf{B} - \nabla^2 \mathbf{B}$$

we get

$$-\nabla^2 \mathbf{B} = \mu_0 \text{ curl } \mathbf{j}_s \tag{15.12}$$

Substituting curl $\mathbf{j}_s$ from (15.10) into (15.12), we have

$$\nabla^2 \mathbf{B} = \frac{\mu_0 n_s e^2}{m} \mathbf{B}$$

$$= \frac{1}{\lambda_L^2} \mathbf{B} \tag{15.13}$$

with

$$\lambda_L = \left(\frac{m}{\mu_0 n_s e^2} \right)^{1/2} \tag{15.14}$$

known as the London penetration depth.

The solution of (15.13) is not uniform in space. It predicts that the magnetic field **B** can be present only within a layer of λ_L of the surface. At the same time, the condition (15.3) requires **B** to be a constant. Therefore, a constant **B** can be a solution of (15.13) only if **B** is equal to zero. The Meissner effect is thus predicted by (15.13), as obtained from the second London equation (15.10), which proves to be more restrictive than (15.8). It is, however, vital to emphasize that the above derivation of the second London equation is based on two assumptions:

(i) The superconductor is resistanceless in the presence of an applied magnetic field.

(ii) The constant of integration emerging from the integration of (15.8) is zero.

Consider, for example, a semi-infinite superconductor occupying the space on the positive side of the x-axis. Then, relation (15.13) implies that the magnetic field in the superconductor decays exponentially:

$$B(x) = B(0) \exp\left(-\frac{x}{\lambda_L}\right) \tag{15.15}$$

The magnetic field decays exponentially over a surface thickness, equalling the penetration depth, λ_L (Fig. 15.4). It is not the mere smallness of the penetrating field, but this exponential dependence, which is characteristic of the Meissner effect. Experiments on thin films as thin as or thinner than λ_L confirm that the penetration is not complete. Since the value of λ_L is found to be very small (10^{-6} to 10^{-5} cm) in most of the superconductors, the penetration law (15.15) amounts to an almost complete expulstion of the magnetic field from a superconductor of ordinary size.

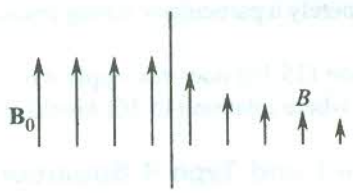


FIG. 15.4 Penetration of the external magnetic flux into a semi-infinite superconductor decreases exponentially with distance from the surface of the superconductor. The distance over which the field drops to 1/e of its value outside the superconductor is defined as the penetration depth λ.

For estimating the order of magnitude of the London penetration depth†, we put in (15.14) m equal to the electron mass and take n_s as the atomic density (i.e. we assume that each atom contributes one superconducting electron). For example, the estimate gives $\lambda_L = 260$ Å for Sn. The value of λ_L can be considerably greater near T_c where n_s approaches zero.

Diamagnetism of a new kind

The description of the static magnetic properties of most of the normal magnetic materials is based on the following relations, introduced in Chapter 13:

$$\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M}) \qquad \text{[see Eq. (13.2)]}$$

$$\mathbf{B}_0 = \mu_0\mathbf{H} \qquad \text{[see Eq. (13.3)]}$$

† Relation (15.14) does not describe λ_L precisely, since the London equations are oversimplified. According to the BCS theory the supercurrent is carried by pairs of electrons (Cooper pairs) and not by single electrons. This alters the values of m, n_s and e in (15.14).

$$\chi = \frac{\mu_0 \mathbf{M}}{B_0} = \frac{\mathbf{M}}{\mathbf{H}} \qquad \text{[see Eq. (13.5)]}$$

From these, we obtain

$$M = KB \tag{15.16}$$

with $K = \dfrac{\chi}{\mu_0(1+\chi)}$: a material constant

The persistent current $\mathbf{j}_s$ in a superconductor, placed in magnetic field, is thought to arise from a spatially non-uniform magnetization $\mathbf{M}(r)$ over the penetration depth. The current density $\mathbf{j}_s$ is related to the effective magnetization $\mathbf{M}$ by the following equation*:

$$\mathbf{j}_s = \text{curl } \mathbf{M} \tag{15.17}$$

or

$$\text{curl } \mathbf{j}_s = \text{curl curl } \mathbf{M} \tag{15.18}$$

Using (15.18) in the second London equation (15.10), we obtain

$$\text{curl curl } \mathbf{M} = -\left(\frac{1}{\mu_0 \lambda_L^2}\right)\mathbf{B} \tag{15.19}$$

It is very significant to notice that relation (15.19) gives a different dependence of $\mathbf{M}$ on $\mathbf{B}$ than that represented by relation (15.16), which is true for a normal magnetic material. This underlines the fact that superconductivity is not merely a particularly strong diamagnetism but rather a diamagnetism of a quite new kind.

It may be recalled that relation (15.16) does not apply even to ferromagnets. Superconductors are identified as another example where relation (15.16) breaks down.

15.1.4 Critical Field: Type I and Type II Superconductors

Let us consider a superconductor at a temperature $T < T_c$ in a magnetic field. As the magnetic field is increased to a certain critical value B_c, the superconducting state is destroyed. The critical field limits the amount of current that can flow in a superconductor. As soon as the field owing to the current exceeds B_c, the specimen reverts back to normal state since it is no more free of the magnetic flux. The change of state, in fact, is also advantageous on energy consideration. The value of $B_c(T)$ determines the minimum strength of the magnetic field that destroys superconductivity at a temperature $T < T_c$. It is known as the *critical field.* The value decreases as T_c is approached from below T_c and finally drops to zero at T_c (see Fig. 15.5).

On reducing the field below B_c, the specimen transforms back into the superconducting phase. Annealed samples when tested show only a negligible hysteresis, showing the reversible nature of the superconducting transition. For long thin cylindrical specimens, with the magnetic field applied along their long axis, there are two dintinct types of flux penetration:

---

* M. Abraham and R. Becker, *The Classical Theory of Electricity and Magnetism,* 2nd. ed., I, 4, 6 (Hafner, 1949). In view of the usual relationship, $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$, relation (15.17) would require the Maxwell relation curl $\mathbf{B} = \mu_0 \mathbf{j}_s$ to be replaced by curl $\mathbf{H} = 0$.

Type I. There is a certain minimum value of the magnetic field $B_c(T)$ below which there is a complete expulsion of the magnetic flux. At this value, the flux abruptly penetrates perfectly into the entire specimen, reverting the specimen to its normal state. The superconducting materials that show this behaviour are called 'Type I Superconductors'. A schematic of the temperature variation of the critical field $B_c(T)$ is shown in Fig. 15.5. The curve represents a phase boundary between the normal and superconducting states. The parabolic curve is reasonably well represented by the empirical relation:

$$B_c(T) = B_c(0)\left[1 - \left(\frac{T}{T_c}\right)^2\right] \tag{15.20}$$

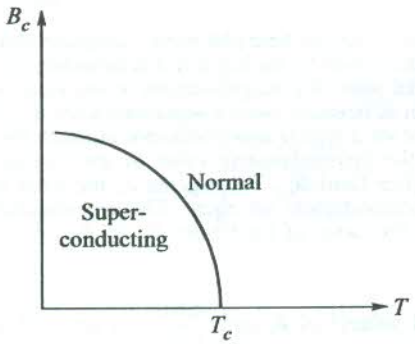


FIG. 15.5 Schematic of the variation of critical magnetic field as a function of temperature. The break near the field axis simply signifies that absolute zero cannot be approached.

The magnetization of these superconductors is illustrated in Fig. 15.6(a).

Type II. There is no penetration of the magnetic field below a certain lower critical field $B_{c_1}(T)$ and the penetration begins at this value, and grows further till an upper critical field $B_{c_2}(T)$ is reached and the penetration is complete. In the region of the partial penetration from $B_{c_1}(T)$ to $B_{c_2}(T)$, the specimen assumes a complicated mixed structure of the normal and superconducting states. The specimen is said to be in a *mixed state,* commonly known as a *vortex state* [see Fig. 15.6(b)]. Superconductors with these features are called 'Type II Superconductors'.

The values of B_c for type I superconductors are typically of the order of 10^2 gauss. These values are too low for any useful application as coils of superconducting magnets. But the so-called *hard* or type II superconductors show high values of B_{c_2}. For example, a value of 41 tesla for an alloy of Nb, Al and Ge at 4.2 K and 54 tesla for $PbMo_6S_8$ are reported in literature* (see also Table 15.3). The hard or type II superconductors have been used as solenoids in superconducting magnets that are capable of producing steady fields of over 200 kG. With these magnets it is possible to carry out the magnetic resonance imaging (MRI), an important application in medical diagnosis.

15.1.5 Thermodynamic Properties

The reversible nature of the superconducting transition enables us to apply thermodynamics for its study. If we neglect the volume changes and consider only the magnetic work term, the Gibbs free energy can be written as

* C. Kittel, *Introduction to Solid State Physics,* 7th ed., p. 340 (John Wiley, 1996).

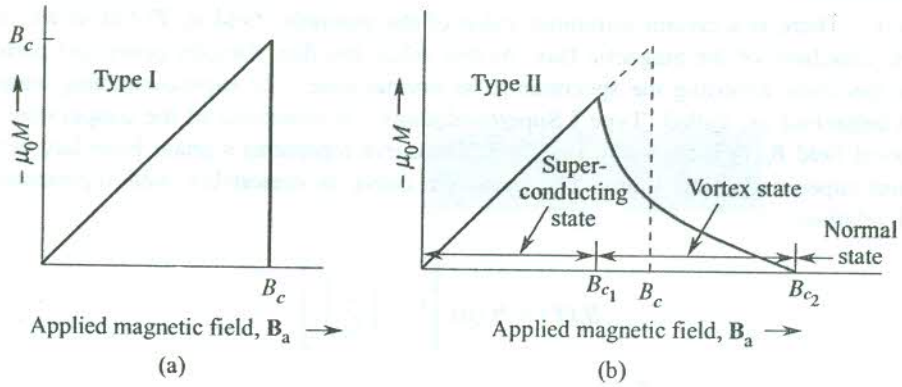


FIG. 15.6 (a) Magnetization versus magnetic field plot shows complete Meissner effect (perfect diamagnetism) for a bulk superconductor which on the basis of this behaviour is labelled as a type I superconductor. Above B_c in the normal state, the magnetization is too small to be measured on this scale. The magnetization is shown as negative since a superconductor is in the diamagnetic state. (b) It shows the magnetization curve for a type II superconductor in which the flux penetration begins at magnetic field B_{c_1}, lower than the thermodynamic value of the critical field B_c. The flux penetration is complete at much higher field B_{c_2}. For a given B_c the areas under the magnetization curves for type I and type II superconductors are equal. The superconductor is said to be in a vortex state (mixed state) between the range of field from B_{c_1} to B_{c_2}.

Table 15.3 Measured values* of B_c and $\dfrac{\Delta(0)}{k_B T_c}$ for some elemental superconductors

Element	B_c (gauss)	$\Delta(0)/k_B T_c$
Al	99	1.7
Cd	30	1.6
Hg(α)	411	2.3
In	293	1.8
Nb	1980	1.9
Pb	803	2.15
Sn	305	1.75
Ta	830	1.8
Tl	171	1.8
V	1020	1.7
Zn	53	1.6

* Sources as mentioned on pages 729 and 745 of Ref. No. 1 listed for further reading.

$$G = U - TS - \mathbf{M} \cdot \mathbf{B}_a \tag{15.21}$$

and any small change in the free energy owing to a small change in applied field $\mathbf{B}_a$ at a constant temperature is

$$dG = -\mathbf{M} \cdot d\mathbf{B}_a \tag{15.22}$$

where all extensive quantities are defined for a unit volume.

Integrating (15.22), we get

$$\int_{T,0}^{T,B_a} dG = \int_0^{B_a} \left(\frac{\mathbf{B}_a}{\mu_0}\right) \cdot d\mathbf{B}_a$$

$$\left(\text{since for a superconductor, } \mathbf{M} = -\frac{\mathbf{B}_a}{\mu_0}\right)$$

or

$$G_s(\mathbf{B}_a, T) = G_s(0, T) + \frac{B_a^2}{2\mu_0} \tag{15.23}$$

At the critical field B_c, where the normal and superconducting states are in equilibrium:

$$G_n(T) = G_s(B_c, T) \tag{15.24}$$

Here, we ignore any weak magnetism in the normal state and assume that the free energy G_n is independent of field. Then, from (15.23)

$$G_n(T) - G_s(0, T) = \frac{B_c^2}{2\mu_0} \tag{15.25}$$

The above relation shows that in zero field the superconducting state is lower in free energy by $B_c^2/2\mu_0$ per unit volume. For a typical field of 0.1 – 1.0 kG, this is 10^3 J m^{-3}. In the presence of weak fields below T_c, the specimen has to choose between gaining in energy by forcing all the magnetic flux out (retaining superconductivity) and gaining in energy by letting the flux in (going to the normal state). The superconducting state is found to be energetically favoured for small fields, but not for large fields. The experimental behaviour of free energy as a function of temperature in the two states is shown in Fig. 15.7. Making use of (15.25), we can estimate the critical field $B_c(T)$ from this graph.

The difference of entropies is determined from (15.25) with entropy S defined as $S = -(\partial G/\partial T)$:

$$S_n - S_s = -\frac{B_c}{\mu_0}\frac{dB_c}{dT} \tag{15.26}$$

Since the slope dB_c/dT of the critical field curve (Fig. 15.5) is negative, $S_n \geq S_s$, revealing that the superconducting state is a 'more ordered state' than the normal state. Also, the slope dB_c/dT approaches zero at absolute zero, leading to the result $S_n \to S_s$ as $T \to 0$, which is consistent with the requirement of the third law of thermodynamics. Figure 15.8 presents a view of the variation of entropy in the two states.

If U_n and U_s denote the internal energy in the normal and superconducting states respectively, then

$$U_n - U_s = T(S_n - S_s) \tag{15.27}$$

This gives the difference in heat capacities for a unit volume as

$$C_s - C_n = T\frac{d}{dT}(S_s - S_n)$$

$$= \frac{TB_c}{\mu_0} \cdot \frac{d^2B_c}{dT^2} + \frac{T}{\mu_0}\left(\frac{dB_c}{dT}\right)^2 \tag{15.28}$$

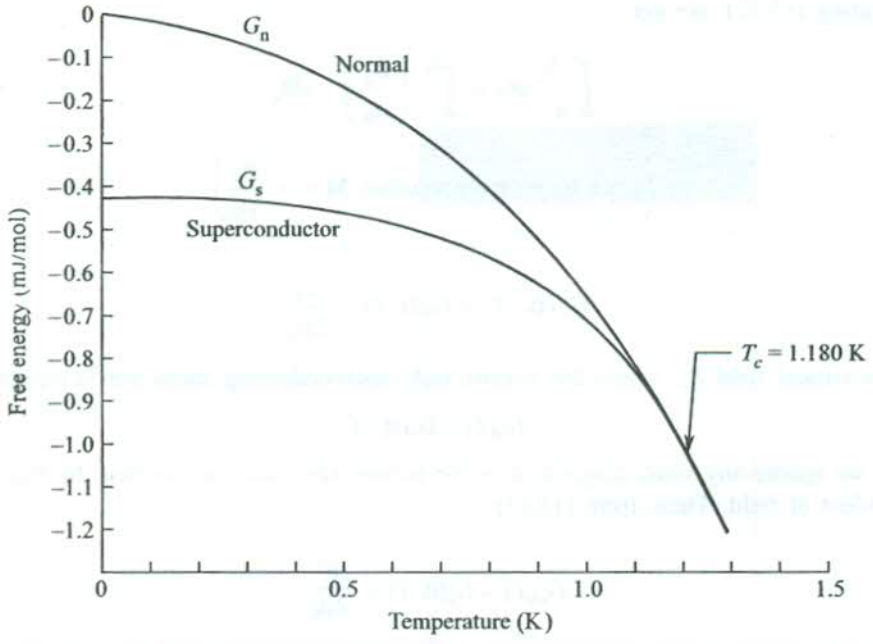


FIG. 15.7 Behaviour of Gibbs energy as a function of temperature in the superconducting and normal states of aluminium. Below the transition temperature, T_c = 1.180 K, the free energy is lower in the superconducting phase. The energies of the two phases merge at T_c. [After N.E. Phillips, *Phys. Rev.*, **114**, 676 (1959).]

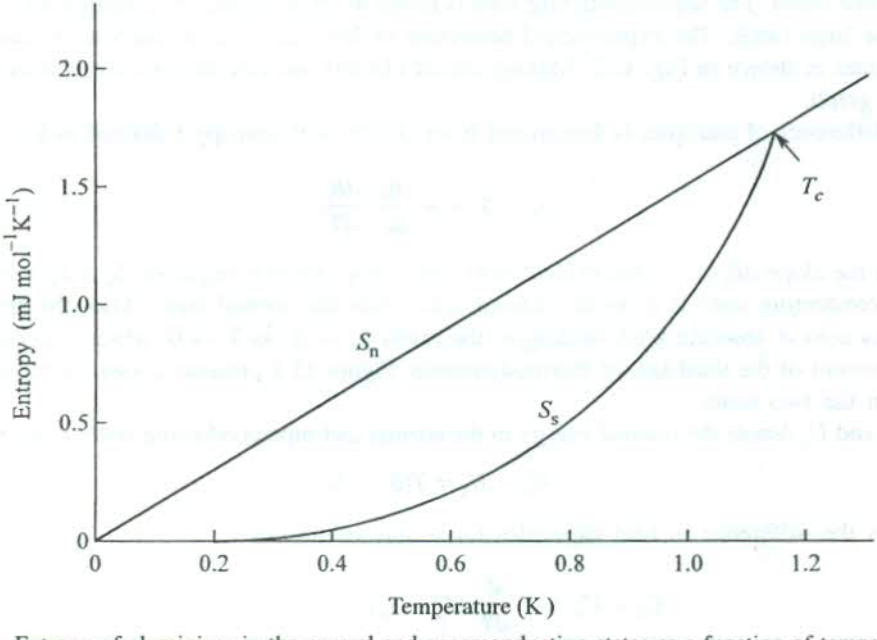


FIG. 15.8 Entropy of aluminium in the normal and superconducting states as a function of temperature. The ordering of electrons lowers the entropy of the superconducting state. [Derived from N.E. Phillips, *Phys. Rev.*, **114**, 676 (1959).]

At $T = T_c$, $B_c = 0$, and then the above relation reduces to

$$C_s - C_n = \frac{T_c}{\mu_0}\left(\frac{dB_c}{dT}\right)^2_{T=T_c} \tag{15.29}$$

This relation is known as the Rugers formula. Estimates from this formula together with the critical field curve (Fig. 15.5) agree reasonably well with the experimental data.

15.2 ENERGY GAP

The observed temperature dependence of electronic heat capacity below T_c gave evidence for a forbidden energy gap between the normal and superconducting states. At temperatures $T < T_c$, the normal state is created by applying a magnetic field that is equal to or greater than the critical field $B_c(T)$. The measured electronic contribution to the heat capacity at very low temperatures in the superconducting state is represented by a term of the form $\exp\left(-\frac{\Delta}{k_B T}\right)$ (see Fig. 15.9). This is the characteristic thermal behaviour of a system whose excited levels are separated from the ground state by an energy 2Δ. Figure 15.10(a) shows the conduction band in the normal state while Fig. 15.10(b) shows an energy gap of width 2Δ centred about the Fermi level in the set of allowed one-electron levels in the superconducting state. Electrons in the excited states above the gap behave as normal electrons. The energy gap, however, is not a universal feature of superconductivity. Under a variety of conditions superconductivity occurs even without a gap. For example, the presence of a suitable concentration of magnetic ions causes gapless superconductivity.

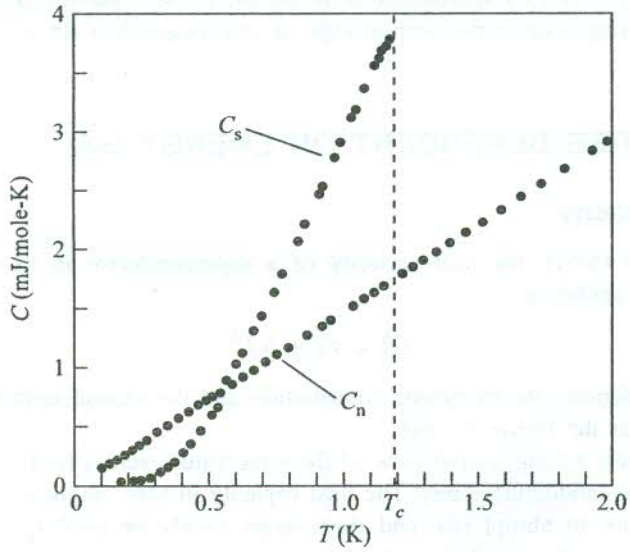


FIG. 15.9 Temperature variation of heat capacity of aluminium in its normal and superconducting states. The Debye temperature for aluminium is quite high. Therefore, electron contribution dominates in the low temperature range shown. At temperatures well below T_c, C_s falls far below C_n which indicates the existence of an energy gap. [After N.E. Phillips, *Phys. Rev.*, **114**, 676 (1959).]

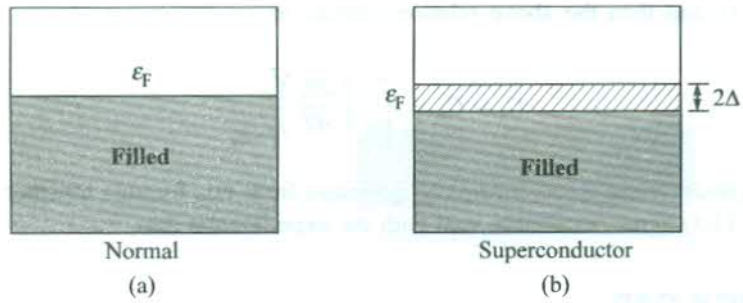


FIG. 15.10 (a) Filled Fermi sea of non-interacting normal electrons. (b) Energy gap (2Δ) centred at the Fermi level ε_F in the superconducting state. Electrons occupying excited states above the gap behave as normal electrons at radio frequency fields and a non-zero resistance occurs. Under the action of dc fields, the superconducting electrons eliminate the effect of normal electrons by current along a parallel resistanceless path.

The theoretical analysis reveals that the energy gap in a superconductor occurs due to the electron-electron interaction that orders the electrons in the k-space with respect to the Fermi gas of electrons. This energy gap is entirely of different origin and nature than the gap in insulators where it is caused by the electron-lattice interaction that ties the electrons to the lattice. Theory and experiment both suggest that the energy gap is of the order of $k_B T_c$. Nevertheless, it is found to be a function of temperature such that its size $2\Delta(T)$ increases with the fall of temperature, levelling off to a maximum value $2\Delta(0)$ ($\sim 10^{-4}$ eV) at very low temperatures. This behaviour can be verified from the plot of the reduced gap $\Delta(T)/\Delta(0)$ versus the reduced temperature T/T_c in Fig. 15.11.

The transition from the superconducting state to the normal state does not incur latent heat but gives a discontinuity in the heat capacity (Fig. 15.9). The energy gap (regarded as an order-parameter) falls continuously to zero as T_c is approached from below. These features characterize the transition from the superconducting state to the normal state in zero magnetic field as a second-order phase transition.

15.3 PROPERTIES DEPENDENT ON ENERGY GAP

15.3.1 Heat Capacity

According to relation (6.65), the heat capacity of a superconductor in its normal state at low temperatures may be written as

$$C_V^{\mathrm{n}} = \gamma T + \alpha T^3 \tag{15.30}$$

where the first term denotes the electronic contribution and the second term represents the lattice contribution, famous as the Debye T^3 term.

Figure 15.9 presents a comparative view of the temperature versus heat capacity of aluminium in its normal and superconducting states. The heat capacity in zero magnetic field at the transition temperature T_c exhibits an abrupt rise and then drops slowly on cooling further. At very low temperatures the values fall even below the normal state values, obtained in the presence of a suitable field. A comparison of values observed in the two states in this temperature range shows that the electronic contribution approaches zero much more rapidly in the superconducting state

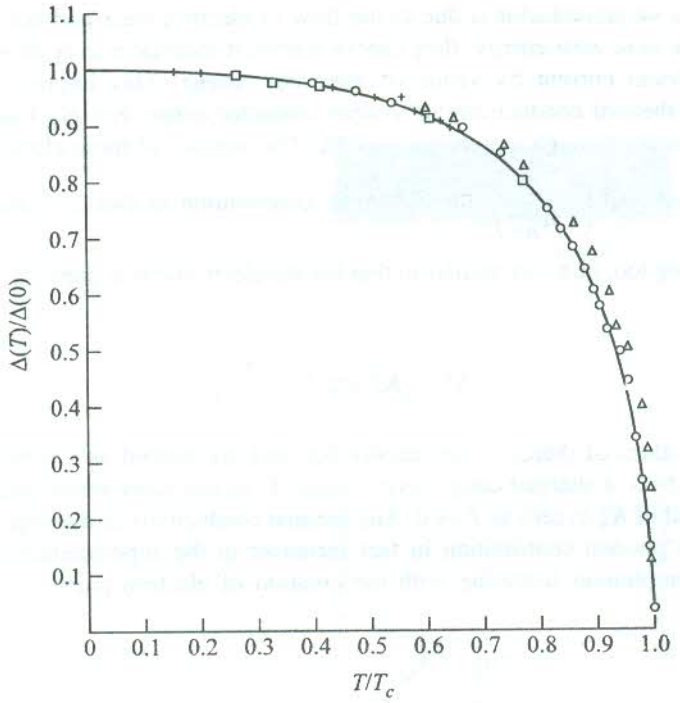


FIG. 15.11 Plot of reduced energy gap $\Delta(T)/\Delta(0)$ versus reduced temperature T/T_c, obtained from tunnelling experiments. The curve represents the BCS prediction. Experimental points: Δ for tin, $\bigcirc$ for tantalum, + for lead and $\square$ for niobium. Estimates of $2\Delta/k_BT_c$ from experimental data on these metals are 3.5, 3.6, 4.3 and 3.8 taken in that order. [After P. Townsend and J. Sutton, *Phys. Rev.*, **128,** 591 (1962).]

than in the normal state where the drop is linear. The drop can be fitted by an exponential term, entailing the expression in the superconducting state to be of the form

$$C_V^s = A \exp\left(-\frac{\Delta}{k_B T}\right) + \alpha T^3 \tag{15.31}$$

where A is a constant and 2Δ is the energy gap.

The electronic contribution represented by the first term in (15.31) depends only on the number of normal electrons. The paired electrons, considered as basis for superconductivity in the BCS model, have zero energy in the ground state and, therefore, do not contribute to the heat capacity.

15.3.2 Thermal Conductivity

Poor thermal conductivity of superconductors is yet another proof of the fact that the one-electron approximation (or the independent-electron approximation) is not applicable to superconductors. As a consequence of this approximation, good conductors of electricity should also be good conductors of heat which is really true with normal conductors. But superconductors contrast the behaviour of normal conductors by exhibiting low values of thermal conductivity, notwithstanding their excellent electrical conductivity. The interpretation to this behaviour is again found in the philosophy that the

electric current in a superconductor is due to the flow of electron pairs and not of single electrons. Since electron pairs have zero energy, they cannot transport thermal energy though their flow does constitute an electrical current by virtue of their net charge. This implies that the electronic contribution to the thermal conductivity of a superconductor comes entirely from normal electrons whose excitation occurs through the energy gap 2Δ. The number of these electrons being given by the Boltzmann factor $\exp\left(-\frac{\Delta}{k_B T}\right)$, the electronic contribution to thermal conductivity K_e^s in the superconducting state too, in a way similar to that for the electronic heat capacity, has an exponential dependence:

$$K_e^s \approx K_e^n \exp\left(-\frac{\Delta}{k_B T}\right) \tag{15.32}$$

Experimental values of thermal conductivity for lead are plotted as a function of temperature in Fig. 15.12. The drop of thermal conductivity below T_c in the superconducting state is due to the rapid exponential fall of K_e^s to zero as $T \to 0$. Any thermal conductivity in this region is predominantly from phonons. The phonon contribution in fact increases in the superconducting state because of the reduced electron-phonon scattering with the creation of electron pairs.

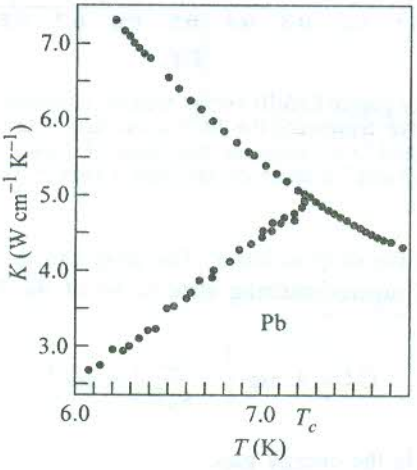


FIG. 15.12 Thermal conductivity of lead as a function of temperature. Below T_c, the upper curve represents the normal state and the lower curve represents the superconducting state. [After J.H.P. Watson and G.M. Grahm, *Can. J. Phys.*, **41**, 1738 (1963).]

In contrast to the behaviour of normal conductors, an electric current at a uniform temperature in a superconductor is not accompanied by a thermal current. This points to the absence of Peltier effect in superconductors, meaning thereby that electrons participating in the persistent current carry no entropy.

15.3.3 Absorption of Electromagnetic Radiation

The energy gap, 2Δ, determines the threshold photon energy, $\hbar\omega$, below which no absorption occurs in a superconductor. For photons with $\hbar\omega < 2\Delta$, a superconductor has zero resistance at absolute

zero. But photons of energy higher than 2Δ cause excitation to the unoccupied normal energy states above the gap and the resistance approaches its normal state value. The resistance in the superconducting and normal states is essentially the same in this case. With rise in temperature in the superconducting state, the resistance no more remains zero for $\hbar\omega < 2\Delta$, though it is still zero at zero frequency, i.e. in a dc field. The inductive component caused by the inertia of superconducting electrons at finite frequencies does not let them screen the photon electric field completely. This allows the electric field to interact with thermally excited normal electrons, leading to the absorption of incident electromagnetic radiation.

In addition to the inductive component caused by the inertia of superconducting electrons, there is also in general a resistive component of the surface impedance. This causes the power absorption to set in near absolute zero itself in the microwave region where the threshold frequency ($2\Delta/\hbar$) happens to occur. The superconducting transition becomes finally undetectable in the infrared region, as shown in Fig. 15.13 for aluminium.

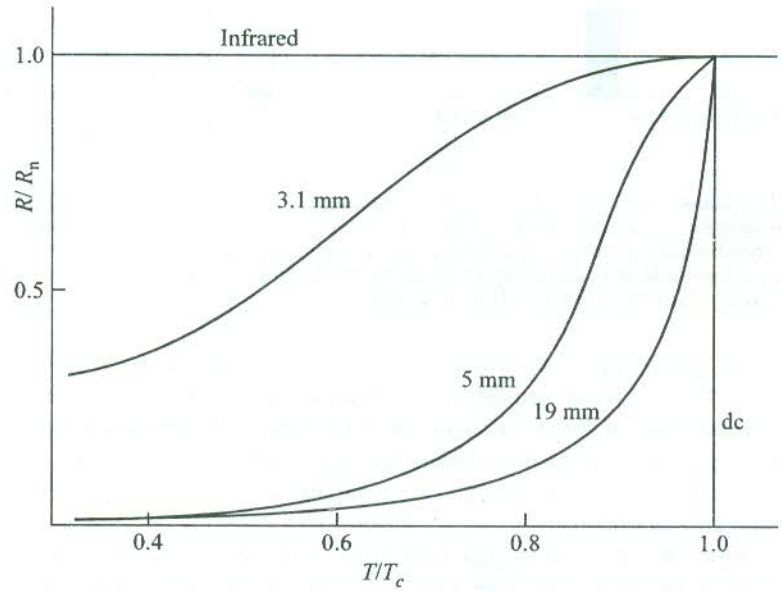


FIG. 15.13 Ratio of the resistance in the superconducting state to that in the normal state as a function of the reduced temperature T/T_c for aluminium in the microwave region. In the infrared region, the superconducting and normal states are not distinguishable ($R/R_n = 1$). [After M.A. Biondi and M.P. Garfunkel, *Phys. Rev. Lett.*, **2**, 143 (1959).]

15.3.4 Normal Tunnelling

The idea of electron tunnelling was introduced in Section 9.10.1 where it was shown as the basis for working of a tunnel diode. A similar tunnelling is known to take place between two metals closely separated by a thin insulating layer (< 10–20 Å) which the electrons can cross by quantum mechanical tunnelling. On applying a difference of potential, the chemical potential of one metal is raised with respect to the other and more electrons tunnel through the insulating film, such that the voltage–current characteristics obey Ohm's law [Fig. 15.14(a)]. Giaever (1962) observed that on replacing one of the metals by a superconductor well below its transition temperature, the voltage–

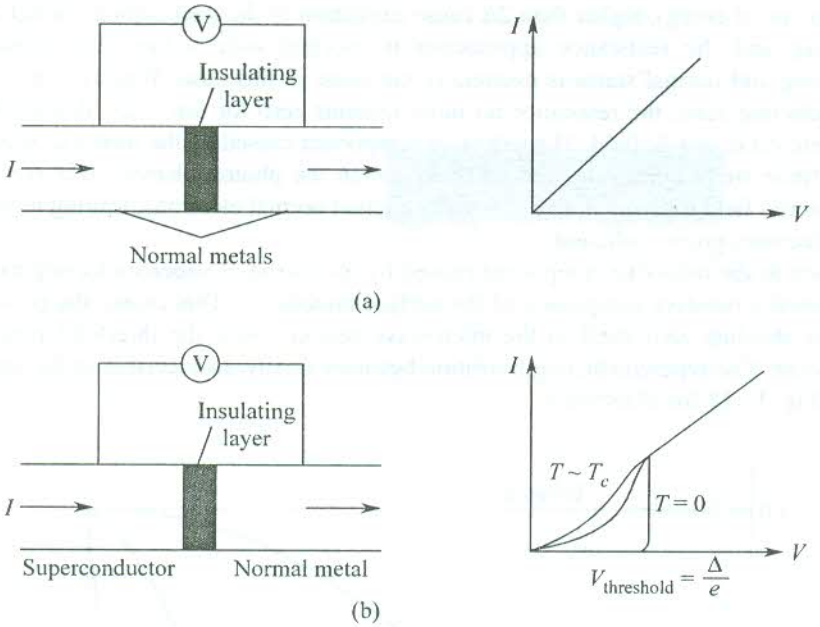


FIG. 15.14 (a) *V–I* characteristic curve for electron tunnelling from one metal through a thin insulating barrier into another metal. For small voltages, the behaviour is linear. (b) *V–I* characteristic curve for electron tunnelling from a superconductor through a thin insulating barrier into a metal. It shows a sharp threshold at $T = 0$; at higher temperatures because of thermal excitation of electrons across the energy gap, the blurring effect is evident.

current characteristic appears not as a straight line but as a curve shown in Fig. 15.14(b). At absolute zero no current can flow until the applied potential difference reaches a threshold value $V = 2\Delta/2e = \Delta/e$. At finite temperatures below T_c, a small current is observed even below this voltage because of electrons in the superconductor getting excited through the energy gap. It is an easier experimental method to determine the energy gap. The gap is essentially a measure of the excitation energy that is needed to break an electron–pair.

Josephson's discovery of electron–pair tunnelling between two superconductors sandwiching a thin insulating layer added an altogether new dimension to the applications of superconductors. It is, however, appropriate to defer its description at this stage for developing the theoretical ideas that would enable us to conduct a complete discussion of properties without any repetition. The BCS theory enjoys the status of a successful theory with regard to the interpretation of the properties of liquid helium superconductors. An account of the salient features of this theory follows the description of isotopic effect that gave clue to the basic tenet of the theory.

15.4 ISOTOPE EFFECT

Realizing that solids in which the normal conductivity is not large are more prone to become superconductors, Fröhlich (1950) proposed a model for superconductivity based on electron–phonon interaction. The model assumes that an electron while moving through a crystal lattice continuously emits and reabsorbs virtual phonons. By analogy with his work on polarons (introduced in Section 10.10), Fröhlich pleaded that an electron deforms the lattice in its surrounding

neighbourhood, causing it to oscillate. The electron is then considered to interact with the lattice through an electron–phonon interaction that might produce a ground state of energy lower than the completely filled Fermi sea of non-interacting electrons. The lower ground state, identified with the superconducting state, would then be separated from the normal conducting state by an energy gap. Though Fröhlich's efforts to treat the electron–phonon interactions by perturbation theory failed to make any headway, he was able to suggest on the basis of these interactions that the transition temperature T_c should be proportional to the Debye characteristic temperature θ_D for the isotopes of a given element. The suggestion implies that $T_c \propto M^{-1/2}$, where M is the isotopic mass. This phenomenon, known as the *isotope effect,* serves as an invaluable link in the theory of superconductivity.

Two independent experimental studies*,† on isotopes of mercury, about which Fröhlich was unaware, supported his contention. The T_c in mercury changes from 4.185 to 4.146 K for a variation in the average mass M from 199.5 to 203.4 amu. The isotope effect in its generalized form that fits experimental data (see Fig. 15.15) is expressed by

$$T_c M^{\alpha} = \text{constant} \tag{15.33}$$

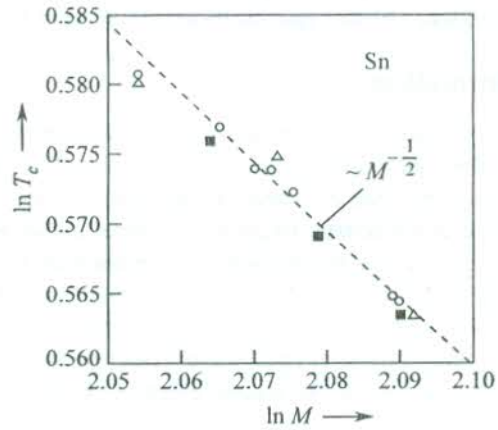


FIG. 15.15 Isotope effect for tin as observed by several workers: Maxwell (○); Lock, Pippard, Schoenberg (■); Serin, Reynolds and Lohman (Δ). [After E. Maxwell, *Phys. Rev.,* **86,** 235 (1952); B. Serin, C.A. Reynolds, C. Lohman, *Phys. Rev.,* **86,** 162 (1952); J.M. Lock, A.B. Pippard, D. Schoenberg, *Proc. Cambridge Phil. Soc.,* **47,** 811, (1951).]

Mostly, the experimental value of α is about (1/2) with Nb_3Sn ($\alpha = 0.08 \pm 0.02$) and Os ($\alpha = 0.15 \pm 0.05$) as the extreme cases. With Coulomb interaction between electrons not taken into consideration, the BCS theory too predicts a value of (1/2). The case of tin is an example in which the BCS–$M^{1/2}$ dependence of T_c is in fine agreement with experiment. The isotope effect is, however, not a universal feature of superconductors. Even some of the liquid helium superconductors (Ru, Zr) to which the BCS is properly applicable do not show isotope effect. The absence of isotope effect in these cases is explained on the basis of their electron band structure.

The isotope effect revealed that electron–phonon interactions are deeply involved in superconductivity. It gave the basic clue to the modelling of the BCS theory.

---

* E. Maxwell, *Phys. Rev.,* **78,** 477 (1950).

† C.A. Reynolds et al., *Phys. Rev.,* **78,** 487 (1950).

15.5 BCS THEORY: A QUALITATIVE APPROACH

The microscopic theory of superconductivity, famous as the BCS theory, is attributable to Bardeen, Cooper and Schrieffer.*

In this section, we describe the basic principles of this theory in a qualitative way. The all important clue to this theory came from the isotope effect which indicated that the vibrational motion of heavy nuclei must play an essential role in the mechanism of superconductivity. The missing link was established by Cooper. He remodelled Fröhlich's ideas on electron–phonon interaction into the philosophy of an electron–phonon–electron interaction which is now confirmed as the cause of superconductivity in liquid helium superconductors. Cooper demonstrated that with the creation of a condition favourable for a net attractive interaction between two electrons in a conductor, the conductor is transformed from its normal conducting state to its superconducting state. The current in a superconductor is attributed to the flow of the band electron pairs, called *Cooper pairs.* At finite temperatures below the transition temperature, the Cooper pairs find themselves immersed in the Fermi sea of non-interacting electrons. At absolute zero all the electrons in states near the Fermi surface are bound into Cooper pairs and are in the ground state. The existence of Cooper pairs is strongly implied by the experimental evidence of flux quantization [see Section 15.7.1(a)].

15.5.1 Cooper Pair Formation

The theory of superconductivity requires a net attractive interaction between electrons near the Fermi surface. Cooper showed that the ionic motion can overscreen the Coulomb repulsion between two electrons to produce a net attraction if the difference in energies of the two electrons is of the order of $\hbar\omega_D$, where ω_D is the Debye characteristic frequency. Cooper explained the formation of bound electron pairs on the basis of an electron–phonon–electron interaction which materializes through a sequence of events. An electron deforms the lattice in its vicinity, exciting a phonon that travels through the crystal. This phonon is absorbed by a second electron, getting thus coupled to the first electron. Let $\mathbf{k}_1$ and $\mathbf{k}'_1$ denote the initial and final wavevectors of the first electron, and $\mathbf{k}_2$ and $\mathbf{k}'_2$ denote the initial and final wavevectors of the second electron. Then, the momentum conservation requires,

$$\mathbf{k}_1 - \mathbf{k}'_1 = \mathbf{k}_2 - \mathbf{k}'_2 = \mathbf{k}_{ph} \tag{15.34}$$

where $\mathbf{k}_{ph}$ is the wavevector of the phonon involved in the process.

It must be, however, observed that the condition of energy conservation is met for the crystal on the whole and not for the electron pair because the phonon involved in the exchange is a virtual phonon.

It is easy to show that pairs with zero momentum are most likely to be formed. Consider two electrons with wavevectors $\mathbf{k}_1$ and $\mathbf{k}_2$ on the Fermi sphere (Fig. 15.16) so that $k_1 = k_2 = k_F$. The total momentum of the pair is $\mathbf{K} = \mathbf{k}_1 + \mathbf{k}_2$. When $\mathbf{K} \neq 0$, its magnitude is equal to $2k_F \cos\theta$ and only those electrons that lie on the ring shown in the figure can form pairs. On the other hand, for $\mathbf{K} = 0$, i.e. $\mathbf{k}_1 = -\mathbf{k}_2$, all electrons on the Fermi surface can be paired because of the finite phonon energy $\sim \hbar\omega_D = \dfrac{\hbar^2 k_F \Delta k}{m}$. These electrons are confined in the thickness Δk of the shell on the Fermi sphere. But this does not affect the inference that pairs with $\mathbf{K} = 0$ have the maximum chance of formation.

---

* J. Bardeen, L.N. Cooper and J.R. Schrieffer, *Phys. Rev.*, **108**, 1175 (1957).

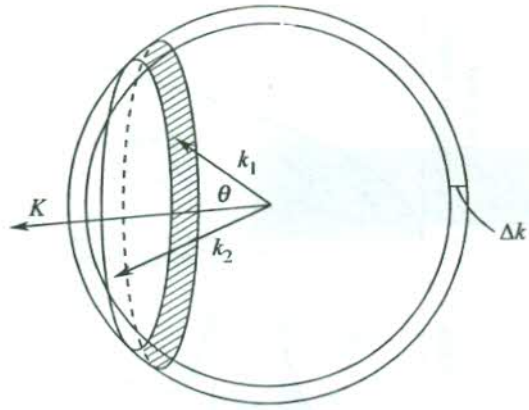


FIG. 15.16 A pair of electron states $\mathbf{k}_1$ and $\mathbf{k}_2$ on a shell of a spherical Fermi surface. When $\mathbf{K} = \mathbf{k}_1 + \mathbf{k}_2 = 0$, all states can be paired.

Capitalizing on Cooper's ideas, Bardeen, Cooper and Schrieffer developed a complete microscopic theory according to which favourable conditions for a condensation into bound pairs can exist between electrons having exactly opposing values of wavevectors and opposing spins, i.e. $\mathbf{k}\uparrow$ and $-\mathbf{k}\downarrow$, which is consistent with the condition $\mathbf{K} = 0$.

We find it rather difficult to appreciate that the attractive interaction between any two electrons differing in energy roughly by $\hbar\omega_D$ would be strong enough to cause pairing, especially when the electrons were in isolation. Cooper was able to demonstrate that in the presence of Fermi sphere of additional electrons, the Pauli exclusion principle significantly modified the two-electron problem so that the net attraction need not have a minimum strength to bind a pair. He could also justify very low superconducing transition temperatures compared to other characteristic temperatures of solids. His calculations showing that the binding energy is too small in comparison with the potential energy of attraction, when the attraction is weak, form the basis of these observations.

15.5.2 BCS Ground State

A Fermi gas of non-interacting electrons in its ground state is entirely confined within the Fermi sphere. For exciting the ground state electrons, little energy is required as they need be raised just above the Fermi surface determining the highest populated level. Bardeen, Cooper and Schrieffer extended Cooper's theory to hypothesize a ground state in which all electrons form bound pairs. With an appropriate attractive interaction, this ground state is preferred to correspond to the superconducting state which is separated from its lowest excited state by a finite energy 2Δ.

Figure 15.17 presents a simple visualization of the normal state at $T = 0$, and that of the superconducting state at $T = 0$ and at temperatures in the range $0 < T < T_c$. It is, however, not an exact depiction of the BCS ground state because the illustration has been derived from the Fermi–Dirac representation of density of states which is not applicable to electron pairs behaving statistically as bosons by virtue of their spin value. Nevertheless, this approximate picturization is very helpful in appreciating how in some experiments (e.g. Giaever tunnelling) the density of states are scanned with respect to energy. Figure 15.17(b) shows a zero density of states within an energy range 2Δ centred at the Fermi level ε_F and a piling of the displaced states on either side of the gap. It must be noticed that all electrons form Cooper pairs at $T = 0$ because of zero electron occupancy above

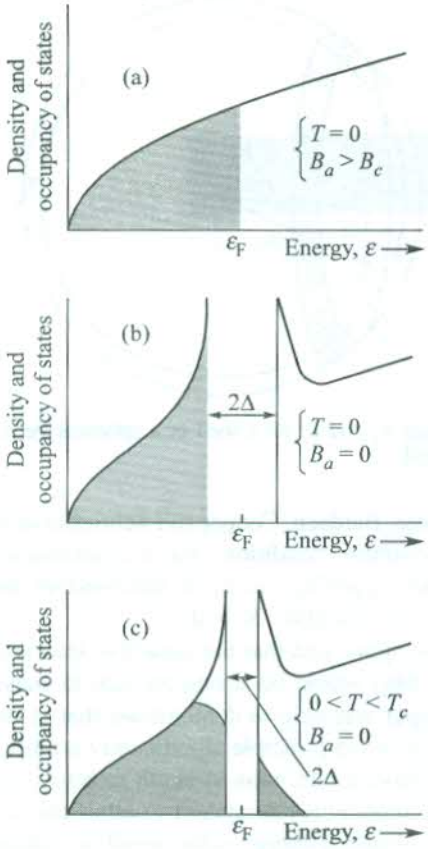


FIG. 15.17 A comparative view of the density of states for a normal conductor (a) and a superconductor [(b), (c)]. In (c), for $0 < T < T_c$ some Cooper pairs break. The released electrons cross the reduced gap.

the gap. At finite temperatures below T_c, the gap decreases [Fig. 15.17(c)] and the Cooper pairs coexist with the non-interacting electrons released by Cooper pairs that break up because of the thermal excitation across the gap.

The knowledge of the ground state wavefunction is of central importance in making theoretical predictions. It is approximated in a manner described below. In a system of N electrons, the electrons are grouped into $N/2$ bound pairs with the state of each pair being described by a wavefunction $\psi(\mathbf{r}\, s, \mathbf{r}'\, s)$, where $\mathbf{r}$ is the electronic position and s is the spin quantum number. The product of these $N/2$ identical two-electron wavefunctions may be used to describe the N-electron wavefunction:

$$\phi(\mathbf{r}_1 s_1, \mathbf{r}_2 s_2, \ldots, \mathbf{r}_N s_N) = \psi(\mathbf{r}_1 s_1, \mathbf{r}_2 s_2) \ldots, \psi(\mathbf{r}_{N-1}\, s_{N-1}, \mathbf{r}_N s_N) \tag{15.35}$$

This represents a state in which all electrons are present as identical bound electron-pairs. Since the wavefunction ϕ as expressed by (15.35) lacks the symmetry required by the Pauli principle, we must antisymmetrize it so that with the interchange of the space and spin coordinates of any two electrons the antisymmetrized wavefunction changes its sign. Such a wavefunction represents the BCS ground state:

$$\phi_{BCS} = \mathbf{a}\phi \tag{15.36}$$

where **a** is the antisymmetrizer.

The wavefunction ϕ_{BCS} satisfies the Pauli principle, though it is constructed from identical pair function ψ. In contrast to the case of single-electron states, the requirement of antisymmetry does not imply any restriction on the occupancy of two-electron levels in states which are antisymmetrized products of two-electron levels. It shows why it is possible for a pair of fermions to behave statistically like a boson, thus allowing all electron-pairs to condense coherently into a single quantum state.

15.6 IMPORTANT PREDICTIONS OF THE BCS THEORY AND COMPARISON WITH EXPERIMENT

In a quick display of success of the BCS theory, the actual net electron–phonon–electron interaction near the Fermi surface is grossly simplified to a translationally invariant effective interaction U, with conduction electrons being treated in the free-electron approximation. The matrix element of U between two-electron states $\psi(\mathbf{k}_1, \mathbf{k}_2)$ and $\psi(\mathbf{k}_3, \mathbf{k}_4)$ over a volume V is expressed as

$$\langle \psi(\mathbf{k}_1, \mathbf{k}_2) \, |U| \, \psi(\mathbf{k}_3, \mathbf{k}_4) \rangle = -\frac{U_0}{V} \tag{15.37}$$

with

$$\mathbf{k}_1 + \mathbf{k}_2 = \mathbf{k}_3 + \mathbf{k}_4$$

and

$$| \varepsilon(\mathbf{k}_i) - \varepsilon_F | < \hbar\omega_D; \qquad i = 1, 2, 3, 4.$$

The matrix element vanishes if these conditions of energy and momentum are not met. For a net attraction, therefore, all the four electron energies are required to have values within an amount $\hbar\omega_D$ of the Fermi energy. This puts a restriction on the values of wavevectors involved in a translationally invariant potential.

The most striking feature of the BCS theory is that it yields certain relations in which its two phenomenological parameters U_0 and ω_D do not figure. Barring a few, such as lead and mercury, to all other liquid helium superconductors these relations apply reasonably well. The exceptions mark their identity with a special class known as 'strong coupling superconductors'. The theory explains the properties of even these superconductors, when applied more rigorously by doing away with the simplifications inherent in (15.37) and using an improved version of phonon effects. We discuss below some of the main quantitative predictions of the theory.

Coherence of the BCS state also follows from the fact that a large number of Cooper pairs are found to be interlocked over the spatial extent of the BCS wavefunction. It is only in the energy range from $(\varepsilon_F - 2\Delta)$ to $(\varepsilon_F + 2\Delta)$ that one-particle occupation in the superconductor is modified compared to that of a normal conductor. The one-particle wavefunctions that constitute the identical two-particle wavefunctions, therefore, originate from this region. This concept helps us in making a crude estimate of the spatial extension of ϕ_{BCS} in the following way.

If we take the energy uncertainty for electrons in Cooper pairs as $2(2\Delta)$,

$$2(2\Delta) \sim \delta\left(\frac{p^2}{2m}\right) \simeq \frac{p_F}{m}\,\delta p$$

where δp denotes the momentum uncertainty.

Using the uncertainty relation, the spatial extension of ϕ_{BCS} is obtained as

$$\xi_0 \sim \delta x \sim \frac{\hbar}{\delta p} \simeq \frac{\hbar p_F}{m \cdot 2(2\Delta)}$$

$$= \frac{\hbar^2 k_F}{2m \cdot (2\Delta)}$$

$$= \frac{1}{k_F}\left(\frac{\varepsilon_F}{2\Delta}\right)$$

Since $\varepsilon_F/2\Delta$ is typically of the order of 10^3 or 10^4 and $k_F \sim 10^8$ cm^{-1}, $\xi_0 \approx 10^3$ to 10^4 Å.

This magnitude turns out to be of the order of the Cooper pair size, interpreted as the intrinsic coherence length. With $\varepsilon_F/2\Delta \approx 10^4$, we can assume that out of 10^{23} electrons per cm^3 about 10^{19} form Copper pairs. Therefore, within the space occupied by a single Cooper pair ($\sim 10^{-12}$ cm^3), there lie centres of gravity of about 10^6 to 10^7 other Cooper pairs. It is thus highly misleading to view Cooper pairs as independent bosons. They are in fact spatially anchored to one another in a very intricate manner. This leads to high coherence and stability of the many-body BCS state, analogical to the photon coherence in a laser beam. Coherence of the BCS state makes it possible for some of the quantum mechanical effects (e.g. flux quantization) to become macroscopically observable.

15.6.1 Critical Temperature

We discussed the superconducting transition more often in terms of the transition temperature which as per the terminology of phase transition should have been mentioned as the critical temperature T_c. On heating a superconductor to its critical temperature, the superconductivity is destroyed; the energy gap vanishes and the conductor reverts back to its normal conducting state. For $\Delta = 0$ and $U_0 D(\varepsilon_F) << 1$, the BCS calculations give

$$k_B T_c = 1.14\ \hbar\omega_D \exp\left(-\frac{1}{U_0\, D(\varepsilon_F)}\right) \tag{15.38}$$

where $D(\varepsilon_F)$ denotes the density of electronic levels of a single spin population at the Fermi energy in the normal conducting state. The interaction parameter U_0 is usually estimated from the electrical resistivity which at room temperature is essentially a measure of the electron-phonon interaction. Relation (15.38) most interestingly reveals that the higher the resistivity (amounting to the higher U_0), the higher is the likelihood for the concerned metal or alloy to become a superconductor on cooling.

The exponential dependence in (15.38) accounts for the observed very low T_c values that are typically one to three orders of magnitude below the Debye characteristic temperature θ_D. It implies that no matter how weak the coupling U_0, the theory predicts a finite T_c, though it may be even unobservably low in the event of U_0 being very small. The T_c values estimated from (15.38) by and large agree with those observed experimentally. When the T_c value is known, it is possible to evaluate $U_0 D(\varepsilon_F)$ which is regarded as the superconducting coupling constant.

15.6.2 Energy Gap

The BCS theory confirms the existence of an energy gap (2Δ) in the excitation spectrum of a superconductor. As discussed earlier in Section 15.3, it forms the basis for numerous features of superconductors. The experimental methods based on electron tunnelling and optical absorption are usually employed to measure the energy gap.

An expression for the gap at absolute zero, similar to (15.38), is predicted by the theory as

$$\Delta(0) = 2\,\hbar\omega_D \exp\left(-\frac{1}{U_0 D(\varepsilon_F)}\right) \tag{15.39}$$

with $U_0 D(\varepsilon_F) << 1$.

Dividing (15.39) by (15.38), we obtain a basic relation from which the phenomenological parameters ω_D and U_0 are absent:

$$\frac{\Delta(0)}{k_B T_c} = 1.76 \tag{15.40}$$

Data in Table 15.3 indicate that this estimate is valid in respect of most of the superconducting elements within about 10 per cent. Large deviations in the case of lead and mercury (~ 30 per cent) should be viewed together with disagreements on other predictions of the simple theory for strong coupling superconductors in general. By abandoning the gross oversimplifications in the interaction Hamiltonian (15.37) and using a more elaborate analysis, the deviations are substantially reduced and the scope of the BCS theory widened. The properties of even strong coupling superconductors are well interpreted in the framework of this broad-based theory.

The simplified BCS theory yields another universal law, describing the temperature dependence of the energy gap near T_c in zero magnetic field:

$$\frac{\Delta(T)}{\Delta(0)} = 1.74\left(1 - \frac{T}{T_c}\right)^{1/2}, \qquad T \approx T_c \tag{15.41}$$

The BCS plot of the above relation together with the experimental points for certain elemental superconductors is shown in Fig. 15.11. The success of the simplified theory in accounting for the continuous drop of the gap to zero at T_c is very much in evidence.

15.6.3 Critical Field

We refer to relation (15.25) which shows that the difference between the free energies per unit volume of the normal and superconducting states is equal to $B_c^2(T)/2\mu_0$, where $B_c(T)$ is the critical magnetic field that destroys superconductivity at a temperature T (of course below T_c). This difference in energy arises since, on the formation of a gap, there occurs a change of energy by 2Δ for a number $\Delta D(\varepsilon_F)$ of electrons. Using (15.25), the calculation for a type I superconductor gives

$$\frac{B_c^2(T)}{2\mu_0} = \frac{1}{2} D(\varepsilon_F)\Delta^2 \tag{15.42}$$

Fetching the value of $D(\varepsilon_F)$ from the expression for the electron heat capacity† and then eliminating the Boltzmann constant with the use of (15.40), we find that near absolute zero

$$\frac{B_c^2(T)}{\mu_0 T_c^2} \approx \frac{1}{2}\gamma \tag{15.43}$$

where γ is the Sommerfeld constant defined as the ratio of the electron heat capacity to the temperature of measurement ($= C_{el}/T$). It may be checked that γ is a function of $D(\varepsilon_F)$ which might differ from metal to metal.

The BCS result (15.43) holds good for a large number of superconductors whose T_c values are spread over a fairly wide range. The simple BCS prediction on $B_c(T)$ is often tested in terms of deviation from the empirical law (15.20), which we rewrite below:

$$\frac{B_c(T)}{B_c(0)} = 1 - \left(\frac{T}{T_c}\right)^2$$

The deviation

$$\frac{B_c(T)}{B_c(0)} - \left[1 - \left(\frac{T}{T_c}\right)^2\right]$$

from the above law is plotted in Fig. 15.18 along with the corresponding quantity as predicted by the BCS theory for a number of superconductors. The limitation of a simplified BCS theory to deal with the strong coupling superconductors is reflected by the large deviations appearing for lead and mercury.

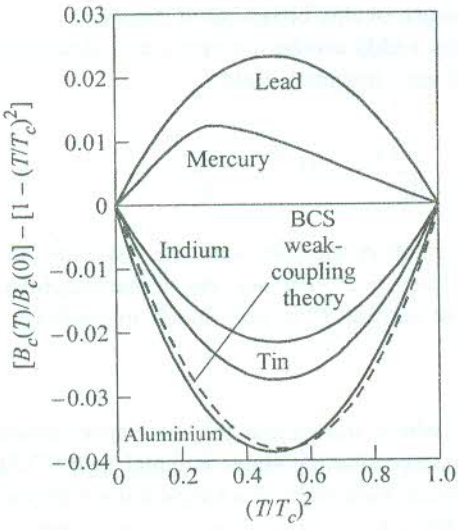


FIG. 15.18 Comparison between experiment and theory (BCS) on the measure of quantity $\left[\frac{B_c(T)}{B_c(0)} - \left(1 - \frac{T^2}{T_c^2}\right)\right]$, representing deviation from the crude empirical relation (15.20) in several metals.

† From (6.62) and (6.65) $C_{el} = \frac{\pi^2}{3} D(\varepsilon_F)\ k_B^2 T = \gamma T$.

15.6.4 Heat Capacity

The relation for heat capacity as obtained in the framework of the elementary BCS theory supports the observed discontinuity at the critical temperature in zero magnetic field. It can be cast in a form independent of the phenomenological parameters of the interaction Hamiltonian (15.37):

$$\left.\frac{C_s - C_n}{C_n}\right|_{T=T_c} = 1.43 \tag{15.44}$$

This estimated ratio is in good agreement with the measured value for a great number of superconductors. Similar to other predictions, for this one also there is an appreciable departure in the case of strong coupling superconductors (e.g. the measured values of the ratio for mercury and lead are 2.4 and 2.7, respectively).

The simple BCS theory also predicts the exponential drop of the electron heat capacity at low temperature as actually revealed by experiments. The parameter independent form of the prediction is

$$\frac{C_{el}}{\gamma T_c} = 1.34 \left(\frac{\Delta(0)}{T}\right)^{3/2} \exp\left(-\frac{\Delta(0)}{T}\right) \tag{15.45}$$

15.6.5 Acoustic Attenuation

The propagation of a sound wave in a metal causes displacement of ions as a consequence of which microscopic electric fields are produced. These fields are capable of exciting electrons near the Fermi level. The energy is thus removed from the sound wave, resulting in its attenuation. But the attenuation of a sound wave is drastically reduced in the superconducting state because the energy carried by the wave is less than the energy required for the promotion of electrons to their lowest excited state (i.e. $\hbar\omega < 2\Delta$). The BCS theory predicts a reduction in the attenuation coefficient, α, of acoustic waves. The ratio of the coefficients in the superconducting and normal states is expressed as

$$\frac{\alpha_s}{\alpha_n} = \frac{2}{1 + \exp\left(\dfrac{\Delta}{k_B T}\right)} \tag{15.46}$$

At low temperatures, the above ratio reduces to

$$\frac{\alpha_s}{\alpha_n} = 2 \exp\left(-\frac{\Delta}{k_B T}\right) \tag{15.47}$$

This ratio for elemental superconductors is plotted as a function of temperature in Fig. 15.19. It can be observed that the experimental plots compare reasonably well with the BCS curve.

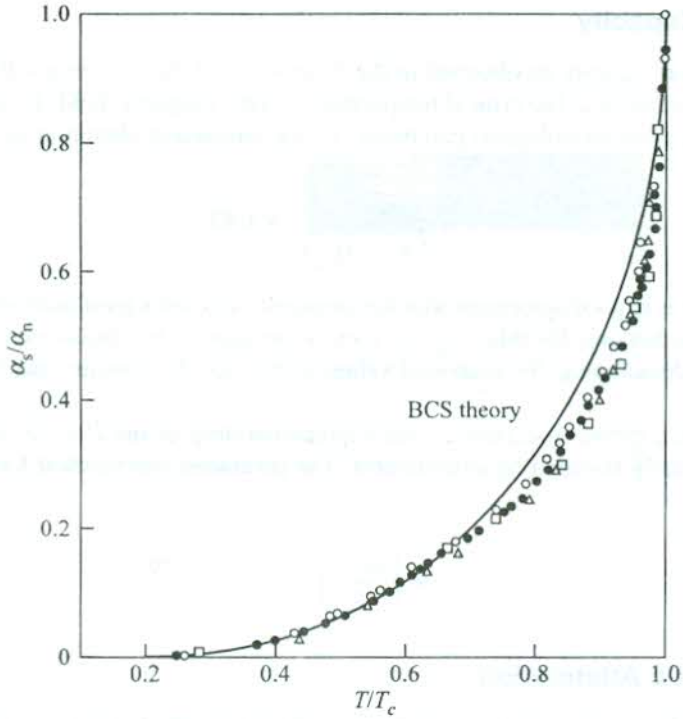


FIG. 15.19 Ratio of attenuation coefficients for acoustic waves in superconducting and normal states as a function of the reduced temperature for tin (○ at 33.5 MHz; ● at 54.0 MHz) and indium (□ at 28.5 MHz; Δ at 35.2 MHz). The curve represents the BCS result. [After R.W. Morse and H.V. Bohm, *Phys. Rev.*, **108,** 1094 (1957).]

15.7 GINZBURG–LANDAU THEORY

Our descriptions of various superconducting phenomena have thus far been conducted under two theoretical approaches: (i) a phenomenological theory effective through the London relations and (ii) the microscopic theory (the BCS theory). But the comprehensive theoretical picture remains incomplete without identifying the impact and consequences of another theory known as the Ginzburg–Landau theory* that had been proposed seven years earlier than the BCS theory. Using the Landau's idea that an ordered thermodynamic phase is characterized by a non-zero order parameter, the theory describes the superconducting state in terms of a complex order parameter $\psi(\mathbf{r})$ that vanishes at T_c. The magnitude of $\psi(\mathbf{r})$ gives a measure of the degree of order at position $\mathbf{r}$ at temperatures below T_c. The theory is accorded a semi-phenomenological status.

The order parameter $\psi(\mathbf{r})$ is viewed as a single-particle wavefunction describing the position of the centre of mass of a Cooper pair in the BCS picture. In a rare exhibition of intuitive skill, Ginzburg and Landau correctly guessed a quantum mechanical equation satisfied by $\psi(\mathbf{r})$. This turned out to be a direct and comprehensive way of describing the characteristics of superconducting phenomena such as the Meissner effect, zero resistivity, flux quantization, type II superconductivity and Josephson tunnelling.

---

* V.L. Ginzberg and L.D. Landau, *Zh. Eksp. Teor. Fiz.*, **20,** 1064 (1950).

The order parameter is a constant in the ground state of a superconductor where each Cooper pair occupies a translationally invariant state which is independent of the centre of mass coordinate. An applied field or a flow of current would, however, change the order parameter. The Ginzburg–Landau fundamental equation expressing the current density $\mathbf{j}_s$ as a function of the order parameter $\psi(\mathbf{r})$ in a superconductor placed in a magnetic field of vector potential $\mathbf{A}(\mathbf{r})$ is

$$\mathbf{j}_s = -\frac{e}{2m}\left[\psi^* \left\{\left(\frac{\hbar}{i}\nabla + 2e\mathbf{A}\right)\psi\right\} + \left\{\left(\frac{\hbar}{i}\nabla + 2e\mathbf{A}\right)\psi\right\}^* \psi\right] \tag{15.48}$$

Here, the current is attributed to the flow of Cooper pairs, i.e. particles of mass $2m$ and bearing a charge—$2e$. If we express the order parameter as $\psi = |\psi| \exp[i\theta(\mathbf{r})]$, assuming that its magnitude $|\psi|$ remains constant and its significant spatial variation occurs entirely through the phase $\theta(\mathbf{r})$, it is straightforward to show that the London equation (15.10) follows from (15.48). A constant value of $|\psi|$ means that the degree of order remains unaffected. This in turn implies that only such disturbances which do not change the Cooper pair density much from its uniform value at thermal equilibrium should be taken into consideration. It is a valid assumption in all those phenomena in which Cooper pairs are in flow but do not accumulate or break up. The flow of current being such an example, relation (15.48) reduces to the form

$$\mathbf{j}_s = -\left[\frac{2e^2}{m}\mathbf{A} + \frac{e\hbar}{m}\nabla\theta(\mathbf{r})\right]|\psi|^2 \tag{15.49}$$

or

$$\text{curl}\ \mathbf{j}_s = -\frac{2|\psi|^2 e^2}{m}\mathbf{B} \tag{15.50}$$

(because curl $\nabla\,\theta(\mathbf{r}) = 0$ and curl $\mathbf{A} = \mathbf{B}$)

On identifying the superfluid density n_s with $2|\psi|^2$, relation (15.50) becomes the London equation (15.10). Since ψ represents the wavefunction of particles carrying the charge $2e$, the identity of n_s with $2|\psi|^2$ sounds perfectly logical.

We describe below other features of superconductors whose interpretations are based essentially on the principles of the Ginzburg–Landau theory.

15.7.1 Magnetic Flux Quantization

Consider a superconducting ring placed in a magnetic field $\mathbf{B}$ of vector potential $\mathbf{A}$ (Fig. 15.20). Imagine a closed loop C deep inside the ring. The magnetic field is zero near this loop and so is the supercurrent because any appreciable current can flow only near the surface. Putting $\mathbf{j} = 0$ and replacing $2e$ by q in (15.49), we have

$$\nabla\,\theta(\mathbf{r}) = \frac{q}{\hbar}\mathbf{A}(\mathbf{r}) \tag{15.51}$$

The order parameter has a unique value at each point which must be such as to minimize the energy. On going round the closed path C once entirely within the ring, the phase $\theta(\mathbf{r})$ can change at most by 2π. Therefore, taking the line integral of (15.51), we get

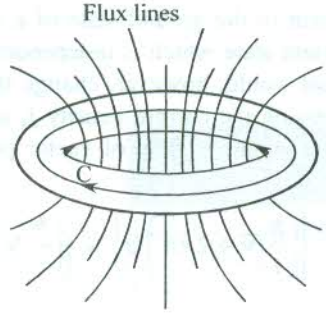


FIG. 15.20 A superconducting ring in a static magnetic field. The total magnetic flux $\Phi(= \phi_{ext} + \phi_{SC})$ is quantized; ϕ_{SC} adjusts itself to meet this condition because there exists no quantization condition for the external flux ϕ_{ext}.

$$\frac{q}{\hbar} \oint \mathbf{A} \cdot d\mathbf{l} = \oint \nabla\theta(\mathbf{r}) \cdot d\mathbf{l} = 2n\pi \tag{15.52}$$

Applying the Stokes theorem, we have

$$\oint \mathbf{A} \cdot d\mathbf{l} = \int_c \text{curl}\, \mathbf{A} \cdot d\mathbf{S} = \int_c \mathbf{B} \cdot d\mathbf{S} = \Phi \tag{15.53}$$

where Φ is the total magnetic flux.

Combining (15.52) and (15.53), we obtain

$$\Phi = n\left(\frac{2\pi\hbar}{q}\right) = n\Phi_0 \tag{15.54}$$

with

$$\Phi_0 = \frac{2\pi\hbar}{q} \tag{15.55}$$

The measured values of the magnetic flux show that $q = 2e$, putting the formation of Cooper pairs in the superconducting state on a firm footing. The fact that the order parameter is single-valued, thus produces the flux quantization in units of $2\pi\hbar/2e$. This unit is called a *fluxoid* or *fluxon*, measuring about 2.0678×10^{-15} tesla m^2.

15.7.2 Coherence Length

The concept of coherence length is deeply involved in the interpretation of some features of superconductivity such as type II superconductivity, introduced in Section 15.1.4. It was shown in Section 15.1.3 that in a pure semi-infinite superconductor the magnetic field is exponentially damped as one moved from its external surface to its interior. The field decays over a length, the so-called London penetration depth.

Suppose we have a thin superconductor that occupies space on both sides of the x-axis. In this case the magnetic field, applied parallel to the external surface of the superconductor and perpendicular to the x-axis, would not vanish at any point within the superconductor. Instead, the field will have

a minimum at $x = 0$ if the superconductor is placed symmetrically with respect to the origin. This picture is useful in accounting for the behaviour of type II superconductors in a magnetic field. A massive type II superconductor when placed in a magnetic field is considered as a collection of filaments (parallel to the magnetic field) of a normal conductor and a superconductor in an alternating pattern. Whereas energy considerations favour the expenditure of energy for the creation of interfaces between the normal and superconducting regions in type II superconductors, the same is not favourable in type I materials. In this model, the distance from the interface between a normal conductor and a superconductor over which the density of Cooper pairs rises from zero to its maximum value is termed *coherence length.*

Although the coherence length is best introduced into the theory through the Ginzburg–Landau equations which follow also from the BCS theory, its idea was first given by Pippard independently. Pippard defined an intrinsic coherence length ξ_0 as the size of the Cooper pair, i.e. the characteristic distance over which there must exist a correlation of the phases of the paired electrons. He measured the penetration of spatially varying magnetic fields in superconducting tin alloys as a function of alloying and suggested the modification of the London equations. Consider a cylindrical flux tube associated with the supercurrent $\mathbf{j}_s$. The current density $\mathbf{j}_s$, in particular, is found to depend on the value of $\mathbf{A}$ in a region of volume and not simply on the value at the same point in space as demanded by the London equation (15.10). These studies required the London equation (the second) to be replaced by a non-local equation

$$\mathbf{j}_s(\mathbf{r}) = \int f(\mathbf{r} - \mathbf{r}')\, \mathbf{A}(\mathbf{r}')\, d\mathbf{r}' \tag{15.56}$$

Equation (15.56) states that the supercurrent at the point $\mathbf{r}$ is contributed by electrons from different points $\mathbf{r}'$ within a volume, approximately $\xi^3{}_0$, where ξ_0 is a measure of the range $f(\mathbf{r} - \mathbf{r}')$. It is also regarded as a measure of the natural scale length over which the spatial variation of the order parameter (i.e. the wavefunction) requires a certain amount of kinetic energy. The increase in kinetic energy must not be more than the energy gap (2Δ) in order to preserve the superconducting state.

The natural scale length (ξ_0) is one of the fundamental lengths (such as the size of the Cooper pair), characterizing a superconductor, all of which are indiscriminately referred to as the coherence length. Well below T_c all such coherence lengths are equal in pure materials. But near T_c or in impure specimens where the mean free paths are small, these coherence lengths may be at variance in different contexts. The value of ξ_0 for pure metals can be as large as a few thousand angstroms at temperatures well below T_c (16,000 Å for aluminium at $T = 0$ K). On alloying, ξ_0 decreases rapidly and is of the order of 100 Å in disordered alloys.

The distance over which the electrons can behave in a coherent manner cannot be larger than their mean free path l. On alloying or adding impurity the mean free path decreases, resulting in the reduction of the effective volume of coherence. The broadening of the superconducting transition in an impure sample (see Fig. 15.1) may be viewed as a consequence of this fact. Both the coherence length ξ and the actual penetration depth λ depend on the mean free path l of electrons measured in the normal state. The effective coherence length in very impure superconductors, where $l << \xi_0$, may be expressed by

$$\xi = (\xi_0 l)^{1/2} \tag{15.57}$$

15.7.3 Type II Superconductivity

We refer to Section 15.1.4 where type I and type II superconductors were introduced and it was demonstrated that they show a different Meissner effect [see Fig. 15.6(b)]. It is very much in order to observe that the two types of materials have the same mechanism of superconductivity and differ only in the form of Meissner effect they exhibit. Let us now examine if it is energetically favourable for magnetic fields to partially penetrate a superconductor as is really observed in type II superconductors. Ginzburg, Landau, Abrikosov, and Gorkov have contributed significantly to the theoretical aspects of type II superconductivity. Abrikosov realized that it is the sign of the normal-superconducting interface energy that decides the partial penetration of the magnetic flux. He developed the concept of an interface between a region in the superconducting state and a region in the normal state. He went on to show that the interface energy is negative in type II superconductors above a certain value of the applied magnetic field and that the magnetic field penetrates in the form of quantized flux tubes.

Consider a flux tube with which only one quantum of flux ($\Phi_0 = h/2e$) may be associated. Let it be a cylinder with its axis along the field direction. The field is maximum along the tube axis and decreases to zero along the radius in a distance of the order of the penetration depth λ [see Fig. 15.21(b)]. The order parameter is shown [see Fig. 15.21(a)] to vanish at the axis and acquire its equilibrium value in a radial distance ξ. Figure 15.21(c) shows that the supercurrent density grows to the maximum value over the same distance. When a region of volume $\pi\xi^2$ per unit tube length goes to the normal state, there is a loss of superconductivity energy $\left(\frac{B_c^2}{2\mu_0}\pi\xi^2\right)$. But there is a gain in the magnetic field energy $\left(\frac{B_a^2}{2\mu_0}\pi\lambda^2\right)$ when the magnetic field penetrates a region of volume $\pi\lambda^2$ per unit tube length. Therefore, the interface energy or the energy of the flux tube per unit tube length may be written as

$$\varepsilon \equiv \frac{B_c^2}{2\mu_0} \cdot \pi\xi^2 - \frac{B_a^2}{2\mu_0} \cdot \pi\lambda^2 \tag{15.58}$$

where B_c is the thermodynamic critical field, B_a is the applied field and $B_c^2/2\mu_0$ represents the stabilization energy density of the superconducting state.

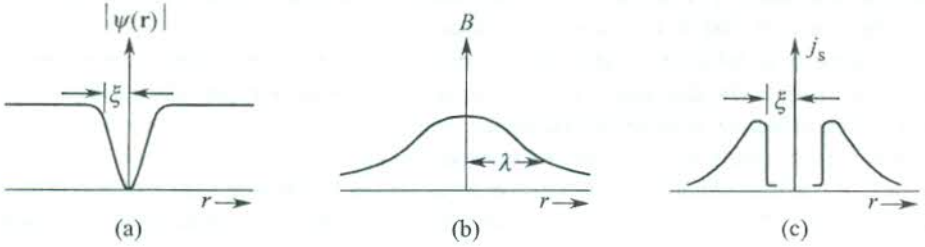


FIG. 15.21 (a) Variation of the order parameter $|\psi(\mathbf{r})|$ as a function of the radial distance from the axis of a cylindrical flux tube. (b) Variation of the magnetic field within a superconductor as a function of the radial distance from the axis of a cylindrical flux tube. (c) Variation of the supercurrent density as a function of the radial distance. It grows to its maximum value (so does the Cooper pair density) over a distance equalling coherence length.

The energy ε is always positive for $B_a < B_c$ if $\xi > \lambda$. It refers to the case of type I superconductors. The energy can be negative, for $\xi < \lambda$ if

$$B_a > B_{c_1} \cong B_c\left(\frac{\xi}{\lambda}\right) < B_c \tag{15.59}$$

The above condition describes the state of condition in a type II superconductor when it is energetically favourable for the magnetic flux to partially penetrate the superconductor. It is quite reasonable to regard the normal core of radius ξ as a hole in the superconductor and exploit the concept of flux quantization in a superconducting ring. At the interface the field would extend out from the normal core to a distance λ into the superconducting region. Thus the field associated with a tube of radius λ is $\pi\lambda^2 \cdot B_{c_1}$, which must equal the flux quantum Φ_0, yielding

$$B_{c_1} \approx \frac{\Phi_0}{\pi\lambda^2} \tag{15.60}$$

On raising the field to this value in a type II superconductor, a single fluxoid is nucleated. At a much higher magnetic field B_{c_2}, the flux tubes are packed to the maximum possible density in the superconducting state as may be permitted by the coherence length ξ. With a small fluctuation on the scale of this flux tube order, the field penetrates the superconductor almost uniformaly and the superconductivity is destroyed. In some conventional superconductors (liquid helium superconductors), for example in Nb_3Sn wires, large supercurrents have been observed up to fields in the range of 10 tesla. Since each flux tube bears a flux approximately equal to $\pi\xi^2 \cdot B_{c_2}$, we have

$$\pi\xi^2 \cdot B_{c_2} \approx \Phi_0$$

or

$$B_{c_2} \approx \frac{\Phi_0}{\pi\xi^2} \tag{15.61}$$

From relations (15.60) and (15.61), we get

$$B_{c_2} \approx B_{c_1}\left(\frac{\lambda}{\xi}\right)^2 \tag{15.62}$$

It is now possible to relate B_{c_2} with the thermodynamic critical field B_c [see Fig. 15.6(b)], which is crucial to the evaluation of the stabilizing energy per unit volume of the superconductor. Placing the value of B_{c_1} from (15.59) in (15.62), we get

$$B_{c_2} \approx \left(\frac{\lambda}{\xi}\right) B_c$$

$$\approx \kappa B_c \tag{15.63}$$

where κ is a parameter $\left(= \frac{\lambda}{\xi}\right)$, introduced by Ginzburg and Landau.

The lower critical field B_{c_1} can then be rewritten as

$$B_{c_1} \approx \frac{B_c}{\kappa} \tag{15.64}$$

and from (15.63) and (15.64), we have

$$(B_{c_1} B_{c_2})^{1/2} \approx B_c \tag{15.65}$$

The Ginzburg–Landau parameter κ is small for type I superconductors ($<1/\sqrt{2}$); for example 0.03 for aluminium and 0.15 for indium. It is much larger for the type II superconductors where the coherence length is small because of the characteristic short mean free path.

As shown in Fig. 15.6(b), a type II superconductor is said to be in the vortex state when $B_{c_1} < B_a < B_{c_2}$. The flux tubes are also addressed as 'vortices' on account of a vortex-like pattern of supercurrent associated with each tube. The vortex state can have zero resistivity if the vortices do not move under the action of an applied electric field. But if the vortices move, their normal core regions too undergo a displacement which is crudely equivalent to the flow of electrons of a normal metal under the influence of an electric field. Since such a current is dissipative, the system in this condition will be endowed with a non-zero electrical resistance. The movement of the flux tubes is effectively prevented in a system with plenty of such regions in which the presence of the flux tubes may be energetically favourable. For example, a cold-worked metal has a large density of dislocation lines. It may be energetically favourable for the flux tubes to be along or near these lines which can act as pinning centres. This is known as *flux pinning* since it prevents the vortices from moving. There can occur a large current density if there is effective pinning in the superconductor. The maximum current density that can be sustained in a superconductor without making it resistive is called the *critical current.* As the current tends to exceed this value, the superconductivity is destroyed (Silsbee effect). A critical current density of the order of 10^{10} A m^{-2} is possible in good quality films or crystals of some type II materials. The wires of these materials are used in superconducting magnets, producing large fields (~ 10 tesla).

15.7.4 Josephson Tunnelling

Earlier in Section 15.3.4 we discussed Giaever tunnelling that demonstrates single-electron tunnelling from a metal through a thin insulating layer into a superconductor. In 1962, Josephson theoretically showed that an electron pair (the Cooper pair) can also tunnel from one superconductor through a fine insulating layer into another superconductor as shown in Fig. 15.22(a). This is known as *Josephson tunnelling*. The coupling between the two superconductors provided by their insulating barrier must be very weak so that there is a very low probability of finding a Cooper pair in the insulating region. In practice, this is achieved by restricting the barrier thickness to about 10–20 Å. An arrangement such as that shown in Fig. 15.22(a) is referred to as a *Josephson junction*. The significant phenomena associated with Cooper pair tunnelling in a Josephson junction are identified as: (i) Josephson Effect and (ii) Supercurrent Quantum Interference.

Josephson effect

We present a basic theoretical description of Josephson effect for an experimental geometry given by Fig. 15.22(b). The interpretation of general expressions reveals two types of effects: the dc Josephson effect and the ac Josephson effect. We will dwell on them at the proper stage of the theoretical development.

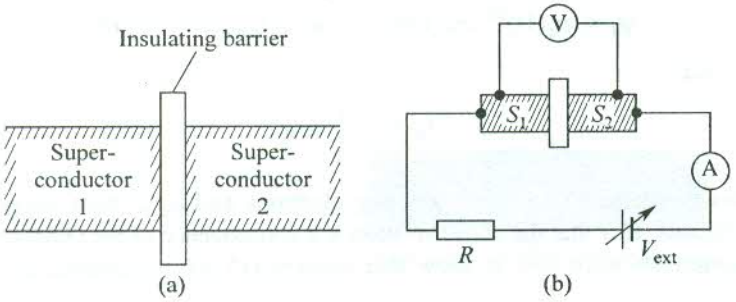


FIG. 15.22 (a) Combination of two superconductors so as to show Josephson tunnelling. (b) Circuit for studying Josephson effect.

For simplicity, we assume that the superconductors S_1 and S_2 in our model are of the same material. The temperature being so low, we need to consider the electrons only in their superconducting ground state. The magnetic field near the junction must be screened completely because the highly coherent Cooper pair states are strongly influenced by magnetic fields. Even the magnetic field due to earth may be sufficient to completely obscure the Josephson effect. Let ψ_1 and ψ_2 denote the many-particle wavefunctions of the Cooper pair states in the two superconductors and $\mathcal{H}_1$ and $\mathcal{H}_2$ be the hamiltonians of the respective isolated superconductors. Then, the time-dependent Schrödinger equation of the system gives

$$i\hbar \frac{\partial \psi_1}{\partial t} = \mathcal{H}_1 \psi_1 + T\psi_2 \tag{15.66a}$$

$$i\hbar \frac{\partial \psi_2}{\partial t} = \mathcal{H}_2 \psi_2 + T\psi_1 \tag{15.66b}$$

where T is the characteristic coupling constant for the junction and the energy contributions $T\psi_1$ and $T\psi_2$ simply signify that the many particle wavefunction ψ_1 or ψ_2 of one superconductor does not completely vanish in the other. If the junction is characterized with an impenetrable barrier, relations (15.66) reduce to two uncoupled equations for the isolated superconductors because, then, $T = 0$.

The circuit in Fig. 15.22(b) shows that when an external difference of potential V_{ext} is impressed, a difference of potential V appears across the tunnel junction. When the two superconductors S_1 and S_2 are of the same material, as assumed for simplicity, following the perturbation approach, ψ_1 can be regarded as an approximate solution with respect to $\mathcal{H}_1$. This allows us to replace $\mathcal{H}_1$ by the energy of the superconducting state. The voltage V that appears across the junction creates a relative shift of (2 eV) in the energies of the superconductors S_1 and S_2. With zero of the energy scale taken at the midlevel of the energies of S_1 and S_2, relations (15.66) transform to:

$$i\hbar \frac{\partial \psi_1}{\partial t} = \mathrm{eV}\psi_1 + T\psi_2 \tag{15.67a}$$

$$i\hbar \frac{\partial \psi_2}{\partial t} = -\mathrm{eV}\psi_2 + T\psi_1 \tag{15.67b}$$

The wavefunctions ψ_1 and ψ_2 are usually regarded as the order parameters in the two superconductors:

$$\psi_1 = (n_c(1))^{1/2} \exp(i\theta_1); \quad \psi_2 = (n_c(2))^{1/2} \exp(i\theta_2) \tag{15.68}$$

Here, we have used

$$|\psi|^2 = \frac{n_s}{2} = n_c$$

where n_s denotes the density of superconducting electrons and n_c is the Cooper pair density.

We saw in Section 15.7 that the above relation is a consequence of the Ginzburg–Landau theory. With this realization we were able to show that relation (15.50) represented the London equation (15.10).

Treating (15.68) as solutions, we place them in (15.67a). The separation of real and imaginary parts gives the following relations:

$$-\frac{\hbar}{2}\frac{\partial n_c(1)}{\partial t}\sin\theta_1 - \left(\hbar\frac{\partial\theta_1}{\partial t} + eV\right) n_c(1)\cos\theta_1 = T[n_c(1)\, n_c(2)]^{1/2}\cos\theta_2 \tag{15.69a}$$

$$\frac{\hbar}{2}\frac{\partial n_c(1)}{\partial t}\cos\theta_1 - \left(\hbar\frac{\partial\theta_1}{\partial t} + eV\right) n_c(1)\sin\theta_1 = T[n_c(1)\, n_c(2)]^{1/2}\sin\theta_2 \tag{15.69b}$$

On multiplying (15.69a) by $\sin\theta_1$ and (15.69b) by $\cos\theta_1$ and subtracting one from the other, we obtain

$$\frac{\partial n_c(1)}{\partial t} = \frac{2}{\hbar} T[n_c(1)\, n_c(2)]^{1/2}\sin(\theta_2 - \theta_1) \tag{15.70a}$$

$$\frac{\partial\theta_1}{\partial t} = -\frac{1}{\hbar} T\left(\frac{n_c(2)}{n_c(1)}\right)^{1/2}\cos(\theta_2 - \theta_1) - \frac{eV}{\hbar} \tag{15.70b}$$

Similarly, from (15.67b), we get

$$\frac{\partial n_c(2)}{\partial t} = -\frac{2}{\hbar} T[n_c(1)\, n_c(2)]^{1/2}\sin(\theta_2 - \theta_1) \tag{15.71a}$$

$$\frac{\partial\theta_2}{\partial t} = -\frac{1}{\hbar} T\left(\frac{n_c(1)}{n_c(2)}\right)^{1/2}\cos(\theta_2 - \theta_1) + \frac{eV}{\hbar} \tag{15.71b}$$

Since we have assumed that the two superconductors are of the same material, $n_c(1) = n_c(2) = n_c$; using this condition in the relations (15.70) and (15.71), we get the following two equations:

$$\frac{\partial n_c(1)}{\partial t} = -\frac{\partial n_c(2)}{\partial t} = \frac{2T}{\hbar} n_c \sin(\theta_2 - \theta_1) \tag{15.72a}$$

$$\hbar\left(\frac{\partial\theta_2}{\partial t} - \frac{\partial\theta_1}{\partial t}\right) = 2eV \tag{15.72b}$$

These equations describe the *V–I* characteristics of a Josephson tunnel diode as depicted in Fig. 15.23. For a comprehensive understanding of the observed characteristic curve shown in Fig. 15.23, we conduct the discussion in two steps.

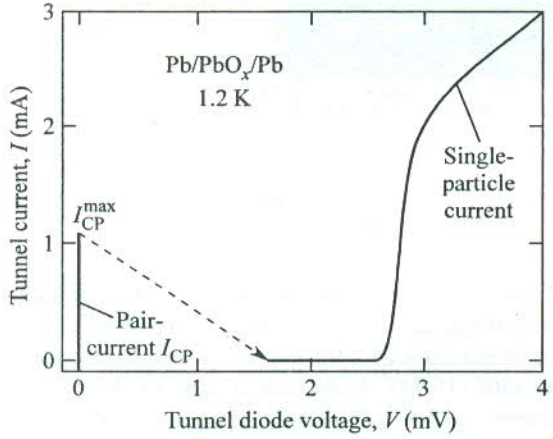


FIG. 15.23 *V–I* characteristic curve for a Pb/PbO$_x$/Pb Josephson tunnel diode. [After D.N. Langenberg, D.J. Scalapino and B.N. Taylor, *Proc. IEEE,* **54,** 560 (1966).]

Step I: The dc Josephson effect. It is interesting to notice that relations (15.72) have a non-zero solution for current ($\partial n_c/\partial t \neq 0$) even when there is no voltage drop across the diode ($V = 0$). This implies that as soon as the contact across the junction is complete, there flows a tunnel current of Cooper pairs in a direction determined by the phase difference ($\theta_2 - \theta_1$) between the wavefunctions on the two sides of the tunnel barrier. It is a significant observation that the voltage drop V across the junction is zero though a voltage V_{ext} has been applied externally. The current flowing in the circuit in this condition is purely the tunnel current of Cooper pairs I_{CP}. This phenomenon is called the *dc Josephson effect*. The tunnel current is collected and replaced at the terminals of the external source of voltage in the circuit. This rules out any charge asymmetry that might arise on account of the flow of current. The dc Josephson current I_{CP} is thus kept constant and it has only one value initially at $V = 0$.

But with increase in V_{ext}, the I_{CP} also rises to a maximum value I_{CP}^{max}. If V_{ext} is increased further, the state becomes unstable and a potential difference now appears ($V \neq 0$) across the diode. The current I_R in this state depends on V_{ext}, V, and the external resistance R; and the characteristic curve (beyond the arrowhead in Fig. 15.23) is determined by the single electron tunnelling [see Fig. 15.14(b)]. The electrons owe their presence to the breaking of Cooper pairs.

Step II: The ac Josephson effect. As far as the phase difference ($\theta_2 - \theta_1$) is time-invariant, in accordance with (15.72a) I_{CP}^{max} remains stable. But (15.72b) tells us that this state is maintainable only if there is no voltage drop across the diode ($V = 0$). We saw above that an increase in V_{ext} that causes I_{CP} to exceed I_{CP}^{max} results in the collapse of I_{CP} and the creation of a potential difference V (purely corresponding to the single-electron tunnelling current) across the diode. In the event of this voltage drop, the phase difference between the two states of the superconductor begins to grow with time according to relation (15.72b). We integrate (15.72b) to obtain

$$\theta_2 - \theta_1 = \frac{2eVt}{\hbar} + \Delta\theta_0 \tag{15.73}$$

where $\Delta\theta_0$ represents the initial phase difference.

On placing the value of $(\theta_2 - \theta_1)$ from (15.73) in (15.72a), we obtain a relation for current that represents an alternating current:

$$\tilde{I}_{CP} \approx \frac{\partial n_c(1)}{\partial t} = \frac{2Tn_c}{\hbar} \sin(\omega_{CP} t + \Delta\theta_0) \tag{15.74}$$

where

$$\omega_{CP} = \frac{2eV}{\hbar} \tag{15.75}$$

The phenomenon of generating an alternating current by producing a dc voltage drop across a Josephson junction is called the *ac Josephson effect*. It must be clear that the alternating current $\tilde{I}_{CP}$ is in addition to the direct current contributed by single-electron tunnelling.

A simple calculation with (15.75) shows that a drop of mere 1 mV across the tunnel diode produces an alternating current of 3×10^{12} oscillations per second which correspond to a frequency in the infrared region of the electromagnetic radiation. The relation (15.75) implies that a photon energy $\hbar\omega_{CP} = 2$ eV must be absorbed or emitted as a Cooper pair tunnels through a Josephson junction. It provides a basis for an accurate measurement of the ratio h/e.

There are several applications of the Josephson effect. For example, the construction of binary switching devices in microelectronics is based on the two states of the Josephson junction, i.e. $I_{CP} \neq 0$ at $V = 0$, and single electron tunnelling at $V \neq 0$. These devices are very fast and are employed for computer data storage.

Supercurrent quantum interference

The Josephson tunnelling in the presence of a magnetic field provides strong evidence for the highly coherent nature of the superconducting state. Two Josephson junctions arranged in a parallel combination are placed in a region in which a magnetic field **B** is impressed as shown in Fig. 15.24. A supercurrent starting in region I is divided into two parts and made to flow along parallel paths,

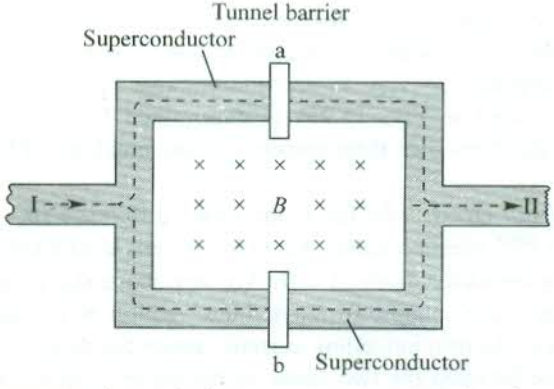


FIG. 15.24 Experimental geometry for producing supercurrent quantum interference. The phase difference of the total Cooper pair wave functions in region I and region II is monitored by the magnetic flux density **B** that penetrates the ring.

each of which contains a tunnel junction. The currents I_a and I_b crossing the tunnel barriers 'a' and 'b', respectively, reunite in region II. The combined current shows oscillations characteristic of an interference pattern produced by two coherent sources. By analogy with the interference of light, I_a and I_b are regarded as two coherent sources of current whose disturbances, when superposed by the way of recombination, produce an interference pattern.

In view of relation (15.73), the tunnelling of Cooper pairs causes a phase shift of the total wavefunction of the superconducting state in region II relative to that in region I. If the phase shifts at the two barriers in the absence of the magnetic field be δ_a and δ_b, the supercurrents through the two junctions can be written as

$$I_a = I_0 \sin \delta_a \tag{15.76a}$$

$$I_b = I_0 \sin \delta_b \tag{15.76b}$$

We refer to relation (15.49) to determine the phase difference between the two regions I and II in the presence of a magnetic field of vector potential **A**. The relation expresses the supercurrent density $\mathbf{j}_s$ which can be taken as zero, deep inside the superconductor. Placing $\mathbf{j}_s = 0$ in (15.49), we have

$$\nabla\theta = \frac{2e}{\hbar}\mathbf{A}$$

Taking a line integral of the above equation along a path between two points P and Q deep inside the superconductor,

$$\Delta\theta \Big|_P^Q = \int_P^Q \nabla\theta \cdot d\mathbf{l} = \frac{2e}{\hbar}\int_P^Q \mathbf{A} \cdot d\mathbf{l} \tag{15.77}$$

Therefore, the total phase shifts in the wavefunction along the two paths from region I to region II can be expressed as

$$\Delta\theta \Big|_I^{II} = \delta_a + \frac{2e}{\hbar}\int_a \mathbf{A} \cdot d\mathbf{l} \tag{15.78a}$$

$$\Delta\theta \Big|_I^{II} = \delta_b - \frac{2e}{\hbar}\int_b \mathbf{A} \cdot d\mathbf{l} \tag{15.78b}$$

The two phase shifts must be identical because the wavefunction has a unique value at every point. Furthermore, the two line integrals in (15.78a) and (15.78b) are taken in opposite directions. Therefore, when taken together they give an integral over a closed path. Using this property and taking the phase shifts (15.78) for the two parallel paths as identical, we obtain

$$\delta_b - \delta_a = \frac{2e}{\hbar}\oint \mathbf{A} \cdot d\mathbf{l} = \frac{2e}{\hbar}\int \mathbf{B} \cdot d\mathbf{S} \qquad \text{(using the Stokes theorem)} \tag{15.79}$$

The above relation states that the total phase difference around the loop can be controlled by varying the magnetic field. The general expressions for δ_a and δ_b in (15.76) may, however, be put as

$$\delta_a = \delta_0 - \frac{e}{\hbar} \int \mathbf{B} \cdot d\mathbf{S} \tag{15.80a}$$

$$\delta_b = \delta_0 + \frac{e}{\hbar} \int \mathbf{B} \cdot d\mathbf{S} \tag{15.80b}$$

where for the sake of generality we have introduced an arbitrary phase shift δ_0 that depends on the nature of the tunnel barrier and the applied voltage.

Inserting the values of δ_a and δ_b from (15.80) in (15.76), we obtain the total recombined supercurrent:

$$I = I_a + I_b = 2I_0 \sin \delta_0 \cos \left(\frac{e}{\hbar} \int \mathbf{B} \cdot d\mathbf{S} \right) \tag{15.81}$$

The cosine interference term characterizes the total current. This phenomenon is called the *supercurrent quantum* interference. Its maxima are determined by the condition,

$$\int \mathbf{B} \cdot d\mathbf{S} = m \frac{\pi \hbar}{e} = m \left(\frac{h}{2e} \right); \qquad 1, 2, 3, \ldots \tag{15.82}$$

$$= m \text{ (flux quantum)}$$

This condition states that for every addition of a flux quantum to the enclosed flux, a new maximum appears. The total supercurrent is plotted as a function of magnetic field in Fig. 15.25 to demonstrate the quantum interference where each oscillation corresponds to a change of flux quantum. Based on this principle, extremely sensitive magnetometers have been developed. Even extremely weak magnetic fields such as those produced by currents in human brain can be measured with these magnetometers. This magnetometer is called a SQUID (Superconducing Quantum Interference Device).

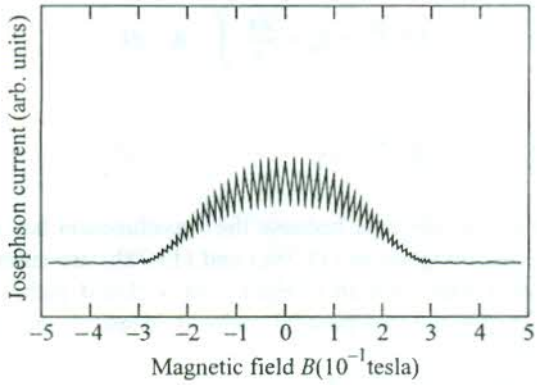


FIG. 15.25 Total supercurrent in region II as a function of the magnitude of the magnetic field, penetrating the loop shown in Fig. 15.24. The plot is typically an interference pattern produced by the superposition of two supercurrents I_a and I_b that have tunnelled two separate Josephson junctions 'a' and 'b'. [After R.C. Jaklevic, J. Lambe, J.E. Mercereau, A.H. Silver; *Phys. Rev.*, **140**, A1628 (1965).]

15.8 HIGH TEMPERATURE SUPERCONDUCTORS (HTS)

Extremely low critical temperatures of conventional superconductors (the low T_c type) put the most serious limitation on their use in technological applications. Working with devices that have to be cooled to temperatures in the range of liquid helium temperature (4.2 K) is obviously not viable on any count. This has kept the scientists world over relentlessly trying to discover superconductivity near room temperature. A decisive boost to this optimism came in 1986, when Bednorz and Müller synthesized metallic oxygen-deficient copper oxide compounds of La-Ba(Sr)-Cu-O system with the transition temperature of about 30 K. A vigorous activity towards the search for materials with higher critical temperatures ensued following this nobel prize winning announcement. It has resulted in the development of a variety of materials with the highest critical temperature T_c observed in the vicinity of 135 K in a mercury cuprate. Under pressure this T_c value approaches 165 K which is not far away from the temperature of the coldest regions on Earth (183 K). The T_c values being so high compared to those of conventional superconductors, these materials are called *high temperature superconductors* or *high T_c superconductors* (HTS).

The scope of this book does not permit us to do justice to the explosive development of HTS and their properties. Nevertheless, we give below an account of the structures and other salient features of some materials that represent the main classes of HTS in the order of their discovery.

15.8.1 Rare-earth Cuprates: Structural Aspect

Chu and Coworkers (1987) earned the distinction of raising T_c to 90 K in ceramics of the $Ba_{1-x}Y_xCuO_{3-y}$ system. With fastly improving methods of preparation of characterization, a ceramic alloy $Y_1Ba_2Cu_3O_{7-x}$ could be prepared even in single crystal form. In all respects including application this has emerged as the most thoroughly studied and tested system, often referred to as YBCO. A series of this class of HTS has been produced with the Y atom being replaced by other rare-earth elements such as Eu and Gd. On the basis of their stoichiometry, these types of ceramics are commonly called 123 systems.

The crystal structure of the YBCO system is illustrated in Fig. 15.26(a). It can be represented by an orthorhombic primitive cell in the superconducting state. The structure is essentially an oxygen-defect modification of the perovskite structure with about one-third oxygen positions vacant. All members of this series are axial crystals with alternating CuO_2 planes [Cu(2), O(2)] and oxygen atoms in both pyramid-type and rectangular coordination along the *c*-axis. Oxygen chains are formed along the *b*-axis with the involvement of atoms in the rectangular planar structure. We will see a little later that the oxygen vacancies in this chain may be interpreted to be actively involved in the mechanism of superconductivity.

15.8.2 Bi-based and Tl-based Cuprates: Structural Aspect

This class of HTS emerged within a year of the synthesis of 123 systems. These materials, typically represented by $Bi_2Sr_2Ca_2Cu_3O_{10}$ and $Tl_2Ba_2Ca_2Cu_3O_{10}$ systems, show still higher T_c. The main classes of ceramic superconductors with $T_c > 90$ K are compiled in Table 15.4. In accordance with their stoichiometry, Bi- and Tl-based HTS are named as 2212 and 2223 systems, respectively.

Similar to 123 systems, 2212 and 2223 systems too have a layered structure along the substantially larger *c*-axis. This layered structure is again considered to play a crucial role in the mechanism of superconductivity. The unit cell shown in Fig. 15.26(b) has two distinct regions, separated by two Bi–O (or Tl–O) planes. In the upper-half region, the copper atoms are located at centres while in the lower-half region they are at corners of the Cu–O planes. The T_c value is strongly controlled by the

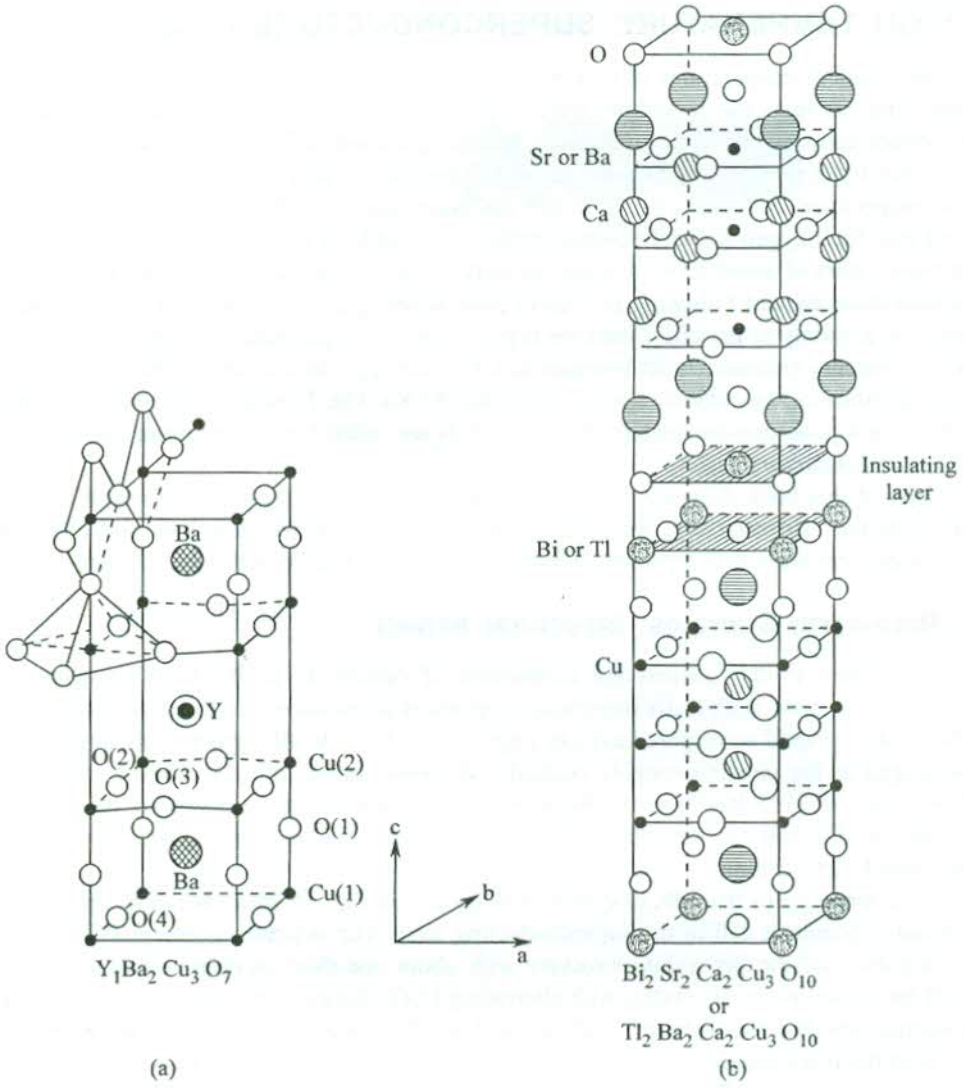


FIG. 15.26 (a) Unit cell structure of a $Y_1Ba_2Cu_3O_7$ crystal. The numbers in brackets represent the special sites of oxygen and copper atoms in CuO_2 layers. (b) Unit cell structure of $Bi_2Sr_2Ca_2Cu_3O_{10}$ or $Tl_2Ba_2Ca_2Cu_3O_{10}$ crystal.

number of CuO_2 layers in the unit cell. These ceramics differ from one another only in the number of CuO_2 layers per unit cell.

15.8.3 Significant Properties of Cuprate HTS

Consider the example of YBCO which is the most thoroughly researched system. Its resistivity around 90 K falls most sharply to an immeasurable value for $x = 0 - 0.1$, where x denotes oxygen deficiency. On increasing x, the transition temperature decreases. For $x > 0.7$, YBCO ceramics cease to be superconductors and behave as antiferromagnetic insulators. On account of their strongly anisotropic

Table 15.4 Prominent families of high temperature superconductors with highest T_c values reached. [After R. Hott, G. Rietschel, M. Sander; *Phys. Bl.*, **48,** 355 (1992).]

Formula	*Short name*	*Highest T_c observed* (K)
$REBa_2Cu_3O_7$ (RE = Rare Earths) = Y, Eu, Gd, ...)	RE BCO or 123	92 (YBCO)
$Bi_2Sr_2Ca_{n-1}Cu_nO_{2n+4}$ (+ Pb dopping)	BSCCO or Bi–22 (n – 1)n	90 (Bi–2212) 122 (Bi–2223)
$Tl_2Ba_2Ca_{n-1}Cu_nO_{2n+4}$	TBCCO or Tl–22 (n – 1)n	110 (Tl–2212) 127 (Tl–2223)
$Tl_2Ba_2Ca_{n-1}Cu_nO_{2n+3}$ (A = Sr, Ba)	Tl–12 (n – 1)n	90 (Tl–1212) 122 (Tl–1223) 122 (Tl–1234) 110 (Tl–1245)
$HgBa_2Ca_{n-1}Cu_nO_{2n+2}$	Hg–12 (n – 1)n	128 (Hg–1212) 135 (Hg–1223)

* approaches 165 K under pressure.

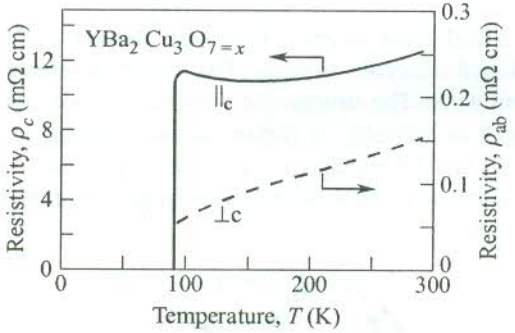


FIG. 15.27 Measured resistivity of YBCO along and perpendicular to the *c*-axis (ρ_a and ρ_{ab} respectively) as a function of temperature. [After S.J. Hagen, T.W. Jing, Z.Z. Wang, J. Horvath, N.P. Ong, *Phys. Rev.*, **B37,** 7928 (1988).]

crystal structure, the ceramic superconductors show highly anisotropic electronic properties. There is a large difference in the resistivities of YBCO, measured along and perpendicular to the *c*-axis (ρ_c and ρ_{ab} in Fig. 15.27). All the ceramic superconductors known to date show type II superconductivity for which B_{c_1} is usually less than 10 mT and the largest estimates of B_{c_2} are around 340 T.

A few extraordinary features of these HTS that might provide clue to the mechanism of superconductivity are as under:

1. The resistivity in the normal state varies linearly with temperature.
2. A near zero oxygen isotope effect is observed ($\alpha \sim 0$–0.2). The vanishingly small isotope effect is considered an important evidence for non-phononic superconductivity in cuprates.

3. The observed energy gaps are large, nearly 20–30 meV, and $\frac{\Delta(0)}{k_B T_c} = 3$ to 4, which is appreciably greater than the BCS estimate equalling 1.764.
4. The thermoelectric power shows a universal behaviour as a function of hole concentration.
5. The Hall coefficient is temperature dependent.
6. An inverted parabolic relation between T_c and the hole concentration is observed.

From the data on Hall coefficient it is inferred that a Cooper pair in YBCO type and Bi- and Tl-based superconductors is a pair of holes resulting in the *p*-type superconductivity in these materials. Because of their high electronegativity, oxygen atoms act as electron acceptors. For example, in YBCO, both Y and Ba ions contribute two electrons separately to the bonding in CuO_2 layers where the oxygen atoms trap these electrons. For small x (i.e. for a less oxygen-deficient composition), there are enough oxygen atoms to swallow the electrons. This way more holes are made available in the CuO_2 planes to get bound into hole Cooper pairs. These observations point to a quasi two-dimensional charge transport in CuO_2 planes by means of holes bound in Cooper pairs. These ideas are also applicable to the Bi and Tl superconductors. Although most of the cuprates show *p*-type superconductivity, there exist a couple of systems, namely Nd_2CuO_4 and $Nd_{2-x}Ce_xCuO_4$ in which the conventional *n*-type superconductivity has been confirmed.

15.8.4 Fullerenes

The novel superconductors added most recently (1991) to the list of HTS are fullerenes whose prominent members are C_{60}-based molecular crystals. The transition temperatures of materials of this class range from 15 K to about 48 K. The structure of a single C_{60} molecule, as shown in Fig. 15.28, consists of 60 carbon atoms. It is a cluster of carbon atoms arranged in the shape of a truncated icosahedron with 20 hexagonal and 12 pentagonal faces (as in graphite, benzene and other organic molecules). The pentagons occur on account of the topological requirement for producing a closed structure that resembles a football.

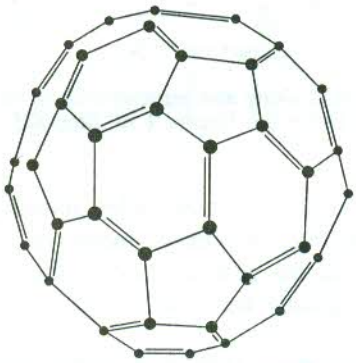


FIG. 15.28 The structure of a C_{60}-molecule.

The C_{60} clusters form the basis for a three-dimensional crystal structure of C_{60} which is characterized by an FCC unit cell. Some other interesting fullerenes have been derived from C_{60} by crystallizing C_{60} along with alkali metals whose atoms are placed in gaps between the C_{60} spheres. These are found to show superconductivity with values of T_c that justify to put them in the category of HTS. A list of such fullerenes with their T_c values is given as follows:

Fullerene	T_c(K)
K_3C_{60}	19.3
Rb_3C_{60}	28
$RbCs_2C_{60}$	33
$Rb_{2.7}Tl_{0.2}C_{60}$	48

Similar to ceramic HTS, fullerenes demonstrate the gradual onset of diamagnetism when cooled in zero magnetic field to temperatures below T_c. It has not been possible so far to carry out other experiments on fullerenes on account of their strong tendency to react with atmospheric oxygen.

Despite the availability of quality data of enormous volume and getting an exclusive attention of the entire community of solid state physicists, the theory of HTS continues to be in an uncertain shape. In view of this fact, it is appropriate to dispense with the theoretical treatment of HTS in this introductory version of the subject.

15.9 APPLICATIONS

Superconductors have tremendous potential for application. It is extremely advantageous to use them as conductors of electric current because of their nondissipative property. Applications are mostly based on conventional (liquid helium) superconductors. Although they have already found numerous applications, their utilization has been much restricted by the requirement of extremely low temperatures. It is not possible to give here a complete list of applications. The main applications are: high field magnets, SQUIDS, and some other electronic and rf devices. The NMR imaging technique employing high magnetic fields has proved far superior to CATSCAN used in medical diagnostics. High Q-cavities and various sensors are some other important devices where superconductors are used. In addition to being employed in SQUIDS, there are a large number of applications of Josephson junctions in electronic circuitry.

In view of their higher T_c, the use of HTS in a much larger number of applications seems imminent. With considerable advancement in the methods of sample preparation (e.g. laser ablation, MOCVD, etc.), good quality wires, tapes and films of cuprates have been produced. But the low critical current and ceramic nature of HTS have seriously limited their applications. The degree of success achieved has been highest with the YBCO system, which is utilized for the fabrication of Josephson junctions and SQUIDS. With revolutionary advancement in material research, a very large number of applications spread over different areas can be realized. A high speed (~ 550 km per hour) leviating train has been tested using a conventional superconductor. But these leviating vehicles and long distance non-dissipative power transmission systems have yet to be realized in practice.

SUMMARY

1. The dc resistance of a superconductor is zero.
2. The Meissner effect ($\mathbf{B} = 0$) is complete in bulk metallic samples in the superconducting state. The London penetration depth λ_L gives the extent over which the magnetic flux penetrates from the surface of a superconductor.

3. The London equation

$$\text{curl } \mathbf{j}_s = -\frac{n_s e^2}{m}\mathbf{B}$$

leads to the equation

$\nabla^2\mathbf{B} = \frac{1}{\lambda_L^2}\mathbf{B}$, which interprets the Meissner effect with $\lambda_L = \left(\frac{m}{\mu_0 n_s e^2}\right)^{1/2}$

4. The application of magnetic field B_c (the critical field) destroys superconductivity. The B_c at any temperature T is given by the empirical rule,

$$B_c(T) = B_c(0)\left[1 - \left(\frac{T}{T_c}\right)^2\right]$$

5. For a type I superconductor, there is a single critical field B_c for which the superconductivity is lost. There are some type II superconductors in which at a lower field B_{c_1} the magnetic flux begins to penetrate (Meissner effect is only partial). At a much higher field B_{c_2}, there is no Meissner effect; the field penetration is complete and the superconductivity is destroyed. The thermodynamic critical field B_c is in between B_{c_2} and $B_{c_1}(B_{c_1} < B_c < B_{c_2})$.

6. The difference between Gibbs free energies per unit volume of the normal and superconducting states of a superconductor is

$$G_n(T) - G_s(T) = \frac{B_c^2}{2\mu_0}$$

7. The specific heat of a superconductor is

$$C_V^s = \text{A} \exp\left(-\frac{\Delta}{k_B T}\right) + \alpha T^3$$

where 2Δ is the energy gap.

8. Isotope effect: $T_c M^\alpha$ = constant, where $\alpha \approx 0.5$.

9. In the superconducting state the current carriers are Cooper pairs, i.e. electron pairs. In a Cooper pair the wavevector and spin of the two electrons are represented as $(\mathbf{k}\uparrow, -\mathbf{k}\downarrow)$.

10. In a superconductor, the magnetic flux is quantized in units of $h/2e$.

11. For a type II superconductor

$$B_{c_2} \approx \kappa B_c \quad \text{and} \quad (B_{c_2}B_{c_1})^{1/2} = B_c$$

$$B_{c_1} = \frac{B_c}{\kappa} \quad \text{where } \kappa \text{ is the Ginzburg–Landau parameter.}$$

12. The SQUID (Superconducting Quantum Interference Device) is an extremely sensitive magnetometer.

PROBLEMS

15.1 Show that, because of field penetration, the critical field of a superconducting slab of thickness t is of the order of $B_c\left(1+\frac{\lambda}{t}\right)$, where B_c is the critical field for the bulk sample. (Assume that the free energy difference between the normal and the superconducting states is independent of the specimen size.)

15.2 A thin superconducting ring of inner diameter 1 cm has 10 flux quanta trapped in its central hole. Estimate the value of the persistent current, assuming the penetration depth as 50 nm. [The self-inductance of a loop of radius R made of wire of cross-sectional radius r is, $\mu_0 R \ln(8R/r)$ henry (for $R >> r$)].

15.3 Show that the destruction of superconductivity by applying a magnetic field produces cooling. Calculate the final temperature of a tin sample, first thermally isolated at 2 K and then placed under a magnetic field that exceeds the critical field value. [Neglect latent heat and take $C_n = \gamma T$ with $\gamma = 1.75 \times 10^{-3}$ J mole^{-1} K^{-2})].

15.4 Taking the critical magnetic field for aluminium as 105 gauss at absolute zero, calculate its stabilization energy density at absolute zero. Calculate the heat capacity discontinuity for a unit volume of an aluminium specimen at the critical temperature T_c.

15.5 Calculate the power dissipation in 1 km of a superconducting Nb cable, 10 mm in diameter, maintained at 8 K when it is carrying a current of 1000 A at 50 Hz. Assume that the cable is in a perfect superconducting state. (For Nb, electrical resistivity = 3×10^{-9} ohm m, $T_c = 9.5$ K, $\lambda = 47$ nm).

15.6 The penetration depth of mercury at 3.5 K is about 750 Å. Calculate (a) the penetration depth and (b) the superconducting electron density as the temperature tends to absolute zero.

15.7 The BCS gap of lead at 4 K is about 2 meV. Calculate the frequency of the tunnel current in a lead Josephson junction when it is biased with a voltage equal to one half of the gap potential. Calculate the gap energy at 4.2 K to two significant figures.

15.8 When the current through a superconducting solenoid immersed in liquid helium is increased slowly to 40 A, the solenoid becomes a normal conductor. If the solenoid is 15 cm long with a mean diameter of 3 cm and has 2000 turns, calculate the volume of the liquid that is evaporated in the process.
(Latent heat of helium = 2.5 J cm^{-3})

15.9 How is the current started in a superconducting magnet? Suppose a large superconducting solenoid magnet is below T_c. What would happen if a small hearing aid battery is connected across the solenoid? What will be the result if a large car battery were used?

15.10 When a current I passes through a cylindrical superconducting wire of radius r, the magnetic field produced immediately outside the wire is B_c. If the current is measured in A, radius in cm and the field in gauss, show that

$$I = 5rB_c$$

15.11 Consider a superconducting plate perpendicular to the x-axis and of thickness d placed in a region of an externally imposed magnetic field **B**. If the penetration of field inside the plate

is described by $\nabla^2 \mathbf{B} = \dfrac{\mathbf{B}}{\lambda}$, show that the magnetic field at any point x inside the plate can be expressed as

$$B(x) = B_0 \frac{\cosh (x/\lambda)}{\cosh (d/2\lambda)}$$

where the centre of the plate is at $x = 0$ and λ denotes the penetration depth. Show that the effective magnetization $\mathbf{M}(x)$ in the plate is given by

$$\mu_0 \mathbf{M}(x) = \mathbf{B}(x) - \mathbf{B}_0$$

$$= -\frac{\mathbf{B}_0 (d^2 - 4x^2)}{8d^2}$$

provided $d << \lambda$.

15.12 A Josephson junction of rectangular cross-section and of thickness d is placed in the region of a magnetic field B_0, applied normal to an edge of width b. If the phase difference between the two superconductors be $\pi/2$ when $B_0 = 0$, prove that the dc current in the presence of the field is given by

$$I = I_0 \frac{\sin (bdB_0 e/\hbar)}{bdB_0 e/\hbar}$$

SUGGESTED FURTHER READING

Aschroft, N.W. and N.D. Mermin, *Solid State Physics* (Saunders College, 1988).

Burns, G., *High-Temperature Superconductivity* (Academic, 1991).

Ibach, H. and H. Lüth, *Solid State Physics* (Springer, 1995).

Ramakrishnan, T.V. and C.N.R. Rao, *Superconductivity Today* (Wiley Eastern, 1992).

Schrieffer, J.R., *Theory of Superconductivity* (Benjamin/Cummings, 1983).

Tinkham, M., *Introduction to Superconductivity* (Krieger, 1980).

Chapter 16

Nanomaterials

16.1 INTRODUCTION

Materials with at least one of the dimensions measuring less than 100 nm (1 nm = 10^{-9} m) are known as nanomaterials. Alternatively, we define nanomaterials as those which have a characteristic length scale within about 100 nm. A particle diameter, grain size, layer thickness, or width of a conducting line on a device are some examples of length scale. Examples of some common nanostructures with their approximate sizes are listed in Table 16.1. The properties of bulk materials are mostly retained till the reduction of their dimensions to the micrometre range. But materials in the nanometre scale may show remarkably new properties. For example, a nanocrystal of gold serves as an excellent low temperature catalyst whereas the bulk gold does not. Bulk semiconductors transform into insulators if their dimensions shrink to a couple of nanometres. The melting temperature of crystals in the nanometre scale is very low, lower by 1000°C in some cases. The lattice constants are, too, reduced. The role of surface energy in thermal stability assumes a great significance, because in this scale the number of atoms or ions occupying the surface turns out to be a big fraction of the total number of atoms or ions. The crystalline order that is stable at a high temperature in bulk stabilizes at a much lower temperature in the respective nanomaterial. As a result, the properties such as ferroelectricity

Table 16.1 Examples of some common nanostructures with their approximate sizes*†

Nanoparticle	*Size* (nm)	*Nanoparticle*	*Size* (nm)
Gas ion salts	0.1–1	Paint pigment	10^2–5×10^3
Sugars	~ 1	Pet dander	10^3–5×10^3
Quantum dot	1–5	Red blood cells	~10^4
Micelles	1–10	Microparticle	10^3–10^4
Pyrogens	1–15	Bacteria	5×10^2–5×10^4
Micromolecules	1–80	Yeast cells	5×10^3–5×10^4
Colloid silica	5–80	Pollens	10^4–10^5
Virus	10–100	Macroparticle	10^4–10^5
Cooking smoke	10–10^3	Sand	~ 10^5
Macromolecules	10^2–10^3	Human hair	10^4–10^5
Clay	10^2–10^3		

* Microscopy and Histology Catalog, Polysiciences, Washington, PA, 1993–1994

† N. Itoh, in Functional Thin Films and Functional Materials: New Concepts and Technologies, ed. D.L. Shi, Tsinghua Univ. Press and Springer-Verlag, Berlin, p.1 (2003).

and ferromagnetism may disappear when the dimensions of materials having these properties are cut down to fall in the nanometre scale. It should surprise none that a nanoscale wire or circuit component does not necessarily obey the venerable law of electricity, the Ohm's law.

With the realization of the characteristic properties of nanomaterials, not seen in bulk materials, a revolution in the growth of nanoscience and its technology has occurred. Nevertheless, the historical survey reveals that the existence of nanomaterials and functional devices based on them dates back to the inception of life itself. For example, the cells near the nose of trout fish are believed to possess a fluid of magnetic nanoparticles. The fish takes advantage of the feel of change in magnetization when it deviates from the direction of the Earth magnetic field. It changes its orientation according to need (say, for avoiding the enemy) under the guidance of brain that stores the information on the change of magnetization. Humans too started exploiting the properties of nanosized materials as early as the fourth century A.D. when Roman glass-makers fabricated the classic Lycurgus cup, named after a king and now displayed in the British Museum in London. The cup contains nanosized metal particles and changes its colour from green to deep red on shining its inner surface with light. Even though the results of initial human effort might not have been always fortuitous, there is no denying the fact that the characteristic properties of nanoparticles remained less understood for a long time. The scenario was almost static for three decades even after the advent of quantum mechanics whose application has ultimately yielded all the exclusive information on nanomaterials. It was the brilliance and imagination of Richard Feynman that broke the ice in 1960 at a meeting* of the American Physical Society. Perhaps, in view of his reputation for jokes cracking the audience did not take him seriously when he stated that a cube of side 0.02 inch could store all the books ever written if the material of the cube were constructed by adding individual atoms with the use of a finely tuned technique. He gave the thought of electron beam lithography indirectly while professing for etching lines a few atoms wide with electron beam. Even then, the spin-off in the fabrication of nanostructures could occur only after two decades of this prediction, albeit a beginning made at Bell Labs and IBM in 1970. Feynman is known primarily for laying the foundations of quantum electrodynamics that fetched him Nobel Prize in the year 1965. His contribution of giving the concept of nanotechnology has too been suitably acknowledged by creating THE FEYNMAN PRIZE to recognize the achievements of nanoscientists.

The last two decades have seen a remarkable growth of knowledge in nanoscience and its applications. As a result of this buzzing activity, the nano has moved from the world of the future to the world of the present. It is exemplified by the arrival of computers that compute algorithms to mimic human brain, biosensors that detect the onset of disease and traject drugs attacking the diseased cells on site, and nanorobots capable of repairing internal damages in our body and extricating chemical toxins. These simple-to-use devices successfully detect even the dangerous biotoxins such as anthrax. Other fascinating achievements of nanotechnology include the development of new superconductors with higher transition temperatures and the emergence of nanoelectronics. The latter has led to a drastic miniaturization of devices and is being used to constantly monitor our environment. On popular commercial front, nano skin creams and scentan lotions are already in the market whereas tennis balls with nanoenhanced bounce arrived at the 2002 Davis Cup. The wide range of these applications unambiguously shows that the current advanced form of nanotechnology is essentially the outcome of an intense interdisciplinary and multi-disciplinary effort. A coordination of so many wide-ranging disciplines is not known in the history of science and technology. With such a united force behind to propel it, nanotechnology holds the

* R.P. Feynman, *Eng. Sci.* **23**, 22 (1960).

promise to produce unforeseen commercially viable technology and satisfy the everlasting human curiosity at the same time.

16.2 NANOPARTICLES

A group of 10^6 or less number of atoms or molecules bonded together in a cluster with the radius of about 100 nm (1000Å) or less forms a nanoparticle. The role of the small colloidal particles of silver in image formation on a photographic film is common knowledge. The presence of silver and gold nanoparticles in the material of Lycurgus cup, referred earlier, accounts for the change in its colour on shining its inner surface with light. The metal oxide nanoparticles of different sizes in stained glass windows produce different beautiful colours because a particle scatters only that wavelength which compares with its size. This goes on to corroborate the observation that nanoparticles show unique properties when their size is smaller than certain critical lengths corresponding to respective physical phenomena. To quote an example, the electron mean free path that strongly controls the electrical conductivity of a metal is a critical length.

The understanding of structure and properties of nanoparticles is crucial for the interpretation of the characteristic phenomena exhibited by them. A brief description of some well-studied nanoparticles is given below.

16.2.1 Metal Nanoclusters

The procedure of making the clusters of metal atoms is demonstrated in Fig. 16.1. The metal surface is melted and the atoms evaporated by making a high power laser beam fall on the metal surface. A burst of helium gas is used to carry the evaporated atoms through a narrow opening into an evacuated chamber where the vapour of atoms expands and cools to form clusters. A mass spectrometric study of these clusters is carried out to determine their stability in terms of the number of atoms contained in a cluster. For this purpose, the clusters are first ionized by exposing them to a UV radiation and then pushed into a mass spectrometer. The data reveal that the number of ions (counts) is large for certain numbers of atoms in a cluster. These numbers indicate the number of

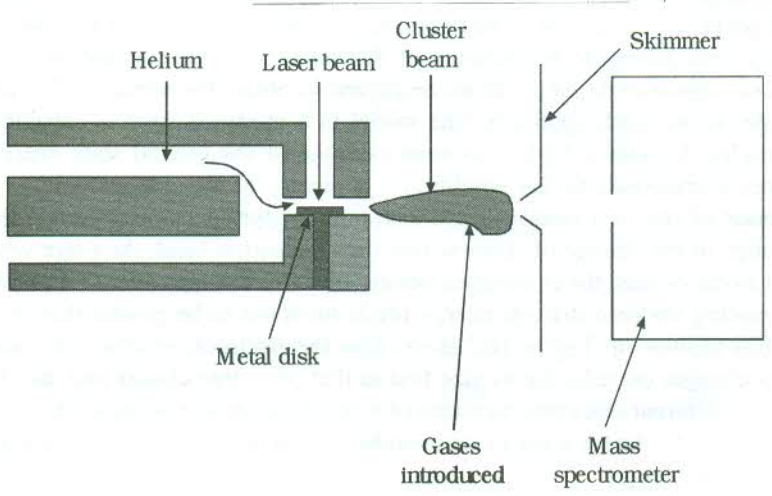


FIG. 16.1 Illustration of the experimental set-up for making metal clusters. (Reprinted from *Introduction to Nanotechnology*, C.P. Poole, Jr., and F.J. Owens, John Wiley, 2003, with permission from John Wiley).

atoms that combine to form a stable cluster. The atoms combine in different numbers to produce stable clusters of different sizes. These different numbers are called *magic numbers*. The magic numbers result mostly from the enhanced stability of closed-shell electronic configurations. In rare gas clusters the magic number arises from geometric effect.

There exist two well-known theoretical models to calculate the properties of clusters: the jellium model and the density functional model (based on the molecular orbital theory). In jellium model the clusters are viewed as superatoms. A sphere approximates the size and the shape of a certain cluster and the positive nuclear charge of each atom of the cluster is considered to be uniformly distributed over the sphere. The interaction between the electron and the positive spherical charge distribution is denoted by a spherically symmetric potential. The solution of Schrödinger equation yields a different order of energy levels for clusters as compared to that for the hydrogen atom. The electronic magic numbers (the total number of electrons in a superatom) correspond to those cluster sizes where all the energy levels are completely occupied.

In the other model the electronic wavefunction for the cluster is constructed on the principles of molecular orbital theory. The simplest of clusters, one could think of, is a molecule such as H_2^+ because the Schrödinger equation can be solved exactly for a hydrogen atom. The ground state wavefunction of the hydrogen atom is expressed as

$$\Psi_{1s}(\mathbf{r}) = Ae^{-\mathbf{r}/a_0} \tag{16.1}$$

where $\mathbf{r}$ gives the position of the electron, relative to that of the nucleus and a_0 is the Bohr radius.

The wavefunction of the electron in molecule is taken as a linear combination of the ground state wavefunctions (16.1) of the two individual H atoms. The problem now reduces to solving the Schrödinger equation using the Hamiltonian

$$\mathcal{H} = \left(-\frac{\hbar^2}{2m}\right)\nabla^2 - \frac{e^2}{r_a} - \frac{e^2}{r_b} \tag{16.2}$$

where r_a and r_b are the distances of the electron from the two nuclei 'a' and 'b'.

The Schrödinger equation with the above Hamiltonian is solved to calculate the actual electronic and geometric structures of small clusters. As the cluster size grows, the exercise assumes a complex shape and getting to the solution of Schrödinger equation ceases to be straightforward. However, several approximate methods are employed to obtain the solution. The use of the density functional model is one such approach. The model in a modified form is capable of dealing with metal nanoparticles. It gives a highly accurate estimate of the ground state energy. The structure with this energy corresponds to the equilibrium geometry of the nanoparticle.

The reduction of size of a metal particle with bulk properties below a certain level brings about a dramatic change in the density of states within the conduction band. At a size when a particle has a few hundred atoms or less, the continuous density of states changes over to a set of discrete energy levels. If the spacing between discrete energy levels turns out to be greater than the thermal energy k_BT, a gap begins to show up. Figure 16.2 shows how the electronic structure of a metal particle with bulk properties changes on reducing its size first to that of a large cluster and then to one of a small cluster. Radically different electronic structure of a small cluster is essentially the quantum size effect that can be observed for larger sizes in semiconductors on account of the wavelength of carriers in semiconductors being longer.

Barring small changes in lattice parameters, there is generally no difference in the crystal structure of a large nanoparticle and that of its bulk. But particles having diameters less than 5 nm may have

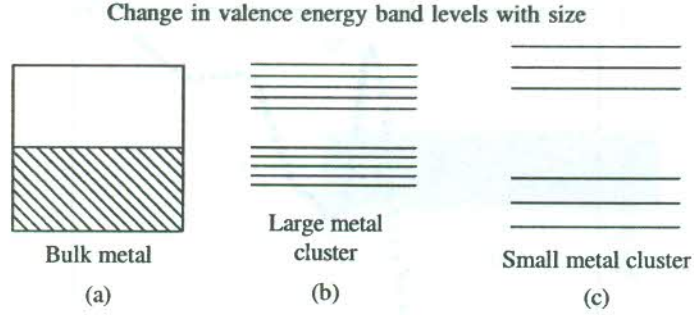


FIG. 16.2 Effect of reducing the number of atoms in a metal sample on the distribution of the valence band energy levels. The bulk sample (a) does not show the gap; in a large cluster (b) containing 100 atoms, the gap becomes distinct and it grows further for a small cluster (c) composed of only three atoms.

a different crystal structure. For example, gold particles of 3–5 nm sizes are shown to have an icosahedral structure and not the bulk FCC structure. Additionally, a high catalytic is switched on in gold nanoparticles as they approach less than 3 to 5 nm size. However, the size at which the transition to bulk behaviour occurs seems to vary according to the property under study.

Also, the evidence for the magnetic behaviour of nanoparticles is well documented. But the observation of a net magnetic moment in clusters comprised of non-magnetic atoms is, perhaps, one of the most remarkable properties of nanoparticles. As the number of atoms in a cluster falls below 20, there occurs a pronounced enhancement in the magnetic moment of these clusters.

16.2.2 Semiconductor Nanoparticles

Here, we refer to nanoparticles that are made of the normal constituents of semiconductors. The constituents are semiconducting elements (Si, Ge, etc.) or the semiconducting compounds (GaAs, CdSe, etc.). The nanoparticles themselves may not behave as semiconductors as is exemplified by the behaviour of the nanoparticles of Ge, Si and Cd. The method of synthesis used for metal nanoparticles works equally well for these particles.

The optical properties of these particles are most interesting. The properties undergo a remarkable change as the transition from bulk to nanoscale occurs. The excitation of excitons strongly modifies the absorption spectra of these particles. When the particle radius is larger than the exciton size, with the exciton motion confined to a limited range, a blue shift (shift towards higher energies) in the spectrum occurs. On the other hand, when the particle size is smaller than the orbital radius of the electron-hole pair, the electron-hole coupling vanishes and the exciton exists no more, releasing the electron and the hole for independent movement. Because the electron and the hole have separate schemes of energy levels, new lines appear in the absorption spectrum, showing again a blue shift. The optical absorption spectra of CdSe nanoparticles recorded at 10 K for two particle sizes are shown in Fig. 16.3. For the smaller particle, we observe that the absorption is more intense and the absorption edge is blue shifted, which is the indication of a larger band gap in this particle. The spectra also reflect the effects of the confinement of exciton. The exciton lines being represented by peaks at higher energies shift towards further higher energies on reducing the particle size. In a particle of smaller size, the electron and the hole are drawn closer and the energy level separations change.

The nanoparticles of Ge and Si are found to undergo fragmentation on being irradiated with a Q-switched Nd-YAG laser light. For example, a Si_{12} particle breaks into two Si_6 particles. The emerging products depend on the cluster size, the wavelength and the intensity of laser beam. A

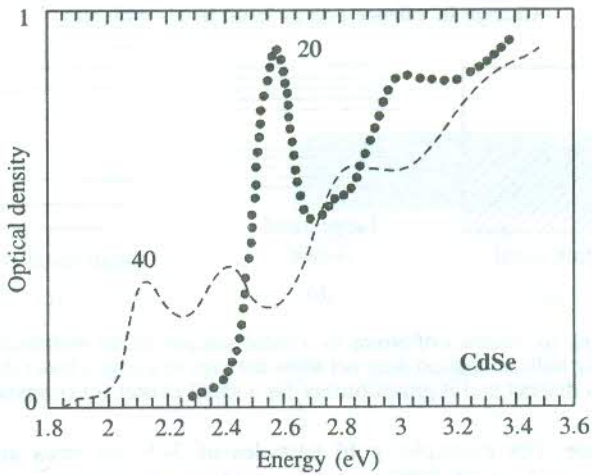


FIG. 16.3 Size effect on the optical absorption spectrum of CdSe at 10 K. The figure shows the spectra for two nanoparticles (sizes: 20 Å, 40 Å) [Reproduced from D.M. Mittleman, *Phys. Rev.*, **B49**, 14435 (1994)].

cluster with more than 30 atoms tends to fission explosively. This is most likely to happen in multiply ionized clusters where each atom gets the positive charge through the quick redistribution of charges on the atoms following the process of ionization. When the Coulombic repulsion between the atoms in a cluster grows beyond the binding energy between the atoms, the cluster explodes and the atoms fly apart with great velocities. The phenomenon is popularly identified as 'Coulombic explosion'.

16.2.3 Other Nanoparticles

Nanoparticles can be formed from rare gases and molecules such as water and carbon dioxide. Some of these are stable even in multiply charged form. Examples of such largest clusters are: $(Kr_{73})^{2+}$, $(Xe_{206})^{4+}$ and $(CO_2)^{4+}_{216}$. Water is comprised mostly of its molecular clusters where the hydrogen atom of one molecule bonds with the oxygen atom of another molecule, such that the hydrogen atom is not equidistant from the two oxygen atoms. Under normal conditions about 80% of water molecules are clustered.

Nanoparticles are also known to show some unusual properties such as superfluidity. It has been observed in He clusters with 64 and 128 atoms. In the state of Bose–Einstein condensation (the state where all the bosons occupy the lowest energy level), the liquid He^4 at 2.2 K becomes a superfluid, losing its viscous property completely.

16.3 NANOSTRUCTURES

16.3.1 Carbon Clusters

In 1985, Kroto and Smalley with their collaborators recorded mass spectra of laser ablated graphite that showed several intense lines corresponding to the clusters of different sizes. Calculations on small clusters based on the density functional theory demonstrated that the clusters form in linear or closed non-planar monocyclic structures. The linear clusters have sp hybridization with carbon

bond angle as 180°. These clusters contain an odd number of atoms and those with 3, 11, 15, 19 and 23 are more stable as reflected by the large intensity of respective lines in the mass spectrum (Fig. 16.4). The two other forms of hybridization occurring in carbon molecules are sp^2 and sp^3, with carbon bond angles measuring 120° and 109° 28′ respectively. Clusters with closed structures contain an even number of atoms and the carbon bond angles are different from the three standard hybridization values given above. One such cluster composed of 60 atoms is found to show the highest intensity in the mass spectrum and hence the most stable. The evidence for the existence of this molecule was first found while measuring the intensity of light coming from stars at different wavelengths. A pronounced reduction in the intensity of light at 220 nm falling on Earth was detected. The phenomenon referred to as 'optical extinction' was attributed to scattering by small graphite particles believed to be present in the interstellar dust filling the regions of the outer space between stars and galaxies. The IR studies involving the use of graphite arcs were carried out for further clarification and confirmation. Huffman and Kratschmer observed four IR lines that correspond closely to the lines of a molecule C_{60}, predicted theoretically to exist several years earlier. Other investigations using techniques including the mass spectrometry were then made that led to the confirmation of the existence of C_{60} molecule in 1990.

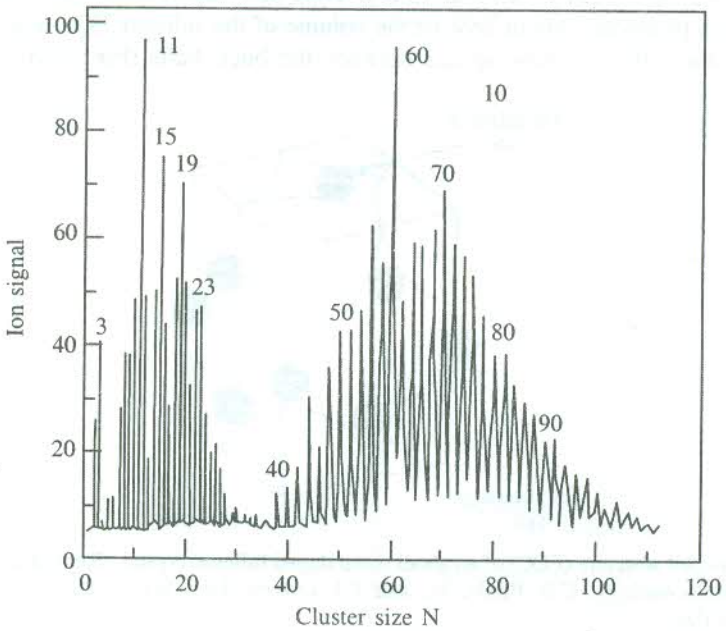


FIG. 16.4 A mass spectrogram of carbon clusters (Reproduced from *Microcluster Physics*, S. Sugano and H. Koizumi, Springer-Verlag, 1998 with permission from Springer).

On the basis of the analysis of their mass spectrum, the team of Kroto, Smalley and collaborators proposed a soccer ball shape and the icosahedral symmetry for C_{60} molecule. According to this model, the three-fold coordinated carbon atoms are bonded to form a spherical surface composed of 12 pentagons and 20 hexagons. The formation of pentagons lends the otherwise graphite sheet bend into a closed surface having no dangling bonds. It resembles geodesic domes whose structure was studied by the American architect Buckminister Fuller. In view of this similarity, the C_{60} molecule was named 'Buckministerfullerene'. The similarity of its geometric configuration with that of a soccer

ball forms the basis for its familiar name 'Buckyball'. The C_{60} molecule shows a single line C^{13} NMR spectrum which strongly confirms the buckyball model. Simply, because the mass spectra showed the evidence for many other even-numbered carbon clusters as well (e.g. C_{70}), the general name 'fullerenes' has been adopted to represent the group of these cage-shaped carbon clusters.

The structure of C_{60} molecule (the buckyball) is shown in Fig. 15.28. Crystalline C_{60} is formed from an ordered array of the tightly bound buckyballs, which are bound together by weak van der Waals forces. The crystals have FCC symmetry and retain many of the molecular properties contrary to the behaviour of other crystalline forms of carbon:diamond and graphite. The buckyballs are highly resistant to photofragmentation and very resilient to collisions against surfaces. The mechanical properties of solid C_{60} (also called 'fullerite') are characterized light, weak and soft while comparing with other forms of carbon. The electronic energy bands show a slight broadening. The fullerite is dubbed as a large bandgap semiconductor. It does not absorb in the near IR and into the visible region where the thin films show orange colour due to the transitions near the bandgap energy (minimum 1.5 eV). In the powder form, the colour depends on the crystallite size and varies from grey to black.

The fullerite is a poor conductor of electricity. But it turns into a superconductor when doped with alkali metals as discussed in Section 15.8.4. Thus new superconductors with T_c in the range 20–50 K have been produced. About 26% of the volume of the fullerite FCC unit cell is empty and the alkali metal atoms fit into these spaces between the buckyballs (Fig. 16.5).

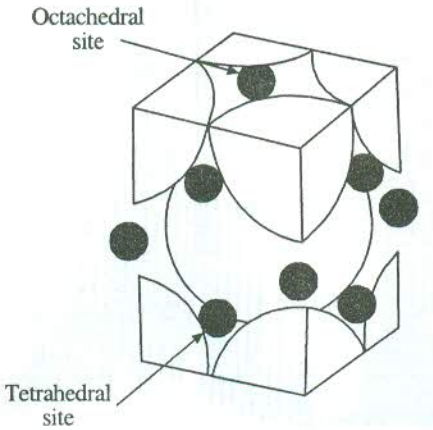


FIG. 16.5 The unit cell structure (FCC) of an alkali metal doped fullerite crystal. (Reproduced from *Introduction to Nanotechnology*, C.P. Poole, Jr., and F.J. Owens, John Wiley, 2003, with permission from John Wiley).

Till date, the fullerenes have not been applied on a large scale. Their biggest potential lies in their use as building blocks for synthesizing new diverse organic chemicals and solid materials. Some of the other projected applications are: photochromic goggles, xerographic materials, lubricants, sensors, ion rocket propulsion, diamond seeding, new semiconductors, patterned superconductors, patterned diamond films and anti-viral agents.

16.3.2 Carbon Nanotubes

Carbon nanotubes have the potential for wide-ranging applications due to their interesting properties. The nanotubes form generally as multiple-walled nanotubes (MWNT) whereas the single-

walled nanotubes (SWNT) are fabricated by mixing traces of some catalysts—cobalt, nickel or iron with the graphite target.

There are three well-known techniques of synthesizing carbon nanotubes: laser evaporation, carbon arc and chemical deposition method. The set-up used in evaporation method is sketched in Fig. 16.6. Argon gas and a graphite target is enclosed in a quartz tube. The tube is centred inside a horizontal furnace maintained at 1200°C with portions toward the ends falling outside the furnace. The target contains small amounts of cobalt and nickel that serve as the catalytic nucleation centres. Carbon atoms are evaporated from the graphite target with the help of an intense laser beam incident on the target. A water-cooled copper collector is mounted within the tube at one of its ends, lying outside the furnace. The large temperature difference between the middle and the end portions of the tube facilitates the argon gas to sweep the carbon atoms to the cold copper collector, resulting in the formation of nanotubes by condensation. The tubes are typically 10–20 nm in diameter and 100 μm long. This method as well as the carbon arc method produces tubes with closed ends. The chemical vapour deposition method, on the other hand, gives tubes with open ends.

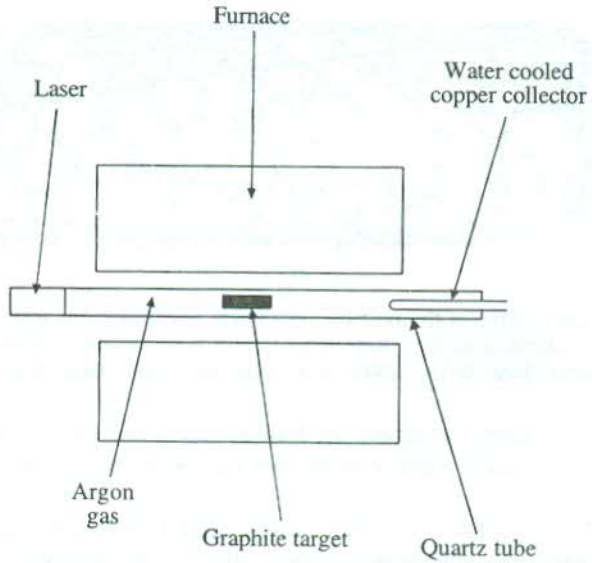


FIG. 16.6 Illustration of the laser evaporation method of synthesizing carbon nanotubes.

The carbon nanotubes are imagined to have been formed by rolling the graphite sheet the bonds at whose ends may join to close the tube. In the actual process of formation, however, no rolling of the sheet occurs. The above imagination only serves to explain various observed structures of nanotubes. The axis about which the graphite sheet is rolled is represented by the axis vector lying on the sheet plane. For example, three possible structures of carbon nanotubes are shown in Fig. 16.7. The structure shown in Fig. 16.7(a) (and referred to as the armchair structure) results when the axis vector is parallel to the C–C bonds of the carbon hexagons. The zigzag structure [Fig. 16.7(b)] and the chiral structure [Fig. 16.7(c)] are formed for two other orientations of the axis vector.

The molecular symmetry of carbon nanotubes classifies their vibrations into two normal modes denoted by A_{1g} and E_{2g} representations. A_{1g} mode represents the 'in and out' oscillations of the diameter of the tube. In the other normal mode (E_{2g}), the tube's cross section oscillates between the

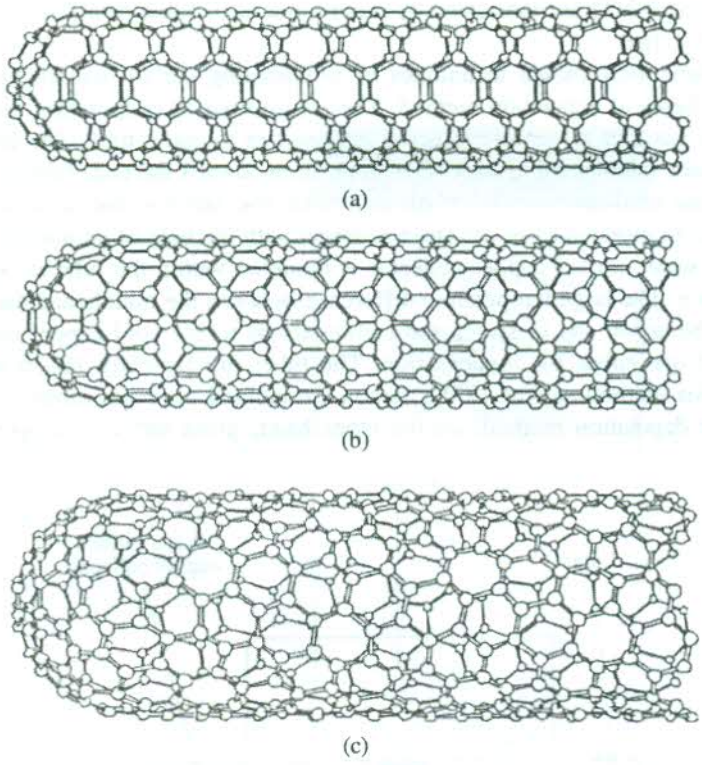


FIG. 16.7 Common structures (three of the possible) of carbon nanotubes: (a) armchair structure, (b) zigzag structure and (c) chiral structure. (Reproduced from *Introduction to Nanotechnology*, C.P. Poole, Jr. and F.J. Owens, John Wiley, 2003; with permission from John Wiley).

circular and elliptical shapes. Both the modes are Raman active and their frequency depends on the tube diameter. It is now a standard practice to exploit this property for measuring the diameter of carbon nanotubes.

The Young's modulus of carbon nanotubes ranges from 1.28 to 1.8 terapascal (one terapascal (Tpa) $\approx 10^7$ atmospheric pressure), a value that is about 10 times that for steel. Thus carbon nanotubes are very stiff and resilient to bending. They buckle without breaking. Since their wall structure has a very low density of defects, they do not fracture when bent severely. The tensile strength is very high (about 45 billion pascals), making them about 20 times stronger than steel. The single-walled nanotubes are stronger than their multiple-walled counterparts.

Generally, carbon nanotubes form as a mixture of semiconducting and metallic tubes in the ratio 2:1. The diameter and the chirality of a tube determine the conducting behaviour of the tube. Chirality describes the manner in which the graphite sheet is rolled with respect to the axis vector. The metallic nanotubes have the armchair structure shown in Fig. 16.7(a). The bandgap of semiconducting chiral nanotubes decreases with increasing diameter. The electronic structure is studied using scanning tunnelling microscopy (STM). The data show that the electronic states split into one-dimensional sub-bands instead of forming a single wide band. The electronic states turn out to be the solutions of the Schrödinger equation where the length of the nanotube is taken as the depth of the involved potential well.

Individual single-walled nanotubes show interesting electronic transport. The voltage-current characteristics of a single metallic nanotube placed across two metal electrodes show step-like features which is similar to that of a field effect transistor (FET) constructed from a carbon nanotube. The metallic nanotubes can carry extremely high current with density approaching 10^9 A/cm^2. The density of defects in carbon nanotubes is so low that the defects cannot cause electron localization and the electrons do not suffer much scattering from defects. This is the cause for high electrical conductivity of nanotubes. The current does not heat carbon nanotubes the same way as in copper, which fails at 10^6 A/cm^2. Carbon nanotubes are also very good conductors of heat, the thermal conductivity being almost twice as large as that of diamond.

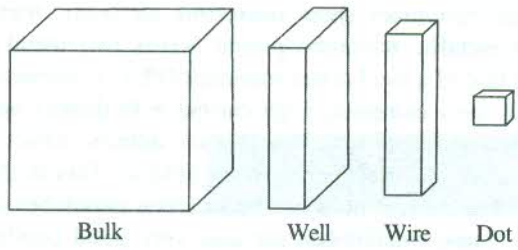
Carbon nanotubes are used to enhance the strength of plastic composites. They serve as good catalysts to speed up the rate of some chemical reactions. They are also applied in battery technology. The field-effect transistor made of chiral semiconducting nanotubes has proved to be a sensitive detector of many gases. Most of the challenging applications are in the fields of electronics and computers. Some have already been realized and several more are envisioned. For example, electrons are emitted at a high rate from the ends of a nanotube when a small electric field is applied along its axis. The effect, known as 'field emission', has contributed to the development of flat panel displays. Based on this effect, vacuum tube lamps are being made in Japan which are equally bright as the common light bulbs and have a higher efficiency and a longer life.

Very low resistance of carbon nanotubes makes them ideally suitable for use as interconnects and their high thermal conductivity for use as heat sinks for chips. The fabrication of field effect transistors from carbon nanotubes has raised vision for very fast computer switches. One of the major challenges of nanotubes has been physical assembly because they tend to stick to one another and do not exhibit the molecular recognition properties more generally associated with organic molecules. Elsewhere, the computer scientists are striving to increase the number of switches on a chip, pressing for large-scale production of thinner nanotubes. Methods for large-scale production of multiple-walled nanotubes are well developed. But those for single-walled nanotubes have very low yields, raising their cost exorbitantly. The application potential of carbon nanotubes can be truly realized with the development of methods for large-scale production of single-walled nanotubes, in view of their more remarkable features.

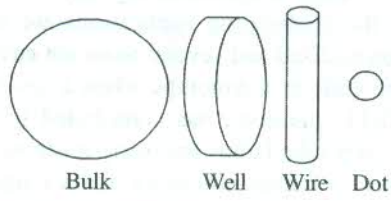
16.3.3 Quantum Nanostructures

Three limiting classes of nanostructures emerge on the basis whether only one, only two or all three dimensions fall in nanometre scale. When only one dimension is reduced to nanometre range and two remain large, we get the material called *quantum well*. If two dimensions are in nanometre range and one large, the structure is known as *quantum wire*. But if all three dimensions are made to be in nanorange, we refer it to as *quantum dot*. The rectangular and curvilinear quantum nanostructures are shown in Fig. 16.8. The word 'quantum' here signifies that the materials with these structures possess new physical properties or exhibit new physical phenomena which their bulk counterparts do not. As such, the word quantum draws parallel to getting new results with the use of quantum mechanics in situations involving small (quantum) numbers and small (physical) quantities.

There are two approaches to the synthesis of nanomaterials and the fabrication of nanostructures: bottom-up and top-down. The bottom-up approach generally uses a sequence of catalysts controlled chemical reactions. The procedure involves the collection of individual atoms and molecules and making them condense into a specific pattern or structure. The top-down approach employs the opposite method where the size of a large-scale object or pattern is reduced gradually.



(a)



(b)

FIG. 16.8 Formation of (a) rectangular and (b) curvilinear quantum structure by progressive reduction in size.

The colloidal dispersion is the simplest example of the bottom-up method whereas attrition or milling describes best the top-down approach to making nanoparticles. Though nanolithography and nanomanipulation are commonly taken as the bottom-up approach, the lithography should in fact be treated as a hybrid approach since the growth of thin film is bottom-up and the etching is top-down.

The quantum wires and quantum dots are fabricated by the attrition of a thin film (quantum well) of the sample material supported by its substrate. The procedure followed in the electron-beam lithography is illustrated in Fig. 16.9. In the first step, the sample quantum well is coated with a radiation-sensitive resist (e.g. polymethyl methacrylate). Then, the central region on the surface of the resist, under which some specific nanostructures is desired, is irradiated by an electron beam through a template. The irradiation transforms the resist material in this region chemically to a form that is soluble in a developer. The irradiated part of the resist is removed, using the developer. Next, an etching mask is planted into the groove or pit formed as a result of the removal of the resist. The remaining part of the resist is now taken out and the region of the quantum well uncovered by the etching mask is chemically etched away. The resulting structure is a quantum nanostructure which may be a quantum wire or a quantum dot, depending on the shape and the size of the groove. If required, the mask may be lifted. Figure 16.9 shows the fabrication of a quantum dot.

Quantum structures of more complex type have shown some remarkably improved properties, making them attractive for device fabrication. The lithographic technique has generally been used to get these structures. For example, a quantum dot array derived from a multiple quantum well structure produces a very enhanced photoluminescent output that is about hundred times greater than that of the initial multiple quantum wells. The property has been exploited in the fabrication of quantum dot lasers. The genesis of the remarkably different properties of quantum nanostructures lies in their drastically reduced dimensions. As the size on nanoscale reduces, the fraction of atoms on the surface of the nanocrystal grows larger and larger. Consequently, the number of conduction electrons

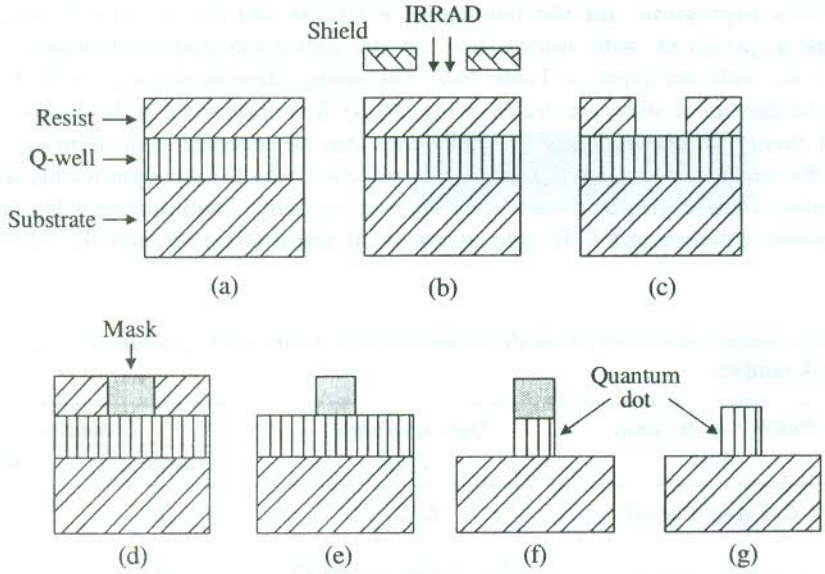


FIG. 16.9 Stages of formation of a quantum structure in electron-beam lithography: (a) a substrate supported quantum well, covered by a resist, (b) irradiation of the sample that is shielded by a template, (c) form after dissolving the irradiated portion of the resist, (d) placing the etching mask, (e) configuration after removing the unirradiated resist, (f) structure after etching away the undesired portion of the quantum well, (g) quantum dot structure on substrate obtained by removing the etching mask. (Reproduced from *Introduction to Nanotechnology*, C.P. Poole, Jr. and F.J. Owens, John Wiley, 2003; with permission John Wiley).

in metals and semiconductors drops considerably and the electron mean free path becomes size-dependent. When the diameters of microcrystallites are much larger than the mean free path, a network of interconnected resistors would be the appropriate representation for the resistivity of polycrystalline sample. But if the dimensions match the mean free path, the scattering from the boundaries between crystallites controls the resistivity in a big way.

It has been shown that an n-type semiconductor quantum dot cube of side 10 nm with 10^{18} donors per cm^3 will have only one conduction electron. Similar estimates have been made for quantum wells and quantum wires. This leads to drastically different forms of density of states that are so vital to the determination of many physical properties of solids. When the delocalization of electrons in a free electron Fermi gas is considered, the density of states in one, two and three dimensions of bulk metals are found to differ very much from each other. However, the consideration of the electron confinement in broad potential wells yields qualitatively a common discrete energy spectrum. The energy spectrum is independent of dimensionality and shape of the potential wells. The presence of both the potential well confinement and the Fermi gas delocalization in the same nanostructure has been found, an evidence for the partial confinement. As a result of confinement in one or two directions, the electrons move along the confinement coordinate directions to occupy the respective potential well levels (i), lying below the Fermi level and tied to the degeneracies d_i. At the same time, the electrons in each case delocalize in the remaining dimensions as they occupy the Fermi gas levels in the delocalization direction of the reciprocal lattice.

Quantum dots exhibit total confinement as against no confinement in bulk materials. Quantum wire and quantum well structures provide only partial confinement to electrons. This behaviour

leads to different expressions for the number of electrons and the density of states for these structures. The expressions with information on delocalization and confinement in quantum nanostructures and bulk are given in Table 16.2. The energy dependence curves of the number of electrons and the density of states are drawn in Fig. 16.10. It is observed that the number of electrons increases with energy for all structures. The behaviour may be accepted as qualitatively similar. But the forms of the density of states are dramatically different which may significantly affect certain properties. Some of these properties are: electron heat capacity, Pauli susceptibility, the intensity of x-ray emission, electron and hole concentrations in semiconductors and the superconducting energy gap.

Table 16.2 Some features of electrons delocalized/confined in quantum wells, quantum wires, quantum dots and bulk samples

Sample	*Number of electrons* $N(\varepsilon)$	*Density of states* $D(\varepsilon)$	*Dimensions*	
			Confined	*Delocalized*
Q-well	$C_1 \Sigma d_i(\varepsilon - \varepsilon_i(w))$	$C_1 \Sigma d_i$	1	2
Q-wire	$C_2 \Sigma d_i(\varepsilon - \varepsilon_i(w))^{1/2}$	$\frac{1}{2} C_2 \Sigma d_i(\varepsilon - \varepsilon_i(w))^{-1/2}$	2	1
Q-dot	$2 \Sigma d_i \Theta(\varepsilon - \varepsilon_i(w))$	$2 \Sigma d_i \delta(\varepsilon - \varepsilon_i(w))$	3	0
Bulk	$C\varepsilon^{3/2}$	$\frac{3}{2} C\varepsilon^{1/2}$	0	3

$$C = \frac{V}{3\pi^2}\left(\frac{2m}{\hbar^2}\right)^{3/2}; \quad C_1 = \frac{L^2}{2\pi}\left(\frac{2m}{\hbar^2}\right); \quad C_2 = \frac{2L}{\pi}\left(\frac{2m}{\hbar^2}\right)^{1/2}$$

$\Theta(x) = 0$ for $x < 0$ and 1 for $x > 0$
$\delta(x) = 0$ for $x \neq 0$, and ∞ for $x = 0$, and integrates to a unit area.
$\varepsilon_i(w)$ refers to the energy of the level i in the potential well.

There are a large number of applications of quantum well structures in electronics, quantum electronics and photonics where they form heterojunctions. Discrete energy levels of quantum wells have been used for the fabrication of modern infrared detectors. The quantum well and quantum wire-based lasers are other important arrivals where the confinement and the localization of conduction electrons in discrete energy levels is finely tuned. The production of quantum dot lasers is one of the most recent advancements in this direction. In view of their increasing utility, a brief discussion on their design and principle of working is very much in order.

A schematic diagram showing the structure of a quantum dot laser as fabricated by Park et al.* is given in Fig. 16.11. A quantum dot laser diode is assembled on a p-type GaAs substrate. The top layer is derived from a p-type metal with the GaAs contact-layer immediately underneath. There is a 190 nm thick waveguide of $Al_{0.05}Ga_{0.95}As$ that centres a 30-nm thick GaAs region (marked QD) comprising 12 monolayers of $In_{0.5}Ga_{0.5}As$ quantum dots of surface density 1.5×10^{10} per cm^2. A

* G. Park, O.B. Shchekin, S. Csutak, D.L. Huffaker and D.G. Deppe, *Appl. Phys. Lett.*, 75, 3267 (1999).

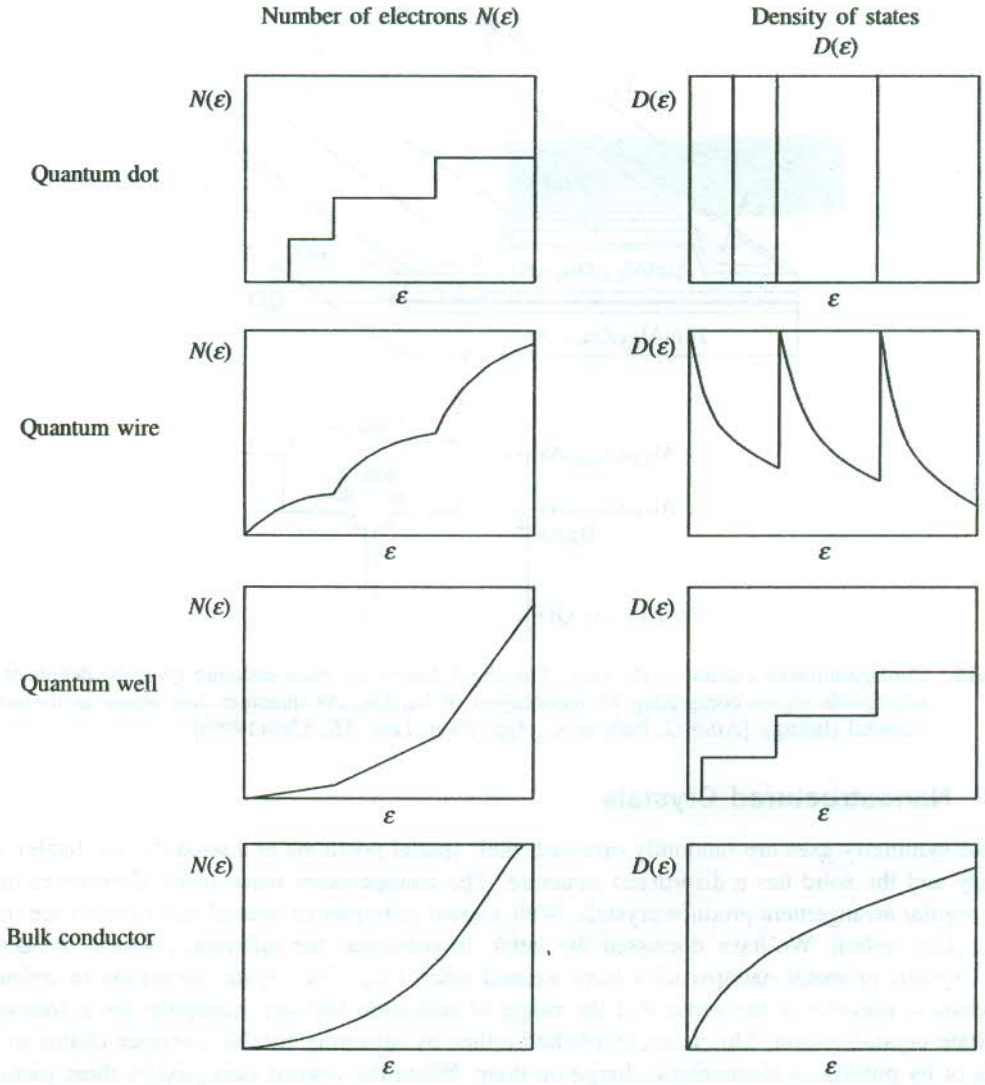


FIG. 16.10 Variation of number of electrons $N(\varepsilon)$ (left column) and of density of electron $D(\varepsilon)$ (right column) with change in energy in square well—Fermi gas approximation for three quantum structures and their bulk counterpart.

pair of 2 μm thick $Al_{0.85}Ga_{0.15}As$ layers is used to bound the waveguide. The pair lies between the contact layer and the substrate (not shown). The coatings of $ZnSe/MgF_2$ at the faces of the laser cause the internal back and forth reflection of light to step-up the stimulated emission. The cavity is typically 1 to 5 mm long and 4 to 60 μm wide. The quantum dots assume the role of active atoms as the neon atoms do in a helium-neon laser. The lasing action is excited through the lateral edge of the device. The threshold current for CW operation at room temperature is close to 4 mA and the output radiation happens to fall in near infrared region around 1.32 μm.

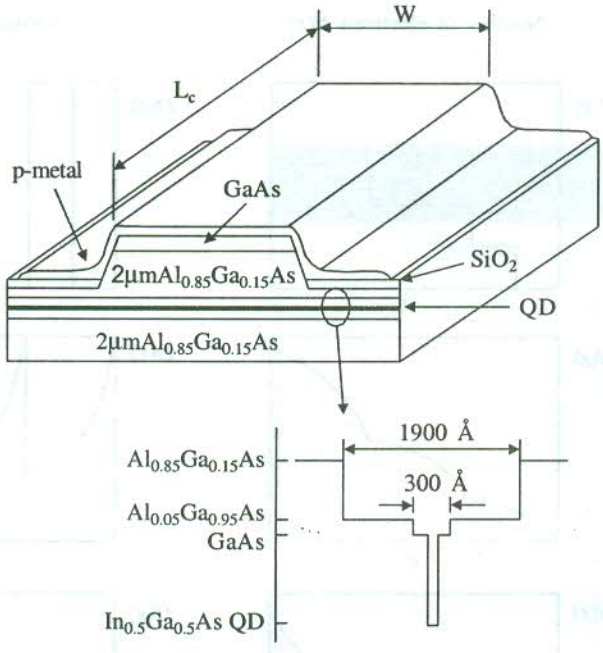


FIG. 16.11 Configuration of a quantum dot laser. The sketch below the main structure gives the details of the waveguide region comprising 12 monolayers of $In_{0.5}Ga_{0.5}As$ quantum dots acting as the active material (lasing). [After G. Park et al., *App. Phys. Lett.*, **75**, 3267(1999)].

16.3.4 Nanostructured Crystals

When the symmetry axes are randomly oriented, their spatial positions in a solid do not display any symmetry and the solid has a disordered structure. The nanoparticles when order themselves in an infinite regular arrangement produce crystals. Well-known examples of natural nanocrystals are those of boron and carbon. We have discussed the latter, in particular the fullerene crystals, in Section 16.3.1. Crystals of metal nanoparticles have existed since long. The crystal formation in colloidal suspensions is peculiar in the sense that the range of repulsion between nanoparticles is increased to facilitate crystallization. This is accomplished either by attaching soluble polymer chains to the particles or by putting an electrostatic charge on them. When the volume occupied by these particles becomes more than half the total volume, the particles start ordering themselves in a regular pattern and crystallize. In view of the limited scope of the book, only two very special classes of these crystals are treated below.

Nanostructured zeolites

The cubic mineral faujasite $(Na_2, Ca)\ (Al_2Si_4)O_{12} \cdot 8H_2O$ is a typical representative of zeolites. These materials are porous with pores arranged in a regular array. Small clusters can be accommodated in these pores where the van der Waals force between the clusters and the zeolite provide stabilization to the clusters. The material desired to fill the pores is added to the melt of the zeolite. Zeolites are often referred to as molecular sieves since their structure enables them to sift materials. Their structure resembles nanoscopic galleries or chambers interconnected by nanoscopic tunnels or pores (see Fig. 16.12).

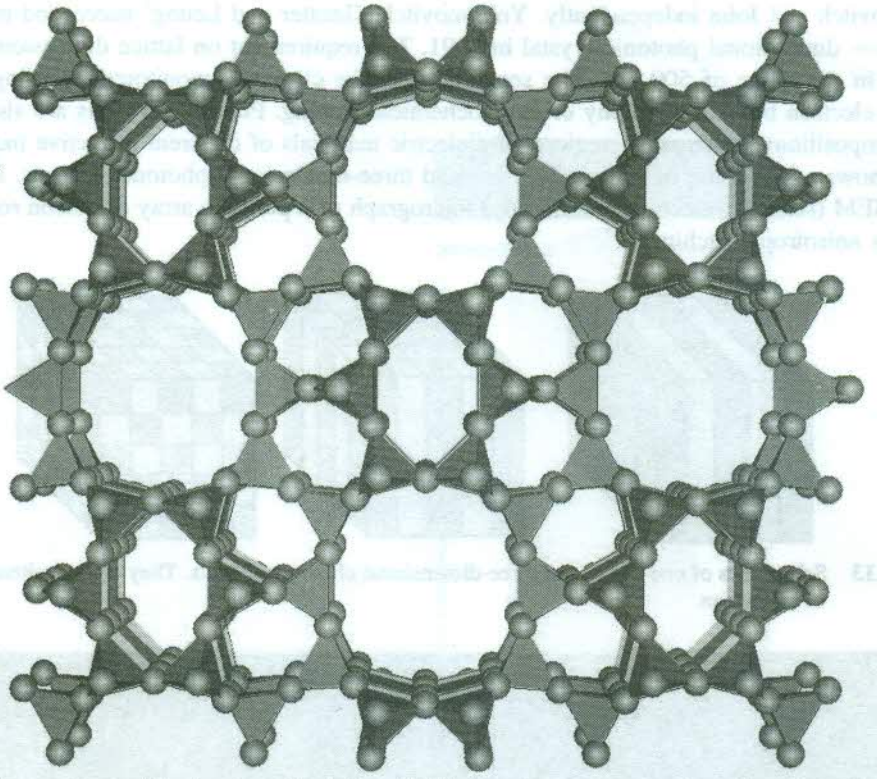


FIG. 16.12 A complex zeolite structure. The holes of different size represent various channels and galleries. (Courtesy Geoffrey Price, Unitersity of Tulsa)

Tremendous interest has been shown in using zeolite structures for directed catalysis whereby petroleum is produced more efficiently and one can select desired products from the broad distribution of petroleum components. The particles of a catalyst are found within the galleries of the zeolite in the catalysis. The locally controlled reactivity by the catalyst particles combined with the physical constraints of the gallery and the pore size selects the formation of hydrocarbons of certain specific shape and composition. More efficient use of feedstocks, less wastage and low cost are the salient features of the process of catalysis by design.

The property of zeolites to facilitate the exchange of calcium ions by sodium ions is made use of to reduce water hardness, promoting them as domestic water softeners. Zeolites, in fact, are one of the first broad-scale, highly profitable applications of nanotechnology.

Photonic crystals

Dielectric nanoparticles arranged on a lattice form photonic crystals. The separation between the nanoparticles is artificially adjusted to match the wavelength of the electromagnetic waves (or photons) that the crystal is meant to deal in an application. For this lattice dimension, the electromagnetic waves will suffer Bragg reflection at the Brillouin zone boundaries where an energy gap opens up (refer to Fig. 7.6). The idea of a photonic crystal was first given in 1987 by

Yablonovitch and John independently. Yoblonovitch, Gmitter and Leung[†] succeeded in getting the first three- dimensional photonic crystal in 1991. The requirement on lattice dimensions for visible light is in the range of 500 nm. The separation can be clinically monitored by using techniques such as electron beam lithography or electrochemical etching. Photonic crystals are also fabricated with composition of alternating regions of dielectric materials of different refractive indices. Figure 16.13 shows a schematic of such one-, two- and three-dimensional photonic crystals. Figure 16.14 shows SEM (scanning electron microscopy) micrograph of a periodic array of silicon rods produced by deep anisotropic etching.

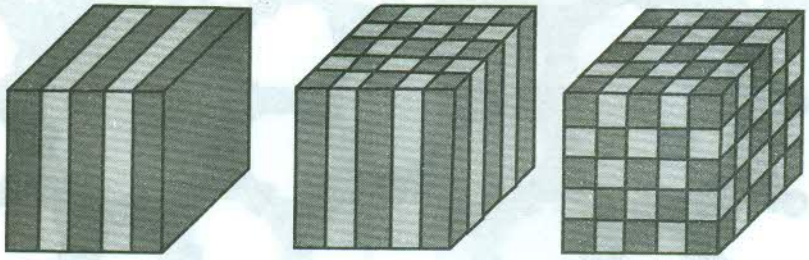


FIG. 16.13 Schematics of one-, two- and three-dimensional photonic crystals. They contain alternating regions of dielectrics.

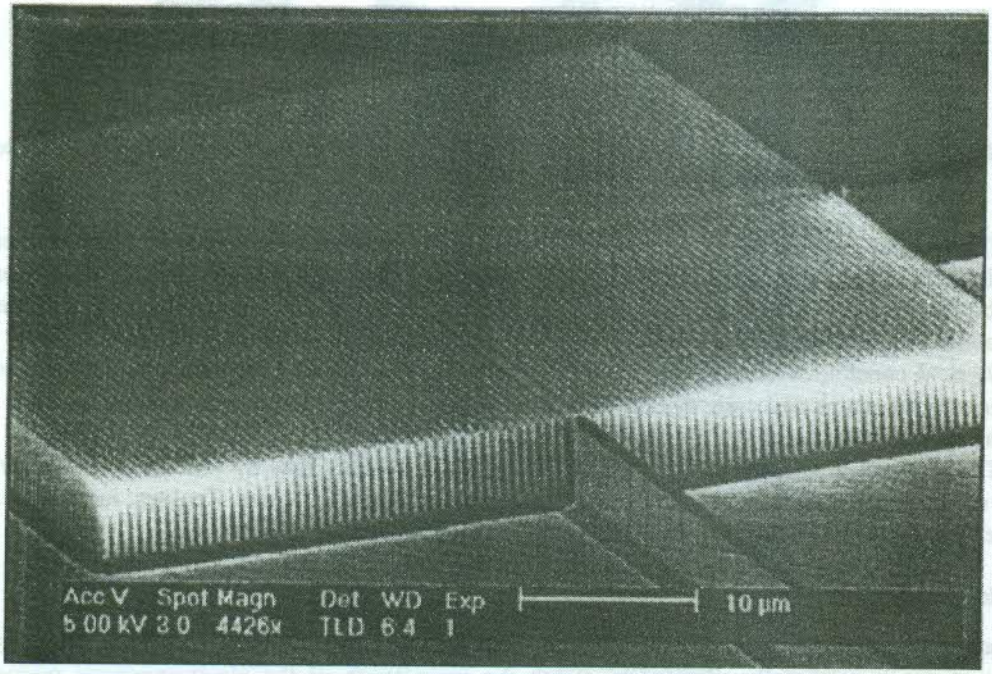


FIG. 16.14 SEM micrograph of a periodic array of silicon pillars made by using deep anisotropic etching. Pillars diameter is 205 nm and height 5 µm. Band gap ~ 1.5 µm for TM polarization. [After T. Zijlstra et al., *J. Vac. Sci. Technol.*, **B17**, 2734 (1999)].

[†] E. Yablonovitch, T.J. Gmitter and K.M. Leung, *Phys. Rev. Lett.*, **67**, 2295 (1991).

Features of electromagnetic wave propagation in photonic crystals are obtained by solving Maxwell's equations in a periodic dielectric medium. In absence of external current sources, the associated Helmholtz equation reads as

$$\nabla^2 \mathbf{H}(\mathbf{r}) + \epsilon \left(\frac{\omega^2}{c^2} \right) \mathbf{H}(\mathbf{r}) = 0 \tag{16.3}$$

where ϵ is the macroscopic dielectric constant of the medium and $\mathbf{H}$ is the magnetic vector associated with the electromagnetic wave.

Because the photon–photon interaction is negligibly small, Equation (16.3) can be solved exactly and the dispersion relation determined accurately. Figure 16.15 shows the dispersion curves for the transverse magnetic modes in a square lattice of aluminium rods designed for microwaves. The symbols Γ and X correspond to special symmetry points in k-space for the square lattice. It is expected that the visible light would be treated similarly at the needed smaller rod separations. In the photonic band structure, there exists a forbidden gap, marked as the photonic band gap in Fig. 16.15. The gap excludes the existence of optical modes within the frequency range spanned by the gap.

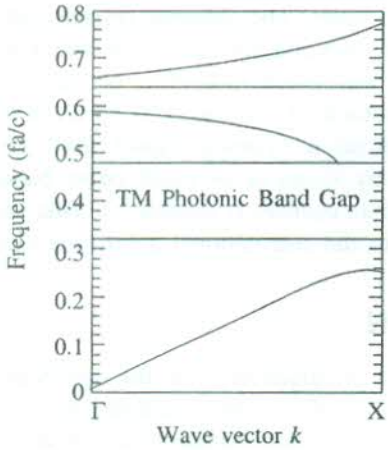


FIG. 16.15 Dispersion curves for a photonic crystal TM mode in a square lattice of aluminium rods designed for microwaves; f stands for frequency and a is the lattice constant. [Taken from J.D. Joannopoulos, *Nature*, **386**, 143 (1997)].

Creating a line imperfection in a photonic crystal amounts to forming a waveguide that may enable the crystal to transmit or focus the light of certain frequency. This is done in the lattice of aluminium rods by pulling a row of rods out of the structure. In the modified dispersion curves, a localized energy level (similar to the donor level or the accepter level in semiconductors) appears in the forbidden gap (see Fig. 16.16). The level, in the present case, is often referred to as 'the guided mode'. The frequency of the mode depends on the radius of the rod. The guided mode acts as a pipe which confines the electromagnetic waves and allows them to move along one direction. The confinement in the waveguide is total whereas the light may escape from an optical fibre when the fibre is too much bent, and the requirement on the angle of incidence for total internal reflection is not met. The waveguide may as well be used as a resonant cavity. The frequency of the cavity can be tuned by monitoring the gap.

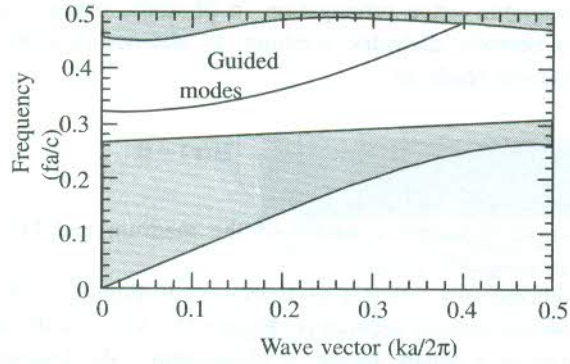


FIG. 16.16 Creation of a level in the forbidden gap (the guided mode) on removing one row of rods from a square lattice of an aluminium photonic crystal. f stands for frequency and a is the lattice constant. [Taken from J.D. Joannopoulos, *Nature 386*, 143(1997)].

These properties provide for blocking the propagation of photons irrespective of polarization direction, localizing photons to a specific area at restricted frequencies, for example using photonic crystals as filters and couplers in lasers. The photonic crystals can also be made to manipulate the dynamics of the spontaneous emission and the stimulated emission independently for lasing action. They serve as an ideal lossless waveguide to confine or direct the propagation of light along a specific channel. Because the photonic crystals work at all wavelengths, they have strong potential for application in telecommunication systems. Apart from providing lossless and noiseless propagation, an optical channel by virtue of its much larger bandwidth carries many times more information than what an electronic channel is capable of. This projects photonics as a technically much advantageous alternative to the conventional solid state electronics.

16.4 POROUS SILICON

Although porous silicon was first produced in 1956, the significant interest in the material became evident only in 1984 with the report of its visible photoluminescence. Porous silicon is also known to display electroluminescence and cathodoluminescence. It appears very surprising because the ordinary silicon has a band gap of 1.125 eV at 300 K and shows emission in the near infrared region. In 1990, Canham brought forward the importance of this phenomenon with special reference to its possible role in the improvement of electronics, based on reducing the dimensions and increasing the functional speed. In a device with reduced dimensions, the interconnects may become three to four times larger than the actual device whose features then get dominated by properties and failure modes in interconnects. The emission of visible light (above 1.4 eV at room temperature) by porous silicon raised the vision of replacing metal interconnects by optical interconnects to overcome this problem. Prospects brightened with the record of high quantum efficiencies (1–10%) in contrast to the low values ($\sim 10^{-4}$%) for ordinary silicon, an indirect gap material involving phonons in interband transitions. The grossly modified luminescence of porous silicon is, however, interpreted mainly as a consequence of the presence of quantum well, quantum wire and quantum dot structures and the surface defect states.

The current electronic technology is heavily based on silicon because of several advantages it holds over other semiconductor materials. Its abundance, low cost, good thermal and mechanical

properties and excellent passivating characteristics provide silicon a distinct edge over widely used group III–V and II–VI compounds. In addition, the porous silicon light emitting structures can be fabricated using a relatively simple and inexpensive etching technique. Also, the structures can be easily integrated with thriving silicon technology. These have created an extraordinary interest in the study of porous silicon in the recent past.

Electrochemical etching is one of the simplest techniques used to produce porous silicon. A silicon wafer is deposited on a metal at the bottom of the cell made of teflon or polyethylene. The silicon wafer is etched by dilute HF solution contained in the cell at a constant current. The silicon is used as the positive electrode and platinum forms the other electrode (see Fig. 16.17). Canham used a two-step etching process, involving the anodization in dilute HF at low current densities in the first step and the open-circuit pore widening in the next. The open-circuit etching carried out in the dilute HF increases the pore size and produces a blue shift in the photoluminescence. But it is now settled that the pore widening is not required for visible luminescence, though the emission peak is blue shifted in the sample etched over a longer duration. Factors determining the nature and the size of the pore have been identified as: the type of extrinsic silicon (*n*-type or *p*-type), the concentration of HF in the electrolyte, the order of current and the presence of a surface-active agent (surfactant). A p-type silicon after etching contains a fine network of pores measuring typically less than 10 nm.

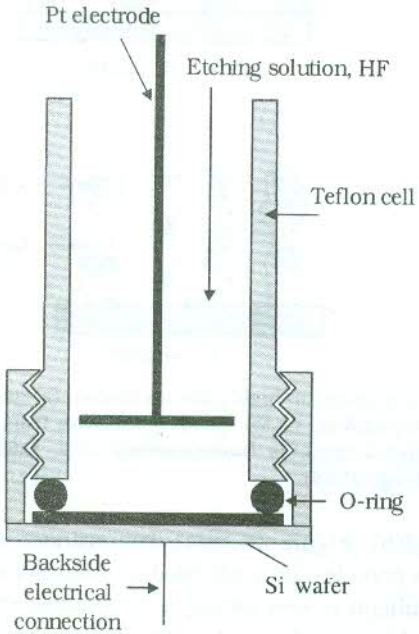


FIG. 16.17 Structure of the cell used for etching a silicon wafer in HF solution to make it porous. (Reprinted from *Handbook of Nanostructured Materials and Nanotechnology*, H.S. Nalwa (Ed.), Academic Press (2000), Vol. 4, Chapter 3, p. 173; with permission from D.F. Thomas (an author) and Elsvier).

Depending on doping and on etching conditions to some extent, specimens of porous silicon can have significantly different structures and surface properties. Some of the common structures of porous silicon are shown in Fig. 16.18. The first of these is the quantum wire structure [Fig. 16.18(a)] with 78% porosity, formed by open-circuit etching. Undulating wires approximate the alternative

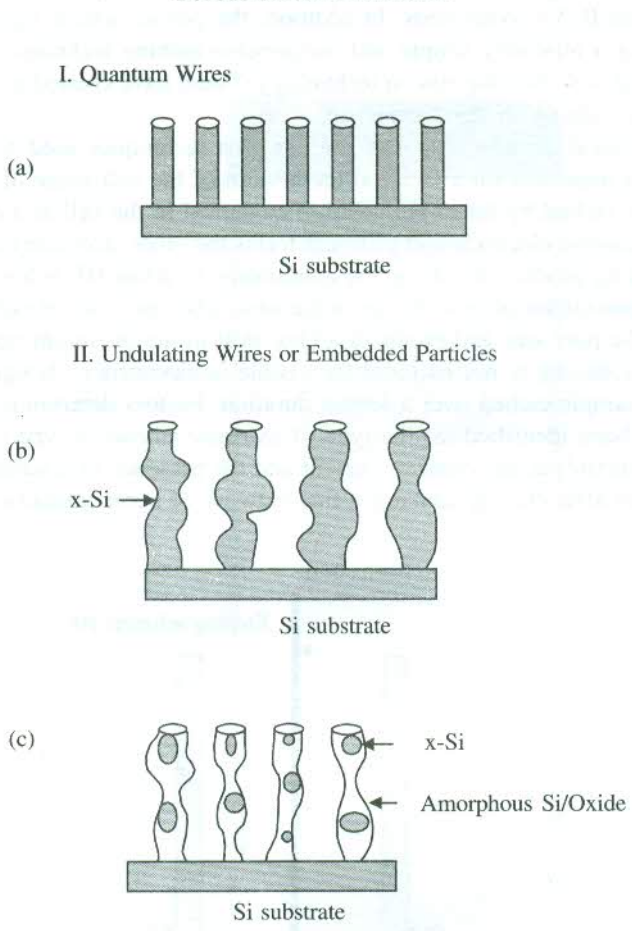


FIG. 16.18 Some possible microstructures of highly porous silicon: (a) quantum wires, (b) 'undulating' wires, and (c) embedded crystallites, (x-Si). (With permission from S.M. Prokes (the author) and IPB, *Porous silicon nanostructures, in Nanomaterials*, A.S. Edelstein and R.C. Cammarata (Eds.), Inst-Physics Publishing, 1998).

structure shown in Fig. 16.18(b). Figure 16.18(c) demonstrates another highly porous structure wherein the crystalline silicon particles are embedded in amorphous silicon or oxide.

Technologically, porous silicon is very attractive for photoelectronic applications. The stability of luminescence under atmospheric conditions and the way it changes on prolonged irradiation are crucial to the fabrication of devices. The exposure to atmosphere and prolonged irradiation both are known to affect the intensity of luminescence in porous silicon adversely. At the present stage, considerable development is required to minimize the effect of these parameters. However, the production of efficient electroluminescent devices holds the key to the ultimate adoption and success of porous silicon. The devices produced so far have unattractive efficiencies that follow from the large requirement on current densities, because the particles are isolated and the electrons must tunnel through oxide.

16.5 PHYSICAL TECHNIQUES OF FABRICATION

Many synthesis routes have been used for the fabrication of nanostructures. Some of them are based on chemical approaches while in others, physical techniques are applied. In this section, we describe a couple of physical techniques, which have assumed importance based on their use for microelectronic fabrication in semiconductor industry. The techniques selected for a proper discussion are photolithography and scanning probe microscopy.

16.5.1 Photolithography

As per our interpretation in Section 16.3.3, photolithography is a hybrid of the bottom-up and top-down approaches. It is also referred to as "photoengraving", essentially a process of transferring a pattern into a reactive polymer film, called resist. The resist is used to replicate the pattern into an underlying film or substrate. Lithographic approaches using various lens systems and exposure radiation sources have been applied, all working on similar principles. Among these, photolithography happens to be the most commonly applied technique in the mass production of integrated circuits.

The basic steps of photolithographic process are illustrated by Fig. 16.19. The base is first thin-coated with a resist whose selected areas are then exposed to light through a mask in an image-wise fashion. The developer is now brought in contact with the exposed resist. A positive tone or a negative tone image of the mask is obtained after the developing process, depending on whether the exposed areas of the resist are more or less soluble in the developer as compared to the unexposed areas. The three-dimensional relief image thus produced in the resist material is a replica of the opaque and transparent areas of the mask. The parts of the resist that remain after developing mask the underlying areas of the substrate and prevent them from attack during the subsequent

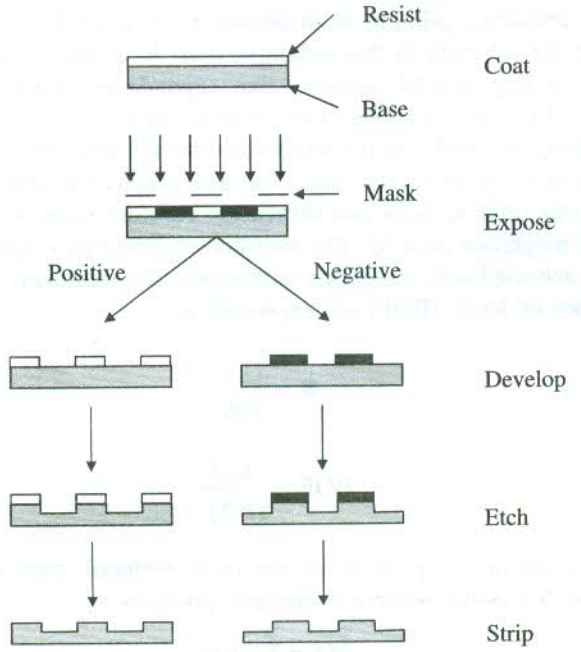


FIG. 16.19 Steps in the photolithographic process of transferring images in the mask to the substrate surface. (After G. Cao, *Nanostructures & Nanomaterials*, Imperial College Press, 2004)

etching operation. When the resist is removed by stripping, the relief image emerges in the underlying substrate. In addition to polymer films, Langmuir–Blodgett films and self-assembled monolayers are other resists that have been used in photolithography. In these applications, photochemical oxidation, generation of reactive groups or cross-linking are used to transfer patterns from the mask to the monolayers. Photolithography is practised in two basic approaches: (1) shadow printing which itself may be of two types, contact-mode printing and proximity printing and (2) projection printing.

The minimum size of elements that can be obtained by photolithography is determined by the limit of resolution in the diffraction pattern extended into the geometrical shadow. For shadow photolithography using a mask consisting of equal lines and spaces of width e, the theoretical limit of resolution is given by,

$$2e_{\min} = 3\sqrt{\lambda\left(w + \frac{t}{2}\right)} \tag{16.4}$$

where $2e$ is the grating element, w the gap width maintained between the mask and the photoresist surface, λ the wavelength of the exposure radiation and t the photoresist thickness. In the case of hard contact printing ($w = 0$) for 1 μm thick resist film and the light of wavelength 400 nm, the maximum resolution obtained in accordance with eqn. (16.4) is slightly less than 1 μm. The resolution provided by other photolithographic techniques can at best approach but not exceed the value obtained with a contact-mode photolithography, because the mask and the wafer are in intimate contact in this method, enabling the transfer of a mask pattern into a photoresist with utmost accuracy. The presence of dust on the substrate and the non-uniformity in thickness of the photoresist and the substrate degrade the resolution which in practice hardly approaches its maximum value. The proximity printing is an answer to these problems. But the gap introduced between the mask and the substrate in this technique must be controlled with precision for desired results. The increase in gap beyond certain value expands the penumbral region, created by diffraction and worsens the resolution. The diffraction in shadow printing is typically Fresnel type.

In projection printing, the mask and the wafer substrate are separated by several centimetres and lens elements are used to focus the mask image. The diffraction concerned in the technique belongs to Fraunhofer class. Lens imperfections and diffraction considerations account for relatively lower resolution provided by projection printing. The well-known Rayleigh's equation is used to determine the limit of resolution achieved with projection techniques. The minimum resolution feature (R) and the corresponding depth of focus (DOF) are expressed as:

$$\mathrm{R} = \frac{C_1\lambda}{\mathrm{NA}} \tag{16.5}$$

$$\mathrm{DOF} = \frac{C_2\lambda}{(\mathrm{NA})^2} \tag{16.6}$$

where C_1 and C_2 are constants, depending on the resist material, process technology and image formation method, and NA is the numerical aperture given by

$$\mathrm{NA} = \bar{n} \sin \alpha \tag{16.7}$$

Here $\bar{n}$ is the index of refraction of the image space and α is the maximum cone angle of the exposure light beam.

Using shorter wavelengths of light and larger numerical apertures, sufficiently high resolutions can be achieved. Generally, the minimum feature size that can be obtained with large numerical features (greater than 0.5) matches the wavelength of light used or is slightly less than this.

16.5.2 Scanning Probe Microscopy (SPM)

SPM forms the basis for nanomanipulation and nanolithography. The technique differs from other imaging techniques like scanning electron microscopy (SEM) and transmission electron microscopy (TEM) in its ability to manipulate molecules and nanostructures on a surface.

SPM is a combination mainly of the scanning tunnelling microscopy (STM) and the atomic force microscopy (AFM) which treat two extreme classes of materials when used independently. STM is applied to conductors whereas AFM is ideally suited to deal with dielectric materials.

Scanning Tunnelling Microscopy (STM)

STM has evolved on the principle of electron tunnelling that is perceived as a quantum mechanical phenomenon. Consider two flat surfaces of a metal or semiconductor, separated by an insulator or vacuum. Generally, the transfer of electrons between the flat surfaces would not be possible due to the presence of an energy barrier as shown in Fig. 16.20(a). But when a difference of potential is applied between the surfaces, the shape of the energy barrier changes [see Fig. 16.20(b)]. The electrons on one surface are now under the influence of a force and they tunnel through the separating region in compliance with the quantum mechanical diktat that the electron wavefunction on one side of the energy barrier has a non-vanishing amplitude on the other side of the barrier. Specially, when the thickness of the separating region is very small, the electron wavefunctions (electron waves) on two sides of the separating region may overlap, making tunnelling feasible. The tunnelling is exhibited by the flow of a small current (I), expressed as

$$I \propto e^{-2kz} \tag{16.8}$$

where z is the thickness of the separating region and

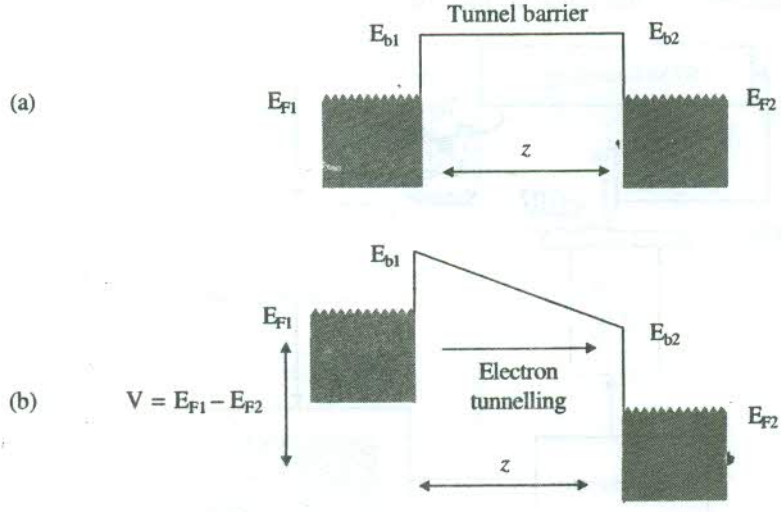


FIG. 16.20 Energy levels in two solids (metals or semiconductors) separated by an insulating or vacuum barrier. (a) in the absence of the electric field (b) when an electric field applied between the solids. E_{F1} and E_{F2} indicate the positions of Fermi level in the two solids.

$$k = \frac{\sqrt{2m(V-E)}}{h} \tag{16.9}$$

Here V is the height of the potential barrier, m the electron mass, E the electron energy and h denotes the Planck's constant. According to relation (16.8), a decrease in the separation (z) by 0.1 nm would increase the tunnelling current by one order of magnitude. This property has been exploited in STM. Binning and Rohrer developed the technique and demonstrated its functioning for the first time in 1982.

In the STM assembly, a conducting wire having a very sharp tip is positioned above the surface of the sample under study. The conducting tip is usually derived from a metal or metal alloy such as tungsten or platinum-iridium alloy. The tip is positively charged and held at a distance about 2–6 nm above the sample surface. Being positively charged, the tip exerts a strong force on electrons in the individual atoms at the surface. Under the influence of this force, the electrons tunnel into the tip, creating the flow of a feeble current (1–10 nA). Thus the tip acts as a probe. The probe tip is moved back and forth in either a constant current imaging mode or an alternating imaging mode. In constant current mode, a feedback loop maintains a constant probe height above the sample profile, involving up/down movement of the probe. The alternating imaging mode is essentially a constant height operation in which the feedback establishes the initial height and the feedback is switched off. As the tip moves, the distance between the tip and the sample surface in this mode keeps varying and so does the current. The alternating current mode provides for faster scan rates whereas the constant current mode is known for producing a contrast directly related to electron charge density profiles.

The tip positions are used to construct a topographic map of the surface. The probe tip is often mounted on a piezoelectric tripod scanner. The piezoelectric arrays accurately control the movement of such a tip in three-dimensions above the sample surface. The method offers scanning resolution of about 0.01 nm in x-y direction and 0.002 nm in z direction. This results in the atomic scale resolution in the three-dimensional image. The first atomic scale resolution was achieved in a STM image of silicon 7 × 7 restructured (111) surface by Binning and collaborators in 1983. The schematic of a STM assembly is illustrated in Fig. 16.21.

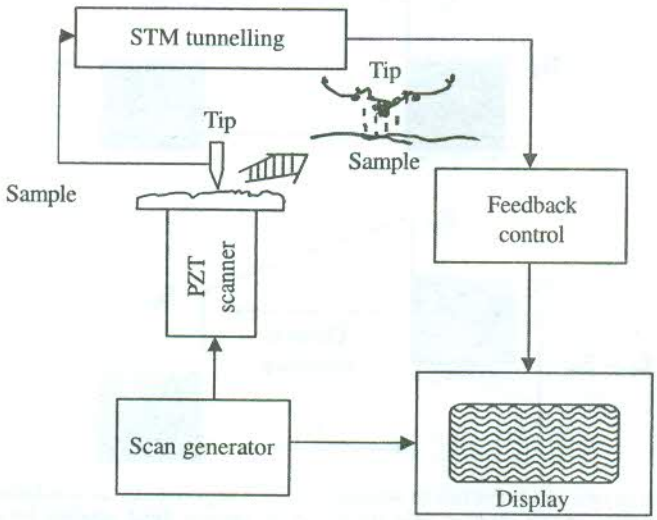


FIG. 16.21 Block diagram of a scanning tunnelling microscope. (After G. Cao, *Nanostructures & Nanomaterials*, Imperial College Press, 2004)

Atomic Force Microscopy (AFM)

AFM is a modification of STM, suitable for application to dielectric materials. The basic difference between the two is that STM monitors the tunnelling current between the surface and the probe tip whereas AFM monitors the force between the surface and the probe tip. Thus AFM is capable of measuring various tip-sample interactions, dominant at respective tip-sample separations. For example, the van der Waals interactions are predominant at small separations and long range forces such as electrostatic attraction or repulsion are significant at large separations. Even capillary forces, in case of condensation of water between the tip and the sample, dominate when the tip to sample distance is increased. The AFM technique is about measuring the motion of a cantilever beam with an ultra small mass and a nanoscale tip. Typically, a force of 10^{-18} N is required for a movement through about 10^{-4} Å. Piezoelectric scanners are used and the images are also created by scanning the tip across the surface. An interference pattern of light beam through an optical fibre serves to monitor the up/down motions of the tip (see Fig. 16.22). The alternative method uses the reflection of a laser beam. The schematic of the technique is depicted in Fig. 16.23. In addition to the laser whose beam is directed at the end of the cantilever, other components of the assembly are—a mirror, a photodiode and an array of piezoelectrics, forming a three-dimensional positioning sample stage.

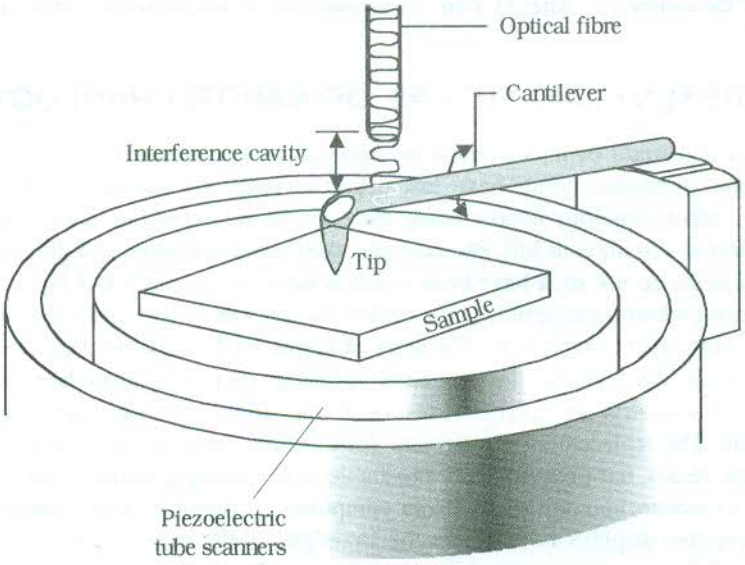


FIG. 16.22 Schematic of an atomic force microscope. A cantilever arm is shown to have the probe tip that traverses the sample surface in coordination with a piezoelectric scanner. (Reprinted from *Introduction to Nanotechnology*, C.P. Poole, Jr. and F.J. Owens, John Wiley 2003; with permission from John Wiley).

The atomic force microscopy is sensitive to the vertical component of surface forces. However, both normal and lateral components of surface forces on the tip can be simultaneously measured with a more versatile microscope called '*the friction force microscope*'. STM and AFM happen to be the prominent members of SPM. The SPM has proved most useful in imaging of surfaces of all kinds of solids and the fabrication and processing of nanostructures.

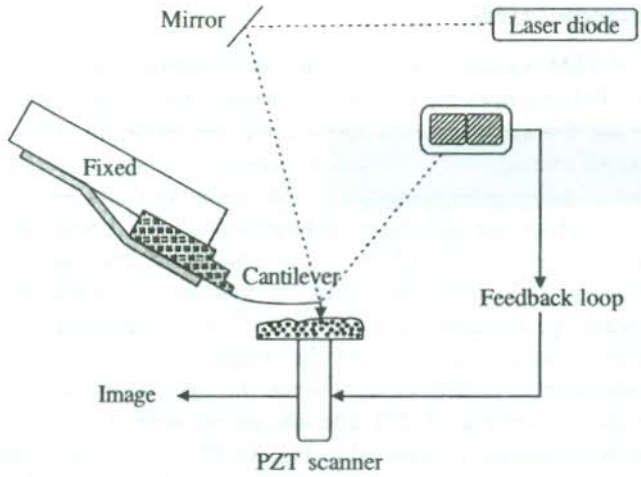


FIG. 16.23 Working of an atomic force microscope. A microscopic positioning device is employed in the assembly where the sample is mounted on a scanner; and the cantilever and tip have to be positioned near the surface. A photo diode records the position of a laser beam, reflected off the top of the cantilever. (After G. Cao, *Nanostructures & Nanomaterials*, Imperial College Press, 2004).

16.6 THE PRESENT AND FUTURE OF NANOTECHNOLOGY

Although the original thought of nanotechnology came about two decades ago, the first few nano-based appliances and consumer goods have just begun to reach the market. Apart from its major impact in industry, nanotechnology is now being felt even in our home and daily life. Coatings and laminates are the first to slip into our life. For example, the coated bathroom and kitchen tiles to which the dirt and grit particles do not stick have been manufactured in Germany and Japan. The tiles have suspensions of antimicrobial nanoparticles that inhibit the growth of rots and fungi, thus helping in maintaining a very high level of sanitation. The entry of harmful UV radiation, a source of skin cancer, is prevented by coating the windows in new Audi A4 series cars with glass laminates.

Applications in industry are fairly widespread and some of them have already become commercially viable. The replacement of catalysts, for example, with zeolites (the molecular sieves), described in Section 16.3.4, has proved highly profitable in extracting gasoline from crude. However, the major support to nanotechnology comes from computers, communications, consumer electronics and medicine. Computer displays have stolen the limelight in the recent past. The heavy TV-like cathode ray tubes (CRTs) are being replaced by more compact flat-panel liquid crystal displays (LCDs) that cause less eye strain and have higher energy efficiency. Samsung and some other companies have started using nanotubes as replacement for scanning electron guns and succeeded in shrinking the screens and reducing the power consumption. Being small, light and efficient, these screens may soon be applied in laptop computers. The nano accounts for the most amazing and spectacular invention in display technology—known as 'digital paper'. It is very likely to create computers with ultrasharp screens. It would be possible to feed digital files in these computers, facilitating the storage of a large number of books and documents that could be read by holding the screen like a book. In order to make up for the poor screen resolutions, new digital displays have been fabricated by using the same chemicals as employed for paper fabrication. This gives the display a paper-like look. The displays employ elements whose bistable characters enable to maintain a programmed

image without consuming additional power. So, the day is not far in signage when the advantages with digital paper, its commercial viability included, would prepare us mentally to lose our love for the touch and feel of a book, raising the vision of a paperless world.

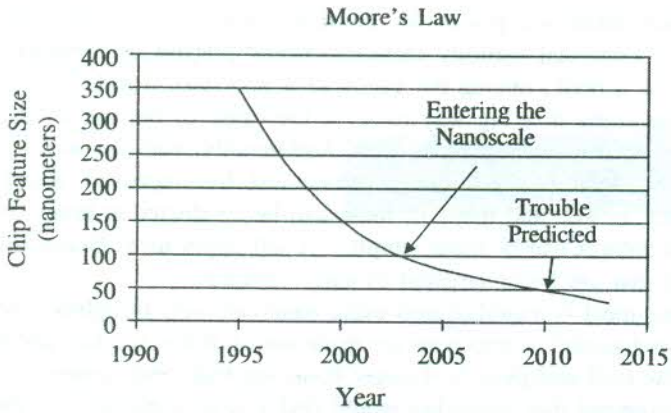
Nanotechnology has come in a big way to serve the cause of human health. Some details to this effect have already been given in introduction. Additionally, with the use of nanoparticles the accuracy of several tests, including the home pregnancy test, has improved remarkably. Tests have become easy and speedy as well and many of these can be conducted at home. It is also believed that tests for AIDS and anthrax can be made simpler for self-application. Braces and Prosthetics are the examples of goods that are being targeted as early ventures.

A lot of sports equipment is manufactured using nanomaterials. In lightweight motorbikes and sailboats, carbon fibre and graphite composites are made use of. Better football and hockey pads have been made using similar stuff and plastics. Longer bouncing balls and stronger rackets are already on tennis courts. It is expected that nanotubes would find a large application in the manufacture of sports goods.

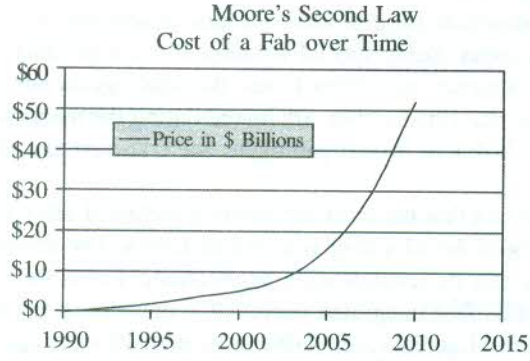
Lastly, our fashion too does not live uninfluenced by nanotechnology. Developments in molecular-scale composite materials have produced a new generation of clothing and brought nano very much into our fashion arena. Some special features of these products are: resistance to stains, strength and durability, convenience and sharp looks. Biocidal agents similar to those incorporated in bathroom and kitchen tiles, described earlier, are mixed during the manufacture of some fabrics that are ideally suited for use in hospitals where pathogens are common and the risk of infection from each other is very high.

There is no denying the fact that the most outstanding technical advancement during the second half of the 20^{th} century has been the advent of silicon electronics. The miniaturization of microchips, the heart of nanoelectronics and its revolutionary applications, forms the centre of development of silicon technology. The world's first integrated circuit (IC), containing one transistor and a few other components, was built by Nobel Laureate Jack Kilby more than 40 years ago on a slice of germanium, measuring (7/16 × 1/16) inches. Today, one can place more than 100 million transistors in less than 1 cm^2 of silicon surface area because of the availability of nano-sized transistors (~ 90 nm long). Such a single IC can implement an entire electronic system and is called *system-on-chip* (SoC). The raw power of semiconductor fabrication technology is only one aspect of challenges. The other equally important aspect concerns the rapid design of a differentiated product and bringing it to the market place.

In 1975, Gordon Moore gave two empirical laws to describe the growth of IC electronics. According to his first law (famous as Moore's law), the amount of space required to install a transistor on a chip shrinks by roughly half every 18 months. Figure 16.24(a) shows how rapidly the size of a feature on a chip is getting smaller with time. While the first law brings joy to smile, the second law paints a gloomy picture as it predicts that the cost of building a chip fabricating plant (also called fab) doubles every 36 months [see Fig. 16.24(b)]. Thus we see that the shrinking is not to come for free. In fact, the soaring manufacturing cost is posing a big challenge. As we shift from microchip to nanochip, it is strongly felt that the basic principles involved in present chip making cannot survive in totality beyond 2010 and a rethinking will become utmost essential. As for present, the revolutionary applications of nanotechnology and its impact on our life clearly indicate that the age of nano is truly upon us. Its great potential is very much likely to instill an instant motivation among scientists for surmounting the future barriers in sight.



(a)



(b)

FIG. 16.24 (a) Illustration of Moore's first law: fastly decreasing chip size with time. Chip size shrinks by roughly half every eighteen months.
(b) Illustration of Moore's second law: increasing cost (Fab) of chip manufacturing plant with time. It doubles every 36 months.

SUMMARY

1. Nanomaterials: at least one dimension < 100 nm.
2. Nanoparticles: metal nanoclusters, semiconductor nanoparticles, rare gas and CO_2 clusters: $(Xe_{206})^{4+}$, $(CO_2)^{4+}_{216}$.

   Cluster Theory: Jellium model, Density functional model.
3. Nanostructures:

   (A) Carbon clusters:

   Linear—sp hybridization, carbon bond angle – 180°, with odd number of atoms

   Closed—carbon bond angles–different from 180°, 120° and 109° 28′ (the standard hybridization values), with even number of atoms

   Example—C_{60} molecule (fullerene)—called 'Buckyball', icosahedral symmetry with spherical surface, composed of 12 pentagons and 20 hexagons.

	Crystalline C_{60} (Fullerite)	:	FCC symmetry; light, weak and soft as compared to other forms of carbon.
	Pure	:	poor conductor.
	Alkali metal dopped	:	superconductor, T_c in 20–50 K range. No application on large-scale till date.
	Biggest potential	:	as building blocks for making new organic materials
(B)	Carbon nanotubes	:	(Single-walled, Multiple-walled) Dia ~ 10–20 nm, 100 μm long
	Armchair structure	:	Axis vector parallel to C–C bonds
	Zigzag structure and Chiral structure	:	For two other orientations of the axis vector Diameter estimated from Raman Spectra

Features:

SWNT are stronger than MWNT ~ 20 times stronger than steal.

Mixtures of metallic and semiconducting nanotubes (2:1)

Defect free metallic nanotubes can carry 10^9 A/cm^2 (copper fails at 10^6 A/cm^2)

For semiconducting chiral tubes: bandgap decreases with increasing diameter.

Challenges: (i) physical assembly, as they tend to stick,
(ii) large-scale production of SWNT at a reasonable cost.

Common applications: to increase strength of plastic composites, catalysts, FET as gas detector in electronics and computers.

(C) Quantum nanostructures

Quantum well	:	only one dimension in nanometre scale
Quantum wire	:	two dimensions in nanometre scale
Quantum dot	:	all three dimensions in nanometre scale
Q-nanostructures	:	have different forms of density of states and exhibit new physical phenomena—not shown by bulk counterparts.
Common applications	:	quantum wells in heterojunctions used in electronics, quantum electronics, photonics; most recent—Q-dot lasers.

(D) Nanostructured crystals

Examples: boron, carbon, zeolites, photonic

Zeolites: porous, pores arranged in a regular array (serve as molecular sieves) used for directed analysis (production of petroleum more efficiently)—the first broad-scale profitable application of nanomaterials.

Photonic crystals: (Dielectric nanoparticles arranged on a lattice)

(i) Separation between particles clinically monitored using techniques such as electron beam lithography or by electrochemical etching.
(ii) Separation adjusted to match the wavelength of em waves to be dealt. For visible light ~ 500 nm.

(iii) Creation of a line imperfection produces a localized energy level in the forbidden gap, often referred to as 'the guided mode'. It acts as a pipe or waveguide that confines the em waves of a certain frequency and allows them to move along one direction. The confinement is complete.

Common applications:

(i) Resonant cavity, filters and couplers in lasers
(ii) Lossless waveguide for light

Projected use: in telecommunication (possible to have crystals working at all wavelengths).

4. Porous Silicon (produced by electrochemical etching technique):

Features:

(i) Displays visible photoluminescence (ordinary silicon emits in the infrared region). Also electroluminescence and cathodoluminescence displayed.
(ii) High quantum efficiencies (1–10%), contrasting low values (~10^{-4}%) for ordinary silicon.
(iii) Light emitting structures fabricated, using a relatively simple and inexpensive etching technique.

Luminescence interpreted as a consequence of the presence of quantum structures and surface defect states.

Attractive for photoelectric applications. Porous silicon has raised the vision of replacing metal interconnects by optical interconnects.

5. Physical Techniques of Fabrication

Basic approaches: bottom-up, top-down

(A) Photolithography (hybrid of bottom-up and top-down)

Practised in two approaches:

(i) shadow printing (contact mode and proximity printing)
(ii) projection printing

Shadow printing (Diffraction – Fresnel type)

Limits of resolution

$$2e_{\min} = 3\sqrt{\lambda(w + t/2)}$$

$2e$: grating element

λ : wavelength of exposure radiation

w : gap width

and t : photoresist thickness

Projection printing (Diffraction – Fraunhofer type)

Min. resolution feature

$$R = \frac{C_1 \lambda}{NA}$$

and depth of focus

$$\text{DOF} = \frac{C_1 \lambda}{(\text{NA})^2} \text{ with NA} = \bar{n} \sin \alpha$$

C_1, C_2: constants depending on resist material, process technology and image formation method

NA : numerical aperture

(B) Scanning Probe Microscopy (SPM)

Basis of nanomanipulation and nanolithography. Mainly combination of scanning tunnelling microscopy (STM) and atomic force microscopy (AFM).

(i) STM: (applied to conductors)

Evolved on the principle of electron tunnelling, a quantum mechanical phenomenon.

Tunnelling current

$$I \propto e^{-2kz}$$

with z as the thickness of the separating region

and $k = \frac{\sqrt{2m(V-E)}}{h}$; V is the height of the potential barrier and m is the electron mass and E its energy.

STM offers atomic scale resolution in the three-dimensional image: ~0.01 nm in x-y direction and ~0.002 nm in z direction.

(ii) AFM: (applied to dielectric materials)

Features:

(i) Capable of measuring various tip-sample interactions

(ii) Detects the effect of capillary forces for increased tip-sample separation.

(iii) Involves the measurement of the motion of an ultralight cantilever and a nanoscale tip. (A force of 10^{-18} N required for movement through 10^{-4} Å.)

(iv) Sensitive to the vertical component of surface forces.

Difference between STM and AFM:

STM: monitors the tunnelling current between the surface and the probe tip.
AFM: monitors the force between the surface and the probe tip.

QUESTIONS

16.1 What are nanomaterials? Describe briefly different types of nanoparticles and discuss their structures.

16.2 Describe the structure and properties of various nanoparticles.

16.3 Discuss in brief the essentials of jellium model and its success.

16.4 Describe the structure and properties of carbon clusters.

16.5 What are fullerenes? Describe in detail the structure and properties of fullerite. Mention present and projected applications of fullerenes.

16.6 Describe in brief how carbon nanotubes are formed. Discuss their structure, characteristic properties and applications.

16.7 What are quantum nanostructures? Give a brief description of their fabrication.

16.8 Discuss the salient features of quantum nanostructures and describe their potential applications.

16.9 Describe the properties and applications of zeolites.

16.10 What are photonic crystals and how are they formed? Discuss in detail their characteristic properties and significance in device fabrication.

16.11 Discuss salient features of photonic crystals. Give a systematic account of their present and projected applications.

16.12 Describe special features and properties of porous silicon. Discuss critically how it is advantageous to apply porous silicon in devices.

16.13 Describe in brief how porous silicon is produced and discuss the significance of its visible photoluminescence and its impact on technology.

16.14 Discuss the basic principles of photolithography. Give experimental and theoretical bases of the technique.

16.15 Describe the principle and working of a scanning tunnelling microscope. Discuss briefly its application and the resolution achieved.

16.16 Describe the principle and working of an atomic force microscope. Give a comparative discussion on its application with reference to STM.

16.17 Give a critical review of the present and future of nanotechnology.

SUGGESTED FURTHER READING

Nalwa H.S., ed., *Handbook of Nanostructured Materials and Nanotechnology*, Vols. 1–5 (Academic Press, Boston, 2000).

Poole Jr., C.P. and F.J. Owens, *Introduction to Nanotechnology* (Wiley-Interscience, 2003).

Cao G., *Nanostructures & Nanomaterials* (Imperial College Press, 2004).

de Heer W.A., "Physics of Simple Metal Clusters", *Rev. Mod. Phys.* **65**, 611(1993).

Sugano S. and H. Koizumi, *Microcluster Physics* (Springer-Verlag, Heidelberg, 1998).

Dresselhaus M.S., G. Dresselhaus and P.C. Eklund, *Science of Fullerenes and Carbon Nanotubes* (Academic Press, San Diego, 1995).

Ferry D.K. and S.M. Goodnick, *Transport in Nanostructures* (Cambridge Univ. Press, Cambridge, 1997).

Jacak L., P. Howrylak and A. Wojs, *Quantum Dots* (Springer, Berlin, 1998).

Appendix A

van der Waals Interaction

As mentioned in Section 2.3.1, the liquefaction and solidification of noble gases serve as pointer to a certain attractive interaction among neutral atoms or molecules. It is so, especially because the closed electronic configurations of noble gas atoms can be treated as spherical charge distributions that must be disinclined to interact with their neighbours. The attractive interaction, though weak, was introduced as the van der Waals interaction in Section 2.3.1. One of the simplest methods used to calculate this interaction is to follow the perturbation approach of quantum mechanics. To show how this works, we consider the interaction between two widely separated hydrogen atoms. Let $\mathbf{R}$ denote the distance vector between two protons treated as fixed; and let $\mathbf{r}_1$ and $\mathbf{r}_2$ represent the position vectors of the electrons relative to the positions of the protons in the respective atoms, as shown in Fig. A-1. The interaction between the two atoms is the sum of the various Coulomb interactions between the charges on the first atom and those on the second atom:

$$U = Ke^2 \left[\frac{1}{|\mathbf{R}|} + \frac{1}{|\mathbf{R} + \mathbf{r}_2 - \mathbf{r}_1|} - \frac{1}{|\mathbf{R} - \mathbf{r}_2|} - \frac{1}{|\mathbf{R} - \mathbf{r}_1|} \right] \tag{A.1}$$

where $K = \dfrac{1}{4\pi\epsilon_0}$, assuming that the atoms are placed in vacuum.

It is possible to expand the denominators in (A.1) in powers of $\mathbf{r}_1/R$ and $\mathbf{r}_2/R$, if $R >>$ the Bohr radius a_0. In this expansion, the first nonvanishing terms are second-order terms, leading to

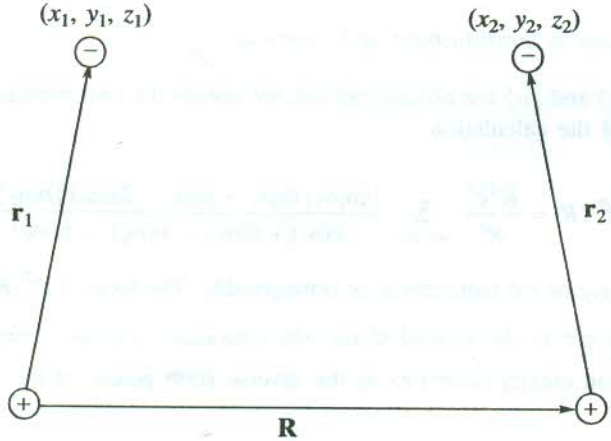


FIG. A-1 Two hydrogen atoms at separation $|\mathbf{R}|$: the protons are shown as (+) and the electrons as (–). $\mathbf{r}_i (i = 1, 2)$ represents the position vector of the electron with respect to the position of the proton in atom i.

$$U = -Ke^2 (\mathbf{r}_1 \cdot \nabla)(\mathbf{r}_2 \cdot \nabla) \frac{1}{R}$$

$$= Ke^2 \left[\frac{\mathbf{r}_1 \cdot \mathbf{r}_2}{R^3} - \frac{3(\mathbf{r}_1 \cdot \mathbf{R})(\mathbf{r}_2 \cdot \mathbf{R})}{R^5} \right] \tag{A.2}$$

The above result assumes importance since its form reveals that for large separations R, the interaction between the two atoms is simply represented by the interaction between the two electric dipoles of dipole moments $e\mathbf{r}_1$ and $e\mathbf{r}_2$. If $\mathbf{R}$ happens to be directed along the z-axis,

$$U = \frac{Ke^2}{R^3} (x_1x_2 + y_1y_2 - 2z_1z_2) \tag{A.3}$$

Let us consider a situation when the atoms are in states $|n_1\rangle$ and $|n_2\rangle$, and no interaction between them is effective with the result that the unperturbed energy of the atoms is $[\varepsilon(n_1) + \varepsilon(n_2)]$. The total energy of the atoms when allowed to interact to first order in the perturbation approach is expressed as

$$\varepsilon(R) = \varepsilon(n_1) + \varepsilon(n_2) + \langle n_1n_2|U(R)|n_1n_2\rangle \tag{A.4}$$

Using (A.3), the diagonal matrix element of $U(R)$ is written as

$$\langle n_1n_2|U(R)|n_1n_2\rangle = \frac{Ke^2}{R^3} [\langle n_1|x_1|n_1\rangle \langle n_2|x_2|n_2\rangle + \langle n_1|y_1|n_1\rangle \langle n_2|y_2|n_2\rangle - 2\langle n_1|z_1|n_1\rangle \langle n_2|z_2|n_2\rangle] \tag{A.5}$$

in view of the fact that the wavefunction of the state $|n_1n_2\rangle$ is a simple product of the states $|n_1>$ and $|n_2\rangle$.

In case either of the atoms is in the ground state, (A.5) contributes only in the second-order as the first-order correction vanishes from the reflection symmetry of the ground state wavefunction. But if both the atoms are in degenerate states, the necessary diagonalization process for $U(R)$ shows that the first-order correction is nonvanishing and varies as $\frac{1}{R^3}$.

Assuming that $|n_1\rangle$ and $|n_2\rangle$ are nondegenerate, we obtain the first nonvanishing correction only in the second-order of the calculation:

$$\varepsilon^{(2)}(R) = \frac{K^2e^4}{R^6} \sum_{m_1, m_2} \frac{|\langle n_1n_2|(x_1x_2 + y_1y_2 - 2z_1z_2)|n_1n_2\rangle|^2}{\varepsilon(n_1) + \varepsilon(n_2) - \varepsilon(m_1) - \varepsilon(m_2)} \tag{A.6}$$

with $m_1 \neq n_1$ and $m_2 \neq n_2$, on the requirement of orthogonality. The form of $\varepsilon^{(2)}(R)$ has two vital features:

(i) If both sums are in the ground states, the interaction energy is negative.

(ii) The interaction energy decreases as the inverse sixth power of the separation between the atoms.

We use the substitution

$$\frac{e^2}{a_0^5} \sum_{m_1, m_2} \frac{|\langle n_1 n_2 | (x_1 x_2 + y_1 y_2 - 2 z_1 z_2) | n_1 n_2 \rangle|^2}{\varepsilon(n_1) + \varepsilon(n_2) - \varepsilon(m_1) - \varepsilon(m_2)} = \xi \tag{A.7}$$

where ξ turns out to be a dimensionless constant. Detailed calculations for hydrogen atoms in their ground states give $\xi = 6.5$.

Making use of the substitution (A.7) in (A.6), the total energy of the two atoms is written as

$$\varepsilon(R) = \varepsilon(n_1) + \varepsilon(n_2) - \frac{K^2 e^2}{a_0} \left(\frac{a_0}{R}\right)^6 \cdot \xi \tag{A.8}$$

where the third term on RHS shows up as a second-order correction and represents the effective potential energy for interaction between two widely separated hydrogen atoms in their ground states. This effective potential energy is the 'van der Waals interaction'. It is interpreted as a quantum effect since the interaction energy goes to zero as $h \to 0$; ($a_0 = \hbar^2/me^2$).

The occurrence of the van der Waals interaction between two hydrogen atoms sounds misconceived at first sight because the hydrogen atom has no electric dipole moment in its ground state. In fact, the presence of a second atom in the vicinity of an atom marked as the first atom induces a dipole moment in the first atom whose components are proportional to the dipole moment operator of the second atom times R^{-3}. The van der Waals interaction is essentially an interaction ($\sim R^{-3}$) between the induced dipole moment of the first atom and the dipole moment operator of the second atom (and vice versa). It must be clear that it is certainly not the interaction between the induced dipole moment of the first atom and the induced dipole moment of the second atom because that varies as R^{-9}.

Appendix B

Photoconductivity

Illumination of an insulating crystal by photons of energy equal to or greater than the band gap E_g usually excites electrons from the valence band to the conduction band. It makes available the free electrons in the conduction band and the free holes in the valence band, both of which contribute to the electric current under the influence of an external electric field. This phenomenon is called *photoconductivity*.

In addition to the intrinsic process, impurities and imperfections, too, have significant contributions to photoconductivity. We learnt in Section 9.4 that the presence of certain impurities in semiconductors gave rise to the localized energy levels in the forbidden energy gap. Such an energy level is identified as donor or acceptor level depending on whether the level is closed to the conduction band edge or to the valence band edge (see Fig. 9.9). Just as these levels enhance the electrical conductivity of semiconductors, they also contribute appreciably to the photoconductivity. For example, the impurity induced photoconductivity forms the basis of the working of many infrared detectors that are mostly derived from the semiconducting materials. In this case, photoconduction does not necessarily require the excitation through the forbidden gap; instead, low energy photons as may ionize the donor/acceptor atoms cause photoconductivity.

Certain imperfections such as lattice vacancies also create discrete energy levels in the forbidden gap. The carriers of photocurrent are freed from one or other type of the bound state by incident photons. The electrons and holes spend sometime in the conduction band and the valence band, respectively and their flow during this period of time constitutes the photocurrent. They may finally be captured (trapped) at the discrete levels in the energy gap region. These levels are commonly referred to as *traps* (see Fig. B-1). The photocurrent is, however, sustained owing to almost the continuous generation of carriers by a constant photon flux. There is evidence for another type of traps that act as the centres of electron-hole recombination, leading to the restoration of thermal equilibrium. One such trap is labelled as 'Recombination centre' in Fig. B-1. Therefore, the decay of photoconductivity is a natural consequence of switching the illumination off. This is true even for an intrinsic sample since band-to-band process of electron–hole recombination always has a finite probability.

Let us discuss the process of photoconductivity in an extrinsic material for the sake of generality. Suppose that the incident photons add electrons to the conduction band at a rate r_E and other natural processes (e.g. the thermal ionization of donors in semiconductors) do so at a rate r. Let the electron–hole recombination resulting from a variety of energy transformation mechanisms occur at a rate r_C. We may then express the time rate of change of the population of conduction electrons by the following continuity equation:

$$dn/dt = (r - r_C) + r_E \tag{B.1}$$

The RHS of the above equation would have an additional term, if the photoconductor had a

nonuniform distribution of electrons, requiring the excess electrons to move elsewhere. In such a case, dn/dt would have to be replaced with $\partial n/\partial t$ in (B.1) whose present form ignores the spatial dependence.

Denoting the electron density in the equilibrium state (i.e. in the absence of illumination) by n_0 and at any time t by n, we write the excess density of electrons at time t as $n_e \equiv (n - n_0)$. Supposing that this excess density is not very large, we assume that $(r_C - r)$ is proportional to n_e in a crude approximation. The electron lifetime is then conveniently given by

$$\tau_n = \frac{n_e}{(r_C - r)} \tag{B.2}$$

With use of this relation, (B.1) can be written as

$$\frac{dn_e}{dt} = \frac{dn}{dt} = r_E - \frac{n_e}{\tau_n} \tag{B.3}$$

At a steady state $dn/dt = 0$, and hence the excess electron lifetime is simply n_e/r_E.

Similarly, the rate of change of the hole population (in the valence band) may be expressed as

$$\frac{dp_e}{dt} = \frac{dp}{dt} = r'_E - \frac{p_e}{\tau_p} \tag{B.4}$$

The lifetimes of electrons and holes are equal, if their generation and recombination both occur exclusively through band-to-band processes ($r = r'$, $r_C = r'_C$, $n_e = p_e$). The two lifetimes have different values, if only one carrier density deviates from its equilibrium value. For example, this happens when low energy photons excite electrons from the donor states in semiconductors. Also, the electron and hole lifetimes are generally different when the recombination occurs by successive capture of electrons and holes at localized impurity states (shown as traps in Fig. B-1).

Some of the important applications in which the photoconductive effect is applied are television cameras, lightmetres, photographic processes (indirectly), photoconductive cells and infrared detectors. Of these, the development of infrared detectors has attracted maximum attention on account of difficulties faced with the detection of radiation at long wavelengths (far infrared region). A high level of sensitivity has been achieved for photons at long wavelengths by using extrinsic photoconductors. In this context the transient response of n_e to the transient illumination is a matter of prime concern. Hence, if the excess generation caused by photons is some arbitrary function $r_E(t)$ of time, our immediate task is to calculate $n_e(t)$. But the dependence of τ_n on n_e as shown by (B.2) complicates the calculation. The task becomes simpler if the lifetime did not deviate much from its value τ_0 at equilibrium so that τ_n could be treated as a constant ($\tau_n = \tau_0$) in a crude approximation. We shall see below that it enables us to have knowledge about $n_e(t)$ over the full time interval before time t.

The generation of electrons takes place as a sequence of instantaneous events, each of which is represented by a delta function. Let us denote the generative activity at time t_0 by $N\delta(t - t_0)$, where N determines the density of electrons created at time t_0. Then according to equation (B.3), the density of those electrons that survive at time t out of those with density N at time t_0 is

$$n_e(t) = N \exp\left(\frac{t_0 - t}{\tau_0}\right) \tag{B.5}$$

In view of our assumption that the electron lifetime τ_n is constant, equation (B.3) becomes linear.

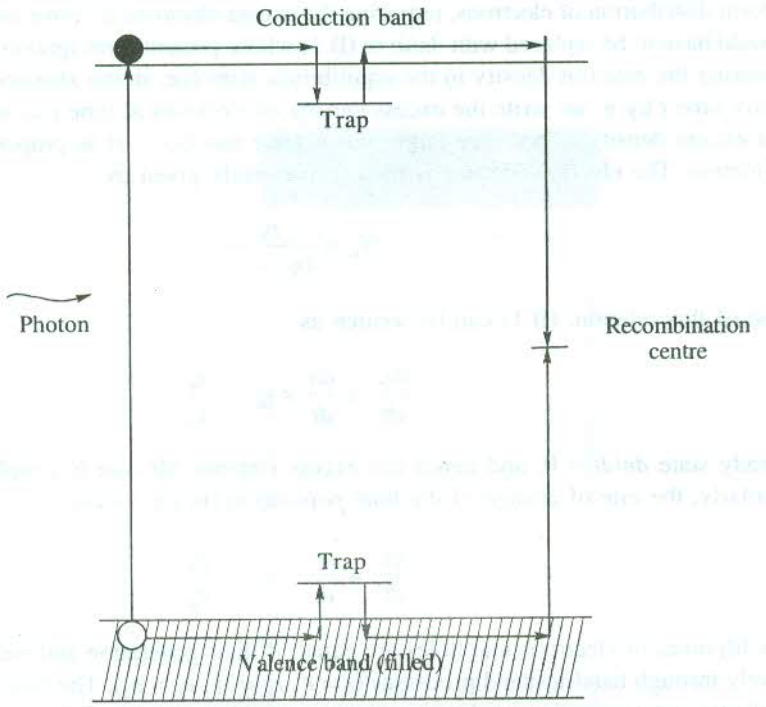


FIG. B-1 The figure shows two different types of traps. One type marked 'trap' affects the freedom of motion of electrons or holes. The other type helps the electron–hole recombination. Such a trap is labelled as 'Recombination centre'.

Therefore, the solution for $n_e(t)$ is simply a sum of the terms of the type in equation (B.5) under the realization that r_E could have been a continuous or discontinuous function of time previously. Hence,

$$n_e(t) = \int_{-\infty}^{t} r_E(t_0) \exp\left(\frac{t_0 - t}{\tau_0}\right) dt_0 \tag{B.6}$$

It is implicit in the validity of the above relation that the lifetime does not depend on n_e. This enables us to find the time dependence of n_e for a generation function which may be an isolated pulse (delta function, sinusoidal, rectangular, etc.) or a series of pulses.

It may be in order to comment briefly on the crucial role of impurity centres or traps in detection at extremes of the infrared region. For extending the long wavelength limit of detection, the detector must be operated at temperatures low enough for the electrons (holes) to be trapped in dark. The operating temperature is reduced as the ionization energy of impurity centres decreases. Te-doped InSb is a typical example with extremely shallow (close to conduction band edge) donor levels ($\sim 10^{-4}$ eV) because of small effective mass ($m^*/m = 0.01$) and large dielectric constant (~ 25).[†] Even at 1 K, the binding of electrons to donors is not noteworthy. The binding energy in this case is enhanced by applying a magnetic field that reduces the electron orbit, constraining it in the plane perpendicular to the field. The orbit size is smaller than even the Bohr radius at high magnetic fields. This method enables an extrinsic photoconductor to be used for detecting photons of wavelengths

[†] Refer to relation (9.25).

as long as 1 mm. For near infrared region (> 0.1 eV), the inconvenience of working at low temperatures is avoided by incorporating deep level impurities. Germanium doped with Cu or Zn is one such example where the state of ionization of Cu or Zn is adjusted by simultaneous doping with a shallow donor (like As).

In all photoconductive devices, a uniform illumination and a free supply of carriers by the electrodes must be ensured. Otherwise, the space charges would reduce the photocurrent severely. The current may be even stopped, if the electric field of the space charges becomes large enough to cancel the electric field produced by the electrodes. Because of this problem, pulse methods are now very often used to study photoconductivity.

Appendix C

Luminescence

The phenomenon of the absorption of energy in matter and its re-emission (mostly in part) as visible or near visible radiation is generally termed *luminescence*. In specific terms the phenomenon is identified as *fluorescence* if the emission occurs during excitation or within 10^{-8} s of excitation. If the emission occurs after 10^{-8} s, it is commonly known as *phosphorescence* or *after glow*. The period of after glow may be between microseconds to hours. The decay time of 10^{-8} s, taken as the demarcation line between fluorescence and phosphorescence, refers to the order of lifetime of an atom in an excited state for which the return to the ground state is accompanied by dipole radiation.

Crystalline luminescent solids are called *phosphors*. Most of the commercially important luminescent materials, however, are used in the form of microcrystalline powders. The ability of phosphors to convert energy into visible radiation has been immensely exploited for application. But the efficiency of conversion in pure materials is poor with a few exceptions. The high efficiency is mostly related to activator atoms that are special impurity atoms introduced intentionally with a small concentration. Examples of phosphors broadly identified as belonging to two different classes are Cu-activated ZnS and Tl-activated KCl. Anthracene activated with naphthacene is a typical example of the organic phosphors.

Some of the prominent applications of luminescent solids are: fluorescent lamps, CRO, radar, television presentation, nucleon and radiation detectors. Light emitting diodes (LEDs) and solid state injection lasers are examples of more sophisticated devices.

Luminescence in solids is excited in a variety of ways. Based on these, the luminescence may be categorized into the following types:

Photoluminescence. Luminescence excited by the absorption of photons is called *photoluminescence*. The absorption of a photon in the valence band leads to the generation of electrons and holes, photoconductivity and optical emission by a process characteristic of the system under study. Photoconductivity must not be taken as a necessary adjunct to luminescence. The luminescence may be entirely within the luminescence centre.

Cathodoluminescence. Electron-beam excited luminescence is known as *cathodoluminescence*. One of its common applications is in the television tube where the primary electrons of about 15 keV energy strike the luminescent screen and generate a shower of electron-hole pairs. Luminescence occurs by the usual recombination processes. This type of luminescence is far less efficient as compared to photoluminescence.

Electroluminescence. It involves the excitation by electric fields. The most common applications of electroluminescence are luminescent panels (e.g. activated ZnS) as traffic lights, LEDs and injection lasers (see Section 9.10.2).

Thermoluminescence. In this case the excitation is accomplished by heating. The heating keeps on

pumping the electrons stored in a metastable state to the conduction band or an excited state of the activator from where they make a radiative transition. It is shown as a second possibility of the radiative emission in Fig. C-1. One fascinating application of thermoluminescence is in personal radiation dosemetres which are heated periodically to determine the total radiation dose received since previous heating.

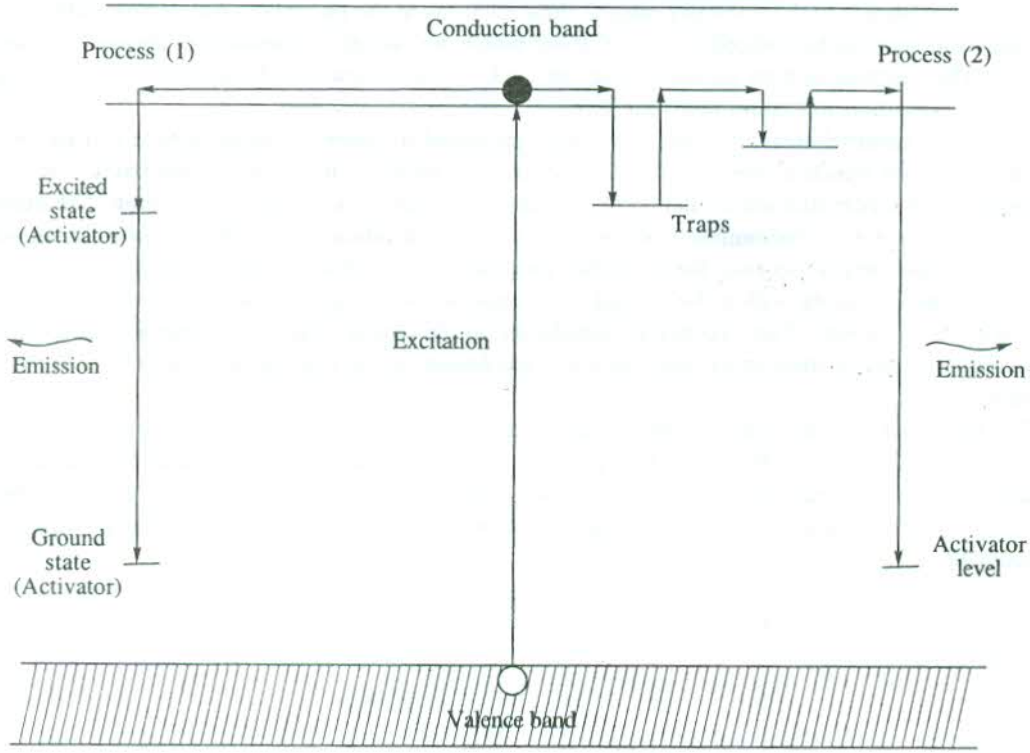


FIG. C-1 Two simple processes of an activated radiative emission. In process (2), electrons at the traps are thermally excited to the conduction band.

Chemiluminescence. It occurs in certain exothermic chemical reactions where the energy is released mostly in the form of the visible radiation.

Activated Luminescence: A Simple Approach

Commercially applied phosphors generally have traces of some impurity or the other to achieve a desired level of efficiency. Controlled amounts of impurities identified as activators are introduced into phosphors to create localized energy levels in the forbidden energy gap of the host phosphor. These localized levels may be of two types: (i) levels of activator atoms; (ii) perturbed energy levels of the host lattice with activators as the source of perturbation. The levels in the latter class may also include those associated with lattice defects such as vacancies. The condition of charge neutrality demands the creation of these vacancies when the oxidation state of an activator atom is higher than that of the atom it substitutes in the host lattice (e.g. Mn^{4+} replacing Zn^{2+} in ZnS). The levels attributed to lattice defects are referred to as *traps*. We will see below that the role of the localized levels is very

vital to the processes of luminescence in real systems. Of these, only those activator levels are crucial which the electrons find easy to enter and leave.

Figure C-1 shows two simple possibilities for a radiative emission. In the first process, an electron in the conduction band (excited from the valence band) drops to an excited activator level and produces radiation by finally descending to the activator ground state. This process is frequent only when the free carrier density is high. The other process involves a trap (located in the forbidden gap), owing its existence to the naturally present imperfections in the phosphor. An electron falls from the conduction band to be trapped at a level from which no radiative transition is allowed (metastable state). The electron is then thermally excited back to the conduction band from where it makes a radiative transition to certain activator level.

Several comprehensive models have been proposed to interpret luminescence in real systems. The process on which a model is based is generally typical of the system under study. As it is not possible to describe all these models here, we restrict ourselves to a general discussion. The treatment shall, however, remain incomplete without mention of an excitation process that involves the movement of an exciton simply because the excitonic processes are central to many optical phenomena. An exciton that is mobile within the crystal may transfer its energy to a luminescence centre that is subsequently excited. The process is significant in the sense that it demonstrates a mechanism whereby the energy from an external source of excitation can be transferred to activators via the host lattice.

The colour and intensity of luminescence are characteristic of the activators used, as is evident from Fig. C-2 for ZnS phosphor. This property has been exploited to increase the 'whiteness' of fluorescence in fluorescent lamps by using two kinds of activators, one for blue and the other for yellow–orange emission. With the overlap of the two broad emission bands, a nearly white light is produced.

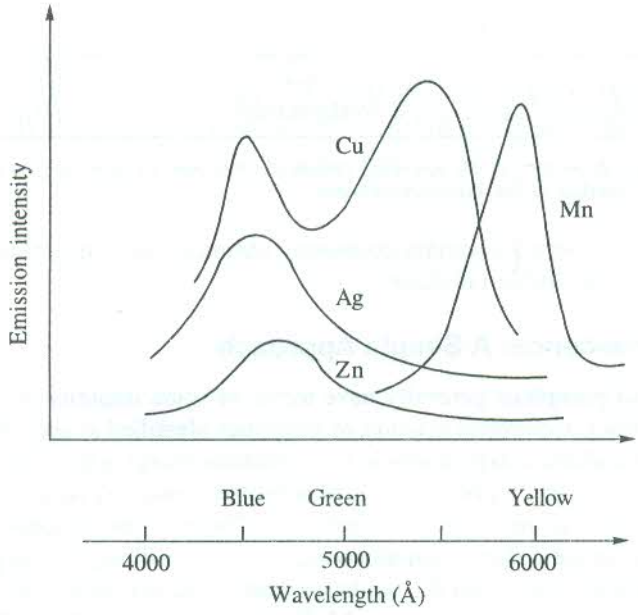
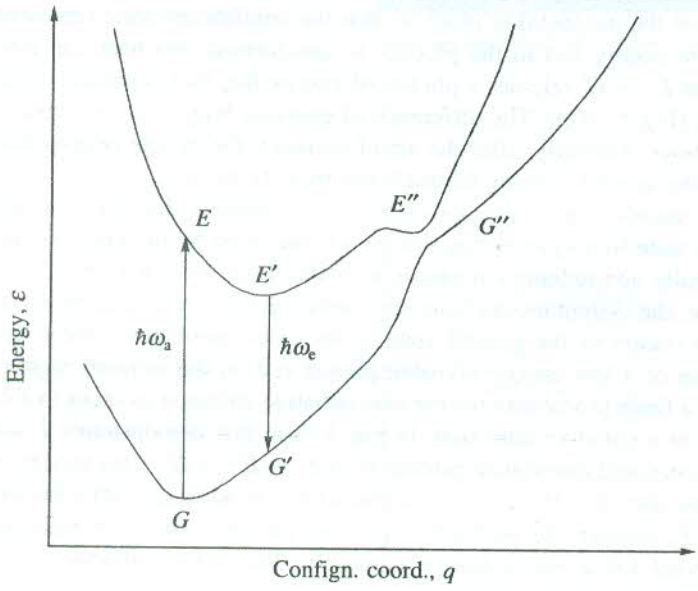


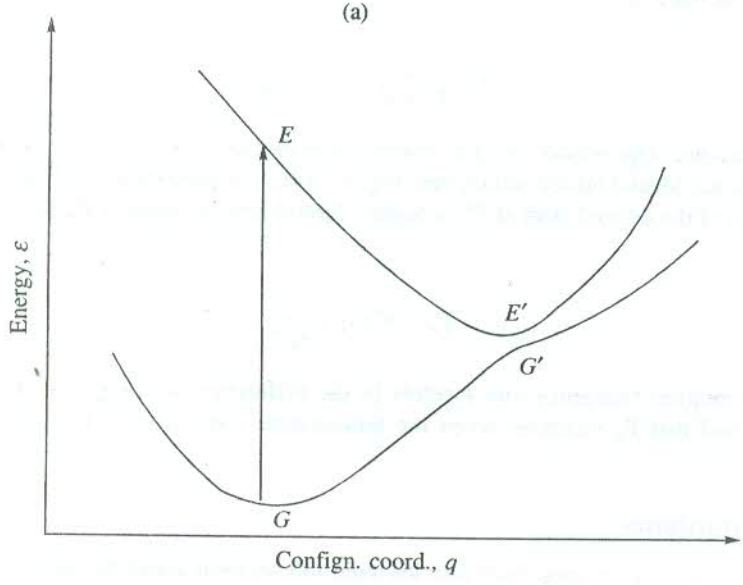
FIG. C-2 The luminescence of ZnS phosphors activated by Cu, Ag, Mn and excess Zn. The source of excitation is ultraviolet light. [After F.A. Kröger, *Proc. IRE*, (Solid State issue), p. 1941 (Dec., 1955).]

Radiative and Non-radiative Emissions

As stated in the beginning itself, the energy of the emitted radiation in luminescence is usually less than the energy absorbed. We use Fig. C-3 to show this. In the figure, the ground state and an excited state of an activator in a phosphor are plotted as functions of the configurational coordinate q whose each value corresponds to a certain configuration of the nuclei around the activator. According to



(a)



(b)

FIG. C-3 Plots of the energy ε in the ground state and in an excited state of an activator versus the configurational coordinate q showing: (a) the radiative emission $E' \rightarrow G'$ and the non-radiative possibility $E'' \rightarrow G''$and (b) purely a non-radiative possibility $E' \rightarrow G'$.

the Frank–Condon principle, the configuration coordinate remains unchanged during a transition. This means that the transitions are vertical, beginning from the minima of the energy curves. The minimum of a curve refers to the configuration at equilibrium in an energy state. With the absorption of energy, the system is excited to the configuration in the excited state. Although the configurational coordinate does not change during the absorption, it does not certainly correspond to the equilibrium in the excited state since E is not at the minimum of the curve. As a result, a small rearrangement in the configuration of the nuclei takes place so that the equilibrium state represented by the point E' is approached. The energy lost in the process is transformed into heat, i.e. into the vibrational energy. The transition $E' \rightarrow G'$ releases a photon of energy $\hbar\omega_e$ that is smaller than the quantum of energy absorbed $\hbar\omega_a$ [Fig. C-3(a)]. The difference of energies $\hbar(\omega_a - \omega_e)$ is consumed in generating phonons, as stated above. Similarly, after the act of emission the system returns finally to the initial equilibrium state in the ground state by thermalizing from G' to G.

Contrary to the situation discussed above, it is also possible that the transition of an excited centre to the ground state in a system may not be accompanied by the emission of a photon. Such a transition is essentially non-radiative in nature, as is illustrated by Fig. C-3(b). After the absorption of energy in this case, the system moves from the configuration E to E' and then a transition through the thin energy gap occurs to the ground state in the non-equilibrium configuration G' with the emission of a phonon or a low energy invisible photon (i.e. in the infrared region).

There is always a finite probability for the non-radiative emission to occur in a configuration that may be most suited to a radiative transition. In Fig. C-3(a) that demonstrates a radiative process, a non-radiative transition could materialize enroute $E \rightarrow E' \rightarrow E'' \rightarrow G''$. This model, proposed by Mott and Gurney[†], implies that the efficiency of luminescence is adversely affected by a non-radiative transition. Suppose P_r denotes the probability per second for a radiative emission and P_n be the probability per second for a non-radiative emission. Then, the efficiency of luminescence is conventionally defined as

$$\eta = \frac{P_r}{P_r + P_n} = \left(1 + \frac{P_n}{P_r}\right)^{-1} \tag{C.1}$$

The temperature dependence of η is mainly determined by P_n, since P_r is nearly temperature independent. In the Mott–Gurney model [see Fig. C.3(a)], the probability of finding an electron in a vibrational level of the excited state at E'' or higher determines the value of P_n. Hence, we may express P_n as

$$P_n = \nu \exp\left(\frac{-\varepsilon}{k_B T}\right) \tag{C.2}$$

where ν is a vibration frequency and ε refers to the difference of energies at E'' and E'.

Thus we find that P_n increases when the temperature rises, leading to a decrease in the value of η.

Decay Mechanisms

Decay characteristics of luminescence provide vital information about the luminescence centres and their energy levels. We discuss below some of the prominent decay mechanisms as are useful in these studies.

[†]N.F. Mott and R.W. Gurney, *Electronic Processes in Ionic Crystals*, 2nd ed., p. 221 (Oxford, 1948).

Temperature-independent exponential decay

Let us suppose that the source of excitation is withdrawn at time $t = 0$ and the number of electrons in an excited state is N_0 at $t = 0$ and $N(t)$ at time t. Also, assume that the probability per second of a radiative transition from an excited state of average lifetime τ is $1/\tau$. If a luminescence centre is shielded well from its surroundings, the lifetime τ does not depend on temperature and on the number of other excited centres. Then, we may write the intensity of luminescence as

$$I(t) = -\frac{dN(t)}{dt} = \frac{N(t)}{\tau} \tag{C.3}$$

Integrating this equation, we get

$$N(t) = N_0 \exp(-t/\tau) \tag{C.4}$$

The substitution of this value of $N(t)$ in (C.3) leads to

$$I(t) = I_0 \exp(-t/\tau) \tag{C.5}$$

where $I_0 = N_0/\tau$, the intensity at $t = 0$.

Temperature-dependent exponential decay

In the luminescence of Tl-activated alkali halides, τ is observed to drop exponentially with increasing temperature though the decay occurs still according to (C.5). Also, the value of τ is of the order of minutes as against 10^{-8} s for a dipole radiation. A model that explains these features of decay involves a metastable state M between the ground state G and an excited state E of an activator as shown in Fig. C.4. In this model the radiative emission (4) starts only when the source of excitation has been removed, i.e. at $t = 0$. During the excitation, i.e. at times prior to $t = 0$, the electrons are collected via

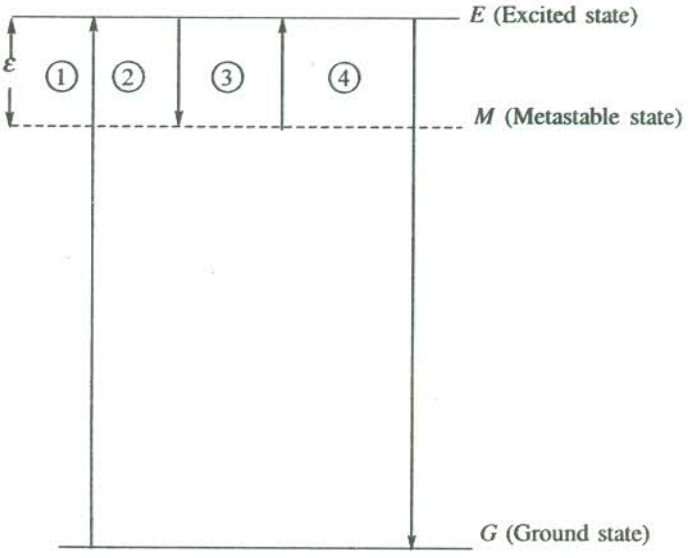


FIG. C-4 Qualitative representation of a delayed emission. The measured lifetime τ of the excited state E is appreciably increased owing to the time spent in storing an electron at the metastable state M and then thermally exciting it to E. The storing is indicated as (2) and the thermal excitation as (3).

(2) at M from E and at $t = 0$ the boiling of electrons [i.e. the thermal excitation (3)] to E begins from where they make the radiative transition to G shown as (4). If ε denotes the energy separation between E and M, then the probability per second for an electron to be excited from M to E is

$$\frac{1}{\tau} = \frac{1}{\tau_0} \exp\left(-\frac{\varepsilon}{k_B T}\right) \tag{C.6}$$

where $1/\tau_0$ is a frequency.

Placing this value of $1/\tau$ in (C.5), we obtain the following form for the decay:

$$I(t) = \frac{N_0}{\tau_0} \exp\left(-\frac{\varepsilon}{k_B T}\right) \exp\left[-\frac{t}{\tau_0} \exp\left(-\frac{\varepsilon}{k_B T}\right)\right] \tag{C.7}$$

This relation shows that the electrons are pumped from the metastable state to the excited state at high rates at higher temperatures. The intensity falls rapidly as the temperature is lowered.

Appendix D

Mössbauer Effect

The resonance phenomenon demonstrated by the emission and absorption of a γ-ray photon without loss of energy owing to recoil of the nucleus and without thermal broadening is known as the Mössbauer effect. This implies that a system in principle would show Mössbauer effect if the nature of atomic binding in the system did not allow a single atom to recoil in the emission and absorption process. Since the thermal broadening turns out to be a function of the recoil energy (to be established in the following section), the recoil-free emission or absorption does not incur thermal broadening. Hence, one of the easiest ways to explain the Mössbauer effect is to demonstrate how the recoil of the nucleus makes the resonance emission and absorption of a γ-ray photon least probable. We approach this exercise by giving below a treatment of nuclear recoil and average thermal broadening for a simple case of γ-ray emission and absorption.

Recoil and Thermal Energies of a Nucleus

Consider the emission of a γ-ray photon from a moving isolated atom. Such a situation is close to realization in a gaseous medium. We assume that the atom moves along a line that coincides with the directions of the γ-ray emission and the recoil of the atom. This simplified approach, however, does not suffer from lack of generality because the motion in other axes remains unaltered.

Denoting the energy of the nucleus at rest in the ground state by E_o and in the excited state of interest by E_e, we write

$$E_e - E_o = E$$

Let the γ-ray photon be emitted in the x-direction and the velocity of the atom in this direction at the time of the emission be denoted by $\mathbf{u}_x$. If the nucleus recoils in the x-direction with velocity $\mathbf{v}_R$, the energy conservation demands that

$$\underbrace{E + \frac{1}{2} M \mathbf{u}_x^2}_{\text{(before the emission)}} = \underbrace{E_\gamma + \frac{1}{2} M(\mathbf{u}_x + \mathbf{v}_R)^2}_{\text{(after the emission)}} \tag{D.1}$$

where M is the nuclear mass and E_γ is the energy of the γ-ray photon.

It should be noticed that $\mathbf{u}_x$ and $\mathbf{v}_R$ are vectors and hence can be oppositely directed. The use of non-relativistic mechanics is valid here as u_x and v_R are both much smaller than the speed of light c.

From relation (D.1), we obtain the difference between the energy of the nuclear transition and the energy of the γ-ray photon:

$$\delta E = E - E_\gamma = \frac{1}{2} M \mathbf{v}_R^2 + M \mathbf{u}_x \cdot \mathbf{v}_R \tag{D.2}$$

$$= E_R + E_{DB}$$

In the above relation, E_R represents the recoil energy of the nucleus and E_{DB} stands for the linewidth contributed by the Doppler effect. The composition of the second term in (D.2) indicates that the Doppler broadening energy depends on the recoil energy and the translational kinetic energy of the atom. The mean kinetic energy in the present case of one translational degree of freedom is

$$\bar{E}_k = \frac{1}{2} M \bar{u}_x^2 \simeq \frac{1}{2} k_B T$$

Making use of the above relation, we write the mean Doppler broadening energy as

$$\bar{E}_{DB} = M v_R \, (\bar{u}_x^2)^{1/2} = 2\sqrt{\bar{E}_k E_R} \tag{D.3}$$

It is practically useful to express E_R and $\bar{E}_{DB}$ in terms of the energy of the γ-ray photon E_γ. Applying the law of momentum conservation to the event of γ-ray emission.

$$\mathbf{p}_R = - \mathbf{p}_\gamma$$

or

$$p_R = - E_\gamma / c \tag{D.4}$$

where $\mathbf{p}_R$ is the momentum of the recoiling nucleus and $\mathbf{p}_\gamma$ is the momentum of the γ-ray photon.

Using (D.4) we express the recoil energy as

$$E_R = \frac{p_R^2}{2M} = \frac{E_\gamma^2}{2 Mc^2} \tag{D.5}$$

On placing this value of E_R in (D.3), we obtain

$$\bar{E}_{DB} = E_\gamma \sqrt{\frac{2 \bar{E}_k}{Mc^2}} \tag{D.6}$$

Now, it can be shown how the nuclear recoil renders the resonance emission and absorption of γ-rays impossible. In a process of γ-ray emission, the energy of the emitted photon is lower than the nuclear energy level separation E by the recoil energy of the nucleus E_R. Similarly, for the reverse process or the reabsorption of γ-rays to occur the photon energy is required to be equal to $(E + E_R)$. Thus the energy of the absorbed photon differs from the energy of the emitted photon by $2E_R$.

For $E_\gamma = 10^4$ eV and $M = 100$ a.m.u., relations (D.5) and (D.6) respectively give $E_R = 5.4 \times 10^{-4}$ eV and $\bar{E}_{DB} \simeq 5 \times 10^{-3}$ eV at 300 K. In this case the amount of overlap between the emission and absorption profiles, shown in Fig. D-1, turns out to be extremely small and resonance does not occur. In contrast, the magnitudes of E_R and $\bar{E}_{DB}$ for ultraviolet radiation are smaller by several orders. For $M = 100$ a.m.u. and a radiation of energy 6.2 eV (2000 Å), E_R is merely 2.1×10^{-10} eV and $\bar{E}_{DB}$ is of the order of 3×10^{-6} eV at 300 K. These magnitudes lead to a large overlap between the emission and absorption profiles, resulting in a strong resonance emission and absorption.

The above discussion suggests that in the presence of recoil, resonance emission and absorption

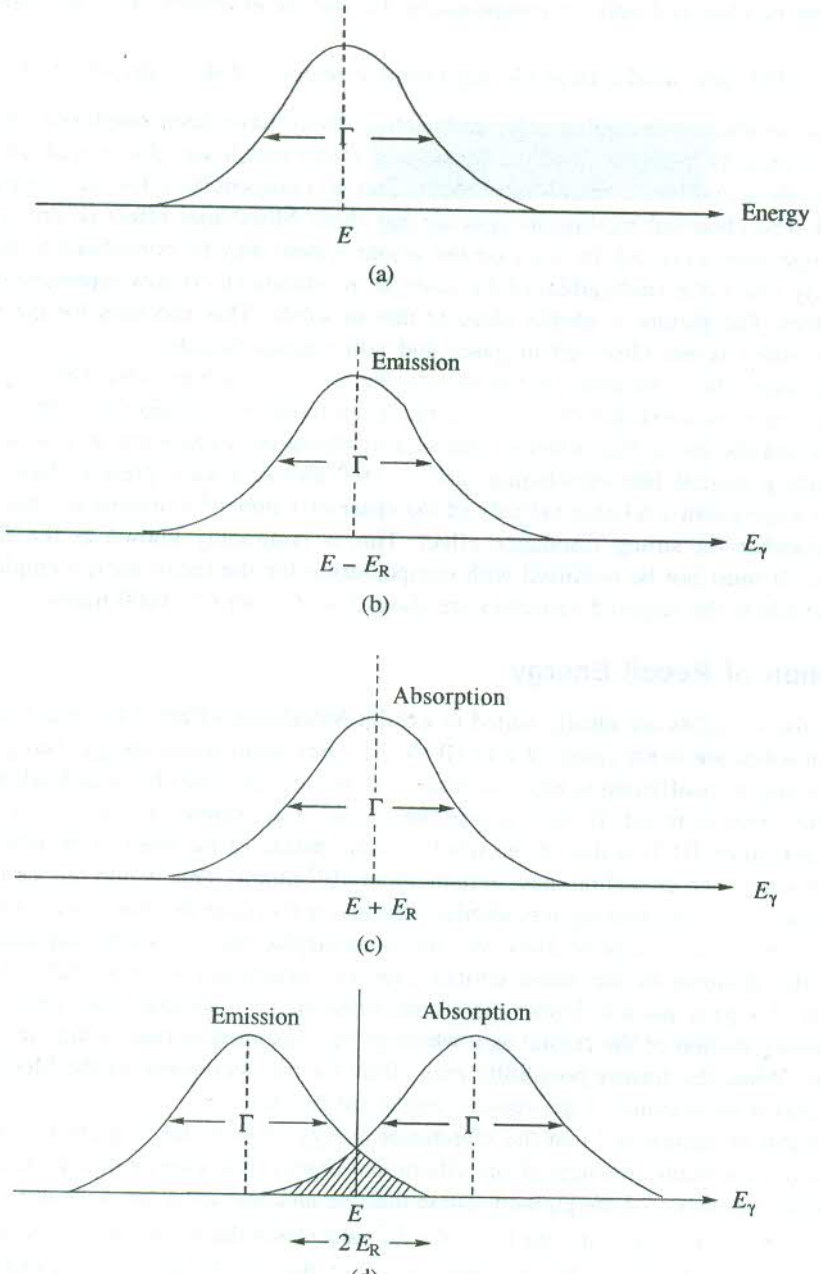


FIG. D-1 (a) Energy distribution of the excited state involved in the emission and absorption of the γ-ray photons. (b) Energy distribution of the γ-ray photons emitted by a free atom. (c) Energy distribution of the γ-ray photons absorbed by a free atom. (d) Overlap of the energy distributions for the emission and absorption of γ-ray photons in a free atom. The overlap region (shaded) is small. This region accounts for the small nuclear gamma resonance.

Symbols:
E is the transition energy; E_γ is the energy of the γ-ray photon; Γ denotes the linewidth, measured at half the height of the maximum.

of γ-rays can be observed only by compensating for the recoil energy. This was first demonstrated by Moon in 1950 who used a large closing Doppler velocity of the order of $2v_R \left(= \frac{2E_\gamma}{Mc^2}\right)$. Since then several techniques including prior radioactive decay have been employed. But this form of resonance is entirely different from the Mössbauer effect which was discovered later in 1957. The Mössbauer effect involves a recoil-free process and no compensation for the energy disparity $2E_R$ is required. The chemical binding in systems that show Mössbauer effect is strong enough not to allow a single atom to recoil. In this case the whole system may be considered to recoil, leading to an extremely low value (negligible) of E_R since M in relation (D.5) now represents the mass of the whole system. The picture is ideally close to that in solids. This accounts for the reason why the Mössbauer effect is not observed in gases and non-viscous liquids.

At this stage a few comments on line broadening are very much in order. The Doppler broadening in solids is almost non-existent owing to the restricted translatory motion of atoms. It can be further suppressed and the recoil-free activity enhanced by doing the experiment at low temperatures. But the Heisenberg natural line broadening (10^{-9} – 10^{-6} eV) is always present. Hence, in the actual Mössbauer experiment a relative velocity of the order of 1 mm s^{-1} is maintained between the source and the absorber for strong resonance effect. This is commonly known as the *Doppler velocity modulation*. It must not be confused with compensation for the recoil energy employed in Moon's experiment where the required velocities are about 5×10^5 mm s^{-1} (1600 mph).

Elimination of Recoil Energy

As stated above, solids are ideally suited to exhibit Mössbauer effect. Chemical binding and lattice energies in solids are in the range of 1 to 10 eV. The free-atom recoil energy, being small compared to these values, is insufficient to eject an atom from the crystal. The chemical binding may not allow the emitting atom to recoil. In such a case the crystal is considered to recoil as a single mass and the mass in relation (D.5) is that of the whole crystal instead of the mass of the emitting atom. Even in a fine powder, one crystallite may contain about 10^{15} atoms. This would reduce the recoil energy E_R by an identical factor making it negligible. But in a real crystal the atoms are not so rigidly bound; instead an atom is free to vibrate. However, the mean displacement of a vibrating atom about its mean position (the position in the static lattice) averages practically to zero during the emission or absorption of a γ-ray photon. Under these conditions the recoil energy may either be contained in the translatory motion of the crystal as a whole or be consumed in raising the vibrational energy of the crystal. While the former possibility may allow for the occurrence of the Mössbauer effect, the latter forbids it on account of the reasons explained below.

We learnt in Section 4.7 that the vibrational energy of a crystal is quantized and it can change only in units of certain quantum of the vibrational energy (the energy of a phonon, $\hbar\omega$), where ω is the angular frequency of the phonon. These discrete amounts are $\pm\hbar\omega, \pm 2\hbar\omega, \pm 3\hbar\omega, \ldots$. For the transfer of the recoil energy into this form, the E_R must match these amounts. It implies that for values of E_R in between 0 and $\hbar\omega$ or $\hbar\omega$ and $2\hbar\omega$, and so on, the transfer will not take place. For the lowest vibrational excitation in which only one phonon is present, E_R must match $\hbar\omega$. But if $E_R < \hbar\omega$, the Mössbauer effect may occur, materializing through a zero-phonon transition.

Recoil-free Fraction

It follows from the above discussion that the probability of the occurrence of the Mössbauer effect

is determined by the extent to which the use of the recoil energy for creating phonons is inhibited. In other words, the fraction of γ-photons that are emitted without exciting phonons must be large for an intense Mössbauer effect. This fraction is commonly known as the *recoil-free fraction* and denoted by f. Accordingly, the emission of a fraction $(1 - f)$ of γ-photons would be accompanied by the excitation of phonons. The simplest vibrating solid is described by the Einstein model in which all atoms are believed to vibrate independently with a common frequency ω_E. If it is assumed that with the emission of a single γ-photon only one phonon is created, in a crude approximation we can write

$$E_R = (1 - f)\, \hbar\omega_E$$

or

$$f = 1 - \frac{E_R}{\hbar\omega_E} \tag{D.7}$$

An emission or absorption line involving purely a recoil-free process is identified as a zero-phonon line. It is imperative to approach a quantitative treatment of the process of emission or absorption in totality for evaluating the recoil-free fraction. We refer back to the case of the γ-ray emission in one dimension. Consider a general event in which with the emission of a γ-photon the crystal containing the emitting nucleus simultaneously undergoes a change in its vibrational state. If the interaction that materializes this event is represented by the Hamiltonian operator H_{int}, the probability P of the event varies according to the following proportionality.

$$P \propto |\langle \psi_f | H_{int} | \psi_i \rangle|^2 \tag{D.8}$$

where ψ_i and ψ_f are wavefunctions describing the initial and final quantum states of the crystal.

The interaction Hamiltonian H_{int} depends on the position of the emitting nucleus, and the momenta and spins of the particles inside the nucleus. The entire process involves the forces within and outside the emitting nucleus. The forces within the nucleus (nuclear forces) that hold the nucleons together are short-range forces in contrast to the long-range nature of crystal binding forces. Therefore, the processes of nuclear decay and the change of the vibrational states of crystal must go independently. Accordingly, the matrix element appearing in (D.8) can be split into two parts. The nuclear part being characteristic of the properties of the nucleus concerned can be treated as a constant. This enables us to represent the matrix element by a single term. The form of H_{int} in the present case leads to

$$P = \text{const.} \times |\langle \psi_f(v) | \exp(i\mathbf{k} \cdot \mathbf{x}) | \psi_i(v) \rangle|^2 \tag{D.9}$$

where $\mathbf{k}$ is the wavevector of the emitted γ-ray photon and $\mathbf{x}$ is the position vector of the centre of the emitting nucleus. The probability for the recoil-free emission f is obtained by replacing $\psi_f(v)$ in (D.9) with $\psi_i(v)$, since the change of the vibrational state is not allowed in a zero-phonon transition:

$$f = \text{const.} \times |\langle \psi_i(v) | \exp(i\mathbf{k} \cdot \mathbf{x}) | \psi_i(v) \rangle|^2 \tag{D.10}$$

Therefore, when the vibrational wavefunction $\psi_i(v)$ is normalized, the probability of the recoil-free fraction is written as

$$f = \exp(-\mathbf{k}^2 \cdot \mathbf{x}^2) \tag{D.11}$$

The vector $\mathbf{x}$ in this relation denotes the random displacement of the emitting atom. Hence, it is appropriate to replace $\mathbf{x}^2$ by $\langle x^2 \rangle$, representing the component of the mean square displacement in the direction of the γ-ray emission. Using

$$k^2 = \frac{4\pi^2}{\lambda^2} = \left(\frac{E_\gamma}{\hbar c}\right)^2$$

relation (D.11) is rewritten as

$$f = \exp\left[-\left(\frac{E_\gamma}{\hbar c}\right)^2 \langle x^2 \rangle\right] \tag{D.12}$$

It follows from this relation that the recoil-free fraction falls exponentially with increasing γ-ray energy and increasing mean square displacement. This puts an upper limit on the values of E_γ beyond which the Mössbauer effect may not be measurable. The highest γ-ray energy for which the Mössbauer effect has been observed is 155 keV, emitted by ^{188}Os. Similarly, the mean square displacement is too required to be restricted so that the Mössbauer effect might remain detectable in the system of interest. This restriction is most effective in solids of strong chemical binding. On the other hand, the magnitude of $\langle x^2 \rangle$ in gases and non-viscous liquids is so large that the Mössbauer effect becomes immeasurably weak. Relation (D.12) demands that for a detectable Mössbauer effect, $\langle x^2 \rangle$ should be small in comparison with the square of the wavelength λ^2 of the γ-ray photon. Practically, the magnitude is reduced by lowering the temperature (i.e. the experiment is performed at low temperatures). The study of the temperature dependence of recoil-free fraction offers itself as one of the accurate methods of probing lattice vibrations.

The precise nature of the temperature dependence of recoil-free fraction is obtained from an appropriate model of vibrational modes of crystals. Adopting the Einstein model for the evaluation of the matrix element in relation (D.10), we obtain

$$f = \exp(-E_R/\hbar\omega_E) = \exp(-E_R/k_B\theta_E) \tag{D.13}$$

where $\hbar\omega_E = k_B\theta_E$; θ_E is called the Einstein characteristic temperature.

It may be observed that for $E_R << \hbar\omega_E$, the above relation reduces to (D.7), obtained by intution. The Einstein model, though simple, is far from being realistic. The Debye model, another important model, describes the vibrations of real crystals more truthfully. Debye treated the crystalline medium as an elastic continuum and argued that atoms do not vibrate with a single common frequency. He proposed that the atoms in a crystal oscillate with frequencies that range from zero to a maximum ω_D (the Debye cut-off frequency). For details on the Einstein and Debye models, the reader is referred back to Section 5.2. The average frequency of vibration $\bar{\omega}$ in the Debye spectrum is given by

$$\hbar\bar{\omega} = \frac{3}{4}\hbar\omega_D$$

The expression for the recoil-free fraction as derived in the framework of the Debye model is

$$f = \exp\left[\frac{-6E_R}{k_B\theta_D}\left(\frac{1}{4} + \left(\frac{T}{\theta_D}\right)^2 \int_0^{\theta_D/T} \frac{x\,dx}{e^x - 1}\right)\right] \tag{D.14}$$

where θ_D is the Debye characteristic temperature. It is also expressed as

$$f = \exp(-2W) \tag{D.15}$$

where W is termed the 'Lamb–Mössbauer factor'. It is often referred to as the 'Debye–Waller factor'. The Debye–Waller factor figures in the theory of Bragg diffraction involving the elastic scattering of x-rays. Thus the Bragg diffraction occurs essentially through a recoil-free process. It is in this spirit that the Bragg diffraction is viewed as the Mössbauer effect for x-rays. The difference between the two forms of the Mössbauer effect lies in the relative time involved in x-ray scattering and emission or absorption of γ-rays. Compared to the characteristic time for lattice vibrations the x-ray scattering is fast whereas the lifetime of a Mössbauer nucleus is long.

It is useful to evaluate the expression for f (D.14) in the two extreme limits of temperature. The forms of the expression in these limits are found as under:

$$f = \exp\left[\frac{-E_R}{k_B \theta_D}\left\{\frac{3}{2} + \frac{\pi^2 T^2}{\theta_D^2}\right\}\right]; \qquad \text{for } T << \theta_D \tag{D.16}$$

(at low temperature)

and

$$f = \exp\left[\frac{-6\,E_R\,T}{k_B \theta_D^2}\right]; \qquad \text{for } T >> \theta_D \tag{D.17}$$

(at high temperature)

A universal feature of these expressions is that the recoil-free fraction increases with the lowering of temperature. The effect is established by experiments. The Mössbauer effect was discovered[†] while the scattering of 129 keV γ-rays from ^{191}Ir was being measured. In this case E_R = 0.05 eV and $E_{DB} \simeq$ 0.1 eV at room temperature. The emission and absorption lines thus have a considerable overlap and the resonance scattering can be observed at room temperature. Mössbauer cooled both the source and the absorber only to be surprised by the enhanced resonance scattering, an observation not expected on classical considerations. The Mössbauer's observation cannot be regarded as a mere accident. Instead, it is an exhibition of rare genius involved in planning the experiment and explaining the effect. Mössbauer was awarded the noble prize for this invaluable contribution in 1961.

Applications

The most useful feature of the Mössbauer effect is the sharpness of the γ-ray emission. The emission gives a typical Lorentzian line with a width of the order of 10^{-8} eV which is consistent with the measure of the natural line broadening predicted by the uncertainty principle. This is understandable because the width is purely due to the width of the nuclear levels (lifetime ~10^{-7} s) involved and no excitation of lattice vibrations (phonons) occurs in the Mössbauer effect. The effective width in the presence of lattice excitation usually conforms to the order of the phonon energies and hence about six orders of magnitude larger.

The width of a spectral line is alternatively expressed in terms of the fractional linewidth obtained by finding the ratio of the linewidth to the energy of the transition. For a γ-emission of energy 100 keV and linewidth 10^{-8} eV in a Mössbauer experiment, the fractional linewidth is 10^{-13}. This implies that the energy of the γ-rays emitted by a Mössbauer nucleus is measurable to an accuracy of 1 part in 10^{13}. Such a precision in measurement is not achievable even with lasers used as common sources

[†] R.L. Mössbauer, *Z. Physik*, **151**, 124 (1958).

of visible and infrared radiations. Thus γ-rays acquire the status of the most stable and accurately defined electromagnetic radiation for use in experiments.

The discovery of the Mössbauer effect has proved to be revolutionary, in particular because of making it possible to use the transition between nuclear levels for studying the magnetic dipole and electric quadrupole interactions (or the hyperfine interactions) of nuclei with surrounding electrons. This is feasible due to the fact that the width of the γ-ray emission lines is smaller than even the characteristic energies of hyperfine interactions. Some of the important applications of the Mössbauer effect are described below.

Study of hyperfine interactions

There are three components of this interaction—the isomer shift (or electric monopole splitting), the nuclear electric quadrupole splitting and the nuclear magnetic dipole splitting. We obtain the charge density of atomic electrons at the nucleus from the measure of the isomer shift. This furnishes useful information about the chemical binding. The gradient of the electric field at the nucleus, created by other ions in the crystal lattice, determines the electric quadrupole coupling. Hence, the quadrupole splitting reveals the symmetry around the nucleus of interest. With no evidence of this splitting the site symmetry of the Mössbauer nucleus is usually taken as cubic or near cubic. The magnitude of splitting indicates the degree of distortion in the cubic symmetry. The measurement of the magnetic dipole splitting is used for studying the properties of magnetically ordered solids. It is an accurate method of determining the Curie and Néel temperatures.

Detection of impurities and short-lived charge states

It is a chemical application having no reference to hyperfine interactions. There is a distinct advantage in using the Mössbauer technique for this purpose on the ground that it provides a non-destructive analysis from which information about the valency and sometimes about the chemical composition of impurities (present in small concentrations) is obtained. It can also be used for detecting elements that have a Mössbauer isotope. But this is possible only for a lattice of atoms of low atomic numbers where the interference of the non-resonant absorption with the transmission of the resonant γ-rays may be ruled out. A slight modification in the technique enables us to study the chemical properties of radioactive atoms present as impurity in a crystal.

Measurement of red shifts

The Mössbauer effect provides the most stable electromagnetic radiation that is measurable most accurately. This property of γ-rays is exploited for the accurate determination of the gravitational red shift and the thermal[†] red shift.

[†] The thermal red shift is a measure of the lowering in the energy of an emitted γ-ray photon caused by heating.

Appendix E

Magnetoconductivity

Magnetoconductivity is the term used to denote the electrical conductivity as modified by the application of a steady magnetic field. On grounds of simplicity, we discuss the case of a metal whose conductivity is interpreted in terms of the nearly free electron motion and examine the same in the classical relaxation time approximation.

The Lorentz force controls the motion of nearly free electrons described by the usual equation of motion:

$$m\frac{d^2\mathbf{r}}{dt^2} + \frac{m}{\tau}\frac{d\mathbf{r}}{dt} = -e\left(\mathbf{E} + \frac{d\mathbf{r}}{dt} \times \mathbf{B}\right) \tag{E-1}$$

The second term on LHS represents the damping force that follows from the collisions with an average relaxation time τ. When the magnetic field is impressed along z-direction ($\mathbf{B} = B\hat{\mathbf{z}}$), the components of electron motion along the three Cartesian axes (x, y, z) are given by the following equations:

$$\begin{aligned} m\ddot{\mathbf{x}} + \frac{m}{\tau}\dot{\mathbf{x}} &= -e\mathbf{E}_x - m\omega_c\dot{\mathbf{y}} \\ m\ddot{\mathbf{y}} + \frac{m}{\tau}\dot{\mathbf{y}} &= -e\mathbf{E}_y + m\omega_c\dot{\mathbf{x}} \\ m\ddot{\mathbf{z}} + \frac{m}{\tau}\dot{\mathbf{z}} &= -e\mathbf{E}_z \end{aligned} \tag{E-2}$$

where ω_c stands for the cyclotron frequency $\left(\frac{eB}{m}\right)$.

We begin with the analysis of the motion along z-direction described by the last of the above equations simply because it is the simplest of the three. For the sake of generality, let us consider the application of an oscillatory electric field,

$$\mathbf{E}(t) = \mathbf{E}(\omega)e^{i\omega t} \tag{E-3}$$

assuming a similar time dependence for the electron instantaneous position ($\sim e^{i\omega t}$).

With these substitutions, the z-component of motion is given by

$$m\left(i\omega + \frac{1}{\tau}\right)\dot{\mathbf{z}} = -e\mathbf{E}_z$$

or
$$\dot{\mathbf{z}} = -\frac{e\mathbf{E}_z}{m\left(i\omega + \frac{1}{\tau}\right)} \tag{E-4}$$

It leads to the following expression for the current density,

$$\mathbf{j}_z(t) = -ne\dot{\mathbf{z}} = \frac{ne^2}{m\left(i\omega + \frac{1}{\tau}\right)}\mathbf{E}_z(t)$$

or
$$\mathbf{j}_z(\omega) = \frac{ne^2}{m\left(i\omega + \frac{1}{\tau}\right)}\mathbf{E}_z(\omega) \tag{E-5}$$

(using $\mathbf{j}(t) = \mathbf{j}(\omega)e^{i\omega t}$)

Comparing (E-5) with the statement of Ohm's Law ($\mathbf{j}(\omega) = \sigma\mathbf{E}(\omega)$), we obtain the z-component of the conductivity tensor

$$\sigma_{zz} = \frac{ne^2\tau}{m(1 + i\omega\tau)} = \frac{\sigma_0}{(1 + i\omega\tau)} \tag{E-6}$$

Similarly, by working with the first two equations of the set (E-2) we can calculate the conductivity in the xy-plane. With the electron instantaneous position and the electric field respectively represented as

$$\mathbf{r}_{xy} = \mathbf{x} + i\mathbf{y}$$

and
$$\mathbf{E}_{xy} = \mathbf{E}_x + i\mathbf{E}_y \tag{E-7}$$

the equation of motion in the xy-plane can be written as

$$m\ddot{\mathbf{r}}_{xy} + \frac{m}{\tau}\dot{\mathbf{r}}_{xy} = -e\mathbf{E}_{xy} + i\omega_c\dot{\mathbf{r}}_{xy} \tag{E-8}$$

The above equation gives

$$m\left[i(\omega - \omega_c) + \frac{1}{\tau}\right]\dot{\mathbf{r}}_{xy} = -e\mathbf{E}_{xy} \tag{E-9}$$

using which we obtain the following expression for the current density

$$\mathbf{j}_{xy}(t) = -ne\dot{\mathbf{r}}_{xy} = \frac{ne^2}{m\left[i(\omega - \omega_c) + \frac{1}{\tau}\right]}\mathbf{E}_{xy}(t)$$

or
$$\mathbf{j}_{xy}(\omega) = \frac{\sigma_0}{[1 + i(\omega - \omega_c)\tau]}\mathbf{E}_{xy}(\omega) \tag{E-10}$$

This gives the component of the conductivity tensor in xy-plane in the form

$$\sigma^{xy}(\omega) = \frac{\sigma_0}{1 + i(\omega - \omega_c)\tau}$$

$$= \frac{\sigma_0[1 - i(\omega - \omega_c)\tau]}{1 + (\omega - \omega_c)^2\tau^2} \tag{E-11}$$

The relation (E-11) indicates a likely resonance effect at $\omega = \omega_c$. This is confirmed by the results of measurement of some important properties such as surface absorption, reflectance and other related optical phenomena. The resonance effect is famous as "Cyclotron Resonance'. It may be remarked here that the use of $(e^{-i\omega t})$ in place of $(e^{i\omega t})$ in solution (E-3) is equally valid. It gives resonance for $-\omega = \omega_c$ which refers to the emission process.

The relationship between the current density and the conductivity tensor in three dimensions may be generalized as

$$\begin{bmatrix} \mathbf{j}_x \\ \mathbf{j}_y \\ \mathbf{j}_z \end{bmatrix} = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix} \begin{bmatrix} \mathbf{E}_x \\ \mathbf{E}_y \\ \mathbf{E}_z \end{bmatrix} \tag{E-12}$$

Starting with relations (E-2), one readily establishes the following relation involving the real part of the conductivity tensor:

$$\begin{bmatrix} \mathbf{j}_x \\ \mathbf{j}_y \\ \mathbf{j}_z \end{bmatrix} = \frac{\sigma_0}{1 + (\omega - \omega_c)^2\tau^2} \begin{bmatrix} 1 & (\omega - \omega_c)\tau & 0 \\ (\omega_c - \omega)\tau & 1 & 0 \\ 0 & 0 & \dfrac{1 + (\omega - \omega_c)^2\tau^2}{1 + \omega^2\tau^2} \end{bmatrix} \begin{bmatrix} \mathbf{E}_x \\ \mathbf{E}_y \\ \mathbf{E}_z \end{bmatrix} \tag{E-13}$$

The above relation also demonstrates Hall effect when $\omega = 0$ (i.e. for static electric fields). Considering the components of current in the xy-plane (with $\mathbf{B} = B\hat{\mathbf{z}}$),

$$\begin{bmatrix} \mathbf{j}_x \\ \mathbf{j}_y \end{bmatrix} = \frac{\sigma_0}{1 + (\omega_c\tau)^2} \begin{bmatrix} 1 & -\omega_c\tau \\ \omega_c\tau & 1 \end{bmatrix} \begin{bmatrix} \mathbf{E}_x \\ \mathbf{E}_y \end{bmatrix} \tag{E-14}$$

which gives the magnitudes

$$j_y = \omega_c\tau E_x + E_y = 0 \tag{E-15}$$

(since $j_y = 0$ in the Hall geometry)

and
$$j_x = \frac{\sigma_0}{1 + (\omega_c \tau)^2} (E_x - \omega_c \tau E_y) \tag{E-16}$$

Substituting E_x from (E-15) in (E-16),

$$j_x = -\frac{\sigma_0}{\omega_c \tau} E_y$$

$$= -\frac{ne}{B_z} E_y$$

or
$$-\frac{1}{ne} = \frac{E_y}{j_x B_z} \tag{E-17}$$

(E-17) is the same as (6.89), representing the Hall coefficient R_H.

Index